Phylodynamique des virus évoluant rapidement : approches par calcul bayésien approché et par vraisemblance

Gonché Danesh

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Évolution des systèmes infectieux

École doctorale GAIA

Unité de recherche MIVEGEC

Phylodynamique des virus évoluant rapidement : approches par calcul bayésien approché et par vraisemblance

Présentée par Gonché DANESH

Le 06 juillet 2021

Sous la direction de Samuel ALIZON et Marc CHOISY

Devant le jury composé de

Morgane ROLLAND, DR, Walter Reed Army Institute of Research	Rapporteur
Roger KOUYOS, Pr, University of Zurich	Rapporteur
Thérèse COMMES, Pr, Université de Montpellier	Examinateur
Rodolphe THIEBAUT, Pr, Université de Bordeaux	Examinateur
Raphael LEBLOIS, CR, INRA	Examinateur
Samuel ALIZON, DR, CNRS	Directeur
Marc CHOISY, CR, Oxford University	Co-encadrant

UNIVERSITÉ DE MONTPELLIER

Résumé

Le contrôle des maladies infectieuses est un enjeu majeur en santé publique humaine, animale et végétale. Aux données épidémiologiques classiques telles que l'incidence ou la prévalence, se sont ajoutées des données génétiques issues des progrès des techniques de séquençage. L'abondance de ces données associée à l'évolution rapide des pathogènes, en particulier des virus à ARN, a entraîné l'essor d'un champ disciplinaire appelé « phylodynamique » dont une des hypothèses sous-jacentes est que la manière dont les virus se propagent laisse des traces dans leurs génomes. Les méthodes phylodynamiques tirent parti des ces informations génomiques pour estimer des paramètres épidémiologiques tels que les taux de croissance d'une population virale, le nombre d'infections ou leur durée moyenne.

Les méthodes d'inférence phylodynamique bayésienne sont les plus couramment utilisées et reposent en général sur des fonctions de vraisemblance. Cependant, elles peuvent rapidement se révéler inadaptées pour des modèles composés de beaucoup de paramètres, par exemple pour refléter des hétérogénéités de transmission. D'autres méthodes d'inférence sont basées sur le calcul bayésien approché (ABC) et ne nécessitent aucune fonction de vraisemblance. Ces approches reposent sur des simulations à partir de modèles épidémiologiques, des statistiques de résumé capturant l'information épidémiologique des phylogénies et des techniques de régression. Jusqu'ici, les approches de phylodynamique avec ABC ont étudié des modèles épidémiologiques relativement simples.

Cette thèse se compose en deux principales parties. La première partie a consisté à développer un outil de simulations rapides de séries temporelles et de phylogénies, qui soit utilisable dans les approches ABC. Ce simulateur, TiPS, présente l'avantage de pouvoir inclure une grande diversité de modèles épidémiologiques tout en étant plus rapide que beaucoup d'outils existants.

La seconde partie de la thèse a consisté à étudier des épidémies dans des contextes différents en utilisant des approches ABC ou de vraisemblance.

La première étude a concerné l'analyse phylodynamique de la propagation du virus de l'hépatite C (VHC) à Lyon. Le conexte épidémiologique nous a conduit à étendre l'application de la méthode ABC à un modèle structuré avec deux types d'hôtes. Pour cela, nous avons développé de nouvelles statistiques de résumé adaptées aux phylogénies dites « labellisées » dont les feuilles sont associées à un groupe à risque. Cette approche ABC a ensuite été utilisée pour étudier l'avantage de transmission que conférerait la présence d'une certaine mutation aux VIH-1/O. La dernière étude est consacrée aux analyses épidémiologiques et phylodynamiques du SARS-CoV-2 à partir de données françaises en utilisant des méthodes d'inférence bayésienne basées sur la vraisemblance.

Bien qu'en plein essor, le domaine de la phylodynamique est encore peu connu et peu étudié en France. En plus de ses développements méthodologiques et de ses analyses d'épidémies en cours, ce travail de thèse illustre, à l'échelle nationale, le potentiel d'analyses de données de séquences génomiques pour aider à la prise de décision en santé publique surtout en temps de crise.

Abstract

Controlling infectious diseases is a major issue for human, animal, and plant public health. In addition to traditional epidemiological data such as incidence or prevalence, genetic data resulting from advances in sequencing techniques are now available. The abundance of these data associated with the rapid evolution of pathogens, in particular RNA viruses, has led to the development of a disciplinary field called "phylodynamics". One of the underlying hypotheses of this field is that the way viruses spread leaves footprints in their genomes. Phylodynamic methods take advantage of this genomic information to estimate epidemiological parameters such as the growth rates of an epidemic, the number of secondary infections caused, or the average infection duration.

Bayesian phylodynamic inference methods are the most commonly used and are generally based on likelihood functions. However, they can be unsuitable for models with many parameters, e.g. those that capture transmission heterogeneity. Other inference methods are based on Approximate Bayesian Computation (ABC) and do not require a likelihood function. These approaches involve simulations from epidemiological models, summary statistics capturing epidemiological information from phylogenies, and regression techniques. So far, phylodynamic ABC approaches have focused on simple epidemiological models.

This thesis work comprises two main parts. The first part consisted in developing a tool for rapid simulations of time series and phylogenies, which can be used in ABC approaches. This simulator, TiPS, has the advantage to include a large diversity of epidemiological models while being faster than many existing tools. The second part of the thesis consisted in studying epidemics in different contexts using ABC or likelihood-based approaches.

The first application consisted in a phylodynamic analysis of the spread of hepatitis C virus (HCV) in Lyon. The epidemiological context led us to extend the application of the ABC method to a structured model with two types of hosts. For this purpose, we developed new summary statistics adapted to labelled phylogenies, where leaves are associated with a risk group. This ABC approach was then also used in another context to study the transmission advantage that the presence of a certain mutation would confer to HIV-1/O. The last study is devoted to epidemiological and phylodynamic analyses of SARS-CoV-2 from French data using Bayesian likelihood-based inference methods.

Although in full expansion, the field of phylodynamics is poorly known and studied in France. In addition to its methodological developments and its analyses on current epidemics, this work illustrates, at a national scale, the potential of genomic sequence data analysis to help public health organisms to establish more efficient control measures, especially in times of crisis.

Remerciements

Je tiens tout d'abord à remercier Morgane Rolland, Roger Kouyos, Thérèse Commes, Rodolphe Thiebaut et Raphael Leblois d'avoir accepté de juger mes travaux de thèse.

Un immense merci à Samuel Alizon pour avoir accepté ma demande de stage de M1 en 2015 et pour m'avoir proposé cette thèse par la suite. Merci d'avoir toujours été disponible et à l'écoute de mes questions. Merci d'avoir été positif et optimiste lorsque je ne l'étais plus. J'admire tes qualités humaines et ta passion pour la science. Ces trois années ont été extrêmement enrichissantes à tes côtés.

Je remercie Marc Choisy pour les nombreux conseils avisés et retours sur les travaux de ma thèse, notamment sur le développement d'outil.

Je remercie Emma Saulnier qui m'a co-encadrée lors de mon stage de M1 et qui m'a accompagnée lors des premiers mois de ma thèse. Merci pour tous tes conseils et le temps que tu m'as accordé.

Je tiens à remercier les membres de l'équipe ETE, en particulier l'équipe de modélisation. Christian, Ramsès, Mircea, Yannis, Samuel, merci de m'avoir montré l'aspect riche en terme d'échanges et l'aspect humain de la recherche scientifique au cours des réunions ou de simples pauses. Je remercie mes collègues doctorants. Jérôme, merci d'être venu au moins une fois au Kayak. Thomas, merci d'avoir été le premier à me suivre dans le clan des doctorants du Kayak et bien sûr merci pour tes délicieux gâteaux. Bastien et Baptiste, merci de venir nous voir au bureau presque tous les matins, je vous nomme responsables de la tâche d'installer un canapé dans le bureau. Basak, thanks for always having snacks at the office.

Merci aux personnes du Kayak notamment Franck, Virginie, Ève et Céline pour les pauses café, parfois longues il faut quand même se l'avouer. Merci pour la bonne ambiance.

Je remercie les plateformes bioinformatiques Genotoul et I-Trop pour l'accès au cluster de calcul, en particulier Ndomassi Tando pour avoir installé tant d'outils que je n'arrivais pas à installer malgré mon entêtement.

Je tiens à remercier mes proches. Mes poulettes Marianne, Laura et Quentin, chaque superbe week-end passé avec vous m'a rempli de joie. Florian, merci pour ton soutien et les super brunch. Jules, merci pour les soirées séries et les glaces mars et twix toujours réconfortantes. Je souhaite exprimer ma gratitude envers toi, Simon. Merci pour ton soutien indéniable, pour m'avoir supportée et encouragée.

Et enfin, je ne saurais trouver les mots pour exprimer ma reconnaissance envers mes parents qui m'ont toujours soutenue et ont cru en moi. C'est à vous que je dois cette thèse.

Contexte de la thèse

Cette thèse a été co-encadrée par Samuel Alizon et Marc Choisy et réalisée au sein du laboratoire MIVEGEC (Maladies Infectieuses et Vecteurs : Écologie, Génétique, Évolution et Contrôle). Cette thèse s'est déroulée autour d'un champ disciplinaire en plein essor, la phylodynamique. La phylodynamique vise à comprendre la dynamique de propagation des infections microbiennes en analysant les données de séquences génétiques de ces microbes.

Cette thèse a bénéficié d'un premier financement par la Fondation pour la Recherche Médicale (FRM) entre 2017 et 2020. Le projet initialement prévu de cette thèse était l'étude d'un jeu de données de VIH-1 récoltées auprès de patients infectés dans 50 hôpitaux en Thaïlande depuis la fin des années 1990, mais aussi des données de co-infections par le VIH et l'hépatite B (VHB) ou C (VHC). Le projet impliquait une étude phylogéographique afin de comprendre la propagation du VIH-1 en Thaïlande, une étude de co-infections pour en comprendre l'origine, et enfin une analyse de l'existence de facteurs viraux affectant la virulence d'une infection par le VIH. Malheureusement, nous n'avons pas eu accès aux données. Néanmoins, nous avons pu étudier la dynamique de propagation de différents virus, en établissant une collaboration avec Laurent Cotte du CHU de Lyon et une collaboration avec Jean-Christophe Plantier du CHU de Rouen et Marie Leoz de l'Université de Rouen.

Suite à la pandémie du COVID-19, le premier confinement de deux mois en 2020 a rendu l'avancement de la thèse difficile. La thèse a bénéficié d'un second financement pour une quatrième année, jusqu'à la fin de l'année 2021, dans le cadre du projet PhyEpi financé par la région Occitanie et l'ANR. Ce projet nous a permis d'analyser les données d'incidence de l'épidémie en France et de séquences du virus du SARS-CoV-2.

Les travaux de cette thèse ont donné lieu à deux articles présentés dans les chapitres 3 et 5 qui ont été recommandés après processus de revue par les pairs par *Peer Community In Evolutionary Biology*, et un article en cours de rédaction présenté dans le chapitre 2.

Table des matières

Chapitre 1

État de l'art

1.1 Épidémiologie mathématique des maladies infectieuses

L'épidémiologie est une discipline qui étudie les facteurs impactant la santé des populations, dans l'espace et dans le temps, et qui cherche des moyens d'atténuer cet impact. Les épidémiologistes cherchent donc à répondre à des questions telles que : *quelles sont les populations touchées ou à risque ? est-ce une maladie infectieuse ? où et quand a eu lieu l'émergence de la maladie ? par quelle route de transmission se propage la maladie ?*

1.1.1 Du « mauvais air » au SARS-CoV-2

Jusqu'au 19^{e} siècle prévalait la théorie des miasmes selon laquelle les maladies étaient provoquées par les « mauvais airs », particules qui émanent de la matière organique décomposée (**?**). Cette théorie a été remise en question par le médecin anglais John Snow. Celui-ci émit l'hypothèse qu'une maladie, le choléra, serait due à un poison se reproduisant dans le corps et qui serait retrouvé dans les selles des patients, depuis lesquelles il se propagerait *via* l'eau contaminée. John Snow a pu valider sa théorie au cours d'une épidémie de choléra à Londres en 1854. En collectant les adresses de domicile des malades et des cas de décès, ainsi que leur fournisseur en eau, il constata que la plupart des patients morts se fournissaient auprès d'une compagnie qui puisait son eau en aval de la ville de Londres où étaient déversées les eaux usées (**?**). Bien qu'ignoré à l'époque, John Snow est aujourd'hui considéré comme fondateur de l'épidémiologie moderne.

Quant à la théorie des miasmes, elle fut enterrée une trentaine d'années plus tard avec la découverte de la bactérie fécale responsable du choléra *Vibrio cholerae* par Robert Koch et les travaux en parallèle de Pasteur. Ces travaux ont permis de comprendre et démontrer que de nombreuses maladies sont provoquées par des micro-organismes (ou microbes) comme les bactéries, les virus, les champignons et autres organismes pluricellulaires.

L'étude des maladies dépend des disciplines. Les médecins, par exemple, s'intéressent aux symptômes cliniques d'un patient (par exemple, la fièvre), alors que les microbiologistes s'intéressent aux agents pathogènes biologiques (virus, bactéries, champignons) et que les épidémiologistes cherchent à comprendre les facteurs affectant la santé d'une population et la transmission des maladies.

On distingue généralement les maladies infectieuses et non infectieuses. Les maladies sont dites infectieuses (comme la grippe) lorsqu'elles sont causées par un organisme parasitaire qui se transmet d'un individu à un autre, contrairement aux maladies non infectieuses (comme le cancer). Cependant, la distinction entre les deux est moins simple qu'il n'y paraît. Par exemple, certaines maladies infectieuses, telles que celles provoquées par le virus de l'hépatite C (VHC) ou le papillomavirus humain, peuvent causer des cancers (**?**).

Les maladies infectieuses peuvent elles-mêmes être classées selon l'agent biologique provoquant l'infection, qui peut être soit un micro-parasite (virus, bactérie, champignon), soit un macro-parasite (helminthe ou vers parasite, trématode). Mais on peut aussi les classer selon leur route de transmission. La transmission peut se produire à la fois au sein de différentes espèces hôtes et entre elles. Beaucoup d'infections humaines, appelées zoonoses, sont des infections où les animaux constituent un réservoir d'agents pathogènes qui peuvent être transmis à l'homme. Certaines infections micro-parasitaires se transmettent indirectement *via* l'environnement ou par l'intermédiaire d'un vecteur comme c'est le cas pour la dengue ou le chikungunya qui se transmettent par des moustiques. La majorité des maladies infectieuses micro-parasitaires, telles que la grippe, la maladie provoquée par le virus de l'immunodéficience humaine (VIH) ou la maladie à COVID-19, sont à transmission directe, c'est-à-dire que le pathogène est transmis par contact étroit avec un individu infecté (contact buccal, sexuel, manuel, sanguin). Comprendre la manière dont se propagent les microbes n'est pas une tâche facile, même aujourd'hui. Par exemple, la transmission aéroportée du virus SARS-CoV-2 a été officiellement reconnue par l'Organisation Mondiale de la Santé (OMS) environ un an après le début de la pandémie. Il est important de déterminer le cycle de vie du pathogène causant l'infection ainsi que son mode de transmission pour réagir en santé publique et apporter des réponses adaptées notamment *via* des traitements ou bien *via* des interventions non

pharmaceutiques (appel au port du masque, campagne de promotion du préservatif, etc.).

1.1.2 Modélisation en épidémiologie

La modélisation mathématique a trois principaux rôles : la description, la compréhension et la prévision. L'idée que la transmission et la propagation des maladies infectieuses obéissent à des lois qui peuvent être formulées mathématiquement est ancienne. Le premier modèle épidémiologique a été formulé par Daniel Bernoulli en 1760 (**?**) afin d'évaluer l'impact de l'inoculation de la variole sur l'espérance de vie humaine. Il a fallu attendre le début du 20^{e} siècle pour que la modélisation épidémiologique réapparaisse avec les travaux de W. H. Hamer (1906) sur la rougeole et de Sir R. A. Ross (1911) sur la malaria. Hamer a été l'un des premiers à suggérer que la propagation d'une infection dépend des taux de contact entre individus susceptibles et infectés. Ross a montré qu'une réduction de la population de moustiques en dessous d'un niveau critique serait suffisante pour éliminer le paludisme. **?** ont développé le théorème de seuil, qui prédit la proportion d'individus susceptibles à dépasser pour qu'une épidémie se propage, en formulant un modèle mécaniste simple de transmission et de guérison, le modèle compartimental SIR (Susceptibles-Infectés-Retirés). La diversité des maladies infectieuses étudiées depuis associée à la disponibilité de données de plus en plus précises a suivi le développement de divers modèles épidémiologiques incluant des différences d'âge, de contacts, de sexe ou différentes phases d'infection.

Les données jouent un rôle central dans l'épidémiologie mathématique. Elles sont généralement basées sur les notifications de maladies rapportées par les médecins libéraux ou hospitaliers. Selon Santé Publique France (SPF), 33 maladies sont à déclaration obligatoire en France dont 32 sont infectieuses. Ce dispositif a été mis en place à la fin du 19^{e} siècle. Les données classiquement utilisées en épidémiologie sont les données d'incidence et de prévalence. L'incidence est définie comme le nombre de nouveaux cas par unité de temps et la prévalence correspond au nombre de cas, à un instant donné. L'incidence reflète donc la dynamique de transmission de la maladie alors que la prévalence est davantage liée aux propriétés statiques de la maladie et dépend de la durée d'infection. Ces données peuvent être stratifiées par âge, sexe ou bien géographiquement. D'autres données peuvent être récoltées et utilisées telles que les données de téléphone portable, les données issues des réseaux sociaux ou encore le séquence génétiques.

Modèles déterministes

Dans les modèles de type SIR, les individus sont répartis dans différents compartiments selon leur état clinique, et les compartiments sont liés entre eux par des flux qui décrivent les transitions d'individus entre les compartiments. Ils peuvent être enrichis à l'infini en rajoutant des états ou des transitions. Nous introduisons ici le modèle déterministe SEAIR, plus détaillé que le modèle classique SIR car il comporte notamment une phase d'exposition où les individus sont infectés mais non infectieux et asymptomatiques (dénotés E pour exposés), et une phase où les individus deviennent infectieux et demeurent asymptomatiques (dénotés A). Biologiquement, il capture le cycle de vie d'une maladie infectieuse telle que celle provoquée par le virus SARS-CoV-2.

Nous pouvons représenter graphiquement ce modèle sous forme de diagramme de flux (Figure 1.1).

Les flux de transmission, les flux de transition d'une phase d'infection à une autre et le flux de guérison dépendent chacun d'un taux. Le flux de transmission, déplaçant des individus du compartiment S au compartiment E, dépend du nombre d'individus susceptibles $S(t)$ et de la force d'infection $\lambda(t)$. Le flux de déplacement d'individus exposés E au compartiment A dépend du nombre d'individus $E(t)$ et du taux ε. Le flux de déplacement d'individus du compartiment A au compartiment I dépend du nombre d'individus asymptomatiques $A(t)$ et du taux σ. Le flux de déplacement d'individus du compartiment au compartiment R correspond à une fin d'infection, suite à une mort ou à une guérison. Ce flux dépend du nombre d'individus infectieux symptomatiques $I(t)$ et du taux de fin d'infection γ. La durée moyenne d'infectiosité, période

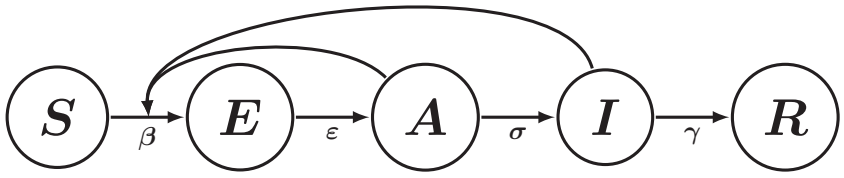


FIGURE 1.1 – **Diagramme de flux du modèle SEAIR.** Les cercles représentent les différents compartiments où sont répartis les individus selon leur état, dénotés S pour susceptibles, E pour exposés, A pour asymptomatiques, I pour infectieux et R pour retirés (par mort ou guérison). Les flèchent représentent les flux de transition des individus d'un compartiment à un autre en-dessous desquelles sont indiqués les taux de chaque transition.

pendant laquelle un individu peut transmettre le pathogène, est $d = 1/(\sigma + \gamma)$. Il est possible d'intégrer une structure démographique avec des taux de natalité et de mortalité qui font varier la taille de la population $N(t)$. Ici, nous supposons que la taille de la population reste constante telle que $N(t) = S(t)+E(t)+A(t)+I(t)+R(t)$.

La force d'infection $\lambda(t)$ exprime le taux avec lequel un individu susceptible devient infecté et varie en fonction du nombre d'individus infectieux $A(t)$ et $I(t)$. Sa valeur dépend du type de transmission. Dans le cas d'une transmission densité-dépendante, par exemple la grippe, l'augmentation de la densité d'individus infectieux entraîne une augmentation du taux de transmission par contact. La force d'infection est alors $\lambda(t) = \beta A(t) + \beta I(t)$. $\beta(t)$ est le taux de transmission et correspond au produit du taux de contact entre les hôtes et de la probabilité de transmission s'il y a contact. Pour modéliser des infections sexuellement transmissibles (IST) telles que celles par le VIH, on fait plutôt une hypothèse de transmission fréquence-dépendante et on obtient alors $\lambda(t) = (\beta A(t) + \beta A(t))/N(t)$. En effet on suppose que le taux de contact reste fixe quelque soit la taille de la population et que le nombre de contacts sexuels n'augmentent pas plus du fait que la population d'hôtes augmente et dépend plutôt des caractéristiques épidémiologiques et sociales de la population. Il est important de définir si la taille de population $N(t)$ varie au cours du temps. Si ce n'est pas le cas, le terme $1/N$ est absorbé dans le taux de transmission et l'expression de la force d'infection devient identique à celle de la densité-dépendante.

Le modèle SEAIR sans démographie peut être décrit par le système d'équations différentielles suivant, exprimant la dynamique des densités d'individus de chaque compartiment en fonction du temps :

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = -\beta\ S(t)\ A(t) - \beta\ S(t)\ I(t) \tag{1.1a}$$

$$\frac{\mathrm{d}E(t)}{\mathrm{d}t} = \beta\ S(t)\ A(t) + \beta\ S(t)\ I(t) - \epsilon\ E(t) \tag{1.1b}$$

$$\frac{\mathrm{d}A(t)}{\mathrm{d}t} = \epsilon\ E(t) - \sigma\ A(t) \tag{1.1c}$$

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = \sigma\ A(t) - \gamma\ I(t) \tag{1.1d}$$

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t} = \gamma\ I(t) \tag{1.1e}$$

Un des concepts clés en épidémiologie est le nombre de reproduction de base, le R_0 (**?**). Il est défini comme le nombre moyen d'infections secondaires engendrées par un hôte infecté dans une population entièrement susceptible. Si on introduit un petit nombre d'infectés n au sein d'une population entièrement susceptible, en moyenne $R_0 n$ nouvelles infections seront engendrées et $R_0^k n$ infections seront engendrées au bout de k générations de transmission. Ainsi, dans un modèle déterministe, si $R_0 < 1$, le nombre moyen d'infection secondaire est inférieur à un et l'infection ne peut envahir la population ; si $R_0 > 1$, l'infection se propage et devient épidémique.

La valeur du R_0 dépend du parasite causant l'infection mais aussi de la population hôte. Selon le modèle mathématique, R_0 aura des expressions différentes. Dans le cas de modèles simples, à un compartiment infecté (SI, SIS, SIR) on trouve facilement une expression du R_0 en considérant que le population est à l'état d'équilibre sans infections, et ainsi en résolvant l'équation $dI/dt > 0$. Lorsque le modèle contient plusieurs compartiments représentant des individus infectés, d'autres méthodes sont nécessaires, telle que la méthode *next-generation matrix* (« matrice de génération suivante ») (**?**).

Le modèle SEAIR possède un équilibre sans infections où $S_0 = N$, $E_0 = 0$, $A_0 = 0$, $I_0 = 0$ et $R_0 = 0$. La stabilité de cet équilibre détermine si l'épidémie peut se propager ou non. Cette stabilité est déterminée par le signe de la partie réelle de la valeur propre dominante de la matrice Jacobienne du système d'équations différentielles 1.1. Cette matrice est obtenue en dérivant chacune des dérivées du système d'équations différentielles concernant les infectés (dE/dt, dA/dt et dI/dt), par rapport à la densité de chaque compartiment infecté (E, A et I). La matrice jacobienne J du système d'équations différentielles 1.1 au point d'équilibre est :

$$J = \begin{bmatrix} -\epsilon & \beta S_0 & \beta S_0 \\ \epsilon & -\sigma & 0 \\ 0 & \sigma & -\gamma \end{bmatrix}$$

L'idée sous-jacente du théorème de next-generation est que la matrice jacobienne J peut se décomposer en deux matrices, l'une des naissances (correspondant aux entrées dans les compartiments d'infectés) F et l'une des morts (correspondant aux sorties des compartiments d'infectés) V, tel que $J = F - V$ avec $F \geq 0$ et $V^{-1} \geq 0$. Ainsi nous définissons ces deux matrices :

$$F = \begin{bmatrix} 0 & \beta S_0 & \beta S_0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$V = \begin{bmatrix} -\epsilon & 0 & 0 \\ \epsilon & -\sigma & 0 \\ 0 & \sigma & -\gamma \end{bmatrix}$$

Si la valeur propre dominante de $-V$ est strictement négative, alors le R_0 est donné par le module (la valeur absolue) de la valeur propre dominante de la matrice $F.V^{-1}$ (**?**). S'il est supérieur à 1 alors l'état d'équilibre sans infections est instable et l'infection peut se propager. Le R_0 obtenu est :

$$R_0 = \frac{\beta\ S_0\ (\gamma + \sigma)}{\gamma\ \sigma} \tag{1.2}$$

La formulation du R_0 varie selon le modèle et peut fournir différentes informations aux épidémiologistes afin d'adapter le type et le niveau d'intensité des mesures de santé publique à mettre en place. Par exemple, nous pouvons déduire par cette expression du R_0 que plus la durée d'infectiosité est longue, donnée par les termes $1/\sigma$ et $1/\gamma$ plus le R_0 est grand.

Le nombre de reproduction effectif $R(t)$ est un moyen de mesurer l'efficacité des mesures de contrôle. Il correspond au nombre moyen d'infections secondaires engendrées par un individu infecté dans une population de susceptibles dont la taille varie au cours du temps. Dans un cas sans démographie, ce nombre vaut :

$$R(t) = R_0 \times \frac{S(t)}{N} \tag{1.3}$$

Modèles stochastiques

La stochasticité peut être présentée sous deux formes : la stochasticité d'observation et la stochasticité du processus. Cette dernière présente deux composantes que sont la stochasticité environnementale et la stochasticité démographique. La stochasticité environnementale reflète l'impact de perturbations environnementales, telles que les variations temporelles de température, sur la dynamique de population, tandis que la stochasticité démographique est liée à l'imprévisibilité du comportement de chaque individu qui a une conséquence au niveau de la population et dépend donc de la taille de la population.

L'aléa joue un rôle important notamment en début d'épidémie, lorsque le nombre d'individus infectés est encore petit. C'est cette part d'aléa qui détermine si une épidémie va se propager ou s'éteindre. Il est donc important de prendre en compte la stochasticité dans les modèles épidémiologiques.

La manière la plus simple d'introduire de la stochasticité en modélisation est d'ajouter un bruit d'observation. On fait alors l'hypothèse que les dynamiques épidémiques sous-jacentes sont déterministes, mais qu'il existe une certaine incertitude dans les données observées et enregistrées. En effet, certaines maladies infectieuses qui sont en grande partie asymptomatiques peuvent faire l'objet d'une incertitude dans leur déclaration. Dans ces modèles, le nombre d'individus infectés observés correspond à la somme de deux distributions binomiales :

$$I_{obs} = \text{Bin}(I, p_{VP}) + \text{Bin}(N - I, p_{FP}) \tag{1.4}$$

où p_{VP} correspond à la probabilité qu'un cas a été correctement identifié (Vrai Positif) et p_{FP} correspond à la probabilité qu'un individu sain ait été identifié comme étant infecté (Faux Positif), I est le nombre d'individus infectés et N est la taille de la population totale. Cependant, l'introduction de bruit d'observation n'a aucun impact sur les dynamiques épidémiologiques.

Une autre manière d'intégrer la stochasticité au modèle est de modifier directement le système d'équations différentielles (Bjørnstad et al, 2002). À chaque pas de temps la dynamique est soumise à une certaine variabilité aléatoire, qui est propagée dans le temps par le système d'équations différentielles. Par exemple, si on intègre le bruit dans les termes de transmission, le système 1.1 devient :

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = -[\beta\ S(t)\ A(t) + \xi(t)\ f(S(t), A(t))] - [\beta\ S(t)\ I(t) + \xi(t)\ f(S(t), I(t))] \tag{1.5a}$$

$$\frac{\mathrm{d}E(t)}{\mathrm{d}t} = [\beta\ S(t)\ A(t) + \xi(t)\ f(S(t), A(t))] + [\beta\ S(t)\ I(t) + \xi(t)\ f(S(t), I(t))] - \epsilon\ E(t) \tag{1.5b}$$

$$\frac{\mathrm{d}A(t)}{\mathrm{d}t} = \epsilon\ E(t) - \sigma\ A(t) \tag{1.5c}$$

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = \sigma\ A(t) - \gamma\ I(t) \tag{1.5d}$$

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t} = \gamma\ I(t) \tag{1.5e}$$

où ξ est une série temporelle de variables aléatoires suivant une loi normale centrée-réduite (donnant $\xi(t)$ variable aléatoire au temps t), $f(S(t), A(t))$ et $f(S(t), I(t))$ sont des fonctions pour mettre à l'échelle l'aléa en fonction des densités $S(t)$, $A(t)$ et $I(t)$. Ainsi, en ajoutant du bruit, la dynamique épidémiologique s'écarte davantage de l'équilibre déterministe et présente une composante oscillatoire autour de cet

TABLE 1.1 – Événements épidémiologiques possibles du modèle, taux auquel se produit chaque événement et la réaction associée.

Événement	**Réaction**	**Taux**
Transmission par A	$S \to E$	βSA
Transmission par I	$S \to E$	βSI
Transition état E à A	$E \to A$	ϵE
Transition état A à I	$A \to I$	σA
Fin d'infection	$I \to R$	γI

équilibre dont l'amplitude est déterminée par $f(S(t), A(t))$ et $f(S(t), I(t))$. Le désavantage de cette méthode est qu'elle ne prend pas en compte le caractère discret et individuel des populations et ne sont donc pas adaptés lorsque le nombre d'individus infectés est petit.

Les approches événement-centrées permettent d'intégrer facilement le caractère individuel et discret de la stochasticité démographique. Les modèles événement-centrés requièrent d'expliciter tous les événements qui peuvent se produire et leur taux. Dans le modèle SEAIR, cinq événements peuvent se produire, chacun engendrant un changement d'état du système. Ils sont listés dans le Tableau 1.1.

Il existe différents algorithmes de simulation à événements discrets notamment l'algorithme de simulation directe de Gillespie (Gillespie's Direct Method) (**?**) qui est présenté dans l'Algorithme 1. Le formalisme de Gillespie permet une correspondance exacte entre un système d'équations différentielles et la simulation stochastique événement-centrée. Ces modèles font l'hypothèse qu'un type d'événement se produit après un temps d'attente qui suit une loi exponentielle de paramètre égal au taux du processus de Poisson. Le temps d'attente jusqu'à ce que tout type d'événement se réalise suit donc une loi exponentielle de paramètre égal à la somme de tous les taux des événements du modèle, sous l'hypothèse que ces événements sont indépendants.

1.1.3 Inférence de paramètres épidémiologiques à partir de données épidémiques

Les modèles mathématiques sont caractérisés par une certaine combinaison de paramètres, chacun ayant une signification biologique, comme la force d'infection. Ils permettent également de dériver des concepts qui ne sont pas directement perceptibles sur les données, comme le nombre de reproduction. L'estimation de paramètres d'un modèle à partir des données correspond à l'un des principaux objectifs de la

Algorithme 1 Méthode directe de Gillespie

Entrée: Le nombre d'événements n, les événements $E_1, ..., E_n$, leur taux $R_1, ..., R_n$, le temps final de simulation t_f

```
Initialiser le temps t = 0
tant que t < t_final faire
    Calculer le taux total R_total = Σ_{i=1}^{n} R_i
    Tirer deux nombres aléatoires dans une loi uniforme RAND_1 et RAND_2
    Tirer un temps d'attente d_t = (-1 / R_total) log(RAND_1)
    Initialiser R_sum = 0.0
    pour j = 0 à n faire
        R_sum = R_sum + R_j
        si R_sum ≥ RAND_2 × R_total alors
            Indice de l'événement suivant k = j - 1
        fin si
    fin pour
    Réaliser événement E_k et mettre à jour le système
    t = t + d_t
fin tant que
```

modélisation, la description qui a un rôle important, en particulier lorsque ces paramètres sont difficiles à mesurer directement à partir des données.

Inférence par maximum de vraisemblance

Les approches d'inférence par maximum de vraisemblance consistent à formuler la fonction de vraisemblance $\mathcal{L}(\theta)$ qui est définie comme la probabilité d'observer un ensemble de données Y sous un modèle $\mathcal{M}$ aux paramètres θ :

$$\mathcal{L}(\theta) = Pr(Y|\theta) \tag{1.6}$$

Prenons l'exemple d'une série n observations $(y_1, ..., y_i, y_n)$ supposées indépendantes. La vraisemblance est le produit des vraisemblances de chaque observation :

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} Pr(y_i|\theta) \tag{1.7}$$

Le maximum de cette fonction de vraisemblance est obtenu lorsque la dérivée de la fonction est nulle. Il est souvent plus pratique d'utiliser le logarithme de cette fonction de vraisemblance, le produit se transformant ainsi en somme, ce qui est plus

simple à dériver. En effet, la vraisemblance et la log-vraisemblance atteignent leur maximum au même point. Ainsi la fonction de vraisemblance devient :

$$\log(\mathcal{L}(\theta)) = \sum_{i=1}^{n} \log(Pr(y_i|\theta)) \tag{1.8}$$

Inférer θ par maximum de vraisemblance consiste donc à trouver l'estimateur θ^* qui maximise cette fonction de vraisemblance.

Approche bayésienne

L'inférence bayésienne repose sur le théorème de Bayes appliqué aux probabilités conditionnelles :

$$p(A|B) = \frac{P(B|A)\ P(A)}{P(B)} \tag{1.9}$$

Dans le cadre de l'inférence de paramètre(s) θ à partir de données D on obtient :

$$P(\theta|D) = \frac{P(D|\theta)\ P(\theta)}{P(D)} \tag{1.10}$$

Ici, les paramètres θ sont considérés comme des variables aléatoires décrites par une distribution de probabilité $P(\theta)$. $P(\theta)$ est la distribution *a priori* qui reflète les connaissances sur les valeurs possibles de θ, $P(\theta|D)$ est la distribution *a posteriori* des paramètres θ sachant les données observées D, $P(D|\theta)$ est la vraisemblance d'observer D sous θ et $P(D)$ la vraisemblance marginale. Le calcul numérique de la vraisemblance marginale $P(D)$ peut être difficile, notamment pour des modèles composés de beaucoup de paramètres. Dans ces cas-là, on a recours en général à des algorithmes basés sur des simulations d'échantillonnage, notamment les algorithmes de Monte-Carlo par chaînes de Markov (dits *Markov chain Monte Carlo* ou MCMC).

Calcul bayésien approché

Pour des modèles détaillés, composés de beaucoup de paramètres, il peut être difficile d'exprimer la fonction de vraisemblance $P(D|\theta)$. De plus, même s'il est possible de l'exprimer, la calculer numériquement et/ou calculer le maximum de vraisemblance peuvent être impossibles pour de grands jeux de données. Une solution est d'approcher la vraisemblance par simulation à l'aide de la méthode de calcul bayésien approché (ou ABC pour *Approximate Bayesian Computation*). La distribution *a posteriori* des paramètres θ est construite par méthode d'acceptation/de rejet basée

sur la comparaison de données simulées et donnée observée au travers de statistiques de résumé.

L'algorithme le plus simple de l'ABC présente plusieurs étapes. La première étape est de tirer aléatoirement un ensemble de paramètres θ à partir de la distribution *a priori*. La deuxième étape est de simuler un jeu de données D_S à partir du modèle et de θ. Les statistiques de résumé des données simulées ($S(D_S)$ et celles des données observées $S(D_O)$ sont ensuite calculées afin d'être comparées selon une mesure de distance telle que la distance Euclidienne. L'acceptation/rejet d'une donnée simulée dépend du paramètre de seuil de tolérance ϵ. Une donnée simulée $D_{S,i}$ à partir de θ_i est acceptée si la distance $d(D_{S,i}, D_O) \leq \epsilon$. Suite à l'acceptation ou au rejet de données simulées et donc de valeurs de paramètres, on obtient une distribution *a posteriori*.

En 2002, un article de **?** a par la suite introduit deux améliorations aux méthodes d'inférence par rejet existantes : la pondération des données simulées selon une fonction de noyau à partir de la distance $D(S(D_S), S(D_O))$ et l'ajustement de la distribution *a posteriori* par régression linéaire. L'algorithme se présente tel que décrit dans l'Algorithme 2.

Algorithme 2 ABC-régression

1: Calculer $S(D_O)$
2: Tirer aléatoirement n valeurs de paramètres θ_i (avec $i = 1...n$) à partir de la distribution *a priori* Θ
3: Simuler une donnée $D_{S,i}$ à partir du modèle pour chaque valeur de paramètre θ_i (avec $i = i...n$)
4: Calculer $S(D_{S,i})$, le vecteur de statistiques de résumé pour chaque donnée simulée $D_{S,i}$ (avec $i = 1...n$)
5: Associer à chaque paire $S(D_{S,i}, \theta_i)$ un poids $w_i = 0$ si la distance $d(S(D_{S,i}), S(D_O))$ n'appartient pas au quantile P_ϵ de la distribution empirique des distances $d(S(D_S), S(D_O))$, et un poids défini par une fonction de noyau statistique de la distance $d(S(D_{S,i}), S(D_O))$ sinon
6: Construire un modèle de régression linéaire $\hat{m}(S(D_S)) = \theta$ à partir des paires pondérées $S(D_{S,i}, \theta_i)$
7: Calculer la distribution *a posteriori* à partir du modèle de régression telle que $\theta_i^* = \theta_i + \hat{m}(S(D_O)) - \hat{m}(S(D_O))$

Dans cet algorithme, $P_\epsilon \in [0; 1]$ correspond à la valeur du quantile de la distribution empirique des distances $d(S(D_{S,i}), S(D_O))$ telle que, par exemple, pour un échantillon de $n = 100$ données simulées et $P_\epsilon = 0.1$, les $n \times P_\epsilon = 10$ valeurs de

distances les plus petites auront un poids non-nul tandis que les autres auront un poids nul.

L'étape 6 de construction d'un modèle de régression de l'algorithme 2 consiste à modéliser une relation, ici linéaire, entre les statistiques de résumé des données simulées acceptées et les valeurs de paramètres à partir desquelles ont été simulées les données.

La méthode ABC est applicable pour des modèles complexes et a un coût, en termes de temps et de calcul, équivalent, voire plus faible que les méthodes basées sur l'expression de la vraisemblance. Cependant certaines limitations peuvent être rencontrées. Comme pour toute méthode bayésienne classique, le choix de la distribution *a priori* est important car l'inférence du paramètre en dépendra directement. Le choix des statistiques de résumé est aussi important dans cette approche. En effet, remplacer la totalité des données par leurs statistiques résumées entraîne nécessairement une certaine perte d'information, et la distribution *a posteriori* du paramètre obtenue peut être faussée si les statistiques sont mal construites. La méthode ABC est utilisée pour l'inférence de paramètres d'un modèle mais aussi pour la comparaison et sélection de modèles.

Exemples de méthodes d'inférence de R_0

Les approches d'inférence de paramètres épidémiologiques tel que le R_0 à partir de données d'incidence sont nombreuses. Nous présentons ici deux méthodes d'inférence de R_0 à partir de données d'incidence et de l'intervalle sériel, ainsi qu'un exemple d'inférence du R_0 par ABC à partir de données d'incidence et d'un modèle épidémiologique.

Le temps de génération correspond à la différence entre la date de transmission vers un premier individu et la date de transmission de ce premier individu vers un autre individu. Cependant, cette information est souvent difficile à obtenir. Les épidémiologistes font en général l'hypothèse que le temps de génération est équivalent à l'intervalle sériel, qui correspond à la différence entre la date d'apparition de symptômes chez un individu et la date d'apparition de symptômes chez un autre individu infecté par le premier individu. En effet, les patients se souviennent plus facilement de la date d'apparition de symptômes ainsi que des informations concernant leurs contacts récents. La récolte de ces informations permet ainsi de construire la distribution empirique du temps de génération. Cependant, cette hypothèse est une simplification et le temps de génération approché par l'intervalle sériel peut prendre des valeurs négatives lorsque l'individu infecté par le premier individu présente des symptômes avant lui. L'estimation de l'intervalle sériel est nécessaire pour l'infé-

rence du R_0 qui en pratique ne peut pas être directement estimé à partir des données d'incidence.

La méthode, proposée par **?**, et implémentée dans le package R `R0` (**?**) repose sur l'hypothèse que, pendant la période de croissance exponentielle d'une épidémie, le nombre d'infections secondaires causées par un cas suit une loi de Poisson d'espérance de valeur R. Étant donné l'observation de $(N_0, N_1, ..., N_T)$ cas incidents sur des unités de temps consécutives jusqu'au temps T, et une distribution du temps de génération w obtenu à partir d'un intervalle sériel, R est estimé en maximisant la log-vraisemblance :

$$\log\mathcal{L}(R) = \sum_{t=1}^{T} \log\left(\frac{e^{-\lambda_t}\ \lambda_t^{N_t}}{N_t!}\right) \qquad \text{avec} \quad \lambda_t = R\sum_{i=1}^{t} N_{t-i}\ w_i$$

Une autre méthode, bayésienne, introduite par **?**, et aussi implémentée dans le package `R0` (**?**), permet l'estimation séquentielle du nombre de reproduction. Il repose sur une approximation du modèle SIR, où l'incidence au temps $t+1$, notée $N(t+1)$, suit une loi de Poisson de moyenne $N(t)\ e^{(\gamma\ (R-1))}$, avec $1/\gamma$ la durée moyenne de la période infectieuse. L'algorithme proposé, décrit dans un cadre bayésien, commence avec un *a priori* sur la distribution du nombre de reproduction R. La distribution est mise à jour au fur et à mesure que de nouvelles données sont observées, en utilisant :

$$P(R|N_0, ..., N_{t+1}) = \frac{P(N_{t+1}|R, N_0, ..., N_t)\ P(R|N_0, ..., N_t)}{P(N_0, ..., N_{t+1})}$$

Ainsi, la distribution *a priori* de R utilisée à chaque nouveau jour correspond à la distribution *a posteriori* du jour précédent. Comme précédemment, la méthode exige que l'épidémie se trouve dans une période de croissance exponentielle.

L'approche d'inférence par ABC est encore peu utilisée en épidémiologie pour des modèles simples car on peut souvent exprimer une fonction de vraisemblance. Mais il existe des exceptions notamment dans une étude sur les données d'épidémie de la grippe en Australie (**?**). Ces données comportaient le nombre de cas confirmés ainsi que des données récoltées par les médecins telles que le nombre de consultations. Les auteurs ont développé un modèle épidémiologique SEIOR, avec un compartiment O (pour Observés) dans le cas où les infectés consultent leur médecin pour recevoir un traitement. Le modèle étant composé de nombreux paramètres, les auteurs ont choisi d'utiliser l'approche par ABC. Ils ont simulé des trajectoires à partir de ce modèle et ont calculé des statistiques résumé sur les données de surveillance et les trajectoires simulées, telles que le nombre moyen de cas par saison et le nombre de

cas par semaine. Ainsi, en combinant ces données avec un modèle épidémiologique ils ont pu estimer par approche ABC le R_0 de l'épidémie.

1.2 Phylogénies des infections

1.2.1 Terminologie

Une phylogénie ou arbre phylogénétique est un objet qui représente les liens de parenté entre des entités telles que des espèces, des populations, des individus ou des gènes (**?**). Dans une phylogénie, les nœuds internes représentent les ancêtres hypothétiques inférés des entités qui sont représentées par des nœuds externes, ou « feuilles ». Dans les phylogénies enracinées, le nœud qui représente l'ancêtre hypothétique de l'ensemble des entités est appelé « racine ». Les relations de parenté sont matérialisées par les branches qui lient les nœuds. La structure des branchements définit la topologie de l'arbre. Un groupe monophylétique ou clade ou *cluster* comprend tous les descendants d'un ancêtre hypothétique et l'ancêtre lui-même.

Un nœud ne met pas toujours en relation trois liens. En effet, certaines feuilles peuvent être connectées à plus de trois liens auquel cas nous parlerons de « multifurcations » par opposition aux bifurcations classiques. Ces multifurcations peuvent provenir de deux causes. La première est due à la méthodologie. Certains jeux de données ou certaines méthodes ne parviennent pas à départager les différentes hypothèses de parenté possibles. Dans ce cas, les relations ne sont pas résolues et cette irrésolution est représentée par une multifurcation. La deuxième raison est biologique et provient d'une irrésolution temporelle. Si un groupe d'entités ancestral se diversifie très rapidement, il est possible que celui-ci soit divisé en plus de deux ensembles.

1.2.2 Liens avec les chaînes de transmission

L'avènement des techniques de séquençage de nouvelle génération a conduit à la génération en grande quantité de données génétiques presque en temps réel. **?** ont pu générer 142 séquences génomiques du virus ébola *via* l'emploi d'un séquenceur portable et ce malgré la crise sanitaire en Guinée. Un exemple plus récent est celui de la génération et la disponibilité de données de séquences de génomes, en masse, du virus SARS-CoV-2 depuis le début de l'épidémie.

L'abondance des données de séquences de micro-organismes associée à leur évolution rapide, en particulier des virus à ARN, a conduit à l'émergence de phylogénies d'infections qui sont construites à partir de séquences récoltées chez différents patients sur de courtes échelles de temps épidémiologiques. Leurs feuilles représentent

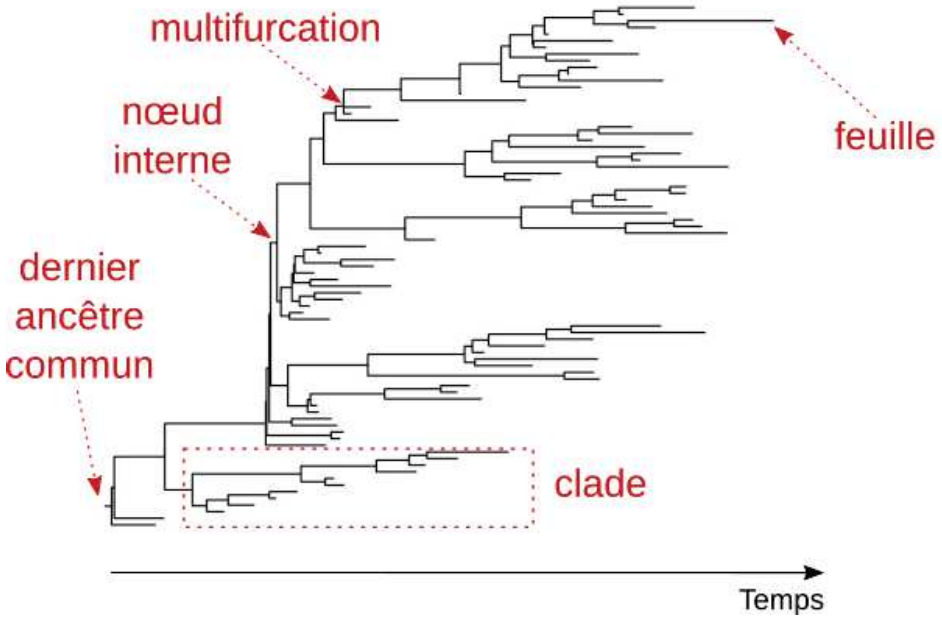


FIGURE 1.2 – **Illustration de la structure d'une phylogénie.** Sur cette figure le temps s'écoule de gauche à droite et les lignes verticales n'ont aucune signification biologique.

des infections, et non pas des espèces comme dans les arbres phylogénétiques classiques. Ainsi, ces phylogénies d'infections permettent de mettre en évidence des liens entre différentes infections par un pathogène, liens qui reflètent des chaînes de transmission entre individus. Un exemple célèbre est l'utilisation, dans un cadre médico-légal, d'une phylogénie du VIH confirmant des contaminations provenant d'un dentiste (**?**).

Les phylogénies d'infections, comme toutes les phylogénies, sont construites à partir d'alignement de séquences par diverses approches d'inférence statistique, sous des hypothèses telles que celles faites sur le modèle d'évolution des séquences et la topologie de l'arbre. Les séquences sont généralement datées, permettant par exemple de dater l'origine d'une épidémie.

1.2.3 Alignement des séquences

Lorsque l'on construit une phylogénie à partir de séquences, on fait l'hypothèse que celles-ci sont homologues, c'est-à-dire qu'elles partagent un ancêtre commun. L'alignement consiste à identifier, pour chaque séquence, les caractères (nucléotides ou acides aminés) homologues et à les positionner en regard. L'alignement de séquences est une étape clé pour toutes les approches d'inférence phylogénétique.

La Figure 1.3 illustre un alignement réalisé à partir de trois séquences nucléotidiques de longueur différente. Un alignement optimal arrange deux séquences ou plus de manière à ce qu'un nombre maximal de caractères, des nucléotides ou des acides aminés, identiques ou similaires soient mis en correspondance en colonne. Une colonne d'un alignement est appelée « site ». Ce processus de réarrangement peut se faire par l'introduction d'un ou plusieurs espaces, appelés *gaps*, représentés par des tirets dans l'alignement. Un *gap* indique une perte ou un gain possible d'un caractère. Ainsi, les alignements de séquences peuvent mettre en évidence des événements d'insertion ou de délétion évolutive, regroupés dans le terme *indels*. L'introduction des *gaps* dans un alignement doit être faite avec parcimonie. Un bon alignement contient le moins d'événements de mutation possibles, pondérés selon l'événement mutationnel (substitution, insertion, délétion, prolongation des *gaps*, etc.). Notons qu'une substitution désigne une mutation, événement lié aux processus biochimiques, qui s'est fixée dans le génome et donc dans la population.

Le nombre d'alignements possibles est important et tous ne sont pas de qualité équivalente. Le meilleur alignement peut être identifié à l'aide d'une fonction de score. Dans la notation nucléotidique, les paires de nucléotides identiques se voient attribuer un score positif et les *gaps*, des scores négatifs, selon une matrice de substitution et un modèle de pénalité associée aux *gaps*.

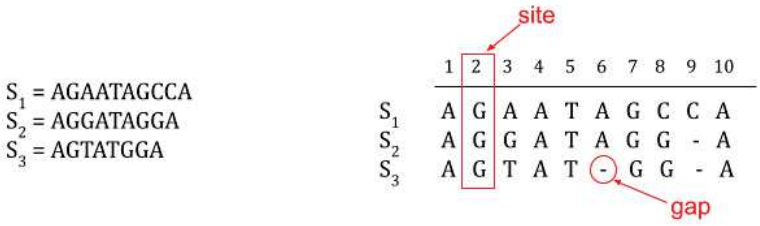


FIGURE 1.3 – **Exemple d'un alignement possible (à droite) réalisé à partir de trois séquences de longueurs différentes (à gauche).** Exemple extrait de **?**.

Il existe deux types d'alignements : l'alignement de séquences par paires, qui ne considère que deux séquences à la fois, et l'alignement de séquences multiples, qui aligne l'ensemble des séquences simultanément. Les méthodes d'alignement par paires sont les plus simples à mettre en œuvre et sont en général utilisées pour rechercher une homologie entre une séquence test et une séquence de référence, souvent extraite d'une base de données. L'alignement multiple est plus avantageux car il prend en compte plusieurs membres d'une famille de séquences et fournit ainsi plus d'informations biologiques. Ce type d'alignement est également un préalable aux analyses génomiques comparatives pour l'identification et la quantification de régions conservées ou de motifs fonctionnels au sein d'une famille de séquences, ainsi que pour l'estimation de la divergence évolutive entre séquences.

Pour l'alignement par paires, la programmation dynamique fournit toujours l'alignement optimal (c'est-à-dire celui qui a le meilleur score) en notant toutes les paires possibles de caractères alignés et en pénalisant les gaps. Cependant, cette méthode étant coûteuse en termes de calcul et de mémoire, elle est rarement utilisée pour plus de quelques séquences. La plupart des approches d'alignement multiple sont donc heuristiques, fournissant une solution d'alignement réalisable dans un laps de temps court et limité. Ces méthodes sont implémentées dans de nombreux logiciels dont les plus connus sont CLUSTAL (**??**), T-Coffee (**?**), MUSCLE (**?**) ou encore MAFFT (**??**). Ces deux derniers logiciels sont les plus précis et les plus rapides. Ils sont donc plébiscités pour aligner des gros jeux de données de séquences.

Cependant, ces approches d'alignement de séquences présentent certaines limites. La première est qu'elles supposent que les séquences sont colinéaires, c'est-à-dire qu'elles conservent un ordre ancestral commun d'éléments, que ce soient des nucléotides, des acides aminés ou des gènes, selon les séquences comparées. Cette colinéarité n'est pourtant pas toujours observée, notamment au sein des génomes viraux qui présentent une grande variation dans le nombre et l'ordre des éléments génétiques en raison de leurs taux de mutation élevés, des événements de recombinaison génétique fréquents, des transferts horizontaux de gènes, des duplications de gènes et des gains/pertes de gènes (**?**). Un autre exemple se présente au sein des génomes bactériens où l'ordre des gènes a tendance à être moins conservé. Une autre limite est qu'un alignement de séquences dépend d'hypothèses concernant l'évolution des séquences *via* les matrices de substitution et le modèle de pénalisation de *gap*. Les paramètres associés à ces hypothèses sont relativement arbitraires dans le sens où chaque méthode a une fonction de score spécifique et un faible changement de ces paramètres peut affecter l'alignement (**?**). Enfin, une des limites de ces méthodes d'alignement basées sur des approches heuristiques est qu'elles sont approximatives et ne garantissent donc pas l'identification de l'alignement optimal avec le score le

plus élevé. Cela peut entraîner des inexactitudes qui peuvent limiter la qualité de nombreuses analyses en aval, telle que la phylogénétique.

Une solution relativement récente, qui s'affranchit de ces limites, consiste à utiliser des approches de comparaison de séquences sans alignement (**?**). Celles-ci sont nombreuses et calculent des mesures de dissimilarité ou de distances, par paires, entre séquences. Parmi ces méthodes, se trouvent celles basées sur le calcul de distance, comme la distance Euclidienne, entre fréquences de mots ou sous-séquences de longueur k (ou k-mer) des séquences comparées par paires. Cette distance représente une bonne mesure de la dissimilarité entre séquences et est ensuite enregistrée dans une matrice de distance qui peut par la suite être utilisée pour reconstruire une phylogénie (**?**).

La plupart des articles publiés impliquant les méthodes de comparaison de séquences sans alignement sont encore principalement techniques, explorant leurs fondements mathématiques et leurs performances théoriques par rapport aux approches basées sur l'alignement. Ils reposent très souvent sur des données simulées. De plus, les méthodes sans alignement publiées ne sont pas toujours implémentées dans des logiciels dédiés et ne peuvent donc pas être facilement être comparées sur des ensembles de données communs. Par conséquent, il est encore relativement délicat d'indiquer si une méthode sans alignement pourrait être particulièrement adaptée à une certaine analyse.

1.2.4 Inférence phylogénétique

Il existe de nombreuses méthodes d'inférence phylogénétique. Nous présentons ici les méthodes probabilistes basées sur le concept de vraisemblance, qui mettent en jeu la méthode du maximum de vraisemblance et une approche bayésienne. Ces méthodes reposent sur un modèle dont les composantes sont la topologie, qui correspond à l'ordre des branchements, les longueurs de branches associées à la topologie, qui correspond au nombre de substitutions accumulées au cours de l'évolution, et les paramètres du modèle d'évolution des séquences.

Modèles d'évolution des séquences

Les modèles d'évolution permettent de calculer les probabilités des substitutions observées entre les séquences. Ces modèles font l'hypothèse que les processus de substitution sont Markoviens, c'est-à-dire que la probabilité de changement d'un état de caractère i à un état de caractère j dépend uniquement de l'état i. Ils s'appuient aussi sur l'hypothèse que les positions dans le génome, ou sites, évoluent indépendamment les uns des autres.

Le modèle de substitution pionnier et le plus simple est le modèle JC69 proposé par **?** qui émet l'hypothèse que les fréquences en nucléotides A, C, G et T sont égales et que les taux de changement d'un nucléotide vers un autre sont identiques. Ce modèle n'a donc qu'un paramètre. De nombreux modèles ont par la suite été introduits pour augmenter le niveau de réalisme, ajoutant ainsi de nouveaux paramètres. Par exemple, le modèle K2P (**?**) incorpore deux types de substitutions, les transitions (changement d'un nucléotide d'une famille vers un nucléotide de la même famille ; A $\leftrightarrow$ G pour les purines et C $\leftrightarrow$ T pour les pyrimidines) et les transversions (changement d'un nucléotide d'une famille vers un nucléotide d'une autre famille). Il introduit pour cela un nouveau paramètre qui correspond au rapport des transitions et des transversions. Lorsque ce ratio est élevé, les transitions sont plus probables que les transversions. Le modèle HKY85 (**?**), lui, suppose trois types de substitutions (2 classes de transitions et 1 classe de transversions) et des fréquences nucléotidiques inégales. Le modèle le plus général est le modèle GTR (pour *General Time Reversible*) aussi appelé REV (pour *Reversible*) (**?**). Il considère que le changement d'un nucléotide i vers un nucléotide j est égal au changement d'un nucléotide j vers un nucléotide i et introduit six taux de substitution ainsi que des fréquences propres à chaque nucléotide.

Cependant, ces modèles font l'hypothèse que les taux de substitution sont les mêmes pour tous les sites et ne varient pas dans le temps. Cette hétérogénéité des vitesses d'évolution relatives entre sites a été modélisée par une distribution gamma (Γ) de paramètre de forme α, avec $\alpha > 0$ (**?**). Plus α est faible, plus l'hétérogénéité d'un site à l'autre est élevée.

Les méthodes probabilistes sont dépendantes du modèle d'évolution sous-jacent. Il est donc important de choisir des modèles qui s'ajustent le mieux aux données analysées. Des tests statistiques tels que le rapport des vraisemblances (LRT pour *Likelihood Ratio Test*) et le critère d'information d'Akaike (AIC pour *Akaide Information Criterion*) sont utilisés pour comparer les différents modèles. Des programmes de sélection de modèles ont été développés tels que jModelTest (**?**) ou SMS (**?**).

Inférence phylogénétique par maximum de vraisemblance

L'application de la méthode du maximum de vraisemblance en phylogénétique a été introduite par **?** dans les années 1960 et développée par **?**.

Appliquée en phylogénétique, la fonction de vraisemblance est la probabilité conditionnelle d'observer des données X, ici un alignement de séquences, étant donné un modèle $\mathcal{M}$ décrit par la topologie d'un arbre τ, des longueurs de branches ν et les paramètres θ d'un modèle décrivant l'évolution des séquences le long des branches

de l'arbre. On la note

$$\mathcal{L}(\tau, \nu, \theta) = Pr(X|\tau, \nu, \theta) \tag{1.11}$$

L'inférence phylogénétique par maximum de vraisemblance consiste à trouver les estimateurs τ^*, ν^* et θ^* qui maximisent la fonction de vraisemblance, c'est-à-dire

$$(\tau^*, \nu^*, \theta^*) = \text{argmax}_{\tau,\nu,\theta}\mathcal{L}(\tau, \nu, \theta) \tag{1.12}$$

Une recherche exhaustive explorant toutes les combinaisons des arbres possibles implique de calculer la fonction de vraisemblance pour chacun d'entre eux, ce qui devient rapidement très coûteux en temps de calcul. En effet, pour n nombre de séquences représentées par des feuilles, le nombre de phylogénies possibles est $\frac{(2n-3)!}{2^{n-2}\ (n-2)}$ avec $n \geq 2$. Ainsi, pour 10 individus ou feuilles, il existe plus de 34 millions de topologies possibles. Diverses méthodes de recherches heuristiques ont été proposées afin de contrer ce problème. Ces méthodes considèrent d'abord un arbre aléatoire comme point de départ, arbre calculé par des méthodes de parcimonie ou de distance qui sont de moins bonne qualité que les méthodes de vraisemblance mais plus rapides. Puis, à partir de cet arbre, ces méthodes essaient de l'améliorer en explorant le voisinage de cet arbre jusqu'à obtenir le meilleur arbre. L'approche heuristique permet ainsi d'inférer, en un temps raisonnable, un arbre satisfaisant mais sans avoir la possibilité de savoir si celui-ci est optimal. Ces méthodes sont implémentées dans des logiciels tels que PhyML (**?**), IQ-TREE (**?**), RAxML (**?**) et FastTree (**?**). Ce dernier est le plus rapide notamment pour des phylogénies de grande taille mais présente de moins bon résultats d'inférence que les trois autres outils (**??**).

Inférence phylogénétique bayésienne

L'approche bayésienne a été appliquée en phylogénétique à partir des années 1990 (**?**). Dans le cadre de l'inférence phylogénétique, en reprenant les notations précédentes, la distribution *a posteriori* $P(\tau, \nu, \theta|X)$ des paramètres sachant les données observées X (ici un alignement de séquences) peut s'écrire :

$$P(\tau, \nu, \theta|X) = \frac{P(X|\tau, \nu, \theta)\ P(\tau, \nu, \theta)}{P(X)} \tag{1.13}$$

où $P(\tau, \nu, \theta)$ constitue la distribution *a priori* à définir au préalable des paramètres relatifs à la topologie de l'arbre, aux longueurs de branches et au modèle d'évolution des séquences. $P(X|\tau, \nu, \theta)$ est la fonction de vraisemblance et $P(X)$ la probabilité des données.

L'inférence bayésienne cherche à estimer une distribution de probabilité *a posteriori* sur l'ensemble des arbres possibles, par opposition à la méthode par maximum de vraisemblance qui cherche à trouver un arbre, le plus vraisemblable. Lorsque les données et/ou la distribution *a priori* sont informatives, la distribution des probabilités *a posteriori* est généralement concentrée sur un arbre (ou un petit sous-ensemble d'arbres dans un grand espace d'arbres). Dans les autres cas, la probabilité *a posteriori* peut vite se retrouver répartie sur un très grand nombre d'arbres. Il faut alors estimer la distribution *a posteriori* à l'aide de l'échantillonnage de Monte Carlo par chaînes de Markov. La méthode la plus couramment utilisée est l'algorithme de Metropolis-Hastings (**??**). L'algorithme se base sur la construction d'une chaîne de Markov dont chaque pas implique une modification aléatoire de la topologie, des longueurs de branches et des paramètres d'évolution des séquences.

Ces approches sont implémentées dans les logiciels MrBAYES (**?**), BEAST (**?**) et BEAST2 (**?**). Ces deux derniers logiciels sont souvent utilisés dans l'analyse de données de séquences de pathogènes.

1.2.5 Calibration temporelle

Dans le milieu des années 1960, **?** ont émis l'hypothèse que le taux d'évolution d'une protéine donnée est constant dans le temps. Ainsi, la distance génétique, c'est-à-dire le nombre de substitutions, entre séquences provenant d'espèces différentes peut être convertie en temps de divergence entre ces espèces. Cette hypothèse de l'existence d'une « horloge moléculaire » a donné naissance à la datation moléculaire qui permet d'estimer le taux de substitution et la date de chaque ancêtre commun le plus récent (MRCA pour *Most Recent Common Ancestor*) d'une phylogénie.

Les phylogénies d'infections présentent une différence majeure avec les phylogénies d'espèces classiquement utilisées en biologie. Ces dernières sont représentées par des arbres ultramétriques, c'est-à-dire que toutes les feuilles sont contemporaines. Des données fossiles peuvent alors être utilisées comme points de calibrations pour dater l'âge des différents nœuds internes, âges souvent exprimés en millions d'années. Dans le cas des phylogénies d'infections, les séquences de génomes microbiens, en particulier ceux évoluant rapidement tels que les virus à ARN, accumulent généralement des mutations sur des échelles de temps épidémiologiques se comptant en années voire en mois, de telle façon que les différences entre les dates d'échantillonnages ne sont plus du tout négligeables par rapport à la date de la racine de la phylogénie. Les arbres ne sont donc plus ultramétriques. Mais cela implique aussi que les dates d'échantillonnage peuvent être utilisées comme points de calibration.

Bien qu'il puisse y avoir un taux moyen d'évolution relativement constant sur des

échelles de temps épidémiologiques, il peut y avoir une variation des taux d'évolution entre les lignées d'un arbre phylogénétique. Ne pas tenir compte de telles variations peut conduire à des inférences incorrectes des taux et des dates d'évolution. Pour pallier à cela, il existe des approches bayésiennes qui supposent un modèle dit d'horloge moléculaire relâchée (par opposition au modèle classique d'horloge moléculaire stricte) où les taux d'évolution varient d'une branche à l'autre de l'arbre phylogénétique (**?**).

Les approches de datation moléculaire basées sur ces deux modèles d'horloge moléculaire ont donné naissance à de nombreuses méthodes. Certaines méthodes sont basées sur le maximum de vraisemblance (méthode implémentée dans le logiciel TreeTime (**?**)), sur la datation par moindres carrées (logiciel LSD (**?**) et package R treedater (**?**)) ou sur l'inférence bayésienne (logiciels BEAST (**?**) et BEAST2 (**?**)). Les deux derniers outils, par opposition aux deux premiers cités, autorisent un modèle d'horloge moléculaire relâchée.

Cependant, avant d'utiliser un modèle d'horloge moléculaire pour calibrer un arbre dans le temps à partir de séquences hétérochrones, c'est-à-dire structuées dans le temps, il est conseillé de vérifier que les séquences étudiées contiennent bien un signal phylogénétique suffisant pour une estimation fiable. Autrement dit, il faut qu'il y ait suffisamment de variations génétiques entre les dates d'échantillonnage pour reconstruire un lien statistique entre la divergence génétique et le temps. En effet, si la fenêtre d'échantillonnage n'est pas suffisamment large, ou si le taux d'évolution est trop faible, ou si les séquences génomiques ne sont pas suffisamment longues, il se peut que le nombre de substitutions soit trop limité. Dans un tel cas où le signal phylogénétique est trop limité, il est préférable de fixer la valeur du taux de substitution.

L'exploration du degré de signal phylogénétique peut être réalisé à l'aide d'une approche simple basée sur la régression linéaire entre les dates d'échantillonnage des feuilles et leurs distances à la racine mesurées en nombre de substitutions. Cette approche est implémentée dans le logiciel TempEst (**?**).

Une autre méthode d'exploration du signal est de construire différentes phylogénies à partir de jeux de données où les dates d'échantillonnage sont randomisées, afin de brouiller la structure temporelle, et d'estimer ensuite le taux de substitution. Si la différence entre le taux de substitution obtenu à partir de la phylogénie réelle et ceux obtenus à partir des phylogénies randomisées est significative alors il existe un signal phylogénétique (**?**).

1.3 Inférence phylodynamique

1.3.1 Phylodynamique

L'utilisation des arbres phylogénétiques en épidémiologie moléculaire est apparue vers la fin des années 1990. Des chercheurs travaillant sur la phylogénétique d'agents pathogènes infectieux, comme le VIH, ont constaté que des variations de tailles de population des microbes affectent la topologie et les longueurs de branches d'une phylogénie d'individus de ces populations permettant ainsi d'identifier des périodes de croissance ou de décroissance de la taille de ces populations (**?**). Puisque les phylogénies et les longueurs de branches peuvent être inférées à partir de séquences, on peut émettre l'hypothèse que les séquences peuvent contenir de l'information sur l'histoire démographique de populations de pathogènes (**??**).

L'étude simultanée des processus épidémiologiques, immunologiques et évolutifs d'un pathogène infectieux à travers les phylogénies est décrite comme la phylodynamique. Ce terme est apparu pour la première fois dans une revue de **?**. Ce champ est en plein essor, notamment grâce au progrès des techniques de séquençage et des puissances de calcul permettant le développement de nouveaux modèles.

La plupart des méthodes d'inférence phylodynamique sont basées sur des modèles de dynamique de populations tels que le modèle de coalescent ou le modèle de naissance et de mort et utilisent une approche d'inférence bayésienne reposant sur les chaînes MCMC. Les implémentations les plus populaires sont sans contestes celles dans les logiciels BEAST (**?**) et BEAST2 (**?**). Dans le contexte phylodynamique, la fonction de distribution *a posteriori* des paramètres de l'arbre phylogénétique T, des paramètres liés au modèle épidémiologique η et des paramètres liés au modèle d'évolution θ sachant les données D est (**?**) :

$$f(T, \eta, \theta|D) = \frac{f(D|T, \theta)\ f(T|\eta)\ f(\eta, \theta)}{f(D)} \tag{1.14}$$

où $f(D|T, \theta)$ est la vraisemblance phylogénétique de l'équation 1.11, $f(T|\eta)$ le modèle de transmission (modèle de coalescent ou modèle de naissance et de mort), $f(\eta, \theta)$ la distribution *a priori* et $f(D)$ un terme de normalisation.

1.3.2 Méthodes basées sur la théorie du coalescent

Les premiers modèles appliqués à la phylodynamique se sont basés sur la théorie du coalescent de **?**, qui décrit le lien entre généalogies et histoires démographiques de populations. Les généalogies, en génétique de populations, répondent à la question *qui*

descend de qui ?, tandis que les phylogénies informent sur la proximité génétique entre entités. Le modèle de coalescent est une approximation du modèle de Wright-Fisher (**??**). Ce dernier représente l'évolution, à des pas de temps discrets, d'une population, où chaque individu « choisit » au hasard un parent de la génération précédente. Les hypothèses du modèle de Wright-Fisher sont que la population est finie et de taille constante, que la reproduction est un processus aléatoire et qu'aucun processus de sélection ou de recombinaison n'est autorisé. L'approximation additionnelle du modèle de coalescent est que la taille de la population est grande et que la taille de l'échantillon est beaucoup plus petite. Le modèle de coalescent décrit ainsi la dynamique d'une population en remontant dans le temps à travers des processus de fusion de branches, processus appelés coalescences. Des modèles ont par la suite été introduits décrivant des tailles de population selon des fonctions de croissance exponentielle ou logistique (**???**).

Les premières approches phylodynamiques basées sur les modèles de coalescent ont permis d'inférer des paramètres démographiques tels que la taille de population et les taux de migration à partir de généalogies. Ainsi, **?** ont introduit une approche basée sur la construction graphique de l'accumulation de lignées d'une phylogénie datée au cours du temps (ou LTT pour *Lineage Through Time*) sous la forme d'une fonction constante par morceaux.

? ont eux introduit la méthode de *Skyline Plot* qui a posé les bases pour un calcul plus précis de la reconstruction de l'histoire démographique donnant naissance à une famille de méthodes dites de *Skyline* permettant d'estimer les variations de la taille de la population de pathogènes au cours du temps. Pour reconstruire l'histoire démographique, ces méthodes supposent que la taille moyenne de la population pour chaque intervalle de coalescence peut être estimée par le produit de la taille de l'intervalle de coalescence γ_i, correspondant à la distance temporelle séparant des nœuds (internes ou externes), et de $i(i-2)/2$ où i est le nombre de lignées dans l'intervalle. Des améliorations ont ensuite été implémentées pour inférer les paramètres démographiques par approche bayésienne utilisant les chaînes MCMC. Par exemple, la méthode *Bayesian Skyline Plot* (**?**) estime simultanément le modèle d'évolution, la phylogénie et l'histoire démographique à partir de séquences hétérochrones. Quant à la méthode *Bayesian Skygrid* (**?**), elle reconstruit, à partir de séquences hétérochrones, le graphe de dynamique de population sous la forme d'une fonction constante par morceaux puis le lisse grâce à une fonction de noyau gaussien, permettant à l'utilisateur de définir au préalable les points de changement. Ces deux modèles sont implémentés dans les logiciels BEAST (**?**) et/ou BEAST2 (**?**).

Les modèles de coalescent sont généralement basés sur des modèles démographiques simples, très différents de ceux utilisés en épidémiologie mathématique. Ils

estiment les variations de taille de populations au cours du temps ou des taux de croissance exponentiels. Certes, on peut faire des parallèles avec des paramètres épidémiologique tels que le R_0, mais cela reste très qualitatif. En effet, ces méthodes supposent que le taux de coalescence est inversement proportionnel à la taille de la population efficace $N_e(t)$. En épidémiologie, on fait généralement l'hypothèse que $N_e(t)$ est l'équivalent du nombre d'individus infectés. Cependant, ce lien n'est pas entièrement clair, bien que la valeur de $N_e(t)$ soit supposée être bien plus petite que la taille de population absolue de pathogènes. Des travaux ont suggéré que le taux de coalescence peut être influencé par plusieurs paramètres démographiques, dont la taille et la structure de la population, ou encore par des facteurs génétiques (**?**).

De récents travaux ont permis d'approfondir le lien entre généalogies et dynamiques de populations de pathogènes (**?**). En particulier, **?** ont démontré que le taux de coalescence est proportionnel à l'incidence, qui, pour un modèle SIR, correspond au produit du taux de transmission (β), de la prévalence (nombre d'individus infectés au temps t, $I(t)$) et du nombre d'individus susceptibles au temps t ($S(t)$). En effet, les auteurs émettent l'hypothèse que chaque lignée de la généalogie correspond à un seul hôte infecté et que les événements de coalescence correspondent donc aux événements de transmissions.

Ces travaux ont conduit au développement d'approche appelée *structured coalescent*. Cette approche d'inférence bayésienne est basée sur la simulation de trajectoires à partir de modèle épidémiologique pour ensuite calculer la vraisemblance de la phylogénie datée sous l'hypothèse d'un modèle de coalescent associé aux dynamiques de populations simulées (**??**). Cette méthode est implémentée dans le package PhyDyn (**?**) du logiciel BEAST2 (**?**). La méthode a donné les meilleurs résultats dans une étude de comparaison de méthodes d'inférence phylodynamique, mais a été la plus coûteuse en terme de ressources computationnelles (**?**).

1.3.3 Méthodes basées sur le processus de naissance et de mort

Ces modèles font appel à des processus de branchement caractérisés par des taux constants de naissance et de mort (**?**). Ils sont utilisés pour capturer des événements de spéciation et d'extinction (**?**) ou pour analyser les dynamiques d'une population (**?**). Ils ont aussi été utilisés dans un contexte épidémiologique, par exemple pour estimer des paramètres tels que le coût des mutations d'antibiorésistance pour la tuberculose à partir de données de génotypage (**?**).

On l'a vu, les modèles de coalescent classiques sont très différents des modèles épidémiologiques. L'essor de la phylodynamique a donc poussé à leur chercher des

alternatives, par exemple les modèles de naissance et de mort ou BD (pour *Birth Death*). En effet, dans sa formulation la plus simple, le modèle BD correspond à un modèle épidémiologique SI où les naissances sont des infections et les morts des guérisons. Le modèle BD peut être comparé au modèle de coalescent classique où la taille de la population est constante, lorsque le taux de naissance et le taux de mort sont égaux (**?**).

Une différence importante entre les modèles BD et les modèles de coalescent classique est que dans ces derniers la taille de population varie de manière déterministe alors que dans les modèles BD cette variation est stochastique. **?** ont comparé le modèle BD avec un modèle de coalescent classique (croissance exponentielle) et avec un modèle de coalescent associé à des dynamiques de trajectoires épidémiologiques sous un modèle stochastique SI, équivalant au modèle BD (**??**). Les auteurs ont montré que le modèle de coalescent classique sous-estime ou sur-estime les temps de coalescence selon la valeur du R_0. Cependant, les versions stochastiques du modèles de coalescent permettent de corriger ces biais pour de faibles valeurs de R_0 et de grandes tailles de population.

Une autre différence entre les modèles BD et les modèles de coalescent est que ces derniers font l'hypothèse d'un faible échantillonnage. **?** a introduit une extension du modèle BD qui prend en compte l'échantillonnage incomplet de la population. En effet, les modèles BD ont l'avantage de pouvoir prendre en compte tout type d'échantillonnage (faible ou élevé) en paramétrant directement la probabilité d'échantillonnage (**?**).

Des approches d'inférence bayésienne utilisant les chaînes MCMC ont été développées et implémentées dans le logiciel BEAST2 (**?**). Par exemple, la méthode du package BD skyline plot (**?**) est basée sur le modèle BD avec échantillonnage incomplet et permet d'estimer des variations des paramètres épidémiologiques, tels que le R_0 ou la taux de fin d'infection, au cours du temps selon une fonction par morceaux. Le modèle implémenté dans le package Multi-Type Birth-Death Model ou bdmm (**?**) décrit des populations structurées permettant d'estimer en plus, des taux de migration, ainsi que des taux de transmission entre sous-populations.

1.3.4 Méthode basée sur le calcul bayésien approché

Les méthodes d'inférence phylodynamique bayésienne citées jusuqu'ici reposent sur la formulation de la fonction de vraisemblance $f(T|\eta)$ de l'équation 1.11. Cependant, celle-ci peut devenir difficile à exprimer pour des modèles détaillés, et son calcul peut nécessiter de grosses ressources computationnelles. C'est pourquoi certaines approches phylodynamiques se tournent vers des approches sans vraisemblance comme

les approches basées sur le calcul bayésien approché.

La méthode basée sur le calcul bayésien approché n'a été appliquée en phylodynamique que dans quelques études.

? ont mis en œuvre une méthode ABC-MCMC pour ajuster un modèle détaillé à la fois à une phylogénie du virus de la grippe A et à des données d'incidence hebdomadaire. La métrique de distance pour l'ajustement de ce modèle était une fonction de neuf statistiques de résumé, dont quatre étaient basées sur des données de surveillance, comme le nombre moyen de cas rapportés par saison, et les cinq autres étaient dérivées de la variation des séquences et de la phylogénie, comme le temps jusqu'à l'ancêtre commun le plus récent de toutes les séquences de la même saison. Les auteurs ont conclu que les principaux paramètres phylodynamiques pouvaient être estimés par ABC, mais que les statistiques de résumé utilisées devraient très probablement être adaptées d'une étude à l'autre, en fonction du type de données disponibles et des caractéristiques du virus.

Dans une autre étude, **?** a implémenté une méthode ABC-MCMC similaire où il utilise une fonction noyau comme mesure de distance, qui opère exclusivement sur les formes des arbres de la phylogénie. Cette méthode ABC-kernel développée était la seule sans vraisemblance dans une étude de comparaison de méthodes d'inférence phylogénétique, mais a produit les résultats avec des intervalles de confiance les plus larges (**?**).

Récemment, **?** ont proposé une nouvelle approche d'inférence phylodynamique par ABC-régression. Les auteurs ont développé un ensemble de 83 statistiques de résumé, notamment sur les longueurs des branches, sur la topologie et sur le graphe du nombre de lignées au cours du temps. L'ajustement par régression a permis de réaliser une sélection des statistiques de résumé, permettant notamment d'identifier des corrélations entre ces variables et certains paramètres épidémiologiques. Les auteurs ont comparé leur méthode à d'autres approches existantes et ont démontré que bien que leurs estimations étaient comparables à celles obtenues par des méthodes basés sur les modèles BD, la précision des estimations augmentait avec la taille de la phylogénie.

1.4 Objectifs de la thèse

Comprendre la dynamique de propagation de pathogènes microbiens est un enjeu majeur en santé publique. Ceci se fait généralement par l'analyse de données épidémiologiques telles que les données d'incidence ou de prévalence récoltées au cours d'épidémies. Depuis la fin des années 1990, les données de séquences et sur-

tout les phylogénies d'infections reconstruites à partir de séquences de pathogènes récoltées chez différents patients sont utilisées comme source d'information sur la dynamique écologique, évolutive et épidémiologique de ces pathogènes mais la phylodynamique est une discipline encore jeune. Même si les prémisses étaient déjà visibles pendant l'épidémie d'ébola en Afrique de l'Ouest, qui a donné lieu à un partage rapide des données de santé publique en particulier des séquences (**?**), l'épidémie du virus SARS-CoV-2 a représenté un changement qualitatif. Depuis le début de l'épidémie, différents types de jeux de données sont partagés sur différentes plateformes. Par exemple, les agences de santé publiques comme Santé Publique France, mais aussi le site web du programme *Our World in Data* (OWID, `https://ourworldindata.org/coronavirus`) communiquent continuellement des données d'incidence. Un exemple encore plus frappant est celui des données génétiques. Depuis le partage des premières séquences de génomes complets du SARS-CoV-2 en janvier 2020, la base de données centralisée par l'initiative GISAID a recueilli plus de 1 000 000 de soumissions en mai 2021, avec beaucoup de différences entre pays puisqu'environ 23 000 proviennent de France, 383 000 du Royaume-Uni et 370 000 des États-Unis. Ce partage de données a permis le développement de nombreux outils de visualisation et de prévision tels que les outils en ligne Rt2, COVIDici ou Nextstrain, déjà existant, qui a étendu ses analyses au virus du SARS-CoV-2. De plus, un nombre croissant de publications, en particulier en phylodynamique, ont suivi l'accroissement du nombre de données de séquences (**???**).

La prise en compte de l'hétérogénéité de la transmission d'un pathogène est un enjeu majeur en épidémiologie, qui se pose aussi pour la phylodynamique. Cette hétérogénéité peut concerner différents aspects et être modélisée de différentes manières. Elle peut être de nature temporelle, lorsque la variabilité de transmission est liée aux températures de saisons ou à l'ouverture d'écoles, et alors modélisée en utilisant des fonctions oscillantes pour les paramètres du modèle (**??**). Mais l'hétérogénéité peut être aussi due au comportement des individus, par exemple en terme de nombre de contacts sexuels dans le cas d'infections sexuellement transmissibles. On parle alors de groupes à risque élevé ou faible, ce qui peut être modélisé en ajoutant de nouveaux compartiments épidémiologiques. L'hétérogénéité spatiale, elle, peut être abordée *via* des modèles de méta-populations (**?**) ou *via* des réseaux de contacts (**?**). Enfin, l'hétérogénéité peut aussi être due à l'évolution constante des virus qui accumulent des substitutions. Certes la majorité sont neutres mais il peut exister des mutations qui modifient le phénotype de l'infection, par exemple en augmentant la virulence ou la contagiosité, et conduire à des problèmes aigus de santé publique, comme observé avec l'apparition des variants du SARS-CoV-2 (**??**).

Dans le cadre de la phylodynamique, de nouvelles approches permettent d'infé-

rer des paramètres de modèles épidémiologiques structurés à partir de données de séquences datées. Une possibilité est d'utiliser la flexibilité du modèle de coalescent structuré (**?**) mais il existe aussi des approches utilisant les modèles de naissances et de mort (**?**). Cependant, ces approches nécessitent d'exprimer la fonction de vraisemblance associée au modèle ce qui est parfois difficile, voire impossible, pour des modèles détaillés. Et même une fois la fonction exprimée, son calcul nécessite de grosses ressources computationnelles et est coûteux en temps de calcul.

L'utilisation des approches basées sur le calcul bayésien approché (ABC) en phylodynamique s'affranchit de ces limites. En effet, ces approches sont sans vraisemblance et se basent sur la comparaison de données observées et des données simulées. Cela implique que tout modèle peut être utilisé à condition que l'on puisse simuler des données comparables aux données observées. Une étude récente a démontré des résultats d'inférence phylodynamique par ABC-régression d'une précision comparable aux approches fondées sur la vraisemblance pour les modèles SIR et BD (**?**). Malheureusement, les statistiques de résumé développées étaient faiblement corrélées aux paramètres d'un modèle structuré, rendant difficile l'inférence des paramètres comme que la durée d'infection. En effet, il n'existe pas d'ensemble de statistiques de résumé de phylogénie universel et il est encore possible d'en développer de nouvelles notamment pour des phylogénies « labellisées », dont les feuilles sont associées à un groupe dont la caractéristique peut être la localisation géographique, ou un comportement (contacts sexuels, injection de drogues) ou une mutation en particulier.

Bien qu'en plein essor, le domaine de la phylodynamique est encore peu connu et peu étudié en France, ce qui s'est reflété par le peu de séquences de SARS-CoV-2 françaises générées et publiées. Cette thèse présente un double objectif. Le premier objectif est de montrer, à l'échelle nationale, le potentiel d'analyses de données de séquences génomiques pour aider à la prise de décision en santé publique surtout en temps de crise, notamment à travers des collaborations avec des cliniciens de centres hospitaliers universitaires (CHU) ou des laboratoires de biologie médicale. Le second objectif se place à l'intersection des problématiques épidémiologiques, notamment en prenant en compte l'hétérogénéité de transmission, et des problématiques bioinformatiques, en développant des outils et méthodes.

La première partie de la thèse a consisté à développer un outil de simulations rapides de trajectoires, c'est-à-dire des séries temporelles, à partir de n'importe quel modèle, et de phylogénies en utilisant un processus de coalescent. Le développement d'un tel outil s'est avéré nécessaire dès le début de la thèse. En effet, les approches basées sur l'ABC nécessitent de simuler de gros jeux de données, ici des phylogénies. Cependant, la plupart des outils de simulations de phylogénies sont basés sur des modèles simples tels que le SIR ou le BD, et sinon sont limités par leur temps de

calcul augmentant avec la taille de la phylogénie à simuler. Le package R TiPS (pour *Trajectories and Phylogenies Simulator*) que nous avons développé présente l'avantage d'être utilisable pour tout modèle et d'être plus rapide que certains packages existants. Le chapitre 2 décrit le package TiPS et présente les algorithmes ainsi que les résultats de comparaison avec d'autres outils.

Le chapitre 3 concerne l'analyse phylodynamique de la propagation du virus de l'hépatite C (VHC) au sein de populations hétérogènes. L'épidémie de VHC est historiquement observée chez des individus infectés *via* transmission nosocomiale ou par injection de drogue. Depuis quelques années une nouvelle épidémie est observée chez des hommes qui ont des rapport sexuels avec des hommes (HSH). L'étude est née d'une collaboration avec des cliniciens du CHU de Lyon et de l'envie de comprendre et quantifier la différence des dynamiques de cette nouvelle épidémie par rapport à celle observée classiquement. La particularité de cette étude est que les paramètres épidémiologiques tels que la durée d'infection et le R_0 spécifique à chaque épidémie sont inférés de manière simultanée à partir d'une phylogénie labellisée. Pour cela, nous avons développé de nouvelles statistiques de résumé et appliqué la méthode d'ABC-regression.

Le chapitre 4 présente l'application de la méthode d'ABC-regression associée aux nouvelles statistiques de résumé aux données de VIH-1 de groupe O. Dans cette étude, nous avons voulu inférer le différentiel de R_0 afin de quantifier l'avantage pour le virus de présenter le résidu 181C par rapport à un autre, le résidu 181Y, ainsi que la date d'origine de l'épidémie au Cameroun.

Le chapitre 5 est consacré aux analyses épidémiologiques et phylodynamiques du SARS-CoV-2 à partir de données françaises. Nous avons réalisé une première étude sur 196 séquences virales afin d'inférer des paramètres épidémiologiques tels que le temps de doublement, temps nécessaire pour que la taille d'une population double, et le nombre de reproduction au cours du temps (R_t) ainsi que la durée d'infection. Nous avons aussi estimé la date d'origine de la première vague de l'épidémie en France.

En Discussion, nous revenons sur les limites rencontrées au cours de cette thèse, comme le biais d'échantillonnage ou l'impact d'un faible signal phylogénétique sur une analyse phylodynamique, ainsi que sur les pistes de futurs travaux de recherche pour intégrer de nouvelles données à analyser, en plus des phylogénies.

Chapitre 2

Simulation de séries temporelles et de phylogénies avec le package TiPS

2.1 Présentation générale de TiPS

2.1.1 Contexte et structure du simulateur

Comme nous l'avons vu, les modèles mathématiques et les simulations stochastiques sont couramment utilisés pour comprendre et prédire la dynamique des populations dans les systèmes biologiques, écologiques ou épidémiologiques (**???**). Avec le nombre croissant des données génétiques échantillonnées à partir de populations évoluant de manière mesurable, en particulier dans le contexte de l'évolution des agents pathogènes, il est possible de lier dynamiques de population et généalogies (**?**). Certaines méthodes d'inférence phylodynamique, comme celles basées sur le calcul bayésien approché (ABC), nécessitent la simulation en grand nombre de trajectoires et de généalogies (**??**). Il existe donc un besoin évident d'outils pour l'analyse simultanée des prédictions démographiques et généalogiques dans le cadre de modèles stochastiques de dynamique des populations.

L'algorithme de **?** est une méthode couramment utilisée pour réaliser des simulations stochastiques de systèmes basés sur des équations différentielles ordinaires (ODE). L'algorithme résulte de la preuve mathématique formelle de **?**. Cet algorithme ainsi que d'autres versions approchées visant à améliorer sa vitesse sont implémentés dans différents outils tels que les packages R GillespieSSA, adaptivetau et epimdr qui fournissent des fonctions pour simuler les trajectoires à partir de tout modèle. Les packages R geiger (**?**), phytools (**?**), ape (**?**) et TreeSim (**?**) permettent de simuler des phylogénies à partir d'un modèle simple de naissance et de mort. Ce-

pendant, aucun de ces outils ne permet de simuler à la fois des trajectoires à partir de modèles détaillés et des généalogies correspondant aux dynamiques de population simulées. Il existe des exceptions comme les packages R rcolgem (**?**) et phydynR (qui est en fait une amélioration du premier outil rcolgem par le même auteur) ou les packages MASTER (**?**) et Phydyn (**?**) associés au logiciel BEAST2 (**?**). Ces deux derniers outils requièrent de spécifier le modèle et les paramètres du modèle à travers une interface qui n'est pas intuitive pour les utilisateurs. De plus, les packages phydynR et rcolgem peuvent être coûteux en terme de temps de calcul.

Nous présentons ici TiPS (pour *Trajectories and Phylogenies Simulator*), un package R flexible et rapide pour simuler des trajectoires à partir de tout modèle et simuler des phylogénies à partir de ces trajectoires en utilisant un processus de coalescent. Nous introduisons aussi un nouvel algorithme approximatif de la méthode directe de Gillespie permettant d'améliorer les performances en temps de calcul des simulations.

Le package TiPS présente deux principales caractéristiques : la flexibilité et la rapidité. Le besoin d'un outil flexible est à l'origine de notre motivation de développer cet outil. En effet, les méthodes d'inférence phylodynamique basée sur l'ABC nécessitent de simuler des données en masse et ce à partir d'un modèle qui peut varier d'une étude à une autre. Le modèle de dynamique de populations peut être utilisé dans différents domaines notamment en épidémiologie ou en écologie.

Nous avons comparé TiPS aux packages adaptivetau et phydynR pour la simulation des trajectoires et des phylogénies en termes de temps d'exécution et d'exactitude. L'analyse comparative montre que TiPS est le plus rapide en utilisant notre nouvel algorithme approché de simulation de trajectoires. Nous montrons que les trajectoires simulées par TiPS et adaptivetau oscillent autour de la trajectoire déterministe, ce qui est attendu, contrairement aux trajectoires simulées avec le package phydynR. Enfin notre analyse montre que TiPS est au moins 10 fois plus rapide que phydynR pour simuler des phylogénies de taille de 10 feuilles, et peut être jusqu'à 1 000 fois plus rapide pour simuler des phylogénies de plus grande taille de 1 500 feuilles.

Nous souhaitons déposer le package TiPS sur le CRAN (pour *Comprehensive R Archive Network*) qui est le dépôt central des packages R. TiPS a passé une première étape de vérification que le package satisfait aux exigences du CRAN ou *check*. Une cinquantaine de tests sont réalisés lors de cette étape qui permet par exemple de détecter des erreurs de programmation, de tester les exemples de documentation, de s'assurer qu'il n'y a pas d'erreur de syntaxe dans les fichiers de configuration du package et que les dépendances à d'autres packages sont bien indiquées.

L'article présentant le package est en rédaction et à soumettre. La structure du package, les méthodes utilisées et les résultats de comparaison sont présentés

dans la section 2.2. Les détails sur l'algorithme de simulation des phylogénies sont présentés dans l'annexe. Bien que conçu dans une optique de phylodynamique, le package peut aussi être utilisé en écologie ou en dynamique des populations et des scénarios d'applications allant dans ce sens sont présentés dans l'annexe. Enfin, une documentation de type « vignette » avec des exemples de l'utilisation de TiPS est présentée en section 2.3.

Avant de présenter l'article, j'ai souhaité développer quelques points concernant le choix du langage de programmation, le nouvel algorithme proposé et l'hypothèse faite en phylodynamique permettant de simuler des phylogénies à partir de trajectoires, en faisant le lien entre phylogénie et chaîne de transmission.

2.1.2 Rcpp : langages de programmation R et C++

L'implémentation d'un simulateur pour chaque modèle devient rapidement redondante et source d'erreur de code alors que la majorité des lignes du code, quel que soit le langage, est en général identique. En effet, les changements concernent seulement les parties du code liées aux paramètres et taux du modèle. L'autre caractéristique, la rapidité, est bien sûr importante. Par exemple, les méthodes ABC peuvent nécessiter de simuler des millions de jeux de données. Le choix du langage de programmation est de ce fait important et l'exécution d'un programme de simulation uniquement codé en un langage interprété tel que R demandera *a priori* plus de temps de calcul (jusqu'à 500 fois plus) que pour un programme développé en un langage Orienté Objet (OO) tel que C++.

R est un langage de programmation de haut niveau, mais aussi un logiciel multiplateforme qui est très utilisé notamment dans la recherche et dans l'enseignement. De plus, il est interactif permettant une visualisation et analyse directe des données. Toutefois, son utilisation alourdit le temps de calcul de certaines boucles, par exemple, qui ne peuvent pas être parallélisées. Les langages de bas niveau tels que C ou C++ sont, eux, très rapides mais ne son pas interactifs. Il est maintenant possible d'avoir facilement recours au langage C++ dans R *via* le package Rcpp (**?**) aujourd'hui utilisé par environ 15% soient 15 000 packages R. Rcpp permet de transformer un code C++ en une fonction lisible en R. Ainsi, les fonctions qui traitent les informations sur le modèle tel que ses paramètres pour générer un simulateur de trajectoires spécifique au modèle, le temps final de la simulation ou les dates d'échantillonnage pour les phylogénies, sont codées en R. Les simulateurs, eux, sont codés en C++ et possèdent une structure Orientée Objet. Ainsi, l'utilisation de Rcpp nous permet d'avoir les avantages en terme de rapidité, de flexibilité et de diffusion dans la communauté scientifique biologique.

2.1.3 Nouvel algorithme de simulation stochastique de trajectoires

Une des méthodes de simulation stochastique la plus couramment utilisée est la méthode directe de **?**, qui est un algorithme exact de simulation à événements discrets (voire **?** pour une preuve mathématique). Cet algorithme génère le temps jusqu'au prochain événement suivant une loi exponentielle, et le type d'évènement à réaliser. Dans cet algorithme, les taux d'événements augmentent linéairement avec les taille de population. Cette augmentation des taux de transition entraîne une diminution du temps jusqu'au prochain événement et donc une augmentation du nombre d'itérations nécessaires pour la simulation. Le temps de calcul augmente donc linéairement avec la taille de population, et lorsque le processus comporte beaucoup d'événements à simuler dans un intervalle de temps restreint, l'algorithme de la méthode directe de Gillespie peut avoir un temps et un coût de calcul élevé. Cela a donné naissance à des méthodes d'approximations comme la méthode tau-leap (**?**), qui consiste à estimer le nombre d'événements ayant lieu au cours d'un intervalle de temps τ déterminé au préalable. Le nombre d'occurrences dans le laps de temps τ dépend des taux associés aux événements et est donné par la loi de probabilité de Poisson. Toutefois, cet algorithme est limité lorsque le pas de temps τ est relativement petit par rapport aux taux d'événements du système. Le choix du pas de temps est donc important (**?**) et certaines méthodes permettent de l'optimiser (**?**).

Nous introduisons ici un nouvel algorithme, l'algorithme de simulation mixte (MSA pour *Mixed Simulation Algorithm*), qui combine ces deux algorithmes. Cet algorithme commence la simulation avec l'algorithme de la méthode directe de Gillespie (GDA) et bascule vers l'algorithme tau-leap de Gillespie (GTA) si pendant plus de 10 itérations consécutives le temps jusqu'au prochain événement généré est plus petit qu'un seuil. Ce seuil prend par défaut la valeur $\tau/10$ mais peut aussi être défini par l'utilisateur. La méthode passe de l'algorithme GTA à GDA si le nombre d'événements à simuler dans un pas de temps est plus petit que le nombre d'événements possibles. Nous avons implémenté ces trois algorithmes dans le package TiPS. Le MSA est présenté dans l'Algorithme 3. Dans cet algorithme, les termes GDA et GTA correspondent respectivement à l'exécution de la méthode directe et de la méthode tau-leap.

2.1.4 Simulations de phylogénies

TiPS simule des phylogénies enracinées et binaires, c'est-à-dire que chaque nœud interne a deux nœuds fils. Nous distinguons ici le processus biologique et le processus

Algorithme 3 Algorithme de simulation mixte

Entrée: Le nombre d'événements n, les événements $E_1, ..., E_n$, leur taux $R_1, ..., R_n$, le temps final de simulation t_f, le pas de temps τ

Initialiser le temps $t = 0$
Initialiser le compteur d'itérations $\epsilon = 0$
Initialiser le nombre d'événements à simuler $n_{ES} = 0$
tant que $t < t_{final}$ **faire**
  Calculer le taux total $R_{total} = \sum_{i=1}^{n} R_i$
  Tirer deux nombres aléatoires dans une loi uniforme $RAND_1$ et $RAND_2$
  Tirer un temps d'attente $d_t = \frac{-1}{R_{total}}\ log(RAND_1)$
  si $d_t < \tau/10$ **alors**
    $\epsilon = \epsilon + 1$
  sinon
    $\epsilon = 0$
  fin si
  si $\epsilon > 10$ **ou** $n_{ES}/n > 1$ **alors**
    $n_{ES} =$ GTA
    $d_t = \tau$
  sinon
    $n_{ES} = 1$
    $d_t =$ GDA
  fin si
fin tant que

d'observation. Le premier correspond aux événements de naissance, de mort ou de migration, tandis que le second correspond aux événements d'échantillonnage ou de capture qui peuvent interrompre le processus biologique et impacter la dynamique. Dans un contexte épidémiologique, lorsqu'un individu est détecté positif (événement d'échantillonnage), celui-ci peut être traité ou bien peut changer de comportement, interrompant ainsi la transmission du pathogène.

La phylodynamique repose sur l'évolution neutre de pathogènes, notamment viraux, au sein de l'hôte (**?**). Sous cette hypothèse, il est possible de faire un parallèle entre une chaîne de transmission et une phylogénie complète de l'épidémie, pendant laquelle chaque individu infecté a été séquencé. Cependant, une épidémie n'est jamais intégralement échantillonnée. Il faut donc ajouter un système d'échantillonnage dans la simulation afin de simuler des phylogénies d'infections échantillonnées. Les données génétiques, notamment celles utilisées en phylodynamique, sont en général associées à une date d'échantillonnage.

TiPS utilise une approche de coalescent pour simuler les phylogénies en combinant ces dates d'échantillonnage et une trajectoire simulée correspondant à une liste d'événements datés. Sous cette approche nous retraçons l'histoire évolutive et épidémiologique des pathogènes échantillonnés. Ainsi, un événement de transmission peut être représenté par la coalescence de deux lignées ou embranchement, et un événement d'échantillonnage, de fin d'infection ou de mort, si observée, est représenté par une feuille dans la phylogénie. Un événement de migration d'un état à un autre, par exemple la transition d'une phase aiguë à chronique, d'un individu échantillonné changera l'état associé au nœud correspondant.

2.1.5 Applications

Le package TiPS a été utilisé dans une analyse phylodynamique utilisant l'approche ABC que nous présentons dans le chapitre 3 ainsi que dans une étude d'estimation la date d'origine et de fin d'épidémie de SARS-CoV-2 (**?**). Dans cette étude, les simulations de trajectoires ont été réalisées *via* TiPS en utilisant un modèle SEAIRHD (Susceptiles - Exposés - Asymptomatiques - Infectés - Retirés par guérison - Hospitalisés - Décédés). Enfin, notre package a été utilisé dans une récente étude illustrant l'impact des événements de superpropagation, où un individu infecté en infecte de nombreux autres (**?**). Dans cette étude, le modèle utilisé est un modèle SEAIR avec hétérogénéité de transmission avec des individus hôtes dits « supercontaminateurs ». Des trajectoires épidémiologiques ont été simulées à partir de ce modèle et ont été utilisées pour simuler des phylogénies dont un exemple est illustré dans la Figure 2.1.

2.2 TiPS : Simulating trajectories and phylogenies from population dynamics models

Gonché Danesh[1,*], Emma Saulnier[1,2], Olivier Gascuel[2], Marc Choisy[3,4], Samuel Alizon[1]

[1] MIVEGEC, CNRS, IRD, Université de Montpellier
[2] Unité Bioinformatique Evolutive, C3BI/DBC USR 3756, Institut Pasteur, CNRS, Paris, France

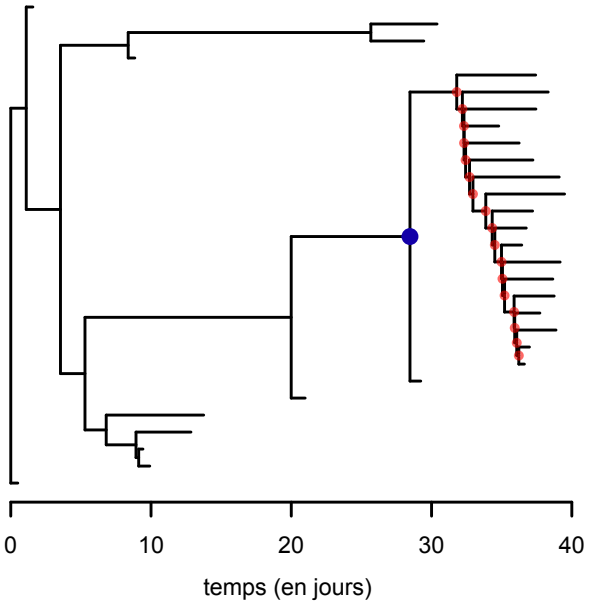


FIGURE 2.1 – **Effet d'un événement de superpropagation sur la topologie d'un arbre représentant une chaîne de transmission simulé avec TiPS.** Chaque nœud interne correspond à un événement de transmission et chaque feuille correspond à une fin d'infection. Le cercle violet représente un événement de superpropagation et les nœuds en rouges correspondent aux infections issues de cet événement.

[3] Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, UK
[4] Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam
* corresponding author : gonche.danesh@gmail.com

2.2.1 Introduction

Stochastic population dynamics simulations are routinely used in biology, ecology, or epidemiology (**???**). These can be used to generate trajectories (i.e. time series of population sizes) and genealogies that capture the relatedness between individuals. The increasing amount of genetic data is fuelling interest in linking population dynamics and genealogies because the former can leave footprints in individuals' genomes (**???**). Such phylodynamics studies involve computer-intensive methods (e.g. Approximate Bayesian Computing (ABC)) that can require the simulation of many trajectories and genealogies (**???**).

One of the most common methods to simulate population dynamics trajectories is Gillespie's exact stochastic simulation algorithm (SSA) (**?**), which derives from the formal result by **?** and has been implemented in a variety of programming languages such as R with packages GillespieSSA (**?**), adaptivetau (**?**), and epimdr (**?**). Still in R, packages geiger (**?**), phytools (**?**), ape (**?**), and TreeSim (**?**) can simulate phylogenies using a birth-death model. However, few software packages simulate both trajectories and genealogies. One exception is the R package rcolgem (**?**), which involves coalescent processes. Another exception are the software packages MASTER (**?**) and Phydyn (**?**) in the BEAST2 platform (**?**). MASTER reads the specification of the model of interest from an XML file, making it difficult to implement detailed models.

We introduce TiPS, a flexible and easy-to-use R package to rapidly simulate population trajectories and phylogenies using a backwards-in-time, i.e. coalescent, process. We also introduce an original approximate version of the Gillespie algorithm to increase calculation speed. A benchmarking analysis shows that TiPS is faster than adaptivetau to simulate trajectories, especially for large populations (Figure 2.3a). It is also at least one order of magnitude faster than phydynR to simulate phylogenies (Figure 2.3b).

2.2.2 Methods

Structure overview

TiPS has two types of stochastic simulation outputs : population dynamics trajectories and phylogenies. These are obtained using a continuous-time individual-based

model defined in R as a system of reactions. The model is first transcribed and compiled in C++, before being linked back to a simulating function in R thanks to the Rcpp package (**?**). The general structure of the pipeline is illustrated by the diagram in the upper box in Figure 2.2 using an epidemiological model as an example.

Model description

We illustrate the use of TiPS with the SIR epidemiological model, where individuals can have three types of status : susceptible (S), infected (I), and removed (R) (**?**). The model can be described as a system of individual-based reactions :

$$S \xrightarrow{\beta SI} I \tag{2.1a}$$

$$I \xrightarrow{\gamma I} R \tag{2.1b}$$

where β and γ are the transmission and recovery rates. The rate of occurrence of each reaction is indicated above the transition arrow. The corresponding population-based ODE system is shown in the top-right panel of Figure 2.2.

To generate a simulating function for a model of interest, the user gives as input the individual-based reactions of the model in a string vector as shown in Figure 2.2.

Simulating trajectories

The simulating function takes as arguments a vector with the initial number of individuals in each status, a named list with the parameter values, a vector of the time limits of the simulation, and the type of algorithm to use for the simulation. Users can also enter a vector of breaking time points, which allows parameter values to vary over time. These breaking time points are indicated in the time limits vector, and the corresponding parameter values are indicated chronologically in the named list of parameter values.

Three simulation algorithms are implemented in the Rcpp simulating function :

1. Gillespie's Direct Algorithm (GDA, the default option) (**?**) is an exact algorithm that simulates the time until the next event by assuming that waiting times are exponentially distributed. A limitation is that its computational complexity scales linearly with the number of events and the population size.
2. Gillespie's Tau-Leap Algorithm (GTA) (**?**) is an approximate algorithm that introduces a fixed time-step during which the number of events of each type is

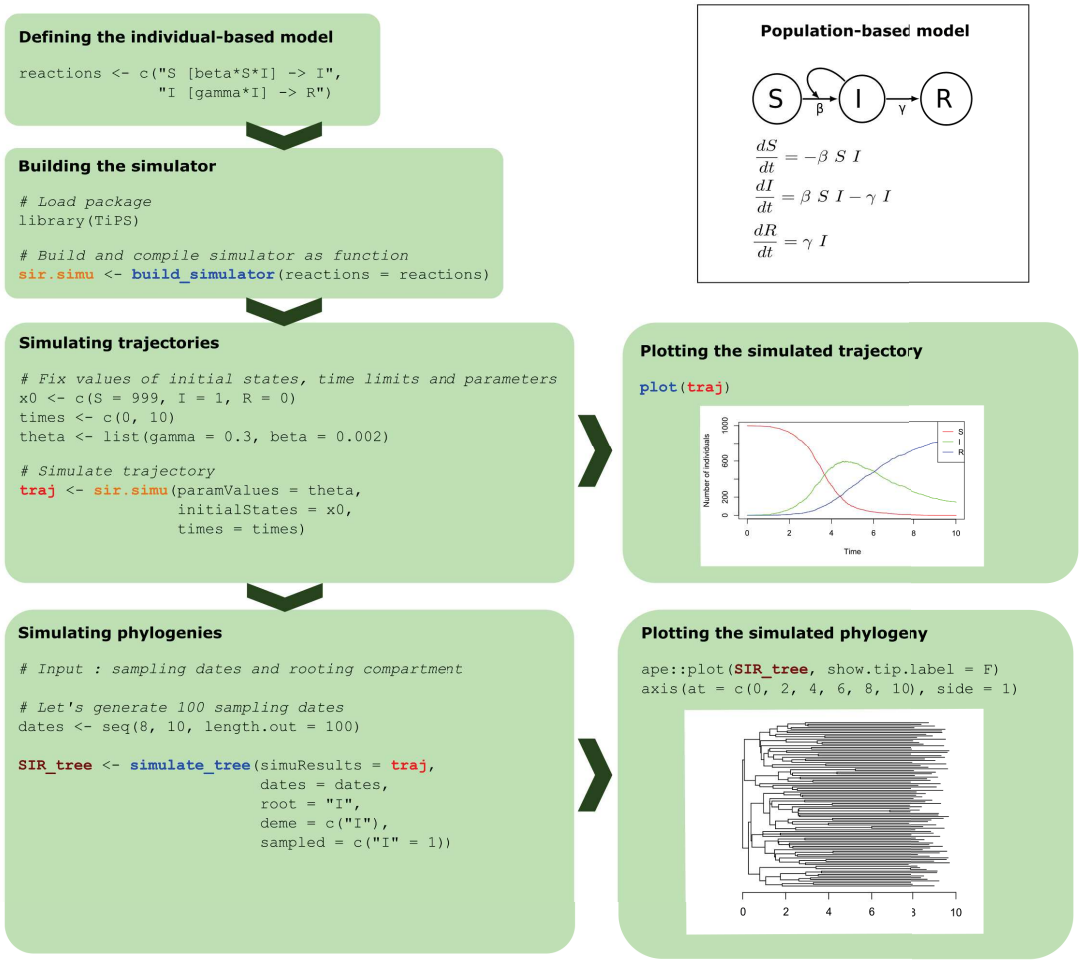


FIGURE 2.2 – **General structure of the TiPS pipeline**. The equations and outputs correspond to the *SIR* epidemiological model (**?**). The functions of the R package are in blue. The simulator of trajectories, which is built as a function, is in orange. The variable *traj*, in red, is the output trajectory and a *simutraj* class object. It can be plotted using our *plot* function. The simulated phylogeny is a *phylo* class object from the ape R package (**?**), which can be used for plotting.

assumed to be Poisson-distributed. This algorithm is limited if the time step is small compared to the rate at which events occur.

3. the Mixed Simulation Algorithm (MSA) is a new algorithm we introduce that

switches from GDA to GTA if over 10 iterations the time until the next event is below a user-defined threshold, and from GTA to GDA if the total number of realised events is lower than the number of possible events. For similar approximations of the GDA see the next reaction method (**?**), the optimized direct method (**?**), the sorting direct method (**?**), or the adaptive explicit-implicit tau-leaping method (**?**).

For an illustration, see the 'Simulating trajectories' box in Figure 2.2.

Simulation of phylogenies

A phylogeny is the representation of the evolutionary history and relationships between organisms or groups of organisms. TiPS simulates rooted and binary phylogenies. The root of the tree represents the ancestral lineage, and the leaves represent the descendants of that ancestor.

We distinguish the biological process from the observation process. The former corresponds to a sequence of individual birth, migration, and deaths events. The latter corresponds to sampling or capture events that can interrupt the biological process.

TiPS uses a coalescent approach (**?**) to simulate phylogenies based on trajectories, i.e. a list of dated events (or 'reactions'), and known sampling dates (typically corresponding to observed data in ABC approaches). A forward-in-time birth event can be represented as a coalescence between two lineages (or branching) under a backwards-in-time process. A sampling event is represented as an external node (or leaf) in the phylogeny, and a death event, if observed, is also represented as a leaf.

TABLE 2.1 – **Main algorithms of the R packages compared.** GDA stands for Gillespie's Direct Algorithm, GTA for Gillespie's Tau-Leap Algorithm, and MSA for Mixed simulation algorithm.

Package	**Methods**	**Features**
TiPS	GDA (exact)	
	GTA (approximate)	Requires a fixed time-step τ
	MSA (mixed)	Requires a fixed time-step τ
adaptivetau	GDA (exact)	
	GTA (approximate)	Requires a fixed time-step τ
phydynR	Euler-Maruyama	Requires a time-step τ

A 'migration' event of a sampled individual from one status to another changes the status of its corresponding node.

TiPS can simulate the phylogeny of the full trajectory but it can also simulate a sampled phylogeny if sampling dates are provided. If sampled individuals can belong to more than one host status, the user can choose between forcing the status associated with each sampling event, define a proportion of sampling events associated with each type of status, or draw the status at random with equal probability. In the last two cases, TiPS randomly associates the sampling dates to a status.

After this pre-processing, the sampling dates are organised as a named vector containing decimal dates and the reaction indicating the class of individuals to sample. TiPS then incorporates this vector into the recorded trajectory (containing also dates and reactions) in a chronological order.

The tree simulation starts from the last (i.e. most recent) sampling date and progresses through the simulated trajectory backwards-in-time. Each of the four types of reactions (birth, death, migration, and sampling) can result in a modification in the simulated tree as explained above.

The number of events that lead to a change in the phylogeny, such as a new leaf or a branching, can be determined using a drawn without replacement from a hypergeometric distribution. Further details can be found in the Appendix.

The output simulated phylogeny is an R object of class *phylo* as defined in the ape package (**?**).

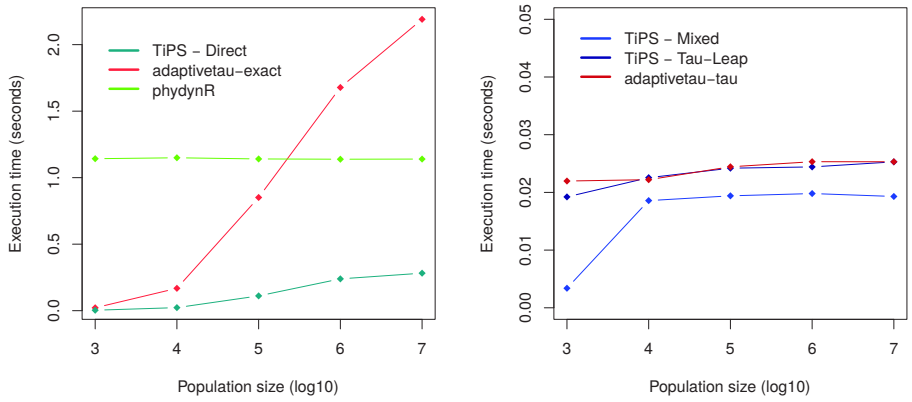
An example of how to run a simulation of a phylogeny is shown in the 'Simulating phylogenies' box in Figure 2.2.

2.2.3 Results

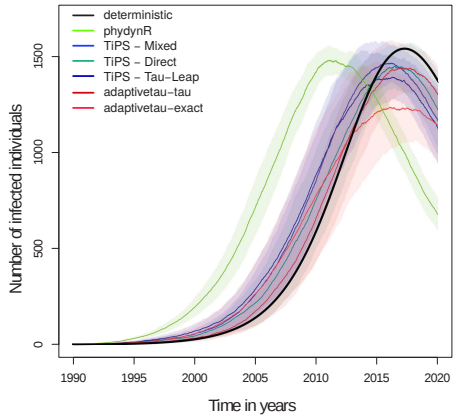
To evaluate TiPS's performances, we performed a benchmarking analysis on both modules of the software package, i.e. the trajectory and the phylogeny simulators, using existing R packages (adaptivetau and phydynR). Table 2.1 summarises the main algorithms and approaches used.

We first evaluated the computational speed and the accuracy of the simulation of trajectories (i.e. populating dynamics). For 5 different initial population sizes, we simulated 1,000 trajectories of the epidemiological Susceptible-Infected-Recovered (SIR) model and measured the execution time using functions from three different R packages.

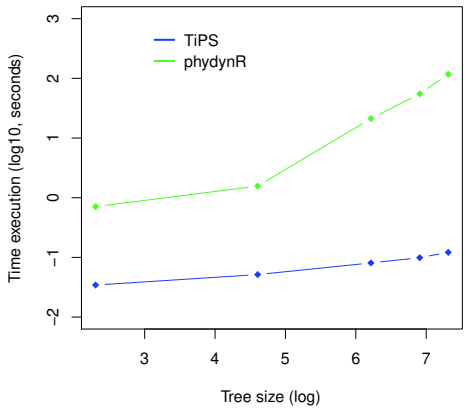
In Figure 2.3(a), we show the mean execution time. For TiPS and adaptivetau, we use Gillespie's Direct method, whereas phydynR uses the Euler-Maruyama integration (EMI). As expected, the population size, and, hence, the number of events per unit



(a)



(b)



(c)

FIGURE 2.3 – Comparison of computation speed when a) simulating trajectories and b) phylogenies. c) Accuracy of the simulated trajectories.

of time increases the execution time for the GDA-based method but not for the one using the EMI. However, TiPS remains faster than the other two software packages for large populations (10^7 individuals). On the right panel of Figure 2.3(a), we perform the same simulations using approximations of the Gillespie algorithm with fixed time-

steps. The computational speed of this GTA implemented in TiPS and adaptivetau is comparable and much faster than the GDA. Furthermore, our new MSA algorithm improves computational speed compared to the GTA, especially for small population sizes.

In Figure 2.3(b), we show the deterministic trajectory, and, for each algorithm used, the mean simulated trajectory and its 90% confidence envelope. Trajectories simulated using TiPS (in blue) are the closest from the deterministic prediction (in black).

We then compared the speed of computation of simulations of phylogenies under an SIS model assuming different sampling proportions, which led to different phylogeny sizes : 10, 100, 500, 1000, and 1500 leaves. We simulated 1,000 phylogenies under each sampling scheme using phydynR and our package. Figure 2.3(c) shows the mean time executions of simulations of phylogenies for each tool and each sampling scheme. We can see that TiPS appears to be one order of magnitude faster, with a more pronounced advantage for large phylogenies.

2.2.4 Discussion

We have developed an R package to simulate trajectories and phylogenies from compartmental models. The trajectories simulation relies on the event-based Gillespie's algorithm, which already allows one to capture several degrees of heterogeneity between individuals or between populations, e.g. metapopulation structures. The simulation of the phylogeny relies on the trajectory and involves a coalescent approach.

Benchmarking analyses show that TiPS is comparable our outperforms existing R packages in terms of speed when generating numerous trajectories or phylogenies. The accuracy of the simulations is also equivalent or better than that of the other software packages.

The two great advantages of the method are its flexibility, it can readily capture any ODE-based population dynamics model, and its speed. These properties have already been used for phylodynamics studies involving approximate bayesian computing (**?**) or to illustrate the effect of superspreading events (**?**).

Future extensions will consist in introducing non-Markovian dynamics and simulating multifurcating phylogenies.

2.2.5 Supplementary Information

Phylogenies simulation algorithm

TiPS simulates rooted and binary phylogenies, meaning that every internal node has exactly two daughter nodes. The root of the tree represents the ancestral lineage, and the tips of the branches (or leaves) represent the descendants of that ancestor. Using the backwards-in-time process, the coalescence between two lineages (i.e. a speciation event or transmission event) corresponds to a branching event in the phylogeny. Note that we will refer to a node's height as its distance to the root.

In this section we distinguish two types of compartment : the deme compartments, where individuals contribute directly or indirectly to the phylogeny, and the non-deme compartments that cannot be placed in the genealogical process. Each deme compartment is denoted X_i, with i ranging from 1 to the number of deme compartments in the model. Sampled individuals belong to the sub-compartment X_i' ($X_i' \subset X_i$) and are all associated with a leaf in the tree. We also introduce X_i'', the sub-compartment of X_i' ($X_i'' \subset X_i'$) corresponding to individuals in X_i' that have not yet been but may be sampled in a backwards-in-time process. The discrete size of compartments X_i, X_i' and X_i'' at time t are denoted as $|X_i|$, $|X_i'|$ and $|X_i''|$, respectively. A non-deme compartment is denoted Z and its discrete size $|Z|$.

The tree simulation starts from the last (i.e. most recent) sampling date and progresses through the simulated trajectory backwards-in-time. The number of events that lead to a change in the phylogeny is drawn without replacement from a hypergeometric distribution. The hypergeometric distribution is appropriate as it describes the number of events k from a sample n drawn from a total population N without replacement. Each of the four types of reactions (sampling, birth, death, and migration) can result in a modification in the simulated tree : a new external node (or leaf) or the coalescence of two lineages.

Sampling event. We define a sampling event as an event that interrupts the biological process of an individual in the population of interest, and a re-sampling event as an observation event. In an ecological context, the sterilisation of an animal would be a sampling event whereas marking the animal would be a re-sampling event. In an epidemiological context, under our definition, when an individual gets vaccinated or when an infected person is detected, we assume that it will lead to the end of infection or the interruption of the spread of the pathogen through isolation or the use of protection, which we will consider as a sampling event. However, if this is not the case, it is considered as a re-sampling event, where the individual will still transmit the pathogen to other individuals.

When we know the sampling dates, we know the number of sampling events occurring at time t (n^S_{rep}) and, therefore, the number of tips to create with height t. However, we allow sampled lineages to continue to have an offspring and therefore need to determine if the nodes we sampled at time t are associated with re-sampling events or not (i.e. if they should be linked to a node that has already been sampled after time t in our coalescent approach). The number of re-sampling (n_{RS}) and sampling events (n_S) at time t is governed by the following relationships :

$$n_{RS} \sim \text{HyperGeom}\left(n^S_{\text{rep}}, |X''_i|, |X_i| - (|X'_i| - |X''_i|)\right) \tag{S1a}$$

$$n_S = n^S_{\text{rep}} - n_{RS} \tag{S1b}$$

where n^S_{rep} is the total number of phylogeny nodes generated at t. Using the hypergeometric distribution, we compute the number of re-sampling events n_{res} if we sample n^S_{rep} times without replacement in a sample of size $|X_i| - (|X'_i| - |X''_i|)$ containing $|X''_i|$ individuals. The number of 'classical' samplings, n_S, is the difference between n_{RS} and the number of tips to be generated (n^S_{rep}).

In the case of a re-sampling, we randomly pick a node from X''_i (the pool of individuals who have not yet been sampled), update its height to t, and link it to a node from X'_i (the pool that has been sampled). Each re-sampling event decreases $|X''_i|$ by one. In the case of a 'classical' sampling event, a new node with height t is created in X'_i and $|X'_i|$ increases by one.

Birth event. A birth reaction can be written as follows : $X_i \to X_j + X_i$ where i and j range from 1 to the number of deme compartments. The birth reaction can also be written as $Z + X_i \to X_j + X_i$ if it includes a non-deme individual. A birth reaction leads to one of two types of modification in the tree : a coalescence or an 'invisible' coalescence. A coalescence corresponds to the coalescence of two sampled lineages from two individuals, the recipient and the donor, into one sampled individual lineage (the donor). An 'invisible' coalescence is defined as the coalescence of the sampled lineage of the recipient (here X'_j) and an unsampled lineage of the donor (X''_j) into one unsampled lineage (the donor, X''_j).

We first need to determine the number of recipient lineages $n_{\text{recipient}}$. Again, assuming a hypergeometric distribution, we compute the number of recipient lineages $n_{\text{recipient}}$ if we sample n_{rep} times without replacement in a sample of size $|X_j|$ containing $|X'_j|$ individuals (see equation S2a). If there are no recipient lineages ($n_{\text{recipient}} = 0$), there is no coalescence or invisible coalescence and, therefore, no change in the tree. Otherwise, we need to determine the number of donor lineages n_{donor}. Since the sampling is performed without replacement, the total sample size is

updated to $|X_j| - n_{\text{rep}}$ and the number of daughter lineages Y represented by nodes that are available for the coalescence becomes $|X'_j| - n_{\text{recipient}}$. Thus, we compute the number of donor lineages n_{donor} if we sample n_{rep} times without replacement in a sample of size $|X_j| - n_{\text{rep}}$ containing $|X'_j| - n_{\text{recipient}}$ individuals S2b. Mathematically, we can write :

$$n_{\text{recipient}} \sim \text{HyperGeom}(n_{\text{rep}}, |X'_j|, |X_j|) \tag{S2a}$$

$$n_{\text{donor}} \sim \text{HyperGeom}(n_{\text{rep}}, |X'_i| - n_{\text{recipient}}, |X_i| - n_{\text{rep}}) \tag{S2b}$$

Knowing the number of recipient lineages, we need to determine the number of coalescence (n_C), *i.e.* the number of lineages among the n_{donor} sampled donor lineages that will coalesce with the recipient ones. The number of visible coalescences is drawn from a hypergeometric law where we sample $n_{\text{recipient}}$ times in a sample of total size n_{rep} containing n_{donor} sampled donor lineages S3a. The number of invisible coalescences, n_{IC}, is the number of remaining recipient lineages that have not coalesced with sampled donor lineages S3b.

$$n_C \sim \text{HyperGeom}(n_{\text{recipient}}, n_{\text{donor}}, n_{\text{rep}}) \tag{S3a}$$

$$n_{IC} = n_{\text{recipient}} - n_C \tag{S3b}$$

Upon a coalescence event, $|X'_j|$ is decreased by one, a node is randomly picked in X'_i, its height is updated to t, and it is linked to another node that is removed from X'_j. For an invisible coalescence event, $|X'_j|$ is decreased by one and both $|X''_i|$ and $|X'_i|$ are increased by one. In this case, a new node from X'' is created with height t and linked to a node randomly picked in X'_j. In both cases, $|X_i|$ is decreased by one (which is already known from the trajectory).

Death event. By default, a death reaction does not require any tree modification. However, if we want to simulate a full tree instead of a sampled one, the sampling events will correspond to the death events, which will therefore lead to the addition of a node. In this case, the number of sampled death events is $n_{SD} = n_{\text{rep}}$.

Migration event. Only migration events involving deme individuals can lead to a modification in the tree. These migration reactions can be written as $X_i \to X_j$, with $j \neq i$. We assume that the number of migrations that lead to a tree modification is given by the following hypergeometric distribution :

$$n_M \sim \text{HyperGeom}(n_{\text{rep}}, |X'_j|, |X_j|) \tag{S4a}$$

A migration increases $|X'_i|$ and decrease $|X'_j|$ by one. A new node is created in X'_i with height t and linked to a node randomly picked in X'_j, which is then removed from X'_j. Furthermore, X_i is incremented by one and X_j is decremented by one.

Application to an epidemiological SI_aI_cR model

We illustrate the functioning of TiPS using an SI_aI_cR epidemiological compartmental model, where individuals can be susceptible (with density S), infected in acute phase (I_a), infected in chronic phase (I_c), and removed (R). The corresponding ODE system is

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = -\beta\ S(t)\ I_a(t) - \beta\ S(t)\ I_c(t) \tag{S5a}$$

$$\frac{\mathrm{d}I_a(t)}{\mathrm{d}t} = \beta\ S(t)\ I_a(t) + \beta\ S(t)\ I_c(t) - \sigma\ I_a(t) \tag{S5b}$$

$$\frac{\mathrm{d}I_c(t)}{\mathrm{d}t} = \sigma\ Ia(t) - \gamma\ I_c(t) - \alpha\ I_c(t) \tag{S5c}$$

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t} = \gamma\ I_c(t) \tag{S5d}$$

where β is the infectious contact rate, γ the recovery rate, α the virulence, and $1/\sigma$ the expected duration of the acute phase.

The model can be described as an individual-based model using a system of reactions :

$$S + I_a \xrightarrow{\beta S I_a} I_a + I_a \tag{S6a}$$

$$S + I_c \xrightarrow{\beta S I_c} I_c + I_a \tag{S6b}$$

$$I_a \xrightarrow{\sigma I_a} I_c \tag{S6c}$$

$$I_c \xrightarrow{\gamma I_c} R \tag{S6d}$$

$$I_c \xrightarrow{\alpha I_c} \emptyset \tag{S6e}$$

where the rate of occurrence of each reaction is indicated above the transition arrow.

Reactions S6a and S6b are birth reactions, S6c and S6d are migration reactions, and finally S6e is a death reaction event.

TiPS will first build and generate a function to simulate trajectories using this system of reactions. Population dynamics are simulated by providing parameter values using one of the algorithm described in the main text. In the following, we focus on the model captured by system of equations S6.

In this approach, under this epidemiological model, we trace back the epidemiological history of the sampled virus. Hence, the deme compartments, i.e. the ones sampled, where the virus is present and then contributing to the phylogeny, are I_a and I_c.

Once the trajectory is simulated, to simulate a sampled phylogeny, TiPS requires the sampling dates and the proportion of the sampling dates to be associated with each type of deme. Let us assume, for example, that 15% of the sampling dates are associated with the I_a deme and 85% with the I_c deme compartment. TiPS randomly assigns each sampling date to a deme compartment based on these ratios and adds the dates to the list of events in the simulated trajectory (see Supplementary Figure S5). The R code to build the simulator, simulate a trajectory and a phylogeny are shown in Supplementary Figure S1.

Deme compartments (I_a and I_c) can be composed of individuals represented by nodes in the simulated tree (belonging to the sub-compartments I'_a and I'_c). These individuals represented by nodes can also be still unsampled (belonging to I''_a and I''_c).

TiPS starts the simulation of the phylogeny from the most recent sampling event and follows the trajectory backwards in time.

If the backward step in the trajectory leads to one or multiple migration events, i.e., in this model, from the acute phase compartment (I_a) to the chronic infectious phase compartment (I_c) as in reaction S6c, the number of tree modifications is given by the following hypergeometric distribution :

$$n_M \sim \text{HyperGeom}(n_{\text{rep}}, |I'_c|, |I_c|) \tag{S7}$$

An illustration of the tree update is shown in Supplementary Figure S2.

If the backward step in the trajectory leads to a birth reaction, two tree modifications are possible : a coalescence and an invisible coalescence. In this model, there are two different demes and two different birth reactions (see reactions S6a and S6b). Given the birth reaction S6a, the number of coalescences n_C and invisible coalescences n_{IC} leading to tree modifications are governed by the following relationships :

$$n_{\text{recipient}} \sim \text{HyperGeom}(n_{\text{rep}}, |I'_a|, |I_a|) \tag{S8a}$$
$$n_{\text{donor}} \sim \text{HyperGeom}(n_{\text{rep}}, |I'_a| - n_{\text{recipient}}, |I_a| - n_{\text{rep}}) \tag{S8b}$$
$$n_C \sim \text{HyperGeom}(n_{\text{recipient}}, n_{\text{donor}}, n_{\text{rep}}) \tag{S8c}$$
$$n_{IC} = n_{\text{recipient}} - n_C \tag{S8d}$$

Given birth reaction S6b, the number of coalescences and invisible coalescences lea-

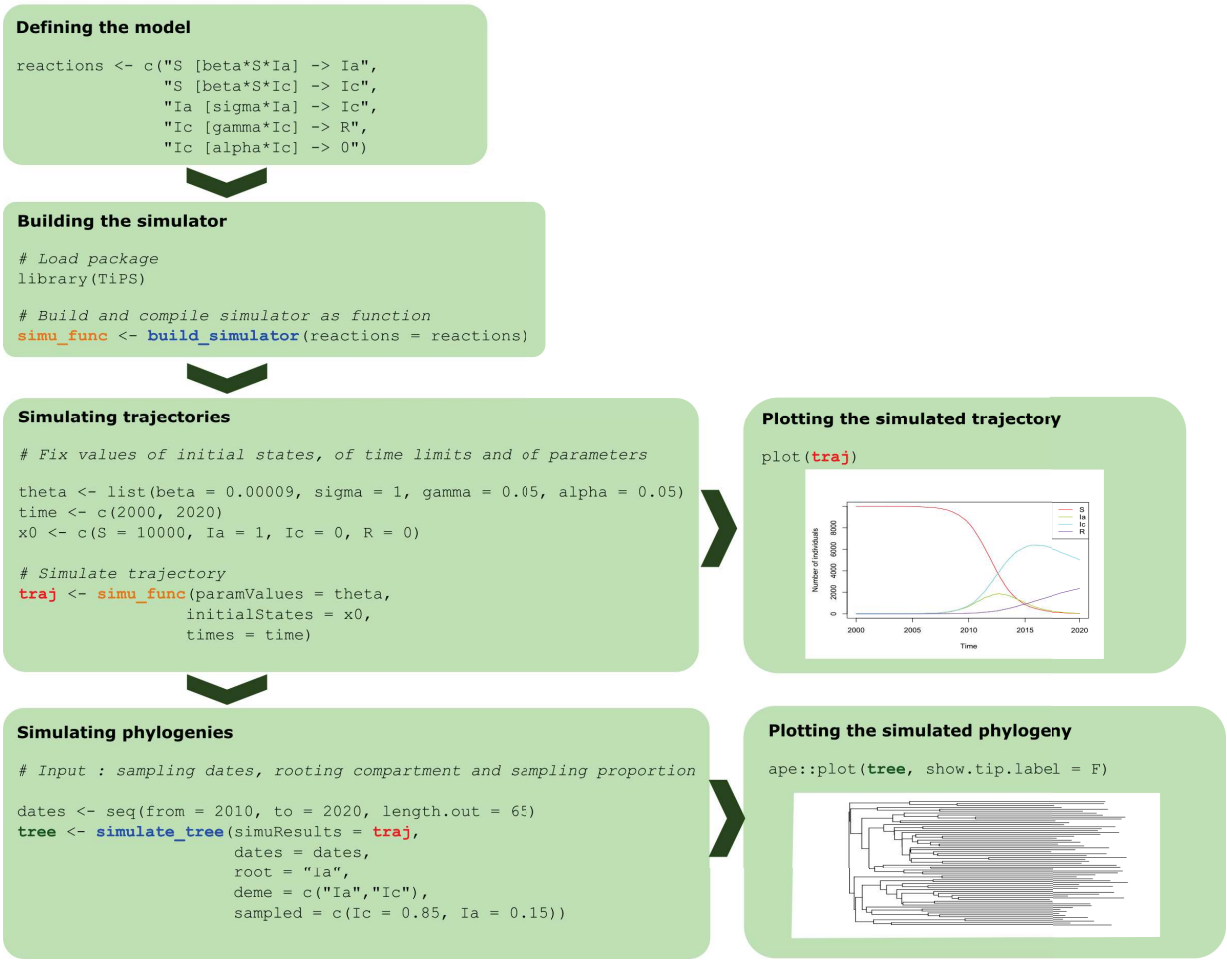


FIGURE S1 – **Simulating a trajectory and a phylogeny using TiPS.** The equations and outputs correspond to the SI_aI_cR model. The functions of the R package are in blue. The simulator of trajectories built as a function is in orange. The variable *traj* in red is the output trajectory. The phylogeny is plotted using a function from the ape R package (**?**).

ding to tree modifications are governed by the following relationships :

$$n_{\text{recipient}} \sim \text{HyperGeom}(n_{\text{rep}}, |I'_a|, |I_a|) \tag{S9a}$$

$$n_{\text{donor}} \sim \text{HyperGeom}(n_{\text{rep}}, |I'_c| - n_{\text{recipient}}, |I_c| - n_{\text{rep}}) \tag{S9b}$$

$$n_C \sim \text{HyperGeom}(n_{\text{recipient}}, n_{\text{donor}}, n_{\text{rep}}) \tag{S9c}$$

$$n_{IC} = n_{\text{recipient}} - n_C \tag{S9d}$$

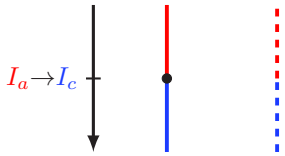


FIGURE S2 – **Migration event in the SI_aI_cR model.** The evolution of a pathogen lineage in an individual during the acute phase of its infection is represented in red. The individual then enters the chronic phase of his infection and the pathogen lineage continues to evolve (represented in blue branches). The solid branch represents the evolution of the sampled pathogen lineage and the dashed branch represents the evolution of an unsampled pathogen lineage. The unsampled lineage is eventually removed and not represented in the final simulated phylogeny.

Supplementary Figure S3 illustrates possible tree updates.

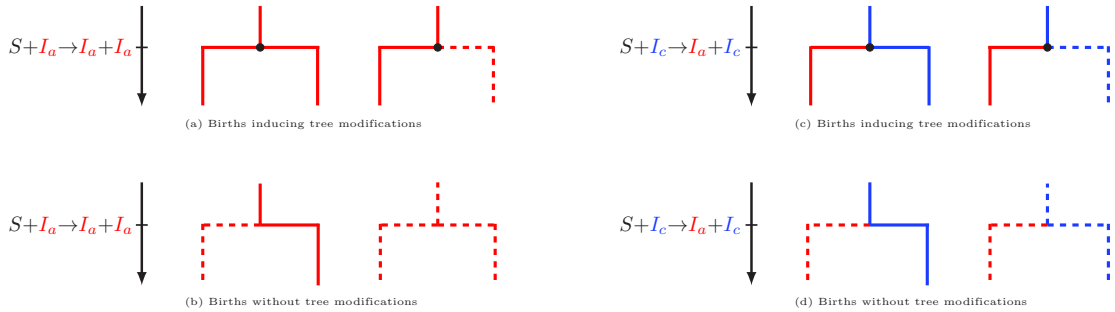


FIGURE S3 – **Births events in the SI_aI_cR model.** Each transmission event from a living infectious individual to a susceptible individual can be represented by a branching. We show the 'donnor' lineage on the right side of the branching and the recipient pathogen lineage (i.e. the newly-infected individual) on the left side. Dots correspond to nodes in the resulting phylogeny. Color and branch line codes are identical to Figure S2.

This model features two types of sampling reactions : the sampling of an I_a individual and the sampling of an I_c individual. If the backward step in the trajectory leads to one or multiple sampling reactions of I_a individuals, the number of re-samplings ($n_{RS(I_a)}$ and $n_{RS(I_c)}$) and classical samplings ($n_{S(I_a)}$ and $n_{S(I_c)}$) are governed by the relationships S10a and S10b, respectively. An illustration of the possible modifications in the tree are shown in Supplementary Figure S4. Note that the user can allow

for re-sampling or not.

$$n_{RS(I_a)} \sim \text{HyperGeom}\left(n_{\text{rep}}^{S(I_a)}, |I_a''|, |I_a| - (|I_a| - |I_a''|)\right) \tag{S10a}$$
$$n_{S(I_a)} = n_{\text{rep}}^{S(I_a)} - n_{RS(I_a)} \tag{S10b}$$
$$n_{RS(I_c)} \sim \text{HyperGeom}\left(n_{\text{rep}}^{S(I_c)}, |I_c''|, |I_c| - (|I_c| - |I_c''|)\right) \tag{S10c}$$
$$n_{S(I_c)} = n_{\text{rep}}^{S(I_c)} - n_{RS(I_c)} \tag{S10d}$$

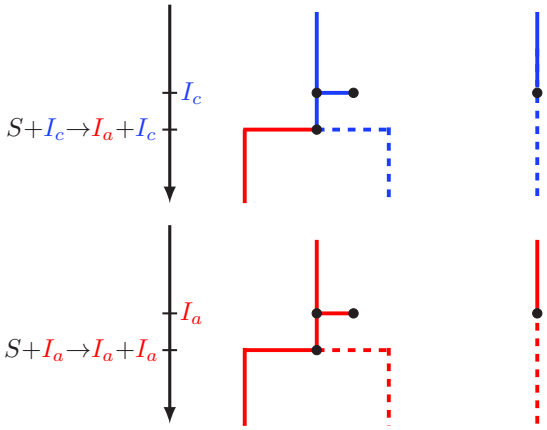


FIGURE S4 – **Samplings in the SI_aI_cR model.** Colors, branches, and dots code are identical to Figure S3. Each sampling event leads to the addition of a node. Re-sampling events (left side of the figure) occur when the pathogen lineage has already been sampled but the individual currently carrying it has never been sampled (individuals can only be sampled once but can transmit after sampling). Otherwise, we have a classical sampling event (right side of the figure), i.e. the pathogen has not been sampled yet.

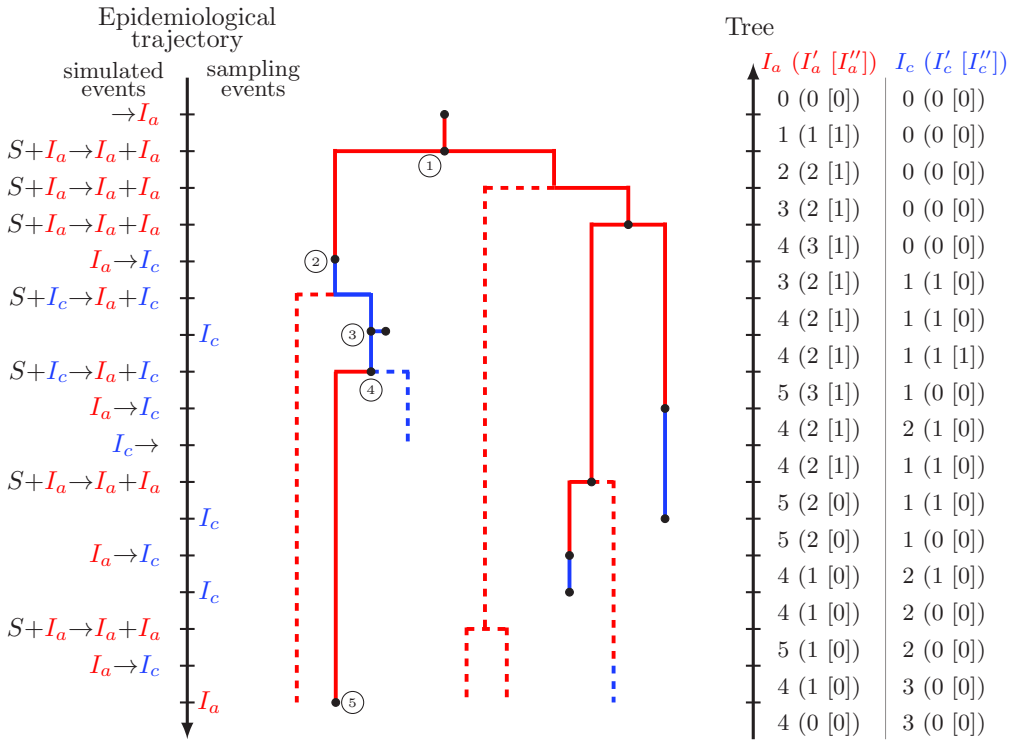


FIGURE S5 – **Tree simulation.** We represent the epidemiological history of the individuals carrying sampled viruses in solid lines and the rest in dashed lines. Colors are identical to Figure S3. In this representation, at each birth event (branching), the donor is deviated to the right side and the recipient to the left side. All possible tree modifications are represented in this figure : (1) Coalescence ; (2) Migration ; (3) Re-sampling ; (4) Invisible Coalescence ; (5) Sampling.

Application to a logistic growth model

In this section we use the African Savannah elephant (*loxodonta africana*) population in the Kruger National Park as an example that has demonstrated an important increase in the last two decades. A study has shown that a high number of the elephants can damage the Park's ecosystem (**?**). To prevent that, until the early 90's, a solution was the culling of elephants. Due to ethical issues, another solution has been proposed where female elephants would be on contraceptives (**?**).

Here we use a logistic growth model to study the population dynamics. The ODE

system is :

$$\frac{\mathrm{d}N}{\mathrm{d}t} = r\ N\ (1 - \frac{N}{K}) \tag{S11}$$

$$\tag{S12}$$

where N is the elephant population density, K is the carrying capacity of the environment, and r is the intrinsic growth rate of the population where $r = b - d$ with b the birth rate and d the death rate.

The model can be described as an individual-based model using a system of reactions :

$$N \xrightarrow{bN} N + N \tag{S13a}$$

$$N \xrightarrow{(d+(b-d)N/K)/N} \emptyset \tag{S13b}$$

where the rate of occurrence of each reaction is indicated above the transition arrow. Reaction S13a is a birth reaction and S13b a death reaction.

TiPS allows to simulate a phylogeny without sampling dates, where events interrupting the biological process, such as deaths events or here sterilisation events, simulated in the trajectory are represented as leaves in the phylogeny. To illustrate this module of the tool, we add a sterilisation rate to the model to simulate the events in the trajectory. The system of reactions becomes :

$$N \xrightarrow{bN} N + N \tag{S14a}$$

$$N \xrightarrow{d+(b-d)N/K} \emptyset \tag{S14b}$$

$$N \xrightarrow{\epsilon N} \emptyset \tag{S14c}$$

where reaction S14c is the sterilisation reaction with a rate of its occurrence indicated above the transition error.

Using this system of reactions, TiPS will first build and generate a function to simulate trajectories. Population dynamics are simulated by providing parameter values using one of the algorithm described in the main text. The R code to build the simulator, simulate a trajectory and a phylogeny are shown in Supplementary Figure S6.

We assume that a sampling event is an event that interrupts a biological process. In this example, where no sampling dates are required, the death events and the elephant sterilisation events will be considered as the sampling events and will be represented as leaves in the simulated phylogeny.

We introduce here the compartments N, N' and N'' where $N'' \subseteq N' \subseteq N$. All the elephants are in compartment N, the sampled elephants are sub-compartment N' and the elephants that in N' that have not yet been but may be sampled are in sub-compartment N''.

TiPS starts the simulation of the phylogeny from the most recent sampling event and follows the trajectory backwards-in-time.

When the backward step in the trajectory at time t leads to n death or sterilisation events, n new nodes are created with height t in N' and $|N'|$ increases by one. Note that, each node is labelled with the corresponding reaction, so we can distinguish and visualise them when plotting the phylogeny.

If the backward step in the trajectory leads to a birth reaction, two tree modifications are possible : a coalescence and an invisible coalescence. A coalescence in the phylogeny corresponds to the coalescence of two sampled lineages each representing an elephant, into one sampled individual (the donor). An 'invisible' coalescence is the coalescence of the sampled lineage of the recipient individual (N') and an unsampled lineage of the donor individual (N'') into one unsampled lineage (the donor, N''). Given the birth reaction S14a, the number of coalescences n_C and invisible coalescences n_{IC} leading to tree modifications are governed by the following relationships :

$$n_{\text{recipient}} \sim \text{HyperGeom}(n_{\text{rep}}, |N'|, |N|) \tag{S15a}$$
$$n_{\text{donor}} \sim \text{HyperGeom}(n_{\text{rep}}, |N'| - n_{\text{recipient}}, |N| - n_{\text{rep}}) \tag{S15b}$$
$$n_C \sim \text{HyperGeom}(n_{\text{recipient}}, n_{\text{donor}}, n_{\text{rep}}) \tag{S15c}$$
$$n_{IC} = n_{\text{recipient}} - n_C \tag{S15d}$$

where n_{rep} is the number of birth events in the trajectory at time t of the trajectory, $n_{recipient}$ is the number of recipient lineages and n_{donor} the number of donor lineages. Upon a coalescence, $|N'|$ is decreased by one, a node is randomly picked from N' with height t and is linked to another node that is removed from N'. For an invisible coalescence, $|N'|$ is decreased by one both and N'' and N' are increased by one. A new node from N'' is created with height t ans is linked to a node randomly picked in N'. In both cases, $|N|$ is decreased by one as recorded already in the simulated trajectory.

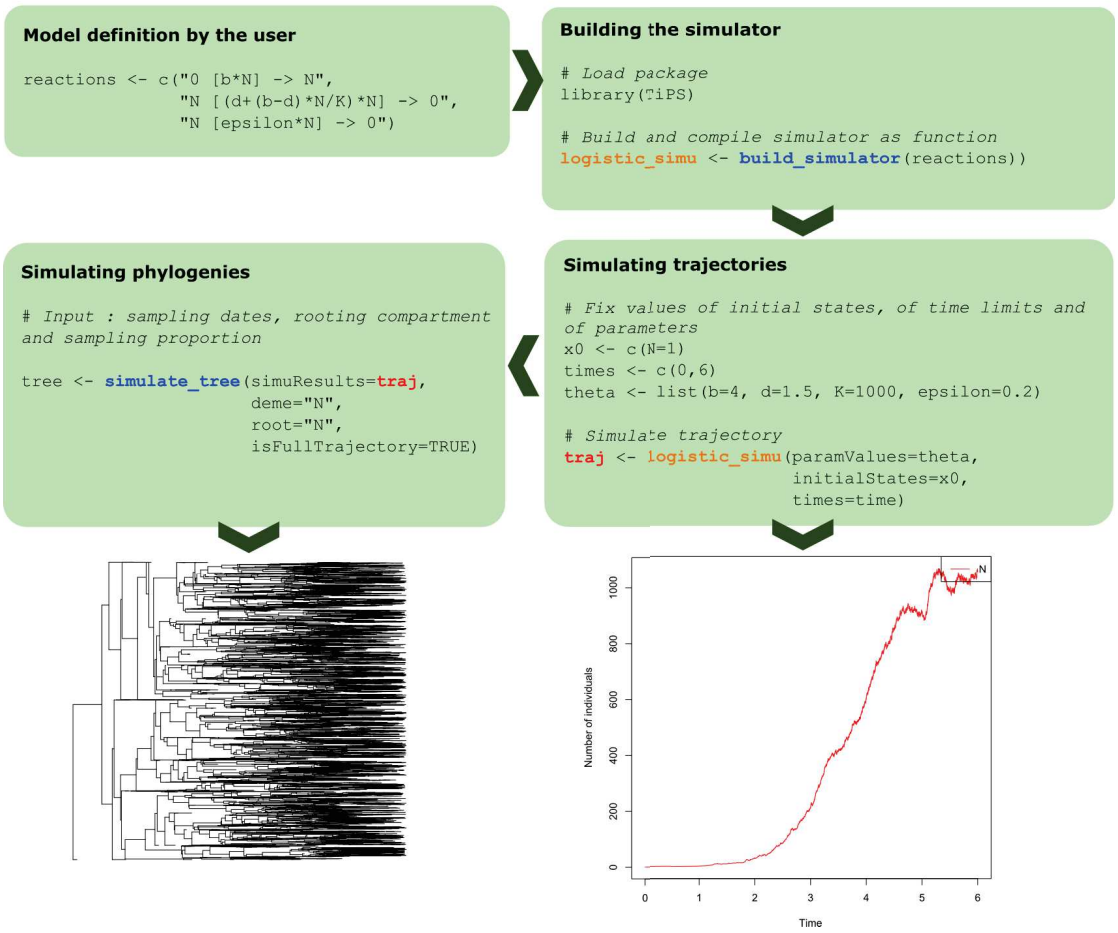


FIGURE S6 – **Simulating a trajectory and a phylogeny using TiPS given a logistic growth model.** The equations and outputs correspond to the logistic growth model. The functions of the R package are in blue. The simulator of trajectories built as a function is in orange. The variable *traj* in red is the output trajectory. The phylogeny is plotted using a function from the ape R package (**?**).

2.3 Vignette du package TiPS

```
library(TiPS)
library(ape)
```

Simulating trajectories

Building the simulator

We use the classical SIR epidemiological model to illustrate the functionning of `TiPS`.

This SIR model can be described by a system of reactions such as

$$S + I \xrightarrow{\beta SI} I + I$$

and

$$I \xrightarrow{\gamma I} R$$

or by a system of differential equations such as

$$\frac{dS}{dt} = -\beta SI$$

and

$$\frac{dI}{dt} = \beta SI - \gamma I$$

In R, with `TiPS`, this model can either be written as

```
reactions <- c("S [beta*S*I] -> I", "I [gamma*I] -> R")
```

Let's now build the simulator:

We then build the simulator that will allow us to run multiple trajectories:

```
sir_simu <- build_simulator(reactions)
```

This typically takes 10-15' as it involves compilation.

Defining simulations parameters

To run numerical simulations, we define the initial values of the state variables,

```
initialStates <- c(I = 1, S = 9999, R = 0)
```

the time range of the simulations,

```
time <- c(0, 20)
```

and the parameters values

```
theta <- list(gamma = 1, beta = 2e-04)
```

For the τ-leap and mixed algorithms, a time step is also required:

```
dT <- 0.001
```

Running simulations

In some simulations, the population size of a deme compartment may be zero before the upper time limit is reached, because of stochasticity or parameter values. In this case, the simulation is considered to have failed and is halted.

To bypass these failures, we can define the following wrapper:

```
safe_run <- function(f, ...) {
    out <- list()
    while (!length(out)) {
        out <- f(...)
    }
    out
}
```

A safe version of our simulator `sir_simu()` is then:

```
safe_sir_simu <- function(...) safe_run(sir_simu, ...)
```

Direct method

The default mode of the simulator is Gillespie's direct method. A trajectory is obtained by

```
traj_dm <- safe_sir_simu(paramValues = theta, initialStates = initialStates,
    times = time)
```

The output consists of a named list containing the reactions of the model (with `$reactions`), the parameter values (with `$values`), the time range (with `$times`), the algorithm used to simulate (with `$algo`), the time-step in case the algorithm is τ-leap or the mixed algorithm (with `$dT`) and finally the simulated trajectory (with `$traj`) :

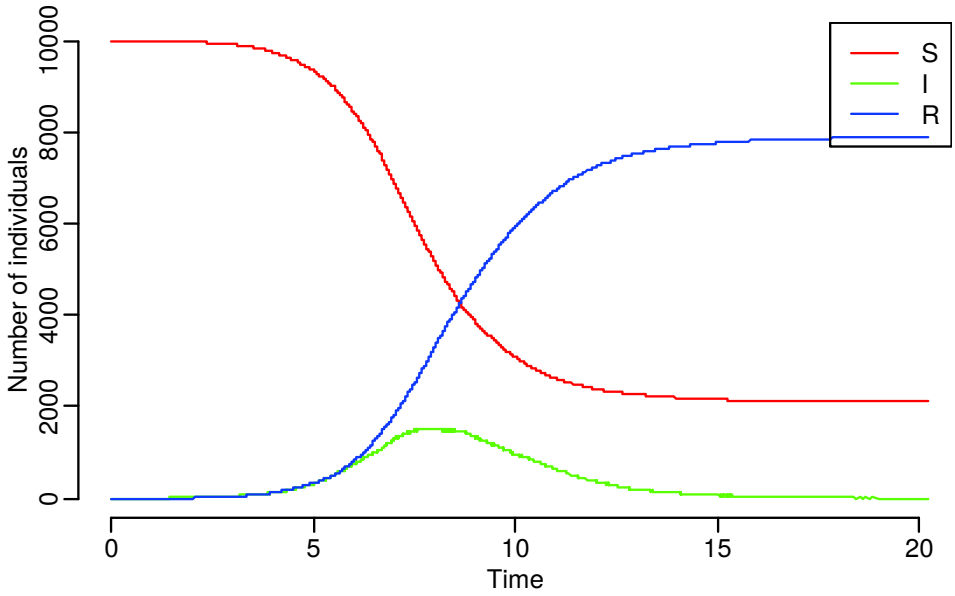
```
names(traj_dm)
#> [1] "reactions" "values"    "times"     "method"    "tau"       "traj"
```

The simulated trajectory is also a named list, where each simulated reaction event is recorded `$Reaction`, along with the time at which it occured `$Time`, the number of times it occured `$Nrep` (if τ-leap or mixed algorithm chosen), and the size of each compartment througt time, here `$I` `$R` `$S`.

```
head(traj_dm$traj)
#>        Time          Reaction Nrep    S I R
#> 1 0.0000000              init    1 9999 1 0
#> 2 0.5781659 S [beta*S*I] -> I    1 9998 2 0
#> 3 0.6882365 S [beta*S*I] -> I    1 9997 3 0
#> 4 0.7014271 S [beta*S*I] -> I    1 9996 4 0
#> 5 0.7268401  I [gamma*I] -> R    1 9996 3 1
#> 6 1.2214185  I [gamma*I] -> R    1 9996 2 2
```

The trajectory can readily be plotted using the `plot()` function:

```
plot(traj_dm)
```



τ-leap method

To use the τ-leap method, the `method` argument must be specified. The time-step `tau` is by default set to 0.05. Let's fix its value to 0.009:

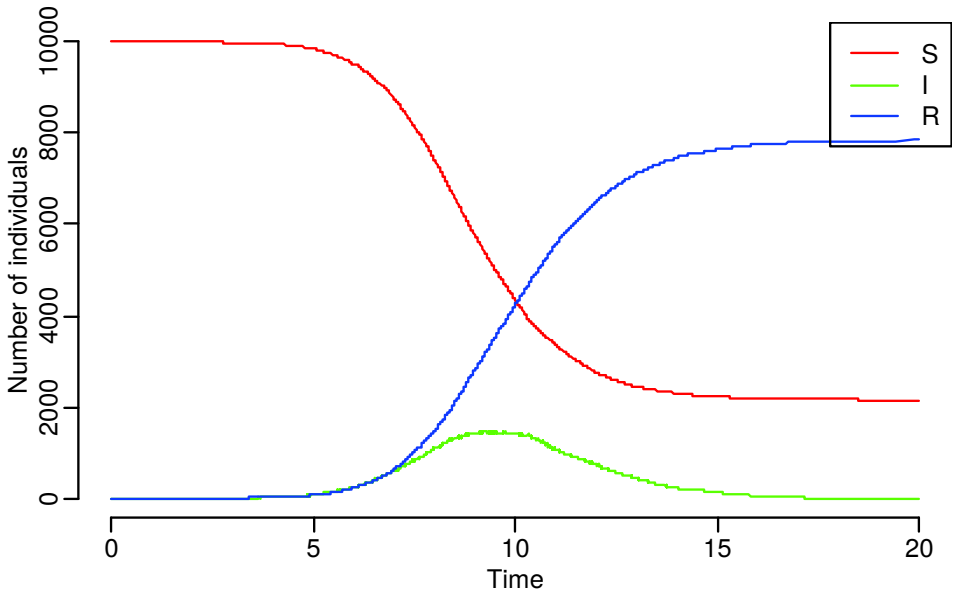
```
traj_tl <- safe_sir_simu(paramValues = theta, initialStates = initialStates,
    times = time, method = "approximate", tau = 0.009)
```

We obtain the same type of output as with the direct method:

```
head(traj_tl$traj)
#>    Time          Reaction Nrep    S I R
#> 1 0.000              init    1 9999 1 0
#> 2 0.054 S [beta*S*I] -> I    1 9998 2 0
#> 3 0.207  I [gamma*I] -> R    1 9998 1 1
#> 4 0.603 S [beta*S*I] -> I    1 9997 2 1
#> 5 0.774 S [beta*S*I] -> I    1 9996 3 1
#> 6 0.801 S [beta*S*I] -> I    1 9995 4 1
```

The trajectory can also be plotted:

```
plot(traj_tl)
```



Mixed method

To run simulations with If a time-step the `switchingMode` algorithm (basically switching between the direct method and the τ-leap method depending on the number of reactions occurring per unit of time), the `method` argument must be specified and fixed to `mixed`. The time-step `tau` is by default set to 0.05. Let's fix its value to 0.009 :

```
traj_mm <- safe_sir_simu(paramValues = theta, initialStates = initialStates,
    times = time, method = "mixed", tau = 0.009, nTrials = 2)
```

Outputs are similar to the other methods:

```
head(traj_mm$traj)
#>         Time          Reaction Nrep    S I R
#> 1 0.00000000              init    1 9999 1 0
#> 2 0.02156502 S [beta*S*I] -> I    1 9998 2 0
#> 3 0.09517016 S [beta*S*I] -> I    1 9997 3 0
#> 4 0.14461423 S [beta*S*I] -> I    1 9996 4 0
#> 5 0.18061818 S [beta*S*I] -> I    1 9995 5 0
#> 6 0.26428439 S [beta*S*I] -> I    1 9994 6 0
```

Comparing the three methods

We can compare the outputs obtained using the three different methods.

For this, we generate 100 stochastic trajectories using each method.

```
nb <- 100
```

```
traj_dm100 <- purrr::rerun(nb, safe_sir_simu(paramValues = theta,
    initialStates = initialStates, times = time)$traj)
```

```
traj_tl100 <- purrr::rerun(nb, safe_sir_simu(paramValues = theta,
    initialStates = initialStates, times = time, method = "approximate",
    tau = dT)$traj)
```

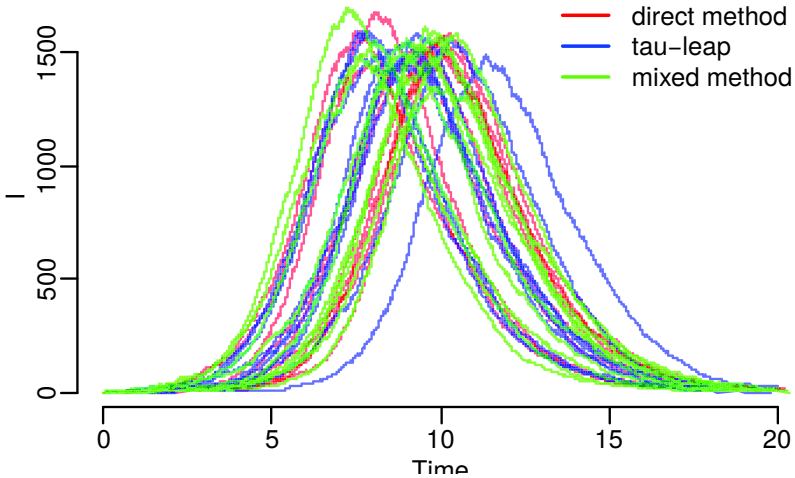
```
traj_mm100 <- purrr::rerun(nb, safe_sir_simu(paramValues = theta,
    initialStates = initialStates, times = time, method = "mixed",
    nTrials = 5, tau = dT)$traj)
```

```
frame <- function(...) {
    plot(NA, xlab = "Time", ylab = "I", xlim = c(0, 20), ...)
}
```

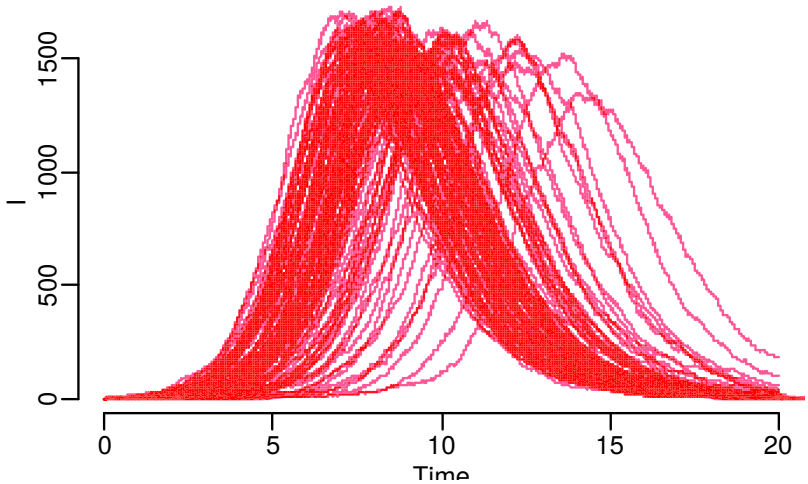
```
ylim <- c(0, max(unlist(purrr::map(c(traj_dm100, traj_tl100, traj_mm100),
    ~.[["I"]]))))
```

```
alpha <- 0.5
```

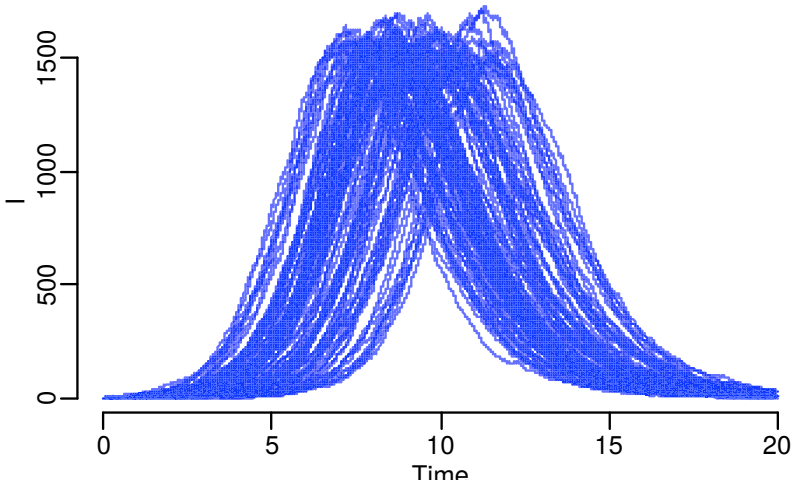
```
frame(ylim = ylim)
purrr::walk(traj_dm100[1:10], with, lines(Time, I, col = adjustcolor("red",
    alpha)))
purrr::walk(traj_tl100[1:10], with, lines(Time, I, col = adjustcolor("blue",
    alpha)))
purrr::walk(traj_mm100[1:10], with, lines(Time, I, col = adjustcolor("green",
    alpha)))
legend("topright", c("direct method", "tau-leap", "mixed method"),
    col = c("red", "blue", "green"), bty = "n", lty = 1, lwd = 2)
```



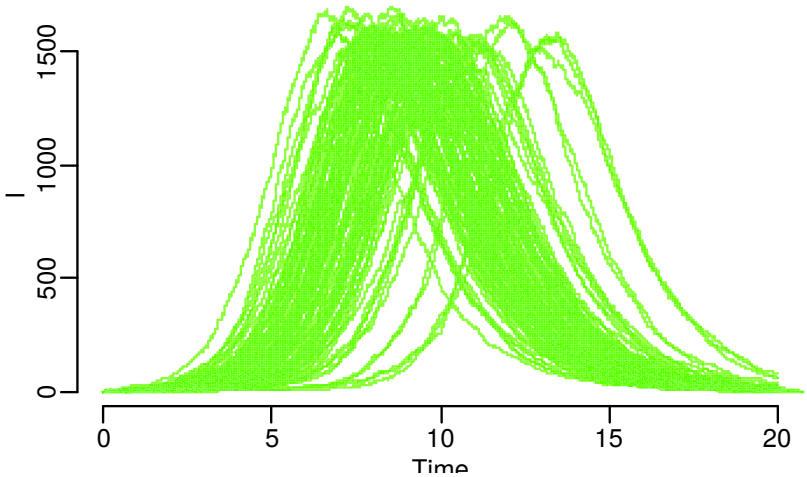

```
frame(ylim = ylim)
purrr::walk(traj_dm100, with, lines(Time, I, col = adjustcolor("red",
    alpha)))
```




```
frame(ylim = ylim)
purrr::walk(traj_tl100, with, lines(Time, I, col = adjustcolor("blue",
    alpha)))
```




```
frame(ylim = ylim)
purrr::walk(traj_mm100, with, lines(Time, I, col = adjustcolor("green",
    alpha)))
```



Simulating phylogenies

A great advantage of `TiPS`, besides its compulational efficiency, is that it can generate phylogenies from the population dynamics trajectories using a coalescent approach.

For this, we may need a vector of sampling dates.

For the SIR example, these are stored here:

```
dates <- system.file("extdata", "SIR-dates.txt", package = "TiPS")
```

The `simulate_tree` function simulates a phylogeny from a trajectory object and using a set of sampling dates:

```
sir_tree <- simulate_tree(
  simuResults = traj_dm,
  dates = dates,
  deme = c("I"), # individuals that contribute to the phylogeny
  sampled = c(I = 1), # individuals that are sampled and proportion of sampling
  root = "I", # type of individuals at the root of the tree
  isFullTrajectory = FALSE, # deads do not generate leaves
  nTrials = 5,
  addInfos = FALSE) # additional info for each node
```

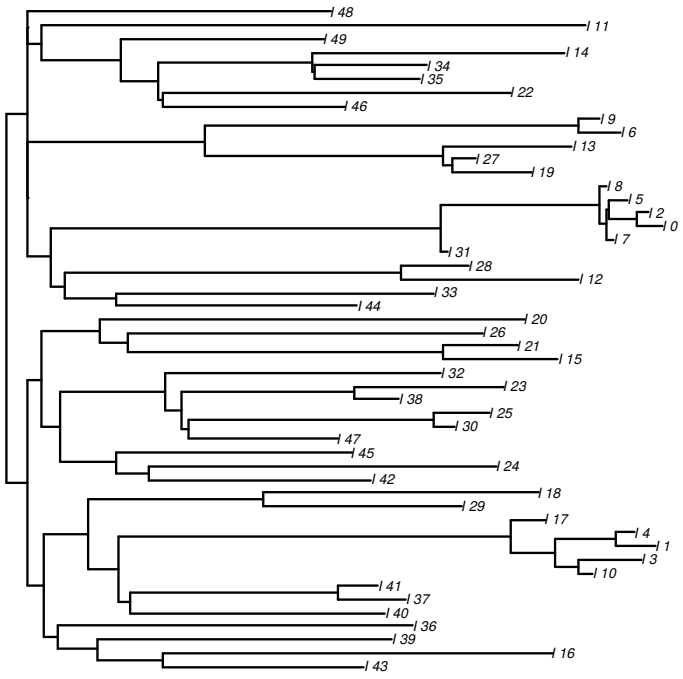
The `sampled` option can be used for labelled phylogenies (i.e. with multiple host types) but it requires specifying the proportion of each label type. The `root` option indicates the state of the individual initiating the dynnamics. See the next section for details.

The full phylogeny can be obtained (therefore neglecting the sampling dates) with the option `isFullTrajectory`.

Finally, some runs may fail to simulate a phylogeny from a trajectory for stochastic reasons. The `nTrials` parameter indicates the number of unsuccessful trials allowed before giving up.

The simulated phylogeny can be visualised using:

```
ape::plot.phylo(sir_tree, cex = 0.5)
```



Multiple demes

We sometimes have multiple demes, i.e. different types of individuals that contribute to the pylogeny or that can be sampled (e.g. juveniles vs. adults or acute vs. chronic infections).

We illustrate this example using an SIR model with two patches (labelled 1 and 2) and migration between these patches (at a rate μ).

Initialising the system

$$\frac{dI_1}{dt} = \beta_1 I_1 - \gamma_1 I_1 - \mu_1 I_1 + \mu_2 I_2 \frac{dI_2}{dt} = \beta_2 I_2 - \gamma_2 I_2 - \mu_2 I_2 + \mu_1 I_1$$

The associated reactions are:

```
reactions <- c("0 [beta1 * I1] -> I1", "I1 [gamma1 * I1] -> 0", "I1 [mu1 * I1] -> I2",
    "0 [beta2 * I2] -> I2", "I2 [gamma2 * I2] -> 0", "I2 [mu2 * I2] -> I1")
```

We then build the simulator:

```
bd_simu <- build_simulator(reactions)
```

The initial state variables values are

```
initialStates <- c(I1 = 0, I2 = 1)
```

The time range of the simulation is between 1975 and 2018:

```
time <- c(1975, 1998, 2018)
```

the parameters values are

```
theta <- list(gamma1 = c(0.2, 0.09), gamma2 = 0.1, mu1 = 0.025, mu2 = 0.021,
    beta1 = c(0.26, 0.37), beta2 = 0.414)
```

and the time step (for the τ-leap and mixed algorithms) is:

```
dT <- 0.01
```

A safe version of the simulator `bd_simu()` is:

```
safe_bd_simu <- function(...) safe_run(bd_simu, ...)
```

Tau-leap trajectory simulation

We perform the simulations using:

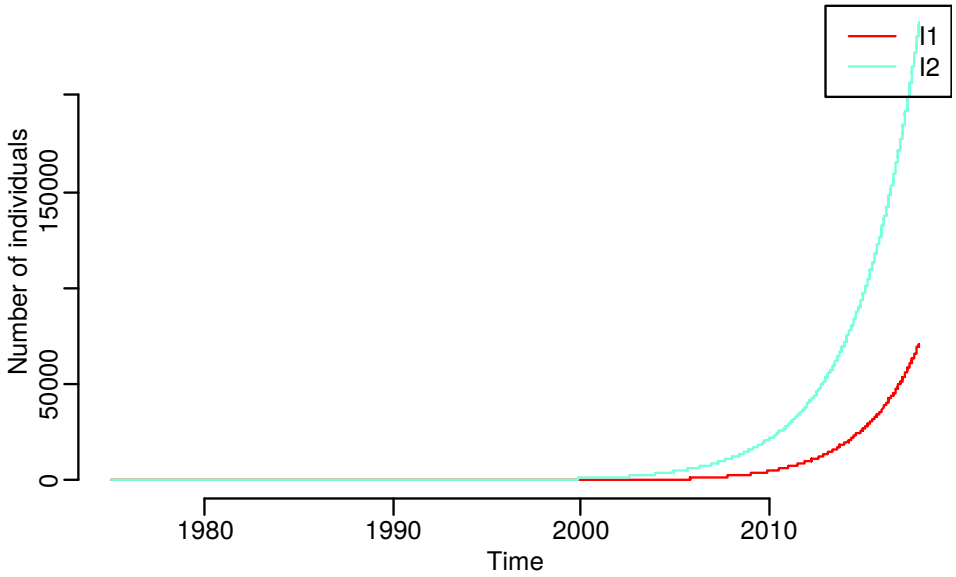
```
trajbd_tl <- safe_bd_simu(paramValues = theta, initialStates = initialStates,
    times = time, method = "approximate", tau = 0.001)
```

We obtain:

```
head(trajbd_tl$traj)
#>       Time              Reaction Nrep I1 I2
#> 1 1975.000                  init    1  0  1
#> 2 1977.592  0 [beta2 * I2] -> I2    1  0  2
#> 3 1977.909 I2 [gamma2 * I2] -> 0    1  0  1
#> 4 1981.946  0 [beta2 * I2] -> I2    1  0  2
#> 5 1982.568  0 [beta2 * I2] -> I2    1  0  3
#> 6 1982.706  0 [beta2 * I2] -> I2    1  0  4
```

Graphically, we get:

```
plot(trajbd_tl)
```



Phylogeny simulation

With known sampling dates and known proportion of sampling

Instead of loading a vector, we assume we have 100 samples at 100 sampling dates between 2015 and 2018. We can generate the dates vector as:

```
dates_bd <- seq(from = 2015, to = 2018, length.out = 100)
```

We then simulate a phylogeny where 20% of the sampling dates correspond to the I1 compartment, and 80% to the I2 compartment:

```
bd_tree <- simulate_tree(
  simuResults = trajbd_tl,
  dates = dates_bd,
  deme = c("I1", "I2"),
  sampled = c(I1 = 0.2, I2 = 0.8), # individuals that are sampled and proportion of sampling
  root = "I2", # type of individual at the root of the tree
  isFullTrajectory = FALSE, # deads do not generate leaves
  nTrials = 3,
  addInfos = TRUE) # additional info for each node
```

This is done using a coaslescence process informed by the trajectory. Therefore, each internal node of the phylogeny corresponds to a coalescence event and is associated with a label stoed in `$node.label`.

In our two-patches example, there are two types of coalesence: I2 individuals creating a new I2 individual, and I1 individuals creating a new I1 individual.

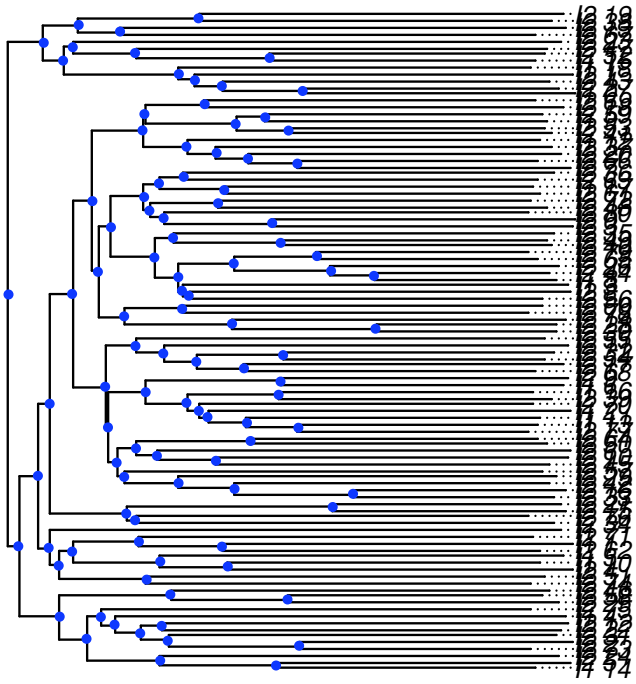
We can plot the phylogeny and color the internal nodes based on the type of coalescence.

First we generate a vector of colors for the nodes (if we find `I2` in the node label we color it in blue, otherwise in red):

```
inode_cols <- ifelse(grepl(x = bd_tree$node.label, pattern = "I2"),
    "blue", "red")
```

Then we plot the phylogeny:

```
ape::plot.phylo(bd_tree, root.edge = T, no.margin = F, align.tip.label = T)
nodelabels(pch = 20, col = inode_cols)
```



With known sampling dates, each assigned to a compartment by the user

One can give as input, sampling dates assigned to a compartment, in which case the option `sampled` is not required.

```
dates_bd <- seq(from = 2015, to = 2018, length.out = 100)
dates_bd <- data.frame(Date = sample(dates_bd), Comp = c(rep("I1",
    20), rep("I2", 80)))
head(dates_bd)
#>       Date Comp
#> 1 2017.212   I1
#> 2 2017.818   I1
#> 3 2015.788   I1
#> 4 2015.939   I1
#> 5 2015.667   I1
#> 6 2016.030   I1
```

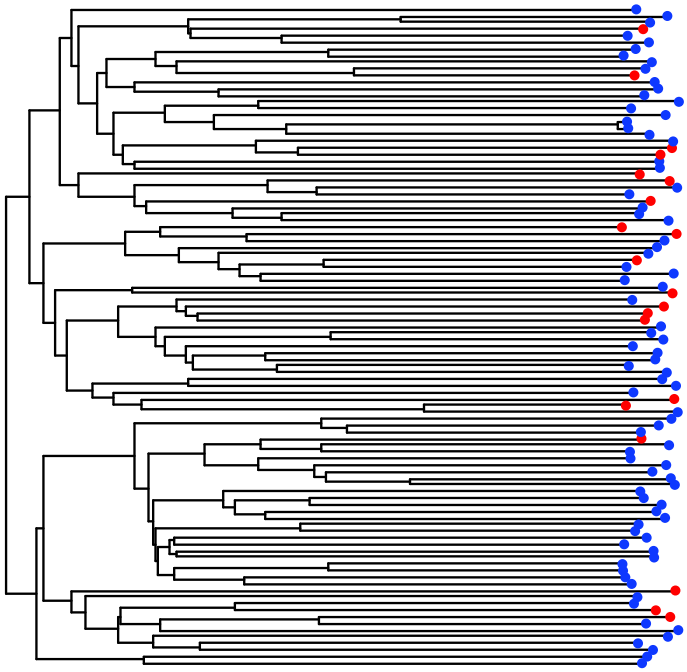
Now let's simulate a phylogeny with sampling dates assigned to a compartment by the user.

```
bd_tree <- simulate_tree(
  simuResults = trajbd_tl,
  dates = dates_bd,
  deme = c("I1", "I2"),
  root = "I2", # type of individual at the root of the tree
  nTrials = 3,
  addInfos = TRUE) # additional info for each node
```

We can plot the phylogeny and color the external nodes given the compartment.

```
tips_cols <- ifelse(grepl(x = bd_tree$tip.label, pattern = "I2"),
    "blue", "red")
```

```
ape::plot.phylo(bd_tree, root.edge = T, no.margin = F, show.tip.label = F)
tiplabels(pch = 20, col = tips_cols)
```



With only known sampling dates

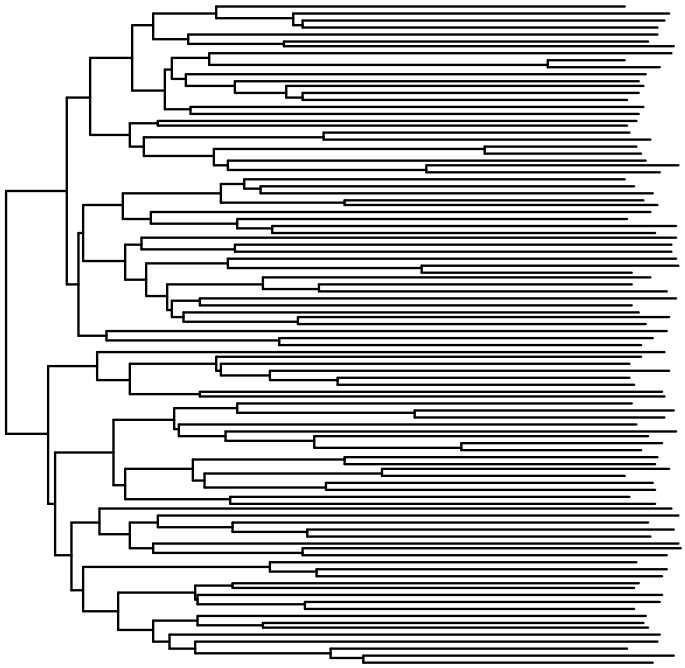
In the case where the user has no information on the sampling proportions or the assignment of sampling dates on any compartment, the algorithm will randomly assign each sampling date to a compartment. The user gives as input sampling dates:

```
dates_bd <- seq(from = 2015, to = 2018, length.out = 100)
```

Now let's simulate a phylogeny with sampling dates and no information about the sampling schemes :

```
bd_tree <- simulate_tree(
  simuResults = trajbd_tl,
  dates = dates_bd,
  sampled = c("I1","I2"),
  deme = c("I1", "I2"),
  root = "I2", # type of individual at the root of the tree
  nTrials = 10,
  addInfos = TRUE) # additional info for each node
```

```
ape::plot.phylo(bd_tree, root.edge = T, no.margin = F, show.tip.label = F)
```

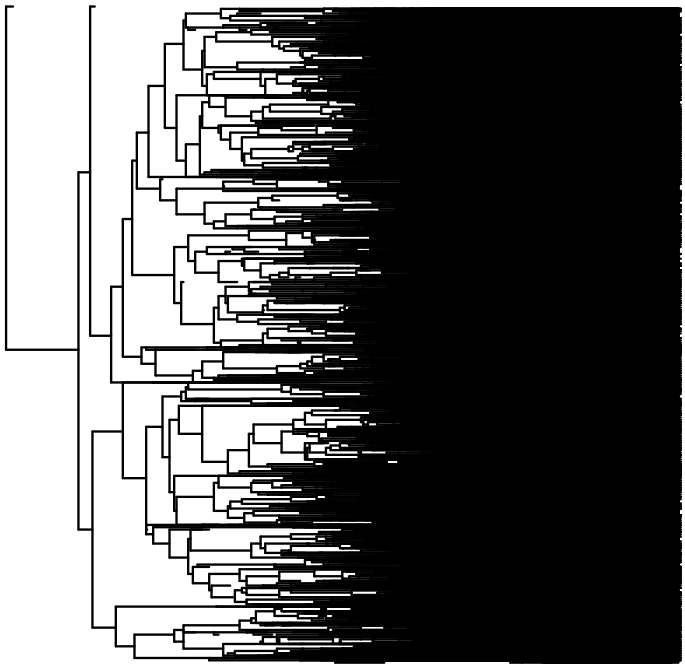


Without given sampling dates

Let's simuate a phylogeny where simulated deaths in the trajectory generate leaves. This can be done with the `isFullTrajectory` option.

```
bd_tree <- simulate_tree(
  simuResults = trajbd_tl,
  deme = c("I1", "I2"),
  root = "I2", # type of individual at the root of the tree
  nTrials = 10,
  isFullTrajectory = TRUE, # deads generate leaves
  addInfos = TRUE) # additional info for each node
```

```
ape::plot.phylo(bd_tree, root.edge = T, no.margin = F, show.tip.label = F)
```



References

- For further details see Danesh G, Saulnier E, Gascuel O, Choisy M, Alizon S. 2020. Simulating trajectories and phylogenies from population dynamics models with TiPS. *bioRxiv*, 2020.11.09.373795. DOI: 10.1101/2020.11.09.373795.

- This work was supported by a doctoral grant from la Fondation pour la Recherche Médicale (FRM grant number ECO20170637560) to Gonché Danesh.

Chapitre 3

Phylodynamique du virus de l'hépatite C au sein d'une population hétérogène, par ABC-régression

3.1 Résumé

L'hépatite C est une maladie infectieuse du foie causée par le virus de l'hépatite C (VHC). Ce virus peut provoquer des infections aiguës et des infections chroniques. La phase aiguë de l'infection est généralement définie comme les 6 premiers mois suivant l'exposition au virus. Bien que certains patients présentent des symptômes d'hépatite aiguë, la plupart des personnes infectées sont asymptomatiques. Par conséquent, de nombreux patients ne sont pas conscients de l'infection jusqu'à ce qu'elle évolue vers une infection chronique et peuvent ne développer des symptômes que des décennies plus tard avec l'apparition d'une cirrhose. Environ 30% des personnes infectées se débarrassent spontanément du virus dans les 6 mois qui suivent l'infection, sans recevoir de traitement. Par conséquent, le traitement n'est pas systématique chez un patient détecté en phase aiguë, mais ne s'impose que lorsque l'infection par le virus devient chronique. Le traitement actuel par des antiviraux à action directe (AAD), d'une durée de 12 à 24 semaines, permet d'éradiquer le virus de l'hépatite C dans 90% des cas. Cependant ce traitement reste relativement cher et difficilement accessible aux populations les plus vulnérables. De plus, les réinfections restent possibles, notamment dans les populations à risque.

Le virus de l'hépatite C se transmet par le sang. Il peut se transmettre par l'intermédiaire d'aiguilles, de seringues ou autre matériel non (ou pas complètement) stérilisé utilisé pour le tatouage, l'injection de drogues ou en milieu hospitalier. Le

virus peut aussi se transmettre par voie sexuelle. Bien que ce mode de transmission soit rare, le risque augmente lorsque du sang est échangé (menstruations, blessures au niveau des voies génitales ou anales).

L'infection chronique par le virus de l'hépatite C touche plus de 70 millions de personnes dans le monde. L'organisation mondiale de la santé (OMS) et certains pays ont mis en place des comités nationaux pour viser l'élimination du VHC d'ici 2030 (**?**). Les objectifs de l'OMS comprenaient une réduction de 90% des nouvelles infections par le VHC, le traitement de 80% des infections chroniques par le VHC et une réduction de 65% de la mortalité causée par l'hépatite. La co-infection par le VHC est fréquente chez les patients infectés par le VIH, les virus ayant des modes de transmission en commun. Cette co-infection engendre une accélération de la progression de la maladie entraînant des taux de mortalité plus élevés. Par conséquent, cette population a été considérée comme prioritaire pour l'élimination du VHC. En France, et d'autres pays Européens, les interventions de réduction des risques liés à l'échange de matériel non stérilisé et un accès facilité aux traitements de substitution aux opiacés ont entraîné, au cours des dernières années, une diminution drastique de la prévalence des infections par le VHC, notamment chez les patients co-infectés (**?**). Cependant, depuis le début des années 2010 une nouvelle émergence de la transmission sexuelle du VHC est observée, notamment chez des hommes ayant des rapports sexuels avec des hommes (HSH), population auparavant peu touchée. Cette émergence est due à des rapports sexuels non protégés, à des pratiques sexuelles provoquant des saignements et à la consommation de drogues dans le cadre de rapports sexuels. Cette émergence entraîne une augmentation de l'incidence du VHC, notamment dans la région de la ville de Lyon (**?**). La dynamique de l'épidémie classiquement observée chez les usagers de drogues et dans la population générale a été largement étudiée (**??**). Cependant, les résultats de ces populations ne sont pas facilement applicables à d'autres populations comme les HSH.

Comprendre et quantifier les caractéristiques épidémiologiques liées à cette émergence de transmission du VHC chez les HSH est un enjeu de santé publique pour pouvoir mettre en place des interventions. Cependant, le suivi de la population à risque étant limité, une analyse quantitative est difficile à réaliser. De plus, les données d'incidence classiquement utilisées en épidémiologie ne sont pas assez informatives pour comprendre la propagation du virus dans une population hétérogène présentant des profils différents. Pour contourner ce problème, nous avons réalisé une étude phylodynamique en analysant des données de séquence du VHC.

Cette étude est née d'une collaboration avec des cliniciens du CHU de Lyon qui ont partagé avec nous 213 séquences, chacune associée à une date d'échantillonnage et un type d'hôte : soit un hôte dit classique soit un hôte dit nouveau. Les hôtes

classiques sont des patients infectés par transmission nosocomiale ou qui ont des antécédents d'usage de drogue. Les nouveaux hôtes sont des hommes qui ont des rapports sexuels avec des hommes, séropositifs ou non, faisant potentiellement usage de drogues et dont l'infection a été détectée pendant ou peu après la phase aiguë. Ces profils ont été établis par des épidémiologistes de terrain à partir de questionnaires et de facteurs de risque. Les dates d'échantillonnage s'étendent de 2011 à 2018.

La particularité de cette étude phylodynamique est qu'elle prend en compte l'hétérogénéité de la population et estime simultanément, à partir d'une seule phylogénie, des paramètres épidémiologiques, tels que le nombre de reproduction effectif ou la durée de la période d'infection, pour chacune des deux épidémies observées.

Notre analyse phylodynamique s'appuie sur la méthode d'inférence par ABC-régression introduite par (**?**). Cette méthode a été validée pour un modèle simple SIR. Nous avons étendu cette approche pour un modèle structuré en développant de nouvelles statistiques de résumé à calculer sur des phylogénies labellisées. En effet, chaque feuille de la phylogénie construite à partir des séquences est associée à un type d'hôte. Les nouvelles statistiques de résumé permettent de capturer des informations concernant par exemple l'hétérogénéité de la population ou le niveau de brassage entre les types d'hôtes. Une analyse de validation croisée combinée à une analyse de *bootstrap* paramétrique nous a permis de valider notre méthode ainsi que nos résultats.

Les résultats de l'étude suggèrent que la propagation actuelle du virus est dix fois plus rapide chez les nouveaux hôtes que chez les hôtes classiques. La durée de la période d'infection chez les nouveaux hôtes a été estimée à 5 mois, ce qui suggère que la majorité des transmissions se produit pendant la phase aiguë de leur infection, avant le traitement. Ces résultats soulignent la nécessité d'agir rapidement lors de la détection, par exemple en soulignant l'importance des mesures de protection telles que l'utilisation du préservatif et en initiant le traitement même pendant la phase aiguë.

Une limite de l'étude est liée au schéma d'échantillonnage de la population. En effet, la proportion de nouveaux hôtes infectés échantillonnés est inconnue mais pourrait être élevée. Pour les hôtes classiques, nous avons sélectionné un sous-ensemble représentatif des patients détectés dans la région de Lyon mais il se peut que cet échantillonnage soit faible. Pour analyser l'effet du biais d'échantillonnage, nous avons réalisé des analyses sur différentes phylogénies en utilisant la moitié des séquences associées aux nouveaux hôtes. Nous avons trouvé des résultats similaires à ceux obtenus avec l'ensemble de la phylogénie, suggérant que notre méthode par ABC-régression n'est pas sensible aux biais d'échantillonnage.

Cette étude constitue l'objet d'un article qui a été recommandé après processus de

revue par les pairs par *Peer Community In Evolutionary Biology*. Il est actuellement en processus de revue dans PLOS Pathogens. L'article est présenté dans la section 3.2.

3.2 Quantifying transmission dynamics of acute hepatitis C virus infections in a heterogeneous population using sequence data

Gonché Danesh[1,*], Victor Virlogeux[2], Christophe Ramière[3], Caroline Charre[3], Laurent Cotte[4], Samuel Alizon[1]

[1] MIVEGEC, CNRS, IRD, Université de Montpellier
[2] Clinical Research Center, Croix-Rousse Hospital, Hospices Civils de Lyon – Lyon, France
[3] Virology Laboratory, Croix-Rousse Hospital, Hospices Civils de Lyon – Lyon, France
[4] Infectious Diseases Department, Croix-Rousse Hospital, Hospices Civils de Lyon – Lyon, France
[*] corresponding author: gonche.danesh@gmail.com

3.2.1 Background

It is estimated that 71 million people worldwide suffer from chronic hepatitis C virus (HCV) infections (**??**). The World Health Organisation (WHO) and several countries have issued recommendations towards the 'elimination' of this virus, which they define as an 80% reduction in new chronic infections and a 65% decline in liver mortality by 2030 (**?**). HIV-HCV coinfected patients are targeted with priority because of the shared transmission routes between the two viruses (**?**) and because of the increased virulence of HCV in coinfections (**???**). Successful harm reduction interventions, such as needle-syringe exchange and opiate substitution programs, as well as a high level of enrolment into care programs for HIV-infected patients, have led to a drastic drop in the prevalence of active HCV infections in HIV-HCV coinfected patients in several European countries during the recent years (**????**). Unfortunately, this elimination goal is challenged by the emergence of HCV sexual transmission, especially among men having sex with men (MSM). This trend is reported to be driven by unprotected sex, drug use in the context of sex ('chemsex'), and potentially traumatic practices such as fisting (**???**). The epidemiology of HCV infection in the Dat'AIDS cohort has been extensively described from 2000 to 2016 (**???**). The incidence of acute HCV infection has been estimated among HIV-infected MSM between 2012 and 2016, among HIV-negative MSM enrolled in PrEP between in 2016-2017 (**?**) and among HIV-infected and HIV-negative MSMs from 2014 to 2017 (**?**). In the area of Lyon (France), HCV incidence has been shown to increase concomitantly with a shift in the profile of infected hosts (**?**). Understanding and

quantifying this recent increase is the main goal of this study.

Several modelling studies have highlighted the difficulty to control the spread of HCV infections in HIV-infected MSMs in the absence of harm reduction interventions (**??**). Furthermore, we recently described the spread of HCV from HIV-infected to HIV-negative MSMs, using HIV pre-exposure prophylaxis (PrEP) or not, through shared high-risk practices (**?**). More generally, an alarming incidence of acute HCV infections in both HIV-infected and PrEP-using MSMs was reported in France in 2016-2017 (**?**). Additionally, while PrEP-using MSMs are regularly screened for HCV, those who are HIV-negative and do not use PrEP may remain undiagnosed and untreated for years. In general, we know little about the population size and practices of HIV-negative MSM who do not use PrEP. All these epidemiological events could jeopardize the goal of HCV elimination by creating a large pool of infected and undiagnosed patients, which could fuel new infections in intersecting populations. Furthermore, the epidemiological dynamics of HCV infection have mostly been studied in intravenous drug users (IDU) (**????**) and the general population (**??**). Results from these populations are not easily transferable to other populations, which calls for a better understanding of the epidemiological characteristics of HCV sexual transmission in MSM.

Given the lack of knowledge about the focal population driving the increase in HCV incidence, we analyse virus sequence data with phylodynamics methods. This research field has been blooming over the last decade and hypothesizes that the way rapidly evolving viruses spread leaves 'footprints' in their genomes (**???**). By combining mathematical modelling, statistical analyses and phylogenies of infections, where each leaf corresponds to the virus sequence isolated from a patient, current methods can infer key parameters of viral epidemics. This framework has been successfully applied to other HCV epidemics (**????**), but the ongoing one in Lyon is challenging to analyze because the focal population is heterogeneous, with 'classical' hosts (typically HIV-negative patients infected through nosocomial transmission or with a history of opioid intravenous drug use or blood transfusion) and 'new' hosts (both HIV-infected and HIV-negative MSM, detected during or shortly after acute HCV infection phase, potentially using recreational drugs such as cocaine or cathinones), where host profiles have been established by field epidemiologists based on interviews and risk factors. Our phylodynamics analysis relies on an Approximate Bayesian Computation (ABC, (**?**)) framework that was recently developed and validated using a simple Susceptible-Infected-Recovered (SIR) model (**?**).

Assuming an epidemiological transmission model with two host types, 'classical' and 'new' (see the Methods), we use dated virus sequences to estimate the date of onset of the HCV epidemics in 'classical' and 'new' hosts, the level of mixing between hosts types, and, for each host type, the duration of the infectious period and the effective reproduction ratio (i.e. the number of secondary infections, (**?**)). To validate our results we performed a parametric bootstrap analysis, we tested the sensitivity of the method to differences in sampling proportions between the two types of hosts. We also tested the sensitivity of the method to phylogenetic reconstruction uncertainty, and we performed a cross-validation analysis to explore the robustness of our inference framework. We find that the doubling time of the epidemics is one order of magnitude lower in 'new' than in 'classical' hosts, therefore emphasising the urgent need for public health action.

3.2.2 Results

The phylogeny inferred from the dated virus sequences shows that 'new' hosts (in red) tend to be grouped in clades (Figure 3.1). This pattern suggests a high degree of assortativity in the epidemics (i.e. hosts tends to infect hosts from the same type). The ABC phylodynamics approach allows us to go beyond a visual description and to quantify several epidemiological parameters.

As for any Bayesian inference method, we need to assume a prior distribution for each parameter. These priors, shown in grey in Figure 3.2, are voluntarily designed to be large and uniformly distributed to be as little informative as possible. One exception is the date of onset of the epidemics, for which we use the output of the phylogenetic analysis conducted in Beast (see the Methods) as a prior. We also assume the date of the 'new' hosts epidemics to be after 1997 based on epidemiological data.

The inference method converges towards posterior distributions for each parameter, which are shown in red in Figure 3.2. The estimate for the origin of the epidemic in 'classical' hosts is $t_0 = 1957.47$ $[1948.61; 1961.96]$ (numbers in brackets indicate the 95% Highest Posterior Density, or HPD). For the 'new' host type, we were not able to estimate when the epidemic (t_2) has started.

We find the level of assortativity between host types to be high for 'classical' ($a_1 = 0.94$ $[0.83; 1.0]$) as well as for 'new' hosts ($a_2 = 0.92$ $[0.81; 0.99]$). Therefore, hosts mainly infect hosts from the same type.

The phylodynamics approach also allows us to infer the duration of the infectious period for each host type. Assuming that this parameter does not vary over time, we estimate it to be 3.85 years $[1.09; 8.33]$ for 'classical' hosts (parameter $1/\gamma_1$) and 0.45

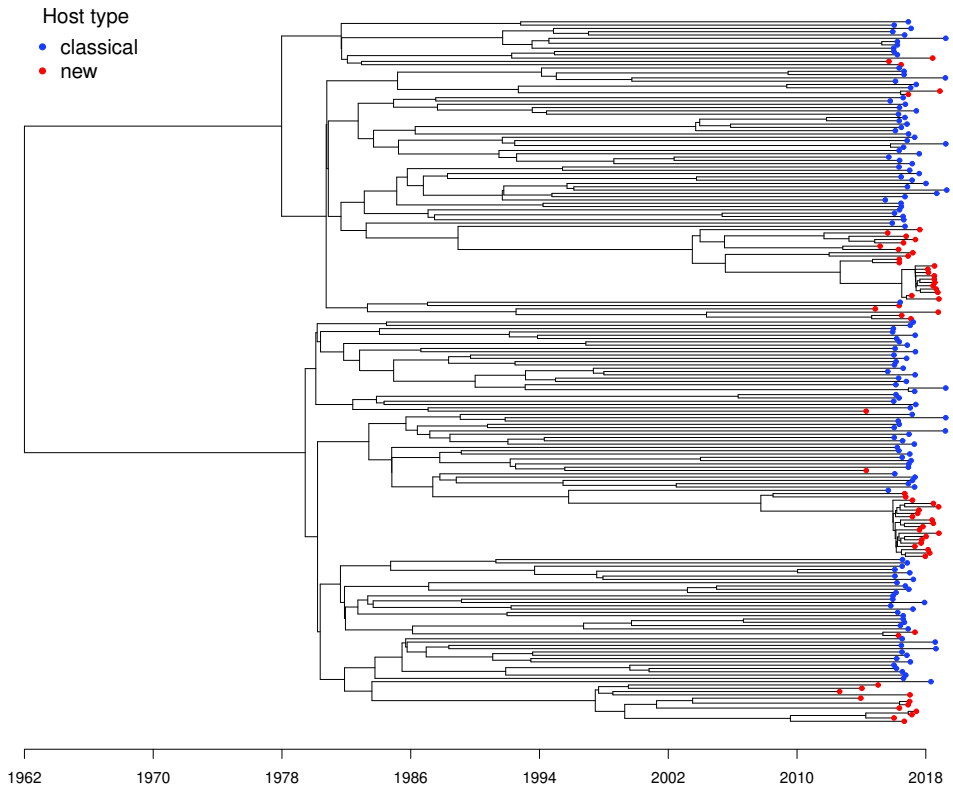


Figure 3.1 – **Phylogeny of HCV infections in the area of Lyon (France).** 'Classical' hosts are in blue and 'new' hosts are in red. Sampling events correspond to the end of black branches. The phylogeny was estimated using Bayesian inference (Beast2). See the Methods for additional details.

years $[0.30; 0.77]$ for 'new' hosts (parameter $1/\gamma_2$). We compute the ratio of γ_2/γ_1 and the 95% credibility interval does exclude 1.

Regarding effective reproduction numbers, i.e. the number of secondary infections caused by a given host over its infectious period, we estimate that of 'classical' hosts to have decreased from $R^{(1),t_1} = 1.96$ $[1.45; 3.29]$ to $R^{(1),t_2} = 1.61$ $[1.05; 2.08]$ after the introduction of the third-generation HCV test in 1997. The inference on the

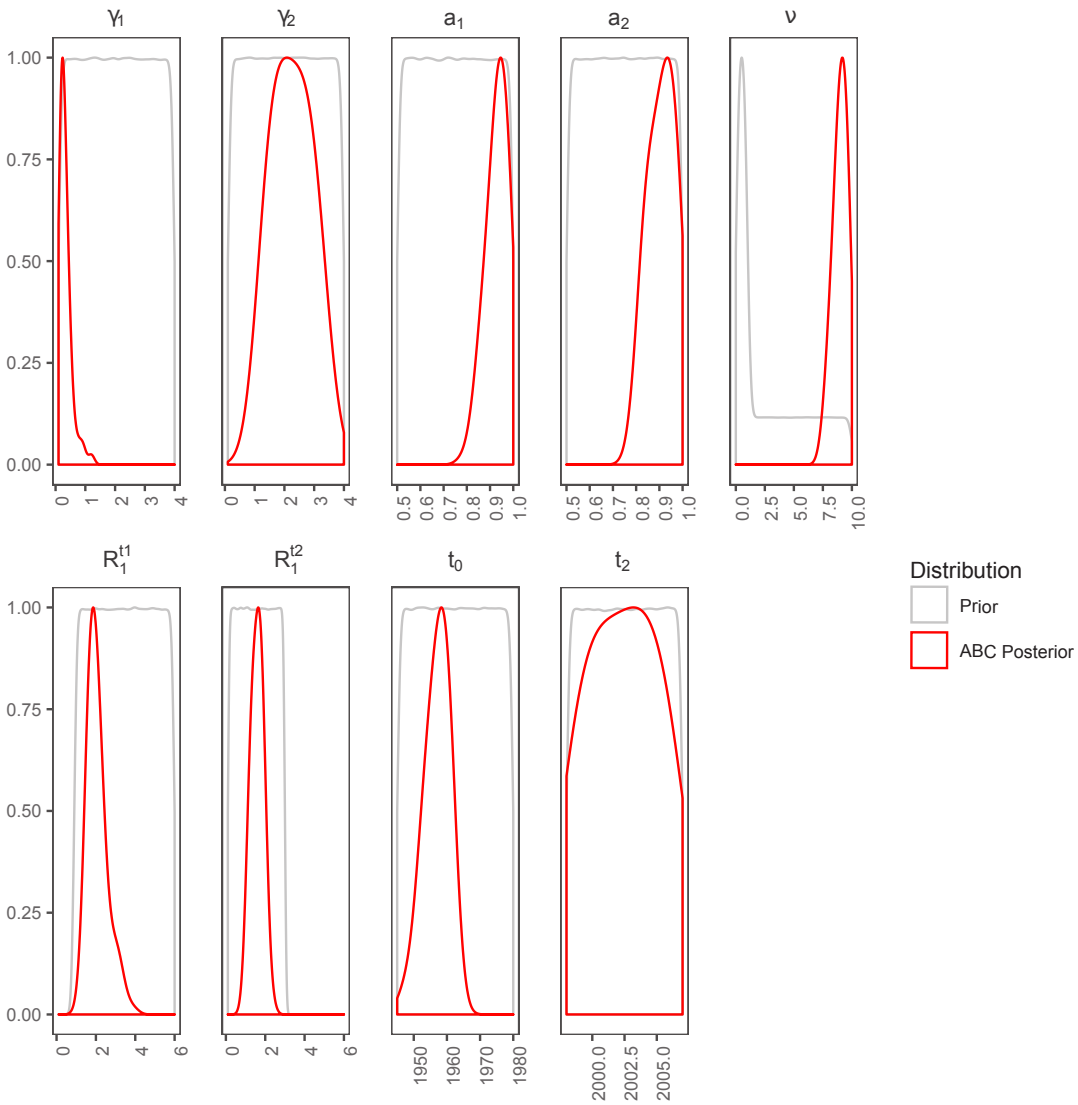


Figure 3.2 – **Parameter prior and posterior distributions.** Prior distributions are in grey and posterior distributions inferred by ABC are in red. The thinner the posterior distribution width, the more accurate the inference. Posterior distributions are truncated based on the prior distribution.

differential transmission parameter indicates that HCV transmission rate is $\nu = 9.0\ [7.7; 9.9]$ times greater from 'new' hosts than from 'classical' hosts. By combining

these results (see the Methods), we compute the effective reproduction number in 'new' hosts and find $R^{(2),t_3} = 1.73$ $[1.03; 4.32]$. We compute the ratio of the $R(t)$ of 'new' hosts over the $R(t)$ of 'classical' hosts after 1997 and, the median value is 1.14 and the 95% credibility interval is $[0.56; 3.25]$.

To better understand the differences between the two host types, we compute the epidemic doubling time (t_D), which is the time for an infected population to double in size. t_D is computed for each type of host, assuming complete assortativity (see the Methods). We find that for the 'classical' hosts, before 1997 $t_D^{(1),t1} \approx 2.8$ years ($[1.1; 5.0]$ years). After 1997, the pace decreases with a doubling time of $t_D^{(1),t2} \approx 4.4$ years ($[2.0; 20.8]$ years). For the epidemics in the 'new' hosts, we estimate that $t_D^{(2),t3} \approx 0.44$ years ($[0.09; 8.84]$ years). When computing the ratio of the doubling times of classical hosts after 1997 over the doubling times of the new hosts ($t_D^{(1),t2}/t_D^{(2),t3}$) to estimate the current difference we find that $t_D^{(1),t2}$ is 10 times higher than $t_D^{(2),t3}$ with a 95% credibility interval of $[0.62; 149.99]$. However, the 75% credibility interval does exclude 1 and is $[3.39; 25.61]$. Distributions for theses three doubling times are shown in Supplementary Figure S2.

Supplementary Figure S3 shows the correlations between parameters based on the posterior distributions. We mainly find that the R_t in 'classical' hosts after the introduction of the third generation of HCV detection tests (i.e. $R^{(1),t_2}$) is negatively correlated to ν and positively correlated to γ_2. In other words, if the epidemic spreads rapidly in 'classical' hosts, it requires a slower spread in 'new' hosts to explain the phylogeny. $R_0^{(1),t_2}$ is also slightly negatively correlated to γ_1, which most likely comes from the fact that for a given R_0, epidemics with a longer infection duration have a lower doubling time and therefore a weaker epidemiological impact. Overall, these correlations do not affect our main results, especially the pronounced difference in infection periods (γ_1 and γ_2).

To validate these results, we performed a goodness-of-fit test by simulating phylogenies using the resulting posterior distributions to determine whether these are similar to the target dataset (see the Methods). In Figure 3.3, we see that the target data in red, i.e. the projection of the observed summary statistics from the phylogeny shown in Figure 3.1, is contained in the envelope containing 90% of the simulations drawn from the posterior distributions. If we use the 95% HPD of the posterior but assume a uniform distribution instead of the true posterior distribution, we find that the target phylogeny is not contained in the envelope. These results confirm that the posterior distributions we infer are highly informative. In Supplementary Figure S4 we show that for 77 summary statistics out of 101, the target value is in the 95% highest posterior distribution of summary statistics computed from the 10,000 simulated phylogenies from the posterior distribution used for the goodness-of-fit

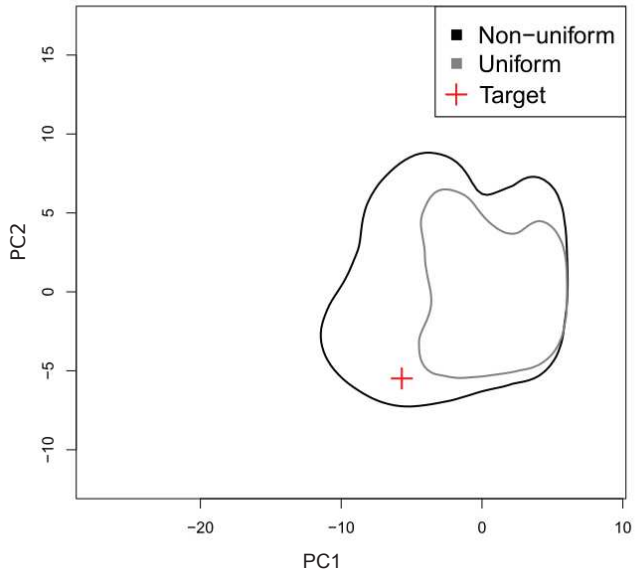


Figure 3.3 – **Goodness-of-fit estimated using parameter bootstrap.** The graph displays envelopes containing 90% of the 10,000 simulations for each distribution. The envelope in black results from the posterior distribution, in grey, results from the uniform distribution drawn from the 95% HPD distribution. The target data is represented by a red cross. Axes units are based on the outcome of principal component analysis using the simulated summary statistics.

test.

To further explore the robustness of our inference method, we use simulated data to perform a 'leave one out' cross-validation (see the Methods). As shown in Supplementary Figure S5, the mean relative error made for each parameter inference is limited and comparable to what was found using a simpler SIR model (**?**). One exception is for the 'new' hosts' level of assortativity (a_2). This is likely due to the poor signal given the small size of the observed phylogeny.

A potential issue is that the sampling rate of 'new' hosts may be higher than that of 'classical' hosts. To explore the effect of such sampling biases on the accuracy of our results, we sub-sampled the 'new' hosts population by pruning the target phylogeny, i.e. randomly removing 50% of the 'new' hosts' tips. In Supplementary Figure S6 we show the posterior distributions estimated by our ABC method using the different

pruned phylogenies. We find that although the confidence intervals are wider, the posterior distributions are all similar with the posterior distributions estimated using the target phylogeny. Finally, to evaluate the impact of phylogenetic reconstruction uncertainty, we analysed 100 additional trees from the Beast posterior distribution. In Supplementary figure S7, we show that the estimates from our ABC method are qualitatively similar for all these trees.

3.2.3 Discussion

Over the last years, the area of Lyon (France) witnessed an increase in HCV incidence both in HIV-positive and HIV-negative populations of men having sex with men (MSM) (**?**). This increase appears to be driven by sexual transmission and echoes similar trends in Amsterdam (**?**) and Switzerland (**?**). A quantitative analysis of the epidemic is necessary to optimise public health interventions. Unfortunately, this is challenging because the monitoring of the population at risk is limited and because classical tools in quantitative epidemiology, especially incidence time series, are poorly informative with such a heterogeneous population. To circumvent this problem, we used HCV sequence data, which we analysed using phylodynamics. To account for host heterogeneity, we extended and validated an existing Approximate Bayesian Computation framework (**?**).

From a public health point of view, our results have two major implications. First, we find a strong degree of assortativity in both 'classical' and 'new' host populations. The virus phylogeny does hint at this result (Figure 3.1) but the ABC approach allows us to quantify the pattern and to show that assortativity may be higher for 'classical' hosts. The second main result has to do with the striking difference in doubling times. Indeed, the current spread of the epidemics in 'new' hosts appears to be five times more rapid than the spread in the 'classical' hosts in the early 1990s before the advent of the third generation tests in 1997, and ten times more rapid that the spread in the 'new' hosts after 1997. That the duration of the infectious period in 'new' hosts is in the same order of magnitude as the time until treatment suggests that the majority of the transmission events may be occurring during the acute phase. This underlines the necessity to act rapidly upon detection, for instance by emphasising the importance of protection measures such as condom use and by initiating treatment even during the acute phase (**?**). A better understanding of the underlying contact networks could provide additional information regarding the structure of the epidemics and, with that respect, next-generation sequence (NGS) data could be particularly informative (**???**).

Some potential limitations of the study are related to the sampling scheme, the

assessment of the host type, and the transmission model. Regarding the sampling, the proportion of infected 'new' host that is sampled is unknown but could be high. For the 'classical' hosts, we selected a representative subset of the patients detected in the area but this sampling is likely to be low. However, the effect of underestimating sampling for the new epidemics would be to underestimate its spread, which is already faster than the classical epidemics. When running the analyses on different phylogenies with half of the 'new' hosts sequences, we find results similar to those obtained with the whole phylogeny, suggesting that our ABC framework is partly robust to sampling biases. In general, implementing a more realistic sampling scheme in the model would be possible but it would require a more detailed model and more data to avoid identifiability issues. Regarding assigning hosts to one of the two types, this was performed by clinicians independently of the sequence data. The main criterion used was the infection stage (acute or chronic), which was complemented by other epidemiological criteria (history of intravenous drug use, blood transfusion, HIV status). Finally, the 'classical' and the 'new' epidemics appear to be spreading on contact networks with different structures. However, such differences are beyond the level of details of the birth-death model we use here and would require a larger dataset for them to be inferred.

To test whether the infection stage (acute vs. chronic) can explain the data better than the existence of two host types, we developed an alternative model where all infected hosts first go through an acute phase before recovering or progressing to the chronic phase. As for the model with two host types, we used three time intervals. Supplementary Figure S9 shows the diagram of the model as well as the corresponding equations. Interestingly, it was almost impossible to simulate phylogenies with this model, most likely because of its intrinsic constrains on assortativity (both acute and chronic infections always generate new acute infections).

To our knowledge, few attempts have been made in phylodynamics to tackle the issue of host population heterogeneity. In 2018, a study used the structured coalescent model to investigate the importance of accounting for so-called 'superspreaders' in the recent Ebola epidemics in West Africa (**?**). The same year, another study used the birth-death model to study the effect of drug resistance mutations on the R_0 of HIV strains (**?**). Both of these are implemented in Beast2. We ran an analysis using the BEAST 2 package bdmm with our data. We were unable to conclude anything from this analysis. However, this is probably due to difficulties in estimating both evolutionary and epidemiological parameters, when in this ABC inference study we inferred epidemiological parameters using a fixed phylogeny.

Overall, we show that our ABC approach, which we validated for simple SIR epidemiological models (**?**), can be applied to more elaborate models that current

phylodynamics methods have difficulties to capture. Further increasing the level of details in the model may require to increase the number of simulations but also to introduce new summary statistics. Another promising perspective would be to combine sequence and incidence data. Although this could not be done here due to the limited sampling, such data integration can readily be done with regression-ABC.

3.2.4 Material and methods

Epidemiological data

The Dat'AIDS cohort is a collaborative network of 23 French HIV treatment centres covering approximately 25% of HIV-infected patients followed in France (Clinicaltrials.gov ref NCT02898987). Host profiles have been established by field epidemiologists based on interviews and risk factors.

HCV sequence data

We included HCV molecular sequences of all MSM patients diagnosed with acute HCV genotype 1a infection at the Infectious Disease Department of the Hospices Civils de Lyon, France, and for whom NS5B sequencing was performed between January 2014 and December 2017 ($N = 68$). HCV genotype 1a isolated from $N = 145$ non-MSM, HIV-negative, male patients of similar age were analysed by NS5B sequencing at the same time for phylogenetic analysis. This study was conducted following French ethics regulations. All patients gave their written informed consent to allow the use of their personal clinical data. The study was approved by the Ethics Committee of Hospices Civils de Lyon.

HCV testing and sequencing

HCV RNA was detected and quantified using the Abbott RealTime HCV assay (Abbott Molecular, Rungis, France). The NS5B fragment of HCV was amplified between nucleotides 8256 and 8644 by RT-PCR as previously described and sequenced using the Sanger method. Electrophoresis and data collection were performed on a GenomeLab^{TM} GeXP Genetic Analyzer (Beckman Coulter). Consensus sequences were assembled and analysed using the GenomeLab^{TM} sequence analysis software. The genotype of each sample was determined by comparing its sequence with HCV reference sequences obtained from GenBank.

Nucleotide accession numbers

All HCV NS5B sequences isolated in MSM and non-MSM patients reported in this study were submitted to the GenBank database. The list of Genbank accession numbers for all sequences is provided in Appendix.

Dated viral phylogeny

To infer the time-scaled viral phylogeny from the alignment we used a Bayesian Skyline model in BEAST v2.5.2 (**?**). The general time-reversible (GTR) nucleotide substitution model was used with a strict clock rate fixed at $1.3 \cdot 10^{-3}$ based on data from Ref. (**?**) and a gamma distribution with four substitution rate categories. The MCMC was run for 100 million iterations and samples were saved every 100,000 iterations. We selected the maximum clade credibility using TreeAnnotator BEAST2 package. The date of the last common ancestor was estimated to be 1961.95 with a 95% Highest Posterior Density (HPD) of [1941.846; 1975.516]. When performing the same inference without the new hosts, we found a similar estimate (1960) and the same 95% HPD of [1942; 1975], which we used as a prior distribution to estimate the origin of the classical hosts t_0 (Table 3.1).

Epidemiological model and simulations

We assume a Birth-Death model with two hosts types (Supplementary Figure S1) with 'classical' hosts (numbered 1) and new hosts (numbered 2). This model is described by the following system of ordinary differential equations (ODEs):

$$\frac{dI_1}{dt} = a_1\beta I_1 + (1-a_2)\nu\beta I_2 - \gamma_1 I_1 \tag{3.1a}$$

$$\frac{dI_2}{dt} = a_2\beta\nu I_2 + (1-a_1)\beta I_1 - \gamma_2 I_2 \tag{3.1b}$$

In the model, transmission events are possible within each type of hosts and between the two types of hosts at a transmission rate β. Parameter ν corresponds to the transmission rate differential between classical and new hosts. Individuals can be 'removed' at a rate γ_1 from an infectious compartment (I_1 or I_2) via infection clearance, host death or change in host behaviour (e.g. condom use). The assortativity between host types, which can be seen as the percentage of transmissions that occur with hosts from the same type, is captured by parameter a_i.

The effective reproduction number (denoted R_t) is the number of secondary cases caused by an infectious individual in a fully susceptible host population (**?**). We seek

Table 3.1 – **Prior distributions for the birth-death model parameters over the three time intervals.** t_0 is the date of origin of the epidemics in the studied area, t_1 is the date of introduction of 3$^{\text{rd}}$ generation HCV tests, t_2 is the date of emergence of the epidemic in 'new' hosts and t_3 is the time of the most recent sampled sequence.

Interval	γ_i	ν	$R^{(1)}$	a_i
$[t_0, t_1]$	Unif$(0.1, 4)$	0	Unif$(0.9, 6)$	Unif$(0.5, 1)$
$[t_1, t_2]$			Unif$(0.1, 3)$	
$[t_2, t_3]$		Unif$(0, 1)$ & Unif$(1, 10)$		

to infer the R_t from the classical epidemic, denoted $R^{(1)}$ and defined by $R^{(1)} = \beta/\gamma_1$, as well as the R_t of the new epidemic, denoted $R^{(2)}$ and defined by $R^{(2)} = \nu\beta/\gamma_2 = \nu R^{(1)}\gamma_1/\gamma_2$.

The doubling time of an epidemic (t_D) corresponds to the time required for the number of infected hosts to double in size. It is usually estimated in the early stage of an epidemic when epidemic growth can assumed to be exponential. To calculate it, we assume perfect assortativity ($a_1 = a_2 = 1$) and approximate the initial exponential growth rate by $\beta - \gamma_1$ for 'classical' hosts and $\nu\beta - \gamma_2$ for 'new' hosts. Following **?**, we obtain $t_D^{(1)} = \ln(2)/(\beta - \gamma_1)$ and $t_D^{(2)} = \ln(2)/(\nu\beta - \gamma_2)$.

We consider three time intervals. During the first interval $[t_0, t_1]$, t_0 being the year of the origin of the epidemic in the area of Lyon, we assume that only classical hosts are present. The second interval $[t_1, t_2]$, begins in $t_1 = 1997.3$ with the introduction of the third generation HCV tests, which we assume to have affected $R^{(1)}$ through the decrease of the transmission rate β. Finally, the 'new' hosts appear during the last interval $[t_2, t_3]$, where t_2, which we infer, is the date of origin of the second outbreak. The final time (t_3) is set by the most recent sampling date in our dataset (2018.39). The prior distributions used are summarized in Table 3.1 and shown in Figure 3.2. Given the phylogeny structure suggesting a high degree of assortativity, we assume the assortativity parameters, a_1 and a_2, to be higher than 50%. For the prior distribution of parameter ν, we combined a uniform distribution from 0 to 1 with a uniform distribution from 1 to 10. This was done to ensure that the probability to have $\nu < 1$ is equal to the probability to have $\nu > 1$.

To simulate phylogenies, we use our TiPS simulator (**?**) implemented in `R` via the `Rcpp` package. This is done in a two-step procedure. First, epidemiological trajectories are simulated using the compartmental model in equation 3.1 and Gillespie's

stochastic event-driven simulation algorithm (**?**). The number of individuals in each compartment and the reactions occurring through the simulations of trajectories, such as recovery or transmission events, are recorded. Using the target phylogeny, we know when sampling events occur. For each simulation, each sampling date is randomly associated to a host compartment using the observed fraction of each infection type (here 68% of the dates associated with 'classical' hosts type and 32% with 'new' hosts). Once the sampling dates are added to the trajectories, we move to the second step, which involves simulating the phylogeny. This step starts from the last sampling date and follows the epidemiological trajectory through a coalescent process, that is backwards-in-time. Each backward step in the trajectory can induce a tree modification given a probability and the population size: a sampling event leads to a labelled leaf in the phylogeny, a transmission event can lead to the coalescence of two sampled lineages or to no modification of the phylogeny (if one of the lineages is not sampled).

We implicitly assume that the sampling rate is low, which is consistent with the limited number of sequences in the dataset. We also assume that the virus can still be transmitted after sampling.

We simulate $60,000$ phylogenies from known parameter sets drawn in the prior distributions shown in Table 3.1. These are used to perform the rejection step and build the regression model in the Approximate Bayesian Computation (ABC) inference.

ABC inference

Summary statistics Phylogenies are rich objects and to compare them we break them into summary statistics. These are chosen to capture the epidemiological information of interest. In particular, following an earlier study, we use summary statistics from branch lengths, tree topology, and lineage-through-time (LTT) (**?**), and summary statistics based on the Laplacian spectrum using the spectR function of the RPANDA R package (**?**).

We also compute new summary statistics to extract information regarding the heterogeneity of the population, the assortativity, and the difference between the two R. To do so, we annotate each internal node by associating it with a probability to be in a particular state (here the host type, 'classical' or 'new'). We assume that this probability is given by the ratio

$$P(Y) = \frac{\text{number of descendent leaves labelled } Y}{\text{number of descendent leaves}} \tag{3.2}$$

where Y is a state (or host type). Each node is therefore annotated with n ratios, n being the number of possible states. Since in our case $n = 2$, we only follow one of the labels and use the mean and the variance of the distribution of the ratios (one for each node) as summary statistics.

In a phylogeny, cherries are pairs of leaves that are adjacent to a common ancestor. There are $n(n + 1)/2$ categories of cherries. Here, we compute the proportion of homogeneous cherries for each label and the proportion of heterogeneous cherries. We also consider pitchforks, which we define as a cherry and a leaf adjacent to a common ancestor, and introduce three categories: homogeneous pitchforks, pitchforks whose cherries are homogeneous for a label and whose leaf is labelled with another trait, and pitchforks whose cherries are heterogeneous.

The Lineage-Through-Time (LTT) plot displays the number of lineages of a phylogeny over time. In this plot, the number of lineages is incremented by one every time there is a new branch in the phylogeny and is decreased by one every time there is a new leaf in the phylogeny. We use the ratios defined for each internal node to build an LTT plot for each label type, which we refer to as 'LTT label plot'. After each branching event in phylogeny, we increment the number of lineages by the value of the ratio of the internal node for the given label. This number of lineages is decreased by one every time there is a leaf in the phylogeny. In the end, we obtain $n = 2$ LTT label plots.

Finally, for each label, we compute some of our branch lengths summary statistics on homogeneous clades and heterogeneous clades present in the phylogeny. Homogeneous clades are defined by their root having a ratio of 1 for one type of label and their size being greater than N_{min}. For heterogeneous clades, we keep the size criterion and impose that the ratio is smaller than 1 but greater than a threshold ϵ. After preliminary analyses, we set $N_{\text{min}} = 4$ leaves and $\epsilon = 0.7$. We then obtain a set of homogeneous clades and a set of heterogeneous clades, the branch lengths of which we pool into two sets to compute the summary statistics of heterogeneous and homogeneous clades. Note that we always select the largest clade, for both homogeneous and heterogeneous cases, to avoid redundancy.

Regression-ABC We first measure multicollinearity between summary statistics using variance inflation factors (VIF). Each summary statistic is kept if its VIF value is lower than 10. This stepwise VIF test leads to the selection of 101 summary statistics out of 330.

We then use the abc function from the abc R package (**?**) to infer posterior distributions generated using only the rejection step. Finally, we perform linear adjustment using an elastic net regression.

The abc function performs a classical one-step rejection algorithm (**?**) using a tolerance parameter P_δ, which represents a percentile of the simulations that are close to the target. To compute the distance between a simulation and the target, we use the Euclidian distance between normalized simulated vectors of summary statistics and the normalized target vector.

Before linear adjustment, the abc function performs smooth weighting using an Epanechnikov kernel (**?**). Then, using the glmnet package in R, we implement an elastic-net (EN) adjustment, which balances the Ridge and the LASSO regression penalties (**?**). Since the EN performs a linear regression, it is not subject to the risk of over-fitting that may occur for non-linear regressions (e.g. when using neural networks, support vector machines or random forests).

In the end, we obtain posterior distributions for t_0, t_2, a_1, a_2, ν, γ_1, γ_2, $R^{(1),t_1}$ and $R^{(1),t_2}$ using our ABC-EN regression model with $P_\delta = 0.05$.

Parametric bootstrap and cross-validation Our goodness-of-fit validation consists in simulating $10,000$ additional phylogenies from parameter sets drawn in posterior distributions. We then compute summary statistics and perform a goodness of fit using the gfitpca function from the abc R package (**?**). The function performs principal component analysis (PCA) using the simulated summary statistics. It displays envelopes containing a given percentage, here 90%, of the simulations. The projection of the observed summary statistics is displayed to check if they are contained or not in the envelopes. If the posterior distribution is informative, we expect the target data to be contained in the envelope. This analysis was performed either on the posterior distribution, or on a uniform distribution based on the 95% HPD posterior distribution of each parameter, the latter being less informative.

To assess the robustness of our ABC-EN method to infer epidemiological parameters of our BD model, we also perform a 'leave-one-out' cross-validation as in (**?**). This consists in inferring posterior distributions of the parameters from one simulated phylogeny, assumed to be the target phylogeny, using the ABC-EN method with the remaining $59,999$ simulated phylogenies. We run the cross-validation 100 times with 100 different target phylogenies. We consider three parameter distributions θ: the prior distribution, the prior distribution reduced by the feasibility of the simulations and the ABC inferred posterior distribution. For each of these parameter distributions, we measure the median and compute, for each simulation scenario, the mean relative error (MRE) such as:

$$MRE = \frac{1}{100} \sum_{i=1}^{100} \mid \frac{\theta_i}{\Theta} - 1 \mid \tag{3.3}$$

where Θ is the true value.

Data accessibility

Data are available online: link or DOI of the webpage hosting the data

Supplementary material

Script and codes are available online: link or DOI of the webpage hosting the script and codes

Acknowledgements

We thank Jūlija Pečerska for her help with Beast2.

GD is funded by the Fondation pour la Recherche Médicale (grant ECO20170637560).

GD and SA acknowledge further support from the CNRS, the IRD and the itrop HPC (South Green Platform) at IRD Montpellier, which provided HPC resources that contributed to the results reported here (`https://bioinfo.ird.fr/`).

Version 5 of this preprint has been peer-reviewed and recommended by Peer Community In Evolutionary Biology (`https://doi.org/10.24072/pci.evolbiol.100117`).

Conflict of interest disclosure

The authors of this preprint declare that they have no financial conflict of interest with the content of this article. SA is a recommender for PCI Evol Biol.

3.2.5 Appendix

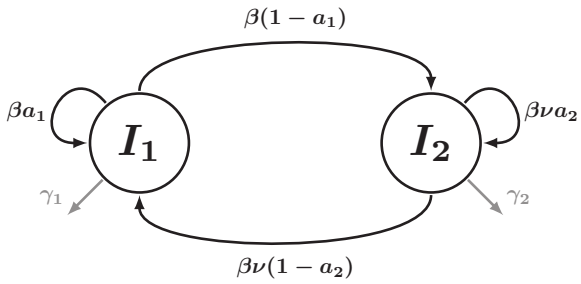


Figure S1 – **Diagram of the birth-death model with host heterogeneity.** The intensity of the colour is proportional to the correlation coefficients.

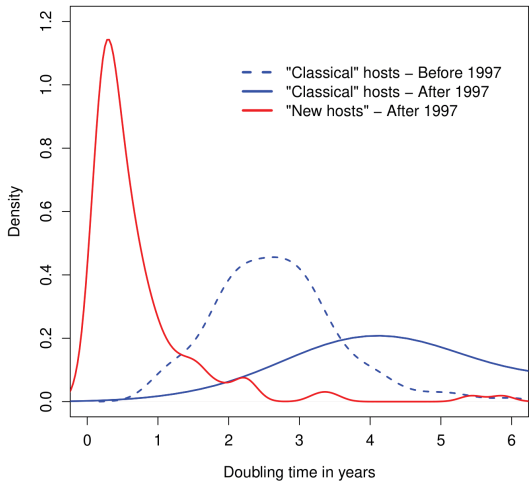


Figure S2 – **Densities of the inferred doubling times.** The density of the doubling time for the 'classical' hosts before 1997 is in blue dashed line, and after 1997 in blue solid line. The density of the doubling time for the 'new' hosts is in red. $(t_D^{(2),t3})$.

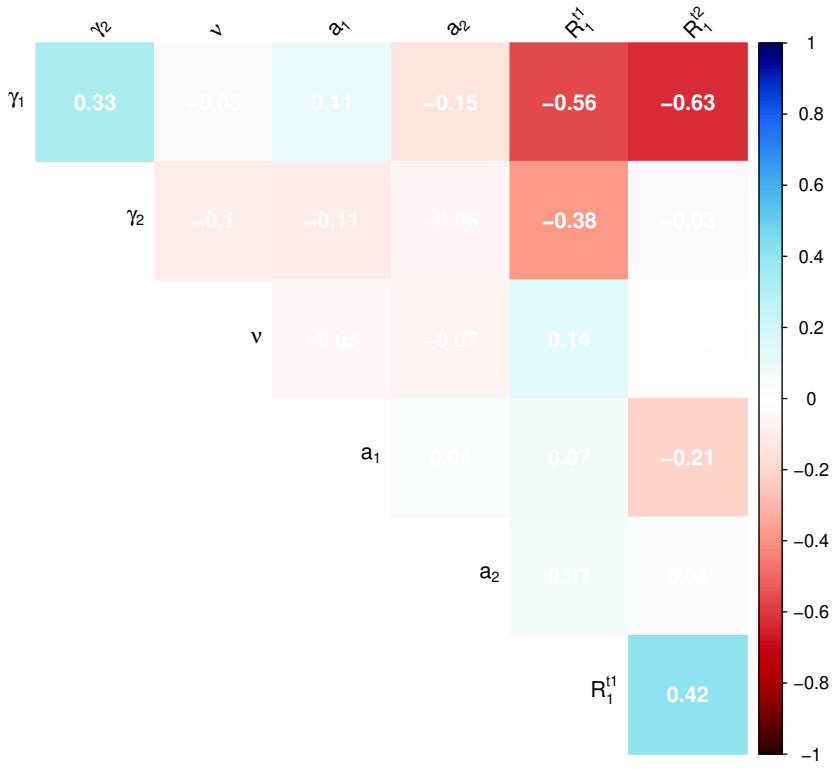


Figure S3 – **Correlation heat map between the posterior distributions for the model parameters.** The intensity of the colour is proportional to the correlation coefficients.

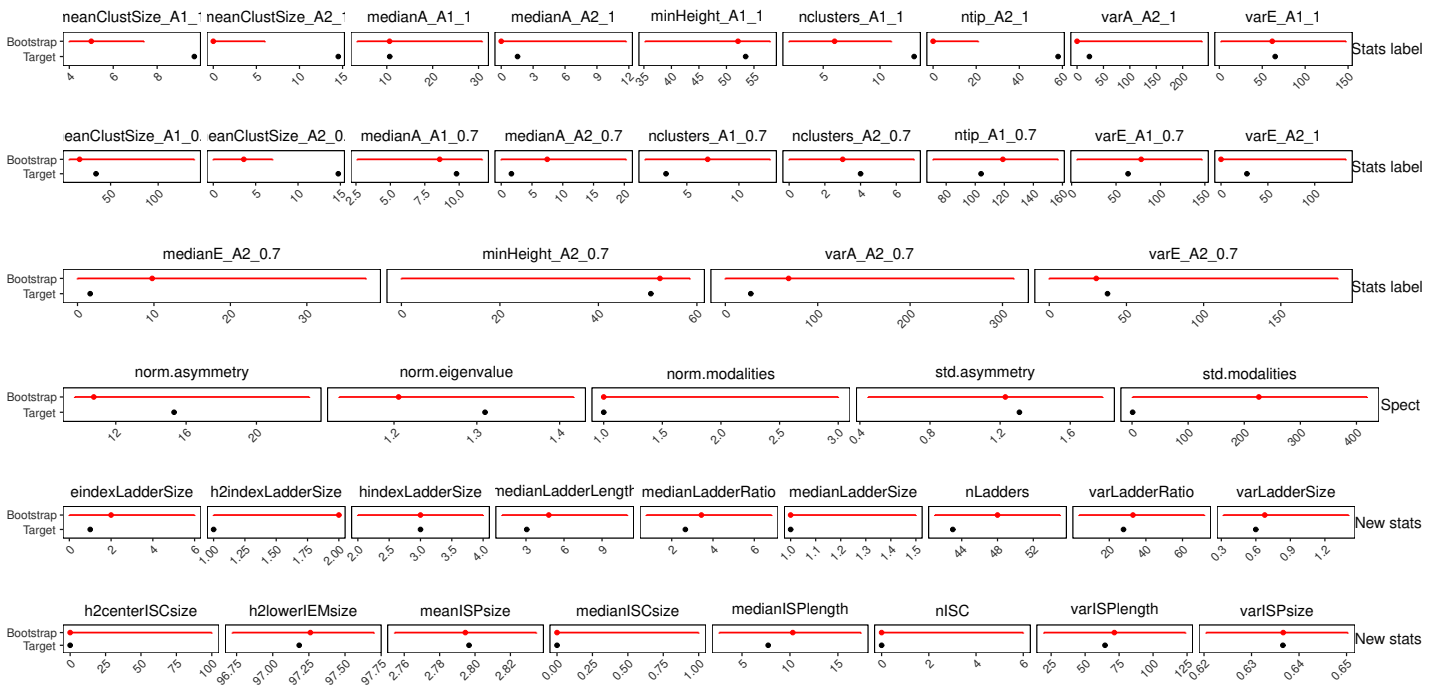


Figure S4 – **Distributions of selected summary statistics.** The dots represent the median and the horizontal lines represent the 95% HPD. Red distributions correspond to the summary statistics computed from the 10,000 phylogenies simulated from the posterior distribution. Black dots represent the values of selected summary statistics computed from the target phylogeny. Summary statistics are represented by group.

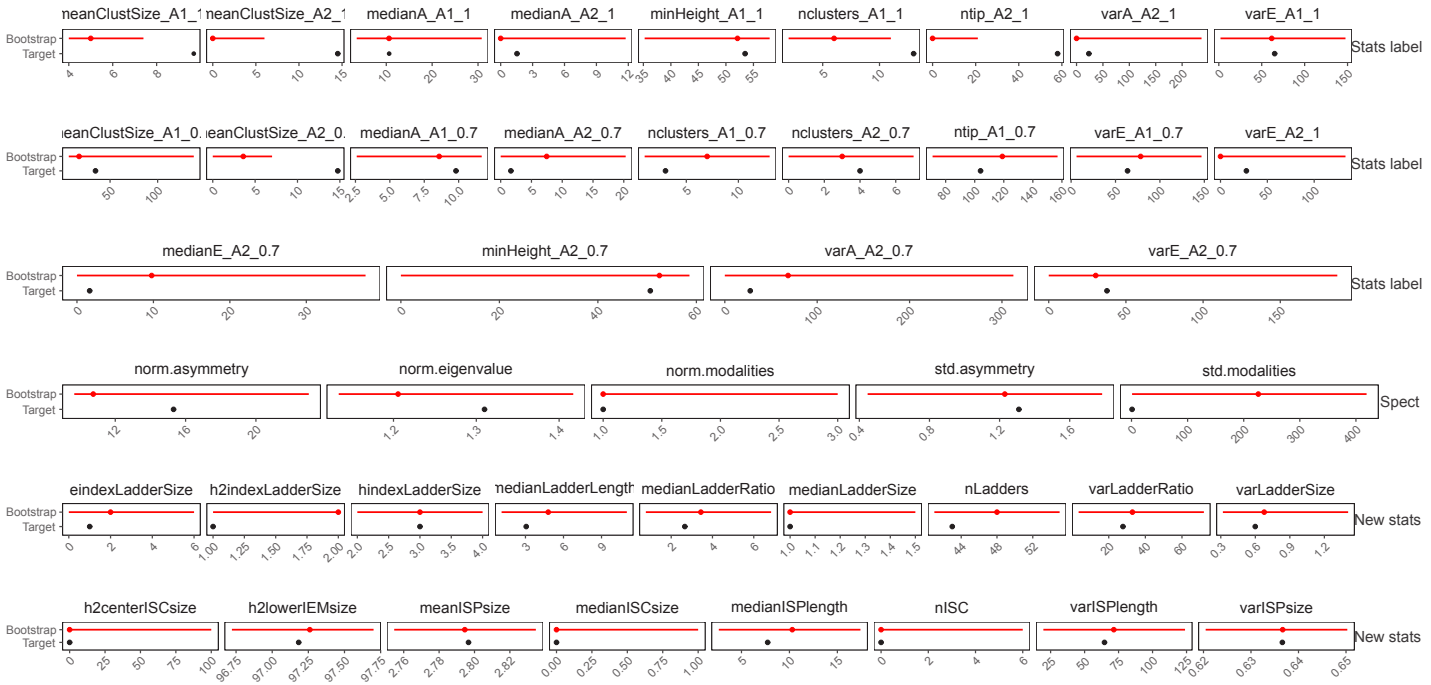


Figure S4 – **Distributions of selected summary statistics.** The dots represent the median and the horizontal lines represent the 95% HPD. Red distributions correspond to the summary statistics computed from the 10,000 phylogenies simulated from the posterior distribution. Black dots represent the values of selected summary statistics computed from the target phylogeny. Summary statistics are represented by group.

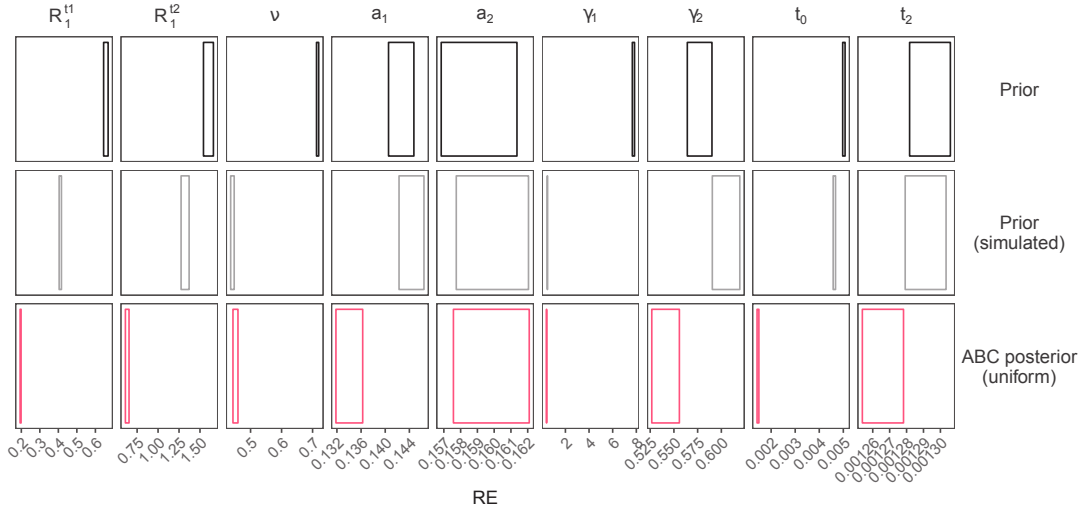


Figure S5 – **Cross-validation results.** Each column corresponds to one of the inferred parameters. The first line shows the prior distribution. The second line shows the distribution of values for which a phylogeny could be simulated. The third line shows the inference after then ABC. For the rejection step of the ABC, the tolerance level was set to $P_\delta = 0.05$. The rectangles show the mean relative errors and their standard errors computed for 100 target sets with known values (see the Methods).

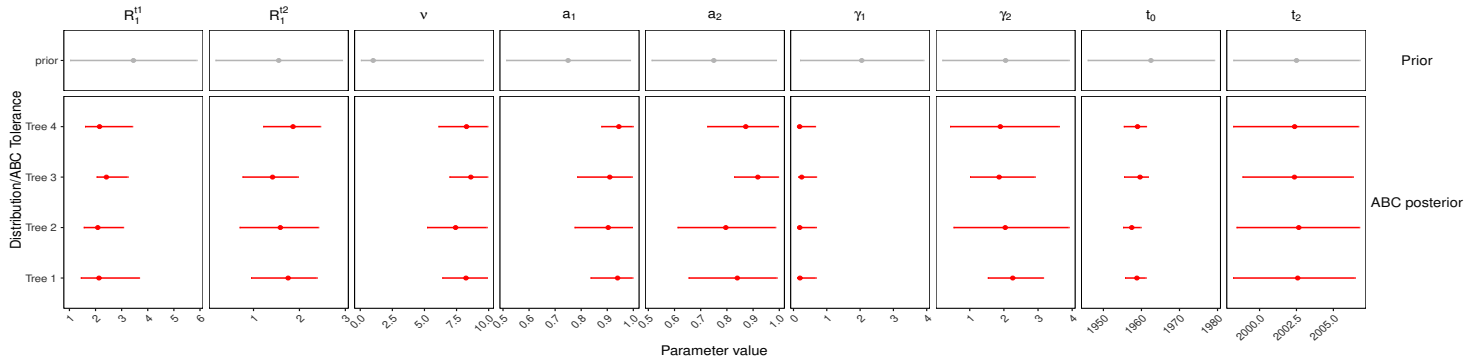


Figure S6 – **Posterior distributions estimated from different phylogenies inferred using half of the 'new' hosts' sequences.** The first line represents the prior (in grey), the last line the full target tree (in red), and all the intermediate lines phylogenies where half of the 'new' host sequences were removed at random.

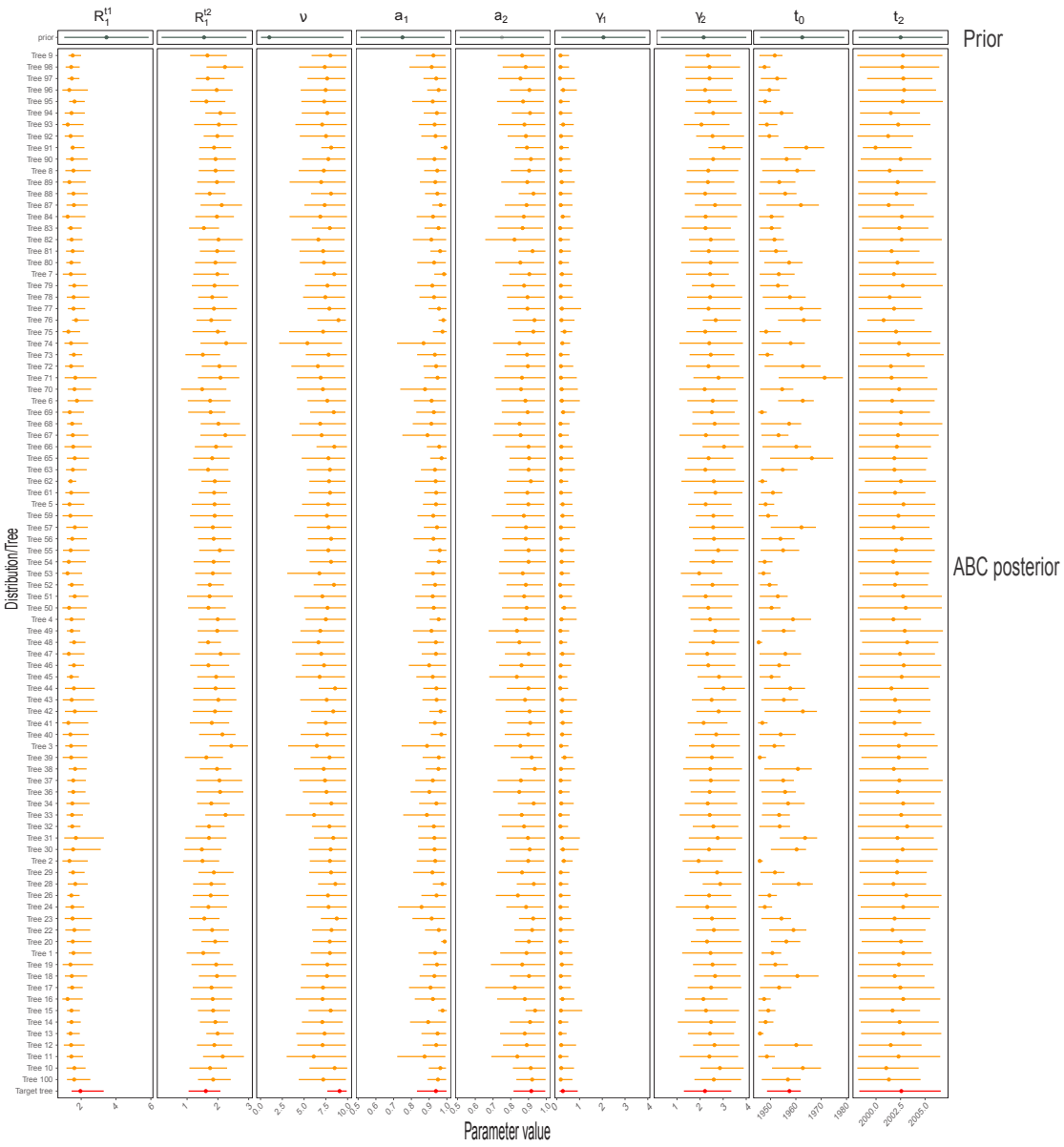


Figure S7 – **Variation in posterior distribution estimated from different inferred phylogenies.** The dots represent the median and the horizontal lines represent the 95% highest posterior density (HPD) of each distribution. Grey distributions correspond to the prior, orange distributions correspond to the different posterior distributions computed from 100 phylogenies drawn at random in the posterior distribution of trees inferred by Beast2 and red distributions correspond to the ABC-EN posterior distributions.

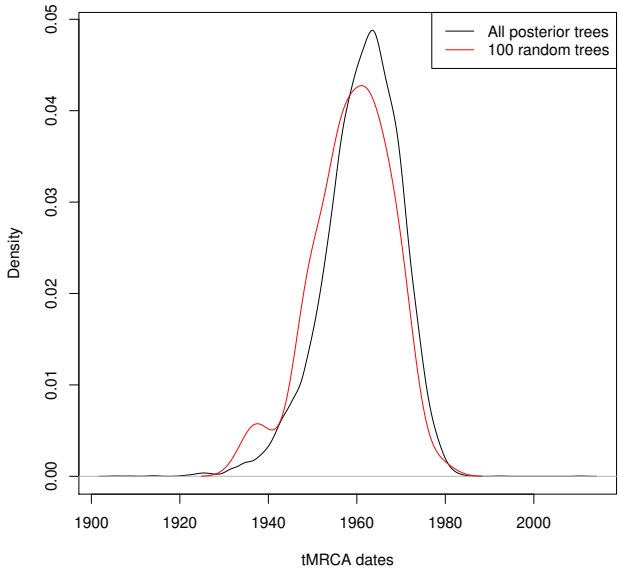


Figure S8 – **Density distributions of the t_{MRCA} for the observed Beast2 phylogeny (in black) and for the 100 phylogenies drawn at random in the posterior distributions of trees inferred by Beast2 (in red).**

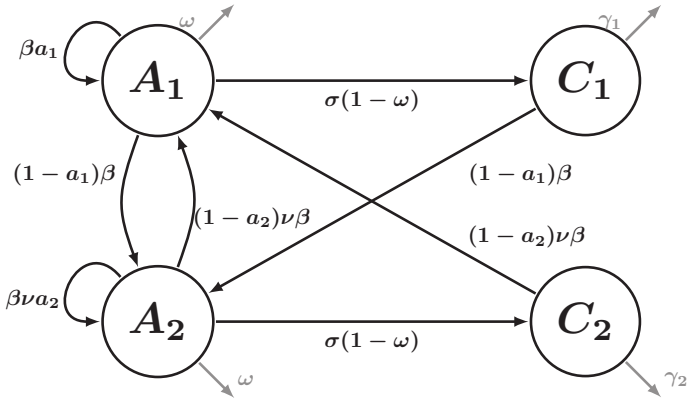


Figure S9 – **Diagram of the alternative model where all infected hosts first go through an acute phase (A_i) before recovering or progressing to the chronic phase (C_i).** ω is the proportion of infections that clear before becoming chronic, σ is the rate at which acute infections become chronic, and other parameters are identical to those in the main text. The equations governing the dynamics of the system can be written as $\frac{dA_i}{dt} = a_i\beta_i(A_i + C_i) + (1 - a_j)\beta_j(A_j + C_j) - \sigma A_i$ and $\frac{dC_i}{dt} = \sigma(1-\omega)A_i - \gamma_i C_i$ with $i \neq j$, $\beta_1 = \beta$ and $\beta_2 = \nu\beta$.

GenBank accession numbers of sequences used to infer the viral phylogeny.

Accession number	Host type	Sampling date
KY928360	classical	31/07/2014
KY928361	classical	24/09/2014
KY928362	classical	06/10/2014
KY928363	classical	17/10/2014
KY928364	classical	20/11/2014
KY928365	classical	01/12/2014
KY928366	classical	05/01/2015
KY928367	classical	13/01/2015
KY928368	classical	16/01/2015
KY928369	classical	16/01/2015
KY928370	classical	20/01/2015
KY928371	classical	30/01/2015
KY928372	classical	04/02/2015
KY928373	classical	06/02/2015
KY928374	classical	10/02/2015
KY928375	classical	11/02/2015
KY928376	classical	13/02/2015
KY928377	classical	17/02/2015
KY928378	classical	17/02/2015
KY928379	classical	18/02/2015
KY928380	classical	23/02/2015
KY928381	classical	25/02/2015
KY928382	classical	02/03/2015
KY928383	classical	04/03/2015
KY928384	classical	05/03/2015
KY928385	classical	09/03/2015
KY928386	classical	16/03/2015
KY928387	classical	18/03/2015
KY928388	classical	27/03/2015
KY928389	classical	30/03/2015
MH885712	classical	31/03/2015
MH885713	classical	31/03/2015
MH885714	classical	10/04/2015
MH885715	classical	14/04/2015
MH885716	classical	15/04/2015
MH885717	classical	21/04/2015
MH885718	classical	28/04/2015
MH885719	classical	30/04/2015
MH885720	classical	30/04/2015
MH885721	classical	30/04/2015
MH885722	classical	07/05/2015
MH885723	classical	20/05/2015
MH885724	classical	26/05/2015
MH885725	classical	28/05/2015
MH885726	classical	03/06/2015
MH885727	classical	05/06/2015
MH885728	classical	08/06/2015
MH885729	classical	09/06/2015
MH885730	classical	10/06/2015
MH885731	classical	11/06/2015
MH885732	classical	12/06/2015
MH885733	classical	12/06/2015
MH885734	classical	17/06/2015
MH885735	classical	18/06/2015
MH885736	classical	19/06/2015
MH885737	classical	03/07/2015
MH885738	classical	08/07/2015
MH885739	classical	22/07/2015
MH885740	classical	24/07/2015
MH885741	classical	28/07/2015
MH885742	classical	29/07/2015
MH885743	classical	30/07/2015
MH885744	classical	12/08/2015
MH885745	classical	17/08/2015
MH885746	classical	17/08/2015
MH885747	classical	18/08/2015
MH885748	classical	21/08/2015
MH885749	classical	24/08/2015
MH885750	classical	27/08/2015
MH885751	classical	07/09/2015
MH885752	classical	08/09/2015
MH885753	classical	08/09/2015
MH885754	classical	09/09/2015
MH885755	classical	10/09/2015
MH885756	classical	16/09/2015
MH885757	classical	17/09/2015
MH885758	classical	22/09/2015
MH885759	classical	29/09/2015
MH885760	classical	29/09/2015
MH885761	classical	06/10/2015
MH885762	classical	12/10/2015
MH885763	classical	19/10/2015
MH885764	classical	20/10/2015
MH885765	classical	21/10/2015
MT108308	classical	28/10/2015
MT108309	classical	28/10/2015
MT108310	classical	06/11/2015
MT108311	classical	20/11/2015
MT108312	classical	25/11/2015
MT108313	classical	02/12/2015
MT108314	classical	04/12/2015
MT108315	classical	07/12/2015
MT108316	classical	11/12/2015
MT108317	classical	11/12/2015
MT108318	classical	14/12/2015
MT108319	classical	30/12/2015
MT108320	classical	04/01/2016
MT108321	classical	05/01/2016
MT108322	classical	05/01/2016
MT108323	classical	08/01/2016
MT108324	classical	15/01/2016
MT108325	classical	15/01/2016
MT108326	classical	19/01/2016
MT108327	classical	03/02/2016
MT108328	classical	05/02/2016
MT108329	classical	05/02/2016
MT108330	classical	16/02/2016
MT108331	classical	16/02/2016

Accession number	Host type	Sampling date
MT108332	classical	19/02/2016
MT108333	classical	26/02/2016
MT108334	classical	07/03/2016
MT108335	classical	07/03/2016
MT108336	classical	30/03/2016
MT108337	classical	01/04/2016
MT108338	classical	01/04/2016
MT108339	classical	07/04/2016
MT108340	classical	18/04/2016
MT108341	classical	25/04/2016
MT108342	classical	26/04/2016
MT108343	classical	19/05/2016
MT108344	classical	25/05/2016
MT108345	classical	26/05/2016
MT108346	classical	27/05/2016
MT108347	classical	03/06/2016
MT108348	classical	06/06/2016
MT108349	classical	07/06/2016
MT108350	classical	13/06/2016
MT108351	classical	21/06/2016
MT108352	classical	01/07/2016
MT108353	classical	04/07/2016
MT108354	classical	07/09/2016
MT108355	classical	08/09/2016
MT108356	classical	06/01/2017
MT108357	classical	08/02/2017
MT108358	classical	01/06/2017
MT108359	classical	05/09/2017
MT108360	classical	22/09/2017
MT108361	classical	10/10/2017
MT108362	classical	24/04/2018
MT108363	classical	24/04/2018
MT108364	classical	02/05/2018
MT108365	classical	02/05/2018
MT108366	classical	02/05/2018
MT108367	classical	03/05/2018
MT108368	classical	24/05/2018
MH885654	new	27/07/2011
MH885655	new	21/01/2013
KY928329	new	15/02/2013
KY928344	new	23/05/2013
KY928348	new	24/05/2013
KY928322	new	16/12/2013
MH885656	new	18/02/2014
MH885657	new	04/04/2014
KY928330	new	22/09/2014
MH885658	new	22/10/2014
KY928355	new	13/02/2015
MH885659	new	26/05/2015
KY928356	new	27/05/2015
MH885660	new	10/06/2015
KY928352	new	12/06/2015
MH885661	new	12/06/2015
KY928314	new	06/08/2015
MH885663	new	16/09/2015
MH885664	new	05/10/2015
KY928336	new	14/10/2015
KY928338	new	05/11/2015
KY928350	new	16/11/2015
MH885665	new	29/12/2015
MH885666	new	30/12/2015
MH885667	new	05/01/2016
MH885668	new	03/02/2016
MH885669	new	10/02/2016
MH885670	new	03/03/2016
KY928357	new	23/03/2016
KY928341	new	24/03/2016
MH885671	new	06/04/2016
KY928354	new	07/04/2016
KY928311	new	15/04/2016
KY928335	new	31/05/2016
KY928346	new	02/06/2016
MH885672	new	13/06/2016
MH885673	new	05/07/2016
KY928349	new	03/08/2016
KY928331	new	01/09/2016
MH885674	new	14/09/2016
KY928324	new	21/09/2016
KY928353	new	19/10/2016
MH885675	new	04/11/2016
KY928351	new	30/11/2016
KY928343	new	23/01/2017
MH885676	new	10/02/2017
KY928325	new	22/03/2017
MH885677	new	27/03/2017
KY928333	new	05/04/2017
KY928345	new	03/05/2017
KY928315	new	27/06/2017
MH885693	new	10/07/2017
MH885695	new	19/07/2017
MH885697	new	24/07/2017
MH885698	new	25/07/2017
MH885699	new	16/08/2017
MH885700	new	17/08/2017
MH885701	new	21/08/2017
MH885702	new	30/08/2017
MH885703	new	27/07/2017
MH885704	new	03/11/2017
MH885705	new	14/11/2017
MH885707	new	27/11/2017
MH885711	new	20/12/2017
MT108306	new	22/11/2017
MT108307	new	27/11/2017
MH885662	new	15/06/2015
KY928313	new	24/07/2015

Chapitre 4

Phylodynamique du VIH-1 de groupe O par ABC-régression

4.1 Introduction

Human immunodeficiency viruses (HIV) belong to two main types, HIV type 1 (HIV-1) and HIV type 2 (HIV-2), which are both causal agents of acquired immunodeficiency syndrome (AIDS). HIV-1 is comprised of four groups (M to P), each originating from an independent cross-species transmission event from Simian Immunodeficiency Virus (SIV) from non-human primates to human. Although the origin of group O is estimated to be around the same time as group M, HIV-1/M (M for 'major') has become pandemic and accounts for more than 99% of HIV infections, representing more than 38 million cases worldwide, while HIV-1/O (O for 'outlier') has remained endemic in Cameroon, accounting for about 1% of HIV infections in the country. However, little is known about the genetic evolution of group O and the causes of this difference in terms of viral spread between HIV-1/O and HIV-1/M.

HIV-1/O virus population have developed high genetic diversity. The limited amount of sequence data and the phylogenetic structure of HIV-1/O presenting few clusters have made it difficult to define a nomenclature. One of the three proposed nomenclatures is based on the nature of the residue present at position 181 of the reverse transcriptase. The Y181C polymorphism observed in some HIV-1/O appears to be related to two distinct lineages named 181Y-*like* and 181C-like. **?** confirmed the existence of a predominant population that was strongly associated with the C mutation at position 181 of the reverse transcriptase. This mutation could thus confer a replicative advantage to those in which it was naturally present.

Here, we seek to quantify the potential virus transmissions advantage associated

with the presentation of one residue over another at the position 181 of the reverse transcriptase. We performed a phylodynamic study by analyzing 75 each with either residue 181Y or residue 181C. Our analysis is based on a method that has already been used and validated on a labelled phylogeny (see chapter 3). This method is based on Approximate Bayesian Computation combined with a regression model (ABC-regression). We inferred the differential of the reproduction number associated with the 181Y and 181C mutation, the substitution rate and we dated the origin of the epidemic.

4.2 Methods

4.2.1 Times-scaled viral phylogeny

We obtained an alignment of 75 HIV-1/O sequences from **?** and the associated sampling dates, which range from 1992 to 2013.

To infer the time-scaled phylogeny from the alignment, we used four different models with different assumptions regarding the molecular clock (strict and relaxed) and the demographic model (constant and exponential population size) using BEAST v2.5.5 (**?**). The general time-reversible (GTR) nucleotide substitution model was used and a Gamma distribution with four substitution rate categories. For each model, the MCMC was run for 100 million iterations and samples were saved every 100,000 iterations. We performed a model comparison using Tracer v1.7.1 (**?**) based on the Akaike information criteria (AIC). The model with a relaxed clock model and a constant population size had the lowest AIC value and was considered to be the most suitable model for this dataset. From the *posterior* distribution of this model, we selected the maximum clade credibility using TreeAnnotator BEAST2 package. The date of the last common ancestor was estimated to be 1929 with a 95% Highest Posterior Distribution (HPD) of [1890;1953], which we used as a prior distribution to estimate the origin of the HIV-1/O epidemic in Cameroon (Table 4.1).

4.2.2 Epidemiological model and simulations

We assume a Birth-Death (BD) model with two hosts types (see Figure 4.1) with hosts infected by the 181Y virus (denoted I_Y) and hosts infected by the 181C virus (denoted I_C). This model can be described by the following system of ordinary differential equations (ODEs):

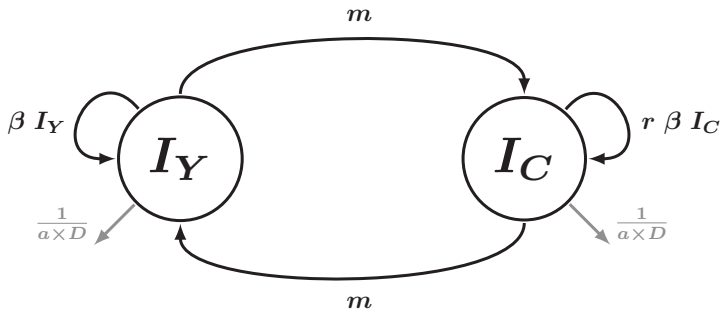


Figure 4.1 – **Flow diagram of the 2-types BD model.**

$$\frac{dI_Y}{dt} = \beta\ I_Y - \frac{1}{aD}\ I_Y - m\ I_Y + m\ I_C \tag{4.1a}$$

$$\frac{dI_C}{dt} = r\ \beta\ I_C - \frac{1}{aD}\ I_C - m\ I_C + m\ I_Y \tag{4.1b}$$

This model comprises five parameters: r, m, R_0, a, and D. R_0 is the number of secondary infections caused by a I_Y individual.

In this model, transmission events are possible within the same type of hosts at a transmission rate β. Parameter r corresponds to the transmission rate differential between the types of hosts. The infectious period duration for both types of hosts is given by the parameter D. This duration can be decreased following treatment for example. The parameter a captures this decreasing. Hence, the removal rate from an infectious compartment (I_Y or I_C) is given by $1/(a\ D)$. The reproduction number R_0 is the number of secondary cases caused by an infectious individual in a fully susceptible host population (**?**). Here, we seek to infer R_0 associated with the presence of residue 181C (R_0^C) where $R_0^C = \beta\ a\ D$ and the R_0 associated with the presence of the 181Y residue (R_0^Y) where $R_0^Y = r\ R_0^C$. The flow of an individual I_C into the compartment I_Y, or inversely, is due to a mutation occurring at rate m. We assume that the mutation is symmetric.

We consider two time intervals : from t_0 to t_1 and from t_1 to t_f, where t_0 is the date of the origin of the epidemic, t_1 correponds to the time when treatments were available and given. Due to the lack of information on this parameter, we used a large prior for t_1 from 1990 and 2010. During the first interval, $a = 1$, while during the second time interval $[t_1, t_f]$, we suppose that the R_0 decreases with $a > 0$.

The prior distributions used are summarized in Table 4.1.

To simulate phylogenies, we use our package TiPS (**?**) implemented in R. This is

done in a two-step procedure. First, epidemiological trajectories are simulated from the compartmental model described in 4.1 using Gillespie's stochastic event-driven simulation algorithm (**?**). These trajectories correspond to a series of dated simulated events (transmission, mutation or end of infection). Phylogenies are then simulated using known sampling dates and simulated trajectories using a backwards-in-time approach. For each simulation, each sampling date is randomly associated with a host compartment using the observed fraction of each infection type. Here 37% of the dates are associated with the 181Y infection (I_Y host compartment) and 63% with the 181C infection (I_C host compartment).

We simulate 85,000 phylogenies from know parameter sets drawn from the prior distributions shown in Table 4.1. These are used to perform the rejection step and build the regression model in the Approximate Bayesian Computation (ABC) inference.

4.2.3 Regression-ABC inference

We first compute all summary statistics listed in **?** on the HIV-1/O time-scaled phylogeny and on each simulated tree.

We then measure multicollinearity between summary statistics using variance inflation factor (VIF). This VIF test leads to the selection of 126 summary statistics out of 330.

We use the abc function from the abc R package (**?**) to infer posterior distributions generated using only the rejection step. The abc function performs a classical one-step rejection algorithm (**?**) using a tolerance parameter P_δ, which represents a percentile of the simulations that are close to the target. To compute the distance between the simulated trees and the target phylogeny for the rejection step, we use the Euclidian distance between normalized vectors of summary statistics of the simulated and the target data.

Table 4.1 – **Prior distributions for the parameters of the model over two time intervals.** t_0 is the date of origin of the epidemic, t_1 corresponds to the introduction and usage of treatment and t_f is the time of the most recent sampled sequence.

Interval	D	r	m	a	R_0
$[t_0, t_1]$	Unif$(1, 20)$	Unif$(0.3, 6)$	Unif$(0, 0.8)$	1	Unif$(0.5, 8)$
$[t_1, t_f]$				Unif$(0.1, 1)$	

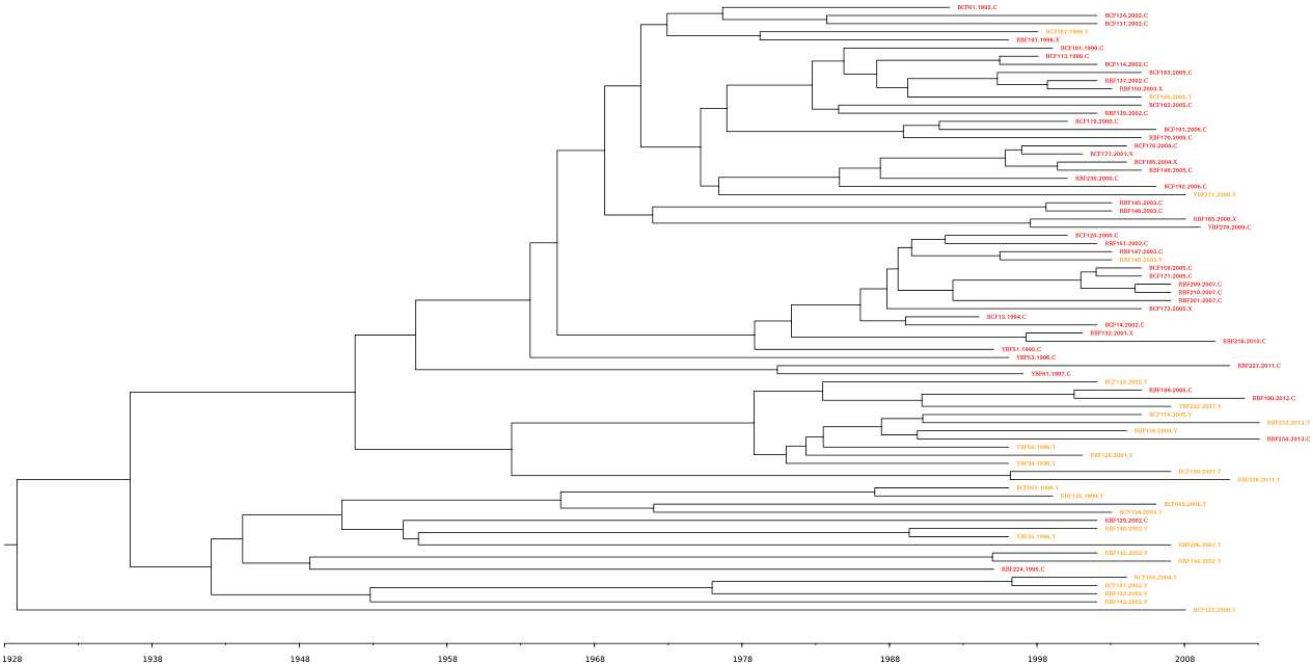


Figure 4.2 – **Viral phylogeny of HIV-1/O.** Sequences with the 181Y residue are represented in orange and those with the 181C residue mutation are in red. The phylogeny was estimated using Bayesian inference (BEAST2). See the Methods for additional deatils.

Finally, we perform linear adjustment using an elastic net (EN) regression (**?**). In the end, we obtain posterior distributions for t_0, t_1, m, r, D, a and R_0 using our ABC-EN regression model with $P_\delta = 0.1$.

4.3 Results

The phylogeny inferred from the dated viral sequences is shown in Figure 4.2. We can clearly distinguish the two main phylogenetic lineages (181C-like and 181Y-like).

Our ABC-regression method relies first on the comparison between simulated data and target data. Here we simulated phylogenies using known sampling dates from the observed phylogeny (Figure 4.2), where each leaf is randomly associated with a type of host following the frequency of the observed fractions of each type of host. Forcing the dates and a host type to a leaf reduces the feasibility of simulations. Indeed, part of simulations may fail, for example, because of a low number of simulated sampled

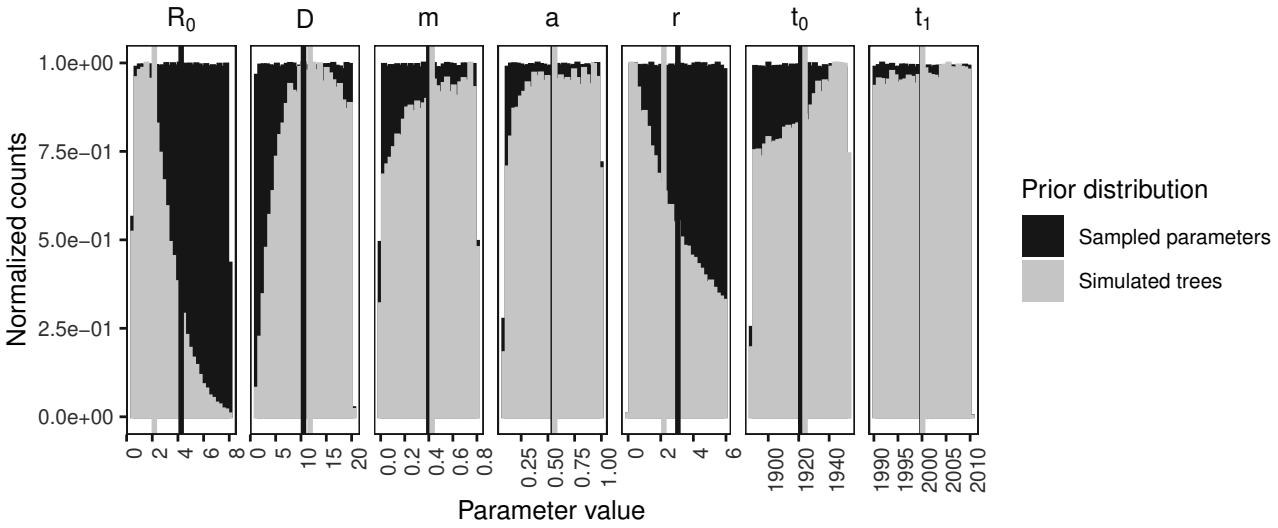


Figure 4.3 – **Histograms of the prior distributions.** Prior distributions are shown for each epidemiological parameter of the model. The initial prior distributions are in black and the final prior distributions after tree simulations are in grey. The vertical lines correspond to the medians of the prior distributions.

individuals in a compartment represented by nodes. Therefore, the feasibility of simulations reduces the prior distribution. Figure 4.3 shows that parameter values, which were initially sampled from a uniform prior distributions (in black), were modified after the tree simulations (in grey).

From the initial prior distribution shown in grey in Figure 4.4, our inference method converges towards posterior distributions for each parameter, shown in red in Figure 4.4.

Using our ABC-EN phylodynamics approach and assuming a two-types BD model, we estimated the origin of the HIV-1/O epidemic in Cameroon (parameter t_0) in 1921 (95% HPD $\in [1911.72; 1929.20]$). This estimation is close to a previous study by **?** which estimated the origin to be in 1920 (95% HPD $\in [1890; 1940]$). However, we were not able to estimate when treatment availability (t_1) affected viral spread, and hence the reproduction numbers R_0^C and R_0^Y and duration of the infectious period.

Our approach allows us to estimate the rate of the mutation of interest m to be 0.021 (95% HPD $\in [0.005; 0.28]$) wich is higher than HIV-1 evolutionnary rate being around $10^{-}3$ given the region of its genome.

The mean duration of the infectious period D is estimated to be 7.80 years

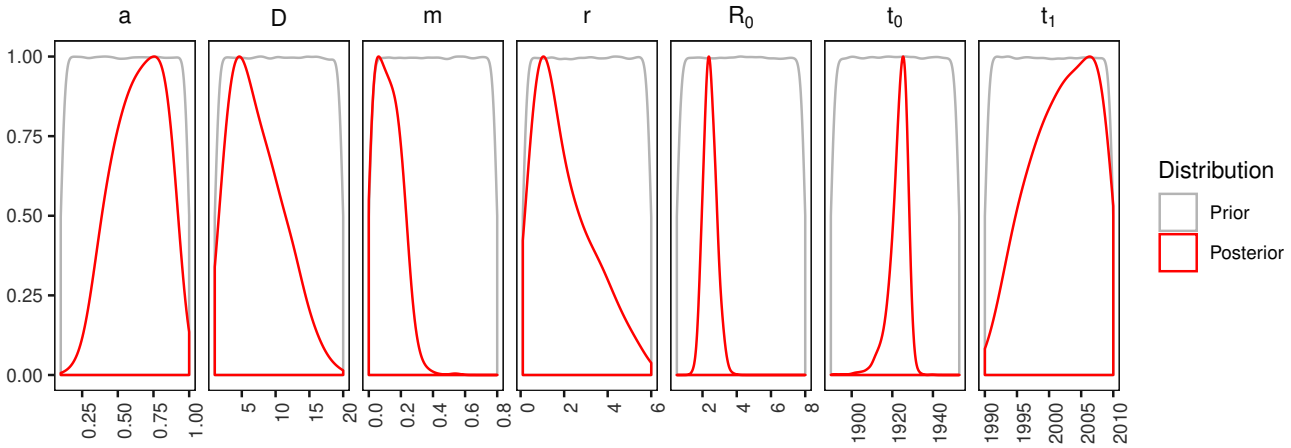


Figure 4.4 – **Parameter prior and posterior distributions.** Prior distributions are in grey and posterior distributions inferred by regression-ABC are in red.

[1.57; 15.70] before the availability of treatments. The differential transmission ratio r quantifying the effect of the mutation 181C on the reproduction number is estimated to be 2.06 [0.27; 5.11]. The R_0 of individuals I_C with $R_0^C = rR_0^Y$ is therefore around 4.87. However, the 95% HPD is large and does include values below 1.

Finally, the parameter a leading to the decrease of D and therefore of both R_0^C and R_0^Y following the introduction of the treatments (after t_1) is estimated to be 0.64 [0.32;0.92]. Thus, following the introduction of treatments, the R_0 of individuals I_Y (R_0^Y) decreases from 2.37 to 1.53, the R_0 of individuals I_C (R_0^C) decreases from 4.87 to 3.14 and finally the mean infectious period duration D decreases from 7.8 years to 5.0 years.

Using our ABC-EN approach we were able to date the origin of the HIV-1/O epidemic in Cameroon and to estimate the R_0. However, we were not able to make any conclusion about the advantage for the virus to present the 181C mutation at the reverse transcriptase. HIV-1/O accounts for nearly 1% of HIV infections (10,000 to 100,000 cumulative cases estimated) in Cameroon. Our set of 75 sequences is therefore not representative of the epidemic. These results would be strengthened by the addition of more sampled sequences related to this epidemic that would contribute to refining the posterior distributions, in particular for the infectious period duration.

Chapitre 5

Phylodynamique du SARS-CoV-2 en France

5.1 Résumé

En décembre 2019, une épidémie de pneumonies virales a émergé dans la ville de Wuhan dans la province de Hubei en Chine (**?**). Le 9 janvier 2020, l'Organisation mondiale de la santé a annoncé la découverte d'un nouveau betacoronavirus présenté comme l'agent provoquant ces pneumonies. Dès janvier 2020, l'épidémie se répand hors de Chine. En France, les trois premiers cas sont recensés le 24 janvier 2020. Le 11 mars 2020, l'épidémie est déclarée pandémique par l'OMS. Au fur et à mesure que le virus se propage, comprendre et contrôler la propagation du virus devient alors un enjeu au niveau mondial.

Les données épidémiologiques telles que les données d'incidence mais aussi les données de séquences génomiques virales ont rapidement été partagées et rendues accessibles au public. Par exemple, le site web du programme *Our World in Data* (OWID, `https://ourworldindata.org/coronavirus`) qui se concentre sur des problèmes comme la pauvreté, les droits de l'homme mais aussi la santé, propose la visualisation et le téléchargement de données d'incidence (nombre de cas d'hospitalisations, cas de mortalité) et sur les vaccinations, et ce pour quasiment tous les pays. Un autre exemple est la base de données en ligne GISAID, qui en 2021 a dépassé le 1 000 000 de séquences du virus SARS-CoV-2 partagées. Cependant, une disparité est observée entre pays en termes de séquençage et/ou de partage de séquences. Ceci peut être expliqué du fait que les technologies de séquençage génomique ne sont pas largement disponibles, notamment dans les pays en développement où les dépenses sont limitées en éducation et en recherche (**?**). Cependant, les différences en nombre

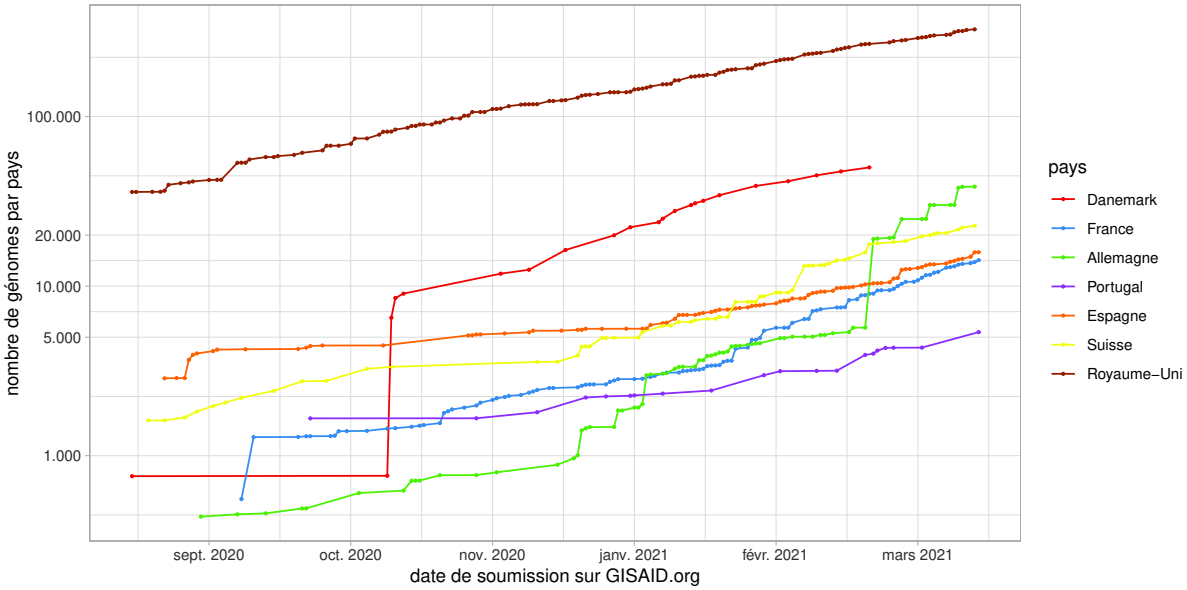


FIGURE 5.1 – Nombre de séquences génomiques du SARS-CoV-2 soumises sur la base de données GISAID, par pays, entre début septembre 2020 et mi mars 2021.

de séquences soumises à GISAID sont très marquées même entre pays européens voisins puisqu'en au mois de mai 2021 environ 17 000 proviennent de Belgique, 23 000 de France, 77 000 d'Allemagne et 383 000 du Royaume-Uni (Figure 5.1).

Ce partage de données a permis le développement de nombreux outils de visualisation et de prévision. Un exemple est l'outil en ligne Rt2 (**?**) que nous avons développé `https://cloudapps.france-bioinformatique.fr/covid19-rt2/`. Cet outil repose sur l'analyse statistique des variations temporelles des données épidémiologiques telles que le nombre cumulé de cas confirmés et de décès, le nombre de personnes hospitalisées ainsi que le nombre de personnes hospitalisées en unité de soins intensifs. L'outil permet donc de visualiser la dynamique épidémiologique à travers l'estimation du nombre de reproduction au cours du temps et ce, selon la localité géographique (pays, région ou département). Un autre exemple est l'outil de visualisation temporelle et spatiale de la propagation de virus, tels que les virus Zika ou de la grippe, à travers des analyses phylogénétiques, Nextstrain, qui a étendu ses analyses au virus du SARS-CoV-2.

L'essor de la phylodynamique a suivi l'accroissement du nombre de données de séquences. Parmi les premières études, **?** a rapidement estimé un taux de substitution à $8,8 \cdot 10^{-4}$ substitutions par site par an. **?** ont estimé la date d'apparition du premier

cas de SARS-CoV-2 dans la province du Hubei en Chine. **?** ont étudié l'origine et l'impact des événements de super-transmission sur la propagation locale du virus dans la région de la ville de Boston. **?** ont estimé l'origine géographique et temporelle des lignées de transmission génétiquement distinctes observées au Royaume-Uni. **?** ont réalisé une première étude phylodynamique à partir de 97 séquences génomiques récoltées dans les régions du Nord de la France entre le 24 janvier et le 24 mars 2020 et ont identifié plusieurs introductions virales en France.

Nous avons réalisé une étude phylodynamique de la première vague en France en analysant 196 séquences génomiques virales récoltées chez des patients détectés en France, disponibles au 4 avril 2020 sur la base de données GISAID. Nous avons généré différents sous-ensembles de séquences à analyser du fait que certaines dates d'échantillonnage du jeu de données étaient plus représentées que d'autres, mais aussi pour analyser l'effet du signal temporel que peuvent contenir des jeux de données de taille différente et qui s'étendent sur des périodes différentes, sur l'estimation des paramètres épidémiologiques.

L'étude s'appuie sur l'utilisation de deux approches d'inférence phylodynamique : le modèle de coalescent avec une taille de population exponentielle (**?**), implémenté dans le logiciel BEAST (**?**), et le modèle de naissance et de mort permettant la variation des paramètres au cours du temps, implémenté dans le package BDSKY (**??**) du logiciel BEAST2 (**?**).

En fixant le taux de substitution à $8,8 \cdot 10^{-4}$, nous avons daté l'origine de la première vague épidémique en France entre la mi-janvier et le début de février. Cette incertitude vient du nombre limité de séquences disponibles, notamment au tout début de l'épidémie, mais aussi de la diversité génétique relativement faible. Ces résultats sont cohérents avec les modèles de l'équipe basés sur les données d'incidence (**?**).

Grâce à la première méthode, nous avons estimé le temps de doublement de l'épidémie, défini par la relation $t_d = r/\ln(2)$ où r est le taux de croissance exponentiel de l'épidémie, pour deux jeux de données. À partir de 61 séquences récoltées entre le 21 février et le 12 mars on trouve une valeur médiane de $t_d = 2,5$ jours avec un intervalle de crédibilité à 95% entre $1,19$ et $4,94$ jours. En ajoutant plus de séquences dont les dates d'échantillonnage s'étendent jusqu'au 24 mars 2020, cette valeur devient $t_d = 3,7$ jours (IC 95% $[2,39; 5,47]$). Ainsi, nos résultats montrent que plus le nombre de données est grand et s'étend sur une plus grande période, moins l'épidémie double en taille rapidement, suggérant donc un ralentissement après le 12 mars.

Nous avons aussi pu estimer, en utilisant la seconde méthode, la vitesse de propagation du virus en estimant le nombre de reproduction effectif (R_t) sur trois périodes

temporelles : avant le 19 février 2020, du 19 février au 7 mars et du 7 mars au 24 mars. Le peu de signal sur la première période fait que les valeurs de R_t ne diffèrent pas du *prior* qui suit une loi log-normale de paramètres d'espérance $\mu = 0$ et de variance $\sigma^2 = 1.2$ avec une valeur maximale de 10. Sur la seconde période, le R_t estimé est élevé avec une valeur médiane de $2,6$ (IC 95% $[1,66; 4,74]$). Sur la troisième période, il est de $1,4$ (IC 95% $[1,13; 2,03]$). Cette diminution de R_t est cohérente avec la mise en place du premier confinement national au 17 mars 2020. Enfin, nous avons estimé la durée du la période d'infectiosité, période pendant laquelle un individu infecté peut engendrer d'autres infections, à environ 5 jours (IC 95% $[3; 7]$ jours).

Une des limites de cette étude tient à la faiblesse du signal phylogénétique, qui peut biaiser les estimations temporelles. Cette étude illustre aussi l'importance de dépister et séquencer dès le début d'une épidémie afin d'avoir de meilleures estimations de paramètres tels que le nombre de reproduction ou la date d'émergence de la première vague en France, et ainsi comprendre la dynamique de propagation du virus.

Cette étude constitue l'objet d'un article qui a été recommandé après processus de revue par les pairs par *Peer Community In Evolutionary Biology*. L'article est présenté dans la section 5.2.

5.2 Early phylodynamics analysis of the COVID-19 epidemic in France

5.2.1 Introduction

On Jan 8 2020, the Chinese Center for Disease Control announced that an outbreak of atypical pneumonia was caused by a novel coronavirus (**?**). The genetic sequence of what is now known as SARS-Cov-2 was released on Jan 10 (**??**). This was less than two weeks after the initial report of the outbreak by the Wuhan Health Commission, which took place on Dec 31 2019. Never has a novel pathogen been sequenced so rapidly.

The number of sequences in the databases grew rapidly thanks to an altruistic and international effort of virology departments all around the world gathered via the Global Initiative on Sharing All Influenza Data (GISAID, `https://www.gisaid.org/`). Early results allowed better understanding the origin of SARS-Cov-2 and identification of a bat coronavirus (SARSr-CoV RaTG13) as its closest relative with more than 96% homology, as well as some potentially adaptive mutations (**???**).

The available sequences were also analysed using the field of phylodynamics (**???**), which aims at inferring epidemiological processes from sequence data with known sampling dates. Most of these analyses were shared through the website `virological.org`. In particular, using 176 genomes from which he extracted 85 representative sequences (to avoid a potential cluster effect), **?** estimated the molecular clock to be approximately $8 \cdot 10^{-4}$ substitutions per position per year, with a 95% Highest Posterior Density (HPD) between $1.4 \cdot 10^{-4}$ and $1.3 \cdot 10^{-3}$ subst./pos./year, which yielded a date of origin of the outbreak mid-Nov 2019, with a 95% HPD spanning from Aug 27 to Dec 19. Further analysis with more recent sequences found a median estimate of $1.1 \cdot 10^{-3}$ subst./pos./year with a similar HPD (**?**). In their work, **?** explored a variety of priors for the analysis and found similar orders of magnitude for the molecular clock estimate. They also applied a birth-death model to estimate several parameters including the temporal reproduction number ($\mathcal{R}_t$) but a difficulty is that not all sequences originated from China and the sampling rate could also vary. Finally, **?** performed one of the early analyses of the outbreak using coalescent models, allowing them to estimate the date of the origin of the epidemic in early Dec 2019 (with a 95% CI: between 6 Nov and 13 Dec 2019) and the doubling time of the epidemic to be 7.1 days (with a 95% CI: 3.0-20.5 days). These reports mention several caveats, which are due to the limited number of sequences, the limited amount of phylogenetic signal, the potentially unknown variations in sampling rates and the sampling across multiple countries.

The first COVID-19 cases were detected in France from Jan 24, 2020, mostly from travellers, but these remained isolated until Feb 27, when the national incidence curve of new COVID-19 cases started to increase steadily. Limited measures were announced on Feb 28, but schools were closed from Mar 16, and a nationwide lock-down was implemented from Mar 17. On Apr 19, the prime minister gave the first official estimate of the basic reproduction number ($\mathcal{R}_0$), which was 3.5, and of the temporal reproduction number after the lock-down, which was 0.5 (**?**).

We study the COVID-19 epidemic in France by analysing 196 genomes sequenced from patients diagnosed in France that were available on Apr 4, 2020 thanks to the GISAID and to French laboratories (see the online Supplementary Table for the full list). **?** provided a first picture of the general genomic structure of French epidemic using 97 genomes from samples collected in the north of France between Jan 24 and Mar 24, 2020. They identified several independent introductions of the virus in France but also found that the majority of the sequences belong to a major clade. This clade belongs to a larger clade labelled as G by GISAID, A2 by the nextstrain (`http://nextstrain.org/`) platform and B.1 following the dynamical taxonomy introduced by **?**. We refer to it as the clade related to the epidemic wave.

Our early phylodynamics analyses focus on the epidemic doubling time, the generation time, and the temporal reproduction number $\mathcal{R}(t)$. Current data does not allow us to perform a phylogeographic study and future work will investigate the structure of the epidemic within France, as well as potential dispersion between regions.

5.2.2 Materials and Methods

Data and quality check

On Apr 4, 196 sequences were available from samples originating from France via the Global Initiative on Sharing All Influenza Data (GISAID, `https://www.gisaid.org/`) thanks to the work of the two Centre National de Référence and local virology laboratories. These sequences only provide a partial view of the epidemic as they originate from 8 the 18 French regions (Figure S1). Sequences were aligned and cleaned using the Augur pipeline developed by nextstrain (**?**). One sequence was removed due to low quality. The list of the sequences used is shown in the Online Supplementary Table.

We screened the dataset with RDP4 (**?**) using default parameter values and did not detect any recombination events.

Phylogenetic inference

We first performed a maximum likelihood inference of the phylogeny using SMS (**?**) and PhyML (**?**). The mutation model inferred by SMS was GTR and was used as in input in PhyML. Other PhyML parameters were default. The resulting phylogeny was time-scaled and rooted using the software LSD (**?**) using a constrained mode with the sampling dates and a molecular clock rate fixed to $8.8 \cdot 10^{-4}$ substitutions/position/year (the tree is provided in a Newick format in the Online Supplementary Materials).

We then used Beast 1.8.3 (**?**) to perform inference using a Bayesian approach. More specifically, we assumed an exponential coalescent for the population model (**?**). We used the default settings for the model, which correspond to a gamma distribution for the growth rate prior $\Gamma(0.001, 1000)$ and an inverse prior for the population size $1/x$ (see Supplementary Methods 5.2.5).

We also used Beast 2.3 (**?**) to estimate key parameters using the birth-death skyline (BDSKY) model (**??**). One of these parameters is the temporal reproduction number ($\mathcal{R}_t$) and we here assume three periods in the epidemic (which means we estimate 3 values $\mathcal{R}_1$, $\mathcal{R}_2$ and $\mathcal{R}_3$). Another parameter is the the recovery rate, i.e. the rate at which the infectiousness ends. The final key parameter is the sampling rate, the inverse of which corresponds to the average number of days until an infected person is sampled. The ratio between the sampling rate and the sum of the sampling and the recovery rates indicates the fraction of infections that are actually sampled. By sampled, we mean that the patient is identified and the virus population causing the infection is sequenced. Note that we assume sampled hosts are not infectious anymore. We considered multiple priors for the rate of end of the infectious period by setting a lognormal prior LogNorm$(90, 0.5)$ and a uniform prior Unif$(5, 350)$. We assumed a beta prior $\beta(1.0, 1.0)$ for the sampling rate (see Supplementary Methods 5.2.5). As in previous models (**?**), we set the sampling rate to 0 before the first infected host is sampled (here on Feb 21, 2020).

For both analyses in Beast, we assumed a GTR mutation model, following the results of SMS. We also assumed a uniform prior $U(0, 1)$ for the nucleotide frequencies and a lognormal prior for parameter κ, LogNorm$(1, 1.25)$.

Regarding the molecular clock, earlier studies have reported a limited amount of phylogenetic signal in the first sequences from the COVID-19 pandemic. Given that we here focus on a subset of these sequences, we chose to fix the value of the strict molecular clock to $8.8 \cdot 10^{-4}$ substitutions/position/year, following the analysis by **?**. In Appendix, we study the influence of this value on the results by setting it to a lower ($4.4 \cdot 10^{-4}$ subst./pos./year) or a higher ($13.2 \cdot 10^{-4}$ subst./pos./year) value. Finally, we also estimate this parameter assuming a strict molecular clock. The most

recent estimates suggest that the intermediate and high value are the most realistic ones (**?**).

Data subsets

We analysed subsets of the whole data set. Our largest subset excluded 10 sequences that did not belong to the French epidemic wave clade and therefore contained 186 sequences. Figure S1 shows the sampling date and French region of origin for each sequence. In general, the proportion of infections from the French epidemic that are sampled is expected to be in the order of 0.01%.

Some sampling dates are over-represented in the dataset, which could bias the estimation of divergence times (**??**). To correct for this, we sampled 6 sequences for each of the days where more than 6 sequences were available. This was done 10 times to generate 10 datasets with 122 sequences (France122a to France122h).

To investigate temporal effects using the coalescent model, we created three other subsets of the France122a dataset: "France61-1" contains the 61 sequences sampled first (i.e. from Feb 21 to Mar 12), "France61-2" contains the 61 sequences sampled more recently (i.e. from Mar 12 to Mar 24), and "France81" contains the 81 sequences sampled first (i.e. from Feb 21 to Mar 17).

With the exponential coalescent model (denoted DT for "doubling time"), we analysed all subsets of data (France61-1, France61-2, France81, and all the 10 France122 datasets), whereas for the BDSKY model we show the main dataset (France186) and analyse the 10 subsets with 122 leaves in Appendix.

5.2.3 Results

Phylogeny and regional structure

Figure 5.2 shows the regional structure of the French epidemic. Sequences corresponding to black leaves were ignored in the subsequent analyses because they do not belong to the main clade. Most of these originate from travelers isolated upon arrival in France, which explains their under-representation in the ongoing epidemic wave.

Focusing on the main clade, we see that all the leaves originate from a common branching event, which is approximately half-point of the phylogeny. The polytomy in this point likely indicates a lack of phylogenetic signal. Addressing this issue will require more sequences from the early stages of the epidemic wave since currently the earliest sequence in this major clade is from Feb 21, 2020.

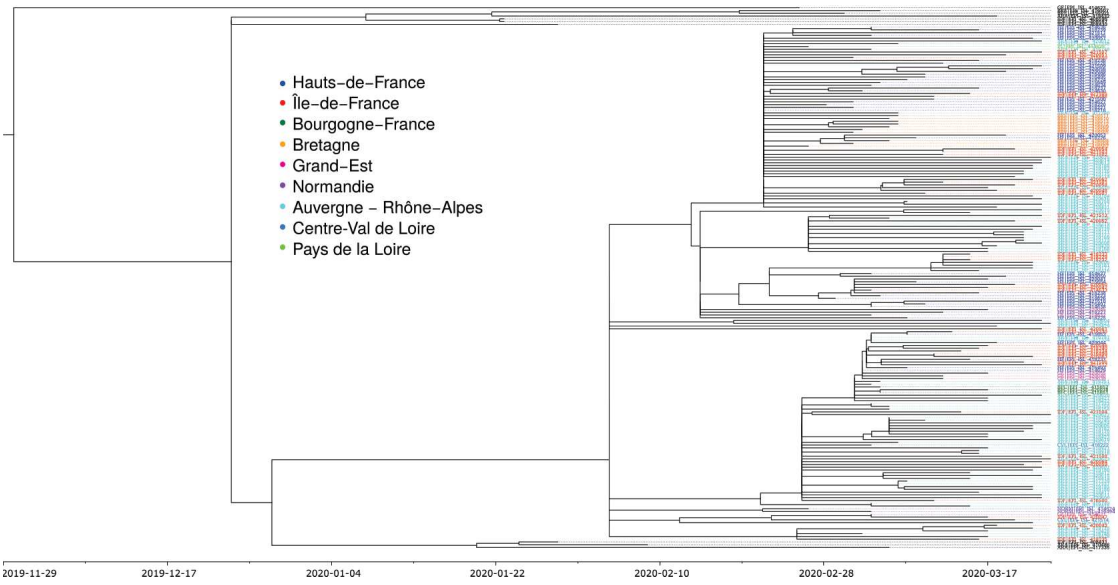


Figure 5.2 – **Phylogenetic structure of 196 SARS-Cov-2 genomes from France.** Color shows the French region of sampling. Sequences in black were removed from the analysis because they fall outside the main clade corresponding to the epidemic wave. The phylogeny in a Newick format is available in the online Supplementary Materials.

Colors indicate the regional structure of the French epidemic. As expected, we see some regional clusters. We also see that sequences from the same region belong to different subclades of the major clade, which is consistent with multiple introductions or dispersal between regions. Several French regions are not represented in the analysis. This is largely reflects the nature of the French COVID-19 epidemic, which has been stronger in the East of France and in the Paris area. This is also why this work focuses more on the speed of spread of the epidemic than on its general structure, which will be the focus of a future study (see also the work by **?**).

In the following, we focus on the main clade associated with the epidemic wave.

Dating the epidemic wave

We first report the estimation of the time to the most recent common ancestor (TMRCA) of the 186 sequences that belong to the epidemic wave. Although this is the ancestor of the vast majority of the French sequences grouped in the B.1 clade

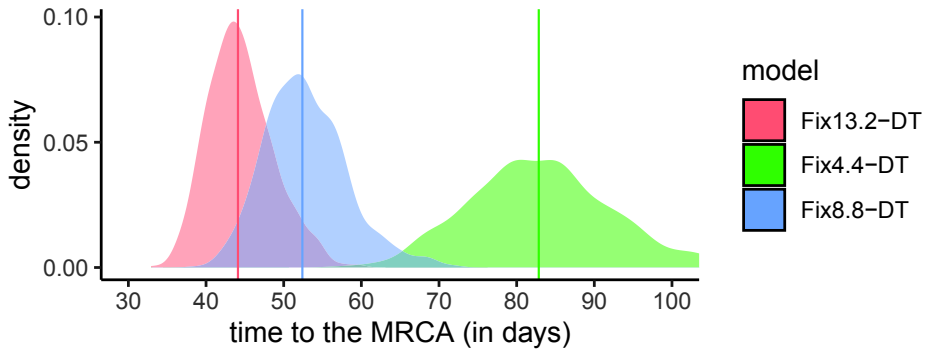


Figure 5.3 – **Time to the origin of the French epidemic wave as a function of the molecular clock.** This estimate was obtained assuming an exponential growth coalescent population model and a fixed molecular clock (see Figure S4 for the BDSKY model). Colors indicate substitution rates and numbers (4.4, 8.8, 13.4) refer values to be multiplied by 10^{-4} subsitutions/site/year. The slower the clock rate, the further away in time the most recent common ancestor (MRCA). Vertical lines show the distribution medians. The most recent sample dates from Mar 24, 2020.

(also referred to as G or A2 clade), the associated infection may have taken place outside France because the epidemic wave may be due to multiple introduction events (although from infections caused by similar viruses given the clustering).

Estimates of SARS-Cov-2 molecular clock should be treated with care given the limited amount of phylogenetic signal (**??**). This is particularly true in our case since we are analysing a small subset of the data. In the Appendix, we present the analysis of the temporal signal in the data using the TempEst software (**?**) and show that it strongly relies on early estimates that do not belong to the epidemic wave clade (Figure S2).

As shown in Figure 5.3, the molecular clock value directly affected the time to the most recent common ancestor for the coalescent model. This was also true for the BDSKY model, where the prior shape for the recovery rate, lognormal or uniform, had little impact compared to the assumption regarding the molecular clock (Figure S3). For both models, sampling of the 122 sequences amongst the 186 has a much smaller impact (Figure S5).

Table 5.1 shows the dates for models with different evolution rates and different population models (exponential coalescent or BDSKY). Note that smaller datasets may not include the most recent samples.

For most of our datasets and models, the origin for the clade corresponding to the sequences from the French epidemic wave is dated between mid-Jan and early Feb. This large interval is due to the scarcity of "old" sequences (the first one collected in this clade dates from Feb 21) and on the fact that this clade averages the epidemic in several regions of France, which could have been seeded by independent introductions from outside France. The date provided by the slowest molecular clock (Fix4.4-DT) seems at odds with the data as we will see below.

To evaluate the effect of a potential sampling bias, we also estimate the time to the MRCA for 10 different sets of 122 sequences (Figure S5). We found similar median values for 9 of these 10 random datasets. Notice that the value of their parent dataset (France186), was slightly larger. For the BDSKY model, the effect was even less pronounced (Figure S4).

Overall, these dates (except for the slowest molecular clock) are consistent with those obtained by **?** regarding the beginning of the epidemic in China, which is dated November 17, 2019 with a confidence interval between Aug 27 and Dec 19, 2020. This interval is highly dependent on the number of available sequences as there are documented (but unsequenced) cases of COVID-19 in China early Dec 2019 (**?**).

Doubling time

Using a coalescent model with exponential growth and serial sampling (**?**), we can estimate the doubling time, which corresponds to the number of days for the epidemic wave to double in size. This parameter is key to calculate the basic reproduction number $\mathcal{R}_0$ (**?**).

In Figure 5.4, we show this doubling time for datasets that cover the whole (France122a), the first three quarters (France81), and the first half (France61-1) of the time period. Since the first dataset includes more recent sequences than the second, which itself includes more recent sequences than the third, our hypothesis is that we can detect variations in doubling time over the course of the epidemic. For completeness, we also show the results for the dataset covering only the second half of time period (France61-2).

Adding more recent sequence data indeed leads to an increase in epidemic doubling time. Initially, with the first 61 sequences (which run from Feb 21 to Mar 12), the epidemic spreads rapidly, with a median doubling time of 2.5 days. With the addition of sequences sampled between Mar 12 and 17, the doubling time increases to 3.3 days. Finally, by adding sequences sampled between Mar 17 and 24, the doubling time rises to 3.7 days.

Importantly, the lower the number of sequences, the more the inferences become

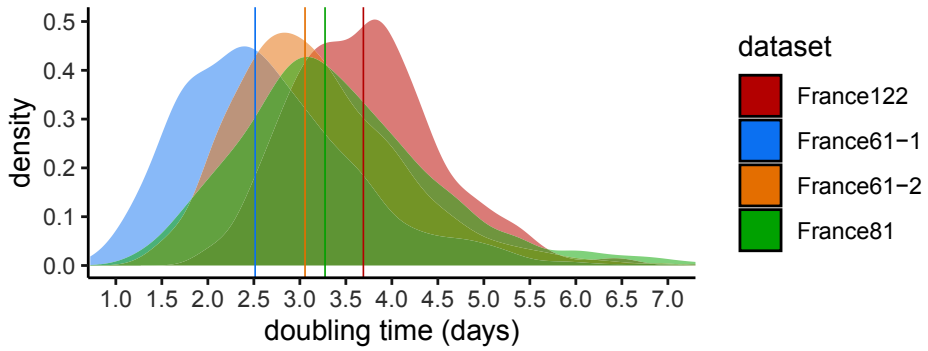


Figure 5.4 – **Epidemic doubling time.** We assume an exponential growth coalescent model with a fixed molecular clock. The four datasets differ in the sequences analysed (see the Methods). Vertical lines show the distribution medians.

sensitive to the sampling scheme. This can be visualised with the fact that the doubling time obtained with the 61 most recent sequences (France61-2), which is, as expected, higher than that obtained using the 61 oldest sequences (France61-1), is lower than that obtained using all sequences (France122). Our interpretation is that phylogenetic signal becomes limited when only 61 sequences are considered. This can also be seen when estimating the date of origin of the epidemic: with the 61 most recent sequences, the date is comparable to that inferred using 122 sequences but assuming a faster evolutionary rate (Table 5.1). A more recent origin of the epidemic estimated with this subset of the data would directly lead to a lower epidemic doubling time.

To further explore the effect of sampling, we estimate the doubling time on 10 different sets of 122 sequences and find a limited effect on the median value (Figure S7). Notice that the value of the parent dataset (France186), is slightly larger.

We also study the effect of the molecular clock, i.e. the substitution rate, on the doubling time (Figure S6). As already mentioned above, the higher the molecular clock value, the lower the doubling time. However, for our realistic molecular clocks, the effect is limited: the median is 3.4 days assuming a high value for the molecular clock and 3.7 days for our default (medium) value. The low value of the molecular clock led to a high median doubling time of 5.6 days. This is at odds with the incidence data in France, which indicates an exponential growth rate of 0.23 days $^{-1}$ which corresponds to a doubling time of 3 days, suggesting that our default molecular clock is more realistic.

In comparison, phylodynamic inferences made from data from China with 86

genomes (**?**) found a median doubling time of about 7 days with a confidence interval between 4.7 and 16.3 days). One reason for the slower growth rate of the epidemic compared to ours is that we have focused on one rapidly expanding clade of the epidemic and neglected the smaller clades. Another possibility could be related to the timing of the sampling (early or late in the infection).

Effective infection duration

The birth-death skyline (BDSKY) model (**?**) allows us to estimate the effective duration of infection, which is defined in the model as the rate of becoming non-infectious (either through recovery, death, or sampling), and the reproduction number of the epidemic (i.e. the number of secondary infections caused by an infected host). The exponential growth coalescent model described above cannot distinguish between these two quantities. However, the BDSKY model requires more parameter values to be estimated.

The BDSKY model estimates separately the recovery rate and the sampling rate, and it is important to account for the latter because patients whose infections are sequenced can be assumed not to transmit the infection after this detection. The sampling rate after Feb 21 (it is set to 0 before that date) is estimated at 0.093 days $^{-1}$ with a (wide) 95% confidence interval between 0.006 and 0.627 days $^{-1}$. If we analyse this in days, the median value of the distribution yields 10.8 days and is consistent with the fact that in the French epidemic most of the screening for SARS-Cov-2 is done on severe cases upon hospital admission.

The distribution of infectious durations is obtained by taking the inverse of the sum of the sampling rate and the recovery rate. The median of this distribution is 5.12 days and 95% of its values are between 2.89 and 7.05 days (Figure 5.5). Note that this is an effective infection duration in that public health interventions can reduce it, e.g. by preventing transmission in the later stages of the infection, such that people can be infected but not infectious.

In Supplementary Figure S8, we show that the estimate for the effective infection duration is sensitive to the shape of the prior assumed for the recovery rate. Indeed, if we use a less informative (uniform) prior then the median sampling rate estimate is larger and the median infectious period estimated is shorter.

Reproduction number

With the BDSKY model, we can estimate the temporal reproductive number, noted $\mathcal{R}(t)$, since the onset of the epidemic wave. Here, given the limited temporal

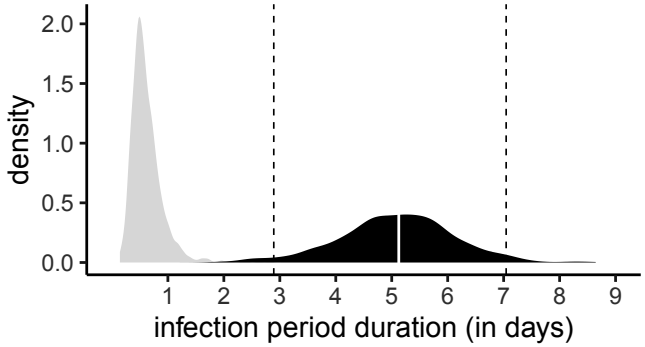


Figure 5.5 – **Distribution of effective infection duration.** The prior distribution is shown in gray, and the posterior distribution in black. The white line shows the distribution median and the dashed line the 95% highest posterior density (HPD), which is between 3 and 7 days in agreement with results obtained using contact tracing data.

signal, we only divided the time into 3 intervals to estimate three reproduction numbers: $\mathcal{R}_1$ before Feb 19, $\mathcal{R}_2$ between Feb 19 and Mar 7, and $\mathcal{R}_3$ between Mar 7 and Mar 24.

These results are very consistent with those obtained for the doubling time, even if the time periods are different. For the period before Feb 19, the estimate is the least accurate with values of $\mathcal{R}_1$ with a median of 1.05 but at 95% Highest Posterior Density (HPD) between 0.13 and 3.22. The lack of information can be seen in Figure 5.6 as the posterior distribution (gray area) is very similar to the prior (dashed curve). This is consistent with the fact that the oldest sequence dates from Feb 21, while the tree root is estimated at the beginning of Feb. Over the second time period (in orange), the distribution shape is similar to that of the prior but the median is very different and rapid growth is detected with a median value of $\mathcal{R}_2$ of 2.56 (95% HPD between 1.66 and 4.74). Finally, the most recent period after Mar 7 is the most accurate and detects a slowing down of the epidemic with a $\mathcal{R}_3$ of 1.38 (95% HPD between 1.13 and 2.03)

In Appendix, we show that these estimates for $\mathcal{R}_t$ are robust to the prior used for the recovery rate (Figure S9). They are also robust to the sampling of 122 of the 186 sequences (Figure S10).

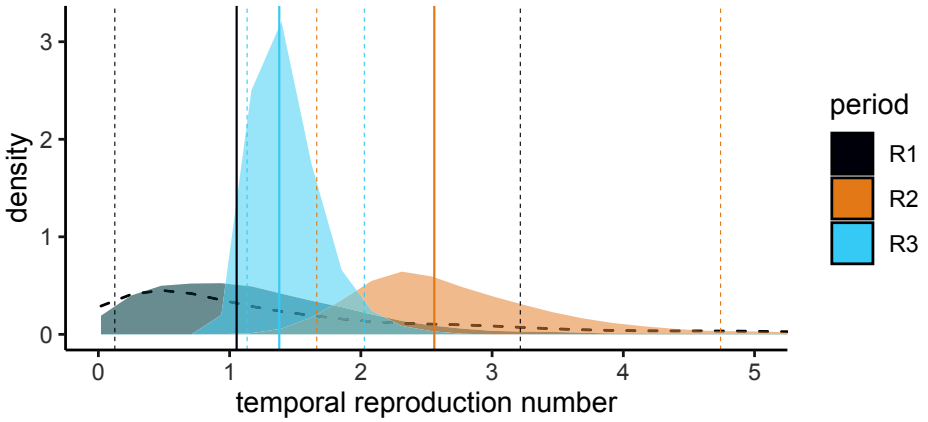


Figure 5.6 – **Temporal reproduction numbers inferred using the BDSKY model.** These results are obtained for the France186 dataset. The black dashed curve show the prior distribution, and the posterior distributions are in color. Vertical plain lines show distribution medians, while vertical dashed lines indicate the 95% highest posterior density (HPD).

5.2.4 Discussion

Analysing SARS-Cov-2 genome sequences with a known date of sampling allows one to infer phylogenies of infections and to estimate the value of epidemiological parameters of interest (**??**). We performed this analysis based on the 196 sequences sampled in France and available on Apr 4, 2020. We focus in particular on the largest clade regrouping 186 of the most recent sequences and likely corresponding to the epidemic wave that peaked in France early Apr 2020.

Before summarizing the results, we prefer to point out several limitations of our analysis. First, the French clade we analysed is in fact an international clade: although most French sequences appear to be grouping into two main subclades within this clade, it is possible that the variations in epidemic growth that we detect are more due to European than French control policies. Second, some French regions (e.g. Auvergne-Rhône-Alpes) are more represented than others (e.g. Occitanie is absent), which could bias the analysis at the national level. However, the coverage is largely proportional to the state of the epidemics in France in March, where the Paris area and the East of France were more heavily impacted. Therefore, we expect the addition of sequences from less impacted regions to have a limited effect on our doubling time and reproduction number estimates. Finally, the molecular clock had to be set in this analysis because we do not have enough samples from the month of

Feb in France.

Despite these limitations, our results obtained early Apr confirm a slowing down of the epidemic in France, where the epidemic peak in terms of ICU admissions was reached on Apr 1. Indeed, by adding sequences sampled between Mar 12 and 24 to the phylogeny, the doubling time of the epidemic estimated by an exponential growth coalescent model increased by 48%. This slowdown is more clearly detected using a birth death model via the temporal reproduction number $\mathcal{R}(t)$: the median value decreased by 41% after Mar 12. This is consistent with the implementation of strict control measures in France as of Mar 17. These variations and even these orders of magnitude are consistent with our estimates based on the time series of incidence of new hospitalizations and deaths (**?**). However, these results were obtained with relatively few sequences and a denser sampling is needed to be more confident in our ability to detect an epidemic slowdown.

Finally, the BDSKY model also provides us with an estimate of the effective infection duration. This can be seen as the generation time of the epidemic, i.e. the number of days between two infections, and is an essential component in the calculation of $\mathcal{R}_0$ (**?**). The result we obtain, with a 95% Highest Posterior Distribution between 3 and 7 days and a median of 5.2, is highly relevant biologically and comparable to results obtained using contact tracing data. For instance, **?** estimated a serial interval, which corresponds to the time between the onset of the symptoms in a 'donor' host and that in a 'recipient' host, with a median of 5 days and a standard deviation of 1.9 days. To date, there is no estimate of the serial interval in France.

By increasing the number of SARS-CoV-2 genomic sequences from the French epidemic (and the number of people working on the subject), in particular sequences collected at the beginning of the epidemic, it would be possible to better estimate the date at which the epidemic wave took off in France, improve the estimate for the infection generation time and the reproduction number, better understand the spread between the different French regions, and estimate the number of virus introductions into the country.

Finally, it is important to set these results into their context. As acknowledged in the introduction, the French state only acknowledged the magnitude of the COVID-19 epidemic on the last days of Feb 2020 and these genomes were mostly collected between Feb 21 and Mar 24. Most of this analysis was published on Apr 6. At this time, the epidemic peak was barely noticeable in the incidence data. Furthermore, the serial interval, which is used to estimate the generation time of the infection and classically measured from contact tracing data, is still unknown in France. These results illustrate the contribution phylodynamics can make to public health during a crisis.

Data accessibility

Data are available online at https://platform.gisaid.org/ with prior registration to the GISAID at www.gisaid.org

Online supplementary material

A TSV file listing the sequences used and a text file containing the unrooted phylogeny in a Newick format can be found on www.medrxiv.org/content/10.1101/2020.06.03.20119925v2.supplementary-material

Acknowledgements

We thank the patients, nurses, doctors, and all the French laboratories who made this work possible by generating and sharing the virus genome sequences. We also thank the ETE modelling group for discussion.

Gonché Danesh is supported by a doctoral grant from the Fondation pour la Recherche Médicale (FRM grant number ECO20170637560).

This work was partly supported by the *Urgence Recherche Covid-19* call from the Occitanie Region (contract 20007477) and the ANR (PhyEpi project).

We acknowledge the IRD itrop high-performance computer (South Green Platform) at IRD montpellier for providing resources that have contributed to the results presented in this work (more details on bioinfo.ird.fr).

Preliminary versions of this work were posted on Apr 9 (in French on http://covid-ete.ouvaton.org) and on Apr 21 (in English on http://virological.org/).

We thank Luca Ferretti and two anonymous reviewers for their carefull reading and numerous suggestions.

Version 3 of this preprint has been peer-reviewed and recommended by Peer Community In Evolutionary Biology (https://doi.org/10.24072/pci.evolbiol.100107)

Conflict of interest disclosure

The authors of this preprint declare that they have no financial conflict of interest with the content of this article. SA is a recommender for PCI Evolutionary Biology and PCI Ecology.

5.2.5 Appendix

BEAST priors

MCMC chains were run for $5 \cdot 10^8$ iterations. The first 10% runs were discarded as a burn in and convergence was assessed using Effective Sample Size (ESS). All parameters had ESS greater than 200.

Original XML files cannot be shared due to the GISAID agreement.

Table 5.1 – **Date of the most recent common ancestor of the clade corresponding to the French epidemic wave.** Unless specified otherwise, the year is 2020. The "model" indicates the value of the molecular clock and the population dynamics model used (DT or BDSKY).

model	size	most recent sample	median value	95% HPD
Fix8.8-DT	122a	24 Mar	31 Jan	[19 Jan - 9 Feb]
Fix8.8-BDSKY	122a	24 Mar	31 Jan	[20 Jan - 11 Feb]
Fix13.2-DT	122a	24 Mar	8 Feb	[30 Jan - 15 Feb]
Fix4.4-DT	122a	24 Mar	1 Jan	[11 Dec 2019 - 17 Jan]
Fix8.8-DT	81	17 Mar	2 Feb	[17 Jan - 11 Feb]
Fix8.8-DT	61-1	12 Mar	03 Feb	[21 Jan - 12 Feb]
Fix8.8-DT	61-2	24 Mar	08 Feb	[25 Jan - 17 Feb]

Table S1 – Prior summary for the exponential coalescent model

Parameter	**Value**
Molecular clock	fixed
Evolution model	GTR
kappa	LogNormal(1,1.25)
frequencies	Uniform(0,1)
popsize	1/x
growth rate	Gamma(0.001,1000)

Table S2 – Prior summary for the BDSKY model

Parameter	**Value**
Molecular clock	fixed
Evolution model	GTR
kappa	LogNormal(1,1.25)
frequencies	Uniform[0,1]
Rate of end of infection	Uniform(1.2, ∞) or LogNormal(0,1.2)
Sampling rate	Beta(1,1)
Reproduction number	LogNormal(0,1.2) with maximum at 10

Supplementary figures

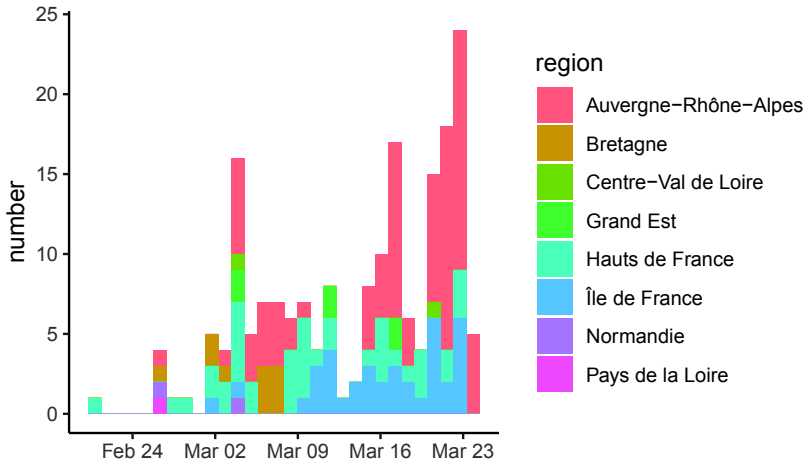


Figure S1 – **Sampling date and region.** List of samples collected, analysed and shared via GISAID by the two French National Reference Centers (CNR) as of Apr 4, 2020.

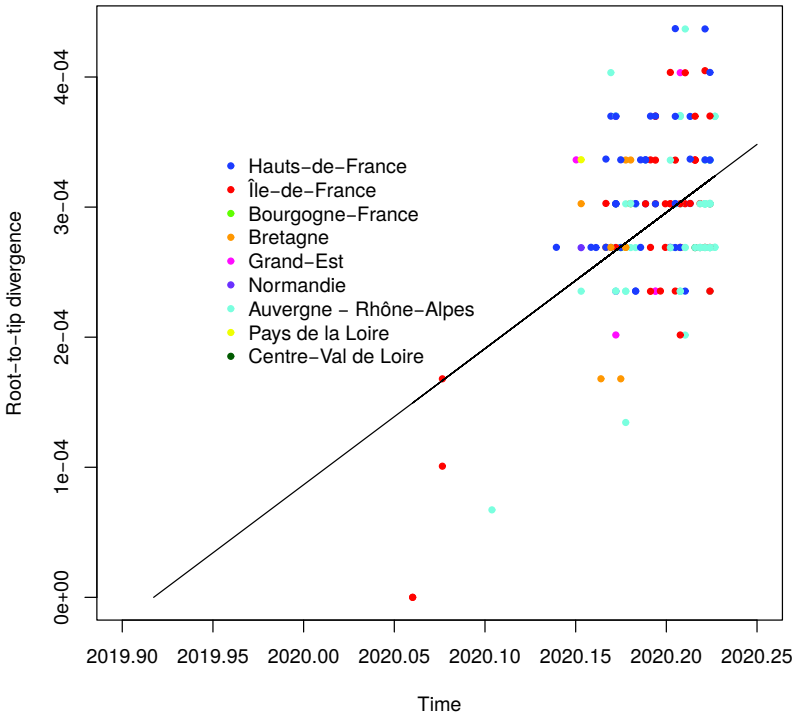


Figure S2 – **Root-to-tip correlation.** We analyse a phylogeny based on all 196 French sequences (i.e. not only that from the epidemic wave). The four earliest cases in Jan and early Feb were all isolated and belong to another clade than the rest of the sequences. The figure was obtained using TempEst (**?**).

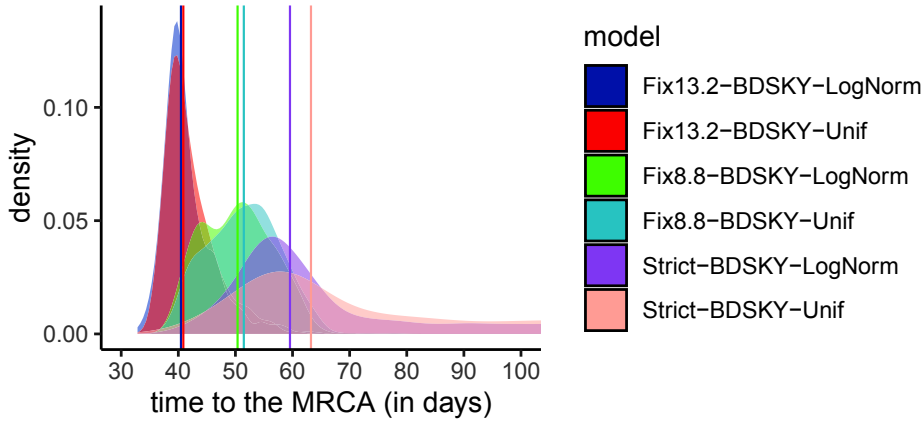


Figure S3 – **Time to the MRCA as a function of the molecular clock and of the recovery rate prior.** Here we assume BDSKY model. Note that the posterior distributions are not very informative when the molecular clock value is estimated (the 'Strict' model).

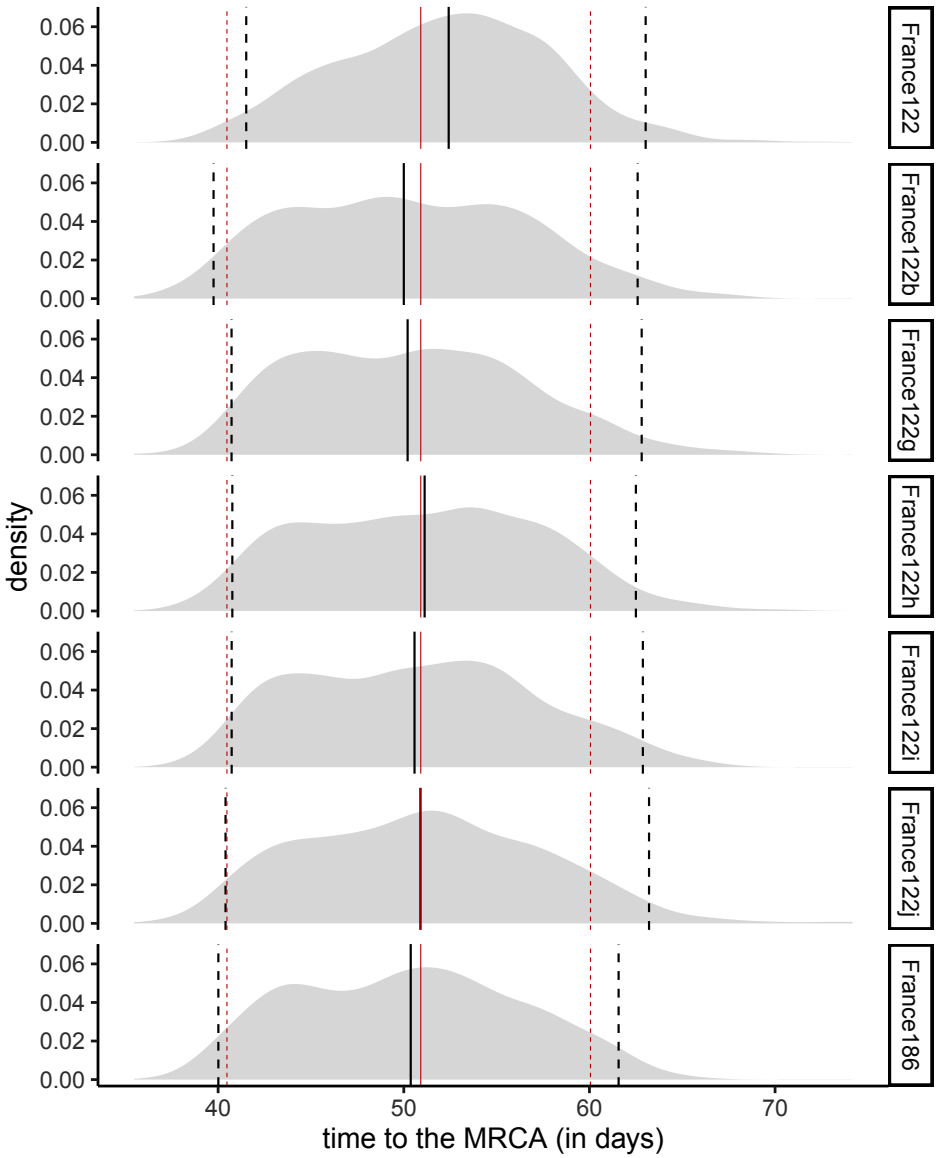


Figure S4 – **Time to the MRCA for the France186 dataset and subsets with 122 sampled sequences assuming the BDSKY model.** The red lines show the quantiles (0.025, 0.5, and 0.975) for the average of the 10 datasets. The black line shows the quantile for each dataset, and we can see how it behaves compared to the average. The last panel shows the largest phylogeny built without sampling, i.e. using all 186 sequences.

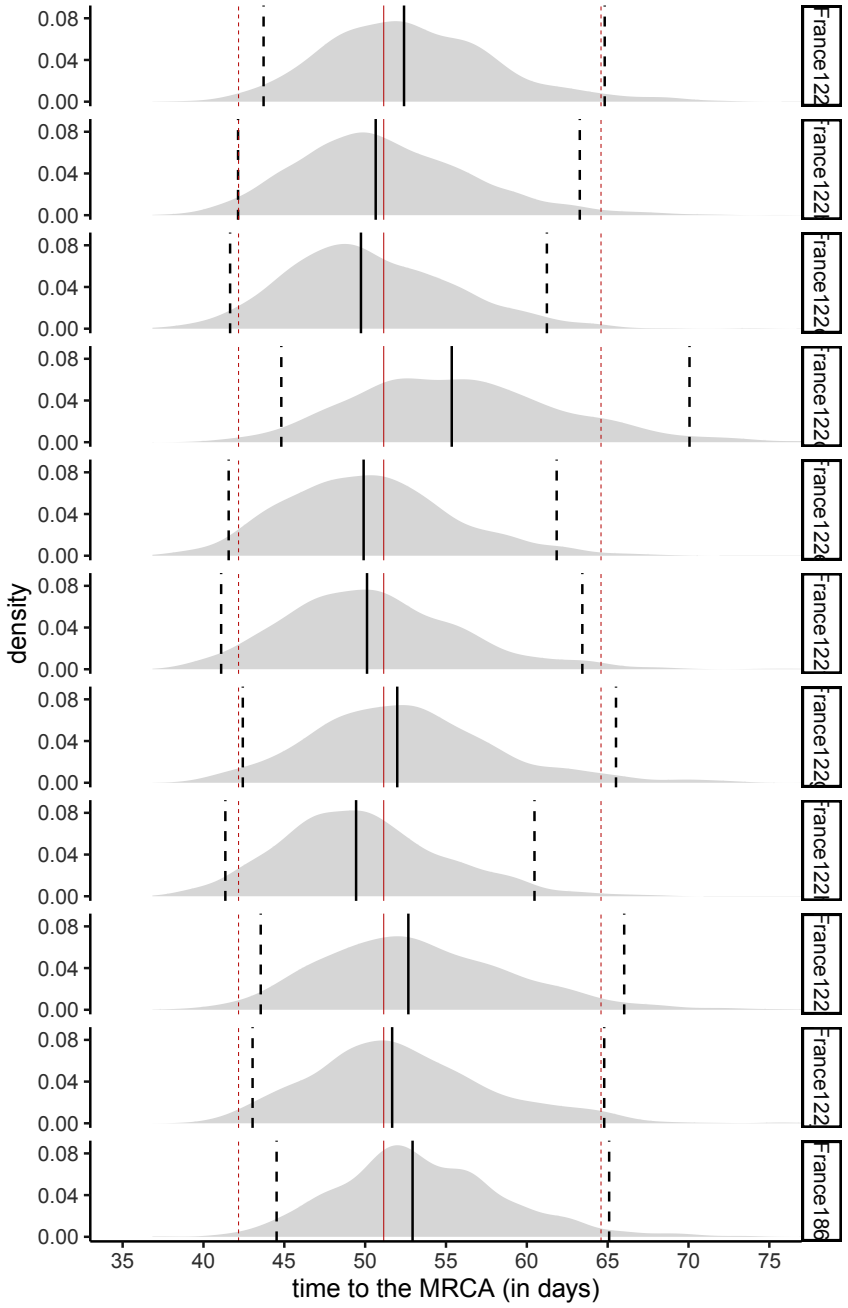


Figure S5 – **Time to the MRCA for the France186 dataset and subsets with 122 sampled sequences assuming an exponential growth coalescent model.** The red lines show the quantiles (0.025, 0.5, and 0.975) for the average of the 10 datasets. The black line shows the quantile for each dataset, and we can see how it behaves compared to the average. The last panel shows the largest phylogeny built without sampling, i.e. using all 186 sequences. The latter phylogeny has a slightly larger number of days to the MRCA.

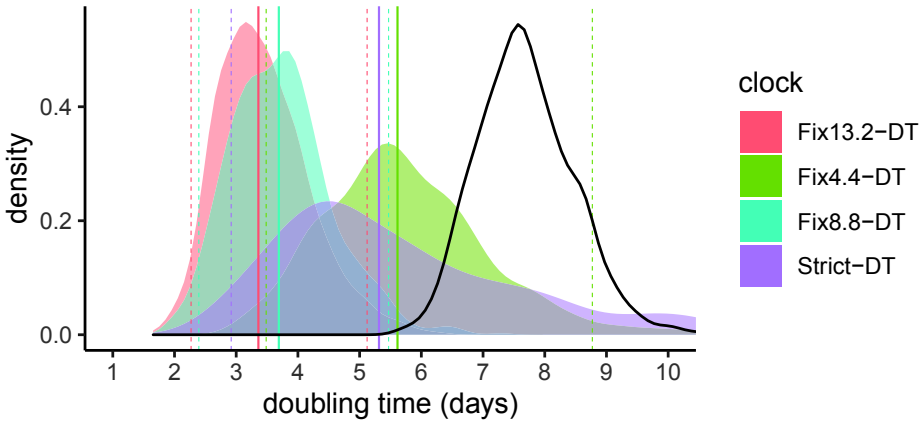


Figure S6 – **Effect of the molecular clock on the doubling time.** Note that in the "Strict" model we infer the value using a strict molecular clock but the width of the posterior distribution is large which indicates a lack of phylogenetic signal. The thick black line shows the prior distribution (the true prior values do not allow to see the regular plot so we use here Gamma(3,100) instead of Gamma(0.001,1000)). Here, an exponential growth coalescent model is assumed and the dataset used is France122a.

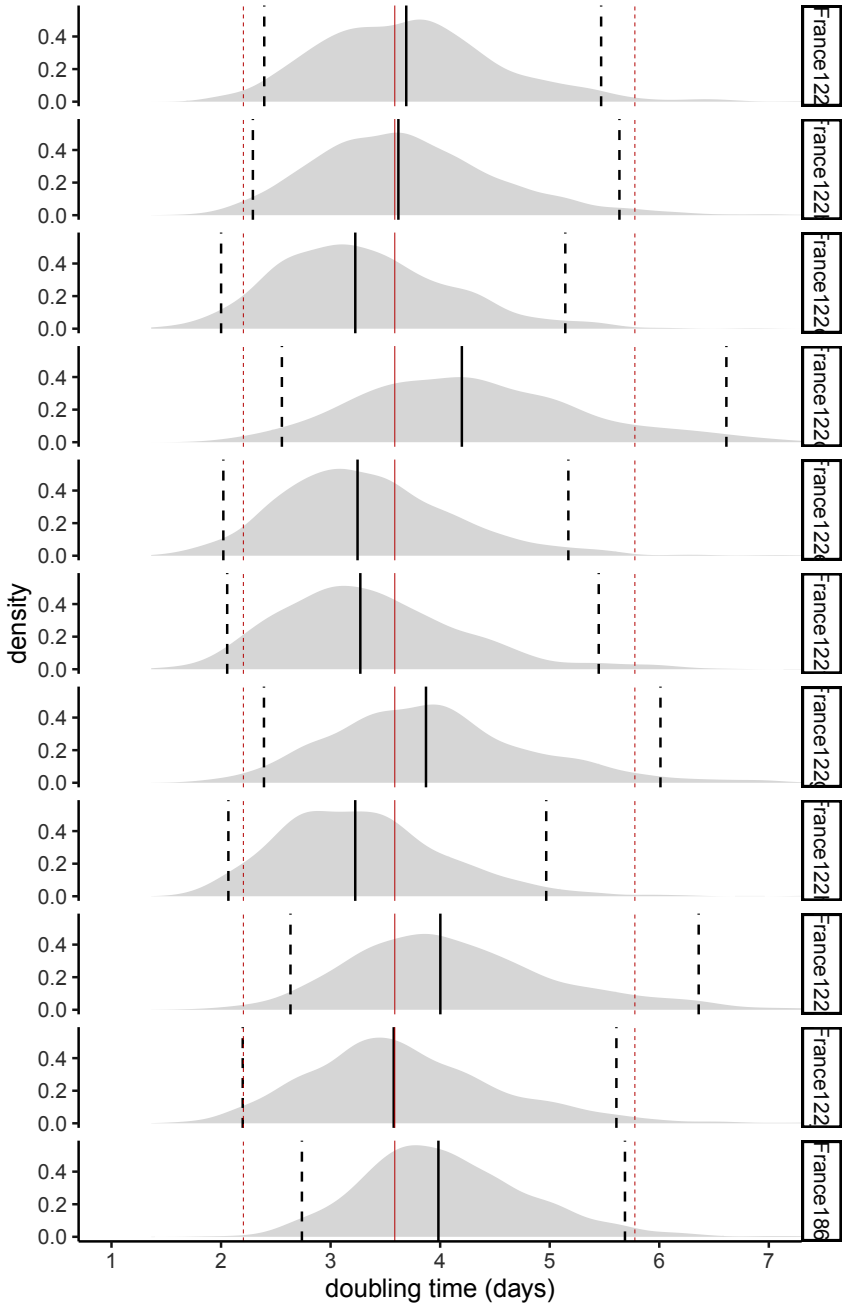


Figure S7 – **Doubling time for the 10 trees with 122 leaves sampled.** The red lines show the quantiles (0.025, 0.5 and 0.975) for the average of the 10 datasets. The black line shows the quantile for each dataset. The last panel shows the largest phylogeny with 186 leaves (which slightly overestimates the doubling time).

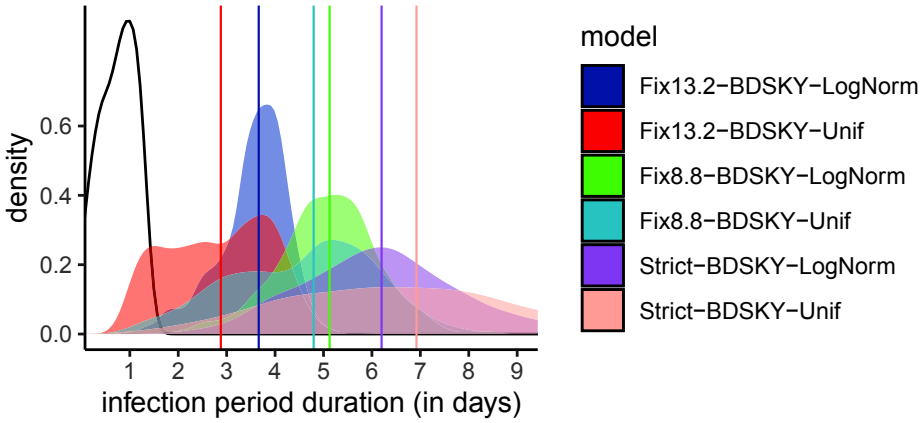


Figure S8 – **Effect of the prior shape and of the molecular clock on the effective infection duration estimate.** The molecular clock value has a stronger effect than the prior shape. The thick black line shows the prior distribution. Note that the posterior distributions are not very informative, i.e. the 95% HPD is very large, when the molecular clock value is estimated (the 'Strict' model).

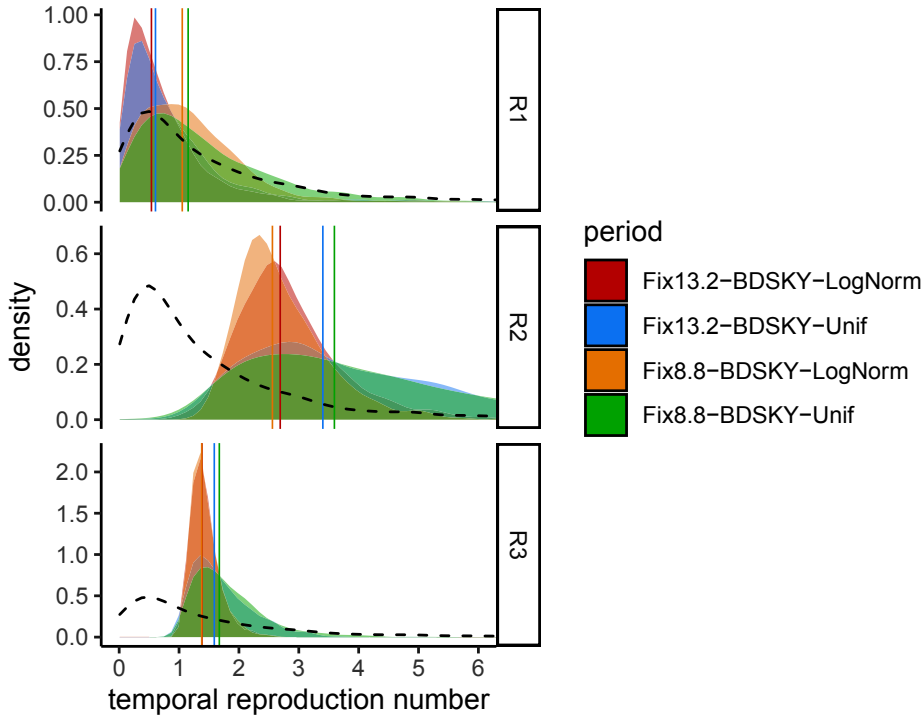


Figure S9 – **Effect of the prior shape and of the molecular clock on the temporal reproduction number estimate.** The molecular clock value has a stronger effect than the prior shape except for $\mathcal{R}_3$, where the effect is limited. For $\mathcal{R}_1$, posterior distributions are close to the prior (black dashed line), as discussed in the main text.

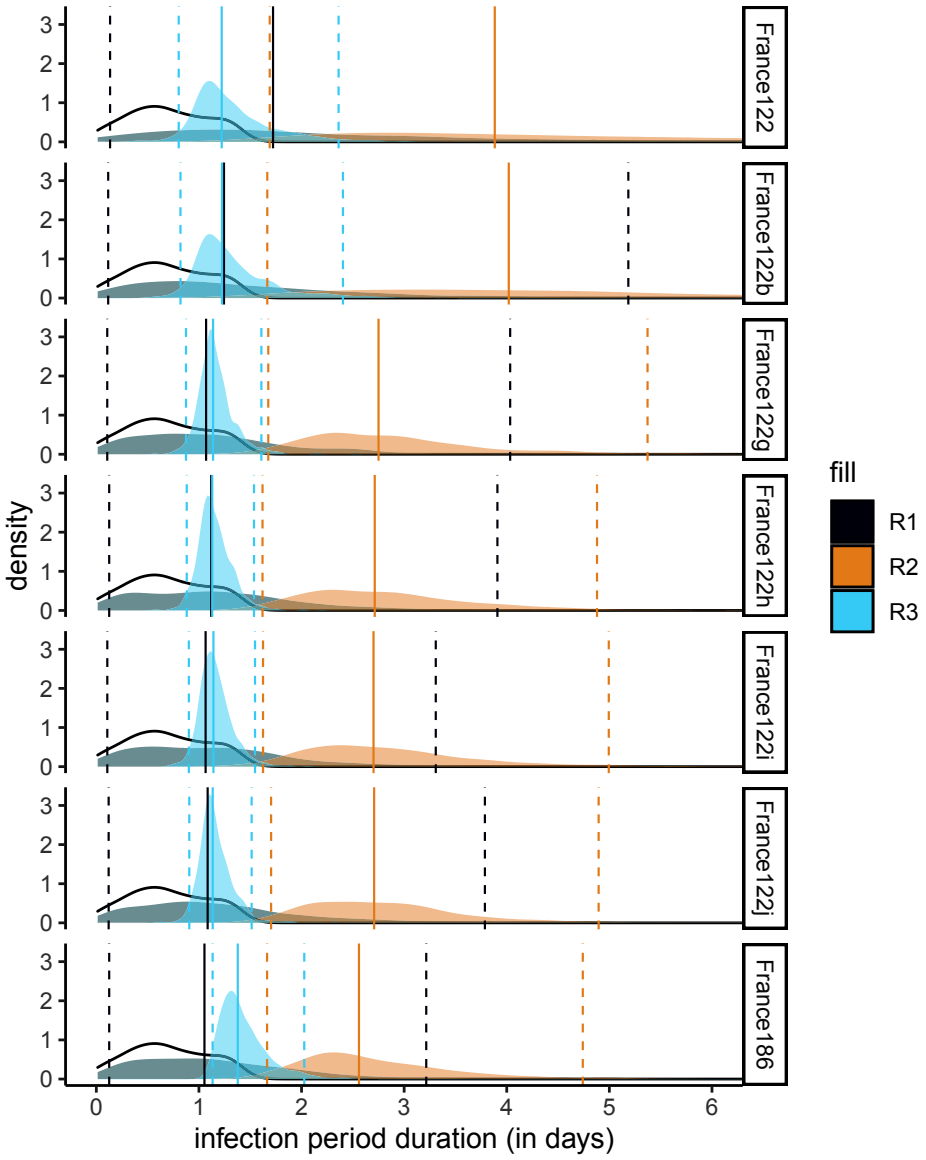


Figure S10 – **Reproduction numbers for the France186 dataset and subsets with 122 sequences.** Here we assume BDSKY model. The thick line shows the prior distribution. For $\mathcal{R}_1$, posterior distributions are close to the prior (black dashed line) indicated limited inference power.

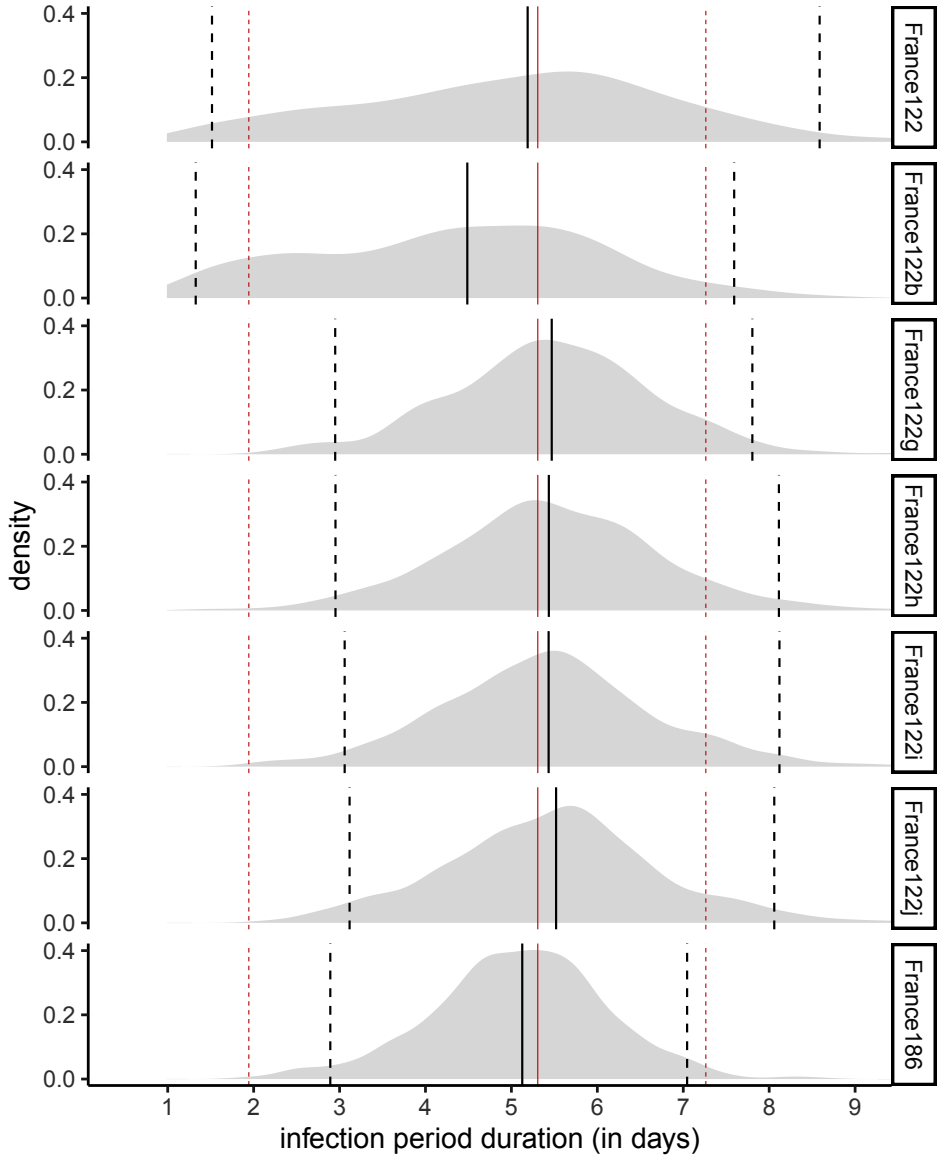


Figure S11 – **Effective infection duration for the France186 dataset and subsets with 122 sequences assuming a BDSKY model.** The median value obtained with the whole dataset (France186, black full line) is close to the average of the median values obtained with the subsets (red full lines).

Chapitre 6

Discussion

Comprendre la dynamique de propagation des infections microbiennes est un enjeu majeur pour contrôler les épidémies qu'elles causent. Ceci se fait généralement par l'analyse de données épidémiologiques telles que les données d'incidence ou de prévalence récoltées à différents niveaux (dépistages, hospitalisations, décès). Au cours des dernières décennies, les progrès des techniques de séquençage et des puissances de calcul ont permis le développement de méthodes d'analyse massive et quasiment en temps réel de données de séquences et surtout de phylogénies pour étudier la dynamique épidémiologique des infections microbiennes.

À travers cette thèse, nous avons développé un nouvel outil de simulations de séries temporelles et de phylogénies. Nous avons étendu l'application de la méthode de calcul bayésien approché (ou ABC pour *Approximate Bayesian Computation*) à un modèle structuré avec deux types d'hôtes. Pour cela, nous avons développé de nouvelles statistiques de résumé adaptées aux phylogénies dites « labellisées », c'est-à-dire dont les feuilles sont associées à un certain trait (groupe à risque, région géographique ou mutation). Nous avons utilisé cette approche ABC pour étudier la propagation du virus de l'hépatite C à Lyon, mais aussi pour étudier l'avantage de transmission que conférerait la présence d'une certaine mutation aux VIH-1/O. Enfin, nous avons réalisé des analyses épidémiologiques et phylodynamiques du SARS-CoV-2 à partir de données de séquences françaises en utilisant des méthodes d'inférence bayésienne basées sur la vraisemblance.

Je discute dans les prochaines sections des limites rencontrées au cours de cette thèse et des pistes de futurs travaux de recherche.

6.1 Comparaison de méthodes

Le concours de phylogénétique et de réseaux pour l'épidémie de VIH généralisée PANGEA-HIV (pour *Phylogenetics And Networks for Generalized HIV Epidemics in Africa*) (**?**) a été mis en place dans le but d'évaluer les méthodes phylodynamiques sur des jeux de données simulées avant que ces méthodes soient appliquées sur des milliers de données de séquences génomiques. Les équipes de recherche en phylodynamique participantes ont travaillé sur des données d'épidémies de VIH simulées grâce à différents modèles individus centrés, caractérisant un village ou une région en Afrique subsaharienne. Bien que clos, les méthodes et résultats de chacune des équipes de recherches ont été publiés et les données du concours sont toujours disponibles (**?**). La méthode obtenant les meilleurs résultats d'inférence est basée sur le coalescent structuré, avec un modèle déterministe à plusieurs compartiments. Cependant, cette méthode a nécessité des ressources computationnelles considérables.

? ont comparé l'approche par ABC-régression avec les approches phylodynamiques basées sur la vraisemblance de modèles BD et ont obtenu une précision comparable pour l'inférence du R_0 et la durée d'infection, précision augmentant avec la taille de la phylogénie. Les travaux de cette thèse ont permis d'étendre l'application de la méthode par ABC-régression à un modèle structuré avec plusieurs types d'hôtes en développant de nouvelles statistiques de résumé adaptées aux phylogénies dites « labellisées ».

Dans le cadre de son stage de licence au sein de l'équipe que j'ai co-encadré, Mahé Liabeuf a développé un modèle stratifié selon le sexe, la progression de l'infection par le VIH et le groupe à risque représenté par le nombre de contacts sexuels par unité de temps. Elle a notamment réalisé une étude bibliographique afin de paramétrer le modèle, en recherchant par exemple la taille de la population totale ou la proportion de femmes ou hommes en Afrique subsaharienne. Une des pistes pour la suite serait de simuler des phylogénies à partir de ce modèle et d'utiliser notre méthode ABC-régression pour analyser les données de PANGEA-HIV et mesurer l'erreur d'inférence.

6.2 Modification du *prior* après simulations

L'approche phylodynamique par ABC-régression nécessite de simuler de nombreux arbres, ce que permet notre outil TiPS. Pour cela, il utilise les trajectoires épidémiologique et les dates d'échantillonnage associées aux séquences microbiennes récoltées chez des individus infectés, qui correspondent aux feuilles de la phylogénie.

Un des défis du champ consiste à ajouter une information supplémentaire, ou « trait » à ces feuilles. Ce trait peut être une région géographique, un comportement, etc. Dans ce cas, pour chacune des simulations de phylogénies, chaque date d'échantillonnage doit être associée à un trait. Ceci peut se faire en suivant les données réelles ou de manière aléatoire, en suivant la proportion observée de chaque trait dans le jeu de données étudié. Nous avons vu dans le Chapitre 4 que la faisabilité des simulations, réduite par cette double contrainte imposée, modifie les distributions *a priori*. Il serait intéressant d'approfondir les biais possibles de ce phénomène sur l'inférence par ABC-régression.

Un autre point à soulever concerne la variabilité des simulations. En effet, il est possible de simuler plusieurs différentes trajectoires à partir d'une seule combinaison de valeurs de paramètres du modèle. De même, il est possible de simuler plusieurs phylogénies à partir d'une même trajectoire et des mêmes dates d'échantillonnage. Cependant, dans notre approche par ABC, nous ne gardons que la première phylogénie simulée obtenue par trajectoire simulée. Il serait intéressant de quantifier cette variabilité des simulations en utilisant, par exemple, les statistiques de résumé et par la suite chercher un moyen d'intégrer cette variabilité dans notre approche.

6.3 Biais d'échantillonnage

L'augmentation continue du nombre de données de séquences génétiques microbiennes et notamment virales a conduit à une situation dans laquelle il est délicat de réaliser une analyse phylodynamique qui inclurait toutes les séquences disponibles, et ce en un temps raisonnable. Le nombre de données génétiques du SARS-CoV-2 disponibles en est un exemple frappant et l'outil Nextstrain réalise maintenant systématiquement un sous-échantillonnage des données. De plus, il permet aussi aux utilisateurs de sous-échantillonner en suivant un nombre de séquences à garder par unité de temps et par région géographique.

Au-delà du problème du temps de calcul, le biais d'échantillonnage, par exemple lorsqu'une certaine population ou région est plus échantillonnée qu'une autre, peut introduire un biais dans l'inférence phylodynamique, quelle que soit la méthode utilisée. Il est donc nécessaire de concevoir de meilleures stratégies d'échantillonnage et/ou d'avoir connaissance des biais dans un jeu de données afin de pouvoir éventuellement les corriger. **?** ont étudié l'effet de différentes stratégies d'échantillonnage sur la qualité de la reconstruction de la dynamique des populations virales *via* des méthodes d'inférence bayésienne basées sur le modèle coalescent. Suite à leurs résultats, les auteurs recommandent l'utilisation d'un échantillonnage des séquences de

manière uniforme sur le temps et l'espace. Une autre solution serait de recalibrer ces données de séquences à partir des données de surveillance (**?**).

À notre connaissance, aucune étude n'a été réalisée afin d'estimer la sensibilité de l'ABC appliquée en phylodynamique au biais d'échantillonnage. Dans notre étude sur la propagation du virus de l'hépatite C au sein de deux populations différentes, nous avons montré qu'en sous-échantillonnant 50 % de données d'une population probablement sur-représentée les résultats étaient similaires à ceux obtenus en analysant le jeu de données complet. Ces résultats suggèrent que la phylodynamique en ABC-régression pourrait être peu sensible aux biais d'échantillonnage. Pour en être plus sûrs, une piste intéressante serait de simuler différentes phylogénies sous différents modèles de paramètres connus, avec un ou plusieurs types d'hôtes, et selon différents schémas d'échantillonnage. En inférant des paramètres tels que le R_0 ou la durée de contagiosité pour chacune des phylogénies avec notre approche ABC, on pourrait estimer l'erreur d'inférence des paramètres associée à la stratégie d'échantillonnage.

6.4 Signal phylogénétique

Nous l'avons vu, les analyses phylodynamiques de données génétiques de pathogènes peuvent être utilisées pour déduire la date et la région géographique d'origine d'une épidémie, le taux d'évolution virale, les dynamiques épidémiologiques et démographiques. Cependant, ces inférences reposent sur le fait que les données génomiques sont suffisamment informatives, et qu'elles contiennent donc suffisamment de signal phylogénétique.

Par exemple, dans notre étude phylodynamique des épidémies du virus de l'hépatite C, une analyse préliminaire des données de séquences a révélé un faible signal temporel. Nous avons donc fixé la valeur du taux de substitution. Cela peut être expliqué par la taille du jeu de données qui est relativement petite, mais aussi par le fait que seule une petite région du génome, correspondant au gène NS5B du virus, a été analysée. Un autre exemple concerne notre étude phylodynamique du SARS-CoV-2 en France, où la phylogénie inférée est mal résolue, c'est-à-dire qu'elle présente des multifurcations du fait du faible signal phylogénétique. Un faible signal peut entraîner un biais dans l'inférence, par exemple, de la date d'origine d'une épidémie virale. Nos résultats montrent qu'en utilisant des séquences récoltées sur une fenêtre d'échantillonnage courte, la date d'origine inférée est similaire à celle inférée à partir d'un plus grand nombre de données répartie sur une fenêtre temporelle plus large mais avec un taux d'évolution plus rapide.

Les résultats soulèvent l'importance d'une exploration du signal phylogénétique

avant toute analyse phylodynamique, l'importance de séquencer dès le début d'une épidémie mais aussi de séquencer le génome entier ou des régions du génome les plus informatives possibles (**?**).

6.5 Sélection du modèle

Les méthodes d'inférence reposent sur des modèles mathématiques. En phylodynamique, ces modèles peuvent être simplement démographiques, des modèles de naissance et de mort ou bien des modèles épidémiologiques détaillés. La formulation d'un modèle nécessite d'établir un ensemble d'hypothèses qui reposent sur les connaissances disponibles, par exemple la physiopathologie des infections (durée des phases aiguës et chroniques), ou encore l'hétérogénéité des populations hôtes, qu'elle soit spatiale ou comportementale avec des groupes à risque. Déterminer l'identifiabilité de paramètres d'un modèle, c'est-à-dire la capacité des paramètres à être identifiés ou non à partir des données, n'est pas facile, surtout lorsque le modèle est complexe avec un grand nombre de paramètres à identifier. Le choix du modèle est donc important car les inférences qui en résultent dépendent de son adéquation avec les données.

Parallèlement à l'étude phylodynamique du SARS-CoV-2 présentée dans la section 5.2, j'ai réalisé une analyse par ABC-régression sur le même jeu de 186 données génomiques (sans fixer la valeur du taux d'évolution pour l'inférence phylogénétique). J'ai travaillé avec trois différents modèles épidémiologiques : un modèle BD (pour *Birth Death*), un modèle EI (Exposés - Infectés) et un modèle EAI (Exposés - Asymptomatiques - Infectés). Les analyses par ABC-régression à partir de la même phylogénie ont généré, pour chacun de ces modèles, des estimations différentes de la durée de la période d'infection.

Un autre exemple est celui de l'analyse phylodynamique de l'hépatite C au Chapitre 3.2 où un modèle composé d'un compartiment où les infectés sont en phase aiguë et d'un compartiment pour la phase chronique s'est avéré moins adapté aux données que le modèle BD à deux types d'hôtes.

Dans les deux cas, une erreur a été de ne pas réaliser de comparaison de modèles au préalable, d'autant qu'une sélection de modèle peut être effectuée dans un cadre ABC. Le but, en général, est d'estimer la probabilité *a posteriori* d'un modèle sachant des données, ici représentées par des statistiques de résumé. Par exemple, avec la méthode de rejet, la probabilité *a posteriori* d'un modèle donné est approchée par la proportion de simulations acceptées pour ce modèle. Une autre méthode est basée sur la validation croisée et utilise les données simulées et leurs statistiques de résumé.

Le package R abc (**?**) propose des fonctions qui permettent de réaliser une étape de sélection de modèle *via* ces méthodes. Le temps de calcul de l'étape de sélection de modèle en utilisant cet outil est raisonnable, ce qui permet d'intégrer cette étape dans les prochaines études phylodynamiques par ABC-régression.

6.6 Perspectives ABC-régression

Les approches ABC ont l'avantage de pouvoir être appliquées à n'importe quel type de modèle et à n'importe quel type de données, à condition que l'on puisse générer à partir du modèle des données comparables aux données observées.

Les travaux de cette thèse reposent uniquement sur l'analyse de phylogénies virales. Une des extensions naturelles de ce travail serait d'intégrer de nouvelles données, notamment des données d'incidence des cas infectieux, de décès ou d'hospitalisations au cours du temps.

De plus, les études phylodynamiques reposent en général sur l'hypothèse que le pathogène évolue de manière neutre au sein de son hôte. Cependant, la plupart des pathogènes, notamment les virus, subissent une pression du système immunitaire de l'hôte qui peut déterminer la structure de la phylogénie (**?**). Une piste intéressante serait de développer et d'intégrer dans notre approche ABC-régression des modèles permettant de prendre en compte cette pression de sélection. Ceci pourrait aussi s'appuyer sur une approche partant directement des séquences génétiques au lieu des phylogénies (**?**). Un des avantage est que cela permettrait d'analyser l'évolution intra-hôte du virus qui n'est pas reflétée par l'alignement des séquences consensus sur lequel repose la construction phylogénétique.

6.7 Qu'est-ce-qu'un bon outil ?

Tout au long de cette thèse, j'ai dû installer, utiliser, mettre à jour ou chercher la documentation de nombreux outils. Le développement du simulateur TiPS n'a pas été tâche facile. De manière générale, le développement d'un outil nécessite de se poser de nombreuses questions telles que : *Quels langage et environnement de programmation choisir ? À quel public l'outil est-il destiné ? Comment l'utilisateur peut-il donner en entrée des informations sur le modèle ? Comment nommer les fonctions et méthodes ? Sous quelle forme sont présentés les résultats ? Quel nom donner à l'outil ?* Toutes ces questions sont liées à une principale question : *Qu'est-ce-qu'un bon outil ?*

Selon moi, un bon outil est bien documenté, facile à installer et à utiliser, produit des résultats cohérents et de bonne qualité, et s'exécute en un temps de calcul

raisonnable (en fonction bien sûr de la tâche à accomplir).

Un autre aspect important de tout outil est, selon moi, la manière dont l'utilisateur peut donner en entrée les informations sur son modèle (épidémiologique, démographique, phylodynamique ou autre). Par exemple, pour réaliser des simulations, le package phydynR requiert de donner en entrée une matrice des naissances, une matrice des morts et une matrice des migrations. Pour des modèles détaillés avec beaucoup de réactions, ceci peut rapidement être non seulement fastidieux mais surtout source d'erreur. À l'inverse, le package adaptivetau est bien plus intuitif et ne requiert qu'un vecteur des différents taux des réactions du modèle et un vecteur avec les flux de transition correspondant. Les packages des logiciels BEAST2 nécessitent quant à eux de paramétrer le modèle *via* un fichier au format XML, un langage de balisage, qui est, selon moi, leur talon d'Achille. Le XML est un langage utilisé pour décrire une structure, ici un modèle, dont la définition peut être lourde. La Figure 6.1(a) illustre cette complexité et présente la description d'un modèle SIR avec deux types d'hôtes en utilisant le langage XML, nécessaire pour l'utilisation de MASTER du logiciel BEAST pour simuler une trajectoire. Certes, le logiciel BEAUTI permet de formater le fichier d'entrée XML pour les logiciels BEAST et BEAST2, et ce pour une large gamme de modèles d'inférence phylogénétique, phylodynamique ou phylogéographique. Cependant, tout n'est par paramétrable à partir de BEAUTI et l'utilisateur doit alors soit chercher dans la documentation (tutoriels, forums ou *mailing list*) comment ajouter manuellement certains paramètres, soit, en désespoir de cause, contacter directement les auteurs pour se rendre compte que la seule méthode consiste à éditer le fichier XML manuellement. Pour TiPS, la manière de paramétrer le modèle a nécessité beaucoup de réflexion et d'avis, y compris de personnes externes au projet. Différents formalismes ont été proposés, de celui qui me paraissait intuitif mais ne l'était pas pour les autres, à un formalisme qui je l'espère est le plus intuitif et compréhensible pour tous (Figure 6.1(b)).

Un bon outil génère des résultats cohérents, d'où l'importance de réaliser des tests comparatifs pour mesurer l'exactitude des méthodes. Ces tests peuvent être réalisés à partir de données de référence comme cela a été le cas dans le cadre du concours PANGEA-HIV. Un autre exemple est l'analyse comparative en terme d'exactitude réalisée entre les trajectoires simulées à partir d'un modèle structuré par TiPS en utilisant l'algorithme approché du tau-leap, et celles simulées par phdynR, et ce en utilisant des pas de temps de valeurs différentes. Les résultats de cette analyse ont montré que phydynR est fortement sensible au pas de temps, nécessitant un pas de temps optimal, qui varie selon les valeurs de paramètres du modèle. Selon, le pas de temps, phydynR avait tendance à surestimer le nombre d'indvidus infectés et donc la dynamique de la population. Dans le cadre de l'ABC, les inférences des paramètres

dépendent de l'exactitude des données simulées et donc du simulateur. Les tests comparatifs sont donc à être considérés avant de publier un outil.

Un autre point important est le temps de calcul. Prenons l'exemple des méthodes d'inférence bayésienne phylodynamiques basées sur la vraisemblance. Dans celles que nous avons utilisées, le temps de calcul dépend du calcul de la fonction de vraisemblance et de la convergence jusqu'à une distribution *a posteriori* conduite par des chaînes MCMC. Le calcul de la fonction de vraisemblance devient difficile et long pour des modèles détaillés, composés de beaucoup de paramètres. La longueur de la chaîne MCMC est paramétrable mais doit être suffisamment large pour obtenir une bonne convergence. Ainsi, pour des modèles détaillés ou pour des grands jeux

```
<beast version='2.0' namespace='master:master.model:master.steppers:master.outputs'>
  <run spec='Ensemble'
       simulationTime='50'
       nTraj="5">

    <model spec='Model' id='model'>
      <populationType spec='PopulationType' id='S' typeName='S' dim="2"/>
      <populationType spec='PopulationType' id='I' typeName='I' dim="2"/>
      <population spec='Population' id='R' populationName='R'/>

      <reactionGroup spec='ReactionGroup' reactionGroupName="Infection">
        <reaction spec='Reaction' rate="0.001"> S[0] + I[0] -> 2I[0] </reaction>
        <reaction spec='Reaction' rate="0.001"> S[1] + I[1] -> 2I[1] </reaction>
      </reactionGroup>
      <reactionGroup spec='ReactionGroup' reactionGroupName="Recovery">
        <reaction  spec='Reaction' rate="0.2"> I[0] -> R </reaction>
        <reaction  spec='Reaction' rate="0.2"> I[1] -> R </reaction>
      </reactionGroup>
      <reactionGroup spec='ReactionGroup' reactionGroupName="Migration">
        <reaction spec='Reaction' rate="0.01"> S[0] -> S[1] </reaction>
        <reaction spec='Reaction' rate="0.01"> S[1] -> S[0] </reaction>
        <reaction spec='Reaction' rate="0.01"> I[0] -> I[1] </reaction>
        <reaction spec='Reaction' rate="0.01"> I[1] -> I[0] </reaction>
      </reactionGroup>
    </model>

    <initialState spec='InitState'>
      <populationSize spec='PopulationSize' size='400'>
        <population spec='Population' type='@S' location="0"/>
      </populationSize>
      <populationSize spec='PopulationSize' size='500'>
        <population spec='Population' type='@S' location="1"/>
      </populationSize>
      <populationSize spec='PopulationSize' size='100'>
        <population spec='Population' type='@I' location="0"/>
      </populationSize>
    </initialState>

    <output spec='JsonOutput' fileName='StructuredSIR_output.json'/>
  </run>
</beast>
```

(a)

```
library(TiPS)

reactions <- c("S1 [beta*S1*I1] -> I1",
               "S2 [beta*S2*I2] -> I2",
               "I1 [gamma*I1] -> R",
               "I2 [gamma*I2] -> R",
               "S1 [epsilon*S1] -> S2",
               "S2 [epsilon*S2] -> S1",
               "I1 [epsilon*I1] -> I2",
               "I2 [epsilon*I2] -> I1")

sir_simu <- build_simulator(reactions)

theta <- list(beta=0.001,gamma=0.2,epsilon=0.01)
x0 <- c(S1=400,S2=500,I1=100,I2=0,R=0)
traj <- sir_simu(
        paramValues = theta,
        initialStates = x0,
        times = c(0,50))
```

(b)

$$\frac{dS_1}{dt} = -\beta S_1 I_1 + \varepsilon S_2 - \varepsilon I_1 \quad \frac{dS_2}{dt} = -\beta S_2 I_2 + \varepsilon S_1 - \varepsilon I_2 \quad \frac{dI_1}{dt} = \beta S_1 I_1 - \gamma I_1 \quad \frac{dI_2}{dt} = \beta S_2 I_2 - \gamma I_2$$

(c)

FIGURE 6.1 – **Présentation des différentes manières de décrire un modèle SIR structuré avec deux types d'hôtes.** Le modèle peut être décrit a) dans un format XML utilisé par MASTER, b) dans un formalisme utilisé par TiPS, c) sous forme d'équations différentielles.

de données, le temps de calcul peut rapidement devenir limitant.

Dans notre approche par ABC-régression, la partie la plus consommatrice en terme de calculs est la simulation de données et les calculs de statistique de résumé de ces données. Nous l'avons vu, TiPS permet de simuler des jeux de données en grand nombre rapidement. De plus, les simulations et le calcul des statistique de résumé peuvent être réalisés en parallèle. Quant à l'algorithme de rejet et la régression, ce sont des étapes rapides. Finalement, l'étape la plus coûteuse en terme de temps de calcul est celle du test de VIF (pour *variance inflation factors*) qui mesure la multicolinéarité entre les statistiques de résumé. Cela suggère que l'approche par ABC-régression pourrait être une alternative aux approches basées sur la vraisemblance, surtout lorsque la taille des données est grande.

Bien qu'il faille encore explorer les limites de cette approche en phylodynamique, telles que la sensibilité au biais d'échantillonnage ou au signal phylogénétique, une piste intéressante serait de développer un outil d'inférence phylodynamique en utilisant une approche par ABC-régression. Chacune des étapes des analyses phylodynamiques réalisées au cours de cette thèse a nécessité l'utilisation d'outils que nous avons développé ou déjà existants. La première étape consiste à simuler des données, ici des trajectoires épidémiologiques et des phylogénies, ce qui est possible avec le package TiPS. L'étape de calcul de statistiques de résumé est réalisée *via* le package RPANDA et notre programme en C++, qui peut être facilement être recodé en Rcpp. Enfin, l'étape de rejet et de régression de l'ABC est réalisée en utilisant les fonctions des packages abc et glmnet. Nous pourrions donc développer une pipeline ou un package R qui reprendrait toutes ces étapes et permettrait à tout utilisateur de réaliser une étude phylodynamique à partir de phylogénies et/ou de données d'incidence, et par la suite possiblement à partir de séquences directement.