Phylodynamique

Samuel Alizon

Phylogénies d'infections et phylodynamique

Samuel Alizon

Laboratoire MIVEGEC (UMR CNRS 5290, IRD 224, UM), Montpellier, France
samuel.alizon@cnrs.fr

Résumé

Au cours des deux dernières décennies s'est mise en place une sorte d'émulation mutuelle entre production de données de séquences génétiques microbiennes et analyses phylogénétiques, les unes encourageant le développement des autres. En effet, les phylogénies d'infections permettent de décrire la structure d'une épidémie et sa propagation géographique. En retour, l'originalité de ce type de données, en particulier leur fort ancrage temporel, a ouvert de nouvelles dimensions pour les analyses phylogénétiques autour d'un champ connu sous le nom de phylodynamique.

citation Alizon S (2020) Phylogénies d'infections et phylodynamique. In *Comprendre l'évolution, approches mathématiques et informatiques*. Didier G, Guindon S, ed. ISTE Editions.

1 Réconcilier écologie, évolution et mathématiques

À la base de l'évolution, il y a la dynamique des populations. Darwin l'avait non seulement compris mais en plus illustré dans son livre « L'Origine des espèces ». Pourtant, les approches mathématiques en biologie de l'évolution ont longtemps négligé ces aspects démographiques (aussi appelée « écologiques »). Ainsi, les modèles fondateurs de la génétique des populations se placent le plus souvent sur des échelles de temps très longs ou considèrent des populations de taille infinies. La réconciliation entre dynamique des populations et génétique s'est opérée à partir des années 70 ; d'abord avec la théorie des jeux évolutive et ensuite, dans les années 90, avec la dynamique adaptative. Aujourd'hui, modélisation en écologie et en évolution sont indissociables, comme le détaille la synthèse récente de Lion (2018).

L'étude des maladies infectieuses a eu un rôle central dans la création de ponts entre démographie, évolution et mathématiques. D'une part, comme pour la biologie de l'évolution, la tradition de modélisation mathématique est très ancienne en épidémiologie et des modèles tels que ceux de Kermack & McKendrick (1927) servent encore de base à des études contemporaines. D'autre part, au niveau biologique, les microbes ont généralement de faibles temps de génération, de grandes tailles de populations et des taux de mutation élevés, ce qui conduit à une superposition entre leurs dynamiques évolutives et écologiques. C'est une des illustrations de la boucle de rétroaction entre écologie et évolution. Par exemple, une mutation de résistance à un antibiotique (évolution) modifie la propagation des bactéries qui la portent (écologie). Toutefois, la sélection de cette mutation (évolution) dépend de l'environnement car en l'absence de traitements antibiotiques (écologie), l'antibiorésistance est plus un coût qu'un avantage. Ceci a été particulièrement bien capturé au niveau mathématique par les approches de dynamique adaptative (Dieckmann *et al.* 2002).

La phylodynamique, définie en 2004 par Grenfell *et al.* a renforcé les liens entre évolution et épidémiologie en y apportant une masse croissante de données. Jusque là, la réconciliation entre écologie et évolution s'était avant tout construite sur des données phénotypiques forcément limitées. La phylodynamique repose en premier lieu sur l'évolution neutre, ce qui lui permet d'exploiter facilement beaucoup de données de séquences génétiques. Le succès de ce champ repose d'ailleurs grandement sur un espoir, qui est d'extraire une grande quantité d'information pertinente en santé publique à partir des milliers de données de séquences générées en routine dans les laboratoires.

Dans ce chapitre, nous retracerons brièvement les évolutions techniques, qui ont contribué à l'essor de la phylodynamique. Puis nous présenterons les phylogénies d'infections et leur utilisation initiale, presque purement illustrative, et qui reste largement majoritaire. Ensuite nous définirons la phylodynamique proprement dite ainsi que ses modèles de base. Puis nous présenterons les analyses spatiales ou, plus généralement, en environnement hétérogène que l'on peut réaliser. Enfin, nous parlerons des extensions plus récentes couplant des phénotypes d'infections (par

exemple charge virale) aux données de séquences. Nous conclurons en listant certains des défis pour le champ.

2 Des données et des processeurs

Les analyses phylogénétiques sont possibles depuis longtemps (Felsenstein 1981) mais ont initialement été freinées par la disponibilité de données biologiques, ainsi que par la puissance de calcul mobilisable. Ces limites sont aujourd'hui presque caduques. L'amélioration des processeurs couplée à leur fonctionnement en réseau permet d'analyser des dizaines de milliers de données de séquences simultanément. En parallèle, les progrès des techniques de séquençage ont augmenté qualité, longueur et rapidité de production des données de séquences génétiques.

2.1 De Sanger à Nanopore et Illumina

La médecine a souvent beaucoup progressé au cours des guerres et autres tragédies sanitaires. Il en est un peu de même pour la phylodynamique car la pandémie meurtrière du virus de l'immunodéficience humaine (VIH) a été l'un des déclencheurs de l'utilisation du séquençage dans la routine clinique. En effet, après la découverte des premières molécules antirétrovirales efficaces contre le virus, s'est rapidement posé le problème de l'évolution d'infections résistantes. En liant des données de séquences virales à des résultats de traitements, il a été possible de mettre au point des algorithmes prédisant le risque d'échec thérapeutique pour une molécule donnée en fonction de la séquence virale (Beerenwinkel *et al.* 2003).

Ces utilisations en santé publique ont conduit à la génération de données de séquences génétiques d'une quantité et d'une couverture spatiale inédite. Ce séquençage de routine était au début limité au VIH et ne concernait que quelques portions du génome. Au cours de la dernière décennie, il s'est étendu à d'autres virus virulents, voire même à des infections bactériennes ou des microbiotes entiers. De plus, le pourcentage du génome séquencé a explosé pour aller maintenant vers des séquençages de génomes entiers. Un des exemples les plus frappants a été celui de Quick *et al.* (2016), qui sont parvenus à générer 142 séquences de génomes entiers du virus ébola via la technologie d'Oxford Nanopore avec un matériel biologique et informatique de base (20 kg) dans un pays en pleine crise sanitaire.

Concernant ces techniques de séquençage dites de nouvelle génération, trois d'entre elles se distinguent. La plus populaire est développée par la compagnie Illumina. Elle permet de générer une grande quantité de séquences courtes (200 à 300 paires de bases) pour un coût bas avec un faible taux d'erreur. Son principal inconvénient est que la taille des séquences ne permet pas de savoir si des mutations sont présentes dans le même génome et pas seulement dans la population de génomes. La seconde technique de Pacific Biosciences (PacBio) résout un peut ce problème en faisant du « Single-molecule real-time sequencing ». Les fragments analysés sont longs (potentiellement plus de 30.000 paires de bases) mais les taux d'erreurs plus élevés que pour l'Illumina. Enfin, la technique Nanopore mentionnée ci-dessus est une solution quasi-portable, contrairement aux deux autres qui nécessitent des équipements coûteux, pour générer des fragments très longs avec là aussi un taux d'erreur relativement élevé. Ces deux dernières techniques permettent plus facilement de reconstruire des haplotypes. Plus généralement, les techniques de séquençage de nouvelle génération permettent d'accéder à bas coût à la diversité génétique microbienne.

2.2 PCR et capture

On se focalise souvent sur la puissance des processeurs ou de celle des séquenceurs, mais les techniques de traitements pré-séquençage des échantillons ont aussi progressé, notamment pour gérer la présence très majoritaire d'ADN de l'hôte. Parvenir à enrichir un échantillons en ADN ou ARN microbien est un travail quotidien de nombreux laboratoires. Si la cible microbienne est connue, on peut réaliser des amplification d'ADN par Polymerase Chain Reaction (PCR) mais celles-ci tronquent la diversité génétique et obligent à travailler en qualitatif. De plus, séquencer un génome entier nécessite la mise au point de plusieurs amorces et la réalisation de plusieurs amplifications par PCR.

Les techniques de nouvelle génération permettent de séquencer tout l'ADN présent sans cible particulière, par exemple en utilisant la technique dite « shotgun », qui utilise des cibles d'ADN aléatoires. Sauf dans des cas particuliers (par exemple une charge virale extrêmement élevée dans un milieu avec peu d'ADN humain), il est nécessaire d'enrichir l'échantillon en ADN cible. Ceci peut se faire par de manière mécanique par ultra-centrifigation (pour récupérer les bactéries ou virus plus petits), chimique par des traitement de DNAase (pour détruire l'ADN humain moins bien protégé), voire biotechnologique en utilisant la technique CRISPR-Cas9 pour spécifiquement détruire l'ADN humain.

Depuis quelques années, une autre alternative s'offre, qui permet de générer des données de séquences d'extrêmement bonne qualité : la capture par hybridation. Ainsi, pour moins de 50 euros, on peut aujourd'hui obtenir à partir d'un prélèvement sanguin la diversité virale du VIH tout au long du génome mais aussi de flavivirus et le tout sans PCR (Bonsall *et al.* 2018). L'absence de PCR signifie aussi que la couver-

ture (nombre de fragments par position ou « read ») peut être utilisée pour estimer la charge virale. De plus, la diversité séquencée est proche de la diversité réelle. Ces techniques sont en train d'ouvrir de nouvelles pistes en santé publique pour analyser les réseaux de transmission à l'aide d'approches phylogénétiques (Wymant *et al.* 2018).

3 Phylogénies d'infections

3.1 Lien avec les chaînes de transmission

Si les phylogénies ont envahi la littérature biomédicale, c'est d'abord pour des raisons de classification. Plus précisément, pour déterminer (le plus souvent visuellement) à quelles souches existantes une souche virale (ou bactérienne) infectant un·e patient·e ressemble le plus. D'un point de vue pratique, ceci peut avoir un intérêt si les souches diffèrent en termes de virulence ou de réponse au traitement (ce qui est le cas pour le virus de l'hépatite C). D'un point de vue historique, ces approches de classification en épidémiologie moléculaire ressemblent beaucoup aux premières phylogénies microbiennes basées sur des séquences génétiques issues d'espèces différentes. C'était le cas pour les endosymbiontes (Lake 1988) mais aussi pour le VIH, pour lequel un des premiers travaux phylogénétique porte sur son histoire longue et inclus d'ailleurs des séquences de son équivalent simiesque (Smith *et al.* 1988).

Avec l'épidémie de VIH justement se produit un basculement. Ce virus diffère notablement de la plupart des organismes séquencés jusqu'alors de par sa rapide vitesse évolutive et son cycle de vie épidémiologique (sans oublier la quantité de séquences disponibles). Ceci permet de générer des phylogénies d'infections. Une des premières est issue de données de 9 patients hémophiles infectés par une même source (Balfe *et al.* 1990). Une des plus célèbres porte elle sur une chaîne de transmission liée à un dentiste (Ou *et al.* 1992). Dans les deux cas, les implications éthiques et juridiques des études sont inédites pour des biologistes de l'évolution, que l'on convoque comme experts dans les tribunaux afin de trancher sur des sources de transmission. Ces problématiques sont d'ailleurs toujours d'actualité en phylodynamique (Coltart *et al.* 2018, Johnson & Parker 2019).

Pour caricaturer, le basculement qui s'opère dans le début des années 90 concerne la nature de l'unité des arbres phylogénétiques où l'espèces est remplacées par l'infection. Décrit de cette manière, le changement conceptuel peut paraître énorme. Pourtant, non seulement la plupart des techniques peuvent se transposer directement, mais en plus travailler sur des infections plutôt que des espèces rend au final les mécanismes sous-jacents plus intuitifs. En effet, on peut facilement faire un parallèle entre une phylogénie et une chaîne de transmission. Si chaque individu de la chaîne transmission est présent dans la phylogénie, alors chaque feuille correspond à la fin de la période infectieuse d'un individu et chaque nouvelle infection créé un embranchement dans la phylogénie. À l'inverse, autant les extinctions d'espèces sont évidentes, autant leur apparition est bien plus floue qu'une nouvelle infection (sans parler évidemment de la difficulté à définir une espèce).

3.2 Datation et vitesse d'évolution

Pour les phylogénéticiens, les phylogénies d'infections ont été une source d'inspiration car elles possèdent une différence majeure avec les phylogénies d'espèces. Classiquement, ces dernières sont représentées par des arbres dit « ultramétriques » car toutes les espèces sont contemporaines. Ancrer une telle phylogénie dans le temps est un défi en soi et nécessite d'utiliser des données fossiles pour calibrer les nœuds internes. Les phylogénies d'infections sont elles plus faciles à ancrer dans le temps car des différences de quelques années entre deux dates d'échantillonnage ne sont plus du tout négligeables par rapport à la date de la racine de la phylogénie. Dans le cas des émergences épidémiques, les vitesses d'évolution sont telles qu'il n'est maintenant pas rare d'avoir des phylogénies de bonne qualité dont certaines feuilles sont relativement proches du cas index de l'épidémie, c'est-à-dire la première personne infectée dans la population étudiée (comme pour la récente épidémie de virus ebola en Afrique de l'Ouest décrit par Holmes *et al.* 2016).

Ce fort ancrage temporel a conduit au développement de nouvelles techniques pour estimer les taux de substitution, c'est-à-dire le nombre de mutations fixées dans le génome par unité de temps. La plus simple consiste à déduire ces vitesses d'une régression entre la date d'échantillonnage des feuilles d'une part et leur distance (en nombre de substitutions) à la racine de l'arbre d'autre part (Rambaut 2000, Drummond *et al.* 2003). Pour un taux de substitution donné, on peut alors facilement dater les nœuds internes de l'arbre ainsi que sa racine. Les techniques d'inférence bayésiennes ont contribué à améliorer cette datation. Ils ont aussi permis de lever l'hypothèse, parfois forte, de vitesse d'évolution constante. Ainsi, le modèle d'horloge moléculaire relâchée permet de faire varier les vitesses d'évolution dans les différentes branches de la phylogénie en les tirant dans une distribution (Drummond *et al.* 2006), ce qui évite d'avoir à estimer une valeur pour chaque région de la phylogénie comme dans les modèles d'horloge moléculaire locale (Yoder & Yang 2000). Une des limites de ces méthodes tient à leurs forts besoins de puissance computationnelle. Plus récemment, des approches basées sur des techniques de maximum de vraisemblance ont permis

d'arriver à des qualités de datation comparables de manière plus rapide (To *et al.* 2015).

Une application emblématique de l'ancrage temporel a été la datation de l'origine des VIH dans les populations humaines. Peu de temps après les premiers séquençages, des études phylogénétiques ont situé l'ancêtre commun de l'épidémie avant 1960 (Li *et al.* 1988). Cette estimation s'est affinée avec l'isolation du virus dans des échantillons issus de patients infectés avant 1970 (Jonassen *et al.* 1997, Korber *et al.* 2000, Worobey *et al.* 2008). Au cours des 10 dernières années, les techniques d'inférence ont progressé et, comme on le verra, on peut maintenant estimer à la fois la date et l'origine géographique. Pour le groupe M du VIH, qui est celui qui a fait le plus de victimes, elles se situent dans les années 20 dans l'actuelle République Démocratique du Congo (Faria *et al.* 2014). Cette datation n'est pas le propre des infections virales. Par exemple, la découverte d'ADN bactérien dans des dents humaines datant de 4.900 ans a permis à Rascovan *et al.* (2019) de revoir l'histoire de la propagation des épidémies de peste en Europe depuis le Néolithique.

Si la puissance de calcul et les techniques d'inférence importent, la découverte d'échantillons anciens reste un moteur des études de datation. Toutefois, la biologie des virus contraint l'inférence phylogénétique. Pour des virus à ADN double brin évoluant à la vitesse de leur hôte Vertébré, on peut remonter des centaines de milliers d'années en arrière, le plus souvent en utilisant des calibrations internes de l'arbre à partir de l'évolution de l'hôte, comme dans le cas d'HPV16 (Pimenoff *et al.* 2016). Pour des virus évoluant plus rapidement en revanche, le nombre de substitutions noie le signal phylogénétique sur le temps long mais, en revanche, il permet des inférences sur des échelles de temps courtes.

3.3 Applications biologiques de la calibration temporelle

Dater des événements ayant eu lieu au cours d'une épidémie a des implications évidentes en santé publique. Par exemple, en Afrique de l'Ouest en 2014-2016, le fait de pouvoir rapidement dater le passage de l'épidémie d'ébola en Sierra-Leone depuis la Guinée a contribué à comprendre l'ampleur de l'épidémie (Gire *et al.* 2014). Au niveau clinique, en combinant des données de séquences de VIH d'une mère et de son enfant infectés, il est possible de déterminer si la transmission a eu lieu *in utero* ou lors de la naissance (Chaillon *et al.* 2014).

Les vitesses d'évolution sont aussi informatives en tant que telles. Dans le cas du VIH, cette vitesse est plus rapide au niveau intra-hôte, qu'au niveau inter-hôte (Pybus & Rambaut 2009). En analysant le génome entier, nous avons montré que cette différence d'un ordre de grandeur en termes de taux de substitution se retrouve tout au long du génome (Alizon & Fraser 2013). L'explication la plus parcimonieuse est que les virus les plus fréquents dans le sang sont en fait dans un cul-de-sac évolutif tandis que ceux stockés dans des cellules de manière plus latente conservent leur capacité à générer une nouvelle infection. Ainsi, les virus « moins évolués » sont plus transmis à la génération suivante.

Ces problématiques s'appliquent aussi pour des virus évoluant moins rapidement comme le virus varicelle-zona, qui est à ADN double brin (dsDNA) et pour lequel les données de séquences ne contiennent *a priori* pas suffisamment de signal pour estimer un taux de substitution fiable. Toutefois, en étudiant des souches vaccinales atténuées, Weinert *et al.* (2015) sont parvenus à estimer ce taux de substitution. La différence entre cette souche et la souche naturelle est que la dernière est connue pour avoir des périodes de latence, pendant lesquelles le virus reste dans la cellule hôte sans se multiplier. Ceci a mis en évidence l'importance de la prise en compte de la latence pour interpréter et dater l'épidémie de ce virus.

4 Phylodynamique

4.1 Champ en quête de définition

Le terme de « phylodynamique » a fait son apparition dans une « Perspective » de Grenfell *et al.* (2004). Leur définition du champ est pour le moins inclusive puisque, selon eux, la phylodynamique a pour but d'expliquer « *la grande diversité de phylogénies de pathogènes observées à des échelles allant d'un individu hôte jusqu'à la population* » et elle se définit comme un « *alliage d'immuno-dynamique, d'épidémiologie et de biologie de l'évolution* ». Ce qui frappe dans cette définition c'est la place marginale des phylogénies. D'ailleurs, dans l'article lui-même, l'utilisation qui en est faite est purement descriptive. Même la question de l'inférence est éludée dans cette revue fondatrice.

Plus récemment, des revues se sont penchées sur ce champ en plein essor mais en éludant le flou qui entoure sa définition. Ainsi, pour Volz *et al.* (2013), la phylodynamique est « *l'étude de la manière dont les processus épidémiologiques, immunologiques et évolutifs agissent et potentiellement interagissent pour façonner les phylogénies [virales]* ». On note cependant qu'en 10 ans, la phylogénie a acquis un rôle central. La question de l'inférence est aussi bien plus prégnante et Volz *et al.* notent par exemple que des processus différents peuvent conduire à la même phylogénie, générant des problèmes d'identifiabilité des paramètres.

Dans un soucis de clarté, nous définirons la phylodynamique comme le lien entre données de séquences génétiques (ou phylogénies) et dynamique des populations.

4.2 Au plus près de l'épidémiologie

La phylodynamique est une des nombreuses prolongations des travaux fondateurs de Felsenstein (1981) pour calculer la probabilité d'observer un alignement de séquences à partir d'un arbre phylogénétique, aussi appelée vraisemblance phylogénétique. Les travaux précurseurs de phylodynamique datent des années 1990 et ont consisté à extraire l'information concernant l'histoire démographique des populations de microbes contenues dans les séquences (Holmes *et al.* 1995, Rodrigo & Felsenstein 1999). Ils ont montré que des variations de tailles de population affectent la distribution des embranchements d'une phylogénie d'individus issus de cette même population (aussi appelée généalogie en génétique des populations) permettant de détecter des périodes de croissance ou de décroissance de la taille population (Nee *et al.* 1995).

Toutefois, l'essor de la phylodynamique s'est faite grâce aux séquences microbiennes. On l'a vu, leur place de choix s'explique vraisemblablement par le parallèle intuitif entre chaîne de transmission et phylogénie d'infection. Pour aller plus loin, il convient de formaliser un peu plus ce que l'on entend par dynamique des populations ou, dans notre cas, dynamique épidémiologique. Le modèle SIR est un des modèles épidémiologiques les plus couramment utilisés (Anderson & May 1991, Keeling & Rohani 2008). Son nom provient des trois états possibles des hôtes : Susceptibles, Infectés et Retirés (c'est-à-dire, immunisés ou morts ou guéris). Il est formalisé à l'aide du système d'équations différentielles suivant

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = -\beta\ S(t)\ I(t) + \sigma\ R(t) \tag{1}$$

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = \beta\ S(t)\ I(t) - \gamma\ I(t) \tag{2}$$

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t} = \gamma\ I(t) - \sigma\ R(t) \tag{3}$$

où $S(t)$, $I(t)$ et $R(t)$ sont les densités d'individus susceptibles, infectés et immunisés au temps t, tandis que β, γ et σ sont les taux de transmission, guérison et perte d'immunité. Les individus susceptibles deviennent infectés suite à leur rencontre avec des individus infectés selon l'hypothèse de loi d'action de masse. Il sont ensuite infectés pendant en moyenne $1/\gamma$ unités de temps, après quoi ils deviennent immunisés. L'immunité est ici supposée durer en moyenne $1/\sigma$ unités de temps.

Pour les épidémiologistes de terrain, un des paramètres clés est le nombre de reproduction de base, dénoté R_0. Il correspond au nombre moyen d'infections secondaires engendrées par un individu infecté dans une population entièrement susceptible (Anderson & May 1991). Dans un modèle déterministe, pour qu'il puisse y avoir une épidémie, il faut que $R_0 > 1$. On peut facilement calculer ce R_0 à partir de tout système d'équations différentielles (Diekmann & Heesterbeek 2000, Hurford *et al.* 2010). Par exemple, d'après l'équation 2, on a $R_0 = \beta S(0)/\gamma$, où $S(0)$ est le nombre d'individus présents initialement dans la population. Les interventions en santé publique ont pour but de diminuer ce R_0 par des méthodes thérapeutiques ou de quarantaine.

Bien que simple, le modèle SIR comporte des phénomènes non linéaires. Ainsi, en début d'épidémie, la quantité de nouvelles infections par unité de temps $\beta\ S(t)\ I(t)$, aussi appelée « incidence », augmente du fait de l'augmentation de $I(t)$, pour progressivement se ralentir avec la diminution du nombre d'hôtes susceptibles ($S(t)$). Les durées d'infections sont aussi difficiles à estimer en utilisant uniquement des données d'incidence car leur composante temporelle est limitée. Ceci illustre à la fois les défis mais aussi le potentiel de la phylodynamique : parvenir à tirer partie du signal temporel présent dans les séquences génétiques tout en tenant compte des contraintes inhérentes à l'épidémiologie, tels que les processus non linéaires.

4.3 Coalescent

La phylodynamique s'est d'abord appuyée sur la théorie du coalescent de Kingman (1982), surtout en analysant le nombre de lignées au cours du temps (ou « Lineage Through Time », LTT, en anglais) (Ong *et al.* 1996, Pybus *et al.* 1999). Ceci revient en pratique à sommer le nombre de branches actives d'une phylogénie datée à un moment donné (Nee *et al.* 1995, Rambaut *et al.* 1997). Ce point est important à souligner car il illustre le fait que la structure de la phylogénie elle-même (notamment d'éventuelles asymétries) entre peu en compte. Pour plus de détails, on pourra se référer à la revue de Kühnert *et al.* (2011) sur l'inférence phylodynamique, qui contient des aspects détaillés sur la théorie du coalescent.

Le développement des méthodes phylodynamiques a suivi la disponibilité croissante des données de séquences moléculaires de virus humains très étudiés comme le VIH, le virus de l'hépatite C (VHC) ou la grippe. On l'a vu, contrairement à la plupart des données génétiques des populations, les séquences virales sont souvent échantillonnées à différents moments dans le temps. Les méthodes d'inférence utilisant de telles de phylogénies datées permettent, entre autres, d'estimer les variations de la taille de la population de pathogènes au cours du temps à partir des séquences et des dates d'échantillonnage (Drummond *et al.* 2002, 2005, Lemey *et al.* 2003, Rambaut *et al.* 2008). Des méthodes plus générales, comme les Bayesian Skyline (Drummond *et al.* 2005) et Skygrid (Gill *et al.* 2013), ont rendu possible la reconstruction de dynamiques de population de plus en plus riches.

Les modèles de coalescent ont longtemps été basés uniquement sur des cas démographiques simples, correspondant au mieux au modèle épidémiologique SI

(pour « Susceptible Infectés ») avec deux classes d'hôtes. Il existe des exceptions avec des modèles de coalescent de phylodynamique analysant des populations structurées (Hudson 1990, Notohara 1990), par exemple pour décrire la dynamique de différentes souches d'influenza (Koelle *et al.* 2006, Vaughan *et al.* 2014).

La pertinence des modèles de coalescent standards vis à vis des pathogènes infectieux a donc été progressivement remise en question. En effet, dans la plupart de ces modèles, le taux de coalescence est inversement proportionnel à la taille efficace de la population ($N_e(t)$), qui est différente de la taille de population absolue. Ce problème se pose même au sein d'un individu infecté par un virus diversifié (comme le discutent Kouyos *et al.* (2006) dans le cas du VIH). Il est d'autant plus aiguë au niveau inter-hôtes que l'on fait l'hypothèse que le $N_e(t)$, qui provient de données de séquences de populations microbiennes, est proportionnel au nombre d'hôtes infectés (Pybus *et al.* 2000, Drummond *et al.* 2002). Des travaux ont d'abord bien pointé que la reconstruction phylodynamique de $N_e(t)$ ne devait pas être interprétée comme une dynamique du nombre d'infections mais comme une mesure de la diversité génétique relative, elle-même influencée par la taille de population absolue, la structure de la population et la variabilité reproductive (Pybus *et al.* 2001, Rambaut *et al.* 2008). Plus récemment, les travaux autour du coalescent structuré (Hudson 1990) ont approfondi le lien entre généalogie et dynamique des infections (Volz *et al.* 2009, Bedford *et al.* 2010, Volz 2012, Volz & Siveroni 2018). L'hypothèse sous-jacente est que chaque lignée de la généalogie correspond à un seul hôte infecté (on néglige donc l'évolution intra-hôte), ce qui implique que les événements de coalescence correspondent aux événements de transmission. Le taux de coalescence est donc proportionnel à l'incidence du microbe étudié (soit $\beta S(t)I(t)$ dans le modèle SIR). Il varie donc de manière non-linéaire au cours du temps, sauf dans des conditions où le nombre d'individus susceptibles peut être considéré comme constant, par exemple en tout début d'épidémie (Frost & Volz 2010, Koelle & Rasmussen 2012).

Ces travaux ont conduit au développement de nouvelles approches d'inférence bayésienne basées sur le calcul de la vraisemblance d'un modèle de coalescent associé au modèle épidémiologique. Par exemple, Rasmussen *et al.* (2011), ont utilisé l'expression de la vraisemblance du modèle de coalescent associé à une épidémie de type SIR établi par Volz *et al.* (2009), pour développer une approche permettant d'utiliser à la fois une phylogénie datée et des données de série temporelle. Ce type d'approche a ensuite été étendu pour une large gamme de modèles épidémiologiques (Volz 2012, Rasmussen *et al.* 2014). Ces approches se basent sur la simulation de trajectoires épidémiologiques pour ensuite calculer la vraisemblance de la phylogénie datée sous l'hypothèse d'un modèle de coalescent en intégrant sur les variations de taille de population possibles au cours du temps décrites par les trajectoires. Néanmoins, elles nécessitent souvent de grosses ressources computationnelles (Ratmann *et al.* 2017).

4.4 Modèles de naissance et de mort

Parmi les approches dérivant une vraisemblance phylogénétique, une alternative aux modèles de coalescent se base sur le processus de naissance et de mort (Kendall 1948). Ces modèles ont été appliqués à l'inférence de paramètres épidémiologiques à partir de données de génotypage (Tanaka *et al.* 2006), ainsi qu'à l'étude de la dynamique antigénique du virus de la grippe (Koelle *et al.* 2009). Dans sa formulation la plus simple, le modèle de naissance et de mort est l'équivalent exact d'un modèle SI où la taille de population d'hôtes susceptibles est constante, ce qui peut être le cas en début d'épidémie (car l'ensemble de la population est susceptible). De plus, si le taux de naissance est supérieur au taux de mort, ce modèle peut être comparé à un modèle de coalescent classique avec croissance exponentielle de la taille de population (Stadler 2009).

Une des différences majeures par rapport aux modèles de coalescent classiques est que dans les modèles BD la taille de population peut varier de manière stochastique au cours du temps alors que, pour les modèles de coalescent classiques, la taille de la population varie de manière déterministe. Stadler *et al.* (2015) ont d'ailleurs comparé des modèles de coalescent (Volz 2012, Rasmussen *et al.* 2014) au modèle BD pour l'inférence phylodynamique et ont montré qu'un modèle de coalescent classique ne parvient pas à inférer les bons temps de coalescence de l'arbre de transmission correspondant à une épidémie de type SI (de dynamique identique au modèle stochastique BD). Toutefois, ce biais est corrigé dans les nouvelles versions du modèle de coalescent qui tiennent compte des trajectoires épidémiologiques, mais seulement pour des petites valeurs de R_0 et des grandes tailles de population.

Tout comme pour le coalescent, les approches de phylodynamique basées sur les modèles de naissance et de mort ont connu des extensions pour capturer des modèles épidémiologiques plus détaillés. Une des avancées a ainsi été d'obtenir une approche de type « skyline » non pas pour estimer les variations de taille de population efficace au cours du temps, mais les variations des paramètres épidémiologique tel que le R_0 ou la durée d'infection (Stadler *et al.* 2013, Kühnert *et al.* 2014). Plus récemment, Kühnert *et al.* (2018) ont généralisé le modèle de naissance et de mort à des populations d'hôtes hétérogènes, permettant par exemple de mesurer l'effet des des mutations de résistance aux antirétroviraux sur la propagation du VIH.

4.5 Limites des approches avec vraisemblance

La plupart des méthodes phylodynamiques existantes nécessitent la dérivation d'une fonction de vraisemblance phylogénétique. Si certaines modèles, comme le coalescent structuré, peuvent capturer des processus épidémiologiques très riches, pour la plupart l'ajout du moindre détail biologique nécessite un énorme travail analytique. Et quand bien même la fonction de vraisemblance peut être exprimée analytiquement, sa résolution numérique n'est pas toujours possible. Enfin, se pose le soucis des temps de calcul nécessaires pour trouver les valeurs de paramètres maximisant la vraisemblance. Si les approches bayésiennes ont permis de gagner du temps, les temps de convergence des méthodes de Monte-Carlo par chaînes de Markov augmentent rapidement avec les détails du modèle et, surtout, avec le nombre de séquences (Ratmann *et al.* 2017).

Plus généralement, le champ semble vivre un basculement. Depuis son origine, la phylodynamique a visé a extraire le maximum de signal d'un nombre de données limitées. Pour ce faire, inférer la phylogénie tout en inférant le modèle démographique à partir d'un alignement de séquences daté permettait de maximiser l'utilisation de l'information. C'est là aussi que réside le succès des logiciels Beast (Drummond & Rambaut 2007) et Beast2 (Bouckaert *et al.* 2014) : permettant de générer des résultats cohérents de manière relativement simple, ils offrent à l'utilisateur de réaliser une estimation de paramètres à partir d'un modèle donné (typiquement impliquant le coalescent ou les processus de naissance et de mort). Actuellement, la problématique s'est inversée et il s'agit de trier les informations issues d'une masse de données. Face à cela, les algorithmes estimant simultanément phylogénie et modèle démographique deviennent lents. Cela a été particulièrement frappant pendant le « concours » PANGEA, qui visait à comparer la capacité de différentes approches à inférer des paramètres épidémiologiques à partir d'alignement de séquences simulés. La première étape pour tous les participants fut d'inférer une phylogénie par des méthodes de maximum de vraisemblance rapides (avec des logiciels tels que PhyML ou fastML) pour ensuite appliquer des méthodes phylodynamique (Ratmann *et al.* 2017).

4.6 Phylodynamique ABC

Une des réponses aux limites listées ci-dessus a été de renoncer à la dérivation analytique d'une fonction de vraisemblance en se tournant vers des techniques de simulations numériques. Ainsi, les méthodes dites ABC (pour « Approximate Bayesian Computation »en anglais) consistent à simuler un grand nombre de jeux de données en utilisant un modèle sous-jacent aux valeurs de paramètres connues pour déterminer les valeurs de paramètres qui permettent de générer les données les plus proches aux données observées. Dans le cas des phylogénies, cette approche est particulièrement adaptée car si les phylogénies sont des objets difficiles à comparer deux à deux, leur simulation s'avère assez aisée du fait de ce parallèle avec une chaîne de transmission. En pratique, un algorithme tel que celui de Gillespie, couramment utilisé en épidémiologie (Keeling & Rohani 2008) suffit. La difficulté réside plus dans la comparaison entre arbres phylogénétiques. Pour cela, une méthode classique consiste à décomposer les arbres en statistiques de résumé telles que la longueur moyenne des branches, la variance de la longueur de branche, ou l'asymétrie de l'arbre. Ensuite, comparer deux arbres au travers de leurs valeurs de statistiques de résumé est aisé.

L'ABC en phylodynamique n'est pas une idée si neuve et Ratmann *et al.* (2012) avaient déjà tenté de l'utiliser pour combiner données d'incidence et données phylogénétique. Le succès mitigé de leur approche peut s'expliquer par le fait que la méthode utilisée imposait une limite stricte au nombre de statistiques de résumé utilisable. Plus récemment, nous avons développé une approche utilisant l'ABC régression (Csilléry *et al.* 2012, Blum 2018), qui s'accommode d'un grand nombre de statistiques de résumé en effectuant une sélection de variables. Les tests sur des modèles SIR ont montré une puissance d'inférence comparable à celles obtenues par des modèles de naissance et de mort dans Beast (Saulnier *et al.* 2017). De plus, une des force de ces approches est que leur puissance augmente avec la taille de la phylogénie, sans ralentir les calculs de manière rédhibitoire. Toutefois, des analyses sur des modèles plus détaillés sont nécessaires pour évaluer le vrai potentiel de la méthode. Un autre avantage est que les modèles de régression utilisés dans ces approches ABC, par exemple les forêts aléatoires, offrent des possibilités renforcées pour la comparaison de modèles (Pudlo *et al.* 2016).

5 Phylogéographie d'infections

La phylogéographie en tant que telle est une discipline ancienne (Avise *et al.* 1987) dont les méthodes statistiques étaient bien développées avant l'avènement de la phylodynamique (Knowles & Maddison 2002). Toutefois, les phylogénies d'infection ont incontestablement renouvelé le champ. Comme pour le calibrage dans le temps, la raison principale est que la rapide évolution microbienne permet de générer des arbres non-ultramétriques, c'est-à-dire pour lesquels certaines feuilles sont contemporaines de nœuds internes. Fidèles à notre définition de phylodynamique, nous ne présenterons ici que les approches qui incorporent une dynamique des populations. La plupart sont

d'ailleurs détaillées dans la synthèse de Lemey *et al.* (2009), avec des travaux plus récents résumés par Baele *et al.* (2018).

Le premier type d'approches considère la localisation spatiale comme un trait discret. D'un point de vue pratique, les modèles sous-jacents développés sont les mêmes chaînes des Markov en temps continu utilisées pour l'évolution des séquences. Ce parallèle avec les mutations de bases nucléotidiques fait que l'on parle souvent de modèle de « mugration », les migrations y étant vues comme des événements aléatoires pouvant se produire à des fréquences différentes. Les approches bayésiennes permettent d'inférer ces fréquences, qui correspondent à la connectivité entre les localisations, tout en inférant l'histoire évolutive (Lemey *et al.* 2009). Ces approches ont été appliquées dans de nombreux contextes, en partie grâce à leur implémentation au sein des logiciels Beast et Beast2. Toutefois, l'épidémie récente du virus ébola a probablement conduit aux développements les plus poussés étant donné le grand nombre de séquences et l'information géographique associée (Dudas *et al.* 2017).

Une des limites de ces modèles discrets est qu'ils sont sensibles à des disparités en termes d'échantillonnage (un thème récurrent en phylodynamique). De Maio *et al.* (2015) ont développé un package de Beast appelé BASTA, qui s'appuie sur une approximation du coalescent structuré afin de rendre l'inférence phylogéographique plus robuste à l'échantillonnage, permettant ainsi de gérer plus de localités et de flux de migration. Récemment, Müller *et al.* (2018) ont proposé un logiciel plus rapide se basant sur une solution exacte du coalescent structuré.

Le second type d'approche consiste à modéliser la localisation géographique au moyen de deux traits continus, typiquement la latitude et la longitude (Lemey *et al.* 2009). Un avantage évident de cette méthode est qu'elle permet d'inférer la présence d'individus dans des endroits non échantillonnés. Elle offre aussi une dimension visuelle non négligeable en termes d'attractivité. Enfin, elle permet de calculer un paramètre de vitesse de propagation de l'épidémie. Cette modélisation spatiale peut facilement s'implémenter à l'aide de modèles de mouvement brownien classiques pouvant varier selon les branches. Ceci a grandement contribué à la populariser (Pybus *et al.* 2012). Toutefois, ce modèle de diffusion peut se révéler simpliste lorsqu'il s'agit d'une épidémie. De plus, des facteurs externes peuvent affecter la vitesse de propagation, ce qui peut être testé a posteriori (Dellicour *et al.* 2016).

Enfin, sans être à proprement parler de la phylogéographie, plusieurs études ont tenté d'inférer la structure d'un réseau de contact sous-jacent entre les individus, à partir des données de séquences génétiques. Ainsi, Leventhal *et al.* (2012) ont réalisé une étude de simulation pour déterminer dans quelle mesure des familles de réseaux différentes pouvaient être caractérisées par des statistiques de résumé différentes. Rasmussen *et al.* (2017) ont eux implémenté un modèle de maximum de vraisemblance en utilisant la technique des approximations par paires pour obtenir une solution analytique. Giardina *et al.* (2017) ont quant à eux utilisé une approche basée sur le calcul ABC décrit ci-dessous pour réaliser cette inférence. Enfin, pour certaines réseaux, tels que ceux de contact sexuels, ces inférences peuvent être compliquées par la dynamique du réseau lui-même (Metzig *et al.* 2019).

6 Traits d'histoire de vie des infections et des virus

La phylogénie comparative permet de déterminer dans quelle mesure la structure phylogénétique entre des populations explique la distribution des traits mesurés dans ces populations (Felsenstein 1985). Ignorer cette non-indépendance (phylogénétique) des données revient souvent à surestimer les corrélations potentielles entre traits. De manière surprenante, cette notion n'a été appliquée que tardivement aux traits d'histoire de vie des infections. En 2010, nous avons analysé la distribution des charges virales mesurées chez des personnes porteuses du VIH avant traitement et trouvé que, dans une sous-population, près de la moitié de la variance était expliquée par la phylogénie (Alizon *et al.* 2010). Pour une maladie infectieuse humaine, ce résultat a une implication particulière car, en faisant l'hypothèse que le seul lien entre les personnes infectées sont les virus, il démontre un contrôle génétique du virus sur le trait en question.

Le lien entre les notions de signal phylogénétique et d'héritabilité avait déjà été souligné pour les phylogénies d'espèces (Housworth *et al.* 2004). Pour les phylogénies d'infections, il prend encore plus de sens dans la mesure où l'on travaille avec une population d'individus et non d'espèces. Toutefois, il existe plusieurs écueils à prendre en compte quand on l'applique à des traits tels que la charge virale notamment l'évolution intra-hôte ou le goulot d'étranglement associé à la transmission (Leventhal & Bonhoeffer 2016). Dans une perspective phylodynamique, Vrancken *et al.* (2014) ont développé une approche bayésienne afin d'estimer ce signal phylogénétique en même temps que la phylogénie. De plus, Mitov & Stadler (2017) ont quand à eux relâché l'hypothèse d'évolution neutre en utilisant un processus de Ornstein–Uhlenbeck.

Un autre application directe de cette évolution des traits a porté non plus sur un trait d'une infection mais bien sur un trait viral. Une des raisons pour lesquelles le virus de la grippe ré-émerge chaque année est liée à l'évolution rapide des protéines de surface présentées au système immunitaire (Alizon 2016).

Cette évolution peut être capturée via des analyses d'inhibition réalisées en infectant des furets, qui sont un animal modèle idéal pour le virus influenza. On obtient alors une sorte de matrice décrivant la capacité des anticorps synthétisés par un furet suite à une infection par un virus grippal X à empêcher l'agglutination des globules rouges par un virus grippal Y. De là, on peut calculer une distance entre virus. On savait depuis les travaux de Smith *et al.* (2004) qu'un clustering en deux dimensions permettait de visualiser l'évolution des antigènes influenza avec une très forte structure temporelle (les virus isolés en 1995 étant plus proches de ceux isolés en 1992 ou 1998 que de ceux isolés en 1987 ou 2002). Le tour de force de Bedford *et al.* (2014) a été de combiner ce trait lié aux analyses d'inhibition avec les séquences génétiques. En combinant les sources d'information moléculaires et antigéniques, ils sont parvenus à mieux expliquer les variations de prévalence observées d'une année sur l'autre.

Au final, la combinaison entre phylogénie et traits d'histoire de vie ressemble beaucoup à la phylogéographie. Toutefois, il faut se méfier de cette analogie car les traits soulèvent des problématiques biologiques propres et leur évolution diffère fortement d'une propagation spatiale.

7 Perspectives et défis

La phylodynamique s'est construite sur des promesses, la principale d'entre elles étant de fournir un débouché concret au milliers de séquences microbiennes générées chaque jour en routine dans les laboratoires. Si en une petite vingtaine d'années les progrès accomplis ont été significatifs, le champ reste toutefois marginal en santé publique. Dans le cas de l'épidémie d'ébola de 2014-2014 en Afrique de l'Ouest, certes les deux premières publications d'envergure contenait des arbres phylogénétique (Baize *et al.* 2014, Gire *et al.* 2014), mais l'aspect phylodynamique y était très limité (pas d'estimation de la dynamique des populations par exemple). De plus, les analyses réalisées par le groupe de travail de l'OMS se sont quant à elles limitées aux méthodes traditionnelles basées sur les relevés d'incidence et les suivis de contacts (WHO Ebola Response Team 2014). Cette frilosité peut avoir deux origines. La première est que la phylodynamique, comme la plupart des inférences phylogénétique, repose sur des hypothèses fortes telles que l'absence de recombinaison ou la neutralité de l'évolution moléculaire. Certes certaines études incorporent la sélection naturelle dans la génération de phylogénies (Neher & Hallatschek 2013). Rasmussen & Stadler (2019) ont eux inclus explicitement la valeur sélective (ou « fitness » en anglais) dans un modèle de naissance et de mort et ainsi à estimer des différences entre lignées et l'effet de mutations sur la fitness. Mais, pour le moment, de telles approches demeurent l'exception.

Une autre limite moins discutée est que la phylodynamique impliquant des fonctions de vraisemblance a du mal à incorporer des données hétérogènes. Certes, elle réalise un tour de force en n'utilisant que les données de séquences génétiques pour réaliser une inférence, mais, du point de vue appliqué, pourquoi se couper d'autres sources de données, en particulier de données d'incidence ? Cette agrégation de données de sources différentes a été un peu explorée, notamment au travers d'approches ABC qui facilitent cette intégration de données hétérogènes (Ratmann *et al.* 2012, Smith *et al.* 2017, Saulnier 2017). Elle pourrait d'ailleurs aider à corriger pour les biais d'échantillonnage (Volz & Frost 2014). Mais beaucoup reste à faire, notamment pour une utilisation large.

Si la phylodynamique rencontre des obstacles pour s'imposer en épidémiologie, elle rencontre plus de succès à des endroits où on l'attendait moins. Ainsi, l'explosion du nombre de séquences a contribué à un essor des phylogénies intra-hôtes (Hartfield *et al.* 2014). Le terme de phyloanatomie a même été introduit pour caractériser ces approches phylodynamiques appliquées à la propagation du virus entre les organes (Bons & Regoes 2018).

Un aspect qui se révèle encore plus prometteurs concerne l'analyse des clusters de transmission, qui correspondent aux zones de la phylogénie densément peuplées en feuilles. Un des apports du NGS a été de permettre l'identification de paires de transmission au sein de ces clusters et même d'inférer la direction de la transmission (Romero-Severson *et al.* 2016, Wymant *et al.* 2018). Cette dernière est particulièrement importante en santé publique car elle permet d'identifier des facteurs de risque associés au donneur ou au receveur (Le Vu *et al.* 2019).

Enfin, bien que jeune, le champ de la phylodynamique a vu sa philosophie bouleversée. Si initialement son but était d'extraire le maximum d'information d'un nombre limité de données génétiques, il se retrouve maintenant confronté à une avalanche de données (génétiques ou non). Dès lors, les outils tels que Beast se retrouvent en porte à faux car leurs temps de calcul augmentent de manière non linéaire avec le nombre et la longueur des séquences, même si des librairies permettent de compenser ces ralentissements (Ayres *et al.* 2019). Face à cela, les techniques basées sur l'intelligence artificielle et l'ABC peuvent fournir une réponse pertinente.

References

Alizon, S., 2016 *C'est grave docteur Darwin ? L'évolution, les microbes et nous*. Paris, France: Le Seuil.

Alizon, S. & Fraser, C., 2013 Within-host and between-host

evolutionary rates across the HIV-1 genome. *Retrovirology* **10**, 49. (doi: 10.1186/1742-4690-10-49).

Alizon, S., von Wyl, V., Stadler, T., Kouyos, R. D., Yerly, S., Hirschel, B., Böni, J., Shah, C., Klimkait, T., Furrer, H., Rauch, A., Vernazza, P., Bernasconi, E., Battegay, M., Bürgisser, P., Telenti, A., Günthard, H. F., Bonhoeffer, S. & Study, t. S. H. C., 2010 Phylogenetic approach reveals that virus genotype largely determines HIV set-point viral load. *PLoS Pathog.* **6**, e1001123. (doi: 10.1371/journal.ppat.1001123).

Anderson, R. M. & May, R. M., 1991 *Infectious Diseases of Humans. Dynamics and Control*. Oxford: Oxford University Press.

Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., Reeb, C. A. & Saunders, N. C., 1987 INTRASPECIFIC PHYLOGEOGRAPHY: The mitochondrial DNA bridge between population genetics and systematics. *Ann Rev Ecol Syst* **18**, 489–522. URL https://doi.org/10.1146/annurev.es.18.110187.002421. (doi: 10.1146/annurev.es.18.110187.002421).

Ayres, D. L., Cummings, M. P., Baele, G., Darling, A. E., Lewis, P. O., Swofford, D. L., Huelsenbeck, J. P., Lemey, P., Rambaut, A. & Suchard, M. A., 2019 BEAGLE 3: improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. *Systematic Biology* **68**, 1052–1061. URL https://academic.oup.com/sysbio/article/68/6/1052/5477405. (doi: 10.1093/sysbio/syz020).

Baele, G., Dellicour, S., Suchard, M. A., Lemey, P. & Vrancken, B., 2018 Recent advances in computational phylodynamics. *Curr Opin Virol* **31**, 24–32. URL http://www.sciencedirect.com/science/article/pii/S187962571830066X. (doi: 10.1016/j.coviro.2018.08.009).

Baize, S., Pannetier, D., Oestereich, L., Rieger, T., Koivogui, L., Magassouba, N., Soropogui, B., Sow, M. S., Keïta, S., De Clerck, H., Tiffany, A., Dominguez, G., Loua, M., Traoré, A., Kolié, M., Malano, E. R., Heleze, E., Bocquin, A., Mély, S., Raoul, H., Caro, V., Cadar, D., Gabriel, M., Pahlmann, M., Tappe, D., Schmidt-Chanasit, J., Impouma, B., Diallo, A. K., Formenty, P., Van Herp, M. & Günther, S., 2014 Emergence of Zaire Ebola virus disease in Guinea. *N Engl J Med* **371**, 1418–25. (doi: 10.1056/NEJMoa1404505).

Balfe, P., Simmonds, P., Ludlam, C. A., Bishop, J. O. & Brown, A. J., 1990 Concurrent evolution of human immunodeficiency virus type 1 in patients infected from the same source: rate of sequence change and low frequency of inactivating mutations. *J Virol* **64**, 6221–6233. URL https://jvi.asm.org/content/64/12/6221.

Bedford, T., Cobey, S., Beerli, P. & Pascual, M., 2010 Global Migration Dynamics Underlie Evolution and Persistence of Human Influenza A (H3n2). *PLoSPath* **6**, e1000918. URL https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1000918. (doi: 10.1371/journal.ppat.1000918).

Bedford, T., Suchard, M. A., Lemey, P., Dudas, G., Gregory, V., Hay, A. J., McCauley, J. W., Russell, C. A., Smith, D. J. & Rambaut, A., 2014 Integrating influenza antigenic dynamics with molecular evolution. *eLife* **3**, e01914. URL https://doi.org/10.7554/eLife.01914. (doi: 10.7554/eLife.01914).

Beerenwinkel, N., Däumer, M., Oette, M., Korn, K., Hoffmann, D., Kaiser, R., Lengauer, T., Selbig, J. & Walter, H., 2003 Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids. Res.* **31**, 3850–5. (doi: 10.1093/nar/gkg575).

Blum, M. G. B., 2018 Regression Approaches for ABC. In *Handbook of Approximate Bayesian Computation* (eds. S. A. Sisson, Y. Fan & M. A. Beaumont), pp. 71–85. (doi 10.1201/9781315117195-3).

Bons, E. & Regoes, R. R., 2018 Virus dynamics and phyloanatomy: Merging population dynamic and phylogenetic approaches. *Immunological Reviews* **285**, 134–146. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/imr.12688. (doi: 10.1111/imr.12688).

Bonsall, D., Golubchik, T., Cesare, M. d., Limbada, M., Kosloff, B., MacIntyre-Cockett, G., Hall, M., Wymant, C., Ansari, M. A., Abeler-Dörner, L., Schaap, A., Brown, A., Barnes, E., Piwowar-Manning, E., Wilson, E., Emel, L., Hayes, R., Fidler, S., Ayles, H., Bowden, R. & Fraser, C., 2018 A comprehensive genomics solution for HIV surveillance and clinical monitoring in a global health setting. *bioRxiv* p. 397083. URL https://www.biorxiv.org/content/10.1101/397083v4. (doi: 10.1101/397083).

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A. & Drummond, A. J., 2014 BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput Biol* **10**, e1003537. URL https://dx.plos.org/10.1371/journal.pcbi.1003537. (doi: 10.1371/journal.pcbi.1003537).

Chaillon, A., Samleerat, T., Zoveda, F., Ballesteros, S., Moreau, A., Ngo-Giang-Huong, N., Jourdain, G., Gianella, S., Lallemant, M., Depaulis, F. & Barin, F., 2014 Estimating the timing of mother-to-child transmission of the human immunodeficiency virus type 1 using a viral molecular evolution model. *PLoS One* **9**, e90421. (doi: 10.1371/journal.pone.0090421).

Coltart, C. E. M., Hoppe, A., Parker, M., Dawson, L., Amon, J. J., Simwinga, M., Geller, G., Henderson, G., Laeyendecker, O., Tucker, J. D., Eba, P., Novitsky, V., Vandamme, A.-M., Seeley, J., Dallabetta, G., Harling, G., Grabowski, M. K., Godfrey-Faussett, P., Fraser, C., Cohen, M. S., Pillay, D., Amon, J. J., Baggaley, R., Bernard, E. J., Burns, D., Cohen, M. S., Coltart, C. C., Dallabetta, G., Dawson, L., Dedes, N., Delpech, V., Eba, P. M., Fraser, C., Geller, G., German, D., Godfrey-Faussett, P., Grabowksi, M. K., Hall,

I., Harling, G., Henderson, G., Hoppe, A., Kozlakidis, Z., Laeyendecker, O., Mwanza, F., Novitsky, V., Parker, M., Pillay, D., Reis, A., Seeley, J., Simwanga, M., Tucker, J. D., Vandamme, A.-M., Wertheim, J. O. & Zimmerman, R., 2018 Ethical considerations in global HIV phylogenetic research. *Lancet HIV* **5**, e656–e666. URL http://www.sciencedirect.com/science/article/pii/S2352301818301346. (doi: 10.1016/S2352-3018(18)30134-6).

Csilléry, K., Olivier, F. & Blum, M. G. B., 2012 abc: an R package for approximate Bayesian computation (ABC). *Method Ecol Evol* **3**, 475–479. (doi: 10.1111/j.2041-210X.2011.00179.x).

De Maio, N., Wu, C.-H., O'Reilly, K. M. & Wilson, D., 2015 New routes to phylogeography: A bayesian structured coalescent approximation. *PLoS Genet* **11**, e1005421. (doi: 10.1371/journal.pgen.1005421).

Dellicour, S., Rose, R., Faria, N. R., Lemey, P. & Pybus, O. G., 2016 SERAPHIM: studying environmental rasters and phylogenetically informed movements. *Bioinformatics* **32**, 3204–3206. URL https://academic.oup.com/bioinformatics/article/32/20/3204/2196575. (doi: 10.1093/bioinformatics/btw384).

Dieckmann, U., Metz, J. A. J., Sabelis, M. W. & Sigmund, K. (eds.), 2002 *Adaptive dynamics of infectious diseases. In pursuit of virulence management*. Cambridge studies in adaptive dynamics. Cambridge, UK: Cambridge University Press.

Diekmann, O. & Heesterbeek, J., 2000 *Mathematical epidemiology of infectious diseases: model building, analysis, and interpretation*. New York: Wiley.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A., 2006 Relaxed phylogenetics and dating with confidence. *PLoS Biol* **4**, e88. (doi: 10.1371/journal.pbio.0040088).

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G. & Solomon, W., 2002 Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**, 1307–20.

Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R. & Rodrigo, A. G., 2003 Measurably evolving populations. *Trends Ecol. Evol.* **18**, 481–488. (doi: 10.1016/S0169-5347(03)00216-7).

Drummond, A. J. & Rambaut, A., 2007 BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214. (doi: 10.1186/1471-2148-7-214).

Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G., 2005 Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* **22**, 1185–92. (doi: 10.1093/molbev/msi103).

Dudas, G., Carvalho, L. M., Bedford, T., Tatem, A. J., Baele, G., Faria, N. R., Park, D. J., Ladner, J. T., Arias, A., Asogun, D. & others, 2017 Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* **544**, 309–315. (doi: 10.1038/nature22040).

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., Tatem, A. J., Sousa, J. D., Arinaminpathy, N., Pépin, J., Posada, D., Peeters, M., Pybus, O. G. & Lemey, P., 2014 The early spread and epidemic ignition of HIV-1 in human populations. *Science* **346**, 56–61. URL https://science.sciencemag.org/content/346/6205/56. (doi: 10.1126/science.1256739).

Felsenstein, J., 1981 Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol* **17**, 368–376. URL https://doi.org/10.1007/BF01734359. (doi: 10.1007/BF01734359).

Felsenstein, J., 1985 Phylogenies and the Comparative Method. *Am. Nat.* **125**, 1–15. (doi: 10.1086/284325).

Frost, S. D. W. & Volz, E. M., 2010 Viral phylodynamics and the search for an 'effective number of infections'. *Phil Trans R Soc Lond B* **365**, 1879–1890. URL https://royalsocietypublishing.org/doi/full/10.1098/rstb.2010.0060. (doi: 10.1098/rstb.2010.0060).

Giardina, F., Romero-Severson, E. O., Albert, J., Britton, T. & Leitner, T., 2017 Inference of Transmission Network Structure from HIV Phylogenetic Trees. *PLoS Comput Biol* **13**, e1005316. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005316. (doi: 10.1371/journal.pcbi.1005316).

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B. & Suchard, M. A., 2013 Improving bayesian population dynamics inference: A coalescent-based model for multiple loci. *Mol Biol Evol* **30**, 713–724. URL https://academic.oup.com/mbe/article/30/3/713/1041171. (doi: 10.1093/molbev/mss265).

Gire, S. K., Goba, A., Andersen, K. G., Sealfon, R. S. G., Park, D. J., Kanneh, L., Jalloh, S., Momoh, M., Fullah, M., Dudas, G., Wohl, S., Moses, L. M., Yozwiak, N. L., Winnicki, S., Matranga, C. B., Malboeuf, C. M., Qu, J., Gladden, A. D., Schaffner, S. F., Yang, X., Jiang, P.-P., Nekoui, M., Colubri, A., Coomber, M. R., Fonnie, M., Moigboi, A., Gbakie, M., Kamara, F. K., Tucker, V., Konuwa, E., Saffa, S., Sellu, J., Jalloh, A. A., Kovoma, A., Koninga, J., Mustapha, I., Kargbo, K., Foday, M., Yillah, M., Kanneh, F., Robert, W., Massally, J. L. B., Chapman, S. B., Bochicchio, J., Murphy, C., Nusbaum, C., Young, S., Birren, B. W., Grant, D. S., Scheiffelin, J. S., Lander, E. S., Happi, C., Gevao, S. M., Gnirke, A., Rambaut, A., Garry, R. F., Khan, S. H. & Sabeti, P. C., 2014 Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369–72. (doi: 10.1126/science.1259657).

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L., Daly, J. M., Mumford, J. A. & Holmes, E. C., 2004 Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–32. (doi: 10.1126/science.1090727).

Hartfield, M., Murall, C. L. & Alizon, S., 2014 Clinical applications of pathogen phylogenies. *Trends Mol Med* **20**, 394–404. (doi: 10.1016/j.molmed.2014.04.002).

Holmes, E. C., Dudas, G., Rambaut, A. & Andersen, K. G., 2016 The evolution of Ebola virus: Insights from the 2013-2016 epidemic. *Nature* **538**, 193–200. (doi: 10.1038/nature19790).

Holmes, E. C., Nee, S., Rambaut, A., Garnett, G. P., Harvey, P. H., Harvey, P. H., Leigh Brown, A. J. & Smith, J. M., 1995 Revealing the history of infectious disease epidemics through phylogenetic trees. *Phil Trans R Soc Lond B* **349**, 33–40. URL https://royalsocietypublishing.org/doi/abs/10.1098/rstb.1995.0088. (doi: 10.1098/rstb.1995.0088).

Housworth, E. A., Martins, E. P. & Lynch, M., 2004 The phylogenetic mixed model. *Am. Nat.* **163**, 84–96. (doi: 10.1086/380570).

Hudson, R. R., 1990 Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* **7**, 1–44. URL https://www.cabdirect.org/cabdirect/abstract/19910191040.

Hurford, A., Cownden, D. & Day, T., 2010 Next-generation tools for evolutionary invasion analyses. *J. R. Soc. Interface* **7**, 561–71. (doi: 10.1098/rsif.2009.0448).

Johnson, S. B. & Parker, M., 2019 The ethics of sequencing infectious disease pathogens for clinical and public health. *Nat Rev Genet* **20**, 313–315. URL https://www.nature.com/articles/s41576-019-0109-3. (doi: 10.1038/s41576-019-0109-3).

Jonassen, T. O., Stene-Johansen, K., Berg, E. S., Hungnes, O., Lindboe, C. F., Frøland, S. S. & Grinde, B., 1997 Sequence analysis of HIV-1 group O from Norwegian patients infected in the 1960s. *Virology* **231**, 43–47. URL http://www.sciencedirect.com/science/article/pii/S004268229798510X. (doi: 10.1006/viro.1997.8510).

Keeling, M. J. & Rohani, P., 2008 *Modeling infectious diseases in humans and animals*. Princeton University Press.

Kendall, D. G., 1948 On the Generalized "Birth-and-Death" Process. *The Annals of Mathematical Statistics* **19**, 1–15. URL https://projecteuclid.org/euclid.aoms/1177730285. (doi: 10.1214/aoms/1177730285).

Kermack, W. O. & McKendrick, A. G., 1927 A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A* **115**, 700–721.

Kingman, J. F. C., 1982 The coalescent. *Stochastic processes and their applications* **13**, 235–248.

Knowles, L. L. & Maddison, W. P., 2002 Statistical phylogeography. *Mol Ecol* **11**, 2623–2635. URL https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-294X.2002.01410.x. (doi: 10.1046/j.1365-294X.2002.01410.x).

Koelle, K., Cobey, S., Grenfell, B. & Pascual, M., 2006 Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3n2) in humans. *Science* **314**, 1898–903. (doi: 10.1126/science.1132745).

Koelle, K., Kamradt, M. & Pascual, M., 2009 Understanding the dynamics of rapidly evolving pathogens through modeling the tempo of antigenic change: Influenza as a case study. *Epidemics* **1**, 129–137. URL http://www.sciencedirect.com/science/article/pii/S1755436509000280. (doi: 10.1016/j.epidem.2009.05.003).

Koelle, K. & Rasmussen, D. A., 2012 Rates of coalescence for common epidemiological models at equilibrium. *J R Soc Interface* **9**, 997–1007. URL https://royalsocietypublishing.org/doi/full/10.1098/rsif.2011.0495. (doi: 10.1098/rsif.2011.0495).

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B. H., Wolinsky, S. & Bhattacharya, T., 2000 Timing the Ancestor of the HIV-1 Pandemic Strains. *Science* **288**, 1789–1796. URL https://science.sciencemag.org/content/288/5472/1789. (doi: 10.1126/science.288.5472.1789).

Kouyos, R. D., Althaus, C. L. & Bonhoeffer, S., 2006 Stochastic or deterministic: what is the effective population size of HIV-1? *Trends Microbiol* **14**, 507–11. (doi: 10.1016/j.tim.2006.10.001).

Kühnert, D., Kouyos, R., Shirreff, G., Pečerska, J., Scherrer, A. U., Böni, J., Yerly, S., Klimkait, T., Aubert, V., Günthard, H. F., Stadler, T., Bonhoeffer, S. & the Swiss HIV Cohort Study, 2018 Quantifying the fitness cost of HIV-1 drug resistance mutations through phylodynamics. *PLoS Pathog* **14**, e1006895. URL http://dx.plos.org/10.1371/journal.ppat.1006895. (doi: 10.1371/journal.ppat.1006895).

Kühnert, D., Stadler, T., Vaughan, T. G. & Drummond, A. J., 2014 Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death SIR model. *J R Soc Interface* **11**, 20131106. (doi: 10.1098/rsif.2013.1106).

Kühnert, D., Wu, C.-H. & Drummond, A. J., 2011 Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Inf Genet Evol* **11**, 1825–1841. URL http://www.sciencedirect.com/science/article/pii/S156713481100284X. (doi: 10.1016/j.meegid.2011.08.005).

Lake, J. A., 1988 Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. *Nature* **331**, 184–186. URL https://www.nature.com/articles/331184a0. (doi: 10.1038/331184a0).

Le Vu, S., Ratmann, O., Delpech, V., Brown, A. E., Gill, O. N., Tostevin, A., Dunn, D., Fraser, C. & Volz, E. M., 2019 HIV-1 Transmission Patterns in Men Who Have Sex with Men: Insights from Genetic Source Attribution Analysis. *AIDS Res Human Retrovir* **35**, 805–813. URL https://www.liebertpub.com/doi/full/10.1089/AID.2018.0236. (doi: 10.1089/aid.2018.0236).

Lemey, P., Pybus, O. G., Wang, B., Saksena, N. K., Salemi, M. & Vandamme, A.-M., 2003 Tracing the origin and history of the HIV-2 epidemic. *Proc Nat Acad Sci USA* **100**, 6588–6592. URL https://www.pnas.org/content/100/11/6588. (doi: 10.1073/pnas.0936469100).

Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A., 2009 Bayesian phylogeography finds its roots. *PLoS Comput Biol* **5**, e1000520. (doi: 10.1371/journal.pcbi.1000520).

Leventhal, G. E. & Bonhoeffer, S., 2016 Potential Pitfalls in Estimating Viral Load Heritability. *Trends Microbiol* **24**, 687–98. (doi: 10.1016/j.tim.2016.04.008).

Leventhal, G. E., Kouyos, R., Stadler, T., Wyl, V. v., Yerly, S., Böni, J., Cellerai, C., Klimkait, T., Günthard, H. F. & Bonhoeffer, S., 2012 Inferring epidemic contact structure from phylogenetic trees. *PLoS Comput Biol* **8**, e1002413. (doi: 10.1371/journal.pcbi.1002413).

Li, W. H., Tanimura, M. & Sharp, P. M., 1988 Rates and dates of divergence between AIDS virus nucleotide sequences. *Mol Biol Evol* **5**, 313–330. URL https://academic.oup.com/mbe/article/5/4/313/1026948. (doi: 10.1093/oxfordjournals.molbev.a040503).

Lion, S., 2018 Theoretical Approaches in Evolutionary Ecology: Environmental Feedback as a Unifying Perspective. *Am Nat* **191**, 21–44. (doi: 10.1086/694865).

Metzig, C., Ratmann, O., Bezemer, D. & Colijn, C., 2019 Phylogenies from dynamic networks. *PLoS Comput Biol* **15**, e1006761. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006761. (doi: 10.1371/journal.pcbi.1006761).

Mitov, V. & Stadler, T., 2017 The Heritability of Pathogen Traits: Definitions and Estimators. *Mol Biol Evol* p. \emphin review.

Müller, N. F., Rasmussen, D. & Stadler, T., 2018 MASCOT: parameter and state inference under the marginal structured coalescent approximation. *Bioinformatics* **34**, 3843–3848. URL https://academic.oup.com/bioinformatics/article/34/22/3843/5001387. (doi: 10.1093/bioinformatics/bty406).

Nee, S., Holmes, E. C., Rambaut, A., Harvey, P. H., Harvey, P. H., Leigh Brown, A. J. & Smith, J. M., 1995 Inferring population history from molecular phylogenies. *Phil Trans R Soc Lond B* **349**, 25–31. URL https://royalsocietypublishing.org/doi/abs/10.1098/rstb.1995.0087. (doi: 10.1098/rstb.1995.0087).

Neher, R. A. & Hallatschek, O., 2013 Genealogies of rapidly adapting populations. *Proc Nat Acad Sci USA* **110**, 437–442. URL https://www.pnas.org/content/110/2/437. (doi: 10.1073/pnas.1213113110).

Notohara, M., 1990 The coalescent and the genealogical process in geographically structured population. *J Math Biology* **29**, 59–75. URL https://doi.org/10.1007/BF00173909. (doi: 10.1007/BF00173909).

Ong, C.-K., Nee, S., Rambaut, A. & Harvey, P. H., 1996 Inferring the population history of an epidemic from a phylogenetic tree. *J theor Biol* **182**, 173–178. URL http://www.sciencedirect.com/science/article/pii/S0022519396901526. (doi: 10.1006/jtbi.1996.0152).

Ou, C.-Y., Ciesielski, C. A., Myers, G., Bandea, C. I., Luo, C.-C., Korber, B. T. M., Mullins, J. I., Schochetman, G., Berkelman, R. L., Economou, A. N., Witte, J. J., Furman, L. J., Satten, G. A., Maclnnes, K. A., Curran, J. W. & Jaffe, H. W., 1992 Molecular Epidemiology of HIV Transmission in a Dental Practice. *Science* **256**, 1165–1171. URL https://science.sciencemag.org/content/256/5060/1165. (doi: 10.1126/science.256.5060.1165).

Pimenoff, V. N., Oliveira, C. M. d. & Bravo, I. G., 2016 Transmission between Archaic and Modern Human Ancestors during the Evolution of the Oncogenic Human Papillomavirus 16. *Mol Biol Evol* **34**, 4–19. (doi: 10.1093/molbev/msw214).

Pudlo, P., Marin, J.-M., Estoup, A., Cornuet, J.-M., Gautier, M. & Robert, C. P., 2016 Reliable ABC model choice via random forests. *Bioinformatics* **32**, 859–866. (doi: 10.1093/bioinformatics/btv684).

Pybus, O. G., Charleston, M. A., Gupta, S., Rambaut, A., Holmes, E. C. & Harvey, P. H., 2001 The epidemic behavior of the hepatitis C virus. *Science* **292**, 2323–5. (doi: 10.1126/science.1058321).

Pybus, O. G., Holmes, E. C. & Harvey, P. H., 1999 The mid-depth method and HIV-1: a practical approach for testing hypotheses of viral epidemic history. *Mol Biol Evol* **16**, 953–959. URL https://academic.oup.com/mbe/article/16/7/953/2925485. (doi: 10.1093/oxfordjournals.molbev.a026184).

Pybus, O. G. & Rambaut, A., 2009 Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* **10**, 540–550. (doi: 10.1038/nrg2583).

Pybus, O. G., Rambaut, A. & Harvey, P. H., 2000 An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* **155**, 1429–1437. URL http://www.genetics.org/content/155/3/1429.short.

Pybus, O. G., Suchard, M. A., Lemey, P., Bernardin, F. J., Rambaut, A., Crawford, F. W., Gray, R. R., Arinaminpathy, N., Stramer, S. L., Busch, M. P. & Delwart, E. L., 2012 Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc Natl Acad Sci USA* **109**, 15066–71. (doi: 10.1073/pnas.1206598109).

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., Bore, J. A., Koundouno, R., Dudas, G., Mikhail, A. & others, 2016 Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232. URL http://www.nature.com/nature/journal/v530/n7589/abs/nature16996.html.

Rambaut, A., 2000 Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16**, 395–399. URL https://academic.oup.com/bioinformatics/article/16/4/395/187233. (doi: 10.1093/bioinformatics/16.4.395).

Rambaut, A., Harvey, P. H. & Nee, S., 1997 End-Epi: An application for inferring phylogenetic and population dynamical processes from molecular sequences. *Bioinformatics* **13**, 303–306. URL https://academic.oup.com/bioinformatics/article/13/3/303/423219. (doi: 10.1093/bioinformatics/13.3.303).

Rambaut, A., Pybus, O. G., Nelson, M. I., Viboud, C., Taubenberger, J. K. & Holmes, E. C., 2008 The genomic and epidemiological dynamics of human influenza A virus. *Nature* **453**, 615–619. URL https://www.nature.com/articles/nature06945. (doi: 10.1038/nature06945).

Rascovan, N., Sjögren, K.-G., Kristiansen, K., Nielsen, R., Willerslev, E., Desnues, C. & Rasmussen, S., 2019 Emergence and Spread of Basal Lineages of Yersinia pestis during the Neolithic Decline. *Cell* **176**, 295–305.e10. (doi: 10.1016/j.cell.2018.11.005).

Rasmussen, D. A., Kouyos, R., Günthard, H. F. & Stadler, T., 2017 Phylodynamics on local sexual contact networks. *PLOS Comput Biol* **13**, e1005448. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005448. (doi: 10.1371/journal.pcbi.1005448).

Rasmussen, D. A., Ratmann, O. & Koelle, K., 2011 Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput Biol* **7**, e1002136. (doi: 10.1371/journal.pcbi.1002136).

Rasmussen, D. A. & Stadler, T., 2019 Coupling adaptive molecular evolution to phylodynamics using fitness-dependent birth-death models. *eLife* **8**, e45562. (doi: 10.7554/eLife.45562).

Rasmussen, D. A., Volz, E. M. & Koelle, K., 2014 Phylodynamic inference for structured epidemiological models. *PLoS Comput Biol* **10**, e1003570. (doi: 10.1371/journal.pcbi.1003570).

Ratmann, O., Donker, G., Meijer, A., Fraser, C. & Koelle, K., 2012 Phylodynamic inference and model assessment with approximate bayesian computation: influenza as a case study. *PLoS Comput Biol* **8**, e1002835. (doi: 10.1371/journal.pcbi.1002835).

Ratmann, O., Hodcroft, E. B., Pickles, M., Cori, A., Hall, M., Lycett, S., Colijn, C., Dearlove, B., Didelot, X., Frost, S., Hossain, A. S. M. M., Joy, J. B., Kendall, M., Kühnert, D., Leventhal, G. E., Liang, R., Plazzotta, G., Poon, A. F. Y., Rasmussen, D. A., Stadler, T., Volz, E., Weis, C., Brown, A. J. L. & Fraser, C., 2017 Phylogenetic tools for generalized HIV-1 epidemics: Findings from the PANGEA-HIV methods comparison. *Mol Biol Evol* **34**, 185–203. URL http://mbe.oxfordjournals.org/content/34/1/185. (doi: 10.1093/molbev/msw217).

Rodrigo, A. G. & Felsenstein, J., 1999 Coalescent approaches to HIV population genetics. In *The evolution of HIV*, pp. 233–274. Crandell, K A, jhu press edn.).

Romero-Severson, E. O., Bulla, I. & Leitner, T., 2016 Phylogenetically resolving epidemiologic linkage. *Proc Nat Acad Sci USA* **113**, 2690–2695. URL https://www.pnas.org/content/113/10/2690. (doi: 10.1073/pnas.1522930113).

Saulnier, E., 2017 Phylodynamique des pathogènes viraux par calcul bayésien approché. Ph.D. thesis, Université de Montpellier.

Saulnier, E., Gascuel, O. & Alizon, S., 2017 Inferring epidemiological parameters from phylogenies using regression-ABC: A comparative study. *PLoS Comput Biol* **13**, e1005416. (doi: 10.1371/journal.pcbi.1005416).

Smith, D. J., Lapedes, A. S., de Jong, J. C., Bestebroer, T. M., Rimmelzwaan, G. F., Osterhaus, A. D. M. & Fouchier, R. A. M., 2004 Mapping the antigenic and genetic evolution of influenza virus. *Science* **305**, 371–376. (doi: 10.1126/science.1097211).

Smith, R. A., Ionides, E. L. & King, A. A., 2017 Infectious Disease Dynamics Inferred from Genetic Data via Sequential Monte Carlo. *Mol Biol Evol* **34**, 2065–2084. URL https://academic.oup.com/mbe/article/34/8/2065/3200416. (doi: 10.1093/molbev/msx124).

Smith, T. F., Srinivasan, A., Schochetman, G., Marcus, M. & Myers, G., 1988 The phylogenetic history of immunodeficiency viruses. *Nature* **333**, 573–575. URL https://www.nature.com/articles/333573a0. (doi: 10.1038/333573a0).

Stadler, T., 2009 On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *J Theor Biol* **261**, 58–66. (doi: 10.1016/j.jtbi.2009.07.018).

Stadler, T., Kühnert, D., Bonhoeffer, S. & Drummond, A. J., 2013 Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc Natl Acad Sci USA* **110**, 228–33. (doi: 10.1073/pnas.1207965110).

Stadler, T., Vaughan, T. G., Gavryushkin, A., Guindon, S., Kühnert, D., Leventhal, G. E. & Drummond, A. J., 2015 How well can the exponential-growth coalescent approximate constant-rate birth–death population dynamics? *Proc B* **282**, 20150420. URL https://royalsocietypublishing.org/doi/10.1098/rspb.2015.0420. (doi: 10.1098/rspb.2015.0420).

Tanaka, M. M., Francis, A. R., Luciani, F. & Sisson, S. A., 2006 Using Approximate Bayesian Computation to Estimate Tuberculosis Transmission Parameters From Genotype Data. *Genetics* **173**, 1511–1520. URL https://www.genetics.org/content/173/3/1511. (doi: 10.1534/genetics.106.055574).

To, T.-H., Jung, M., Lycett, S. & Gascuel, O., 2015 Fast dating using least-squares criteria and algorithms. *Syst Biol* **65**, 82–97. URL http://sysbio.oxfordjournals.org/content/early/2015/11/11/sysbio.syv068.abstract.

Vaughan, T. G., Kühnert, D., Popinga, A., Welch, D. & Drummond, A. J., 2014 Efficient Bayesian inference under the structured coalescent. *Bioinformatics* **30**, 2272–2279. URL https://academic.oup.com/bioinformatics/article/30/16/2272/2748160. (doi: 10.1093/bioinformatics/btu201).

Volz, E. M., 2012 Complex population dynamics and the coalescent under neutrality. *Genetics* **190**, 187–201. (doi: 10.1534/genetics.111.134627).

Volz, E. M. & Frost, S. D. W., 2014 Sampling through time and phylodynamic inference with coalescent and birth–death models. *J R Soc Interface* **11**, 20140945. URL https://royalsocietypublishing.org/doi/full/10.1098/rsif.2014.0945. (doi: 10.1098/rsif.2014.0945).

Volz, E. M., Koelle, K. & Bedford, T., 2013 Viral phylodynamics. *PLoS Comput Biol* **9**, e1002947. (doi: 10.1371/journal.pcbi.1002947).

Volz, E. M., Kosakovsky Pond, S. L., Ward, M. J., Leigh Brown, A. J. & Frost, S. D. W., 2009 Phylodynamics of infectious disease epidemics. *Genetics* **183**, 1421–30. (doi: 10.1534/genetics.109.106021).

Volz, E. M. & Siveroni, I., 2018 Bayesian phylodynamic inference with complex models. *PLoS Comput Biol* **14**, e1006546. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006546. (doi: 10.1371/journal.pcbi.1006546).

Vrancken, B., Lemey, P., Rambaut, A., Bedford, T., Longdon, B., Günthard, H. F. & Suchard, M. A., 2014 Simultaneously estimating evolutionary history and repeated traits phylogenetic signal: applications to viral and host phenotypic evolution. *Meth Ecol Evol* **6**, 67–82. (doi: 10.1111/2041-210x.12293).

Weinert, L. A., Depledge, D. P., Kundu, S., Gershon, A. A., Nichols, R. A., Balloux, F., Welch, J. J. & Breuer, J., 2015 Rates of Vaccine Evolution Show Strong Effects of Latency: Implications for Varicella Zoster Virus Epidemiology. *Mol Biol Evol* p. msu406. URL http://mbe.oxfordjournals.org/content/early/2015/02/11/molbev.msu406. (doi: 10.1093/molbev/msu406).

WHO Ebola Response Team, 2014 Ebola virus disease in West Africa–the first 9 months of the epidemic and forward projections. *N Engl J Med* **371**, 1481–95. (doi: 10.1056/NEJMoa1411100).

Worobey, M., Gemmel, M., Teuwen, D. E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J.-J., Kabongo, J.-M. M., Kalengayi, R. M., Van Marck, E., Gilbert, M. T. P. & Wolinsky, S. M., 2008 Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* **455**, 661–4. (doi: 10.1038/nature07390).

Wymant, C., Hall, M., Ratmann, O., Bonsall, D., Golubchik, T., de Cesare, M., Gall, A., Cornelissen, M. & Fraser, C., 2018 PHYLOSCANNER: Inferring transmission from within- and between-host pathogen genetic diversity. *Mol Biol Evol* **35**, 719–733. (doi: 10.1093/molbev/msx304).

Yoder, A. D. & Yang, Z., 2000 Estimation of Primate Speciation Dates Using Local Molecular Clocks. *Mol Biol Evol* **17**, 1081–1090. URL https://academic.oup.com/mbe/article/17/7/1081/1064709. (doi: 10.1093/oxfordjournals.molbev.a026389).