Reconstruction d'arbres de transmission de la tuberculose bovine dans un système multi-hôtes en France : intégration de données génomiques et épidémiologiques

Hélène Duault

université
PARIS-SACLAY

Reconstruction d'arbres de transmission de la tuberculose bovine dans un système multi-hôtes en France : intégration de données génomiques et épidémiologiques.

Transmission tree reconstruction for bovine tuberculosis in a multi-host system in France: combining genomic and epidemiological data.

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°570 : santé publique (EDSP)
Spécialité de doctorat : Epidémiologie
Graduate School : Santé publique. Référent : Faculté de médecine

Thèse préparée dans l'unité de recherche d'**Epidémiologie des zoonoses virales, bactériennes et parasitaires** (Laboratoire de Santé Animale, Anses), sous la direction de **Benoit DURAND**, Directeur de Recherche, le co-encadrement de **Laetitia CANINI**, Chargée de recherche

Thèse soutenue à Paris-Saclay, le 03 octobre 2023, par

Hélène DUAULT

Composition du Jury

Membres du jury avec voix délibérative

Lulla OPATOWSKI Professeure des Universités, Université Versailles St Quentin	Présidente
Samuel ALIZON Directeur de Recherche, Université PSL	Rapporteur & Examinateur
Pauline EZANNO Directrice de Recherche, INRAE	Rapporteur & Examinatrice
Julien THEZE Chargé de recherche, Université Clermont Auvergne	Examinateur

THESE DE DOCTORAT

NNT : 2023UPASR021

université PARIS-SACLAY | ÉCOLE DOCTORALE Santé Publique (EDSP)

Titre : Reconstruction d'arbres de transmission de la tuberculose bovine dans un système multi-hôtes en France : intégration de données génomiques et épidémiologiques.

Mots clés : tuberculose bovine, phylodynamie, reconstruction d'arbres de transmission, système multi-hôtes

Résumé : La plupart des agents pathogènes peuvent infecter plusieurs espèces. Dans un système multi-hôtes, bien comprendre le rôle de chaque espèce est essentiel afin de mettre en place des mesures de prévention et de contrôle. Cependant, l'étude de systèmes multi-hôtes impliquant la faune sauvage (comme la tuberculose bovine ou bTB en France) peut être compliquée par des biais d'observation. Notre objectif était d'investiguer l'impact de ces biais dans l'estimation de la contribution de chaque espèce dans un système multi-hôtes, à partir de deux approches combinant des données génomiques et épidémiologiques. L'une reconstruit l'histoire évolutive de sous-populations de pathogènes, l'autre l'histoire de transmission individuelle. Nous avons commencé par reconstruire les états ancestraux dans un système bovins-blaireaux de bTB qui nous a permis d'estimer de rares évènements de transitions inter-espèces (à partir du blaireau). Nous avons ensuite répertorié et décrit 22 méthodes permettant la reconstruction d'arbres de transmission. Enfin, nous avons évalué les performances de trois de ces méthodes sur des données simulées dans un système à trois espèces-hôtes, avec ou sans biais d'observation. Les performances des méthodes étaient impactées négativement par le taux d'évolution lent (caractéristique de *Mycobacterium bovis*) et la complexité du système épidémiologique mais peu par les biais d'observation. Ce travail de thèse souligne l'importance de tester et de développer des méthodes de reconstruction d'arbres de transmission adaptées à des systèmes multi-hôtes.

Title : Transmission tree reconstruction for bovine tuberculosis in a multi-host system in France: combining genomic and epidemiological data.

Keywords : multi-host system, phylodynamics, transmission tree reconstruction, bovine tubreculosis

Abstract : The majority of pathogens can infect multiple species. In a multi-host system, understanding the role played by each host-species is necessary to implement preventive and control measures. However, observation biases can complicate the study of multi-host systems implicating wildlife (as is the case for bovine tuberculosis or bTB in France). Our objective was to investigate the impact of these biases on the estimation of host-species contribution in a multi-host system, using two approaches combining genomic and epidemiological data. The first reconstructs the evolutionary history of pathogen subpopulations and the other, the individual transmission history. We started by reconstructing the ancestral states in a badger-cattle system, which allowed us to estimate rare inter-species transition events (from badgers). We then listed and described 22 transmission tree reconstruction methods. Finally, we evaluated the performances of three transmission tree reconstruction methods on data simulated in a system with three host-species, with and without observation biases. Method performance was negatively impacted by the low evolutionary rate (characteristic of *Mycobacterium bovis*) and the complexity of the epidemiological system but less by observation biases. This work underlines the importance of testing and developing transmission tree reconstruction methods adapted to multi-host systems.

REMERCIEMENTS

A Benoit Durand et Laetitia Canini, directeur et co-encadrante de ma thèse. Un grand merci pour votre soutien, votre disponibilité, la richesse de vos enseignements et de vos conseils.

Aux membres du jury de cette thèse, qui me font l'honneur d'évaluer mon travail de thèse. A Lulla Opatowski, merci d'avoir accepté de présider mon jury de thèse. A Samuel Alizon et Pauline Ezanno, merci d'avoir accepté d'être les rapporteurs de ma thèse. A Julien Thézé, merci d'avoir accepté d'en être l'examinateur.

A Maria-Laura Boschiroli et Lorraine Michelet, merci pour votre aide, c'était un plaisir de travailler et de voyager avec vous.

Aux membres de mon comité de thèse : Elisabeta Vergu, Anne Cori, Maria-Laura Boschiroli, Lorraine Michelet, Patrick Hoscheit et Gael Beaunée. Merci pour vos précieux conseils.

A l'Université Paris-Saclay, merci d'avoir financé ce travail de thèse.

A toute l'équipe d'EpiMIM: Maud Marsot, Gina Zanella, Julie Rivière, Barbara Dufour, Valentine Guétin-Poirier, Guillaume Crozet, Viviane Domarin, Alex Drouin, Clémence Nadal, Lenin Vinueza, Erika Ornelas Eusebio, Amaias Avalos et Amalia Rataud. Merci pour votre accueil et votre ambiance de travail chaleureuse.

A toute l'équipe d'UZB, merci pour votre accueil en stage et votre travail de manipulation indispensable.

To my ever-growing family, thank you for everything.

VALORISATION DES TRAVAUX

ARTICLES PUBLIÉS

- H. Duault, L. Michelet, M.-L. Boschiroli, B. Durand, et L. Canini, « A Bayesian evolutionary model towards understanding wildlife contribution to F4-family *Mycobacterium bovis* transmission in the South-West of France », Veterinary Research, vol. 53, no 1, p. 28, avr. 2022, doi: 10.1186/s13567-022-01044-x.
- H. Duault, B. Durand, et L. Canini, « Methods Combining Genomic and Epidemiological Data in the Reconstruction of Transmission Trees: A Systematic Review », Pathogens, vol. 11, no 2, p. 252, févr. 2022, doi: 10.3390/pathogens11020252.

COMMUNICATIONS ORALES

- **PhyloMAP#2** (réunion annuelle du réseau Phylodynamique des Maladies Animales et des Plantes) à Paris, le 19 novembre 2021. "Methods combining genomic and epidemiological data in the reconstruction of transmission trees: a systematic review", H. Duault, B. Durand, et L. Canini.
- **EPIDEMICS8**, en ligne, du 30 novembre 2021 au 3 décembre 2021. "Methods combining genomic and epidemiological data in the reconstruction of transmission trees: a systematic review", H. Duault, B. Durand, et L. Canini.
- **MODAH2** (MODelling in Animal Health conference), en ligne, le 2 juillet 2021. "Evolutionary model for *Mycobacterium bovis* spoligotype SB0821 in the South-West of France, which came first: badger or cattle infection?", H. Duault, L. Michelet, M.-L. Boschiroli, B. Durand, et L. Canini.
- **ISVEE** (International Symposium of Veterinary Epidemiology and Economics) à Halifax (Canada), du 7 au 12 août 2022. "Bovine tuberculosis outbreak reconstruction in a multi-host system: conclusions drawn from simulated data reflecting the inherent sampling bias", H. Duault, B. Durand, et L. Canini.
- **PhyloMAP#3** (réunion annuelle du réseau Phylodynamique des Maladies Animales et des Plantes) à Paris, du 5 au 7 octobre 2022. "Bovine tuberculosis outbreak reconstruction in a multi-host system: what about bias?", H. Duault, B. Durand, et L. Canini.

COMMUNICATIONS AFFICHEES

- **EPIDEMICS8**, en ligne, du 30 novembre 2021 au 3 décembre 2021. "A Bayesian evolutionary model towards understanding wildlife contribution to F4-family *Mycobacterium bovis* transmission in the South-West of France", H. Duault, L. Michelet, M.-L. Boschiroli, B. Durand, et L. Canini,
- ***M. bovis*** 2022 à Galway (Irlande), du 7 au 10 juin 2022. "Bovine tuberculosis outbreak reconstruction in a multi-host system: conclusions drawn from simulated data reflecting the inherent sampling bias", H. Duault, B. Durand, et L. Canini.
- **Journées Scientifiques et Doctorales de l'Anses** (JSDA), à Maisons-Alfort, du 18 et 19 octobre 2022. "Bovine tuberculosis outbreak reconstruction in a multi-host system: conclusions drawn from simulated data reflecting the inherent sampling bias", H. Duault, B. Durand, et L. Canini.

CONTENU

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES SIGLES ET ABBREVIATIONS

A : Adénine (base nucléotidique)

ADN : Acide DésoxyriboNucléique

APDI : Arrêté Préfectoral de Déclaration d'Infection

BadTrIP : Bayesian epidemiological Transmission Inference from Polymorphisms (méthode de reconstruction d'arbres de transmission)

BASTA : BAyesian STructured coalescent Approximation (méthode de reconstruction des états ancestraux)

BDNI : Base de Données Nationale d'Identification

BEAST(2) : Bayesian Evolutionary Analysis by Sampling Trees (logiciel)

BEAUti : Bayesian Evolutionary Analysis Utility (interface)

BF : Bayes Factor ou Facteur de Bayes

BORIS : Bayesian Outbreak Reconstruction Inference and Simulation (méthode de reconstruction d'arbres de transmission)

bTB : bovine TuBerculosis ou tuberculose bovine

C : Cytosine (base nucléotidique)

DR : Direct Repeat (locus dans la region conservée de *M. bovis*)

ESA : Epidémiosurveillance Santé Animale (plateforme)

ESS : Effective Sample Size ou taille effective de l'échantillon

F84 : modèle de substitution de Felstein

FA : Fièvre Aphteuse

G : Guanine (base nucléotidique)

GLM(M) : Generalized Linear (Mixed) Model ou modèle lineaire généralisé (mixte)

GTR : Generalized-Time-Reversible (modèle de substitution)

HKY : Hasegawa-Kishino-Yano (modèle de substitution)

HPD : High Posterior Density (intervalle de crédibilité estimé par inférence Bayésienne)

IC : Intervalle de Confiance

ICM : Institut du Cerveau et de la Moelle épinière

IDC : Intradermo-Tuberculination Comparée

IDS : Intradermo-Tuberculination Simple

JC : Jukes Cantor (modèle de substitution)

LNR : Laboratoire Nationale de Référence

Mascot : Marginal approximation of the structured coalescent (méthode de reconstruction des états ancestraux)

MCC : Maximum de Crédibilité de Clade

MCMC : Markov Chain Monte Carlo ou chaîne de Markov Monte Carlo

MRCA : Most Recent Common Ancestor ou ancêtre commun le plus récent

MRSA : Methicillin-Resistant *Straphylococcus aureus*

NPF : Famille Non Phylogénétique (définie dans l'axe 2)

NS : Nested Sampling (algorithme)

OIT : Officiellement Indemne de Tuberculose bovine

PA : Pyrénées-Atlantiques

pb : paire de bases

PCR : Polymerase Chain Reaction (test diagnostique)

PF : Familles Phylogénétiques (définie dans l'axe 2)

SCOTTI : Structured COalescent Transmission Tree Inference (méthode de

reconstruction d'arbres de transmission)

SeqPF : Famille Phylogénétique Séquentielle (définie dans l'axe 2)

SimPF : Famille Phylogénétique Simultanée (définie dans l'axe 2)

SLAFEEL : Statistical Learning Approach For Estimating Epidemiological Links (méthode de reconstruction d'arbres de transmission)

SNP : Single Nucleotide Polymorphism

T : Thymine (base nucléotidique)

TiTUS : Transmission Tree Uniform Sampler (méthode de reconstruction d'arbres de transmission)

UE : Union Européenne

VNTR : Variable Number of Tandem Repeats (technique de typage)

ZPR : Zone de Prophylaxie Renforcée

1 INTRODUCTION GENERALE

1.1 LES SYSTEMES MULTI-HOTES

1.1.1 Importance et définition des systèmes multi-hôtes

La majorité des agents pathogènes sont multi-hôtes, ce qui signifie qu'ils peuvent infecter plusieurs espèces. En 2001, un recensement de 1922 agents pathogènes (comprenant bactéries, champignons, helminthes, virus et protozoaires) a mis en évidence le fait qu'environ 63 % des agents pathogènes répertoriés pouvaient infecter des hôtes de plus d'une espèce (1,2). Ce pourcentage variait selon les espèces chez qui ces agents pathogènes étaient principalement répertoriés : 62 % des agents pathogènes des humains, 77 % de ceux des animaux de production et 90 % de ceux des carnivores domestiques (1). Certains de ces agents pathogènes multi-hôtes sont transmissibles de l'animal à l'homme (zoonotiques) et présentent donc un intérêt en santé publique. En effet, bien que rares chez l'homme, certaines maladies zoonotiques endémiques dans les populations animales européennes font l'objet d'épidémies sporadiques avec des conséquences importantes en santé humaine (la brucellose en Grèce, la tularémie en Suède, Finlande et Norvège et la fièvre Q aux Pays-Bas) (3).

Lorsque des espèces menacées sont concernées, les agents pathogènes multi-hôtes peuvent également avoir un impact sur la biodiversité. Ainsi, en Tanzanie, une épidémie du *Morbillovirus* responsable de la maladie de Carré a provoqué la mort d'environ 30 % des lions (*Panthera leo*) dans le Parc National de Serengeti dans les années 90 (4). Enfin, ils peuvent également avoir des conséquences économiques lorsqu'ils affectent les animaux de production. Nous pouvons prendre l'exemple de la tuberculose bovine ou encore de l'influenza aviaire, toutes deux des maladies qui affectent à la fois des animaux domestiques, de la faune sauvage et potentiellement l'homme (5). La gamme d'espèces-hôtes pouvant être infectées par des agents pathogènes multi-hôtes ainsi que leurs conséquences sur les espèces-hôtes infectées conditionnent l'intérêt que nous pouvons leur porter.

Un *système multi-hôtes* correspond à l'ensemble des espèces-hôtes pouvant être infectées par un même agent pathogène dans une zone donnée ainsi qu'à leur dynamique de transmission dans cette zone. Dans un système multi-hôtes, nous nous intéressons souvent à une espèce-hôte ou une population en particulier, appelée *population cible*, pour laquelle l'objectif est

d'identifier le *réservoir* et de comprendre son fonctionnement afin de proposer des mesures de contrôle adaptées. Le réservoir se définit comme une ou plusieurs populations dans lesquelles l'agent pathogène peut persister (*populations de maintenance*) et à partir desquelles la population cible peut s'infecter (*populations sources*) (6).

1.1.2 Enjeux et difficultés rencontrées dans un système multi-hôtes impliquant la faune sauvage

Plus les populations sources peuvent être identifiées de manière précise et plus il a été recueilli de données quantitatives par rapport à leur contribution à la transmission, plus les mesures de contrôle pourront être ciblées de manière efficace (6). En effet, dans un système multi-hôtes où le réservoir est bien identifié, trois stratégies de contrôle s'offrent à nous : a) mettre en place des mesures uniquement dans la population cible, b) bloquer les transmissions entre la population cible et le réservoir et enfin, c) contrôler l'infection au sein du réservoir (6). La stratégie a) ne considère pas le réservoir, par exemple lors de la vaccination de la population humaine contre des maladies vectorielles comme la fièvre jaune (7), l'objectif est de protéger la population cible. Cependant, avant le développement de ce vaccin au 20ème siècle, en l'absence de connaissances sur les modes de transmission, des quarantaines étaient mises en place à bord de navires touchés (8), cela correspond également à la stratégie a). De plus, une fois la transmission par les moustiques confirmée en 1901, le recours à l'utilisation de larvicides afin de limiter le nombre de vecteurs (7), et donc une stratégie c), pouvait être implémentée. La stratégie b) dans cet exemple correspondrait à toute mesure visant à prévenir les piqûres de moustiques, *ex.* la mise en place de moustiquaires.

Dans le contexte plus général d'un système multi-hôtes, estimer les taux de transmission inter- et intra-espèces permet d'écarter les stratégies de lutte qui pourraient s'avérer inefficaces. Prenons un exemple historique concret, celui de la transmission de la rage dans un système multi-hôtes composé de chacals à flanc rayés (*Canis adustus*) et de chiens domestiques au Zimbabwe. Le taux de transmission des chiens (réservoir) vers les chacals (population cible) était très élevé, et, du fait d'une faible densité de population, le taux de transmission entre chacals était faible. L'épidémie chez les chacals était maintenue grâce à des réintroductions fréquentes de l'agent pathogène par le réservoir (9). Afin de diminuer la prévalence de l'infection dans la population cible, limiter les interactions chiens-chacals, stratégie b), ou vacciner les chiens, stratégie c), serait plus efficient. En effet, une stratégie a) centrée autour de la population cible uniquement, comme la vaccination

orale des chacals, n'aurait que peu d'impact (10).

Cependant, quand un système multi-hôtes implique une espèce-hôte de la faune sauvage, le choix de l'une de ces trois stratégies de lutte n'est pas toujours évident. L'implémentation et l'adaptation des mesures de contrôle sont alors compliquées par l'impossibilité de restreindre les mouvements de la totalité de la population sauvage, la difficulté d'empêcher leurs interactions avec d'autres espèces-hôtes et l'absence d'estimation exacte de la taille de cette population (11). De plus, avant ou de manière conjointe à l'implémentation de mesures de contrôle (12), le manque d'accessibilité des espèces sauvages complique la collecte de données et donc la surveillance de la maladie dans la population sauvage. Au sein d'un même système multi-hôtes, nous pouvons ainsi être confrontés à un contraste entre le taux d'échantillonnage des individus infectés dans la faune domestique et celui de la faune sauvage, qui demeure inconnu du fait d'une taille inconnue de la population. S'il n'est pas pris en compte, ce biais d'échantillonnage pourrait compromettre les efforts de quantification de la répartition des individus infectés par espèce et ainsi conduire à une mauvaise estimation de la contribution de chaque espèce-hôte à la transmission de l'agent pathogène en cause, puis à la mise en place de mesures de contrôle inadaptées.

1.1.3 Données nécessaires pour l'étude d'un système multi-hôtes

Dans leur étude du système multi-hôtes de rage au Zimbabwe (9), Rhodes *et al.* ont utilisé, en plus de données écologiques générales à propos des deux populations, le nombre de cas incidents par an rapporté chez les chacals et chez les chiens dans la région. D'autres données épidémiologiques sur l'histoire naturelle de la maladie, telles que le temps d'incubation (intervalle entre l'infection et le début des symptômes) ou encore la durée de la phase symptomatique, ont été obtenues grâce à des études expérimentales plus anciennes. A partir de cet exemple, deux types de sources se distinguent : les données issues de la surveillance (incidence), propres au système multi-hôtes étudié, et celles issues d'études expérimentales, plus générales.

Les systèmes de surveillance ont pour objectif d'identifier et de décrire les hôtes infectés. L'identification d'hôtes infectés dans un système multi-hôtes peut faire l'objet de campagnes de dépistage spécifiques dans une population d'intérêt (*ex.* bovins dans le cas de la tuberculose bovine (13)) ou de détection d'agents pathogènes dans des vecteurs (*ex.* l'encéphalite japonaise (14)) (surveillance active). L'agent pathogène peut également être isolé sur des hôtes symptomatiques voire des carcasses animales

(surveillance passive) (*ex.* sangliers dans le cas de la peste porcine africaine (15)). La date de découverte d'infection ou de mort, la présence ou non de lésions ou symptômes, la localisation géographique et l'espèce-hôte sont des données épidémiologiques permettant de décrire les hôtes infectés détectés. Cependant, ces données épidémiologiques peuvent manquer de précision sur l'histoire de transmission, notamment les données géographiques dans le cas d'une maladie évoluant lentement ou d'espèces-hôtes très mobiles. Afin de tirer des conclusions plus précises sur les dynamiques de transmission, il est possible de compléter ces données épidémiologiques par des données génétiques obtenues après isolement puis séquençage des souches (16). Ces données génétiques permettent la reconstruction d'arbres phylogénétiques décrivant l'histoire évolutive des séquences isolées. Le lien entre données de séquences génétiques (ou arbres phylogénétiques) et dynamique des populations s'appelle la phylodynamique (17).

1.2 LA RECONSTRUCTION D'ARBRES PHYLOGENETIQUES DANS UN SYSTEME MULTI-HOTES

1.2.1 Définition d'un arbre phylogénétique

Un arbre phylogénétique est une représentation des liens de parenté qui existent entre des souches échantillonnées. Ces souches échantillonnées sont placées sur des feuilles (nœuds terminaux) alors que les nœuds internes représentent des ancêtres communs hypothétiques. Ces nœuds (internes et terminaux) sont reliés entre eux par des branches dont la longueur représente la distance qui les sépare. Nous allons prendre l'exemple de souches pour lesquelles nous avons les séquences génétiques, cette distance correspond alors à une distance génétique (*Figure 1*).

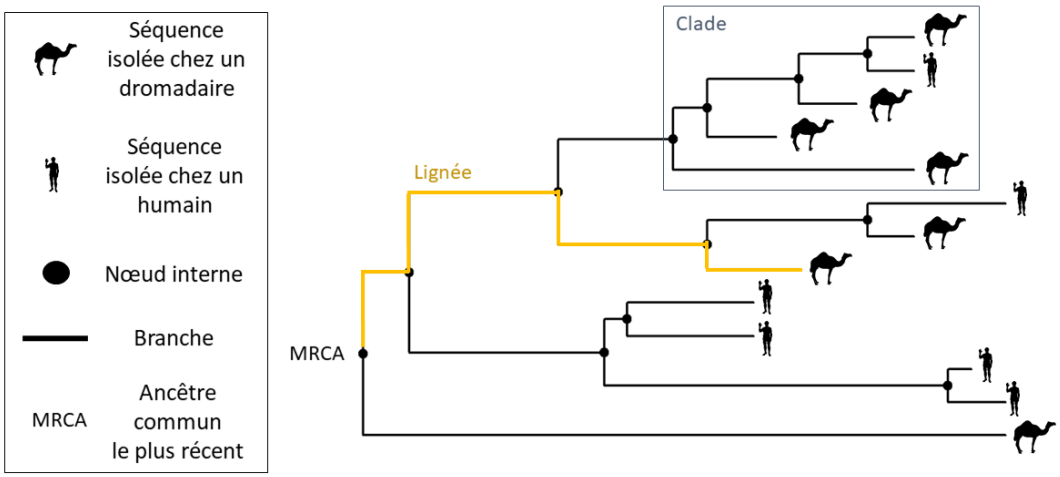


Figure 1 : Arbre phylogénétique reconstruit à partir de séquences de MERS-CoV isolées chez sept dromadaires et six humains (simplification de l'arbre obtenu par Dudas et al. (18)*).* Le MRCA désigne l'ancêtre le plus récent commun à toutes les souches échantillonnées.

Le chemin reliant un ancêtre hypothétique à un nœud terminal s'appelle une lignée (en jaune dans la *Figure 1*). Les descendants d'un même nœud interne correspondent à un clade (le rectangle gris dans la *Figure 1*). Enfin, certaines méthodes de reconstruction permettent l'identification d'un ancêtre commun à toutes les souches (MRCA pour « most recent common ancestor ») et transforment ainsi un arbre non enraciné en un arbre enraciné, le MRCA devient alors la racine de l'arbre. Pour un arbre non enraciné donné, plusieurs arbres enracinés existent, caractérisés par l'emplacement de la racine.

Pour la suite, nous allons considérer uniquement l'utilisation de séquences hétérochrones, c'est-à-dire échantillonnées à des temps différents, dans la reconstruction d'un arbre phylogénétique enraciné.

1.2.2 Reconstruction d'un arbre phylogénétique

1.2.2.1 Modèle de substitution

L'alphabet des séquences nucléotidiques est composé de quatre lettres ou nucléotides {A, C, G, T}, il existe donc trois substitutions possibles pour un site nucléotidique donné. La distance génétique la plus simple à estimer entre deux séquences alignées est le nombre de sites pour lesquels elles ne possèdent pas le même nucléotide. Cette distance correspond à la distance de Hamming qui, divisée par la taille de la séquence considérée,

donne la proportion de sites différant entre les deux séquences (19). Cependant, cette distance ne prend pas en compte la possibilité de mutations reverses ou intermédiaires (*ex*. dans A→G→A et A→G→T, le passage par G n'est pas identifié) et sous-estime ainsi la distance génétique réelle. Il est possible de corriger l'estimation des distances génétiques grâce à des modèles de substitution.

Un modèle de substitution décrit la probabilité qu'un nucléotide se substitue en un autre. Ce modèle précise la fréquence relative de chaque nucléotide et peut prendre en compte les nucléotides considérés ou plus simplement, la nature de leur base : purines (A, G) et pyrimidines (C, T). Le modèle le plus complexe et plus réaliste est le modèle Generalized-Time-Reversible (GTR) (20), il considère que chaque nucléotide a sa propre fréquence et que la probabilité de substitution dépend de la combinaison de nucléotides considérée. Cependant, ce modèle comporte de nombreux paramètres à estimer et n'est donc pas toujours adapté aux données génétiques disponibles. Ainsi, dans le modèle Hasegawa-Kishino-Yano (HKY) (21), on suppose que les nucléotides ont tous la même fréquence et la probabilité de substitution dépend de la nature des bases concernées. Deux types de substitution, définis pour la première fois par Ernst Freese en 1959, sont considérés. L'un correspond à la substitution entre deux nucléotides de même nature (appelé transition), l'autre entre deux nucléotides de natures différentes (appelé transversion) (22) (***Figure 2***). Le modèle HKY définit ainsi un paramètre kappa (κ) qui correspond au ratio entre le taux de transition et le taux de transversion. Dans le modèle le plus simple, celui de Jukes Cantor (23), tous les nucléotides ont la même fréquence et la probabilité pour qu'un nucléotide se substitue en un autre est la même, quelle que soit la combinaison de nucléotides considérée. La fréquence de ces évènements de substitution au cours du temps est définie par le taux de substitution, généralement exprimé en nombre de substitutions par site (ou génome) et par unité de temps.

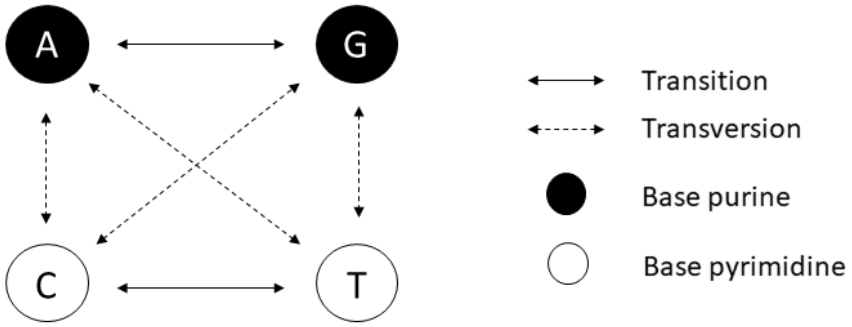


Figure 2 : Représentation schématique des substitutions à considérer dans un modèle de substitution HKY. Les 4 nucléotides possibles {A, C, G, T} sont représentés par des cercles dont la couleur représente la nature de la base.

1.2.2.2 Modèle d'horloge

Le modèle d'horloge décrit la manière dont évolue le taux de substitution au cours du temps. En 1965, en comparant des séquences polypeptidiques provenant de différents mammifères, Zuckerkandl et Pauling ont émis l'hypothèse que le taux d'évolution à l'échelle macromoléculaire est constant au cours du temps et entre lignées (24). Autrement dit, la distance génétique entre deux espèces, déterminée par le nombre de mutations accumulées dans leurs séquences polypeptidiques (ou génétiques), est proportionnelle au temps qui s'est écoulé depuis leur dernier ancêtre commun (25). Cependant, cette hypothèse d'horloge moléculaire est contestée par plusieurs comparaisons de séquences d'ADN concluant à des différences notables de taux de substitution entre lignées de différentes espèces animales (26). Par la suite, les modèles permettant le relâchement de ce modèle d'horloge stricte dans les études phylogénétiques ont été développés à la fin des années 90 et début des années 2000, lors de la popularisation des méthodes Bayésiennes (25). Dans un modèle d'horloge relâchée, le taux de substitution peut varier entre les lignées (modèle autocorrélé) ou entres les branches (modèle non corrélé) (27). Ainsi, dans un modèle d'horloge relâchée non corrélé lognormal (ou exponentiel), le taux de substitution d'une branche est tiré au sort, indépendamment des autres branches, à partir d'une distribution lognormale (ou exponentielle).

Aujourd'hui, il est généralement admis que, bien que cette horloge moléculaire ne soit pas universellement applicable, elle demeure une bonne

approximation de la réalité lors d'études d'espèces apparentées ou de données de population (25).

1.2.2.3 Modèle de population

Dans un arbre phylogénétique reconstruit à partir de séquences génomiques virales ou bactériennes, le modèle de population désigne celui de la population d'agents pathogènes, et non celui des hôtes chez qui les agents pathogènes ont été isolés.

Un des modèles les plus utilisés en phylogénétique est le modèle coalescent (28). Ce modèle décrit les liens de parenté entre *n* échantillons d'une large population évoluant selon le modèle de Wright-Fisher (*Figure 3*). Dans le modèle de Wright-Fisher, la taille de la population est considérée comme constante, aucun évènement de migration n'a lieu et chaque individu descend d'un seul parent (population d'haploïdes). Les générations ne se chevauchent pas et aucune sélection n'est exercée sur la reproduction. En remontant le temps dans une telle population à partir des *n* individus échantillonnés, tout se passe comme si chaque individu choisissait un parent au hasard dans la génération précédente sans tenir compte du choix des autres (28). Quand un individu parent est l'ancêtre de plus d'un individu échantillonné, il coïncide avec un évènement de coalescence. Si, dans ce modèle de Wright-Fisher, nous nous concentrons exclusivement sur les évènements de coalescence, on obtient le modèle coalescent. Quand on applique un modèle coalescent à une population réelle, la taille effective de population estimée correspond à la taille d'une population idéale suivant un modèle de Wright-Fisher pour laquelle on observerait la même variance génétique. L'inverse de cette taille effective de population correspond à la vitesse de coalescence.

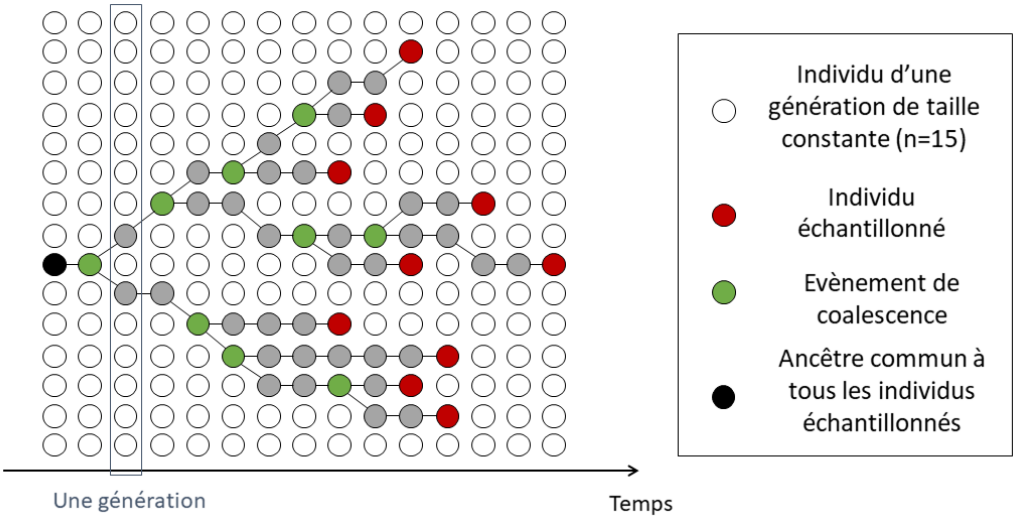


Figure 3 : Représentation schématique du modèle de Wright-Fisher ainsi que l'application du modèle coalescent à un échantillon de 10 individus. La taille de la population est constante (n=15) et les générations ne se chevauchent pas. Chaque individu n'a qu'un seul parent (population haploïde) et quand deux individus ont le même parent, cela correspond à un évènement de coalescence.

Par la suite, l'hypothèse de taille constante de la population dans le modèle coalescent a été relâchée. Il est ainsi possible de reconstruire des arbres phylogénétiques en considérant que la population suit un modèle exponentiel (29). Le modèle non-paramétrique skyline permet quant à lui de considérer une évolution démographique plus complexe ; le temps est alors découpé en intervalles pendant lesquels la taille de la population est constante (30).

Grâce au modèle coalescent, on reconstruit une généalogie en remontant le temps puis on y surimpose les mutations qui ont pu avoir lieu et on modélise ainsi l'obtention de séquences génétiques échantillonnées à partir d'individus clonaux soumis aux effets combinés d'une reproduction aléatoire (pas de sélection) et de mutations neutres (31). Cependant, le modèle coalescent n'est pas le seul modèle de population qui existe, et un autre modèle très utilisé est le modèle « Birth-Death ». Dans ce modèle, une seule lignée est présente au temps zéro, puis les lignées apparaissent (« birth » ou naissance) ou s'éteignent (« death » ou mort) dans les arbres phylogénétiques à des taux constants (32) ou évoluant au cours du temps (33), selon le modèle considéré.

1.2.3 Reconstruction des états ancestraux dans un arbre phylogénétique

Dans la *Figure 1*, le fait qu'il n'y ait pas un regroupement des séquences dans l'arbre selon leur espèce-hôte suggère l'existence d'un ou plusieurs évènements de transmission entre dromadaires et humains. De la même façon, après identification de la présence de multiples espèces-hôtes dans un même clade de l'arbre phylogénétique reconstruit à partir de séquences du virus responsable de la maladie d'Aujeszky, He *et al.* ont conclu à l'existence d'évènements de transmission inter-espèces (34). Cependant, il est possible d'aller plus loin dans l'étude de la dynamique de transmission entre espèces, à partir de l'arbre phylogénétique reconstruit. En effet, en considérant une caractéristique de l'agent pathogène (ici son espèce-hôte), des modèles dit « de trait » permettent de reconstruire l'évolution de cette caractéristique au cours du temps (*Figure 4*). Le changement d'état de cette caractéristique correspond ici à un saut inter-espèces (représenté par une étoile jaune dans la *Figure 4*).

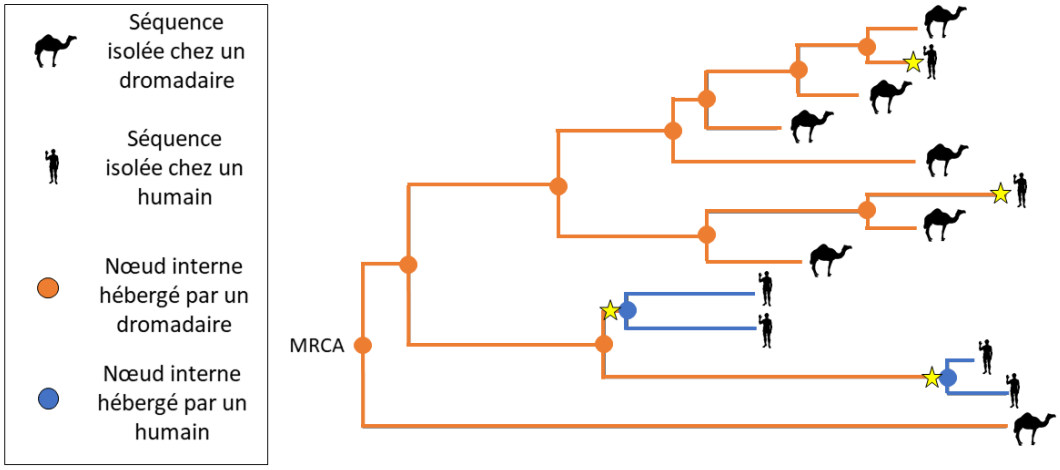


Figure 4 : Arbre phylogénétique dont l'évolution de l'espèce-hôte a été reconstruite au cours du temps (simplification de l'arbre obtenu par Dudas et al. (18)*).* L'arbre a été construit à partir de séquences de MERS-CoV isolées chez sept dromadaires et six humains. Le MRCA désigne l'ancêtre le plus récent commun à toutes les souches échantillonnées. Ici quatre transitions (représentées par des étoiles jaunes) de dromadaire à humain ont été prédites.

La méthode par maximum de parcimonie permet d'assigner des états aux nœuds internes d'un arbre phylogénétique déjà reconstruit. Cette méthode considère que chaque changement d'état ayant lieu dans l'arbre a un coût et que la configuration optimale des états ancestraux minimise la

somme de ces coûts (35). Un des inconvénients majeurs de cette méthode est le fait qu'elle ignore l'incertitude, liée à la reconstruction de l'arbre phylogénétique mais également celle liée à l'évolution des traits dans cet arbre.

L'emploi de méthodes Bayésiennes permet de supprimer la nécessité de fixer l'arbre phylogénétique ainsi que le modèle de transition d'un état à un autre (36). Une méthode Bayésienne fréquemment utilisée, le modèle de « mugration » (36), traite les changements d'états (ou migration quand les états sont des localisations) de la même manière que les mutations. Cette méthode considère donc que la forme de l'arbre phylogénétique n'est en aucune façon influencée par les changements d'états (37). Lorsque l'échantillonnage est biaisé, cette hypothèse peut alors conduire à des biais dans les estimations des taux de transition (38).

Enfin, les modèles de coalescence structurée, considérant des sous-populations définies par la valeur de leur trait (espèce ou localisation) et des phénomènes de migration ayant lieu entre ces sous-populations, ne font pas cette hypothèse d'indépendance entre la reconstruction phylogénétique et celle des états ancestraux (37,38). Avec une de ces méthodes, Dudas *et al.* ont ainsi identifié le dromadaire comme réservoir de MERS-CoV, au sein duquel la plus grand partie de l'histoire évolutive du virus a eu lieu, et la prédominance d'une transmission du dromadaire à l'humain (18) (similaire à la *Figure 4*). Cependant la reconstruction de ces états ancestraux ne permet pas d'identifier des évènements de transmission au sein d'une population d'hôtes (*ex*. au sein de la population de dromadaires), mais uniquement des évènements de transition entre sous-populations de pathogène définies par leur espèce-hôte. De plus, ces modèles de coalescence structurée font l'hypothèse d'une taille constante de la population au cours du temps, ce qui n'est pas forcément compatible avec le processus épidémique. Pour aller à une résolution plus fine de la transmission de l'agent pathogène au sein d'un système multi-hôtes, il faudrait identifier les évènements de transmission entre hôtes infectés. Pour ce faire, il est possible de reconstruire les chaînes de transmission représentées dans un arbre de transmission, retraçant ainsi « qui a infecté qui ? » dans une épidémie.

1.3 NOTIONS PRELIMINAIRES SUR LES ARBRES DE TRANSMISSION

1.3.1 Définition d'un arbre de transmission

Un arbre de transmission est un graphe orienté dont les nœuds représentent des hôtes infectés et les arcs des évènements de transmission

(datés ou non) (*Figure 5*). Un hôte correspond à une entité pouvant héberger l'agent pathogène, comme un individu ou un groupe d'individus (*ex*. un élevage (39–41)). Un évènement de transmission représente généralement l'infection initiale de l'hôte concerné, la possibilité de surinfection ou de réinfection étant négligée. Une surinfection correspond à une infection par une deuxième souche distincte de la première, alors que l'hôte est encore infecté. Ce phénomène a notamment pu être observé chez des patients atteints du VIH (42). Une réinfection se définit comme une infection par une deuxième souche distincte qui survient après le rétablissement de l'hôte. Le virus de l'hépatite C est un exemple d'agent pathogène qui peut donner lieu à des surinfections ou des réinfections (43).

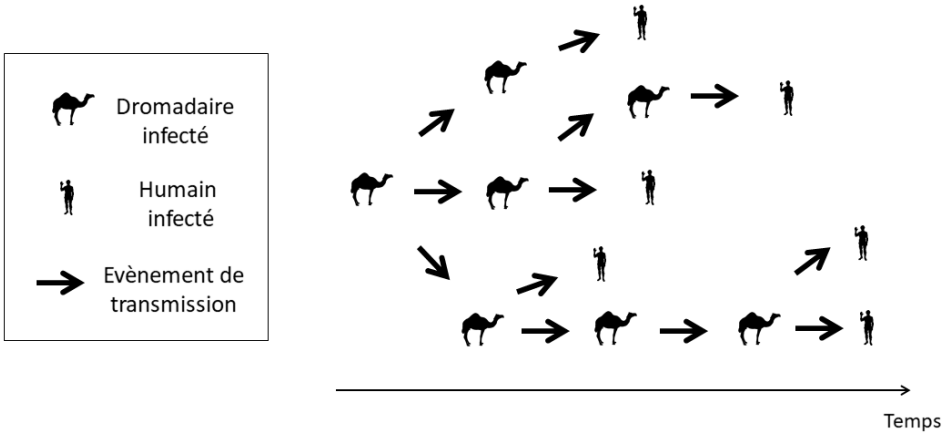


Figure 5 : Arbre de transmission décrivant l'histoire de transmission de MERS-CoV entre sept dromadaires et six humains. Les dromadaires ont été responsables des 12 évènements de transmission reconstruits (six de dromadaires et six d'humains). Aucun évènement de transmission à partir de l'humain n'a été reconstruit.

Reconstruire l'histoire de transmission d'une épidémie permet de mieux comprendre les mécanismes de transmission. Ainsi, afin d'estimer la part de la transmission de H7N7 qui pouvait être attribuée à une diffusion par le vent dans l'épidémie de 2003 aux Pays-Bas, Ypma *et al.* ont comparé des données météorologiques de direction du vent avec les directions de transmission décrites dans l'arbre de transmission (44). Connaître l'histoire de transmission permet également d'estimer des paramètres de transmission, tels que le nombre de reproduction, R, défini comme le nombre moyen d'infections secondaires qui ont été produites par un hôte infecté (45). Par exemple, Faye *et al.* ont cherché à estimer ce R dans l'épidémie d'Ebola en Conakry, Guinée en 2004. En estimant ce R selon trois contextes différents

(dans la communauté, à l'hôpital et aux enterrements), ils ont pu en déduire dans quels lieux les mesures de contrôle avaient été efficaces ainsi que le nombre d'infections qui auraient pu être évitées par une augmentation de l'hospitalisation des personnes infectées (46).

1.3.2 Différences entre arbres de transmission et arbres phylogénétiques

Comme vu précédemment, les séquences sont utilisées pour reconstruire des arbres phylogénétiques décrivant les liens de parenté entre les souches isolées. La proximité des souches dans l'arbre phylogénétique suggère l'existence d'un lien épidémiologique entre ces souches (47). Cette proximité dans l'arbre phylogénétique a notamment été utilisée comme argument en faveur de l'existence d'un évènement de transmission du VIH (*ex.* entre un dentiste et six de ses patients en Floride (48) ainsi qu'entre un chirurgien et un de ses patients en France (49)). Ceci implique la nécessité de définir précisément des « clusters de transmission » de VIH. La majorité des définitions utilisées dans la littérature se base sur la probabilité du nœud ancestral commun associée ou non à un seuil de distance génétique séparant les souches (50).

Certains auteurs ont considéré un arbre phylogénétique comme un arbre de transmission partiellement observé et ce, afin d'inférer des paramètres épidémiologiques tels que la prévalence de la maladie dans une population (51). Cependant, nous allons ici, comme Ypma *et al.* (41) souligner la nette distinction entre les deux. Pour rappel, dans un arbre phylogénétique, les feuilles correspondent aux souches isolées et les nœuds internes à des ancêtres hypothétiques de ces souches. Ceci s'oppose aux arbres de transmission, dans lesquels les nœuds sont tous des hôtes infectés. Sans inférer dans quel hôte se trouvaient les lignées ancestrales présentes dans l'arbre phylogénétique, il n'est pas possible de reconstruire précisément l'histoire de transmission. De plus, dans un arbre phylogénétique, les nœuds ne sont pas reliés par des flèches représentant des évènements de transmission mais par des branches représentant la distance génétique qui les sépare (*Figure 6*).

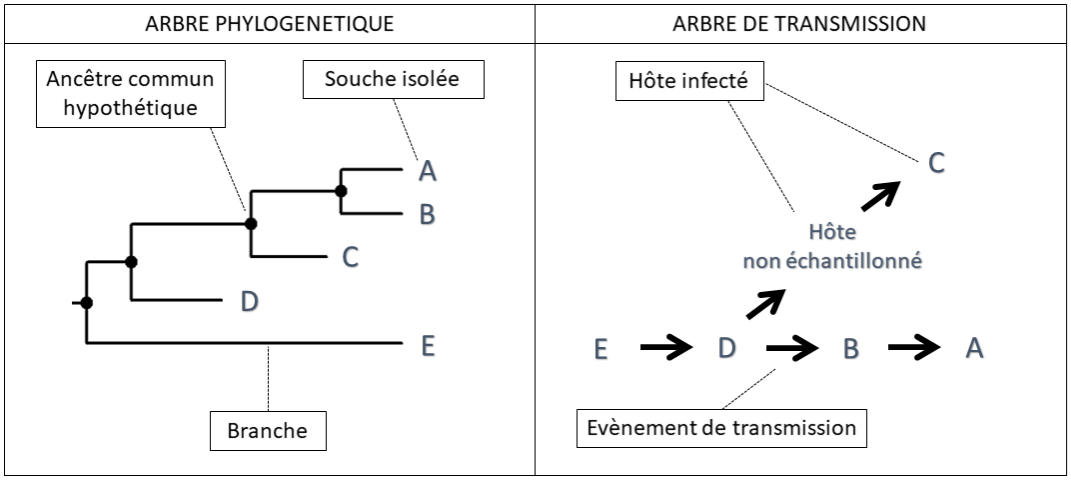


Figure 6 : Différences entre un arbre phylogénétique (à gauche) et un arbre de transmission (à droite). L'arbre phylogénétique et l'arbre de transmission ont été reconstruits à partir de cinq souches isolées (A à E). Cependant, un hôte non échantillonné intermédiaire (entre D et C) a été inféré dans l'arbre de transmission. L'arbre phylogénétique retrace ainsi l'histoire évolutive, et l'arbre de transmission, l'histoire de transmission.

1.3.3 Limites des données épidémiologiques

Le travail de reconstruction de l'arbre de transmission peut être réalisé à partir de données épidémiologiques seules, telles que la date de début des symptômes et des données de contact avec des personnes infectées. Ainsi, afin de reconstruire l'arbre de transmission d'une épidémie de SARS-CoV-1 à Toronto en 2003, un questionnaire portant sur d'éventuelles expositions au virus (*ex.* voyages) était administré aux patients (52). Cependant, ces données ne sont pas toujours disponibles (41) ni d'une résolution suffisante pour élucider l'histoire de transmission. De plus, ces données se basent sur la mémoire des personnes interrogées qui est loin d'être infaillible (53).

Enfin, la possibilité d'interroger ou d'observer les hôtes infectés implique une étape préalable de détection de ces hôtes. Sans la mise en place d'une stratégie de dépistage (surveillance active), les hôtes asymptomatiques ne sont pas échantillonnés. Cependant, la sensibilité imparfaite des tests et les conditions du terrain, notamment lors de surveillance de la faune sauvage, font que même avec la mise en place de tests de dépistage, nous n'allons pas détecter la totalité des hôtes infectés. Prenons l'exemple de la tuberculose bovine, une étude réalisée en France a montré que les vétérinaires ne suivaient pas toujours les recommandations officielles et avaient recours à

des pratiques pouvant diminuer la sensibilité des tests utilisés (*ex.* un tiers des vétérinaires interrogés déclaraient ne pas toujours injecter dans la région cervicale et un quart réalisait une lecture qualitative des résultats) (54).

1.3.4 Limites des données génomiques

Le taux de substitution ainsi que l'intervalle entre infection et échantillonnage d'un agent pathogène conditionne la distance génétique minimale qui permet de différencier les souches isolées dans une épidémie. Si le pathogène évolue lentement, il peut s'avérer difficile de reconstruire l'histoire de transmission à cause du faible nombre de différences entre séquences isolées (55). Prenons le cas de la bactérie *Mycobacterium tuberculosis* dont le taux de substitution s'élève à environ 1×10^{-7} substitutions par site et par an (56). Le nombre de séquences de *M. tuberculosis* identiques, isolées chez des personnes impliquées dans des évènements de transmission au sein d'un même foyer, peut représenter 40 % (15/38) des séquences (57). A l'inverse, un taux de substitution trop élevé, une évolution intra-hôte non-négligeable et/ou un « bottleneck » de transmission incomplet (c'est-à-dire un nombre de souches transmises strictement supérieur à 1) peuvent conduire à une dynamique de transmission trop complexe à reconstruire (58).

Au vu des limites de données épidémiologiques et génomiques, il s'avère intéressant de combiner l'information apportée par ces deux types de données lors de la reconstruction d'arbres de transmission.

1.4 EXEMPLE D'UN SYSTEME MULTI-HOTES IMPLIQUANT LA FAUNE SAUVAGE : LA TUBERCULOSE BOVINE

1.4.1 La constitution du système multi-hôtes de la tuberculose bovine

La tuberculose bovine (bTB) est une maladie généralement asymptomatique qui touche principalement les bovins. Elle peut être due à trois espèces : *Mycobacterium tuberculosis*, *Mycobacterium caprae* et *Mycobacterium bovis* (59). Cependant, *M. bovis*, l'agent pathogène le plus fréquent de la bTB, peut affecter d'autres espèces domestiques (*ex.* petits ruminants et porcins) ainsi que de la faune sauvage (60). Certaines espèces de la faune sauvage ont été impliquées comme réservoirs de *M. bovis* dans le monde. Nous pouvons citer par exemple le phalanger-renard (*Trichosurus vulpecula*) en Nouvelle-Zélande (61) et le cerf de Virginie (*Odocoileus virginianus*) dans le Michigan aux Etats-Unis (62). En Europe, les données recueillies sont en faveur du blaireau (*Meles meles*) en Irlande et Grande-

Bretagne (63) et du sanglier (*Sus scrofa*) en Espagne (64) en tant que réservoirs de bTB. La faune sauvage peut donc jouer un rôle important dans un système multi-hôtes de la bTB, dont la composition varie selon la région étudiée.

En France, la première détection de *M. bovis* dans la faune sauvage date de 2001 et concerne des sangliers et des cerfs élaphes (*Cervus elaphus*) dans la forêt de Brotonne en Normandie (65). Depuis sa création en 2011, le réseau national de surveillance de la faune sauvage, appelé « Sylvatub », rapporte principalement des cas de blaireaux et sangliers infectés dans les Pyrénées-Atlantiques, Landes, Dordogne et Côte-d'Or ainsi que des cerfs élaphes et chevreuils (*Capreolus capreolus*) en Dordogne et Côte-d'Or (66). A la suite de la découverte de renards infectés (4 sur 6 prélevés) en Dordogne en 2015 (67), une étude a eu pour objectif d'estimer le taux d'infection à *M. bovis* chez les renards (*Vulpes vulpes*), espèce non ciblée par « Sylvatub ». Dans cette étude, des taux d'infection comparables avec ceux retrouvés chez les blaireaux (9 à 13,2 %) et les sangliers (4,3 à 17,9 %) ont été estimés chez des renards en Dordogne (7,1 %, intervalle de confiance à 95%, ou IC95%, de [3,8–11,8%]), Charente (9,2 %, IC95% [4,3–16,7 %]) et dans les Landes (5,0 %, IC95% [2,0–10,0 %]) (68). La connaissance d'un système multi-hôtes dans une région donnée peut donc s'avérer incomplète du fait du nécessaire ciblage des espèces soumises à la surveillance.

1.4.2 La surveillance de la tuberculose bovine en France

1.4.2.1 Chez les bovins

La surveillance chez les bovins est constituée de deux volets : a) l'inspection post-mortem systématique des carcasses à l'abattoir en vue de la détection de lésions compatibles avec la bTB, et b) le dépistage par intradermo-tuberculination réalisé de manière périodique en élevage. Ce dépistage en élevage détecte la majorité (actuellement plus de 70 %) des foyers de bTB et ce depuis 2007, avec la mise en place d'une surveillance plus ciblée en élevage qui se concentre sur les cheptels à risque (*ex.* autour de foyers initialement détectés grâce à la surveillance en abattoir) (13).

L'intradermo-tuberculination simple (utilisation de tuberculine bovine seule, IDS) ou comparée (IDC, utilisation conjointe de tuberculine aviaire) est réalisée dans la région cervicale et la lecture des résultats se fait 72 h après injection. Les caractéristiques des tests de dépistage de la bTB varient selon les techniques utilisées et leur mise en œuvre. Ainsi, une étude a rapporté des estimations de sensibilité (Se égale au nombre d'individus infectés

positifs sur le nombre d'individus infectés) variant entre 58,6 et 91 % et de spécificité (Sp égale au nombre d'individus sains négatifs sur le nombre d'individus sains) variant entre 75 et 99,7 % pour l'IDS, ainsi qu'une Se d'environ 50 % et une Sp supérieure à 99,9 % pour l'IDC (69). Le rythme de dépistage en élevage se décide à l'échelle du département et peut varier du contraignant dépistage annuel concernant tous les animaux du troupeau âgés de plus de six semaines, à l'absence de dépistage. L'intradermo-tuberculination est également réalisée avant l'introduction dans un troupeau d'un bovin, dont le transport a duré plus de six jours, provenant d'un troupeau à risque ou d'un département dont l'incidence cumulée à cinq ans est supérieure à la moyenne nationale, ainsi que lorsqu'il existe un lien épidémiologique entre l'élevage d'origine et un foyer déclaré infecté.

Après abattage dû à un test positif ou à la détection de lésions caractéristiques de bTB, un test PCR puis une culture bactérienne sont réalisés afin de détecter des mycobactéries. Les souches sont envoyées au Laboratoire Nationale de Référence (LNR, ANSES Maisons-Alfort) pour spoligotypage et typage VNTR (Variable Number of Tandem Repeats). Le spoligotype et VNTR détermine le génotype, ce génotypage systématique des souches *M. bovis* a permis l'identification de génotypes régionaux (59).

1.4.2.2 Dans la faune sauvage

Sylvatub cherche à détecter la bTB chez les blaireaux, sangliers, cerfs élaphes et chevreuils en France (66). Il existe trois niveaux de surveillance qui définissent le protocole de surveillance de la faune sauvage, les mesures concernées vont du ramassage de carcasses d'animaux retrouvés sur le bord de la route à la mise en place de campagnes de piégeages de blaireaux (*Tableau 1*).

Tableau 1 : Modalités de surveillance en fonction des niveaux de surveillance (note de service DGAL/SDSPA/2018-708).

Type de surveillance	Modalités de surveillance	Niveau 1	Niveau 2	Niveau 3
Événementielle	Recherche de lésions suspectes chez les cervidés et sangliers lors de l'examen de carcasse dans le cadre d'une pratique de chasse habituelle.	X	X	X
	Recherche de lésions évocatrices de tuberculose chez les sangliers, cervidés et blaireaux collectés dans le cadre du réseau Sagir (animaux morts ou mourants) dans son fonctionnement normal.	X	X	X
Événementielle renforcée	Recherche analytique systématique de tuberculose chez les sangliers, cerfs et blaireaux collectés dans le cadre du renforcement du réseau Sagir.		X	X
	Recherche analytique systématique de tuberculose chez les cadavres de blaireaux collectés sur les routes dans le cadre du renforcement réseau Sagir. Ce renforcement des analyses doit s'accompagner d'un renfort de collecte sur l'ensemble des zones de prospection et des zones tampon.		X	X
Programmée	Recherche systématique de tuberculose sur un échantillon de blaireaux prélevés dans les zones infectées de la zone à risque ou en zone de prospection.		X	X
	Recherche systématique de tuberculose sur un échantillon de sangliers prélevés sur l'ensemble de la zone à risque.			X

Ces niveaux de surveillance sont définis à l'échelle départementale, sauf pour certaines zones plus à risque appelées ZPR (Zone de Prophylaxie Renforcée). Le protocole est le même que pour les bovins, à savoir : test PCR, culture bactérienne et génotypage (66). Cependant, la dégradation des carcasses et la possible contamination par le milieu extérieur pourraient diminuer la Se de la culture bactérienne de 35 % par rapport à celles réalisées à partir d'échantillons de bovins (70). De plus, la réalisation de PCR sur des échantillons groupés provenant de la faune sauvage pourrait également diminuer la Se de ce test de 15 % (70). Ainsi, alors que nous détectons sans

doute la plupart des bovins infectés, la proportion des individus infectés dans la faune sauvage reste inconnue et les données provenant des espèces sauvages surveillées ne sont disponibles qu'à partir de 2012. Nous sommes donc bien confrontés dans ce système multi-hôtes à un biais d'échantillonnage entre la faune sauvage et les bovins.

1.4.3 Les enjeux de la tuberculose bovine en France

1.4.3.1 Statut officiellement indemne de la France

Dans l'Union Européenne (UE), la bTB a fait l'objet de programmes de contrôle (EU directive 64/432/EEC). En 1954, un programme d'éradication de la bTB chez les bovins a permis une baisse de l'incidence de la bTB dans les élevages français, de 13 % en 1965 à moins de 0,1 % en 2000 (71). L'obtention du statut officiellement indemne de bTB (OIT) en 2001 a été permis par le maintien de la prévalence de bTB strictement inférieure à 0,1 % pendant une période de six ans. Le succès des mesures de lutte mises en place (72) ainsi que la pasteurisation systématique du lait (décret du 21 mai 1955) limitent le risque zoonotique (73), et les cas humains de *M. bovis* en France demeurent actuellement très rares. Une étude rétrospective de 2018 a en effet mis en évidence le fait que seul 0,02 % (198/9397) des cas de tuberculose humaine de 2011 à 2016 était dû à *M. bovis* (74). Cependant, *M. bovis* est naturellement résistant au pyrazinamide, ce qui empêche l'utilisation de la quadrithérapie de six mois et impose le traitement de neuf mois sans pyrazinamide (d'après l'avis du Haut Conseil de la santé publique du 25 septembre 2020). Enfin, le fait que ce statut OIT facilite le commerce de bovins vivants dans l'UE et avec d'autres pays tiers (EU directive 64/432/EEC, (75)) lui confère un intérêt principalement économique. Cependant, la présence de foyers encore persistants dans le Sud-Ouest de la France (*Figure 7*) menace la conservation du statut OIT (76).

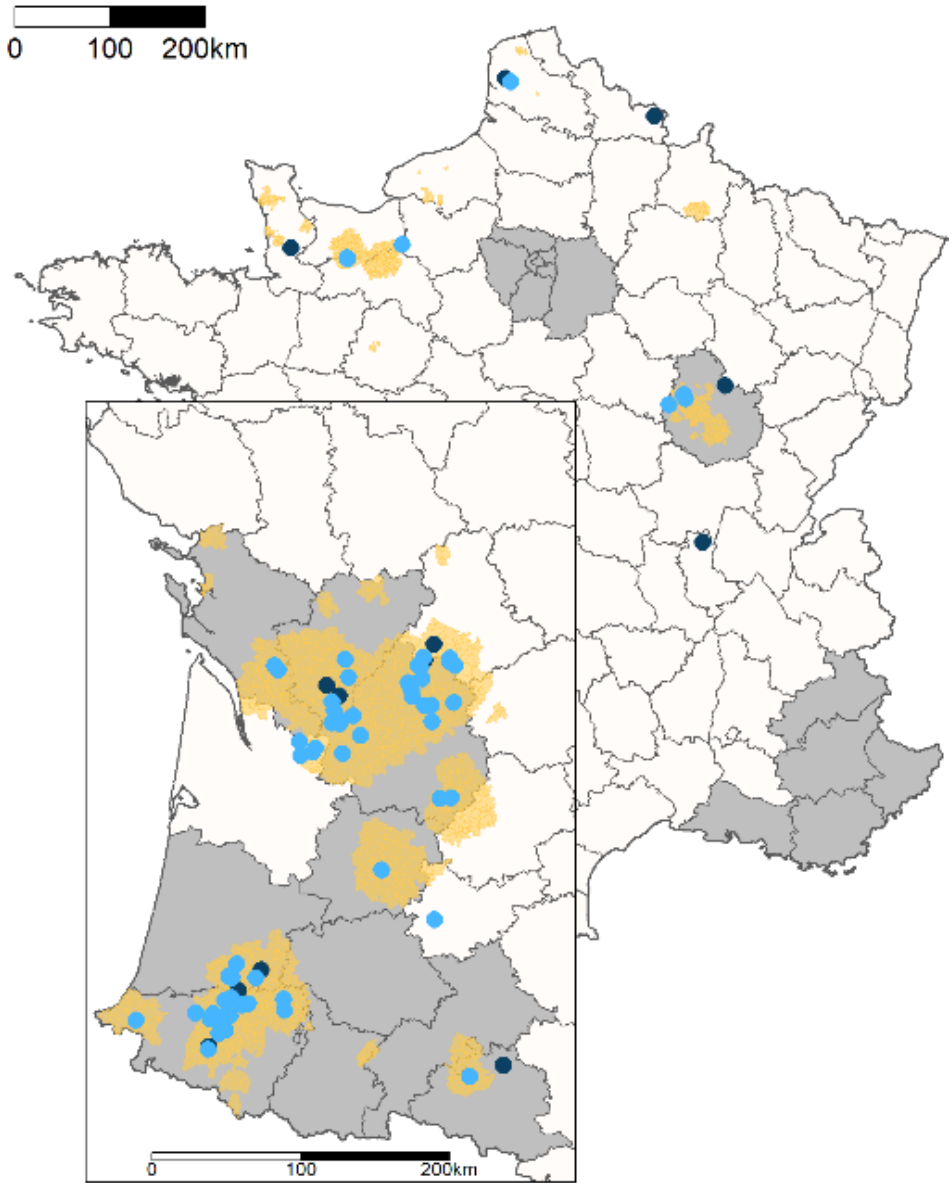


Figure 7 : Distribution des foyers incidents bovins détectés en France en 2019 (carte construite et publiée par Delavenne et al.*, 2021, (13)).* Les départements en gris sont en surveillance programmée, en blanc sans surveillance. Les zones en jaunes correspondent à des ZPR. Les points en bleu clair sont les foyers détectés par la surveillance en élevage, et en bleu foncé, par la surveillance en abattoir.

1.4.3.2 Mesures de contrôles et conséquences au niveau d'un élevage

L'identification d'un cas positif à la bTB dans un élevage conduit à un arrêté préfectoral de déclaration d'infection (APDI). Cet APDI entraîne

l'implémentation de mesures de contrôle au sein de l'élevage déclaré infecté : recensement, isolement et séquestration des bovins puis abattage total (dans un délai de 60 jours maximum). Cependant, si les données épidémiologiques sont en faveur d'une infection récente et transmission faible au sein du troupeau (conditions définies dans l'article 24 de l'arrêté du 8 octobre 2021), il est possible de recourir à un abattage sélectif (seuls les animaux réagissant aux tests sont alors abattus). Enfin, une enquête épidémiologique approfondie, visant à déterminer la source d'infection de l'élevage (amont) ainsi que les possibles transmissions à partir de l'élevage (aval), est mise en place. Cette enquête « en amont » et « en aval » conduit à la mise en place de mesures de dépistage dans les élevages ayant un lien épidémiologique établi (commerce de bovins) avec l'élevage infecté (77).

La récupération de la qualification indemne s'obtient après trois contrôles négatifs du troupeau espacés d'un minimum de deux mois (en cas d'abattage partiel) (78) ainsi qu'un nettoyage-désinfection ; le temps avant obtention de cette requalification peut donc être très long (76). De plus, un troupeau ayant récupéré sa qualification après avoir été reconnu infecté est considéré à risque sanitaire pendant une durée de cinq ans (arrêté du 8 octobre 2021). Une étude, s'intéressant à l'impact de ces mesures de contrôle et d'assainissement sur la survie des élevages, a montré l'augmentation du risque d'arrêt à long terme de l'activité d'élevage (*i.e.* plus de 12 mois consécutifs sans aucun bovin dans l'élevage), après détection de bTB dans les élevages bovins vulnérables (79).

1.4.4 Le séquençage complet des souches de *M. bovis*

1.4.4.1 Le génome de M. bovis

M. bovis fait partie du complexe *M. tuberculosis* et son génome comprend environ 4,3 millions de paires de bases (pb) dont plus de 65 % sont des bases G ou C (80). Du fait d'un taux de mutation faible (inférieur à une substitution par génome et par an (81)) ainsi que l'absence d'échanges de matériel génétique entre mycobactéries, les souches de *M. bovis* sont clonales et peu variables dans le temps (59). Pour cette raison, comme souligné par Michelet *et al.*, l'étude de certains caractères génotypiques peut être très informative et venir compléter les études épidémiologiques classiques.

Dans le contexte de la surveillance de la bTB en France, le LNR procède au génotypage des souches de *M. bovis* isolées, c'est-à-dire au spoligotypage et au typage VNTR. Le spoligotypage est une technique basée sur le polymorphisme d'une région conservée de l'ADN, le locus Direct Repeat. Ce

locus contient un nombre variable de « direct repeats » (DR) qui correspondent à de courtes séquences nucléotidiques (36 pb) fréquemment répétées dans le génome. Ces DR sont séparés entre eux par des espaceurs de 34 à 41 pb de long. Le génome de *M. tuberculosis* possède 43 espaceurs dont la position est fortement conservée et dont la présence ou absence est codée de manière binaire (0/1). Le spoligotypage permet entre autres de distinguer le génome de *M. bovis* de celui de *M. tuberculosis*, car le génome de *M. bovis* ne possède jamais les espaceurs 3, 9, 16 et 39 à 43 (82).

Le principal désavantage du spoligotypage est le fait qu'il se base exclusivement sur un seul locus du génome bactérien. Cependant, ce locus DR étant très stable, « les profils de spoligotypage permettent de reconstruire des évènements évolutifs de manière fiable » (59). A l'inverse, le typage VNTR s'intéresse à des loci répartis dans tout le génome bactérien et identifie le nombre de répétitions en tandem (séquences nucléotidiques qui se répètent à l'identique) qui y sont présentes (83). De plus, ces loci évoluant plus rapidement que le locus DR, « les profils VNTR permettent une analyse plus fine des souches » (59). Cependant, si un seul locus est étudié, il y a de grandes chances pour que le même nombre de répétitions soit retrouvé dans deux souches qui n'ont aucun lien entre elles (homoplasie), la comparaison de multiples loci augmente la résolution et la fiabilité de la méthode (81). Huit loci préconisés par le consortium européen sont utilisés en France (2165, 2461, 0577, 580, 2163a, 2163b, 4052 et 3232) (84).

1.4.4.2 Séquençage complet

Le génotypage systématique des souches de *M. bovis* a permis la mise en évidence du fait que les souches isolées chez les bovins ainsi que celles isolées dans la faune sauvage à proximité de ces foyers bovins partagent souvent le même génotype (85). Du fait de ce partage de génotype entre souches, il est nécessaire de recourir à une méthode plus discriminante afin de les différencier et mieux comprendre les dynamiques de transmission. L'utilisation des données de séquençage complet de *M. bovis* de souches isolées en France a notamment permis une analyse plus fine des liens de parenté entre différents groupes génotypiques (86). Dans cet arbre (*Figure 8*), la famille F4 dont les souches sont isolées dans le Sud-Ouest de la France, est représentée en rouge.

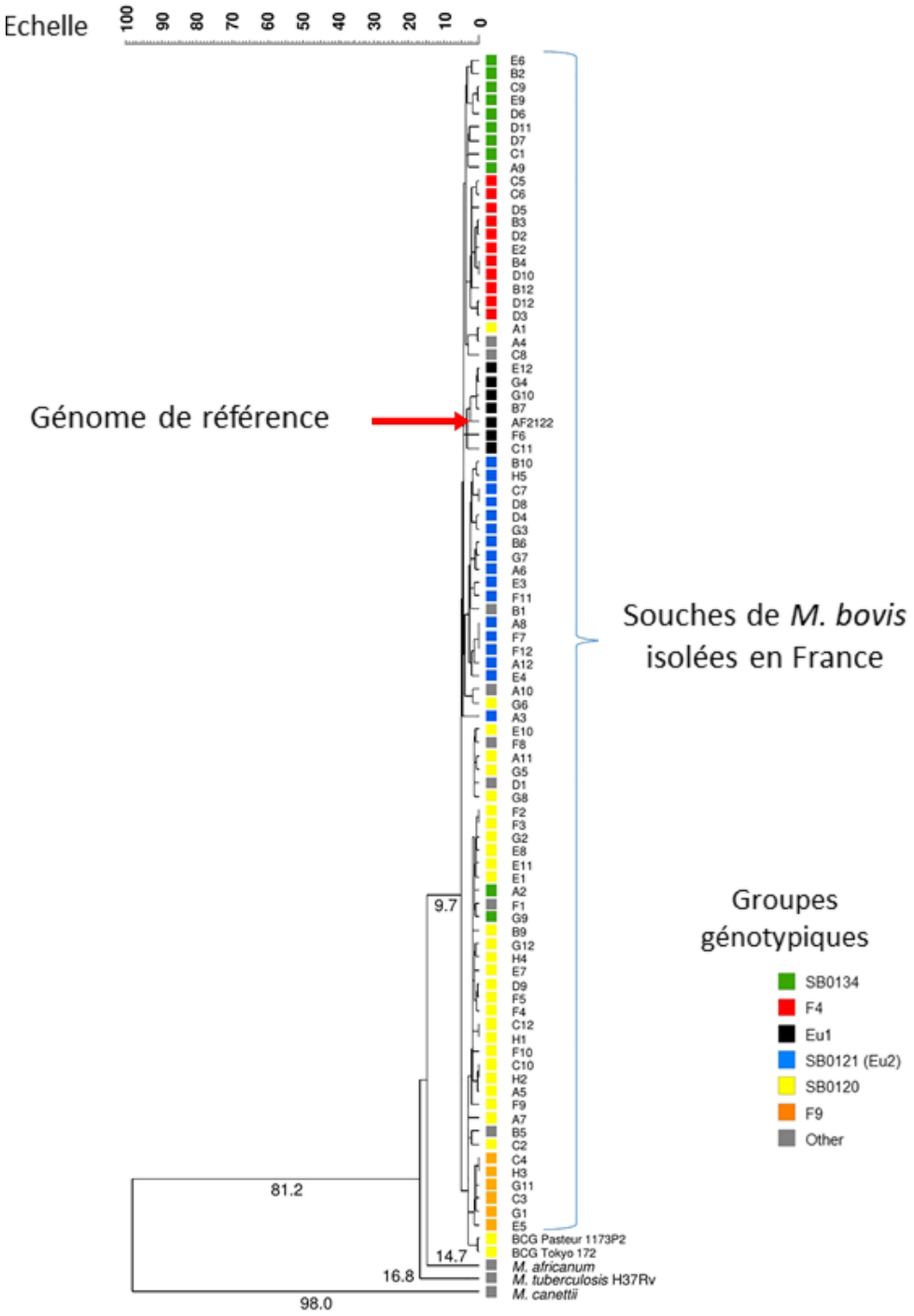


Figure 8 : Arbre phylogénétique reconstruit à partir des données de séquençage complet de souches de M. bovis *isolées en France et de souches de référence (arbre construit et publié par Hauer* et al.*, 2019, (86)).* L'échelle et les nombres situés au-dessus des branches correspondent aux valeurs de coefficients de

similarité entre les souches isolées et leur ancêtre commun.

Le séquençage complet du génome consiste à fragmenter l'ADN, séquencer les fragments d'ADN puis aligner les fragments à partir (ou non, séquençage *de novo*) d'une séquence de référence. La souche de référence dépend du spoligotype, AF2122/97 (souche anglaise isolée en 1997, (80)) représentée par la flèche rouge dans la *Figure 8*, est employée notamment pour les souches de la famille F4. Pour *M. bovis*, du fait de nombreuses zones du génome avec des répétitions, le génome complet n'est jamais obtenu mais ce séquençage permet toutefois l'identification de mutations ponctuelles aussi appelées SNP pour « Single Nucleotide Polymorphism ». Du fait de sa résolution plus fine, le séquençage complet de souches de *M. bovis* a précédemment été sélectionné afin d'investiguer la transmission de la bTB dans un système multi-hôtes (87,88). Pour ce faire, ces études ont eu recours à la reconstruction d'arbres phylogénétiques à partir des SNP identifiés dans les souches échantillonnées.

1.5 OBJECTIFS DE LA THESE

Dans un système multi-hôtes, l'estimation de la contribution de chaque espèce à la transmission est indispensable pour définir des stratégies efficaces de prévention et de lutte. La bTB est un exemple de système multi-hôtes dont la composition et le fonctionnement varie selon la région étudiée. De plus, ce système multi-hôtes peut mettre en jeu à la fois la faune sauvage et la faune domestique, notamment dans des régions d'intérêt pour la conservation du statut OIT, comme le Sud-Ouest de la France. Le contraste entre une surveillance partielle de la faune sauvage mise en place récemment (2012) et une surveillance plus exhaustive des bovins, bien plus ancienne (1954), complique l'étude de ce système multi-hôtes, même quand tous les acteurs semblent bien identifiés dans une région donnée. Nous nous retrouvons donc bien face à un biais d'échantillonnage des individus infectés, entre la faune sauvage et la faune domestique, qui pourrait avoir un impact non négligeable sur un effort de quantification de la contribution de chaque espèce-hôte à la transmission de la bTB.

Notre objectif principal était de mieux comprendre l'impact de ce biais d'échantillonnage dans les études basées sur les données de séquençage recueillis dans un système multi-hôtes, et ce, afin d'améliorer les conclusions tirées sur la contribution de chaque espèce. Dans ce but, notre premier objectif était d'étudier un système multi-hôtes de la bTB avec une méthode déjà bien utilisée dans des systèmes multi-hôtes (87,88) : la

reconstruction des espèces-hôtes des nœuds ancestraux dans un arbre phylogénétique. Cependant, cette méthode étudie la transmission entre populations de pathogènes et non entre individus, elle ne reconstruit donc que partiellement les dynamiques de transmission au sein d'un système multi-hôtes. Pour aller plus loin, nous avons ensuite procédé à l'identification des méthodes permettant de reconstruire plus précisément l'histoire de transmission dans un système multi-hôtes de la bTB. Pour ce faire, nous avons réalisé une revue systématique de la littérature des méthodes de reconstruction d'arbres de transmission, utilisant à la fois des données épidémiologiques et des données génomiques. Nous avons ensuite testé plusieurs de ces méthodes sur des données simulées de bTB afin d'évaluer la fiabilité de leurs résultats dans un tel contexte ainsi que leur robustesse aux biais d'échantillonnage.

2 ETUDE D'UN SYSTEME MULTI-HOTES DE LA TUBERCULOSE BOVINE PAR RECONSTRUCTION D'ARBRES PHYLOGENETIQUES

2.1 UN EXEMPLE DE SYSTEME MULTI-HOTES DE LA TUBERCULOSE BOVINE AVEC UN INTERET MAJEUR EN FRANCE

Une des zones où l'évolution récente de la bTB menace le statut OIT en France se situe dans les Pyrénées-Atlantiques (PA) et les Landes. Les derniers bulletins publiés en 2021 par la plateforme d'Epidémiosurveillance en Santé Animale (ESA) montrent que la majorité des élevages infectés en France (68/92 (74 %) en 2019 et 84/104 (81 %) en 2020) ont été détectés en Nouvelle-Aquitaine. Les zones touchées se situent principalement dans les PA, les Landes, la Charente et la Dordogne. La surveillance des bovins ainsi que celle de la faune sauvage dans les PA et les Landes rapportent la présence de deux spoligotypes, SB0821 et SB0832, appartenant à la famille F4/cluster A (84,86). Ces deux spoligotypes (SB0821, majoritaire qui représente 32 % des foyers bovins de la région sur la période 2002-2017, et SB0832) sont partagés par trois espèces-hôtes : les bovins, les blaireaux et les sangliers (85).

En 2012, année de mise en place de Sylvatub dans les PA et les Landes, le rythme de dépistage en élevage est devenu annuel dans les communes où des foyers de bTB avaient été déclarés l'année précédente. Avant la généralisation de ce rythme annuel à toutes les communes en 2018, le rythme de dépistage, dans les communes sans foyer déclaré l'année précédente, était de tous les deux ans dans les Landes et tous les trois ans dans les PA. Le nombre de foyers incidents annuels était compris entre 16 et 31 de 2012 à 2017 et ne montrait aucune tendance à la hausse ni à la baisse (Boschiroli, communication personnelle). De plus, la prévalence de la bTB chez les blaireaux a été estimée à 5,9 % (IC95% [3,9-6,8 %]) en 2013-2014 (par culture) et 7,9 % (IC95% [5,2-11,2 %]) en 2016-2017 (par PCR) (66).

Notre objectif était de mieux comprendre les dynamiques de transmission des souches *M. bovis* SB0821 dans le système bovins-blaireaux des PA et Landes. Pour ce faire, nous avons reconstruit un arbre phylogénétique à partir de données de séquençage total. Le nombre de séquences isolées chez les sangliers se limitant à six, nous avons dû exclure cette espèce-hôte du système étudié. Puis, nous avons utilisé une méthode

Bayésienne de reconstruction des états ancestraux pour inférer les taux de transitions entre blaireaux et bovins ainsi qu'évaluer l'évolution temporelle des espèces-hôtes dans l'arbre reconstruit.

2.2 MATERIEL ET METHODE

2.2.1 Collecte des données

Notre zone d'étude était composée des communes précédemment sélectionnées dans deux études sur le système bovins-blaireaux dans le Sud-Ouest de la France (89,90). Cette zone était restreinte à 3754 km^2 comprenant 335 communes (*Figure 9*), et se situait à cheval entre les PA et les Landes. Un maximum de trois souches SB0821 par APDI, collectées entre 2002 et 2017 (période d'étude), ont été incluses. Toutes les souches SB0821 isolées chez des blaireaux et collectées par Sylvatub pendant notre période d'étude ont été incluses. La date d'échantillonnage considérée était soit la date d'abattage, pour les souches isolées chez des bovins, soit la date de piégeage enregistrée par Sylvatub, pour les souches isolées chez des blaireaux. Les données sur les bovins provenaient de la « base de données nationale d'identification » (BDNI), dans laquelle tous les mouvements des bovins en France sont enregistrés, de leur naissance (ou importation) à leur mort (ou exportation).

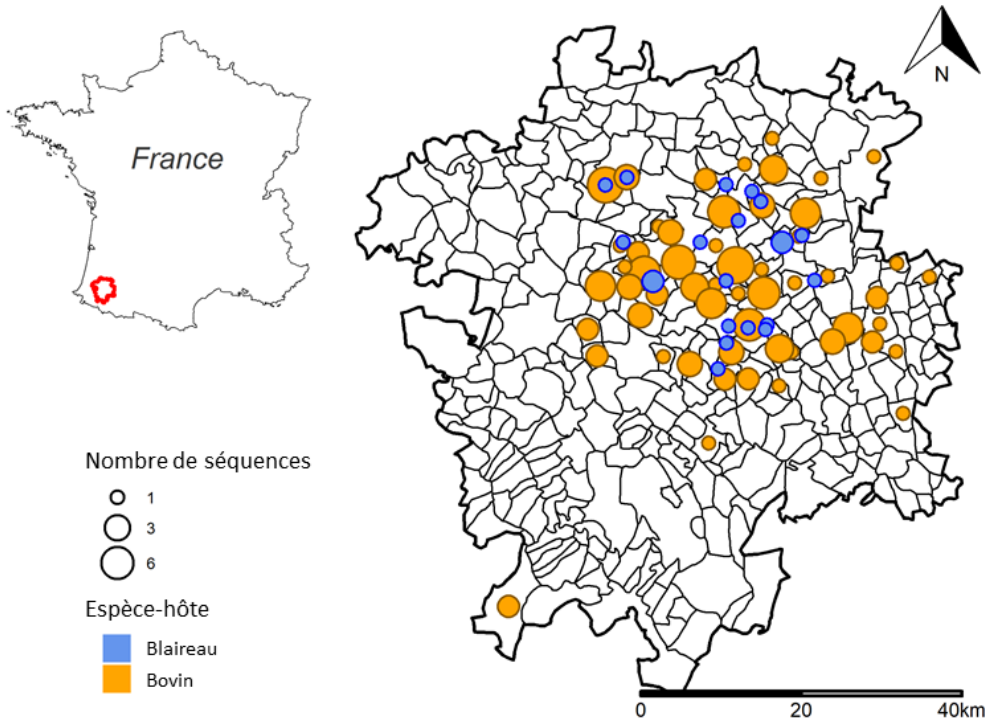


Figure 9 : Nombre et répartition dans la zone d'étude des souches incluses (isolées chez des bovins en jaune et blaireaux en bleu).

2.2.2 Données génomiques

Après réception au LNR, les souches de *M. bovis* ont été mises en milieu de culture liquide (7H9 + ADC). Après chauffage et lavage, le lysat obtenu, composé d'ADN bactérien avec des débris cellulaires, était envoyé pour purification et séquençage Illumina (double lecture de fragments de 150 bp en moyenne) à l'Institut du Cerveau et de la Moelle Epinière (ICM). A l'ICM, la qualité du séquençage était contrôlée avec FASTQC, en considérant comme seuil d'acceptabilité un score de Phred (Q) de 30. Ce score est défini par l'équation :

$$Q = -10 * \log(p)$$

avec p, la probabilité moyenne d'erreur de positionnement d'une base sur un fragment de lecture. Si Q=30, on accepte un risque d'avoir mal identifié une base donnée de 1 pour 1000.

L'alignement et l'assemblage des fragments par homologie à la souche de référence AF2122/97 puis l'identification de SNP ont été réalisés au LNR avec le logiciel Bionumerics, version (v.) 7.6 (AppliedMath, Belgium). Les critères de sélection des SNP étaient : (i) présence sur au moins cinq fragments de lecture, et ce, dans les deux sens de lecture, (ii) séparation de deux SNP par un minimum de 12 paires de base, (iii) localisation en-dehors des régions répétées du génome et (iv) les SNP ambigus (avec au moins une base incertaine (N), ambiguë (qui n'est pas ATCG) ou un trou) n'ont pas été inclus. Ces SNP ont alors été utilisés pour reconstruire un arbre par maximum de parcimonie sur Bionumerics afin d'identifier des « outliers génétiques » (souches aberrantes). Nous avons estimé la distance génétique entre chaque paire de séquences avec la fonction dist.dna disponible dans le package *ape* v.5.0 (91) sur R4.0.3. Le modèle de substitution considéré était le modèle F84 (92) car il ressemble au modèle HKY (21) qui avait déjà été utilisé sur des séquences de *M. bovis* (93,94). Les distances génétiques étaient estimées : entre toutes les séquences, entre celles isolées uniquement chez des blaireaux et enfin entre celles isolées uniquement chez des bovins.

2.2.3 Modèle d'évolution Bayésien

Dans le but d'identifier des transitions entre deux populations de pathogènes définies par leur espèce-hôte, nous avons utilisé un modèle d'évolution Bayésien qui intégrait un modèle de population coalescent structuré dans BEAST2 (pour « Bayesian Evolutionary Analysis by Sampling Trees ») v.2.6.3 (95). Plus précisément, la structure de ces sous-populations de pathogènes était inférée à partir de l'approximation marginale de la coalescence structurée, implémentée dans le package Mascot v.2.1.2 (96).

Nous avons sélectionné le modèle d'évolution (modèle de substitution combiné au modèle d'horloge) le plus approprié, après comparaison deux à deux de leur Facteur de Bayes (BF pour « Bayes Factor »). Le BF correspond au log du rapport des vraisemblances marginales des deux modèles testés. Ce BF était obtenu après estimation des vraisemblances marginales avec l'algorithme de "Nested Sampling" (NS) implémenté dans le package NS v.1.1.0 (97). Dans l'estimation par NS, le nombre de particules était de N=1 (ou 10, si les résultats n'étaient pas concluants avec N=1) et la longueur de la sous-chaîne était fixée à 100 000. Nous avons d'abord testé trois modèles de substitution avec un modèle d'horloge strict : le modèle JC (23), le modèle HKY (21), et enfin le modèle GTR (20). Nous avons ensuite testé trois modèles d'horloge avec le modèle de substitution sélectionné juste avant, l'horloge stricte et deux horloges relâchées : l'horloge non corrélée lognormale et exponentielle (27). Tous les modèles ont été testés

avec le logiciel BEAST2 après annotation sur l'interface BEAUti (pour « Bayesian Evolutionary Analysis Utility ») (95). Nous avons choisi une fréquence empirique des bases dans le modèle de substitution et une distribution lognormale (moyenne : 0, déviation standard : 1) comme distribution à priori de la population constante effective. Les autres paramètres conservaient leur valeur par défaut. Nous avons sélectionné une longueur de chaîne de 300 millions d'itérations, une période de burn-in de 10 % et une fréquence d'échantillonnage de 1 sur 30 000. Quatre chaînes avec ces caractéristiques ont été combinées dans LogCombiner v.2.6.3 avec une fréquence d'échantillonnage de 1 sur 120 000.

Nous avons vérifié la convergence des chaînes MCMC (Markov Chain Monte Carlo), *i.e.* distribution stationnaire (tracé en « chenille poilue ») et indépendance de l'échantillonnage (taille effective d'échantillon, ou ESS pour « Effective Sample Size », supérieure à 200 pour chaque paramètre), sur TRACER v.1.7.1 (98). Afin de résumer les arbres postérieurs échantillonnés, nous avons construit l'arbre avec maximum de crédibilité de clade (MCC) grâce à Tree Annotator, en sélectionnant l'option de la hauteur des ancêtres communs (99). Dans l'arbre MCC, l'espèce-hôte des nœuds internes a été considérée inconnue si la probabilité postérieure de l'espèce-hôte ('Blaireau' ou 'Bovin') était inférieure à 0,7, sinon nous avons distingué arbitrairement trois catégories de probabilités :]0,7; 0,8],]0,8; 0,9] et]0,9; 1]. Nous avons ensuite inféré l'évolution de l'espèce-hôte des lignées au cours du temps, en considérant que la transition d'un nœud à un autre s'effectuait toujours au nœud parent. L'arbre MCC a été visualisé sur R4.0.3 avec les packages *treeio* (100) et *ggtree* (101).

Nous avons ensuite ré-échantillonné les arbres postérieurs à une fréquence de 1 sur 1 200 000 avec LogCombiner. Puis, nous avons importé les 1004 arbres obtenus dans R. Dans chaque arbre phylogénétique, l'espèce-hôte (blaireau ou bovin) est estimée pour chaque nœud. Nous pouvons ainsi compter le nombre de fois où un nœud parent et un de ses nœuds descendants ne sont pas hébergés par la même espèce-hôte. Ce nombre correspond au nombre de transitions inter-espèces au sein d'une lignée (de blaireau à bovin et de bovin à blaireau). Cependant, quand deux nœuds consécutifs sont hébergés par la même espèce-hôte, nous ne pouvons pas inférer avec cette méthode si la lignée réside dans le même animal, dans le même groupe d'animaux (groupe social si on considère des blaireaux ou élevage pour les bovins), ou si un ou plusieurs évènements de transmission intra-espèces ont eu lieu, au sein d'un même groupe ou entre plusieurs groupes d'animaux. De la même manière, entre deux nœuds hébergés dans des espèces-hôtes différentes, au moins un évènement de transmission a eu

lieu (entre un blaireau et un bovin), mais d'autres évènements de transmission (entre animaux d'une même espèce) pourraient également avoir eu lieu. Pour chaque arbre, nous avons compté le nombre de transitions au sein d'une lignée, le nombre de fois où une lignée restait dans la même espèce-hôte (que nous avons appelé persistance intra-espèce) ainsi que le nombre de transitions inconnues, *i.e.* le nombre de fois où au moins l'un des nœuds consécutifs est inconnu (probabilité de l'espèce-hôte inférieure à la valeur seuil : 0,7, 0,8 ou 0,9) (voir *Figure 10*).

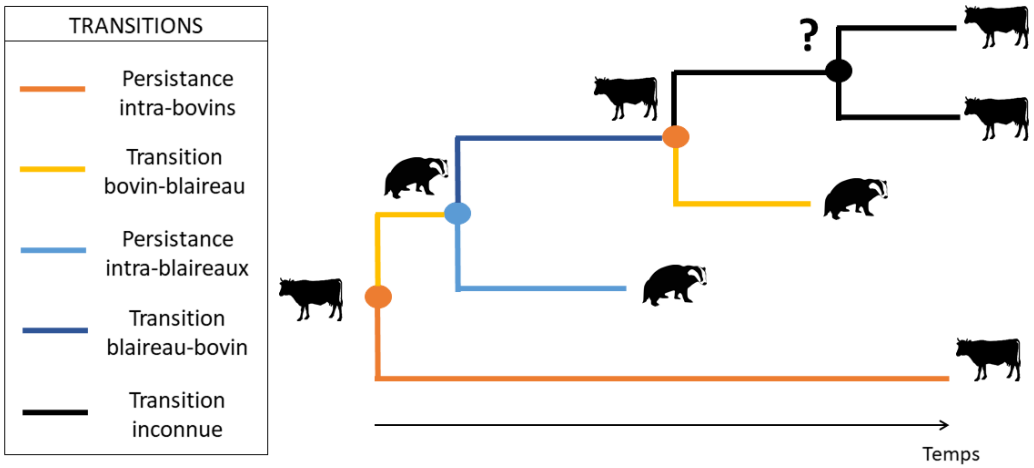


Figure 10 : Types de transitions pouvant être identifiés dans un arbre simple où l'espèce-hôte a été reconstruite pour chaque nœud interne. Le point d'interrogation marque un nœud interne inconnu, c'est-à-dire un nœud où la probabilité de l'espèce-hôte est inférieure à la valeur seuil.

Nous avons ensuite calculé la proportion de transitions au cours du temps, en additionnant le nombre de chaque type de transitions par an puis en le divisant par le nombre de transitions ayant lieu dans l'année. Nous avons considéré que la transition entre deux nœuds avait lieu au niveau du nœud parent, et nous avons ainsi daté ces transitions avec le package *castor* v.1.7.0 (102). Par conséquent, le nombre de transitions au cours du temps correspond à la somme des nœuds internes à un temps donné, multipliée par deux (car un nœud interne donne naissance à deux lignées). De plus, la transition d'un nœud interne à un nœud terminal (représentant une souche isolée) ne date pas de l'année d'échantillonnage de la souche mais de l'année où se situe le nœud interne. La transition d'un nœud interne à un nœud terminal isolé chez un bovin (blaireau) peut soit correspondre à une transition blaireau-bovin (bovin-blaireau), une transition inconnue (si le nœud qui précède est inconnu) ou une persistance intra-bovins (intra-blaireaux).

De même que pour l'arbre MCC, nous avons défini trois seuils de probabilité pour déterminer l'espèce-hôte des nœuds internes : 0,7, 0,8 et 0,9. Cependant, comme nous étudions ici 1004 arbres postérieurs et non pas l'arbre consensus, le MRCA d'un de ces 1004 arbres peut être plus ancien que le MRCA de l'arbre MCC. Par conséquent, la période de temps considérée est supérieure à la hauteur de l'arbre MCC.

2.3 RESULTATS

2.3.1 Données génomiques

Nous avons identifié 171 SNP à partir de 167 souches SB0821. Les distances génétiques entre deux paires de séquences variaient de 0 à 0,145 en considérant toutes les séquences, ainsi que les séquences isolées chez des bovins uniquement (médiane à 0,042) et enfin de 0 à 0,108 en considérant uniquement les souches isolées chez les blaireaux (médiane à 0,036). De plus, 37,1 % (62/167) des séquences isolées étaient uniques. Parmi ces 167 souches SB0821, 146 ont été isolées chez des bovins et 21 chez des blaireaux (*Figure 11*). Les souches SB0821 étaient détectées chez des bovins dès 2002, ce qui précède la première souche SB0821 isolée chez un blaireau (en 2013).

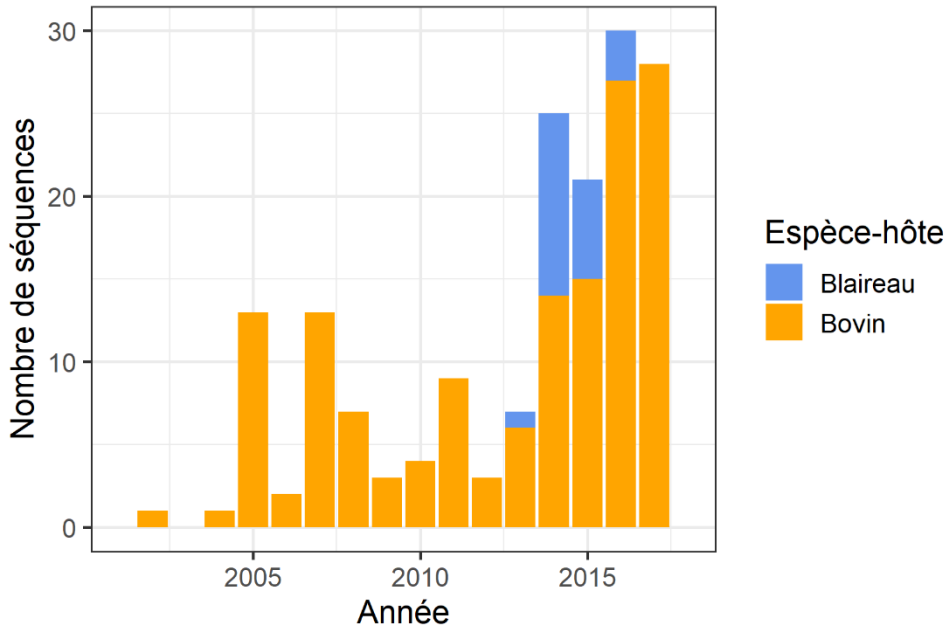


Figure 11 : Année d'échantillonnage et espèce-hôte (bovin en jaune et blaireau en bleu) des souches incluses dans l'étude.

2.3.2 Modèle d'évolution Bayésien

Nous avons sélectionné le modèle d'horloge strict à partir des comparaisons de BF (Annexe 1). Cependant, la comparaison des BF n'a pas permis de distinguer les modèles de substitution HKY et GTR. Nous avons donc choisi le modèle HKY déjà utilisé dans des études sur des séquences de *M. bovis* (93,94).

Tableau 2 : Paramètres estimés et taille effective d'échantillon (ESS) estimée pour chaque paramètre. ESS devait être strictement supérieure à 200.

Paramètre	Médiane	Borne inférieure 95%HPD	Borne supérieure 95%HPD	ESS
Hauteur d'arbre	27,5	21,0	36,6	7995
Taux d'horloge	$2,4 \times 10^{-3}$	$1,7 \times 10^{-3}$	$3,2 \times 10^{-3}$	8967
Taux de substitution = (Taux d'horloge * 171)	0,41	0,29	0,55	/
Kappa	5,9	4,2	8,2	8419
Taille effective de la population (Ne) de blaireaux	34	20	51	7777
Taille effective de la population (Ne) de bovins	1,2	0,27	2,7	7701
Taux de transition de bovin à blaireau	0,042	$7,0 \times 10^{-6}$	0,24	8736
Taux de transition de blaireau à bovin	2,2	0,79	4,7	8060

La médiane du paramètre kappa (ratio transition/transversion) a été estimée à 5,9 (95%HPD pour "High Posterior Density" ou intervalle de crédibilité : [4,2; 8,2]) (*Tableau 2*). Nous avons estimé un taux de substitution médian de 0,41 substitutions/génome/an (95%HPD : [0,29; 0,55]) et une hauteur d'arbre médiane de 27,5 ans (95%HPD : [21,0; 36,6]). Par conséquent, la date de circulation du MRCA était estimée à 1990 (95%HPD : [1980; 1996]). Le taux de transition de blaireau à bovin estimé (médiane de 2,2 transitions par lignée et par an ; 95%HPD : [0,74; 4,5]) était 52 fois supérieur au taux de transition de bovin à blaireau (médiane de 0,042 transitions/lignée/an ;

95%HPD : [3,5 x 10^{-5}; 0,24]). La taille effective de la population estimée pour les blaireaux (médiane de 34 ; 95%HPD : [20; 51]) était supérieure à celle estimée pour les bovins (médiane de 1,2 ; 95%HPD : [0,27; 2,7]) (*Tableau 2*).

Dans l'arbre MCC (*Figure 12* et Annexe *2*), 81 sur 166 des nœuds internes, dont la racine (avec une probabilité postérieure de l'espèce-hôte égale à 0,94) ainsi que les nœuds les plus proches de la racine, ont été identifiés comme étant hébergés par des blaireaux (55 avec une probabilité postérieure >0,9, 15 avec une probabilité postérieure entre 0,8 et 0,9 et 11 entre 0,7 et 0,8). Parmi les 85 nœuds internes restants, l'espèce-hôte identifiée était le bovin pour 57 nœuds (42 avec une probabilité postérieure >0,9, 7 avec une probabilité postérieure entre 0,8 et 0,9 et 8 entre 0,7 et 0,8) et inconnue pour 28 nœuds. La *Figure 13* montre l'évolution au cours du temps de l'espèce-hôte des lignées dans l'arbre MCC. L'espèce-hôte estimée pour toutes les lignées jusqu'en 1996, ainsi que pour 75 % des lignées par an jusqu'en 2000, est le blaireau. De plus, nous avons prédit deux pics de lignées bovines, une au milieu des années 2000 et l'autre au milieu des années 2010, et ces deux pics suivaient chacun un pic dans les lignées circulant chez des blaireaux.

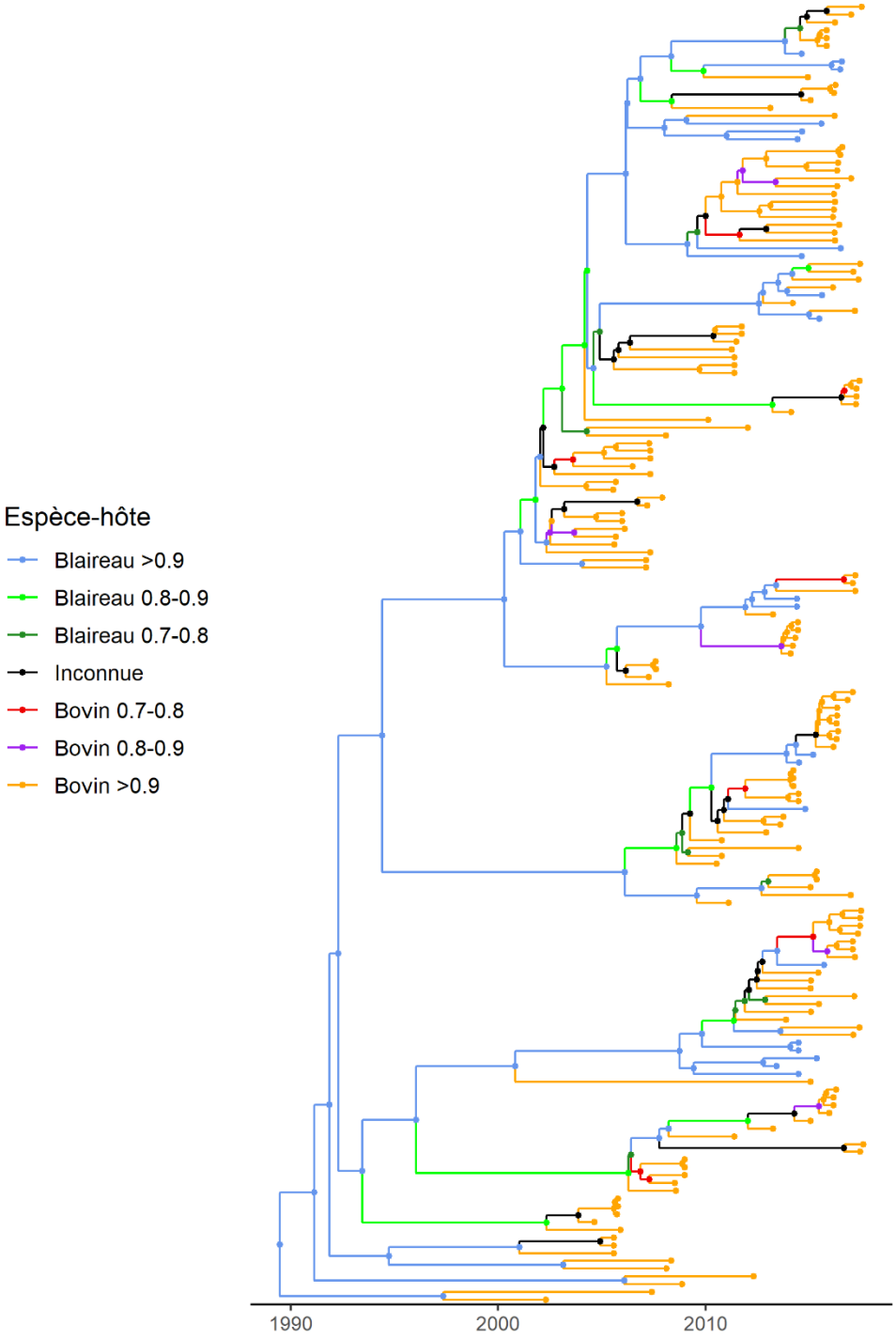


Figure 12 : Arbre consensus MCC où la couleur représente la probabilité de l'espèce-hôte.

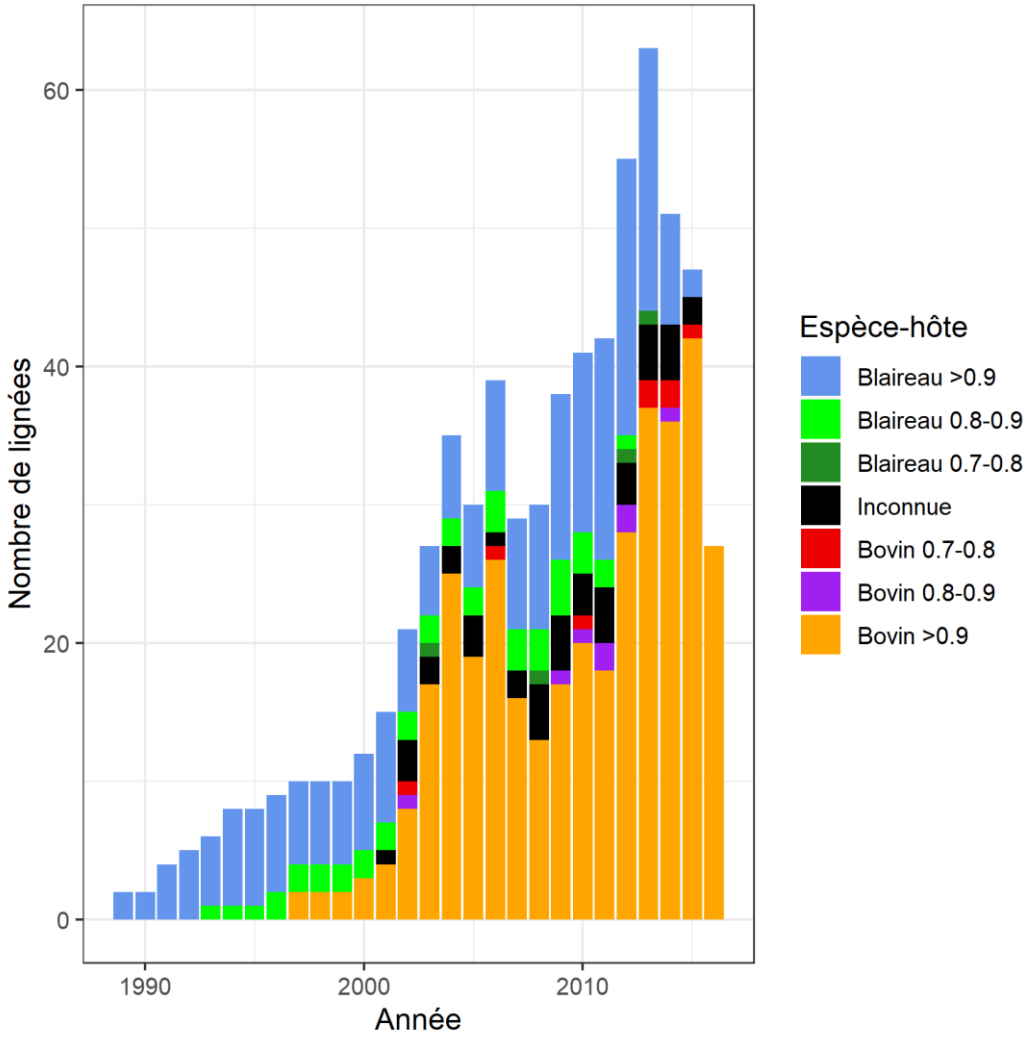


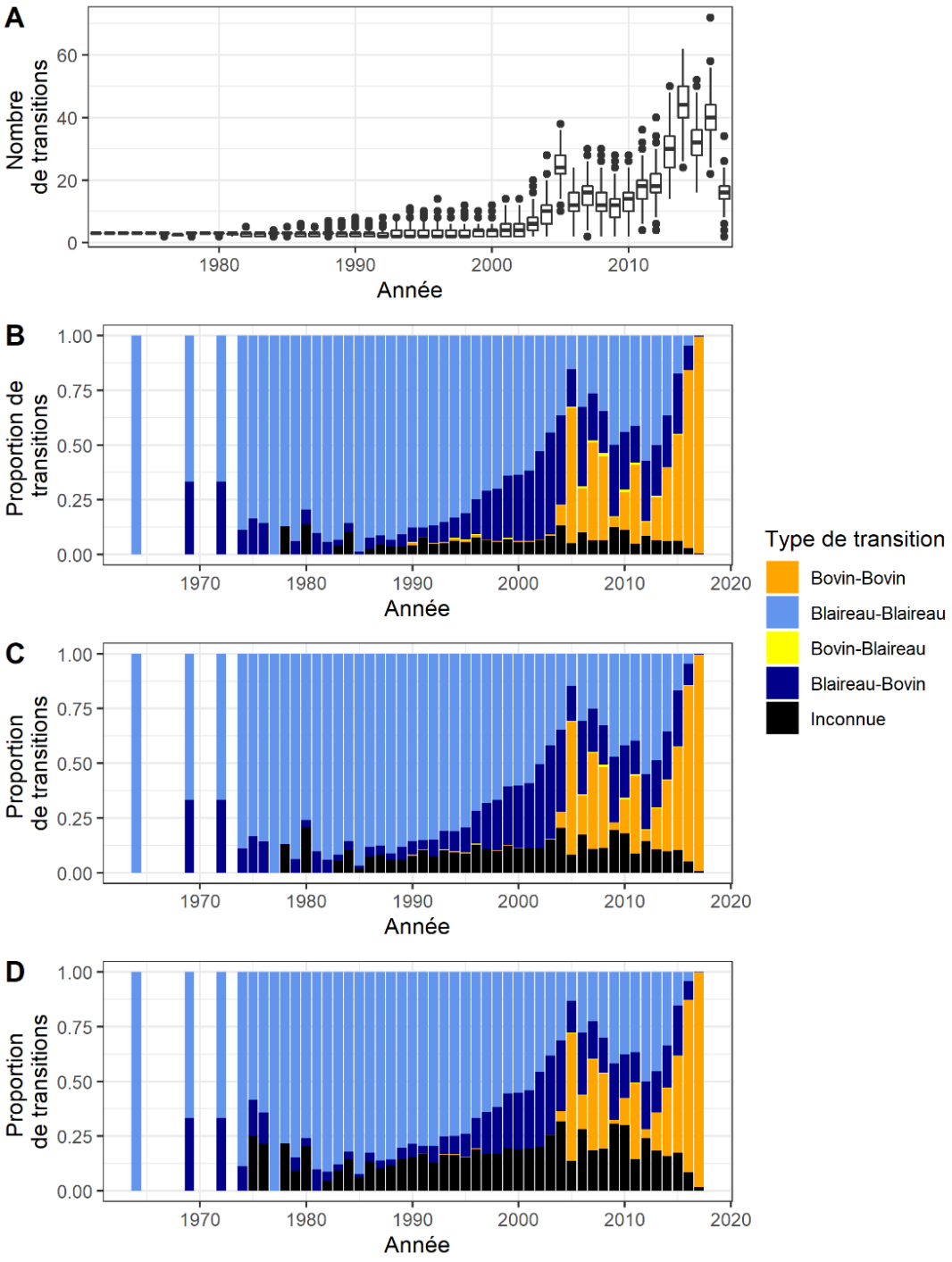
Figure 13 : L'espèce-hôte des lignées dans l'arbre MCC au cours du temps.

Parmi les 1004 arbres ré-échantillonnés, la hauteur d'arbre médiane estimée correspondait à un MRCA circulant en 1990 (1er quartile : 1978, 3ème quartile : 1996). De plus, quand nous avons considéré un seuil de probabilité de 0,9 dans ces arbres (*Tableau 3*), nous avons calculé une médiane de 64 évènements de transition de blaireau à bovin (1er quartile : 10, 3ème quartile : 91) et zéro transition de bovin à blaireau (1er quartile : 0, 3ème quartile : 3). Cependant, le nombre de fois où une lignée persistait dans la même espèce-hôte était similaire que l'on considère la population de blaireaux (médiane : 109, 1er quartile : 14, 3ème quartile : 137) ou celle de bovins (médiane : 112, 1er quartile : 78, 3ème quartile : 158). L'asymétrie entre le nombre de transitions inter-espèces et la similitude entre le nombre de persistances intra-espèces ont été observées quel que soit le seuil de probabilité considéré.

Tableau 3 : Nombre de transitions inter-espèces et de persistances intra-espèces calculé en fonction des seuils de probabilité (0,7, 0,8 et 0,9) parmi les 1004 arbres postérieurs ré-échantillonnés.

Type de transition	Seuil de probabilité	Médiane (1er quartile ; 3ème quartile)	Minimum-Maximum
Persistance intra-blaireaux	0,7	118 (54 ; 140)	11-150
	0,8	115 (38 ; 139)	9-149
	0,9	109 (14 ; 137)	2-148
Blaireau à bovin	0,7	77 (38 ; 97)	2-106
	0,8	72 (26 ; 95)	1-104
	0,9	64 (10 ; 91)	0-103
Bovin à blaireau	0,7	1 (0 ; 8)	0-12
	0,8	1 (0 ; 5)	0-9
	0,9	0 (0 ; 3)	0-7
Persistance intra-bovins	0,7	120 (86 ; 176)	68-273
	0,8	116 (81 ; 169)	66-256
	0,9	112 (78 ; 158)	64-241

La *Figure 14* montre l'évolution au cours du temps du type de transitions en fonction des seuils de probabilités (0,7, 0,8 et 0,9) dans les 1004 arbres ré-échantillonnés. La proportion de persistances au sein de la population de blaireaux constituait plus de 50 % des transitions jusqu'en 2001 (sauf en 1974, où 50 % des transitions dans les lignées étaient inconnues avec un seuil de probabilité de 0,9) alors que la persistance intra-bovins a commencé au plus tôt en 1990. La persistance intra-bovins représentait plus de 50 % des transitions en 2005 puis en 2016 et 2017, ce qui correspond aux deux pics de lignées bovines dans l'arbre MCC. De plus, la proportion de transitions bovin à blaireau n'a jamais excédé 1,3 % des transitions.



<u>Figure 14 :</u> Evolution au cours du temps du nombre (A) ainsi que de la proportion des types de transitions avec un seuil de probabilité de 0,7 (B), 0,8 (C) et 0,9 (D) dans les 1004 arbres ré-échantillonnés.

2.4 DISCUSSION

Dans ce travail, nous avons reconstruit un arbre phylogénétique consensus à partir de données de séquençage total, et ce, afin de mieux comprendre le rôle joué par les blaireaux dans la transmission de souches *M. bovis* SB0821 dans les PA et les Landes. Cette région située dans le Sud-Ouest de la France présente un intérêt majeur dans le contrôle de la bTB en France, du fait de la présence de foyers persistants ces 10 dernières années (d'après les données de surveillance disponibles sur la plateforme ESA).

2.4.1 Estimation du taux de substitution

En ce qui concerne le modèle d'évolution Bayésien, nous avons sélectionné une horloge moléculaire stricte. Nous avons également choisi un modèle de substitution HKY plutôt que GTR en se basant sur des études précédentes de *M. bovis* (93,94). Le taux de substitution que nous avons estimé de 0,41 substitutions par génome et par an (95%HPD : [0,29 ; 0,55]) est supérieur à ceux estimés dans des études précédentes en Irlande du Nord (0,15, 95%HPD : [0,04 ; 0,26] (103) et 0,2, 95%HPD : [0,1 ; 0,3] (104)) et au Michigan, dans les Etats-Unis (0,2, 95%HPD : [0,1 ; 0,3] (94)). Cependant, en 2017, Crispell *et al.* ont estimé un taux supérieur au nôtre, à savoir 0,53 substitutions par génome et par an, 95%HPD : [0,22 ; 0,94] (93).

Les différences observées entre ces études pourraient être dues à la lignée de *M. bovis* considérée. En effet, des caractéristiques de lignées ont été mises en évidence pour *M. tuberculosis* (105). Les souches de *M. bovis* étudiées par Biek *et al.* et Trewby *et al.* font partie du complexe clonal Eu1 (106). En France, les lignées *M. bovis* dépendent de la région étudiée (84). Nos estimations ici sont basées sur des souches SB0821 qui appartiennent à la famille F4/cluster A (86). De plus, les études sur des souches *M. bovis* qui ne partagent pas le même spoligotype ni profil VNTR (93) pourraient estimer un taux de substitution plus élevé. La variation entre les estimations pourrait aussi dépendre des espèces-hôtes échantillonnées. Les espèces sauvages étudiées variaient selon la région étudiée, le cerf de Virginie et l'élan au Michigan (94), le phalanger-renard en Nouvelle-Zélande (93) ou encore le blaireau en Irlande du Nord (103,104).

2.4.2 Estimation des taux de transitions inter-espèces

La méthode d'inférence Bayésienne que nous avons choisie est une méthode de coalescence structurée. Ce choix permettait d'atténuer le biais d'échantillonnage dans l'estimation des taux de transitions inter-espèces.

Plus précisément, nous avons utilisé l'approximation marginale de la coalescence structurée (Mascot) (96) pour modéliser l'évolution de souches SB0821 dans deux sous-populations : les blaireaux et les bovins.

Dans cette méthode, l'hypothèse de la taille effective de population constante au cours du temps est importante à considérer. Dans notre région d'étude, les dynamiques de population de *M. bovis* jusqu'en 2017 sont difficiles à estimer du fait de la mise en place tardive de la surveillance de la faune sauvage par rapport à celle des bovins. De plus, la découverte de la bTB dans la faune sauvage a provoqué une augmentation de la surveillance chez les bovins et de fait, une augmentation de la détection de bTB. Cependant, la prévalence apparente chez les blaireaux n'a pas eu l'air de varier de manière significative entre les deux estimations (2013-2014 et 2016-2017) de Réveillaud *et al.* (66). De même, l'évolution du nombre d'élevages infectés n'a pas montré de tendance particulière sur la même période de temps (Boschiroli, communication personnelle).

Des distances génétiques similaires ont été observées entre les séquences isolées chez les bovins et celles isolées chez les blaireaux. Etant donné le fait que ces distances ont été estimées sur 21 séquences isolées chez des blaireaux et 146 séquences isolées chez des bovins, les séquences isolées chez les blaireaux présentaient donc une plus grande diversité génétique. Ceci est en accord avec le fait que la taille effective estimée de la population de blaireaux est supérieure à celle estimée chez les bovins.

Nous avons inféré un taux de transition de blaireau à bovin 52 fois plus grand que le taux de transition de bovin à blaireau. Nos résultats sont en accord avec ceux d'une étude de Crispell *et al.* datant de 2019. En effet, ils avaient estimé sur un sous-échantillon de 83 souches isolées chez des bovins et 97 souches isolées chez des blaireaux au Royaume-Uni, un taux de transition de blaireau à bovin (0,045 transitions/lignée/an) 10 fois supérieur au taux de transition de bovin à blaireau (0,0044 transitions/lignée/an) (88). A l'inverse, Rossi *et al.* ont estimé un taux de transition de bovin à blaireau supérieur à celui de blaireau vers bovin dans une région nouvellement infectée dans le Nord-Ouest de l'Angleterre. Rossi *et al.* en ont conclu que l'infection doit être bien installée chez les blaireaux avant que les transmissions de blaireau à bovin ne deviennent probables (107).

La densité de blaireaux dans une région donnée pourrait avoir une influence sur les dynamiques de transmission inter-espèces. La densité moyenne de blaireaux estimée dans 13 zones d'étude en France (dont 50 km^2 situés dans notre zone d'étude) était de 3,8 blaireaux par km^2 (plage de

valeurs : 1,7-7,9), ce qui est "relativement plus faible que celles trouvées au Royaume-Uni et concorde avec les estimations globales réalisées en Irlande" (108).

2.4.3 Estimation du nombre de transitions au cours du temps

En ce qui concerne les évènements de transmission inter-espèces, Crispell *et al.* ont eu des résultats semblables à ce que nous avons obtenu, à savoir une majorité de transmissions de blaireau à bovin et une médiane de zéro transmission de bovin à blaireau en Angleterre. De plus, l'évolution des espèces-hôtes des lignées au cours du temps dans notre arbre consensus a montré que l'augmentation des lignées de bovins suivait une augmentation de lignées de blaireaux. Ces résultats suggèrent que la transmission de blaireau à bovin pourrait être amplifiée à son tour par des transmissions entre bovins, une hypothèse proposée par Donnelly et Nouvellet sur des données provenant du Royaume-Uni (109). De plus, une analyse d'un réseau de contacts empirique d'élevages bovins dans la même région a mis en évidence l'importance des contacts médiés par les blaireaux dans la diffusion de la bTB (90).

L'étude de Crispell *et al.* en 2019 avait une approche intéressante dans l'estimation du nombre minimum d'évènements de transmission intra- et inter-espèces à partir d'un arbre phylogénétique dont l'état des nœuds ancestraux était reconstitué (88). Ils ont émis l'hypothèse qu'un évènement de coalescence correspondait à au moins un évènement de transmission. Cela signifie qu'implicitement l'hypothèse de la présence d'une seule lignée de pathogènes au sein d'un hôte infecté était faite. Ici, nous n'avons fait aucune hypothèse quant à l'évolution intra-hôte ni sur les temps de transmission. Nous avons considéré qu'une transition inter-espèces correspondait à au moins un évènement de transmission entre un bovin et un blaireau, mais nous n'avons pas estimé le nombre d'évènements de transmission intra-espèces. Toutefois, le fait que la majorité des transitions soit des persistances suggère l'importance des évènements de transmission intra-espèces qui avait été mise en avant dans le travail de Crispell *et al.* (88). En effet, la majorité des transitions jusqu'à 2001 a été identifiée comme des persistances intra-blaireaux. Ceci peut correspondre soit à l'évolution de lignées de *M. bovis* au sein d'un blaireau ou à de la transmission entre blaireaux appartenant au même groupe social (terrier) ou à des terriers voisins. Cependant, Bouchez-Zacria *et al.* en utilisant un modèle stochastique de transmission de *M. bovis* au sein du système bovins-blaireaux dans les PA et les Landes, ont ainsi déterminé que la transmission de bTB était limitée entre groupes sociaux de blaireaux (110). De fait, la majorité de la persistance

intra-blaireaux jusqu'en 2001 suggère soit le portage à long-terme de bTB dans un blaireau et/ou une longue histoire de transmission au sein d'un groupe social.

De même, la persistance intra-bovins peut soit correspondre à l'évolution de lignées *M. bovis* au sein d'un bovin ou à de la transmission intra- ou inter-élevages. Les données de mouvements de bovins obtenues grâce à la BDNI ont permis de déterminer leur rôle significatif dans la transmission de la bTB dans les PA et les Landes (90). Ces mouvements de bovins pourraient être à l'origine de transmissions inter-élevages. Cependant, la modélisation de la transmission dans ce système bovins-blaireaux a permis de déterminer que 49,3 % des contaminations d'élevages étaient dues à la proximité de pâtures appartenant à un élevage infecté (110).

2.4.4 Limites et perspectives

Les bovins et les blaireaux ne sont pas les seules sources de contamination possibles pour un élevage. Dans les deux modèles précédemment évoqués (90,110) ainsi que dans notre travail, la contribution d'autres espèces-hôtes à la transmission de bTB n'a pas pu être incluse. Cependant, dans les PA et les Landes, « Sylvatub » a détecté depuis son implémentation des sangliers infectés par *M. bovis*. De plus, une étude sur les mouvements de blaireaux en Europe a estimé que la distance moyenne parcourue par des blaireaux était de 1,7 km avec de rares déplacements atteignant les 22 km (111) alors que la distance moyenne parcourue par des sangliers en une journée est estimée à 7-13 km (112). Les sangliers pourraient donc contribuer plus aisément à la diffusion géographique de la bTB que les blaireaux, qui se déplacent généralement sur de plus courtes distances.

Dans notre étude, nous avons utilisé des données génomiques obtenues chez des bovins et des blaireaux. Nous avons vu que ces deux espèces ne sont pas surveillées selon les mêmes modalités. De manière plus générale, pour la bTB, nous avons affaire à un biais d'échantillonnage entre la faune sauvage (surveillance récente et partielle) et les bovins (surveillance ancienne et quasi-exhaustive). De plus, même si la collecte de carcasses de blaireaux ne dépend pas de la présence de bTB dans les élevages, les protocoles de piégeage évoluent au cours du temps et dans l'espace selon la détection de bovins infectés dans les élevages alentours. Le choix de la méthode Mascot était motivé par le fait que cette méthode ne considère pas le nombre d'échantillons de chaque espèce-hôte comme une donnée informative et aide donc à réduire l'impact de ce biais d'échantillonnage (37).

2.4.5 Conclusion

Ce modèle d'évolution Bayésien nous a permis d'inférer des transitions inter-espèces et non intra-espèces contrairement aux études épidémiologiques précédentes, qui considéraient les élevages et les groupes sociaux de blaireaux comme unités épidémiologiques. Nous n'avons pas pu confirmer les groupes sociaux de blaireaux comme de possibles intermédiaires dans la transmission de bTB entre élevages. Cependant, nos résultats ont mis en évidence la présence à long-terme de *M. bovis* au sein de la population de blaireaux et un taux élevé de transitions de blaireau à bovin dans les PA et les Landes. Ceci justifie les mesures de contrôle visant à prévenir les contacts entre les bovins et les blaireaux. L'inclusion de données d'autres espèces-hôtes comme des séquences isolées chez des sangliers (en nombre trop faible lors de la réalisation de ce travail) pourraient améliorer notre compréhension de ce système. La reconstruction d'arbres de transmission permettrait d'inclure non seulement ces autres espèces-hôtes mais également d'explorer les dynamiques de transmission intra-espèces.

3 LES METHODES DE RECONSTRUCTION D'ARBRES DE TRANSMISSION COMBINANT DES DONNEES EPIDEMIOLOGIQUES ET GENOMIQUES

3.1 INTRODUCTION

Avant de reconstruire un arbre de transmission, il est nécessaire de comprendre et choisir une méthode de reconstruction. Une possibilité serait de reconstruire tous les arbres possibles et de sélectionner le plus probable. D'après la théorie des graphes, le nombre d'arbres couvrant de poids minimal (arbre qui connecte tous les sommets, ici les hôtes, et dont le poids des arêtes, ici les liens de transmission, est minimale) qui peuvent être construits pour un ensemble de n nœuds est donné par la formule de Cayley : n^{n-2} où n est le nombre d'hôtes (113). Appliqué aux arbres de transmissions, ce nombre correspond au nombre d'arbres de transmission non enracinés compatibles avec n hôtes infectés. Si par exemple, nous considérons 10 hôtes infectés, cela correspondrait à $10^{10-2}=10^8$ arbres de transmission non enracinés. Il n'est donc pas envisageable d'énumérer tous les arbres de transmission enracinés et orientés quand n est élevé, et d'autres stratégies doivent être utilisées.

Les méthodes de reconstruction d'arbres de transmission qui combinent des données génomiques et épidémiologiques peuvent modéliser quatre processus non observés mentionnés par Klinkenberg *et al.* (114) : la mutation, l'évolution intra-hôte, la transmission et l'observation des cas. Le processus de mutation inclut la délétion ou l'insertion d'un nucléotide (disparition ou addition d'un nucléotide dans la séquence) ainsi que la substitution déjà évoquée dans la reconstruction d'arbres phylogénétiques. L'évolution intra-hôte représente la manière dont le génome du pathogène change au sein d'un individu ou d'un groupe d'individus, conduisant à la diversification du génome. La transmission correspond au passage de l'agent pathogène d'un hôte infecté à un hôte susceptible puis au processus d'infection dans l'hôte nouvellement infecté. Dans les modèles de transmission, des hypothèses sont donc faites à propos de l'introduction de l'agent pathogène au sein de la population d'hôtes, sa diffusion d'un hôte à un autre et l'histoire naturelle de la maladie. L'observation de cas correspond à l'identification et l'échantillonnage d'hôtes infectés dans la population.

Nous avons réalisé une revue systématique de la littérature dans le but d'identifier des méthodes qui combinent des données génomiques et épidémiologiques dans la reconstruction d'arbres de transmission. Dans

l'optique d'aider à sélectionner une méthode adaptée, nous avons analysé ces méthodes selon les données nécessaires à leur implémentation ainsi que selon les modèles sous-jacents de mutation, évolution intra-hôte, transmission et observation des cas.

3.2 MATERIEL ET METHODE

3.2.1 Stratégie de recherche

Nous avons utilisé la méthode PRISMA et avons tout d'abord parcouru deux bases de données électroniques, Pubmed et Scopus, du 13 octobre au 17 novembre 2020. Nous avons choisi des mots clés à propos d'arbres de transmission ("transmission chain", "transmission tree", "transmission reconstruction", "transmission network", "who infected whom") et faisant référence à l'utilisation de données génomiques ("genome", "SNP", "genetic data", "phylogenetic data"). La requête de recherche était la suivante : ("transmission chain" OR "transmission tree" OR "transmission reconstruction" OR "transmission network" OR "who infected whom") AND ("genome" OR "genomic" OR "sequence data" OR "genetic data" OR "phylo* data"). Selon la base de données, cette requête était entrée dans "all fields" (Pubmed) ou dans « Title, abstract, or author-specified keywords » (Scopus). Dans la base de données qui ne supportait pas le *, « phylo* data » était remplacé par « phylogenetic data ». Après sélection des articles qui nous intéressaient, nous avons regardé leur liste des références afin de trouver des études supplémentaires correspondant à nos critères de recherche.

3.2.2 Critères d'éligibilité

Les études étaient incluses quand elles inféraient un arbre de transmission pour l'épidémie d'une maladie infectieuse à partir de données épidémiologiques et génomiques non simulées. Les données génomiques devaient correspondre à des SNP identifiés à partir de séquences consensus ou des séquences consensus de gènes entiers, et non des données de « deep sequencing » qui considèrent de multiples nucléotides pour un seul locus. Nous avons défini un arbre de transmission comme étant un graphe enraciné constitué de nœuds (représentant des cas, c'est-à-dire un individu ou un groupe d'individus infecté) connectés par des flèches (représentant des évènements de transmission). Les arbres de transmission reconstruits avec un seul type de données ont été exclus. Les méthodes qui estimaient les arbres de transmission compatibles avec les données épidémiologiques, séparément de l'estimation des arbres de transmission compatibles avec les données génomiques ont également été exclues. Même si ces méthodes

combinaient graphiquement les évènements de transmission ou comparaient les résultats obtenus par les deux types de données, lorsqu'aucun algorithme ne liait les deux types de données, nous avons considéré que ces méthodes ne combinaient pas formellement les données épidémiologiques et génomiques.

3.2.3 Gestion des données

Les citations ont été exportées depuis les deux bases de données électroniques dans EndNote X9 (2018), où nous avons retiré les doublons et sélectionné les articles sur la base des titres, résumés et quand cela était nécessaire afin de prendre une décision, du matériel et des méthodes. Nous avons ensuite étudié le texte intégral des articles sélectionnés afin d'évaluer leur éligibilité, nous avons procédé par ordre chronologique dans le but de mieux comprendre comment les méthodes sont liées entre elles.

3.2.4 Collecte des données

Nous avons répertorié la méthode d'inférence utilisée (par exemple, Bayésienne ou maximum de vraisemblance) ainsi que les limites d'une méthode de reconstruction quand elles étaient discutées dans un article. Etant donné le fait que la diversité génétique affecte la reconstruction de l'histoire de transmission (55), nous avons cherché de manière systématique les informations suivantes à propos des données génomiques : l'agent pathogène concerné et son taux de mutation, la période de temps considérée, le nombre de séquences génétiques étudiées ainsi que le nombre de SNP ou la taille des séquences utilisées dans la reconstruction des arbres. Quand le taux de mutation d'un agent pathogène n'était pas estimé ni mentionné dans l'article en question, nous avons réalisé une recherche bibliographique afin de trouver cette information.

Nous avons répertorié l'unité épidémiologique considérée : un individu *vs.* un groupe d'individus. Cette information est importante car selon l'unité épidémiologique, l'évolution intra-hôte peut correspondre à l'évolution du pathogène au sein d'un individu ou à son évolution au sein d'un groupe d'individus. Dans ce dernier cas, l'évolution intra-hôte intègre également les dynamiques de transmission entre individus au sein du groupe qui constitue une seule unité épidémiologique. Nous étions également intéressés par le type de données épidémiologiques nécessaires à l'implémentation de la méthode ainsi que le temps de calcul quand il était disponible, ces informations permettant le choix d'une méthode par rapport à des considérations plus pratiques. Les types de données épidémiologiques

incluaient la date de début d'exposition, la date de début d'infectiosité, la date d'échantillonnage, la date de retrait, les données de contact et géographiques, ainsi que des caractéristiques intrinsèques pouvant influencer l'infectiosité ou la susceptibilité. Par exemple, l'espèce-hôte prédominante est une caractéristique intrinsèque d'un élevage qui pourrait être intéressante à inclure dans un modèle de transmission de la fièvre aphteuse (FA) (39). En effet, les porcins sont plus excréteurs que les ruminants, qui sont à leur tour plus susceptibles. La diffusion par aérosols du virus de la FA s'effectue donc plus probablement des porcins vers les bovins et ovins (115).

Enfin, la modélisation explicite ou non des quatre processus non observés (1. mutation, 2. évolution intra-hôte, 3. transmission et 4. observation des cas) a été analysée.

1. Les modèles de substitutions (par exemple le modèle de Kimura (116) ou de Jukes-Cantor (23)) sont souvent utilisés pour décrire la mutation de séquences. Nous avons répertorié le type de modèle de substitution utilisé dans chaque méthode.

2. L'évolution intra-hôte peut être modélisée par les modèles de populations (par exemple le modèle coalescent (28) évoqué précédemment) souvent utilisés pour décrire les liens de parenté entre les agents pathogènes échantillonnés dans la reconstruction d'arbres phylogénétiques.

3. Trois sous-catégories peuvent être considérées pour décrire le modèle de transmission.

   Tout d'abord, comme l'infectiosité d'un individu varie au cours du temps, en fonction de l'excrétion du pathogène (117), les modèles de transmission peuvent prendre en compte différents stades, en fonction du potentiel de transmission, pour une maladie infectieuse. De plus, des paramètres comme la période de latence ou le temps de génération peuvent être fixés en amont, ou bien être estimés lors de l'inférence. La période de latence correspond à la durée entre la date d'infection d'un hôte par un pathogène et le début de la période d'infectiosité. Cette période de latence est suivie par une période infectieuse pendant laquelle l'hôte peut transmettre le pathogène aux autres (118). Le temps de génération (T_g) correspond à l'intervalle de temps entre l'infection d'un cas et l'infection de cas secondaires par ce même cas. Ce T_g dépend donc des périodes infectieuses et de

latence, mais également de la variation de l'infectiosité de l'hôte au cours du temps (119). Pour chaque méthode, nous avons donc identifié les différents états considérés pour un hôte (par exemple, S pour susceptible, E pour exposé, I pour infectieux et R pour retiré) et si des périodes infectieuses et de latence ou des T_g étaient considérés pour modéliser l'histoire naturelle de la maladie.

De plus, un évènement de transmission résultant d'un contact direct ou indirect entre un hôte infectieux et un hôte susceptible, ce contact peut être modélisé en faisant l'hypothèse que tous les individus ont la même probabilité d'entrer en contact (contacts homogènes), en considérant que la probabilité de transmission est fonction de la distance géographique (noyau de transmission spatial) ou en prenant en compte des données de contact. Nous nous sommes ainsi intéressés à la façon dont les contacts entre hôtes étaient modélisés (contacts homogènes, noyau de transmission spatial ou données de contact).

Enfin, nous avons répertorié si une seule ou de multiples introductions de l'agent pathogène dans la population étai(en)t considérée(s) comme à l'origine de l'épidémie étudiée.

4. En ce qui concerne l'observation des cas, nous nous sommes intéressés à la façon dont ces méthodes prenaient en compte la détection imparfaite des cas, si elles avaient besoin que tous les cas observés soient échantillonnés ou si l'existence de données génomiques manquantes pouvait être prise en compte.

3.3 RESULTATS

Après retrait des doublons, 496 articles ont été importés sur EndNote puis nous avons évalué leur pertinence par rapport aux méthodes de reconstruction d'arbres de transmission. Le texte intégral de 98 parmi ces 496 articles a été évalué (Annexe *3*). Les causes d'exclusions d'articles, pour lesquels nous avons évalué le texte intégral, sont détaillées en Annexe *4*. Les principales causes étaient les suivantes : pas d'inférence d'un arbre de transmission selon la définition donnée dans notre matériel et méthodes (n=23), le type de données génétiques considérées (n=12), l'absence de combinaison formelle des deux types de données (n=21).

Vingt-deux méthodes ont été identifiées dans les 41 articles restants, que nous avons regroupées en trois familles : une famille non

phylogénétique, une famille phylogénétique séquentielle et une famille phylogénétique simultanée. Dans la famille non phylogénétique (notée par la suite NPF), les distances génétiques entre chaque paire de séquences étaient estimées mais les arbres phylogénétiques n'étaient pas pris en compte dans la reconstruction des arbres de transmission. A l'inverse, dans les familles phylogénétiques (notées par la suite PF), les arbres de transmission étaient inférés à partir d'arbres phylogénétiques. Cette inférence était réalisée, soit en considérant les arbres phylogénétiques comme une source d'information, soit grâce à un lien établi entre les deux types d'arbres. L'étape de construction de ce lien entre les deux types d'arbres consistait à inférer l'hôte de chaque nœud ou chaque branche dans l'arbre phylogénétique. Dans la famille phylogénétique séquentielle (notée par la suite SeqPF), les arbres phylogénétiques devaient être reconstruits en amont de l'implémentation de la méthode de reconstruction d'arbres de transmission. Dans la famille phylogénétique simultanée (notée par la suite SimPF), les arbres phylogénétiques et de transmission étaient inférés simultanément. Nous avons décidé de faire cette distinction entre les deux familles phylogénétiques car l'approche de la famille phylogénétique séquentielle implique le choix d'une méthode appropriée de reconstruction d'arbres phylogénétiques puis son implémentation en amont de la méthode de reconstruction d'arbres de transmission. La famille phylogénétique séquentielle fait ainsi l'hypothèse que l'arbre phylogénétique ne dépend pas de la transmission.

La *Figure 15* illustre le problème que ces trois familles tentent de résoudre. Prenons l'exemple d'un scénario simple de transmission et d'évolution intra-hôte : D infecte U (un hôte non observé), qui à son tour infecte C et A puis enfin, C infecte B. Dans cette *Figure 15*, la longueur de chaque rectangle représentant un hôte correspond à la durée entre l'infection et le retrait (mort ou guérison) de l'hôte. Considérons les séquences a, b, c et d isolées respectivement chez les hôtes A, B, C et D aux temps T_A, T_B, T_C et T_D. Les temps de retrait des hôtes observés sont également inclus dans les données : R_A, R_B, R_C et R_D. A partir de ces données épidémiologiques et des distances génétiques (NPF) ou de l'arbre phylogénétique (*Figure 15*.B, PF), chaque famille de méthodes cherche à reconstruire l'arbre de transmission (*Figure 15*.C), voire à inférer les temps d'infection t[infecteur, infecté].

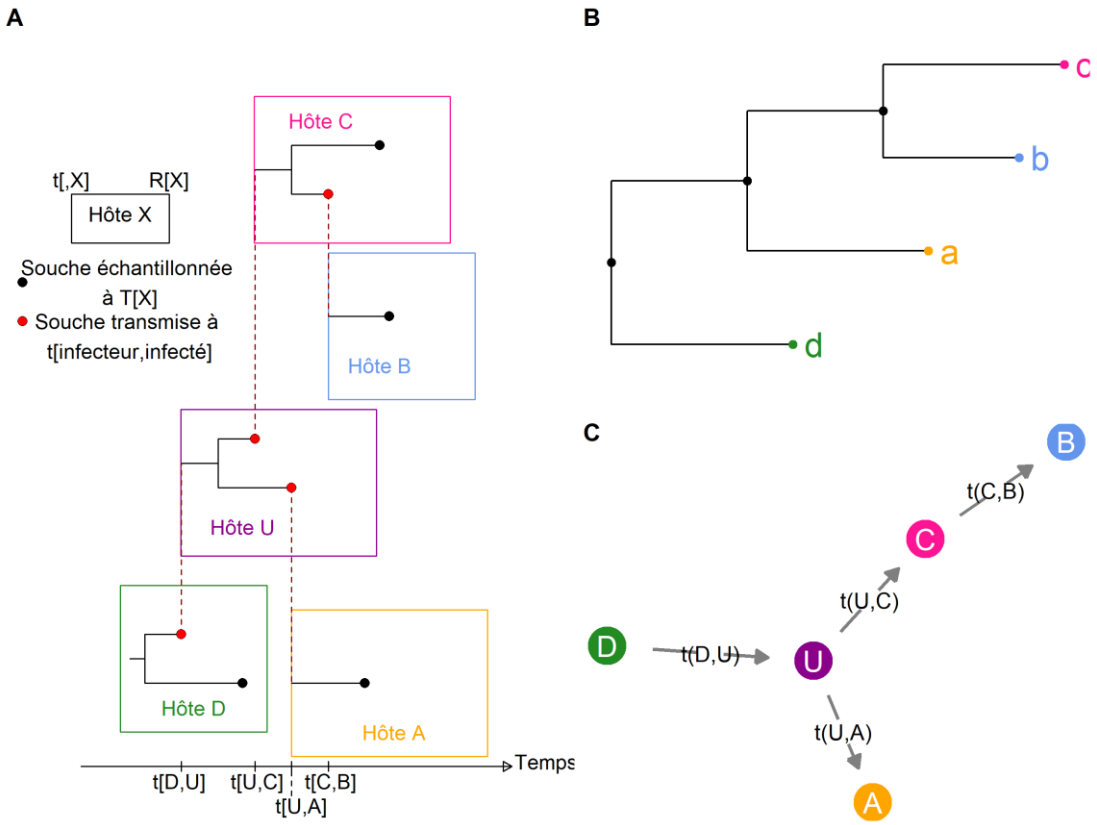


Figure 15 : Transmission d'un agent pathogène et évolution intra-hôte au sein de cinq hôtes (A), l'arbre phylogénétique reconstruit (B) et arbre de transmission inféré (C) à partir des quatre (a, b, c, d) séquences échantillonnées. L'hôte U n'est pas échantillonné. La longueur du rectangle représentant un hôte X correspond à sa durée d'infection (infection à t[,X] jusqu'au retrait à R[X]). Les pointillés rouges correspondent à des évènements de transmission.

Le *Tableau 4* présente les données épidémiologiques et génomiques nécessaires à l'implémentation de chaque méthode. La majorité des méthodes (20/22) avait au moins besoin des dates d'échantillonnages. Onze méthodes prenaient en compte les dates de retrait, sept le début d'infectiosité alors que peu (3/22) prenaient en compte le début d'exposition. De plus, seules deux méthodes (appartenant aux familles NPF et SimPF) respectivement Aldrin 2011 (120) et *BORIS* (pour « Bayesian Outbreak Reconstruction Inference and Simulation ») (39) prenaient en compte les caractéristiques intrinsèques (espèce-hôte prédominante, nombre d'animaux, période de production). Parallèlement, seules deux autres méthodes incluaient des données de contact dans leur modèle de transmission : une dans la NPF, *outbreaker2* (121), l'autre dans la SeqPF, *TiTUS* (pour « Transmission Tree Uniform Sampler ») (122). Dix méthodes sur les 22

sélectionnées étaient implémentées dans des packages, dont sept dans des packages R (cependant, depuis leur publication, deux ont été retirées du CRAN, *bitrugs* et *phybreak*), une méthode avait son code disponible sur Github (TiTUS) et les deux méthodes restantes étaient disponibles sur BEAST (123) (*beastlier*) ou BEAST2 (95) (« Structured Coalescent Transmission Tree Inference » ou *SCOTTI*).

L'évolution intra-hôte était explicitement modélisée dans moins de la moitié des méthodes (8/22) et la plupart des méthodes reposaient sur des hypothèses très contraignantes : tous les cas sont observés et échantillonnés, le bottleneck de transmission est complet (une seule souche transmise) ou encore, une seule introduction de la maladie a eu lieu dans la population d'hôtes (*Tableau 5* à 7). Dans les tableaux, l'observation correspond à la détection d'un hôte infecté. De plus, quand une séquence du pathogène a été isolée chez un hôte, il est dit échantillonné.

3.3.1 Famille non-phylogénétique

La famille non-phylogénétique (*Figure 16*) contient huit méthodes. La majorité de ces méthodes (5/8) attachaient à un modèle explicite de transmission, un modèle génétique qui décrivait la distance génétique entre deux hôtes, en fonction de leur position relative dans l'arbre de transmission. Dans les méthodes Bayésiennes (4/5), ces modèles étaient combinés dans une fonction de vraisemblance qui permettait d'explorer l'espace des arbres de transmission.

Tableau 4 : Données nécessaires à l'implémentation de chaque méthode en fonction de la famille. A ou S signifie arbre phylogénétique ou séquences génétiques. Des packages sont disponibles pour les méthodes en gras. La date de retrait correspond à la fin d'infectiosité (abattage ou fin d'hospitalisation) et les caractéristiques intrinsèques étaient soit le nombre d'individus présents sur site ou l'espèce prédominante.

Famille	Méthode (*nom*)	Début d'exposition	Début d'infectiosité	Date d'échantillonnage	Date de retrait	Données de contact	Données géographiques	Caractéristiques intrinsèques	A ou S
NPF	Aldrin *et al.*, 2011		X		X		X	X	S
	Jombart *et al.*, 2011 (*seqTrack*)			X					S
	Ypma *et al.*, 2012		X		X		X		S
	Jombart *et al.*, 2014 (*outbreaker*)			X					S
	Worby *et al.*, 2014			X					S
	Famulare *et al.*, 2015			X					S
	Worby *et al.*, 2016 (*bitrugs*)	X		X	X				S
	Campbell *et al.*, 2019 (*outbreaker2*)			X		X			S
SeqPF	Cottam *et al.*, 2008		X	X	X				A
	Didelot *et al.*, 2014			X			(X)		A
	Eldholm *et al.*, 2016			X					A
	Didelot *et al.*, 2017 (*TransPhylo*)			X					A
	Sashittal *et al.*, 2020 (*TiTUS*)	X		X	X	X			A

Tableau 4 (suite) : Données nécessaires à l'implémentation de chaque méthode en fonction de la famille. A ou S signifie arbre phylogénétique ou séquences génétiques. Des packages sont disponibles pour les méthodes en gras. La date de retrait correspond à la fin d'infectiosité (abattage ou fin d'hospitalisation) et les caractéristiques intrinsèques étaient soit le nombre d'individus présents sur site ou l'espèce prédominante. Didelot *et al.* ont pénalisé les arbres de transmission en fonction des données géographiques (56) mais ces données n'étaient pas incluses dans l'inférence de ces arbres d'où la présence de parenthèses. Dans *beastlier* (124), des données de contact peuvent être incluses mais des données géographiques ont été utilisées à la place.

Famille		Méthode (*nom*)	Début d'exposition	Début d'infectiosité	Date d'échantillonnage	Date de retrait	Données de contact	Données géographiques	Caractéristiques intrinsèques	A ou S
SimPF	Explicite	Ypma *et al.*, 2013		X	X	X		X		S
		Hall *et al.*, 2015 (*beastlier*)			X	X	(X)	X		S
		De Maio *et al.*, 2016 (*SCOTTI*)	X		X	X				S
		Klinkenberg *et al.*, 2017 (*phybreak*)			X					S
	Implicite	Morelli *et al.*, 2012		X	X	X		X		S
		Mollentze *et al.*, 2014			X			X		S
		Lau *et al.*, 2015		X	X	X		X		S
		Firestone *et al.*, 2020 (*BORIS*)		X	X	X		X	X	S
		Montazeri *et al.*, 2020			X					S

Tableau 5 : Méthode de modélisation des quatre processus non observés et d'inférence dans les méthodes de la NPF. L'évolution intra-hôte inclut également le type de bottleneck (complet ou incomplet). Le processus de transmission est divisé en trois points : les états considérés pour les hôtes infectés (S, E, I ou R), les contacts entre hôtes (homogènes ou non, prise en compte des données géographiques dans un noyau de transmission spatial) et si l'épidémie est due à une unique ou de multiples introductions.

Méthode (*nom*)	**Mutation des séquences**	**Evolution intra-hôte**	**Transmission**	**Observation des cas**	**Inférence**
Aldrin *et al.*, 2011	Modèle de Kimura	Aucun modèle explicite	SIR (période infectieuse)	Tous les cas sont observés mais pas toujours échantillonnés	Maximum de vraisemblance partielle
		Complet	Fonction des distances		
			Multiples		
Jombart *et al.*, 2011 (*Seqtrack*)	Choix	Aucun modèle explicite	Aucun modèle explicite	Tous les cas sont observés et échantillonnés	Algorithme d'Edmonds
		Complet			
Ypma *et al.*, 2012	Délétion + Transition + Transversion	Aucun modèle explicite	SEIR (période de latence/infectieuse)	Tous les cas sont observés mais pas toujours échantillonnés	Bayésienne
		Complet	Noyau spatial		
			Unique		
Jombart *et al.*, 2014 (*outbreaker*)	Taux de mutation	Aucun modèle explicite	SI (temps de génération)	Proportion des cas échantillonnés	Bayésienne
		Complet	Contacts homogènes		
			Multiples		
Worby *et al.*, 2014	Taux de mutation	Taille de la population de pathogènes	Aucun modèle explicite	Tous les cas sont observés et échantillonnés	Distances génétiques observées *vs.* distribution théorique
		Incomplet			

Tableau 5 (suite) : Méthode de modélisation des quatre processus non observés et d'inférence dans les méthodes de la NPF. L'évolution intra-hôte inclut également le type de bottleneck (complet ou incomplet). Le processus de transmission est divisé en trois points : les états considérés pour les hôtes infectés (S, E, I ou R), les contacts entre hôtes (homogènes ou non, prise en compte des données géographiques dans un noyau de transmission spatial) et si l'épidémie est due à une unique ou de multiples introductions.

Méthode (*nom*)	**Mutation des séquences**	**Evolution intra-hôte**	**Transmission**	**Observation des cas**	**Inférence**
Famulare *et al.,* 2015	Taux de mutation	Aucun modèle explicite	Aucun modèle explicite	Aucune hypothèse	Test de rapport de vraisemblance + algorithme de pruning
Worby *et al.,* 2016 (*bitrugs*)	Aucun modèle explicite	Aucun modèle explicite	SEIR (période de latence/infectieuse)	Sensibilité du test <1	Bayésienne
		Aucune hypothèse	Contacts homogènes		
			Multiples		
Campbell *et al.,* 2019 (*outbreaker2*)	Taux de mutation	Aucun modèle explicite	SI (temps de génération)	Proportion des cas échantillonnés	Bayésienne
		Complete	Contact data		
			Multiples		

Tableau 6 : Méthode de modélisation des quatre processus non observés et d'inférence dans les méthodes de la SeqPF. L'évolution intra-hôte inclut également le type de bottleneck (complet ou incomplet). Le processus de transmission est divisé en trois points : les états considérés pour les hôtes infectés (S, E, I ou R), les contacts entre hôtes (homogènes ou non, prise en compte des données géographiques dans un noyau de transmission spatial) et si l'épidémie est due à une unique ou de multiples introductions. * signifie que plusieurs séquences peuvent être considérées pour une seule unité épidémiologique.

Méthode (*nom*)	**Mutation des séquences**	**Evolution intra-hôte**	**Transmission**	**Observation des cas**	**Inférence**
Cottam *et al.*, 2008	Sans objet	Aucun modèle explicite	SEIR (période de latence/infectieuse)	Tous les cas sont observés et échantillonnés	Annotation des nœuds internes
		Complet	Contacts homogènes		Maximum de vraisemblance
			Unique		
Didelot *et al.*, 2014	Sans objet	Processus coalescent	SIR (période infectieuse)	Tous les cas sont observés et échantillonnés	Annotation des branches
		Complet	Contacts homogènes		Bayésienne
			Unique		
Eldholm *et al.*, 2016	Sans objet	Processus coalescent	SEIR (période de latence/infectieuse)	Seuils de probabilité	Source d'information
		Complet	Contacts homogènes		Algorithme d'Edmonds
			Unique		
Didelot *et al.*, 2017 (*Transphylo*)	Sans objet	Processus coalescent	SI (temps de génération)	Proportion des cas échantillonnés	Annotation des branches
		Complet	Contacts homogènes		Bayésienne
			Unique		
Sashittal *et al.*, 2020 (*TiTUS*)	Sans objet	Aucun modèle explicite	Aucun modèle explicite	Tous les cas sont observés et échantillonnés	Annotation des nœuds internes
		Incomplet*			Problème de logique

Tableau 7 : Méthode de modélisation des quatre processus non observés et d'inférence dans les méthodes de la SimPF. L'évolution intra-hôte inclut également le type de bottleneck (complet ou incomplet). Le processus de transmission est divisé en trois points : les états considérés pour les hôtes infectés (S, E, I ou R), les contacts entre hôtes (homogènes ou non, prise en compte des données géographiques dans un noyau de transmission spatial) et si l'épidémie est due à une unique ou de multiples introductions. * signifie que plusieurs séquences peuvent être considérées pour une seule unité épidémiologique.

Méthode (*nom*)	**Mutation des séquences**	**Evolution intra-hôte**	**Transmission**	**Observation des cas**	**Inférence**
Ypma *et al.*, 2013	Taux de mutation	Processus coalescent	SEIR (période de latence/infectieuse)	Tous les cas sont observés et échantillonnés	Bayésienne
		Complet	Noyau spatial		
			Unique		
Hall *et al.*, 2015 (*beastlier*)	Choix du modèle de substitution	Processus coalescent	SEIR (période de latence/infectieuse)	Tous les cas sont observés mais pas toujours échantillonnés	Bayésienne
		Complet*	Noyau spatial		
			Unique		
De Maio *et al.*, 2016 (*SCOTTI*)	Choix du modèle de substitution	Processus coalescent	Modèle de migration	Nombre maximum d''hôtes	Bayésienne
		Incomplet*			
Klinkenberg *et al.*, 2017 (*phybreak*)	Taux de mutation	Processus coalescent	SI (temps de génération)	Tous les cas sont observés mais pas toujours échantillonnés	Bayésienne
		Complet	Contacts homogènes		
			Unique		

Tableau 7 (suite) : Méthode de modélisation des quatre processus non observés et d'inférence dans les méthodes de la SimPF. L'évolution intra-hôte inclut également le type de bottleneck (complet ou incomplet). Le processus de transmission est divisé en trois points : les états considérés pour les hôtes infectés (S, E, I ou R), les contacts entre hôtes (homogènes ou non, prise en compte des données géographiques dans un noyau de transmission spatial) et si l'épidémie est due à une unique ou de multiples introductions. * signifie que plusieurs séquences peuvent être considérées pour une seule unité épidémiologique.

Méthode (*nom*)	**Mutation des séquences**	**Evolution intra-hôte**	**Transmission**	**Observation des cas**	**Inférence**
Morelli *et al.*, 2012	Modèle de JC	Aucun modèle explicite	SEIR (période de latence/infectieuse)	Tous les cas sont observés et échantillonnés	Bayésienne
		Complet	Noyau spatial		
			Unique		
Mollentze *et al.*, 2014	Modèle de Kimura	Aucun modèle explicite	SEIR (période de latence/infectieuse)	Les cas observes contribuent à la transmission après leur retrait	Bayésienne
		Complet	Noyau spatial		
			Multiples		
Lau *et al.*, 2015	Modèle de Kimura	Aucun modèle explicite	SEIR (période de latence/infectieuse)	Tous les cas sont observés mais pas toujours échantillonnés	Bayésienne
		Complet	Noyau spatial		
			Multiples		
Firestone *et al.*, 2020 (*BORIS*)	Modèle de Kimura	Aucun modèle explicite	SEIR (période de latence/infectieuse)	Tous les cas sont observés mais pas toujours échantillonnés	Bayésienne
		Complet	Noyau spatial		
			Multiples		
Montazeri *et al.*, 2020	Modèle de JC	Aucun modèle explicite	Aucun modèle explicite	Tous les cas sont observés et échantillonnés	Bayésienne
		Complet			

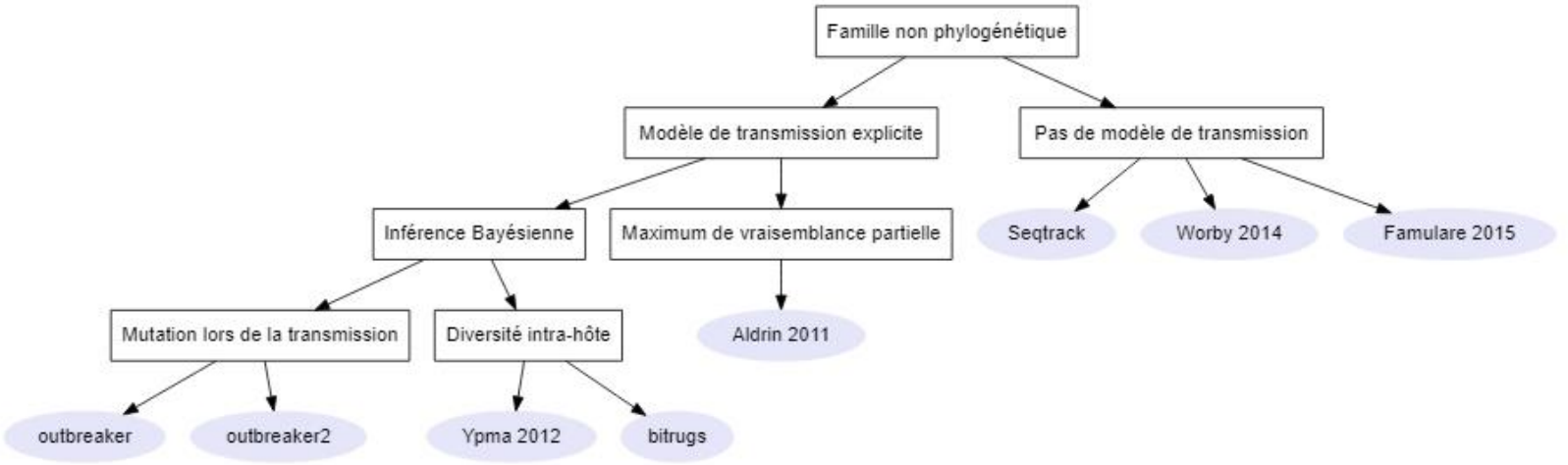


Figure 16 : Liens entre les méthodes de la NPF. Les rectangles représentent les critères de choix pour une méthode, et les cercles gris les noms de package ou le nom du premier auteur et l'année de publication.

3.3.1.1 Méthodes dans lesquelles mutation rime avec transmission

La méthode Bayésienne proposée par Ypma *et al.* (2012) utilisait trois types de données (temporelles, géographiques et génétiques) provenant d'une épidémie de H7N7 dans des élevages aviaires néerlandais. Cette méthode considérait les trois types de données indépendantes les unes des autres. La fonction de vraisemblance était donc le produit des contributions des trois types de données (125). De manière similaire, Jombart *et al.* (2014) ont décomposé la vraisemblance en une vraisemblance génétique et une vraisemblance temporelle dans le package *outbreaker* (126). Campbell *et al.* (2019) ont étendu le modèle de transmission de cette méthode, afin d'inclure des données de contact dans une vraisemblance de contact dans *outbreaker2* (121). La probabilité de transmission entre deux hôtes était inférée à partir de deux distributions, celle du temps de génération et celle de l'intervalle infection-échantillonnage, toutes deux supposées connues. De plus, *outbreaker* et *outbreaker2* considéraient deux paramètres permettant la modélisation des cas non observés : π, la proportion de cas échantillonnés et κ, le nombre maximum de générations entre un hôte échantillonné et son ancêtre échantillonné le plus récent dans l'arbre de transmission (121,126). Des épidémies de SARS-CoV-1 (121,126), de virus de la diarrhée bovine (BVD) (127), de *Klebsiella pneumoniae* (128) et *Acinetobacter baumannii* (129,130) (Annexe 5) ont été étudiées avec l'une de ces deux méthodes, *outbreaker* ou *outbreaker2*, disponibles sur R.

Ces trois méthodes faisaient l'hypothèse que la mutation avait lieu au moment de la transmission et leur vraisemblance génétique dépendait donc uniquement du nombre d'évènements de transmission séparant deux hôtes et non du temps écoulé (126).

3.3.1.2 Méthodes qui considèrent une diversité intra-hôte

Worby *et al.* (2014) ont mis en avant le fait que les méthodes précédentes avaient construit leur modèle génétique sur des hypothèses fortes, telles qu'un bottleneck de transmission complet (121,126) ou encore le fait que les mutations ont lieu au moment de la transmission (121,125,126). La diversité intra-hôte était donc considérée comme nulle dans ces méthodes. A l'inverse, Worby *et al.* (2014) ont d'abord construit une approximation de la distribution des distances génétiques puis ont comparé cette distribution aux distances génétiques observées, et ce, afin de déterminer la probabilité de transmission directe de *Straphylococcus aureus* méthicilline-résistant (MRSA) entre des individus hospitalisés (131). Worby *et al.* (2016) ont ensuite intégré la distribution des distances génétiques à un

modèle de transmission adapté à un contexte d'épidémie nosocomiale. Ils ont ainsi élaboré une méthode Bayésienne disponible dans un package R appelée *bitrugs* (132). Cette approche a permis de considérer la diversité intra-hôte ignorée par les autres méthodes, sans pour autant faire d'hypothèses sur le processus d'évolution intra-hôte (132), contrairement à leur premier travail (131). Le modèle de transmission considérait un hôpital où les patients étaient soit susceptibles (S) soit infectieux (I) un jour après infection et le taux de transmission par patient infecté était constant jusqu'à leur sortie d'hôpital. Les contacts dans la population étaient supposés homogènes, ce qui signifie que chaque individu infectieux avait un contact équivalent avec tous les autres individus susceptibles. De plus, la détection imparfaite des cas était modélisée par un paramètre représentant la Se du test (132).

3.3.1.3 Autres méthodes

Dans leur travail sur l'anémie infectieuse du saumon, Aldrin *et al.* (2011) n'ont pas utilisé d'inférence Bayésienne. Ils ont bien construit un modèle de transmission mais ont effectué le calcul d'un maximum de vraisemblance partielle lors de l'estimation des paramètres du modèle. A partir de ces paramètres estimés, ils ont calculé la probabilité de transmission d'un élevage de saumons à un autre. Dans leur modèle, la probabilité de transmission diminuait de manière exponentielle lorsque les distances génétiques et géographiques augmentaient. La probabilité de transmission dépendait également de caractéristiques d'élevage telles que le nombre maximum de saumons dans une cohorte pendant la période de production ainsi que la période de production (printemps *vs.* automne) (120). Quand les données génétiques n'étaient pas disponibles pour un élevage, la distance génétique inconnue était remplacée par la valeur d'un paramètre estimée à partir des distances génétiques connues (120).

Enfin, la méthode seqTrack (133) et celle proposée par Famulare *et al.* (2015) différaient de toutes les autres et ne modélisaient que le processus de mutation de manière explicite. En effet, Jombart *et al.* (2011) ont reconstruit un arbre de transmission dans lequel « les ancêtres précèdent toujours leurs descendants » (en faisant l'hypothèse que les dates d'échantillonnage suivent le même ordre chronologique que les dates d'infection) et la distance génétique totale entre les nœuds était minimale (arbre couvrant de poids minimal) en utilisant l'algorithme d'Edmonds (134). Cette méthode peut être utilisée uniquement à partir de données génétiques et des dates d'échantillonnage, cependant d'autres données épidémiologiques (*ex.* données géographiques) peuvent être prises en compte quand plusieurs

hôtes ont la même probabilité d'en infecter un autre. L'algorithme seqTrack a été implémenté dans le package *adegenet* et appliqué à la pandémie H1N1 d'origine porcine en 2009 (133) et des épidémies d'influenza équin H3N8 (135), de *M. tuberculosis* (136) et *K. pneumoniae* (137) (Annexe *5*).

A l'inverse, Famulare *et al.* ont identifié des liens de transmission directe en réalisant un test de rapport de vraisemblances afin de déterminer si la date du MRCA (tMRCA) d'une paire d'hôtes était égale à la date d'échantillonnage la plus ancienne des deux hôtes (138). Dans l'estimation de la vraisemblance du tMRCA, Famulare *et al.* ont fait l'hypothèse que le processus de mutation suivait un modèle de Poisson avec un taux constant connu. Les cas où plusieurs hôtes pouvaient avoir infecté le même hôte étaient résolus par un algorithme de « pruning » qui pouvait être spécifié par l'utilisateur, par exemple, en gardant le lien qui minimisait le temps entre le tMRCA et l'échantillonnage. Cette méthode a été appliquée à l'étude de l'épidémie d'Ebola en Sierra Leone, la pandémie à influenza H1N1 de 2001 et l'épidémie de poliomyélite au Nigéria (138) (Annexe *5*).

3.3.2 Familles phylogénétiques

Dans les deux familles phylogénétiques, un lien est établi entre les arbres phylogénétiques et ceux de transmission. A partir d'un scénario simple (*Figure 17*), trois modifications de l'arbre phylogénétique reconstruit permettent d'obtenir un arbre de transmission. La *Figure 17*.A montre un arbre phylogénétique dans lequel des hôtes échantillonnés sont attribués aux nœuds internes. L'arbre de transmission ainsi obtenu contient l'ordre de transmission mais pas les temps de transmission t (sauf si nous considérons que l'évènement de coalescence et la transmission ont lieu en même temps). Dans la *Figure 17*.B, les nœuds sont annotés dans l'arbre phylogénétique. Cependant, la branche entre deux nœuds hébergés par des hôtes différents est considérée comme une « branche d'infection » et l'évènement de transmission a lieu le long de cette branche. Nous obtenons ainsi un arbre de transmission daté dans lequel les évènements de transmission ne coïncident pas forcément avec les évènements de coalescence. Enfin, dans 17C, il est possible d'attribuer un hôte non échantillonné à un nœud interne dans l'arbre phylogénétique, l'hôte U qui n'a pas été observé peut donc être inféré dans l'arbre de transmission.

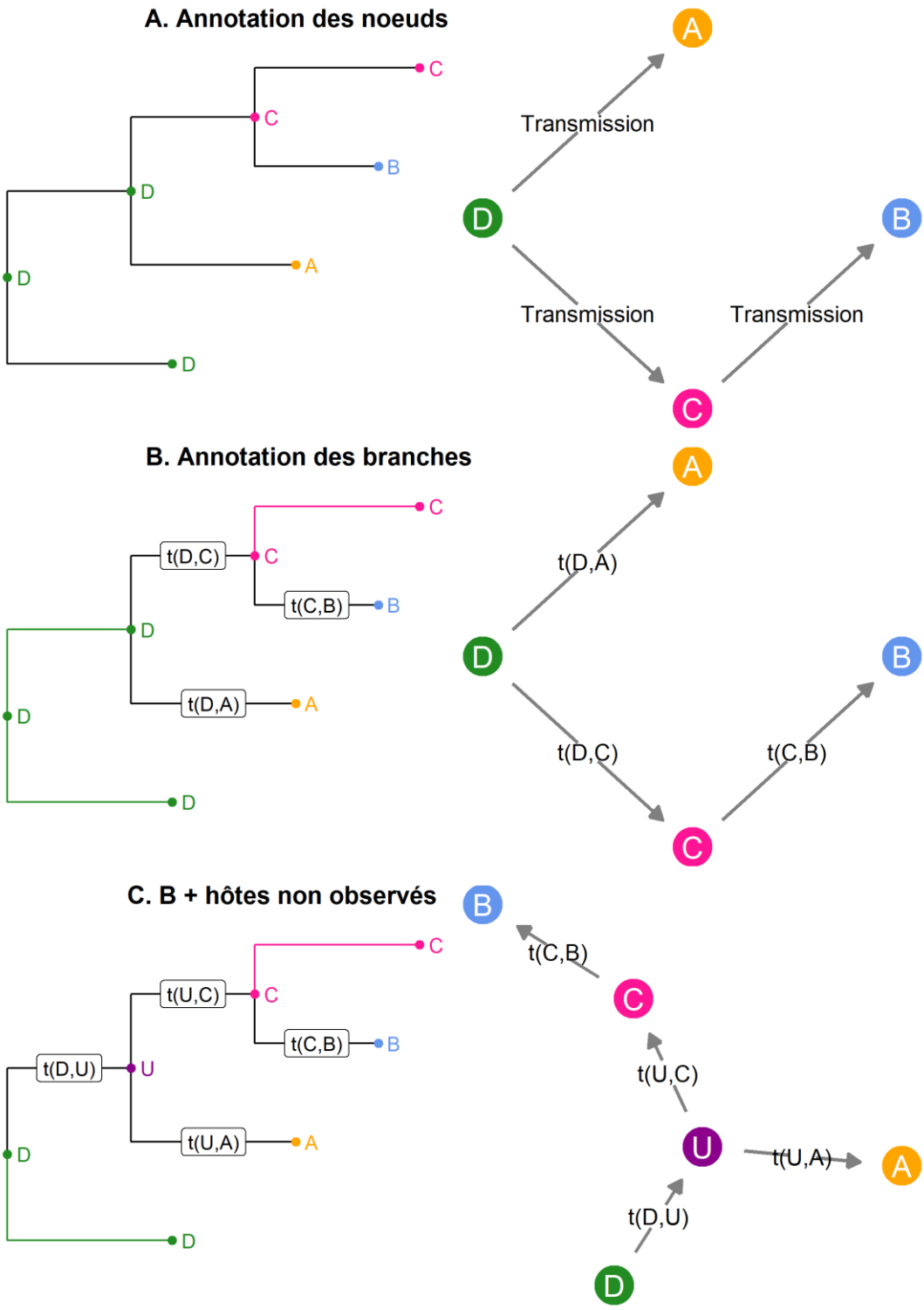


Figure 17 : Trois types de liens entre les arbres phylogénétiques (à gauche) et les arbres de transmission (à droite). L'annotation des nœuds (A) permet

d'identifier des évènements de transmission. L'annotation des branches (B) ajoutent les temps de transmission et la prise en compte des hôtes non observés (C) permet d'identifier l'hôte U.

3.3.2.1 Famille phylogénétique séquentielle

Ces méthodes (*Figure 18*) nécessitent la reconstruction préalable d'un arbre phylogénétique. Dans une seule méthode, l'arbre phylogénétique était utilisé comme une source d'information à propos des temps de coalescence entre deux lignées (139). En effet, Eldholm *et al.* (2016) ont utilisé cette information combinée à un modèle de transmission SEIR afin de calculer la vraisemblance de liens orientés de transmission directe entre des individus échantillonnés lors d'une épidémie de *M. tuberculosis* (139). Une fois la vraisemblance de transmission calculée pour chaque paire d'individus, la direction correspondant à la vraisemblance la plus faible était retirée. Enfin, l'arbre couvrant de poids minimal était construit avec l'algorithme d'Edmonds (134), comme dans seqTrack. Afin de prendre en compte la détection imparfaite des cas, cette méthode a utilisé différents seuils de probabilité de transmission directe en reconstruisant les arbres de transmission (139).

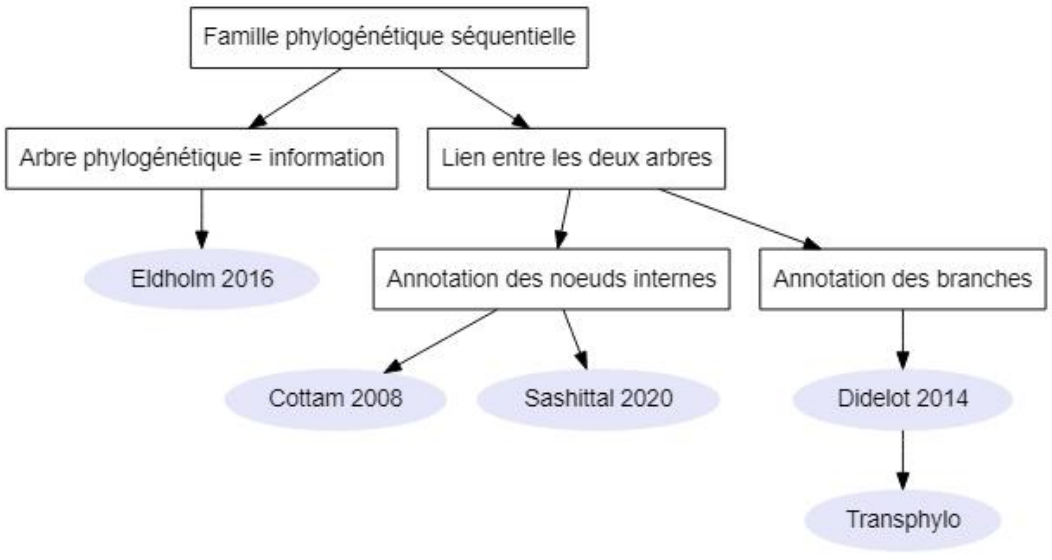


Figure 18 : Liens entre les méthodes de la SeqPF. Les rectangles représentent les critères de choix pour une méthode, et les cercles gris les noms de package ou le nom du premier auteur et l'année de publication.

Les quatre méthodes restantes dans cette famille annotaient l'arbre phylogénétique pour reconstruire un arbre de transmission mais utilisaient

différentes méthodes d'échantillonnage lors de l'exploration de l'espace des arbres de transmission. Dans la méthode de Cottam *et al.* (2008), aucune stratégie d'échantillonnage n'a été mise en place car le nombre d'arbres de transmission compatibles avec leurs données et les connaissances à priori dont ils disposaient sur l'histoire de transmission (grâce à des données de mouvements) était relativement faible (1728 arbres). Chaque arbre de transmission était construit en attribuant à chaque nœud interne un de ses deux descendants tout en remontant l'arbre phylogénétique (140). Cela correspond à l'annotation des nœuds internes décrite dans 17A, mais avec des contraintes supplémentaires. En effet, seuls les deux descendants d'un nœud sont considérés pour son annotation. Puis, la vraisemblance d'un arbre de transmission était estimée à partir de la vraisemblance jointe de chaque paire de transmission, qui était elle-même basée sur la probabilité des données épidémiologiques (dates de retrait et début d'infectiosité) en fonction d'un modèle de transmission SEIR (*Tableau 6*). Cette méthode a été élaborée pour étudier une épidémie de FA en 2001 au Royaume-Uni (Annexe *6*).

De manière similaire, la méthode proposée par Sashittal et El-Kabir (2020) avait pour but d'annoter les nœuds internes dans l'arbre phylogénétique reconstruit à partir de séquences de VIH (122) (annotation des nœuds similaire à 17A). Cependant, cette méthode considérait un bottleneck de transmission incomplet et l'annotation de chaque nœud interne n'était pas restreinte à l'un de ses deux descendants. De plus, la transmission n'était pas modélisée de manière explicite mais l'annotation devait satisfaire des contraintes dérivées de données épidémiologiques : fenêtres de transmission (date de début d'exposition jusqu'à la date de retrait) et données de contact (*Tableau 4*). La reconstruction d'arbres de transmission était traitée comme un problème de logique. Puis, les arbres de transmission étaient échantillonnés de manière uniforme parmi ceux qui satisfaisaient les contraintes temporelles et des données de contact. Enfin, l'arbre le plus parcimonieux était sélectionné parmi les arbres échantillonnés (122).

Les méthodes qui identifient les évènements de coalescence dans un arbre phylogénétique comme autant d'évènements de transmission, ne considèrent pas l'évolution intra-hôte (56). A l'inverse, Didelot *et al.* (2014) ont inféré un arbre de transmission en décrivant les hôtes le long des branches de l'arbre phylogénétique (56,141) (annotation des branches similaire à 17B). Les hôtes pouvant changer le long des branches et pas seulement au niveau des nœuds internes, les évènements de transmission ne sont plus restreints aux temps de coalescence. Dans la méthode de Didelot

et al. (2014), l'inférence Bayésienne permettait d'estimer les paramètres de leur modèle SIR, ceux de leur modèle d'évolution intra-hôte (processus coalescent neutre avec une taille de population constante Ne et un temps moyen de génération de population g, qui ici désigne la durée du cycle de réplication) et enfin, d'inférer l'arbre de transmission. Contrairement à la simple énumération réalisée par Cottam *et al.* (2008) et l'échantillonnage uniforme dans la méthode de Sashittal et El-Kabir (2020), l'échantillonnage par MCMC était ici utilisé dans l'exploration de l'espace des arbres de transmission. De plus, Didelot *et al.* ont utilisé des données géographiques ainsi que des résultats de tests diagnostiques pour valoriser ou pénaliser les arbres de transmission (56).

La limite principale des méthodes précédentes est le fait qu'elles considèrent l'épidémie finie et tous les cas échantillonnés (141). Didelot *et al.* (2017) ont donc implémenté une autre méthode Bayésienne appelée *TransPhylo* dans un package R, où l'utilisateur pouvait définir une distribution à priori de la probabilité qu'un hôte infecté soit échantillonné et sélectionner le scénario d'une épidémie finie ou encore en cours (annotation des branches avec prise en compte des hôtes non observés, comme dans 17C). Contrairement à leur premier travail, le modèle de transmission considéré ici était un processus de branchement (141). Ce processus de branchement était défini par la distribution du nombre de cas secondaires (nombre d'hôtes pouvant être infectés par un seul hôte) et la distribution du temps de génération (141). Le package *TransPhylo* a été choisi pour étudier des transmissions bactériennes (épidémies de *M. tuberculosis* (142–144) et *K. pneumoniae* (145,146)) et virales (pandémie due à SARS-CoV2 (147) et une grande épidémie d'oreillons au Canada (148)). Récemment, le package *TransPhylo* (141) a été étendu pour inférer des arbres de transmission à partir de plusieurs arbres phylogénétiques et non plus d'un seul arbre phylogénétique consensus comme le MCC (149).

Aucune de ces méthodes de reconstruction d'arbres de transmission ne modélisait de manière explicite la mutation des séquences étant donné qu'elles étaient appliquées à un arbre phylogénétique déjà reconstruit (ceci explique le « sans objet » dans le *Tableau 6*). Cependant, certains articles (139,141,142,145,147–149) utilisaient un modèle de substitution dans la reconstruction de leur arbre phylogénétique.

3.3.2.2 Famille phylogénétique simultanée

➢ Méthodes implicitement phylogénétiques

Cinq méthodes dans cette famille (*Figure 19*) considéraient implicitement un arbre phylogénétique où les nœuds internes correspondaient aux évènements de transmission (annotation des nœuds comme dans 17A). Morelli *et al.* (2012) ont tout d'abord construit une fonction de vraisemblance qui prenait en compte les corrélations entre les données génétiques et épidémiologiques dans l'étude des épidémies de FA de 2001 et 2007 au Royaume-Uni (150). En effet, leur pseudo-vraisemblance génétique dépendait de l'intervalle de temps entre infection et détection, permettant donc indirectement l'occurrence de mutations au sein de l'hôte sans modéliser pour autant l'évolution intra-hôte. Ce modèle de transmission a été étendu par Mollentze *et al.* (2014) pour prendre en compte des introductions multiples du pathogène dans la population d'hôtes au lieu d'un seul cas index. La prise en compte d'introductions multiples est plus appropriée lors de l'étude de situations endémiques, ce modèle fut ainsi appliqué à une épidémie de rage canine en Afrique du Sud (151) (Annexe *7*). Ces deux méthodes utilisaient un modèle de transmission SEIR (*Tableau 7*) et estimaient des paramètres dont la période de latence et l'intervalle infection-échantillonnage (150,151). Cependant, Mollentze *et al.* (2014) modélisaient indirectement la présence de cas non observés car ils considéraient qu'un cas observé pouvait continuer à transmettre après sa date de retrait. De plus, ils prenaient en compte deux catégories d'hôtes, ceux qui pouvaient transmettre le virus (chiens) et ceux qui ne le pouvaient pas (151).

Lau *et al.* (2015) ont noté que les méthodes précédentes ne permettaient pas d'inférer les séquences transmises (non observées). En effet, Morelli *et al.* (2012) et Mollentze *et al.* (2014) ont considéré une pseudo-vraisemblance génétique uniquement pour les séquences observées (150,151). A l'inverse, Lau *et al.* (2015) ont proposé une inférence jointe en modélisant les données génétiques manquantes et inférant les séquences non observées en même temps que l'arbre de transmission (40). Dans leur modèle de transmission, deux types d'infections étaient pris en compte : les infections primaires, correspondant aux cas importés dont les séquences étaient dérivées d'une séquence universelle G_M, et les infections secondaires qui suivaient un modèle SEIR (40). Hayama *et al.* (2019) ont appliqué cette méthode à l'épidémie de FA de 2010 au Japon (152) (Annexe *7*). *BORIS* est une extension du modèle de Lau *et al.* qui incorpore des covariables à l'échelle de l'élevage, telles que le nombre d'animaux et l'espèce prédominante (*Tableau 4*), influençant la susceptibilité et l'infectiosité des

élevages dans le modèle de transmission (39).

La méthode la plus récente de cette sous-catégorie (*Figure 19*) ne prenait pas en compte de modèle explicite de transmission (153). En effet, Montazeri *et al.* (2020) ont élaboré deux algorithmes reconstruisant un arbre phylogénétique à partir d'un arbre de transmission possible. Pour ce faire, ces algorithmes considéraient des estimations de dates d'infection des hôtes et une diversité intra-hôte nulle. Montazeri *et al.* ont utilisé cette méthode dans l'étude d'un cluster de transmission du VIH provenant de San Diego, Californie et de l'épidémie d'Ebola de 2014 en Sierra Leone.

➢ Méthodes explicitement phylogénétiques

Contrairement à ces cinq méthodes précédentes, quatre autres méthodes avaient pour objectif d'inférer simultanément les arbres phylogénétiques et les arbres de transmission. Dans ces quatre méthodes Bayésiennes, un lien était établi entre les arbres phylogénétiques et les arbres de transmission et à chaque étape du MCMC, les deux types d'arbres étaient modifiés de manière à conserver leur compatibilité. Ypma *et al.* (2013) considéraient un arbre hiérarchique, où toutes les phylogénies intra-hôtes étaient connectées entre elles par des liens de transmission (114). En s'intéressant à l'épidémie de FA de 2001 au Royaume-Uni précédemment étudiée avec d'autres méthodes (140,150), ils ont considéré que tous les hôtes infectés étaient observés (*Tableau 7*) (41). La méthode de Klinkenberg *et al.* (2017) (114) considérait également un arbre hiérarchique. Cette méthode, implémentée dans le package R *phybreak*, a été appliquée à cinq jeux de données publiés : *M. tuberculosis* (56), MRSA, deux épidémies de FA (41,140,150) et une de H7N7 (124) (Annexe *7*).

Au lieu de modifier individuellement chaque phylogénie intra-hôte comme dans l'approche de l'arbre hiérarchique, la méthode de Hall *et al.* (2015) découpait l'arbre phylogénétique en attribuant des hôtes aux nœuds internes. Puis, un paramètre permettant de déterminer le temps d'infection le long de la branche de l'arbre était estimé pour chaque hôte (annotation des branches similaire à 17B) (124). Hall *et al.* (2015) ont ainsi étudié l'épidémie de H7N7 datant de 2003 en prenant l'élevage comme unité épidémiologique. De plus, ils ont considéré deux catégories d'élevages aviaires (ceux avec un « risque élevé » et ceux avec un « risque faible »), qui se distinguaient par une distribution différente de la période de latence du fait d'implémentations de mesures de contrôle (124). L'observation des cas n'était pas modélisée mais les données génétiques manquantes (cas observés non échantillonnés) étaient remplacées par des séquences non

informatives (répétition du nucléotide « N ») (124). Cette méthode, implémentée dans le package *beastlier* disponible dans BEAST, a également été appliquée à une épidémie de H5N8 (Annexe *7*). L'unité épidémiologique était cette fois-ci l'oiseau (154).

Dans ces trois méthodes, le processus de transmission était modélisé par un des modèles épidémiologiques déjà mentionnés dans la littérature, tels qu'un processus de branchement homogène (114) ou un modèle SEIR individu-centré incluant une fonction décrivant les caractéristiques intrinsèques pouvant influencer la transmission (*ex.* données géographiques dans un noyau de transmission spatial (41,124) ou des groupes basés sur le risque (124)). Cependant, De Maio *et al.* (2016) ont proposé une approche Bayésienne originale (155) et ont utilisé une approximation de la coalescence structurée (BASTA, (38)). Ils ont considéré les hôtes comme des populations distinctes caractérisées par leur intervalle d'exposition (du début de l'exposition jusqu'à leur retrait, *Tableau 4*) et entre lesquelles les pathogènes pouvaient migrer. Ce modèle de transmission autorise les infections multiples au sein d'un même hôte et la transmission de plusieurs souches lors d'une infection. Cette méthode, implémentée dans le package *SCOTTI* (155) disponible dans BEAST2 (95), est donc plus appropriée dans l'étude d'épidémies où les infections mixtes et les inocula de transmission larges sont fréquents ; par exemple, des épidémies de FA et *K. pneumoniae* (Annexe *7*). De plus, cette méthode est la seule dans la SimPF (*Tableau 7*) à modéliser l'observation des cas (l'utilisateur pouvait spécifier un nombre maximum d'hôtes dans l'épidémie) (annotation des branches et prise en compte des hôtes non observés comme dans 17C).

Toutes ces méthodes modélisaient explicitement la mutation des séquences avec un modèle de substitution, et quatre méthodes sur neuf modélisaient l'évolution intra-hôte avec un processus coalescent (*Tableau 7*).

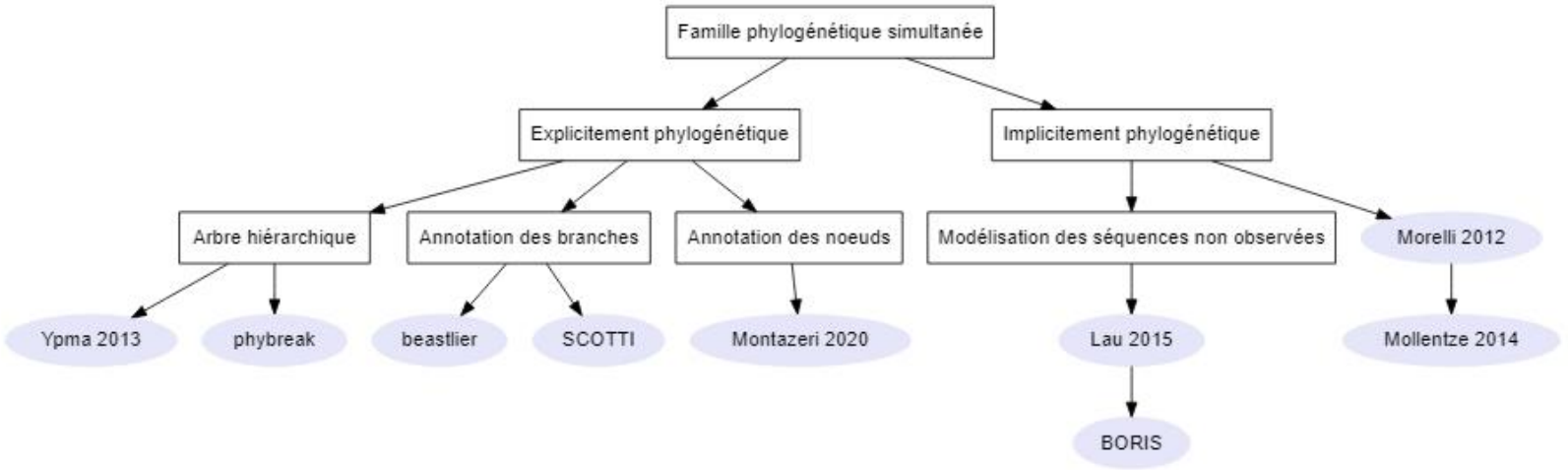


Figure 19 : Liens entre les méthodes de la SimPF. Les rectangles représentent les critères de choix pour une méthode, et les cercles gris les noms de package ou le nom du premier auteur et l'année de publication.

3.3.3 Exemples d'application des méthodes de reconstruction

3.3.3.1 Epidémie de M. tuberculosis

M. tuberculosis est caractérisé par un taux de mutation lent (~1 x 10^{-7} substitutions par site et par an, Annexe *8*) avec une proportion élevée de séquences échantillonnées identiques (57). Une infection à *M. tuberculosis* peut conduire à une longue période de latence et la majorité des cas infectés sont asymptomatiques. Nous ne devrions donc pas faire l'hypothèse que tous les cas sont observés. De plus, ne pas prendre en compte la possible longue période de latence pourrait conduire à inférer un arbre de transmission incorrect. Cependant, l'évolution intra-hôte pourrait être négligée du fait du taux de mutation faible. Les méthodes (disponibles dans un package) qui autorisent la détection imparfaite des cas sont *outbreaker(2)*, *bitrugs*, *TransPhylo* et *SCOTTI* (*Tableau 5* à 7). Parmi ces cinq méthodes, *TransPhylo*, *outbreaker* et *outbreaker2* pouvaient autoriser une longue période de latence en sélectionnant une distribution appropriée pour le temps de génération, comme cela a été montré avec une distribution Gamma dans *TransPhylo* (141). Des épidémies à *M. tuberculosis* ont été reconstruites avec cinq méthodes : seqTrack (NPF), Didelot *et al.* (2014), Eldholm *et al.* (2016) et *TransPhylo* (SeqPF) et *phybreak* (SimPF) (Annexes 5 à 7).

3.3.3.2 Epidémie de FA

A l'inverse, le virus de la FA présente un taux de mutation élevé (~2 x 10^{-5} substitutions par site et par jour, Annexe *8*) et l'unité épidémiologique la plus pertinente dans une épidémie de FA est l'élevage. De plus, le vent peut jouer un rôle dans la propagation du virus de la FA (44) et les porcins sont plus excréteurs que les ruminants, qui sont quant à eux plus susceptibles (115). Quand l'hôte considéré est un élevage et l'agent pathogène présente un taux de mutation élevé, il semble difficile d'ignorer l'évolution intra-hôte. De plus, les élevages ayant des localisations fixes et la transmission indirecte pouvant jouer un rôle, nous ne devrions pas faire l'hypothèse d'un mélange homogène des hôtes ni mettre de côté l'information apportée par les données géographiques. Enfin, prendre en compte l'espèce prédominante dans le modèle de transmission permettrait d'exploiter la dissymétrie des rôles joués par les élevages porcins et bovins. Les méthodes (disponibles dans un package) qui avaient un modèle de transmission individu-centré avec un noyau de transmission spatial sont *BORIS* et *beastlier* (*Tableau 5* à 7). Cependant, alors que *BORIS* prend en compte les caractéristiques

intrinsèques des élevages, *beastlier* modélise l'évolution intra-hôte. Sept méthodes ont été utilisées pour étudier une épidémie de FA : Cottam *et al.* (2008) (SeqPF), Ypma *et al.*, SCOTTI, *phybreak*, Morelli *et al.* (2012), Lau *et al.* (2015) et *BORIS* (SimPF) (Annexes 5 à 7).

3.3.3.3 Epidémie de MRSA

Les MRSA présentent un taux de mutation faible (~4 x 10^{-6} substitutions par site et par an) (Annexe *8*). Cependant, la diversité intra-hôte est importante à prendre en compte lors de l'étude de *S. aureus* (141). Les épidémies étudiées dans les articles sélectionnés avait lieu dans les services de soins intensifs néonatals (121,132,156) (Annexes 5 à 7). Le contexte hospitalier implique une forte proportion de cas échantillonnés ou *a minima* détectés ainsi que de multiples introductions possibles de l'agent pathogène. Des données de contact pourraient être également disponibles et être intéressantes à considérer.

Les méthodes utilisées pour reconstruire une épidémie dans un tel contexte pourraient donc faire l'hypothèse que tous les cas sont observés. Cependant, il semble inapproprié de supposer une seule introduction de l'agent pathogène. Les méthodes (disponibles dans un package) qui ne font pas l'hypothèse d'une introduction unique de l'agent pathogène sont seqTrack, *outbreaker* et *outbreaker2*, *bitrugs* et *BORIS* (*Tableau 5* à 7). Parmi ces méthodes, seule *bitrugs* autorise la diversité intra-hôte et a été élaborée pour étudier une épidémie nosocomiale, alors que *outbreaker2* permet de prendre en compte des données de contact. Le choix entre ces deux méthodes dépend donc des types de données disponibles et si la prise en compte de la diversité intra-hôte est nécessaire afin de répondre aux questions posées à propos de l'épidémie.

3.4 DISCUSSION

Nous avons réalisé une revue systématique de la littérature, à la recherche de méthodes de reconstruction d'arbres de transmission qui combinaient des données génomiques et épidémiologiques. Les données épidémiologiques nécessaires à l'implémentation de chaque méthode étaient un premier critère de différenciation. Les méthodes étaient ensuite divisées en trois familles définies par la manière dont les données génétiques étaient intégrées dans l'inférence de l'arbre de transmission. Nous avons ainsi mis en avant des considérations pratiques permettant de sélectionner la méthode de reconstruction appropriée au contexte. Enfin, nous avons discuté le choix de méthodes appropriées pour trois épidémies bien différentes.

3.4.1 La combinaison des données épidémiologiques et génomiques

Nous nous sommes intéressés à l'intégration des données génomiques et épidémiologiques dans l'inférence d'arbres de transmission. Cependant, deux méthodes (Cottam *et al.* (2008) et seqTrack) ont été critiquées par d'autres car elles n'intégreraient pas toute l'information apportée par les deux types de données. Même si les arbres de transmission étaient reconstruits à partir de l'arbre phylogénétique, Cottam *et al.* (2008) ont calculé la vraisemblance de ces mêmes arbres seulement à partir des données épidémiologiques, mettant de côté toute information supplémentaire que les données génétiques auraient pu contenir (125). A l'inverse, seqTrack ne considère les données épidémiologiques autres que les dates d'échantillonnage seulement pour résoudre les cas où les séquences génétiques des hôtes considérés ne permettent pas de les distinguer (125).

3.4.2 La séparation en trois familles de méthodes

Dans la famille non-phylogénétique, la probabilité de transmission était estimée à partir des distances génétiques entre paires de séquences. Cependant, les méthodes de deux familles utilisaient les arbres phylogénétiques pour reconstruire les arbres de transmission. Ainsi, ces méthodes inféraient l'hôte hébergeant chaque nœud ou branche dans l'arbre phylogénétique, considéraient que les arbres phylogénétiques intra-hôtes faisaient partie d'un arbre hiérarchique ou utilisaient l'arbre phylogénétique comme une source d'information. Dans la famille phylogénétique séquentielle, les arbres phylogénétiques étaient reconstruits avant d'implémenter la méthode et cela nécessitait donc un choix supplémentaire, celui de la méthode de reconstruction de l'arbre phylogénétique. De plus, l'arbre phylogénétique devait être correctement reconstruit pour ne pas conduire à des erreurs dans l'arbre de transmission. Toutes les méthodes phylogénétiques séquentielles ont d'abord utilisé un seul arbre fixe généré au préalable par une méthode de reconstruction standard. Ces méthodes ignoraient donc toute incertitude dans l'estimation de l'arbre phylogénétique (124) et ne prenaient ainsi pas en compte l'incertitude associé au processus évolutif (114). Cependant, le package *TransPhylo* a ensuite été étendu pour reconstruire les arbres de transmission à partir de multiples arbres phylogénétiques (149). Une autre manière de prendre en compte l'incertitude du processus évolutif était d'inférer simultanément les arbres phylogénétiques et de transmission, nous avons regroupé ces méthodes dans la famille phylogénétique simultanée.

3.4.3 Les quatre processus non observés

3.4.3.1 La mutation des séquences

Comme mentionné par Klinkenberg *et al.*, quatre processus non observés pouvaient être pris en compte ou ignorés (114) : la mutation des séquences, l'évolution intra-hôte, la transmission et l'observation des cas. Les modèles de substitution modélisent explicitement la mutation des séquences alors que les distances génétiques calculées sans modèle de substitution ne considèrent pas les mutations intermédiaires (dans A → G → T, le passage par G n'est pas pris en compte) ou reverses (A → G → A) et peuvent donc conduire à des estimations incorrectes. Les méthodes dans la SeqPF modélisaient indirectement ce processus de mutation ou ne le modélisaient pas du tout, selon la méthode utilisée pour générer l'arbre phylogénétique. Cottam *et al.* (2008) ont utilisé une méthode par parcimonie alors que les autres ont généralement opté pour une méthode Bayésienne avec un modèle de substitution. Dans les deux familles restantes (NPF et SimPF), toutes les méthodes proposaient l'option de prendre en compte un modèle de substitution.

3.4.3.2 L'évolution intra-hôte

Nous nous attendons à une évolution intra-hôte non négligeable lors d'infections par des agents pathogènes avec un temps de génération long (141), combiné à un taux de mutation élevé. Ignorer les mutations intra-hôtes (comme c'est le cas dans *outbreaker* qui considère que les mutations ont lieu au moment de la transmission (114,124)) n'est donc pas approprié dans ce cas. De plus, certaines méthodes (Morelli *et al.* (2012), Mollentze *et al.* (2014) et Lau *et al.* (2015)) considèrent que les mutations peuvent avoir lieu au sein de l'hôte mais qu'une seule lignée y est présente (124), ignorant ainsi la diversité intra-hôte. Cependant, Ypma *et al.* ont remarqué qu'ignorer la contribution de la diversité intra-hôte aux différences observées entre les séquences échantillonnées pouvait conduire à une inférence incorrecte de l'arbre de transmission, et ce, quand la proportion d'échantillonnage est élevée dans une épidémie (41). Les méthodes modélisant l'évolution intra-hôte l'assimilaient généralement à un processus coalescent, ce qui requiert l'hypothèse d'une fraction d'échantillonnage faible au sein de l'hôte (124). Cette condition est généralement vérifiée quand l'hôte correspond à un individu mais nous devrions garder à l'esprit cette hypothèse quand l'hôte s'avère en réalité un groupe d'individus comme c'est le cas dans un élevage.

Si nous considérons les élevages comme unités épidémiologiques, il

pourrait être plus difficile de faire l'hypothèse d'une seule infection (de multiples introductions pouvant avoir lieu) et/ou d'une seule lignée intra-hôte. Cependant, les arbres de transmission reconstruits ne considéraient généralement que le premier évènement de transmission ou, quand cela était nécessaire de prendre en compte les évènements de transmission suivants, les hôtes étaient simplement dupliqués et les infections étaient supposées indépendantes (126). Aldrin *et al.* ont complétement ignoré les infections multiples d'un même élevage. En effet, quand plusieurs séquences étaient disponibles pour un élevage, ils sélectionnaient celle qui présentait la distance génétique la plus faible par rapport aux autres élevages (120). La possibilité de transmettre des souches génétiques diverses était négligée dans la majorité des méthodes du fait de l'hypothèse d'un bottleneck de transmission complet (une seule souche transmise) (131). Cette hypothèse n'était relâchée que dans trois méthodes : Worby *et al.* (2014) faisait varier la taille du bottleneck (131) alors que De Maio *et al.* (2016) et Sashittal *et al.* (2020) ne considéraient pas de bottleneck et autorisaient la transmission de plusieurs souches (122,155) voire de multiples infections dans *SCOTTI* (155).

3.4.3.3 La transmission

Les modèles épidémiologiques contribuent à l'estimation de l'arbre de transmission le plus probable. Cependant, un nombre d'hypothèses sur l'histoire naturelle de la maladie et la diffusion de la maladie les régissent et doivent être prises en compte avant de choisir une méthode. Par exemple, les contacts homogènes entre hôtes signifient que tous les hôtes infectés ont autant de chances d'infecter n'importe quel hôte susceptible (comme dans Didelot *et al.* (2014), Eldholm *et al.* (2016) et *bitrugs*). Ceci pourrait poser problème quand nous considérons par exemple une épidémie de FA touchant des élevages et dans laquelle le vent pourrait jouer un rôle dans la transmission (44). La transmission entre élevages n'est plus équivalente mais dépend de la direction du vent et des distances géographiques. Les auteurs de certaines méthodes ont donc opté pour un modèle de transmission individu-centré avec un noyau de transmission spatial voire inclus des caractéristiques intrinsèques d'élevages influençant l'infectiosité et la susceptibilité, telles que l'espèce prédominante et la taille du cheptel (*BORIS*) (39). Enfin, l'hypothèse d'une unique introduction de l'agent pathogène dans la population n'est pas appropriée lors de l'étude d'une situation endémique (151) ou encore d'une infection nosocomiale dans un hôpital où des introductions multiples peuvent avoir lieu (132). Les auteurs de certaines méthodes ont donc identifié les séquences génétiques trop différentes des autres afin de les exclure voire ont inclus l'introduction de l'agent pathogène dans le modèle de transmission. Cinq méthodes (seqTrack, Worby *et al.*

(2014), Famulare *et al.* (2015), Montazeri *et al.* (2020) et *TiTUS*) n'ont pas modélisé le processus de transmission.

3.4.3.4 L'observation des cas

Le dernier processus non observé qui peut être pris en compte est l'observation des cas. Selon Didelot *et al.* (2017), la limite principale de certaines méthodes précédant le développement de *TransPhylo* était l'hypothèse selon laquelle l'épidémie était terminée et tous les cas étaient échantillonnés (141). En effet, considérer que tous les cas peuvent être liés par des évènements de transmission directe peut conduire à des estimations incorrectes de l'histoire naturelle de la maladie ou encore des liens de transmission. Les auteurs de certaines méthodes ont donc modélisé l'observation des cas en estimant la proportion d'hôtes observés (121,126,141), la Se du test utilisé (132) ou encore le nombre maximum d'hôtes dans l'épidémie (155). Mollentze *et al.* (2014) ont indirectement pris en compte les cas non observés en autorisant les hôtes à transmettre le pathogène après leur date de retrait (151). La nécessité de modéliser l'observation des cas dépend de la possibilité de passer à côté d'hôtes infectés dans l'épidémie étudiée. L'histoire naturelle de la maladie, la stratégie de détection des hôtes infectés et son efficacité doivent donc être considérées.

3.4.4 Considérations pratiques dans le choix d'une méthode

Le choix d'une méthode dépend également de sa facilité d'utilisation ainsi que de son applicabilité à un large éventail de jeux de données. Ceci se retrouve dans le nombre d'études, trouvées dans notre recherche, qui utilisaient une méthode disponible dans un package (n=4 pour seqTrack, n=4 pour *outbreaker* et n=9 pour *TransPhylo*). A l'inverse, des méthodes comme celles de Ypma *et al.* (2013) et Morelli *et al.* (2012) étaient élaborées pour un jeu de données spécifique et donc rarement utilisées pour autre chose. Malheureusement, le temps de calcul n'était pas toujours précisé dans les articles sélectionnés, ce qui complique l'estimation de la taille des jeux de données pouvant être étudiés par une méthode.

3.4.5 Limites dans l'inclusion de méthodes dans la revue systématique

Nous avons décidé d'exclure les méthodes qui utilisaient des données de « deep-sequencing » car nous n'avions pas accès à ce type de données pour *M. bovis* en France. Nous allons en citer deux ici pour information : la première, une méthode Bayésienne appelée BadTrIP (pour « Bayesian

epidemiological Transmission Inference from Polymorphisms ») considérait un format de données génétiques (nombre de nucléotides pour une même position du génome) différant grandement des autres méthodes (157). La deuxième appelée SLAFEEL (pour « Statistical Learning Approach For Estimating Epidemiological Links ») considérait un jeu de séquences pour chaque hôte et les données épidémiologiques étaient utilisées pour calibrer une pénalisation de la pseudo-vraisemblance (décrivant la probabilité d'obtenir le jeu de séquences d'un hôte infecté à partir du jeu de séquences présent dans l'infecteur) (158). Ces méthodes (ne constituant pas une liste exhaustive) pourraient être intéressantes quand plusieurs séquences sont disponibles pour un hôte, quand les hypothèses usuelles ne sont pas appropriées (SLAFEEL) ou quand nous ne pouvons pas ignorer la possibilité de recombinaisons (BadTrIP).

3.4.6 Conclusion

Le choix d'une méthode de reconstruction d'arbres de transmission dépend des caractéristiques du pathogène telles que le taux de mutation ou l'histoire naturelle de la maladie, les données génétiques et épidémiologiques disponibles et les questions auxquelles nous souhaitons apporter une réponse. L'impact que pourrait avoir des données biaisées ou l'inadéquation des hypothèses sous-jacentes des modèles d'évolution et épidémiologiques, sur la reconstruction d'arbres de transmission, n'a pas été analysé dans les publications étudiées et serait intéressant à investiguer.

4 EVALUATION DE LA PERFORMANCE DES METHODES DE RECONSTRUCTION D'ARBRES DE TRANSMISSION DANS LE CONTEXTE D'UN SYSTEME MULTI-HOTES

4.1 INTRODUCTION

Certaines méthodes de reconstruction d'arbres de transmission ont été développées afin d'étudier la transmission d'un agent pathogène au sein d'un système multi-hôtes spécifique, comme la méthode de Firestone *et al.* (*BORIS*) pour le virus de la FA (39). Cependant, la plupart des méthodes que nous avons répertoriées se basaient sur un *système mono-hôte* (un seul type d'hôtes). Ces méthodes étudiaient ainsi la transmission d'agents pathogènes évoluant lentement comme *M. tuberculosis* (114,141) ou plus rapidement comme MRSA (114,132) ou SARS-CoV-1 (121), dans une population humaine. Le fait que ces méthodes aient été développées sur des systèmes mono-hôtes ne les empêche pas pour autant de donner des résultats intéressants dans le cas de systèmes multi-hôtes. Par exemple, Willgert *et al.* ont récemment reconstruit l'histoire de transmission du SARS-CoV-2 dans un système humains-cerfs dans l'Iowa, aux Etats-Unis (159). Dans un système multi-hôtes, en plus de la reconstruction exacte d'évènements de transmission entre individus et l'estimation de la taille de l'épidémie (indicateurs épidémiologiques généraux), nous souhaiterions que ces méthodes permettent l'estimation de la contribution de chaque espèce-hôte ainsi que l'identification de l'espèce-hôte du cas index (qui sont des indicateurs épidémiologiques spécifiques aux systèmes multi-hôtes).

Un autre point qui ressort de notre revue systématique de la littérature est la différence dans la prise en compte des cas non observés. Certaines méthodes de reconstruction font l'hypothèse que tous les cas sont connus et échantillonnés (122,133,140), d'autres autorisent l'annotation d'hôtes non observés dans l'arbre phylogénétique (141) ou encore la présence d'hôtes intermédiaires entre deux hôtes échantillonnés (121,126). Si l'échantillonnage est incomplet dans une épidémie, il existe une différence entre l'épidémie qui a eu lieu et l'épidémie qui peut être reconstruite par ces méthodes à partir des séquences échantillonnées. En effet, même si la méthode de reconstruction prend en compte l'observation imparfaite des cas (121,141), les hôtes infectés non échantillonnés ne peuvent être inférés que s'ils ont des descendants échantillonnés (159). L'arbre de transmission ainsi reconstruit est alors une sous-partie de l'arbre réel, induite par le processus

d'échantillonnage.

Notre objectif était d'évaluer et de comparer la performance de trois méthodes de reconstruction dans des systèmes multi-hôtes de bTB ainsi que d'étudier si ces performances étaient influencées par des biais d'échantillonnage. Nous avons donc simulé la transmission de bTB dans le système multi-hôtes des PA et Landes. Nous avons ensuite implémenté plusieurs schémas d'échantillonnage afin de refléter la surveillance tardive de la faune sauvage (biais temporel) et le fait que toutes les espèces ne sont pas surveillées (biais d'espèce). Afin d'évaluer la qualité des arbres de transmission reconstruits, nous avons calculé à la fois des indicateurs épidémiologiques généraux et des indicateurs spécifiques à un système multi-hôtes.

4.2 MATERIEL ET METHODE

4.2.1 Simulation d'arbres de référence

4.2.1.1 Modèle de transmission

Nous avons étendu un modèle simulant des arbres de transmission de bTB pour les 11 génotypes identifiés dans le système bovins-blaireaux des PA et Landes, de janvier 2007 à janvier 2020 (13 ans) (160). Le modèle d'origine était un modèle stochastique avec un pas de temps mensuel qui considérait deux métapopulations (élevages bovins et groupes sociaux de blaireaux), les animaux étaient répartis en compartiments en fonction de l'âge et de l'état de santé. Au sein d'une métapopulation, trois voies de transmission existaient : intra-groupe (élevage ou groupe social), entre groupes voisins (d'après les données de pâtures de 2013 (90) et les domaines vitaux estimés pour les blaireaux), et enfin, entre groupes distants (d'après les données de mouvements de bovins au cours de la période d'étude (90) et la dispersion simulée des blaireaux). Bouchez-Zacria *et al.* ont fait l'hypothèse que la transmission entre élevages de bovins et groupes sociaux de blaireaux était médiée par l'environnement, dans lequel *M. bovis* pouvait survivre pendant trois mois. Le modèle de Bouchez-Zacria *et al.* prenait également en compte les mesures de détection (surveillance) et de contrôle (abattage) mises en place dans la région. Nous avons gardé les paramètres de transmission estimés dans leur étude par ABC (pour « Approximate Bayesian Computation ») (160). Nous avons restreint notre étude à l'un des deux génotypes de *M. bovis* qui étaient isolés à la fois dans la faune sauvage et chez les bovins dans cette région.

Des sangliers infectés ayant été détectés dans la région d'étude (84,86) et notre objectif étant d'étudier un système multi-hôtes complexe, nous avons ajouté une métapopulation de sangliers au modèle épidémiologique. Nous avons simulé les mouvements de cette troisième population en considérant qu'une proportion arbitraire de 0,01 des sangliers migrait d'une commune à une autre chaque mois. Dans un souci de simplicité, nous avons supposé que la taille de la population de sangliers demeurait constante au cours du temps au sein de chaque commune : les mouvements de sangliers étaient donc simulés comme des échanges aléatoires d'animaux entre deux communes voisines. De plus, à partir d'un taux de naissance estimé précédemment (161), le taux de mortalité correspondant et le nombre de sangliers tués à la chasse chaque année (seule cause de mort considérée), nous avons estimé le nombre de sangliers par commune (162). Nous avons fait l'hypothèse que la détection (supposée parfaite) de sangliers infectés lors de la chasse ne conduisait pas à la mise en place de mesures de contrôle ni de surveillance.

Comme dans la population de blaireaux, les sangliers pouvaient être susceptibles (S) ou infectés (I) alors que les bovins avaient un état supplémentaire de latence (E pour exposé), pendant lequel l'infection pouvait être détectée mais les bovins infectés ne pouvaient pas encore transmettre le pathogène (160). Nous avons implémenté des paramètres additionnels de transmission inter- et intra-espèces : $\beta_{C\text{-}Br}$ (de bovins à sangliers), $\beta_{Bg\text{-}Br}$ (de blaireaux à sangliers), $\beta_{Br\text{-}C}$ (de sangliers à bovins), $\beta_{Br\text{-}Bg}$ (de sangliers à blaireaux) et $\beta_{Br\text{-}Br}$ (entre sangliers). Etant donné l'absence de données disponibles concernant cette population de sangliers, nous avons fixé les valeurs de ces paramètres de transmission de façon à obtenir un nombre de sangliers infectés semblable au nombre de blaireaux infectés. Un nombre plus faible aurait compliqué l'implémentation des *schémas d'échantillonnage* (expression définie dans le *Tableau 9*). Le modèle ainsi modifié est représenté schématiquement en *Annexe 9*.

Tableau 8 : Paramètres de transmission concernant les sangliers et leur valeur fixée.

Paramètre	Sangliers à bovins ($\beta_{Br\text{-}C}$)	Sangliers à blaireaux ($\beta_{Br\text{-}Bg}$)	Entre sangliers ($\beta_{Br\text{-}Br}$)	Bovins à sangliers ($\beta_{C\text{-}Br}$)	Blaireaux à sangliers ($\beta_{Bg\text{-}Br}$)
Valeur fixée	0,001	0,01	0,01	0,01	0,001

De plus, les arbres de transmission simulés avec le modèle existant considéraient les élevages et les groupes sociaux de blaireaux comme unités

épidémiologiques, alors que notre but était de reconstruire des liens de transmission individuels. Nous avons donc modifié le modèle pour qu'il sélectionne aléatoirement des animaux infectés au sein de ces groupes selon les dynamiques des systèmes SEI/SI, le modèle simulait ainsi la transmission entre individus. Nous appellerons par la suite ces nouveaux arbres, les *arbres (de transmission) de référence* (*Tableau 9*).

4.2.1.2 Arbres de référence

Nous avons simulé dans un premier temps des bovins en cas index. Nous avons généré 30 arbres de référence afin d'investiguer des épidémies variées tout en limitant le temps de calcul. Ces 30 arbres devaient inclure moins de 500 hôtes infectés au total, pour limiter le temps de calcul, et au moins 15 hôtes infectés par espèce, afin de pouvoir implémenter les schémas d'échantillonnage. Un arbre de référence était représenté par une liste de six variables : identifiant de l'infecteur, identifiant de l'infecté, espèce-hôte de l'infecteur, l'espèce-hôte de l'infecté, date d'infection et date de mort de l'infecté.

Nous avons simulé des séquences génétiques le long des arbres de référence avec un taux de mutation fixe (μ) et selon un modèle de substitution HKY (21), puisque ce modèle avait déjà été utilisé pour reconstruire des arbres phylogénétiques à partir de séquences de *M. bovis* (93,94,163). Nous avons choisi μ égal à 0,0024 substitutions par site et par an et κ (ratio transition/transversion) égal à 5,9. En effet, ces valeurs avaient été estimées dans le premier axe de la thèse à partir des 167 séquences *M. bovis* (171 SNP de long) isolées chez des bovins et des blaireaux dans les PA et Landes (163).

A t=0, nous avons considéré que le cas index était infecté par une seule séquence tirée au sort parmi les 167 séquences isolées dans la zone d'étude du premier axe de la thèse (163). Notre algorithme de substitution était basé sur l'approche de Gillespie (164) implémentée dans le package phastSim (165) (*Figure 20*). Etant donné la faible diversité génétique observée à partir des séquences de *M. bovis* dans la région, nous avons supposé qu'il n'y avait pas de diversité intra-hôte en considérant une seule lignée par hôte tout en autorisant l'évolution intra-hôte (les mutations ont lieu au sein des hôtes infectés et non à la transmission).

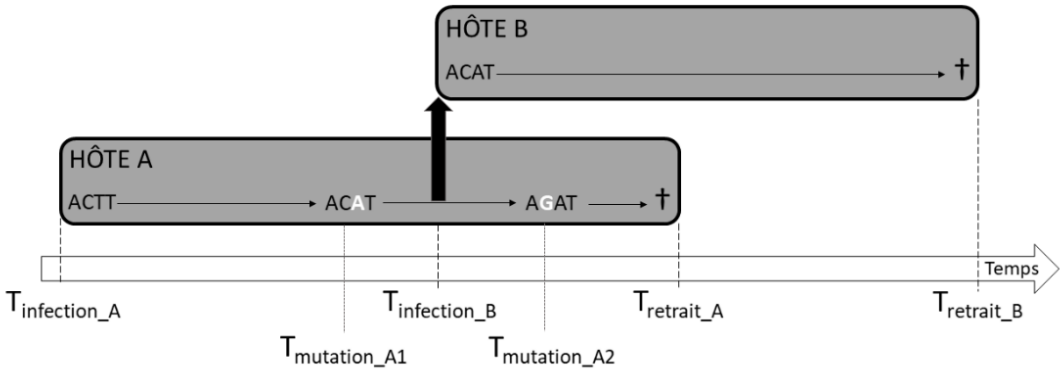


Figure 20 : Simulation de séquences au sein de deux hôtes infectés A et B. L'hôte A, représenté par le rectangle gris à gauche (infecté à $T_{infection\_A}$), transmet le pathogène à l'hôte B à $T_{infection\_B}$. Cet évènement de transmission est représenté par une flèche noire épaisse. Les hôtes étaient retirés de la simulation (représentés par les croix) à $T_{retrait\_A}$ et $T_{retrait\_B}$. Si le temps de mutation était inférieur au temps de retrait (voir $T_{mutation\_A1}$ et $T_{mutation\_A2}$ dans l'hôte A), le nucléotide qui mutait était alors tiré au sort (le 3ème puis le 2ème nucléotide dans l'hôte A, en blanc) et était modifié selon le modèle de substitution. Si le temps de mutation était supérieur au temps de retrait de l'hôte (voir l'hôte B), la séquence restait inchangée jusqu'au retrait, puis elle était échantillonnée.

Nous avons simulé des séquences jusqu'en février 2020, date à laquelle nous avons considéré que tous les hôtes encore vivants étaient échantillonnés. La dernière séquence simulée au sein de chaque hôte était celle présente à leur mort ou celle présente en février 2020, pour les hôtes encore vivants à cette date. Pour chaque arbre de référence, nous avons ainsi obtenu un *ensemble de cas de référence* (*Tableau 9*), c'est-à-dire une liste de quatre variables : identifiant de l'infecté, espèce-hôte de l'infecté, date de mort (ou février 2020 si hôte encore vivant) et séquence échantillonnée.

4.2.2 Schémas d'échantillonnage et arbres de transmission reconstruits

4.2.2.1 Schémas d'échantillonnage

Nous avons d'abord considéré la situation hypothétique où tous les hôtes infectés étaient observés (schéma d'échantillonnage de référence), ce qui correspondait à l'ensemble de cas de référence. Puis, nous avons simulé cinq schémas d'échantillonnage qui reproduisaient des biais d'observation identifiés dans les données épidémiologiques de bTB. Dans le schéma T (« biais temporel »), l'implémentation tardive de la surveillance de la faune

sauvage dans la région d'étude était simulée et nous n'avons considéré de la faune sauvage que les cas échantillonnés après 2012. De plus, le fait que toutes les espèces ne sont pas surveillées était simulé dans les schémas S (« biais d'espèces ») : soit les sangliers n'étaient pas échantillonnés, schéma S_W, soit les blaireaux, schéma S_B. Enfin, dans les schémas $T+S_W$ ($T+S_B$), nous n'avons pas pris en compte les cas échantillonnés avant 2012 dans l'espèce sauvage restante (respectivement les blaireaux et les sangliers).

Pour chaque arbre de référence, nous avons ainsi simulé en plus de l'ensemble de cas de référence, un *ensemble de cas biaisé* (*Tableau 9*) par schéma d'échantillonnage biaisé (T, S_B, S_W, $T+S_B$, $T+S_W$), contenant les mêmes variables. Avec 30 arbres de référence, nous avons ainsi obtenu un total de 30*6=180 ensembles de cas. Pour chaque ensemble de cas, nous avons extrait l'*épidémie reconstructible* (*Tableau 9*) à partir des arbres de référence, c'est-à-dire le sous-arbre ne contenant que les cas échantillonnés et leurs ancêtres.

Nous avons évalué le signal temporel présent dans les séquences simulées pour chaque ensemble de cas (biaisé ou non) en utilisant TempEst v1.5.3 (166). Si la valeur du R^2 (obtenue en sélectionnant la racine qui maximisait la valeur du coefficient de corrélation) était strictement supérieure à 0,1 (par convention) dans chaque ensemble de cas pour un même arbre de référence, nous avons gardé ces séquences. Sinon, nous avons simulé à nouveau les séquences jusqu'à satisfaire cette condition.

4.2.2.2 Reconstruction des arbres de transmission

A partir de la revue systématique de la littérature réalisée dans le deuxième axe de cette thèse (167), nous avons identifié trois méthodes de reconstruction disponibles dans un package R et qui ne nécessitaient que les dates d'échantillonnage et/ou les dates de retrait (seqTrack, *outbreaker2* et *TransPhylo*). Dans seqTrack et *outbreaker2*, l'estimation des liens de transmission se base sur les distances génétiques entre paires d'individus (NPF), alors que dans *TransPhylo*, un lien est établi entre les arbres phylogénétiques et ceux de transmission (PF) (167).

- seqTrack

A partir de l'algorithme d'Edmonds, seqTrack reconstruit l'arbre de transmission dans lequel la distance génétique totale entre nœuds est minimale, en supposant que l'infecteur est toujours échantillonné avant le ou les hôtes qu'il infecte (133). Afin d'utiliser cette méthode, nous avons estimé la distance génétique entre chaque paire d'individus avec la fonction dist.dna

(package *ape* (91)) et le modèle de substitution F84 qui ressemble au modèle HKY (21). Le format de l'arbre reconstruit par seqTrack était un tableau de cinq colonnes correspondant aux variables suivantes : id (indice de l'hôte infecté), ances (indice de son infecteur), weight (nombre de mutations séparant l'hôte infecté de son infecteur), date (date d'échantillonnage de l'hôte infecté), ances.date (date d'échantillonnage de son infecteur).

➢ outbreaker2

outbreaker2 est une méthode Bayésienne qui considère quatre vraisemblances : génétique, temporelle, d'observation des cas et de contact (121). Dans cette méthode, la probabilité de transmission est inférée à partir des distributions supposées connues du temps de génération et de l'intervalle infection-échantillonnage. Ici, nous supposons que les distributions non-paramétriques du temps de génération et de l'intervalle infection-échantillonnage pouvaient être obtenues sans biais à partir des arbres de référence.

Nous avons sélectionné une longueur de chaîne de 100 000 itérations, une fréquence d'échantillonnage de 1 sur 50 et une période de burn-in de 10 % (voir *Annexe 10* à *12* pour détails sur les paramètres et les distributions à priori utilisés). Nous avons vérifié graphiquement que la chaîne avait bien convergé ainsi que l'indépendance de l'échantillonnage (ESS > 200) après estimation des ESS avec le package coda (168). Quand les ESS étaient inférieures à 200, nous avons ajouté 100 000 itérations à la chaîne puis vérifié à nouveau les ESS. Cette étape a été répétée jusqu'à l'obtention d'une valeur supérieure à 200 pour chaque ESS.

Nous avons ensuite construit l'arbre consensus, comme cela était suggéré par les auteurs : l'infecteur le plus fréquent a été estimé pour chaque hôte infecté ainsi que la probabilité postérieure de chaque évènement de transmission retenu. Par construction, des cycles peuvent être présents dans cet arbre consensus (qui devient alors un graphe orienté), ce qui veut dire qu'un hôte infecté peut à la fois être l'ancêtre et le descendant d'un même hôte. De plus, cette méthode prenant en compte l'observation imparfaite des cas, la proportion d'hôtes échantillonnés est estimée et des hôtes non échantillonnés peuvent être présents de manière indirecte dans l'arbre consensus, sous la forme d'un nombre de générations séparant deux hôtes échantillonnés.

Le format de l'arbre consensus reconstruit par *outbreaker2* était un tableau de cinq colonnes correspondant aux variables suivantes : from (indice de l'infecteur), to (indice de l'hôte infecté), support (probabilité de

transmission), time (date d'infection estimée) et generations (nombre d'hôtes intermédiaires + 1).

➢ TransPhylo

TransPhylo, une autre méthode Bayésienne, infère les hôtes infectés le long des branches de l'arbre phylogénétique reconstruit au préalable (SeqPF) (141).

• Étape préliminaire de reconstruction phylogénétique

Nous avons reconstruit les arbres phylogénétiques à partir des séquences simulées en utilisant BEAST2 v2.6.3 (95). Nous avons choisi un modèle d'horloge stricte, un modèle de substitution HKY ainsi qu'un modèle de population skyline (avec quatre intervalles) Bayésien (plus approprié étant donné la multitude des profils épidémiques simulés). Nous avons choisi une fréquence empirique des bases dans le modèle de substitution et les autres paramètres conservaient leur valeur par défaut. Nous avons sélectionné une longueur de chaîne de 500 millions d'itérations, une période de burn-in de 10 % et une fréquence d'échantillonnage de 1 sur 50 000. Nous avons vérifié graphiquement que la chaîne avait bien convergé ainsi que l'indépendance de l'échantillonnage (ESS > 200) sur TRACER v.1.7.1 (98). Quand les ESS étaient inférieures à 200 mais supérieures à 100, nous avons ajouté 300 millions d'itérations à la chaîne puis vérifié à nouveau les ESS. Après cette étape, si la chaîne ne convergeait toujours pas, l'arbre de référence concerné était exclu. Afin de résumer les arbres postérieurs échantillonnés, nous avons construit l'arbre MCC grâce à Tree Annotator, en sélectionnant l'option de la hauteur des ancêtres communs (99).

• Reconstruction de deux types de résultats

Nous avons supposé que le temps de génération et l'intervalle infection-échantillonnage suivaient une distribution Gamma et que la moyenne et la déviation standard pouvaient être obtenues sans biais à partir des arbres de référence en utilisant le package *epitrix* (169). Nous avons sélectionné une longueur de chaîne de 500 000 itérations, une fréquence d'échantillonnage de 1 sur 50 et une période de burn-in de 20 % (voir *Annexe 10* à *12* pour détails sur les paramètres et les distributions à priori utilisés). Nous avons utilisé la même méthode que pour outbreaker2 afin de vérifier la convergence de la chaîne ainsi que l'indépendance d'échantillonnage. Cependant, nous avons considéré un seuil de 100 pour les ESS, comme cela avait été suggéré par les auteurs (170). Quand les ESS avaient une valeur inférieure à 100, nous avons ajouté 500 000 itérations à la chaîne puis vérifié

à nouveau les valeurs des ESS. Cette étape était répétée jusqu'à satisfaction des paramètres de convergence et d'indépendance d'échantillonnage ou jusqu'à l'obtention d'une longueur de chaîne supérieure à 2 500 000 itérations. Si les paramètres de convergence n'étaient pas satisfaits, les arbres de référence étaient exclus.

- *Arbre médoïde*

Puis, comme décrit par Didelot *et al.* (170), nous avons sélectionné l'arbre médoïde (arbre qui différait le moins des autres arbres postérieurs selon une mesure de distances définie par Kendall *et al.* (171)). Puisque cette méthode prend en compte la présence d'hôtes non échantillonnés lors de l'inférence des hôtes le long des branches, des hôtes non échantillonnés peuvent être présents directement dans l'arbre de transmission médoïde. Cela signifie que dans l'arbre médoïde, contrairement à l'arbre consensus d'*outbreaker2*, les hôtes infectés non échantillonnés peuvent être responsables de plus d'un évènement de transmission. De la même manière qu'*outbreaker2* estime une proportion d'hôtes échantillonnés, *TransPhylo* estime la probabilité d'observer et d'échantillonner un hôte infecté. Cette probabilité d'échantillonnage, dans le cas d'une épidémie terminée (comme nous avons considéré ici), est la même pour tous les hôtes infectés et correspond donc à une proportion d'échantillonnage.

Le format de l'arbre médoïde reconstruit par *TransPhylo* était un tableau avec cinq colonnes correspondant aux variables suivantes : tinfection (date d'infection estimée), tremoved (date de retrait de l'hôte infecté), infector_id (identifiant de l'infecteur) et infected_id (identifiant de l'hôte infecté).

- *Matrice des probabilités de transmission*

A partir des arbres postérieurs échantillonnés, nous avons reconstruit une matrice n x n des probabilités de transmission avec la fonction computeMatWIW implémentée dans *TransPhylo*, où n désigne le nombre d'hôtes infectés échantillonnés. Puis, nous avons identifié pour chaque hôte infecté, l'infecteur ayant la probabilité la plus haute de l'infecter. Si cette probabilité était nulle, nous avons considéré que l'infecteur le plus probable était inconnu. Comme avec *outbreaker2*, des cycles peuvent être présents dans la matrice ainsi reconstruite. De plus, la date d'infection n'étant pas estimée, nous ne pouvons pas inférer le cas index à partir de cette matrice.

4.2.3 Information génétique et indicateurs épidémiologiques

4.2.3.1 Information génétique

Afin de comprendre l'influence du modèle de simulation des séquences sur la reconstruction d'arbres de transmission et faciliter la comparaison avec d'autres travaux, nous avons quantifié la diversité génétique présente dans les ensembles de cas. Nous avons estimé la proportion de séquences uniques dans chaque ensemble de cas de référence ainsi que la moyenne de l'écart après transmission (ou « transmission divergence »). L'écart après transmission était défini par Campbell *et al.* (55) comme le nombre de SNP séparant deux individus liés par un lien de transmission direct. A partir des arbres de référence, nous avons identifié tous les hôtes liés par un lien de transmission direct puis nous avons calculé pour chaque arbre de référence, le nombre moyen de SNP séparant deux hôtes ainsi liés.

4.2.3.2 Indicateurs épidémiologiques

TransPhylo produit deux sorties (l'arbre médoïde et la matrice des probabilités de transmission), nous avons considéré la matrice pour évaluer l'exactitude de la méthode et l'arbre médoïde pour tous les autres indicateurs.

➢ Exactitude

Afin d'évaluer la performance des trois méthodes de reconstruction, nous avons d'abord déterminé les évènements de transmission corrects pouvant être reconstruits dans chaque ensemble de cas. Pour un ensemble de cas de référence, ces évènements de transmission corrects étaient ceux présents dans l'arbre de référence. Cependant, pour un ensemble de cas biaisé, nous avons considéré qu'un évènement de transmission correct connectait deux cas observés, ne prenant ainsi pas en compte les cas intermédiaires non échantillonnés. Par exemple, la chaîne de transmission Hôte n°1 (échantillonné) → Hôte n°2 (non observé) → Hôte n°3 (échantillonné) devient Hôte n°1 (échantillonné) → Hôte n°3 (échantillonné). Pour les trois méthodes, nous avons évalué la proportion des paires infecteur-infecté (cela veut dire tous les couples "id"-"ances" pour seqTrack, "from"-"to" pour *outbreaker2* et "infector_id"-"infected_id" pour *TransPhylo*) correspondant à l'un de ces évènements de transmission corrects.

➢ Présence de super-spreaders

Nous avons considéré que des super-spreaders étaient présents dans l'arbre reconstruit quand moins de 10 % des hôtes infectés étaient responsables de plus de 80 % des évènements de transmission. De plus, quand des super-spreaders étaient présents, nous avons identifié le nombre maximum d'évènements de transmission dont un seul hôte pouvait être responsable, ainsi que l'espèce-hôte du super-spreader le plus prolifique.

➢ Espèce-hôte du cas index

Nous avons évalué la capacité de ces trois méthodes à reconstruire l'espèce-hôte du cas index (bovin). Contrairement à l'arbre médoïde de *TransPhylo*, dans lequel il est aisé d'identifier le cas index (« infected_id » avec la date d'infection la plus ancienne), la présence de cycles dans *outbreaker2* et la possibilité de reconstruire de multiples cas index dans seqTrack compliquent l'identification du cas index. Pour seqTrack, nous avons considéré l'espèce-hôte la plus fréquente parmi les cas index reconstruits (« id » pour lesquels « ances » est inconnu). Pour *outbreaker2*, nous avons considéré l'espèce-hôte la plus fréquente parmi les hôtes infectés à la date la plus ancienne ("to" avec le plus ancien "time").

➢ Taille de l'épidémie

Nous avons évalué la capacité de *TransPhylo* et *outbreaker2* à estimer la taille de l'épidémie (seqTrack n'estime pas la taille de l'épidémie et a été exclu pour cet indicateur). La taille simulée de l'épidémie correspondait au nombre d'hôtes infectés présents dans chaque arbre de référence. Nous avons calculé l'estimation correspondante en divisant le nombre d'hôtes échantillonnés dans chaque arbre reconstruit par la médiane de la proportion d'échantillonnage estimée par *outbreaker2* et *TransPhylo*. De plus, nous avons testé si les résultats pour cet indicateur différaient si nous considérions l'épidémie reconstructible ou l'arbre de référence. Nous avons ainsi calculé le nombre d'hôtes infectés présent dans l'épidémie reconstructible puis nous l'avons comparé au nombre d'hôtes (échantillonnés ou non) présents dans les arbres reconstruits par *outbreaker2* et *TransPhylo*.

➢ Contribution des espèces-hôtes

Etant donné l'importance d'identifier l'espèce-hôte ayant contribué le plus à la transmission au sein d'un système multi-hôtes, nous avons évalué la capacité de *TransPhylo* et *outbreaker2* à estimer le nombre d'évènements de transmission dus à chaque espèce-hôte. Comme pour la taille de l'épidémie

et pour les mêmes raisons, seqTrack était exclu. Le nombre d'évènements de transmission dus à chaque espèce-hôte était calculé dans les arbres de référence. Nous avons calculé l'estimation correspondante pour chaque arbre reconstruit en divisant le nombre d'évènements de transmission entre hôtes échantillonnés par la médiane de la proportion d'échantillonnage estimée par *outbreaker2* et *TransPhylo*. Nous avons ensuite calculé le nombre d'évènements de transmission dus à chaque espèce-hôte dans l'épidémie reconstructible. Ce nombre a été comparé au nombre d'évènements de transmissions (vers des hôtes échantillonnés ou non) dus à chaque espèce-hôte, présents dans les arbres reconstruits par les deux méthodes.

➢ Analyses statistiques

Pour la taille de l'épidémie et la contribution des espèces-hôtes, nous avons calculé un intervalle de crédibilité en utilisant les bornes de l'intervalle 95%HPD de la proportion d'échantillonnage. Pour chaque arbre reconstruit, nous avons évalué si l'intervalle calculé contenait la taille réelle de l'épidémie ou le nombre d'évènements de transmission dus à chaque espèce-hôte.

Pour tous les indicateurs épidémiologiques sauf la présence de super-spreaders, nous avons testé l'effet sur la valeur de l'indicateur, de la méthode de reconstruction ainsi que son interaction avec l'effet du schéma d'échantillonnage. Afin de prendre en compte la non-indépendance des arbres reconstruits (six ensembles de cas ayant été construits à partir de chacun des arbres de référence), nous avons utilisé un modèle à effets mixtes (GLMM pour « Generalized Linear Mixed Model »), l'effet aléatoire considéré était l'identifiant de l'arbre de référence. Pour l'exactitude et le cas index, nous avons sélectionné une distribution Binomiale et la probabilité de reconstruire un évènement de transmission correct (ou l'espèce-hôte correcte pour le cas index) était la variable à expliquer. Du fait d'une surdispersion dans les estimations de la taille d'épidémie et du nombre d'évènements de transmission dus à chaque espèce-hôte, nous avons sélectionné une distribution Binomiale Négative pour ces deux indicateurs. Afin de comparer les estimations avec les valeurs de référence (de l'arbre de référence ou de l'épidémie reconstructible), ces valeurs de référence ont été prises comme offset et l'intercept était fixé à zéro. Les IRR (rapport des taux d'incidence) estimés pouvaient alors être interprétés comme des coefficients multiplicateurs de la taille de l'épidémie (ou de la contribution des espèces-hôtes) dans l'arbre de référence (ou l'épidémie reconstructible).

4.2.4 Scénarios de transmission alternatifs

Nous avons testé l'influence du taux de mutation faible, caractéristique de *M. bovis*, sur nos résultats. Pour cela, nous avons simulé de nouvelles séquences le long des 30 arbres de référence en multipliant le taux de mutation par 10 (μ_h = 0,024 substitutions par site par an), puis nous avons implémenté les trois méthodes de reconstruction d'arbres de transmission sur les ensembles de cas de référence.

Afin de tester si l'estimation de la taille de l'épidémie et l'exactitude étaient influencées par la complexité du système épidémiologique, nous avons ensuite simulé 30 nouveaux arbres de référence d'un système mono-hôte en fixant les paramètres de transmission concernant la faune sauvage à zéro. Nous avons ainsi obtenu 30 arbres ne contenant que des bovins. Nous avons ensuite simulé des séquences le long de ces arbres avec μ (0,0024 substitutions par site par an), puis nous avons implémenté les trois méthodes sur ces séquences.

Nous avons ensuite testé si l'asymétrie des rôles des différentes espèces pouvait influencer la reconstruction de la contribution des espèces-hôtes. Avec le même protocole (30 nouveaux arbres de référence et un taux de mutation faible), nous avons testé le scénario de transmission où l'une des espèces-hôtes (le sanglier) pouvait s'infecter mais pas transmettre l'agent pathogène (cul-de-sac épidémiologique), en fixant les paramètres entre et à partir des sangliers à zéro. Nous avons ainsi obtenu un système multi-hôtes dans lequel les sangliers ne jouaient aucun rôle dans la transmission de la bTB.

Finalement, afin d'évaluer la reconstruction de l'espèce-hôte du cas index, nous avons simulé 30 nouveaux arbres de référence avec des blaireaux comme cas index, dans le système multi-hôtes où chaque espèce-hôte contribuait à la transmission.

Tableau 9 : Définition des expressions utilisées dans l'étude (par ordre de présentation dans le matériel et méthodes).

Arbre (de transmission) de référence	Liste de six variables (identifiant de l'infecteur, identifiant de l'hôte infecté, espèce-hôte de l'infecteur, espèce-hôte de l'infecté, date d'infection et date de mort) obtenue avec le modèle de transmission modifié (développé à l'origine par Bouchez-Zacria *et al.* (160)).
Ensemble de cas de référence	Liste de quatre variables (identifiant de l'hôte infecté, espèce-hôte de l'infecté, date de mort et séquence échantillonnée) obtenue à partir de l'arbre de référence après simulation des séquences.
Schéma d'échantillonnage	Un des six processus de sélection appliqué à un ensemble de cas de référence, cinq de ces processus reproduisent des biais d'observation identifiés dans les données épidémiologiques de bTB.
Ensemble de cas biaisé	Ensemble de cas obtenu après application d'un schéma d'échantillonnage biaisé à un ensemble de cas de référence.
Épidémie reconstructible	Une sous-partie de l'arbre de référence qui ne contient que les hôtes infectés échantillonnés et leurs ancêtres.
Scenario de transmission	Combinaison de critères : type de système épidémiologique (multi-hôtes ou mono-hôte), contribution des espèces-hôtes à la transmission (présence ou absence de culs-de-sac épidémiologiques), espèce-hôte du cas index (bovin ou blaireau) et le taux de mutation (faible ou élevé).

4.3 RESULTATS

4.3.1 Reconstruction d'arbres de transmission

La convergence n'a pas été un facteur limitant pour *outbreaker2*. A l'opposé, elle n'a pas pu être obtenue pour tous les ensembles de cas dans BEAST2 ni tous les arbres phylogénétiques consensus avec *TransPhylo*. Nous avons ainsi été restreints à 21 des 30 arbres de référence (126 arbres reconstruits au total). Les arbres de référence pour lesquels nous n'avons pas

pu reconstruire les arbres de transmission avec *TransPhylo* montraient une médiane du nombre d'hôtes infectés plus élevée que les arbres de référence qui ont convergé (*Tableau 10*).

Tableau 10 : Comparaison des arbres de référence en fonction de leur convergence ou non dans BEAST2 et TransPhylo. Valeur médiane (plage des valeurs).

Paramètre	**Arbres ayant convergé**	**Arbres n'ayant pas convergé**
Nombre d'arbres	21	9
Nombre d'hôtes infectés	245 (118-458)	336 (114-376)
Proportion de séquences uniques	6,1 (1,8-8,7)	5,8 (3,4-8,1)
Moyenne de l'écart après la transmission (en nombre de SNP)	0,19 (0,11-0,28)	0,18 (0,09-0,25)

Le temps de calcul variait en fonction de l'ensemble de cas (ou l'arbre phylogénétique) considéré ainsi qu'en fonction de la méthode : moins de 10 minutes pour les 126 arbres reconstruits par seqTrack, de moins de 20 minutes (quand 100 000 itérations avaient suffi) à plus de deux heures par arbre reconstruit par *outbreaker2*, et enfin de moins d'une heure à plus de 12 heures (pour 2 500 000 itérations) par arbre reconstruit par *TransPhylo*. De plus, la reconstruction d'un arbre phylogénétique nécessaire avant l'implémentation de *TransPhylo* variait également en fonction de l'ensemble de cas : de cinq heures à deux jours. Au total, le temps de calcul pour ces 378 (126*3) arbres reconstruits s'est élevé à trois mois environ.

La proportion médiane de séquences uniques dans les ensembles de cas de référence ayant convergé était de 6,1 % (*Tableau 10*). La médiane de l'écart moyen après transmission était de 0,19 (*Tableau 10*) et la majorité des paires de transmission partageait la même séquence (Annexe *13*).

Tous les arbres reconstruits par *outbreaker2* ainsi que toutes les matrices de probabilités de transmission estimées par *TransPhylo* et pour lesquelles nous avons gardé l'infecteur le plus probable, présentaient des cycles.

4.3.2 Indicateurs épidémiologiques

4.3.2.1 Exactitude

Quand toutes les séquences étaient échantillonnées, la proportion médiane d'évènements de transmission correctement reconstruits (*Figure 21*) était de 3,4 % (valeurs allant de 1,3 à 12,1) pour les arbres reconstruits par seqTrack, de 8,0 % (2,2-11,3) pour *outbreaker2* et de 8,9 % (6,0-16,8) pour *TransPhylo* (Annexe *14*).

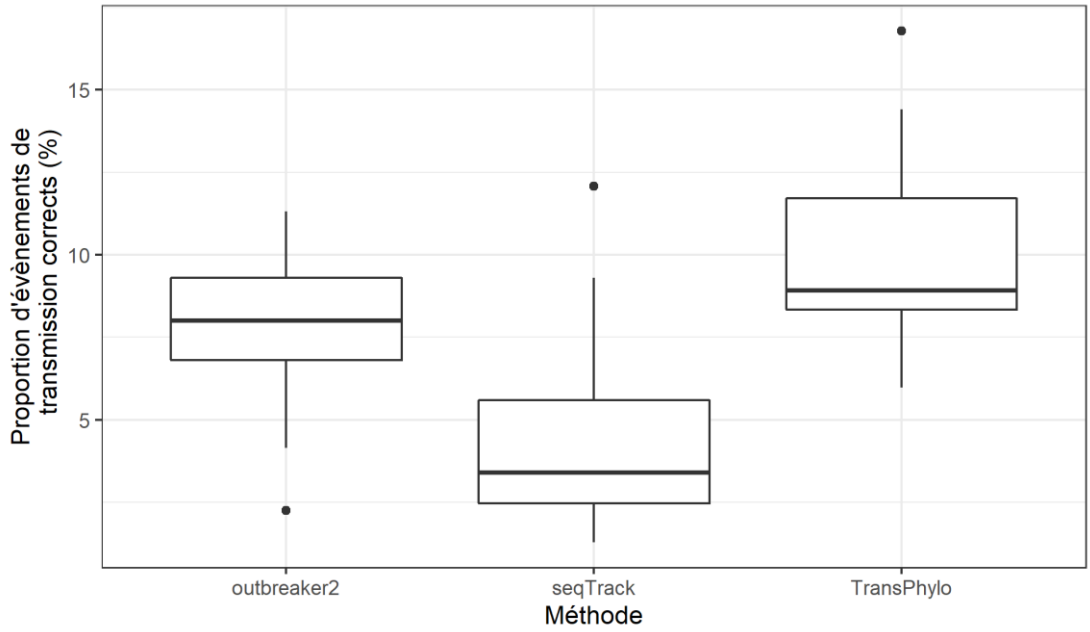


Figure 21 : Proportion d'évènements de transmission reconstruits à partir de toutes les séquences et présents dans l'arbre de référence en fonction de la méthode de reconstruction utilisée.

Comparée à *outbreaker2*, la probabilité de reconstruire un évènement de transmission correct était significativement plus basse pour seqTrack (OR=0,51, p<0,001) mais significativement plus élevée pour *TransPhylo* (OR=1,30, p<0,001) (*Tableau 11*). Dans les arbres reconstruits par seqTrack, quand les sangliers n'étaient pas échantillonnés, la probabilité de reconstruire un évènement de transmission correct augmentait significativement (OR=1,37 et 1,30, p=0,001 et 0,008 pour S_W et $T+S_W$ respectivement). Les résultats n'ont pas montré d'effet significatif du schéma d'échantillonnage pour les deux autres méthodes (*Tableau 11*).

Tableau 11 : Présence des évènements de transmission reconstruits dans les arbres de référence testée avec un GLMM Binomial, en utilisant la méthode ainsi que l'interaction entre la méthode et le schéma d'échantillonnage comme effets fixes. OR signifie « odds ratio ». Les nombres en gras signalent les OR significativement différents de 1 ($p<0,05$). La méthode *outbreaker2* et le schéma d'échantillonnage de référence étaient considérés comme la référence, d'où le "-" présent à la ligne sur les effets des méthodes et l'absence du schéma de référence. T signifie "biais temporel", S_B "biais blaireau" et S_W "biais sanglier". $T+S_B$ ($T+S_W$) combine le biais temporel et le biais blaireau (sanglier).

Effets fixes	*outbreaker2*		seqTrack		*TransPhylo*	
	OR	p	OR	p	OR	p
Méthode	-	-	**0,51**	**<0,001**	**1,30**	**<0,001**
Méthode :T	0,99	0,89	0,94	0,54	0,91	0,17
Méthode :S_B	0,91	0,21	0,95	0,60	1,08	0,30
Méthode :$T+S_B$	0,92	0,32	0,94	0,57	1,11	0,14
Méthode :S_W	1,08	0,33	**1,37**	**0,001**	1,03	0,64
Méthode :$T+S_W$	1,07	0,41	**1,30**	**0,008**	1,01	0,88

4.3.2.2 Super-spreaders

Dans les arbres de référence, aucun super-spreader n'était présent et le nombre maximum d'évènements de transmission dus à un seul hôte infecté était compris entre 9 et 27 (médiane : 14). Tous les arbres reconstruits par seqTrack présentaient des super-spreaders. La médiane du nombre maximum d'évènements de transmission dus à un seul super-spreader était comprise entre 90 et 108, alors que la médiane du nombre d'évènements de transmission présents dans un arbre reconstruit était comprise entre 200 et 244 (Annexe *15*). Le super-spreader le plus prolifique était très fréquemment un bovin (57 % des super-spreaders les plus prolifiques dans le schéma de référence jusqu'à 86 % lorsque le biais temporel et le biais sanglier étaient combinés). Aucun des arbres reconstruits par les deux autres méthodes ne présentait de super-spreaders.

4.3.2.3 Espèce-hôte du cas index

Quand toutes les séquences étaient échantillonnées, la proportion des cas index pour lesquels l'espèce-hôte était correctement reconstruite (bovin) était de 76 % pour seqTrack, 81 % pour *outbreaker2* et 57 % pour *TransPhylo* (*Tableau 12*). Mis à part le cas où l'on considère un biais temporel

seul avec *TransPhylo*, les biais temporel et blaireau (combinés ou non) augmentaient la proportion de cas index correctement reconstruits.

Tableau 12 : Pourcentage (%) de cas index pour lesquels l'espèce-hôte était correctement reconstruite en fonction de la méthode et du schéma d'échantillonnage. T signifie "biais temporel", S_B "biais blaireau" et S_W "biais sanglier". T+S_B (T+S_W) combine le biais temporel et le biais blaireau (sanglier).

Schéma d'échantillonnage	*outbreaker2*	seqTrack	*TransPhylo*
Référence	81	76	57
T	95	100	52
S_B	100	100	91
T+S_B	100	100	86
S_W	81	76	57
T+S_W	95	100	81

4.3.2.4 Taille de l'épidémie

Dans les arbres de référence, le nombre médian d'hôtes infectés était de 245 (Annexe *16*). La taille de l'épidémie simulée était globalement cohérente avec l'intervalle de crédibilité estimé par *outbreaker2* (*Figure 22*). Ainsi, cet intervalle de crédibilité contenait la taille de l'épidémie simulée pour tous les arbres reconstruits avec le schéma d'échantillonnage de référence et le biais temporel. Cependant, un biais d'espèce (combiné ou non à un biais temporel) diminuait le nombre d'arbres qui estimaient correctement la taille de l'épidémie, et conduisait à une majorité d'arbres qui sous-estimaient la taille de l'épidémie (20/21 avec S_B et T+S_B, 16/21 avec S_W et 18/21 avec T+S_W). Cependant, selon le modèle statistique, la taille de l'épidémie estimée par *outbreaker2* n'était pas statistiquement différente de la taille de l'arbre de référence (IRR=1,14, p=0,43) et les schémas d'échantillonnage n'avaient pas d'effet significatif sur la taille de l'épidémie estimée (*Tableau 13*).

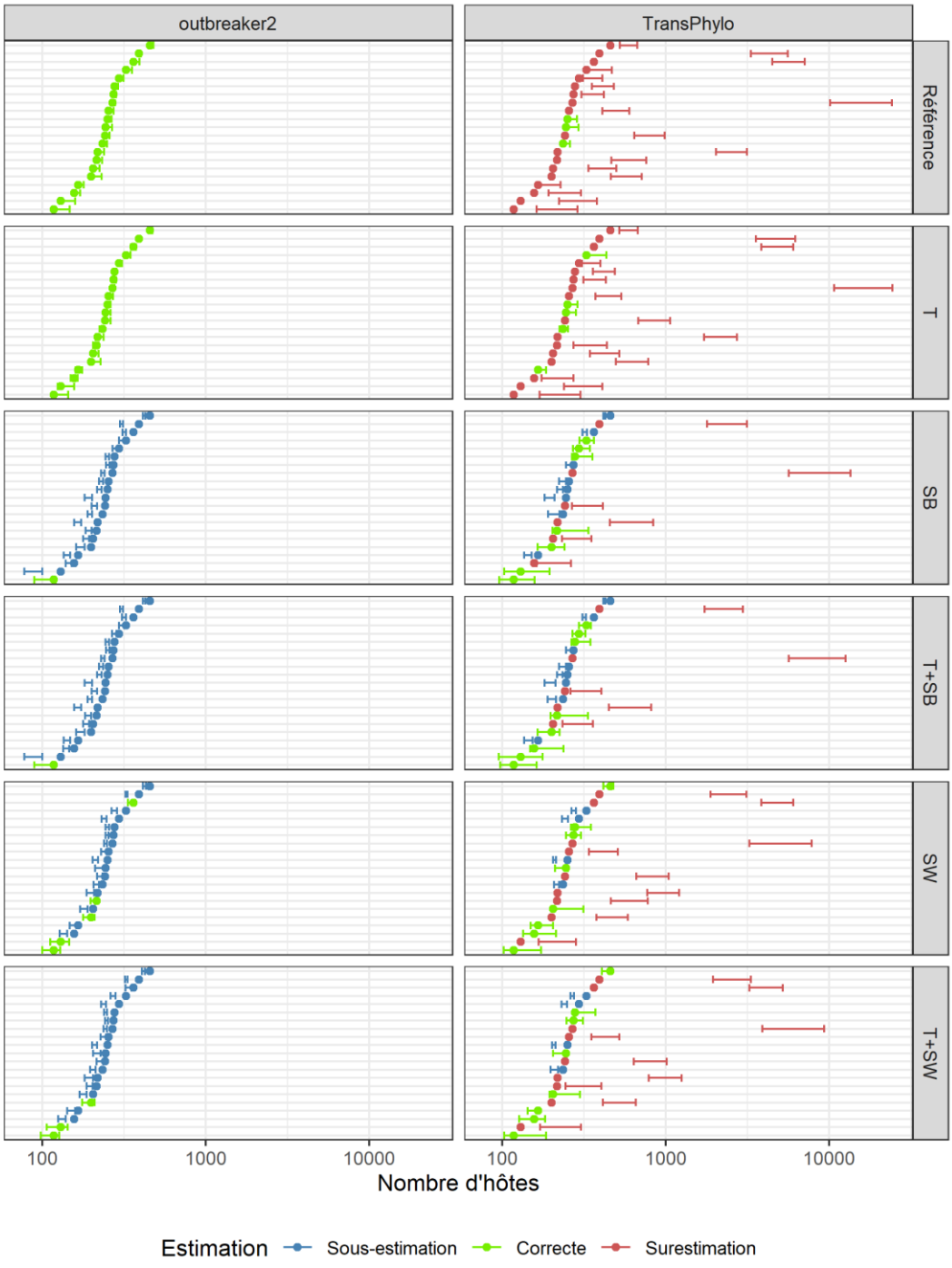


Figure 22 : Intervalle de crédibilité de la taille de l'épidémie estimée par outbreaker2 *et* TransPhylo *comparé (couleur) à la taille de l'épidémie simulée (point), en fonction du schéma d'échantillonnage.* T signifie "biais temporel", S_B "biais blaireau" et S_W "biais sanglier". T+S_B (T+S_W) combine le biais

temporel et le biais blaireau (sanglier). Chaque ligne représente un arbre reconstruit.

TransPhylo pouvait fortement surestimer la taille de l'épidémie et la différence entre la borne inférieure de l'intervalle et la taille de l'épidémie simulée pouvait excéder 10 000 hôtes infectés (*Figure 22*). Le schéma d'échantillonnage de référence et temporel conduisaient à une surestimation de la taille de l'épidémie dans la majorité des arbres reconstruits (19/21 et 16/21) : les intervalles de crédibilité contenaient la taille de l'épidémie simulée dans respectivement trois et cinq des 21 arbres. Le nombre d'estimations correctes restait bas avec les autres types de biais. Le modèle statistique a confirmé ces résultats, puisque la taille de l'épidémie estimée par *TransPhylo* était significativement plus élevée que la taille de l'épidémie simulée (IRR=2,92, p<0,001). De plus, tous les biais d'échantillonnage, à part le biais temporel, diminuaient de manière significative la taille de l'épidémie estimée (IRR compris entre 0,49 et 0,68, p<0,01).

Tableau 13 : Estimation de la taille de l'épidémie testée avec un GLMM Négatif Binomial, en utilisant la méthode ainsi que l'interaction entre la méthode et le schéma d'échantillonnage comme effets fixes. IRR signifie « rapport des taux d'incidence ». Les nombres en gras signalent les IRR significativement différents de 1 (p<0,05). Le schéma d'échantillonnage de référence était considéré comme la référence, d'où son absence. T signifie "biais temporel", S_B "biais blaireau" et S_W "biais sanglier". T+S_B (T+S_W) combine le biais temporel et le biais blaireau (sanglier).

Effets fixes	*outbreaker2*		*TransPhylo*	
	IRR	p	IRR	p
Méthode	1,14	0,43	**2,92**	**<0,001**
Méthode :T	0,98	0,90	0,96	0,80
Méthode :S_B	0,83	0,24	**0,50**	**<0,001**
Méthode :T+S_B	0,83	0,23	**0,49**	**<0,001**
Méthode :S_W	0,87	0,34	**0,68**	**0,01**
Méthode :T+S_W	0,85	0,28	**0,65**	**0,005**

4.3.2.5 Contribution des espèces-hôtes

Le nombre médian d'évènements de transmission dus à chaque espèce-hôte dans l'arbre de référence était de 175 pour les bovins, 24 pour les blaireaux et 40 pour les sangliers (Annexe *18*).

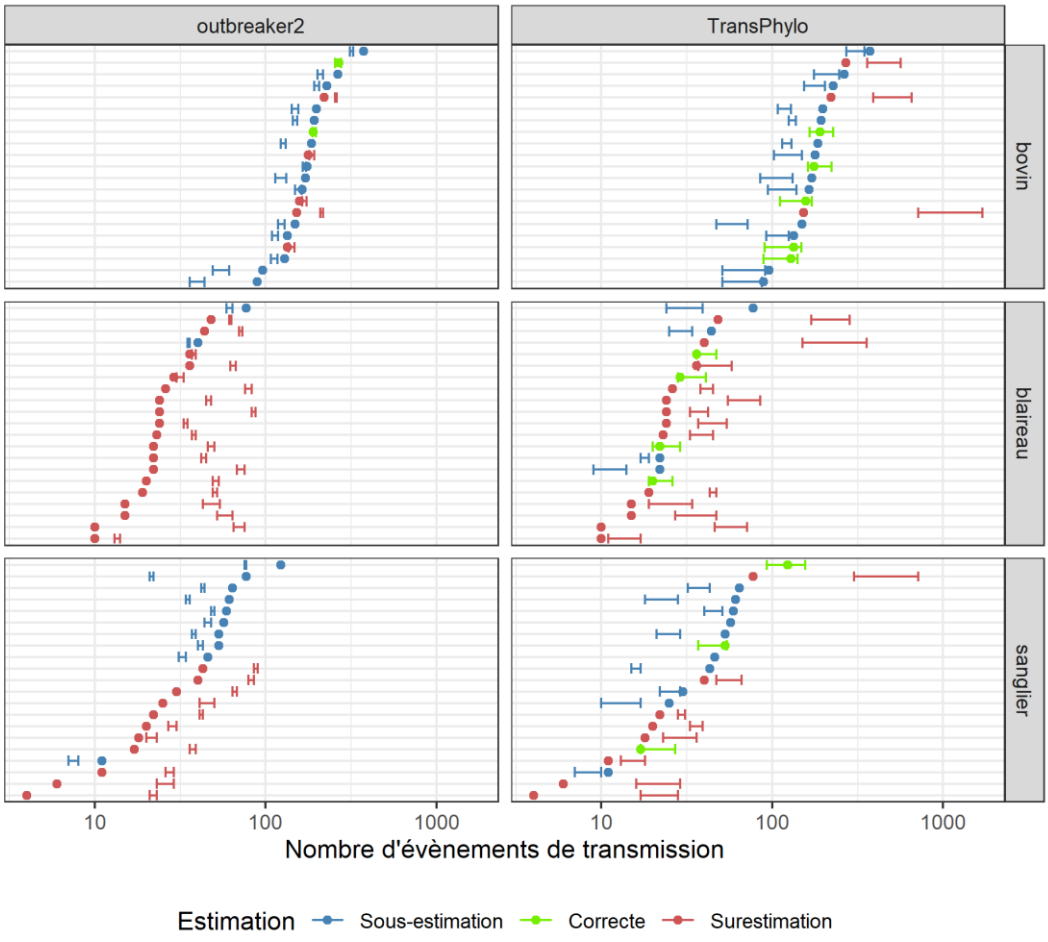


Figure 23 : Intervalle de crédibilité du nombre d'évènements de transmission dus à chaque espèce-hôte estimé par outbreaker2 *et* TransPhylo*, avec le schéma d'échantillonnage de référence, comparé (couleur) au nombre simulé (point), en fonction de l'espèce-hôte.* Chaque ligne représente un arbre reconstruit.

Avec le schéma d'échantillonnage de référence, l'intervalle de crédibilité contenait le nombre simulé d'évènements de transmission dus à chaque espèce-hôte dans peu d'arbres reconstruits par *outbreaker2* (2/21 arbres pour les bovins et aucun pour les blaireaux et les sangliers) et *TransPhylo* (5/21 arbres pour les bovins, 4/21 pour les blaireaux et 3/21 pour les sangliers) (*Figure 23*). Sinon, dans la majorité des arbres de transmission, ce nombre d'évènements de transmission était sous-estimé pour les bovins (14/21 arbres pour *outbreaker2* et 13/21 arbres pour *TransPhylo*), surestimé pour les blaireaux (19/21 arbres pour *outbreaker2* et 13/21 arbres pour *TransPhylo*) et aucune tendance n'était observée pour les sangliers. Des

résultats similaires étaient obtenus pour les cinq autres schémas d'échantillonnage (Annexe *19* à 20).

Tableau 14 : Nombre d'évènements de transmission dus à chaque espèce-hôte testé avec un GLMM Négatif Binomial par espèce-hôte, en utilisant la méthode ainsi que l'interaction entre la méthode et le schéma d'échantillonnage comme effets fixes. IRR signifie « rapport des taux d'incidence ». Les nombres en gras signalent les IRR significativement différents de 1 (p<0,05). Le schéma d'échantillonnage de référence était considéré comme la référence, d'où son absence. T signifie "biais temporel", S_B "biais blaireau" et S_W "biais sanglier". T+S_B (T+S_W) combine le biais temporel et le biais blaireau (sanglier).

Effets fixes	*outbreaker2*		*TransPhylo*	
	IRR	p	IRR	p
1. Contribution des bovins				
Méthode	0,86	0,09	0,95	0,58
Méthode :T	1,04	0,58	1,02	0,77
Méthode :S_B	1,06	0,41	0,95	0,45
Méthode :T+S_B	1,06	0,39	0,91	0,20
Méthode :S_W	1,01	0,91	1,03	0,63
Méthode :T+S_W	1,06	0,37	1,09	0,20
2. Contribution des blaireaux				
Méthode	**2,06**	**<0,001**	**1,70**	**<0,001**
Méthode :T	0,80	0,09	0,84	0,20
Méthode :S_W	1,12	0,39	0,91	0,47
Méthode :T+S_W	0,87	0,27	**0,74**	**0,02**
3. Contribution des sangliers				
Méthode	1,33	0,12	1,04	0,85
Méthode :T	0,92	0,62	1,29	0,13
Méthode :S_B	1,08	0,65	1,06	0,72
Méthode :T+S_B	1,03	0,87	1,05	0,80

D'après le modèle statistique, la sous-estimation du nombre d'évènements de transmission dus aux bovins (*Figure 23*) n'était significative pour aucune des deux méthodes (*Tableau 14*). Le modèle statistique a confirmé les résultats obtenus pour les blaireaux, puisque le nombre d'évènements de transmission estimé par les deux méthodes était significativement plus élevé que le nombre simulé (IRR=2,06 pour *outbreaker2* et 1,70 pour *TransPhylo,* p<0,001) (*Tableau 14*). Les résultats n'ont pas montré d'effet significatif du schéma d'échantillonnage sur la contribution des blaireaux pour *outbreaker2*. Cependant, le biais temporel avec le moins d'hôtes échantillonnés (combinaison des biais temporel et

sanglier) diminuait significativement le nombre d'évènements de transmission dus aux blaireaux estimé par *TransPhylo*. Enfin, le nombre d'évènements de transmission dus aux sangliers estimé par les deux méthodes n'était pas significativement différent du nombre simulé dans l'arbre de référence (*Tableau 14*).

4.3.3 Scénarios de transmission alternatifs

4.3.3.1 Taux de mutation élevé

Les séquences simulées avec un taux de mutation plus élevé ont bien montré une proportion de séquences uniques (médiane : 33,4 %) et un écart moyen après transmission (médiane : 0,69) plus élevés (Annexe *23*).

Un taux de mutation plus élevé augmentait nettement la valeur médiane de l'exactitude pour les trois méthodes : 25,7 % (+17,7, plage de valeurs : 15,9-33,3) pour *outbreaker2*, 15,3 % (+11,9, 8,2-33,3) pour seqTrack et 21,3 % (+12,3, 13,2-29,3) pour *TransPhylo* (Annexe *24*). La majorité des arbres reconstruits par seqTrack contenaient à nouveau des super-spreaders (20/21), cependant la médiane du nombre maximum d'évènements de transmission dus à un seul super-spreader était plus faible avec un taux de mutation plus élevé (35 *vs.* 108, Annexe *26*).

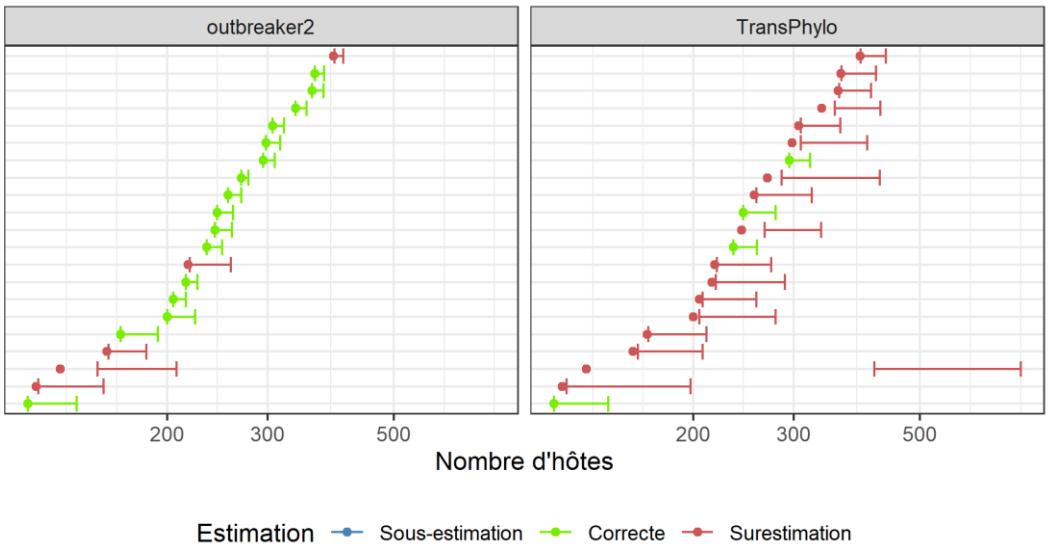


Figure 24 : Intervalle de crédibilité de la taille de l'épidémie estimée par outbreaker2 *et* TransPhylo *comparé (couleur) à la taille de l'épidémie simulée (point), avec le schéma d'échantillonnage de référence et un taux de mutation élevé.* Chaque ligne représente un arbre reconstruit.

L'intervalle de crédibilité contenait la taille de l'épidémie simulée pour 16/21 arbres reconstruits par *outbreaker2* et seulement 4/21 arbres reconstruits par *TransPhylo*, sinon la taille de l'épidémie était surestimée (*Figure 24* et Annexe *28*). Pour les deux méthodes, l'intervalle de crédibilité contenait le nombre d'évènements de transmission dus à chaque espèce-hôte dans seulement 4/21 arbres pour les bovins, 3/21 (*outbreaker2)* et 5/21 (*TransPhylo*) pour les blaireaux et 1/21 pour les sangliers (*Figure 25*). Sinon la contribution des bovins était sous-estimée par *TransPhylo* (16/21 arbres dans la *Figure 25*, Annexe *30*) et celle de la faune sauvage était surestimée par *outbreaker2* (13/21 arbres pour les blaireaux et les sangliers dans la *Figure 25*, Annexe *30*).

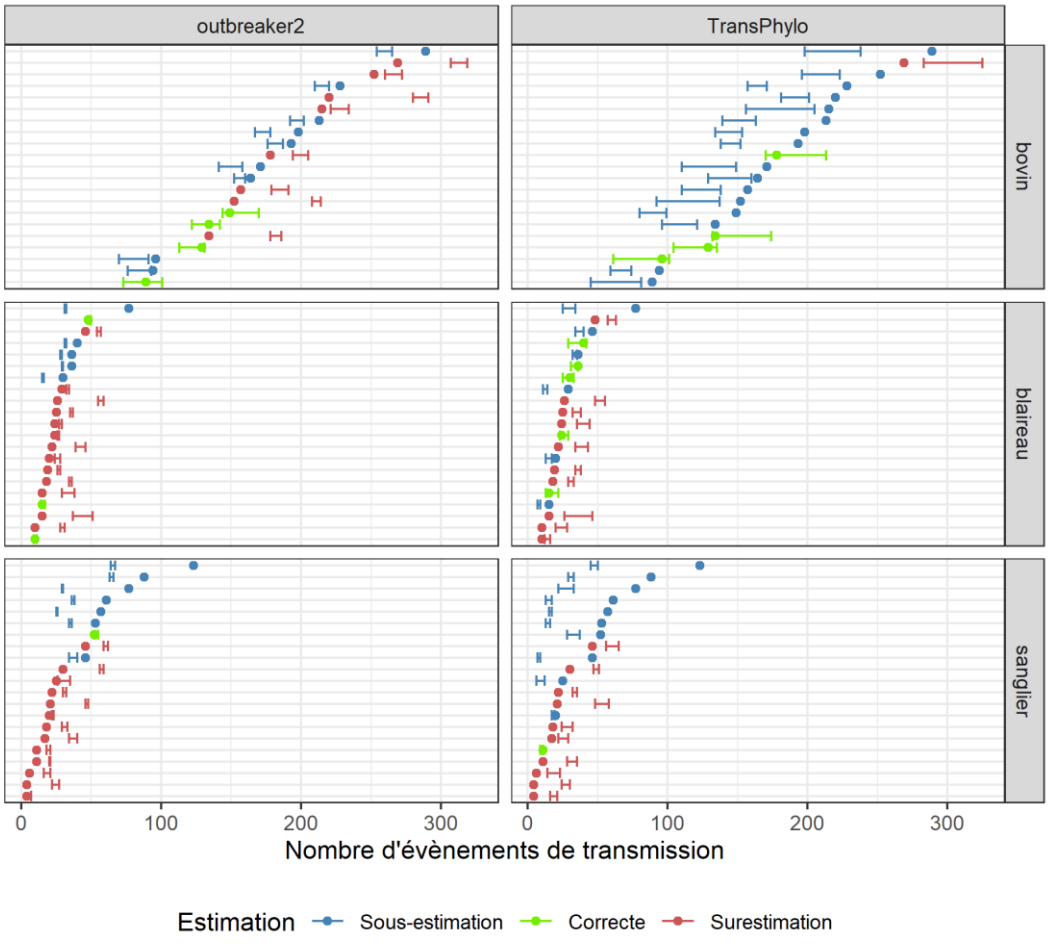


Figure 25 : Intervalle de crédibilité du nombre d'évènements de transmission dus à chaque espèce-hôte estimé par outbreaker2 *et* TransPhylo *comparé (couleur) au nombre simulé (point), en fonction de l'espèce-hôte, avec le schéma d'échantillonnage de référence et un taux de mutation élevé.* Chaque

ligne représente un arbre reconstruit.

4.3.3.2 Système simple composé de bovins

Les séquences simulées dans un système composé uniquement de bovins présentaient une proportion de séquences uniques (médiane : 3,6 %) et un écart moyen après transmission (médiane : 0,14) plus faibles (Annexe *23*).

De même que pour les systèmes multi-hôtes, l'exactitude était plus élevée pour *TransPhylo* (6,5 %), puis *outbreaker2* (5,5 %) et la plus basse pour seqTrack (2 %) (Annexe *24*). Cependant, des super-spreaders étaient présents dans tous les arbres reconstruits par seqTrack mais également dans certains des arbres reconstruits par *outbreaker2* (5/26 arbres, médiane du nombre maximum d'évènements de transmission dus à un seul super-spreaders : 39) et *TransPhylo* (10/26 arbres, médiane : 39,5). (Annexe *26*)

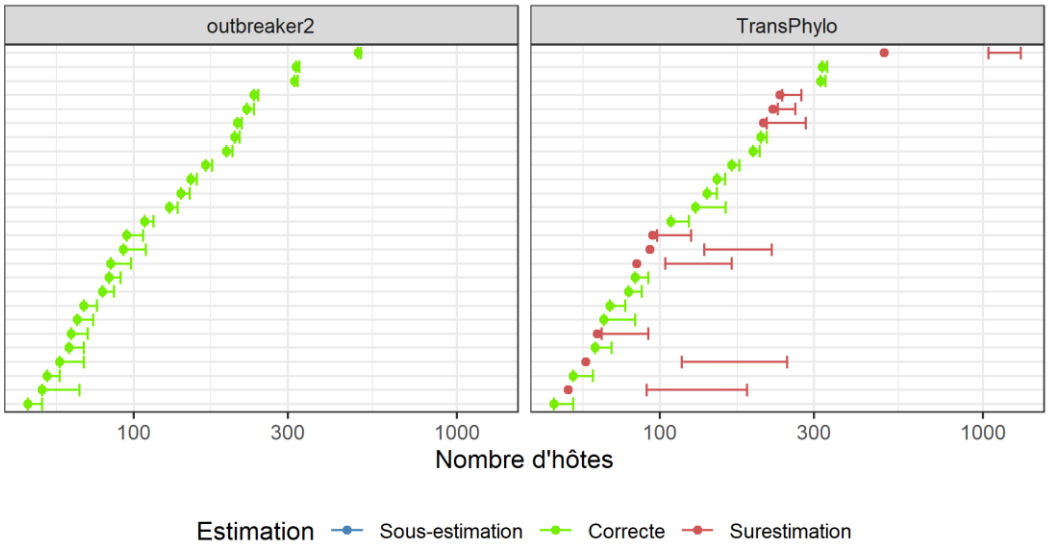


Figure 26 : Intervalle de crédibilité de la taille de l'épidémie estimée par outbreaker2 *et* TransPhylo *comparé (couleur) à la taille de l'épidémie simulée (point), avec le schéma d'échantillonnage de référence, dans le scénario de transmission impliquant uniquement des bovins.* Chaque ligne représente un arbre reconstruit.

L'intervalle de crédibilité contenait la taille de l'épidémie simulée pour les 26 arbres reconstruits par *outbreaker2* et pour 16/26 arbres reconstruits par *TransPhylo*, sinon la taille de l'épidémie était surestimée (*Figure 26*, Annexe *28*).

4.3.3.3 *Les sangliers comme culs-de-sac épidémiologiques*

L'intervalle de crédibilité ne contenait jamais le nombre simulé d'évènements de transmission dus aux sangliers et la contribution des sangliers était surestimée dans les 17 arbres reconstruits (*Figure 27*). Sinon, de même que pour les systèmes multi-hôtes sans culs-de-sac épidémiologiques, la contribution des bovins avait tendance à être sous-estimée par les deux méthodes (17/17 arbres pour *outbreaker2* et 10/17 arbres pour *TransPhylo*) et celle des blaireaux, surestimée par *outbreaker2* (15/17 arbres dans la *Figure 27* et Annexe *30*).

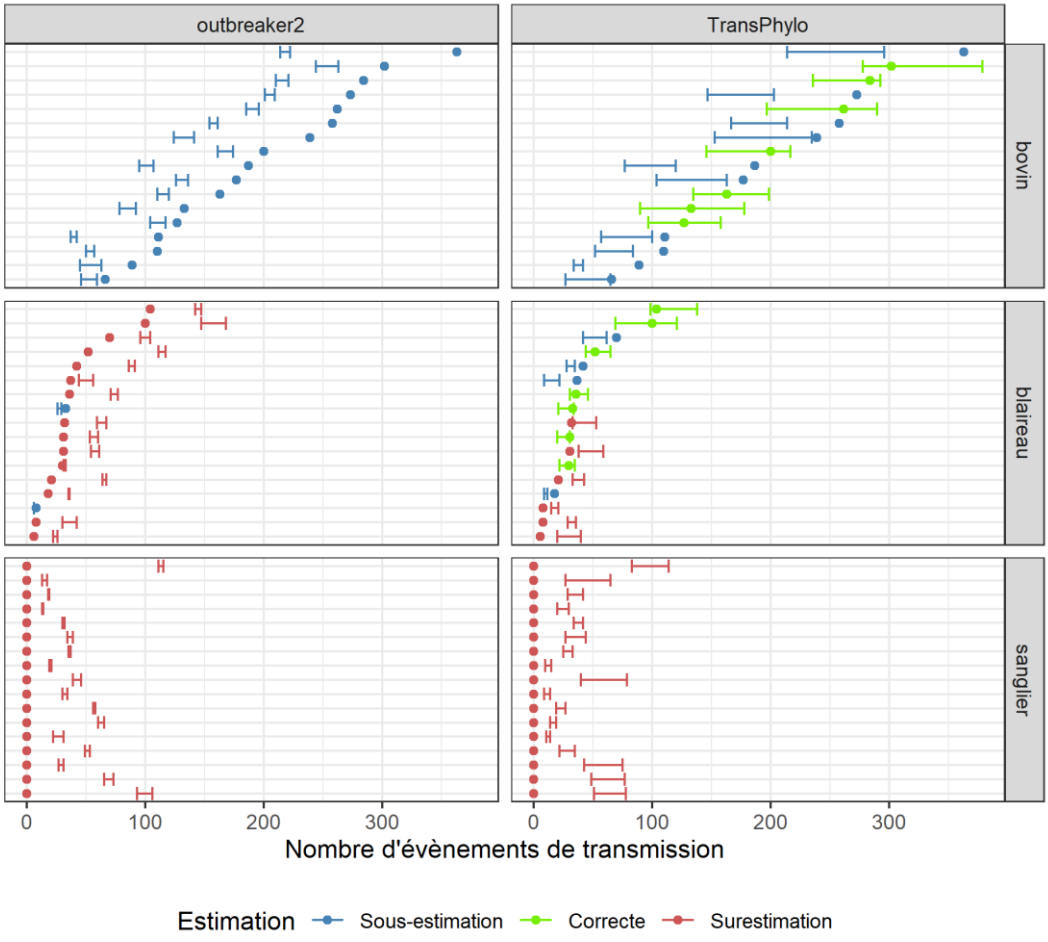


Figure 27 : Intervalle de crédibilité du nombre d'évènements de transmission dus à chaque espèce-hôte estimé par outbreaker2 *et* TransPhylo *comparé (couleur) au nombre simulé (point), en fonction de l'espèce-hôte, avec le schéma d'échantillonnage de référence, dans le scénario de transmission où le sanglier est un cul-de-sac épidémiologique.* Chaque ligne représente un arbre

reconstruit.

4.3.3.4 Des blaireaux en cas index

La proportion de cas index pour lesquels l'espèce-hôte était correctement reconstruite était nettement plus faible pour *outbreaker2* (81 % puis 28 %), seqTrack (76 % puis 28 %) et *TransPhylo* (57 % puis 11 %) avec des blaireaux en cas index (*Tableau 15*).

Tableau 15 : Pourcentage (%) de cas index pour lesquels l'espèce-hôte était correctement reconstruite en fonction du scénario de transmission.

Scénario de transmission	*outbreaker2*	seqTrack	*TransPhylo*
Bovins en cas index (n=21)	81	76	57
Blaireaux en cas index (n=18)	28	28	11

Pour tous les scénarios de transmission, y compris le scénario multi-hôtes de référence, des résultats similaires étaient obtenus, que l'on considère l'arbre de référence ou l'épidémie reconstructible (Annexe *17*, 21, 27 et 29).

4.4 DISCUSSION

Dans ce travail, nous avons évalué et comparé les performances de trois méthodes de reconstruction d'arbres de transmission sur des données de *M. bovis* simulées dans un système multi-hôtes. Nous avons également évalué l'impact de biais d'observation sur ces performances. *M. bovis*, caractérisé par un taux de mutation lent, est un excellent exemple d'agent pathogène multi-hôtes pour lequel des biais d'observation compliquent l'estimation de la contribution de chaque espèce-hôte à la transmission. Cette estimation est nécessaire afin de sélectionner des mesures de contrôle appropriées. Contrairement aux évaluations précédentes de méthodes de reconstruction, le modèle de transmission que nous avons utilisé pour simuler les données n'était pas adapté à une méthode en particulier (55,121,141) mais spécifique de l'agent pathogène multi-hôtes évoluant lentement. De plus, certains des indicateurs épidémiologiques que nous avons estimés étaient spécifiques des systèmes multi-hôtes, et pas seulement des indicateurs généraux de performance (172).

4.4.1 Reconstruction de « qui a infecté qui ? »

4.4.1.1 Exactitude

La reconstruction d'arbres de transmission peut avoir plusieurs

objectifs selon l'agent pathogène et le système épidémiologique étudié, l'objectif premier est la reconstruction exacte de « qui a infecté qui ? ». La proportion d'évènements de transmission correctement reconstruits (que nous avons appelé exactitude) a déjà été utilisée dans l'évaluation des performances de méthodes de reconstruction (121,172). Avec le taux de mutation lent caractéristique de *M. bovis*, nous avons estimé de mauvaises performances pour l'exactitude (médiane inférieure à 9 % pour les trois méthodes). Sobkowiak *et al*. ont comparé la performance de méthodes de reconstruction sur des données réelles de *M. tuberculosis,* qui est également un agent pathogène évoluant lentement. Ils ont estimé la valeur prédictive positive (VPP) qui correspondait à la proportion de paires de cas, ayant un lien épidémiologique, correctement identifiés dans l'arbre reconstruit (pré-print, (173)). Contrairement à l'exactitude que nous avons estimée, les liens n'étaient pas dirigés, nous nous attendions donc à une proportion de cas correctement reconstruits plus élevée dans cette étude que dans la nôtre. La VPP estimée par Sobkowiak *et al*. était de 15 % pour *TransPhylo*, 11 % pour *outbreaker2* et 10 % pour seqTrack. Ces valeurs de VPP étaient comprises dans les plages des valeurs d'exactitude que nous avons estimées et le classement des méthodes était le même que le nôtre (*TransPhylo* en premier, suivi par *outbreaker2*).

L'exactitude n'était que peu influencée par les biais d'échantillonnage ou la complexité du système épidémiologique, cependant elle dépendait grandement du taux de mutation. Quand le taux de mutation était multiplié par 10 (0,024 substitutions par site et par an soit ~6,6 x 10^{-5} substitutions par site et par jour), les valeurs estimées pour l'exactitude étaient plus que doublées. Dans l'étude qui présentait et testait *outbreaker2*, Campbell *et al.* ont estimé une proportion moyenne d'évènements de transmission correctement inférés quand seules des données temporelles et génétiques issues d'épidémies simulées d'Ebola (taux de mutation : 0,31 x 10^{-5} substitutions par site et par jour) et de SARS-CoV-1 (1,14 x 10^{-5} substitutions par site et par jour) étaient utilisées (121). De plus, Firestone *et al.* ont comparé les performances de *TransPhylo* et *outbreaker2* sur des données simulées de FA (2,2 x 10^{-5} substitutions par site et par jour). Ils ont estimé la proportion d'hôtes infectés (élevages) pour lesquels la source d'infection la plus probable était la source simulée (172). Ces deux indicateurs correspondent à ce que nous avons appelé exactitude, nous nous attendions donc à des résultats similaires. Cependant, Campbell *et al.* ont estimé une exactitude moyenne de 29 % (Ebola) et 70 % (SARS-CoV-1). De plus, l'exactitude estimée par Firestone *et al.* était de 4 % pour *TransPhylo* et 35 % pour *outbreaker2*. Par rapport aux plages de valeurs que nous avons estimées, ces valeurs sont respectivement plus élevées pour *outbreaker2* et

plus faibles pour *TransPhylo*. Cependant, le classement des méthodes obtenues par Firestone *et al.* était le même que le nôtre (à savoir *outbreaker2* en premier).

4.4.1.2 Présence de super-spreaders

L'exactitude la plus faible toujours estimée pour seqTrack pourrait être due à l'absence de modèle de transmission dans cette méthode (133), qui considère uniquement des dates d'échantillonnage et des distances génétiques. Comme évoqué par Nigsch *et al.*, seqTrack dépend donc fortement de l'ordre chronologique des dates d'échantillonnage, et lorsque l'ordre d'échantillonnage ne suit pas l'ordre d'infection (ici, du fait de la détection imparfaite des cas et du protocole changeant en fonction de l'espèce-hôte), « l'ordre des ancêtres ne peut pas être inféré avec certitude » (174). Contrairement à ce que nous avons observé avec *outbreaker2* et *TransPhylo,* les arbres reconstruits par seqTrack présentaient des super-spreaders avec des nombres extrêmes d'évènements de transmission dus à un seul hôte infecté (plus d'une centaine de transmissions). Ces nombres ont diminué lorsque nous avons considéré un taux de mutation plus élevé. La faible diversité génétique combinée à l'absence de modèle de transmission pourrait donc expliquer la reconstruction de super-spreaders, qui à son tour pourrait contribuer à la faible exactitude. De la même façon, la faible diversité génétique obtenue dans un système impliquant uniquement des bovins pourrait expliquer la présence de super-spreaders moins prolifiques dans les arbres reconstruits avec *TransPhylo* et *outbreaker2*.

Nous avons donc estimé de faibles exactitudes pour les trois méthodes. Cependant, reconstruire avec exactitude une proportion élevée de liens de transmission dirigés est difficile et n'est pas forcément l'objectif principal lors de l'étude d'un système multi-hôtes impliquant la faune sauvage ou avec une proportion faible d'hôtes échantillonnés. A l'inverse, la présence de super-spreaders est un indicateur important à considérer car il a permis de mettre en évidence des dynamiques de transmission peu réalistes (super-spreaders très prolifiques) reconstruites par seqTrack.

4.4.2 Indicateurs épidémiologiques

4.4.2.1 Taille de l'épidémie

Outre la reconstruction de « qui a infecté qui ? », les méthodes de reconstruction d'arbres de transmission peuvent chercher à estimer des indicateurs épidémiologiques à partir desquels il est possible d'inférer des mesures pratiques à mettre en œuvre. Le premier que nous avons étudié est

la taille de l'épidémie, qui en comparaison avec le nombre de cas échantillonnés peut indiquer par exemple la nécessité d'augmenter les efforts d'échantillonnage (159). L'estimation de la taille de l'épidémie était sensible aux biais d'échantillonnage, à la complexité du système épidémiologique et au taux de mutation. La taille de l'épidémie était correctement estimée par *outbreaker2* mais surestimée par *TransPhylo*, même si nous avons considéré la même distribution à priori non informative pour la proportion d'échantillonnage lors de l'implémentation des deux méthodes. Cette surestimation pourrait donc être due au fait que cette méthode a été développée pour étudier une épidémie de *M. tuberculosis* partiellement échantillonnée tout en prenant en compte la diversité intra-hôte (141). En effet, nous avons à l'inverse fait l'hypothèse que tous les cas étaient échantillonnés dans le schéma de référence et n'avons simulé aucune diversité intra-hôte. De plus, quand toutes les séquences n'étaient pas échantillonnées, de meilleurs résultats étaient obtenus pour *TransPhylo* et la taille de l'épidémie estimée diminuait significativement.

Avec un taux de mutation plus élevé, *TransPhylo* surestimait également la taille de l'épidémie mais dans une moindre mesure. Xu *et al.* ont développé en 2019 une méthode d'inférence simultanée sur plusieurs clusters de *M. tuberculosis* à partir de *TransPhylo*. A partir de cette étude, Xu *et al.* ont discuté le lien entre le taux de mutation et la proportion d'échantillonnage, en expliquant qu'en faisant l'hypothèse d'une horloge plus rapide, les branches dans les arbres phylogénétiques étaient plus courtes et *TransPhylo* était moins susceptible d'y placer des cas non échantillonnés (143). Ceci pourrait expliquer l'effet moindre que nous avons observé.

4.4.2.2 Espèce-hôte du cas index

Certains indicateurs épidémiologiques sont pertinents uniquement dans un système multi-hôtes et révèlent l'espèce-hôte à cibler en priorité, tels que l'identification de l'espèce-hôte à l'origine de l'épidémie et la reconstruction exacte de la contribution à la transmission de chaque espèce-hôte. L'identification de l'espèce-hôte du cas index était sensible aux biais d'échantillonnage et à l'espèce-hôte considérée (bovin ou blaireau). La proportion de cas index pour lesquels l'espèce-hôte a été correctement reconstruite était élevée pour *outbreaker2* et seqTrack (plus de 75 %). Cependant, le fait que *TransPhylo* puisse désigner des hôtes non échantillonnés comme des cas index, combiné à sa tendance à surestimer la taille de l'épidémie, et donc le nombre d'hôtes non échantillonnés, pourrait expliquer sa moins bonne performance. De plus, les biais d'échantillonnage

conduisaient généralement à une meilleure performance, ce qui pourrait être expliqué par le fait que seuls les cas non-index (faune sauvage) étaient concernés par ces schémas d'échantillonnage. Enfin, cet indicateur était sensible à l'espèce-hôte du cas index et une moins bonne performance était observée quand le cas index était un blaireau.

4.4.2.3 Contribution des espèces-hôtes

L'estimation de la contribution des espèces-hôtes était influencée par les biais d'échantillonnage et la complexité du système épidémiologique mais pas par le taux de mutation. Avec les deux taux de mutation, la reconstruction par *outbreaker2* et *TransPhylo* de la contribution des espèces-hôtes était mauvaise et l'espèce-hôte contribuant le plus (bovins) était sous-estimée alors que celles qui contribuaient le moins (faune sauvage) étaient surestimées. Les deux méthodes de reconstruction ont été développées et testées sur des systèmes mono-hôtes (121,141), et pas sur des systèmes multi-hôtes dans lesquels chaque espèce-hôte joue un rôle différent. *TransPhylo* a déjà été appliqué à des systèmes multi-hôtes, un système humains-cerfs de SARS-CoV-2 (159), et deux systèmes bovins-blaireaux de bTB (175,176). Cependant, l'estimation de la contribution des espèces-hôtes dans ces systèmes n'était pas forcément évidente. Le nombre élevé de cas non échantillonnés estimé dans le système humains-cerfs (probabilité d'échantillonnage moyenne de 0,1 %) compliquait l'inférence des évènements de transmission. Les données phylogénétiques semblaient en accord avec des évènements multiples de « spillover » de l'humain au cerf, mais la transmission du cerf à l'humain n'a pas pour autant pu être écartée (159). Dans un système bovins-blaireaux dans le Sud-Ouest de l'Angleterre, van Tonder *et al.* cherchaient à estimer la transmission inter-espèces et ont donc implémenté *TransPhylo* en plus d'une méthode de reconstruction des états ancestraux (BASTA (38)), qui a été principalement utilisée pour estimer le nombre de transitions intra- et inter-espèces (175). Enfin, Akhmetova *et al.* ont également implémenté *TransPhylo* en plus de méthodes Bayésiennes phylogénétiques dans un système bovins-blaireaux en Irlande du Nord. Ils ont ainsi mis en évidence le rôle prédominant des bovins (concernant plus de 90 % des évènements de transmission reconstruits avec une probabilité postérieure élevée) dans la région (176).

Les biais simulés avec les schémas d'échantillonnage entraînaient une diminution du nombre d'hôtes infectés dont la contribution était surestimée. De fait, quand les schémas d'échantillonnages avaient un effet significatif sur l'estimation de la contribution des espèces-hôtes, ils avaient tendance à améliorer les résultats obtenus dans ce système multi-hôtes. Ainsi, ils

diminuaient (pour la faune sauvage) ou augmentaient (pour les bovins) le nombre estimé d'évènements de transmission. De plus, aucune des deux méthodes ne pouvait reconstruire correctement des rôles asymétriques entre espèces-hôtes, c'est-à-dire la présence d'une espèce-hôte cul-de-sac épidémiologique.

Dans le système épidémiologique multi-hôtes que nous avons étendu, le nombre de reproduction effectif variait selon la combinaison d'espèces-hôtes considérée (160). De plus, Bouchez-Zacria *et al.* ont calculé des distributions du temps de génération inter- et intra-espèces, qui montraient une diffusion plus rapide de l'agent pathogène à partir d'élevages de bovins qu'à partir de groupes sociaux de blaireaux. Nous avons ajouté à ce modèle de transmission, une troisième population d'espèce-hôte (les sangliers) qui pouvait transmettre ou non l'agent pathogène. La complexité de ce système multi-hôtes aurait pu contribuer aux mauvais résultats obtenus dans l'estimation de la contribution des espèces-hôtes. En effet, *outbreaker2* et *TransPhylo* considéraient toutes deux une seule distribution du temps de génération et/ou du nombre de cas secondaires pour les trois espèces-hôtes, ne prenant ainsi pas en compte la variation inter-espèces de l'histoire naturelle de la maladie ni les dynamiques de transmission inégales. Un système multi-hôtes dans lequel trois espèces-hôtes contribuent de manière inégale à la transmission n'est pas rare. Les résultats obtenus par reconstruction d'états ancestraux avec une méthode Bayésienne (Mascot, (96)) dans deux autres régions de France semblent indiquer la présence de systèmes multi-hôtes de bTB aussi complexes qui diffèrent selon la région d'étude (177). De plus, l'impact d'une telle complexité sur la performance de la méthode ne concerne pas uniquement les systèmes avec plusieurs espèces-hôtes, les pathogènes pour lesquels différentes catégories d'hôtes (*ex.* en fonction de l'âge et/ou du statut vaccinal (178)) peuvent être définies (par rapport à l'infectiosité et/ou la durée d'infection) constituent également des systèmes épidémiologiques complexes. Enfin, considérer ce qui peut être reconstruit par la méthode (épidémie reconstructible) au lieu de l'arbre de référence n'a pas permis d'améliorer les résultats obtenus pour la taille de l'épidémie ni pour la contribution des espèces-hôtes.

4.4.3 Limites de l'étude

4.4.3.1 Reconstruction des arbres de transmission

Nous avons été limité par des considérations pratiques et les choix qui en ont découlé. Avec les données de *M. bovis* que nous avons simulées, la convergence n'était pas un facteur limitant pour *outbreaker2* mais elle

l'était pour *TransPhylo*. En effet, afin de limiter le temps de calcul, nous avons fixé un nombre maximum d'itérations. Ceci a limité le nombre d'arbres que nous pouvions comparer à ceux qui convergeaient en moins de 48 heures dans BEAST2 et moins de 12 heures dans *TransPhylo*. De plus, afin de mieux comparer les reconstructions, nous avons utilisé le même modèle d'évolution pour tous les ensembles de cas lors de la reconstruction phylogénétique dans BEAST2. Un modèle d'évolution plus adapté à chaque épidémie simulée pourrait conduire à une reconstruction plus exacte de l'arbre phylogénétique, et donc à une meilleure performance de *TransPhylo*.

4.4.3.2 Simulation des séquences

Avec le modèle de simulation de séquences que nous avons implémenté, nous avons simulé une proportion faible de séquences uniques dans ces épidémies de 13 ans, ce qui est cohérent avec le taux de mutation faible de *M. bovis*. Dans l'étude des 167 séquences de *M. bovis* des PA et Landes à partir de laquelle nous avons sélectionné la valeur du taux de mutation (163), la proportion de séquences uniques isolées (37,1 %) était environ six fois plus élevée que la proportion médiane que nous avons simulée. La proportion plus élevée de séquences uniques dans cette étude pourrait être due au fait que toutes les séquences ne sont pas échantillonnées dans des données réelles et que l'épidémie a duré plus longtemps (MRCA datant de 27 ans plus tôt). Dans le modèle de simulation de séquences, nous avons également considéré le même taux de mutation au sein de chaque espèce-hôte. Cependant, le fait que *M. bovis* évolue ou non de la même façon au sein d'espèces-hôtes différentes reste inconnu. En effet, les taux de mutation de *M. tuberculosis* peuvent diminuer chez les humains lors de périodes de latence, ce qui diffère de ce qui a pu être observé chez des primates non-humains (179). Un phénomène similaire pourrait conduire à une variabilité de l'évolution de *M. bovis* au sein d'une espèce-hôte mais aussi entre espèces-hôtes (81), et donc à une différence entre la proportion de séquences uniques observée et celle qui a été simulée.

4.4.3.3 Données épidémiologiques considérées

Nous avons choisi de comparer les résultats obtenus à partir des mêmes données épidémiologiques et génétiques pour les trois méthodes. Dans *outbreaker2*, bien que cela soit difficile à implémenter lors de l'étude d'un système multi-hôtes impliquant la faune sauvage, les données de contact peuvent être directement incorporées dans l'inférence de l'arbre de transmission. L'ajout de données de contact a permis l'obtention de meilleures valeurs d'exactitude dans l'étude d'épidémies simulées d'Ebola et

de SARS-CoV-1 (121). De façon similaire, quand une faible diversité génétique était attendue dans leur étude sur *M. avium ssp paratuberculosis*, Nigsch *et al*. ont exploité le fait que seqTrack puisse prendre en compte des données additionnelles sous forme d'une matrice de poids (174). Ils ont ainsi résolu des liens de transmission équiprobables en utilisant les temps d'exposition et des caractéristiques influençant la susceptibilité. Même quand la méthode ne permettait pas d'ajouter d'autres types de données épidémiologiques, Xu *et al.* ont souligné le fait qu'une des forces de leur étude sur la transmission de *M. tuberculosis* dans une cohorte espagnole résidait dans la validation des résultats génomiques et de *TransPhylo* par des données de contact extensives (142). L'utilisation de ces options ou stratégies aurait pu améliorer les résultats obtenus par *outbreaker2* et seqTrack ainsi qu'évaluer ceux obtenus par *TransPhylo*. Cependant, peu de données de contact peuvent être collectées pour la faune sauvage, nous avons donc choisi de ne pas inclure d'autres types de données épidémiologiques dans cette étude. De plus, nous avons limité notre étude à trois méthodes disponibles dans un package et ne nécessitant que des dates d'échantillonnage. D'autres méthodes seraient intéressantes à tester avec ces données simulées, surtout des méthodes inférant simultanément les arbres phylogénétiques et de transmission (SimPF), comme *phybreak* (114), puisqu'aucune n'a été considérée ici. Enfin, les données que nous avons simulées dans un système multi-hôtes pourraient être utilisées pour tester les performances des méthodes Bayésiennes de reconstruction des états ancestraux, telles que Mascot (96) qui a déjà été utilisée pour étudier des systèmes multi-hôtes complexes de bTB (177).

4.4.4 Conclusion

Les mauvaises performances que nous avons obtenues pour l'exactitude et la contribution des espèces-hôtes, avec ou sans biais, suggèrent la nécessité de combiner des méthodes épidémiologiques et phylogénétiques à des méthodes de reconstruction d'arbres de transmission lors de l'étude d'un pathogène multi-hôtes évoluant lentement. Elles soulignent également le besoin de développer de nouvelles méthodes de reconstruction adaptées à des systèmes épidémiologiques complexes, ainsi que d'évaluer ces méthodes sur des données simulées dans des systèmes multi-hôtes, et non sur des données spécifiques à chaque méthode.

5 DISCUSSION GENERALE

5.1 RAPPEL DES OBJECTIFS

Notre objectif principal était d'investiguer comment estimer la contribution à la transmission de chaque espèce dans un système multi-hôtes. Pour cela, nous avons sélectionné deux types d'approches combinant l'information apportée par des données épidémiologiques et génomiques : la reconstruction des états ancestraux dans un arbre phylogénétique et la reconstruction d'arbres de transmission. Nous avons pris l'exemple de la tuberculose bovine en France, un système multi-hôtes ayant des enjeux économiques et impliquant la faune sauvage. L'implication de la faune sauvage conduit à la présence de disparités dans la surveillance entre les différentes espèces du système multi-hôtes et donc la nécessité de prendre en compte des biais d'observation.

Nous avons commencé par la reconstruction des états ancestraux d'arbres phylogénétiques dans un système bovins-blaireaux d'intérêt en France. Cette méthode a effectivement déjà été utilisée dans des systèmes multi-hôtes avec pour objectif d'estimer le patron de transmission inter-espèces (18,88,93,175–177). Cependant, cette méthode ne permet d'inférer que les transitions entre des sous-populations de pathogènes, définies ici par leur espèce-hôte, et non les transmissions entre individus. Cette reconstruction partielle des dynamiques de transmission nous a conduit à investiguer l'utilisation de méthodes de reconstruction de « qui a infecté qui ? ». Pour cela, nous avons réalisé une revue systématique de la littérature dans le but d'identifier des méthodes de reconstruction d'arbres de transmission qui pourraient être adaptées aux données épidémiologiques et génomiques disponibles et faciles à implémenter. Ces méthodes n'étaient pas développées spécifiquement pour un pathogène évoluant lentement dans un système multi-hôtes partiellement observé. Nous avons donc simulé des données de transmission de *M. bovis* dans un système à trois espèces-hôtes puis avons testé trois méthodes populaires en simulant plusieurs types de biais d'observation.

5.2 RESUME DES PRINCIPAUX RESULTATS OBTENUS

5.2.1 Reconstruction de l'histoire évolutive de *M. bovis* dans un système bovins-blaireaux

Nous avons choisi d'étudier un système bovins-blaireaux dans les PA et Landes, qui sont deux départements où l'évolution défavorable de la bTB pourrait représenter une menace pour le statut officiellement indemne de tuberculose bovine en France. A partir de 167 séquences (171 SNP) et de dates d'échantillonnage nous avons pu reconstruire les espèces-hôtes ('Bovin' ou 'Blaireau') des nœuds ancestraux hypothétiques dans les arbres phylogénétiques. Le taux de transition de blaireau-à-bovin estimé par le modèle était nettement supérieur au taux de transition de bovin-à-blaireau. Cependant, dans les arbres phylogénétiques postérieurs ré-échantillonnés, les évènements de transitions inter-espèces restaient rares et la persistance des lignées au sein d'une espèce-hôte prédominait. Enfin, les pics du nombre de lignées chez les bovins au milieu des années 2000 et 2010 étaient précédés par des pics chez les bovins.

La dynamique de transmission dans ce système bovins-blaireaux semblerait donc dominée par des transmissions, ou du portage long, au sein des deux espèces-hôtes. L'impact de rares transmissions par les blaireaux aux bovins serait amplifié par des évènements de transmission bovin-bovin. Pour aller plus loin, la description des transmissions à l'échelle individuelle et au sein de chaque espèce nécessitait l'utilisation d'autres méthodes d'investigation.

5.2.2 Description de vingt-deux méthodes de reconstruction d'arbres de transmission

En l'absence d'outil permettant le choix d'une méthode de reconstruction d'arbres de transmission, nous avons procédé à une revue systématique de la littérature en ayant pour objectif de donner des critères méthodologiques et pratiques pour faciliter le travail des utilisateurs de ces méthodes. Nous avons regroupé 22 méthodes en trois familles selon l'utilisation de l'information génétique (NPF *vs.* PF) et la nécessité ou non de reconstruire auparavant un arbre phylogénétique (SeqPF *vs.* SimPF). L'existence d'un package et les données nécessaires afin d'implémenter chaque méthode répertoriée ont notamment permis de choisir les méthodes que nous avons testées par la suite.

5.2.3 Evaluation des performances de trois méthodes de reconstruction d'arbres de transmission dans un système *M. bovis* multi-hôtes

Nous avons choisi trois méthodes de reconstruction d'arbres de transmission : *outbreaker2*, seqTrack et *TransPhylo*. Nous avons ensuite étendu un modèle de transmission de *M. bovis* dans le système bovins-blaireaux des PA et Landes (160) en y ajoutant une population de sangliers qui pouvait transmettre (ou non) l'agent pathogène. Les mauvaises performances estimées pour la plupart des indicateurs épidémiologiques, en présence ou non de biais d'observation, pourraient être expliquées par le taux d'évolution lent de *M. bovis* et le peu d'information génétique présent dans les séquences simulées qui en découle. Ce taux d'évolution lent a notamment impacté la reconstruction de « qui a infecté qui ? » et la taille de l'épidémie estimée par *TransPhylo*. A l'inverse, l'estimation de la contribution de chaque espèce-hôte n'était pas influencée par ce taux d'évolution lent mais plutôt par la complexité du système épidémiologique simulé.

5.3 PARTICIPATION A LA COMPREHENSION DES SYSTEMES MULTI-HOTES DE TUBERCULOSE BOVINE EN FRANCE

5.3.1 Rappel bref des études précédentes

La reconstruction de l'histoire évolutive de *M. bovis* dans le système bovins-blaireaux des PA et Landes (premier axe de cette thèse) s'inscrit dans le cadre plus large de multiples travaux ayant pour but de mieux comprendre la dynamique de transmission de la tuberculose bovine dans cette région ainsi que d'autres régions de France.

Bouchez-Zacria *et al.* ont réalisé trois études dans ce système bovins-blaireaux. Ils ont commencé par étudier les facteurs environnementaux associés à la présence concomitante de *M. bovis* chez des blaireaux et des bovins dans les PA et Landes (89). Ils ont ainsi mis en évidence le fait que les facteurs favorisant la survie de *M. bovis* dans l'environnement et/ou la présence de blaireaux sur des pâtures étaient associés à la présence de l'agent pathogène chez les deux espèces-hôtes. Puis, une analyse du réseau de voisinage entre les pâtures utilisées par les élevages bovins et domaines vitaux des blaireaux a permis de souligner l'importance des mouvements de bovins, mais aussi le rôle intermédiaire des blaireaux dans la transmission de bTB entre élevages de la région (90). Enfin, Bouchez-Zacria *et al.* ont simulé la transmission de 11 génotypes de *M. bovis* dans ce système bovins-blaireaux de 2007 à 2019 (160), modèle que nous avons étendu dans ce travail de thèse. Parallèlement, Canini *et al.* ont reconstruit l'histoire évolutive

de *M. bovis* dans deux systèmes à trois espèces-hôtes (bovins, blaireaux et sangliers) en Côte-d'Or et Dordogne/Haute-Vienne (177).

5.3.2 Trois systèmes multi-hôtes en France

5.3.2.1 Importance de la persistance intra-espèce

La comparaison de nos résultats avec ces deux dernières études permet de relever des caractéristiques intéressantes dans les systèmes multi-hôtes de bTB en France. Premièrement, la taille effective de population était la plus élevée chez les blaireaux dans les trois systèmes multi-hôtes : PA et Landes, Côte-d'Or et Dordogne/Haute-Vienne. Cela signifie que la coalescence est plus lente entre deux lignées hébergées par des blaireaux qu'entre deux lignées hébergées par des bovins ou des sangliers (177). Dans le modèle de Bouchez-Zacria *et al.*, les groupes de blaireaux montraient une durée d'infection plus longue que les bovins (160), ce qui pourrait expliquer une diversité génétique plus élevée chez les blaireaux et donc les différences dans les tailles de population effectives estimées.

Ensuite, les évènements de transmission inter-espèces étaient rares dans les trois systèmes multi-hôtes. Ainsi, dans les arbres phylogénétiques postérieurs ré-échantillonnés, la persistance intra-espèces était le type de transition le plus fréquent. Cependant, l'espèce-hôte concernée dépendait du système multi-hôtes ainsi que de la période de temps considérée. Comme dans les PA et Landes, la persistance chez les bovins commençait tardivement (2009) en Dordogne/Haute-Vienne et était précédée par de la persistance dans la faune sauvage. A l'inverse, la persistance chez les bovins commençait dès 1990 dans la Côte-d'Or.

5.3.2.2 Un rôle variable en fonction du système et au cours du temps

Dans une épidémie en pleine expansion (Dordogne/Haute-Vienne) ou n'ayant pas de tendance particulière (PA et Landes), un sens de transmission prédominant inter-espèces similaire a été inféré par une méthode de reconstruction des état ancestraux (Mascot, (96)) : de la faune sauvage aux bovins. A l'inverse, dans une épidémie en déclin (Côte-d'Or), la transmission des bovins à la faune sauvage prédominait. Ces différences dans le sens de transmission prédominant pourraient être dues à des protocoles de biosécurité plus sévères vis-à-vis de la faune sauvage ou encore moins d'interfaces de contacts indirects entre les bovins et la faune sauvage (Côte-d'Or par rapport à Dordogne/Haute-Vienne) (177).

Dans le travail de Bouchez-Zacria *et al.*, au début de chaque

simulation, un bovin était infecté par génotype (sauf un génotype présent dans quatre élevages) alors que seul un blaireau était infecté (160). Dans leur modèle, les auteurs ont reconstruit une inversion au cours du temps du ratio entre le nombre d'évènements des transmissions de bovins à blaireaux et celui de blaireaux à bovin. Ils ont ainsi fait l'hypothèse que la transmission plus rapide à partir de bovins avait conduit à l'installation de l'infection dans les élevages de bovins parallèlement à des évènements de « spillover » des bovins aux blaireaux. Puis, après une longue période de circulation de l'agent pathogène au sein des groupes de blaireaux, le sens prédominant de la transmission inter-espèce s'est inversé, favorisant ainsi la transmission à partir des blaireaux. En plus du système multi-hôtes concerné, le sens de transmission prédominant pourrait donc également être influencé par la phase de l'épidémie dans laquelle le système multi-hôtes se trouve.

5.4 LIMITES DES DEUX APPROCHES SELECTIONNEES DANS L'ETUDE D'UN SYSTEME MULTI-HOTES

5.4.1 Problème de l'espèce-hôte fantôme

Dans le premier axe de cette thèse, nous avons étudié un système bovins-blaireaux dans les PA et les Landes, cependant les données de surveillance ont rapporté la présence de sangliers infectés dans cette région. Le peu de données disponibles chez les sangliers (6 séquences) aurait perturbé l'inférence des paramètres de population, nous avons donc décidé d'exclure cette sous-population dans cette partie de notre travail. Canini *et al.* ont également dû exclure des espèces-hôtes dans l'étude de la Côte d'Or et la Dordogne/Haute-Vienne, le nombre de séquences isolées chez les cerfs, chevreuils et renards étant respectivement de deux, un et cinq (177). Ici, nous avons donc intégré dans le modèle Bayésien uniquement des séquences isolées chez les bovins et les blaireaux. L'état ancestral qui pouvait être reconstruit était alors soit 'Bovin' soit 'Blaireau'. L'addition de données de sangliers dans les systèmes multi-hôtes de la Côte-d'Or et Dordogne/Haute-Vienne a permis de mettre en évidence son rôle d'hôte intermédiaire entre les blaireaux et les bovins (177). Le travail de reconstruction des dynamiques de transmission inter-espèces dans les PA et Landes a donc pu être fortement impacté par l'absence de prise en compte du rôle du sanglier.

Avec certaines méthodes de reconstruction des états ancestraux, il est possible de considérer un « dème fantôme », c'est-à-dire un dème pour lequel aucune donnée génétique n'est disponible. Avec la méthode BASTA, De Maio *et al.* ont ainsi inférer des évènements indépendants de transmissions zoonotiques à l'origine d'épidémies d'Ebola dans la population

humaine (38). Cependant, ils avaient considéré un modèle de transmission unidirectionnelle très simplifié entre deux dèmes. En l'absence de connaissances a priori sur les dynamiques de transmission inter-espèces, il nous était difficile de procéder de la même façon.

Ce problème de l'espèce-hôte fantôme n'est pas résolu pour autant par l'utilisation de méthodes de reconstruction d'arbres de transmission. Ces méthodes ne prenant pas en compte l'espèce-hôte, elles ne demandent pas explicitement une quantité suffisante de données par espèce-hôte. Dans le cas du système épidémiologique des PA et Landes, les six séquences isolées chez des sangliers pourraient être incluses lors de la reconstruction d'un arbre de transmission. Cependant, nous avons vu (troisième axe de cette thèse) que le manque d'information génétique et la complexité du système multi-hôtes pouvaient conduire à une estimation inexacte du rôle des espèces-hôtes. La possibilité d'inclure ces données dans l'inférence d'un arbre de transmission ne garantit en rien la qualité de la reconstruction de cet arbre. Le nombre de données disponibles impacte donc les deux approches que nous avons choisies pour étudier des systèmes multi-hôtes.

5.4.2 Les biais d'observation inhérents aux systèmes multi-hôtes

Inclure toutes les espèces-hôtes contribuant au système multi-hôtes est donc nécessaire pour reconstruire les dynamiques de transmission inter- et intra-espèces. Cependant, dans le système multi-hôtes de la bTB en France impliquant la faune sauvage, nous sommes confrontés à un problème de différences dans le niveau de surveillance entre espèces-hôtes, qui correspond à un biais d'observation. Pour les méthodes de reconstruction d'arbres de transmission, nous avons testé dans le troisième axe comment la performance de trois méthodes était impactée par quelques biais d'observation (temporel et espèce-dépendant) spécifiques d'un système multi-hôtes de la bTB en France. Dans ce système multi-hôtes, les biais tendaient à corriger les erreurs d'estimation inhérentes aux méthodes inadaptées à un système multi-hôtes. Il est difficile de généraliser ces résultats à d'autres systèmes multi-hôtes du fait de la diversité des dynamiques de transmission possibles. En effet, deux systèmes multi-hôtes de la bTB en France impliquant les mêmes espèces-hôtes (Dordogne/Haute-Vienne et Côte-d'Or) montraient deux dynamiques de transmission bien distinctes (177). De plus, nous n'avons testé que trois méthodes, les autres pourraient avoir un comportement différent vis-à-vis des biais d'observation.

Les méthodes de coalescence structurée ne prennent pas en compte le nombre d'échantillons par espèce-hôte dans leur inférence et ne séparent

pas la reconstruction de l'arbre phylogénétique avec celle des états ancestraux. Pour ces raisons, il est admis que leur utilisation par rapport à une méthode de « mugration » diminue l'impact de biais d'observation (38). Cependant, l'évaluation de différentes méthodes de reconstruction de coalescence structurée s'est effectuée sur des pathogènes évoluant rapidement (*ex.* virus de la rage (180), influenza aviaire (38)) au sein d'un système à une seule espèce-hôte (*ex.* population de chiens ou humaine). De plus, quand des schémas d'échantillonnage étaient appliqués sur des données simulées, ils ne reflétaient pas forcément les biais d'observation spécifiques des systèmes multi-hôtes de bTB en France (*ex.* prise en compte d'un biais entre « localisations » mais pas de biais temporel (180)). Nous ne savons donc pas comment ces biais auraient pu impacter la reconstruction des états ancestraux dans le premier axe de cette thèse.

De plus, le choix de cette méthode de coalescence structurée impliquait l'hypothèse d'une taille de population et de taux de transition constants au cours du temps. Il est admis que des variations significatives dans la taille effective de population et les taux de transitions inter-espèces conduisent à une mauvaise reconstruction des dynamiques de population avec ce type de méthodes (96,176). Lorsque cette dynamique de population est inadaptée, la performance des méthodes de coalescence structurée peut même s'avérer inférieure à celle d'une méthode « mugration », en présence ou non de biais (180). Les données dont nous disposions dans le premier axe de cette thèse ne semblaient pas contredire de manière flagrante cette hypothèse, mais elles étaient incomplètes du fait des biais d'observation. Nous ne savons donc pas comment la reconstruction des états ancestraux (premier axe de cette thèse) a pu être impactée par cette hypothèse d'une taille de population constante.

5.5 PISTES D'AMELIORATION DANS L'ETUDE D'UN SYSTEME MULTI-HOTES

5.5.1 Choix des données épidémiologiques

Les données épidémiologiques disponibles dépendent du système multi-hôtes étudié. Dans le premier axe de cette thèse, nous avions à notre disposition des séquences et des dates d'échantillonnage lors de la reconstruction de l'histoire évolutive de *M. bovis* dans les PA et Landes. Pour cette raison, nous avons choisi de considérer ce même type de données lors de la reconstruction d'arbres de transmission dans des épidémies simulées dans la région (troisième axe de cette thèse). Considérer des données épidémiologiques supplémentaires telles que des données de contact,

comme dans le travail de Campbell *et al.* (121), permettrait sans doute d'améliorer la reconstruction de ces arbres de transmission.

Dans le travail de Bouchez Zacria *et al.*, les contacts entre élevages de bovins provenaient des données de mouvements de la BDNI sur la période d'étude ainsi que des données de pâtures du « Registre Parcellaire Graphique » de 2013 (160). Concernant les groupes de blaireaux, les auteurs avaient reconstruit des domaines vitaux, et défini sur cette base une relation de voisinage spatial entre groupes. Enfin, la possibilité de transmission indirecte de *M. bovis* entre les deux métapopulations (bovins et blaireaux) était déterminée par l'intersection de leurs zones d'usage (domaines vitaux pour les groupes de blaireaux et pâtures pour les élevages de bovins). En plus de cette intersection des espaces vitaux, Marsot *et al.* ont considéré l'utilisation de ces espaces et la densité des espèces afin de calculer un indice spatial de contacts indirects entre les bovins et les blaireaux (BACACIX) en France (181). Les sangliers n'étaient pas pris en compte dans ces deux modèles. Cependant, un raisonnement similaire basé sur des données de mouvements de sangliers (112) pourrait être appliqué, et la possibilité de contact pourrait être déterminée par la proximité entre pâtures ou domaines vitaux et les communes dans lesquelles ils étaient simulés. Nous pourrions ainsi considérer ces données de contact/indices spatiaux dans la reconstruction d'arbres de transmission.

5.5.2 Choix de la méthode

Considérer des données épidémiologiques additionnelles permet de prendre en compte ou de croiser les informations apportées par différentes sources. Les performances que nous avons estimées pour trois méthodes de reconstruction d'arbres de transmission (troisième axe de cette thèse) soulignent également l'importance de croiser les résultats de différentes approches. Nous allons donc ici considérer le choix de deux types de méthodes : reconstruction phylogénétique et reconstruction d'arbres de transmission.

5.5.2.1 Reconstruction phylogénétique

Dans la reconstruction des états ancestraux dans le système bovins-blaireaux dans les PA et Landes (premier axe de cette thèse), nous avons choisi d'utiliser la méthode de coalescence structurée Mascot. Cette méthode a notamment été utilisée dans l'étude d'autres systèmes multi-hôtes, de bTB (177) ou encore de MERS-CoV (18). Nous avons déjà évoqué les limites de la méthode de « mugration » qui nous avaient amenés à choisir une méthode

de coalescence structurée. Cependant, d'autres méthodes de reconstruction des états ancestraux différentes de la méthode de « mugration » existent. Ainsi, nous aurions pu choisir une autre méthode de coalescence structurée comme BASTA (38), déjà utilisée pour étudier des systèmes multi-hôtes de bTB (88,175). Enfin, Akhmetova *et al.* ont commencé par une analyse préliminaire phylogénétique de leurs données de *M. bovis*. Cette analyse les a conduit à conclure qu'une méthode de coalescence structurée était inappropriée, du fait de l'inférence d'une augmentation de la taille de population effective de l'agent pathogène, coïncidant avec une augmentation de l'incidence de bTB dans la région (pré-print, (182)). Afin de prendre en compte l'évolution au cours de temps des paramètres de population, Müller *et al.* ont introduit une nouvelle approche combinant un GLM (permettant de prédire des variations dans les tailles efficaces des sous-populations ou des taux de migrations) avec leur méthode (Mascot) (183). Cependant, cette approche nécessite des prédicteurs et donc des données épidémiologiques supplémentaires (tels que des données d'incidence) non biaisées, car la présence de biais dans ces données semblerait biaiser l'inférence des dynamiques de population (180). Akhmetova *et al.* ont quant à eux privilégié une méthode structurée de « Birth-Death » disponible dans le package BDMM de BEAST2. Cette méthode est basée sur le modèle « Birth-Death » skyline et permet donc de prendre en compte des variations de taille effective de population au cours du temps (176).

5.5.2.2 Reconstruction d'arbres de transmission

Dans notre étude sur les performances des méthodes de reconstruction d'arbres de transmission (troisième axe de cette thèse), nous nous sommes limités à trois méthodes adaptées aux données disponibles de bTB en France. Une quatrième méthode répondant à ces critères, appelée *phybreak* (SimPF), mise de côté par manque de temps, serait intéressante à évaluer dans un tel contexte. Dans d'autres systèmes multi-hôtes, les données disponibles peuvent varier et les méthodes que nous avons testées ne sont alors pas forcément les plus adaptées. La revue systématique de la littérature que nous avons effectuée (deuxième axe de cette thèse) permet de choisir une méthode adaptée à des données génomiques basée sur des séquences consensus. Cependant, lorsque des données de « deep sequencing » sont disponibles, les méthodes que nous avons simplement évoquées (SLAFEEL et BadTriP) permettraient d'exploiter la totalité de l'information génomique. Enfin, Akhmetova *et al.* ont fait le choix d'exploiter les résultats de *TransPhylo* obtenus à partir d'arbres phylogénétiques consensus reconstruits avec plusieurs modèles d'horloge (176). Nous pourrions ainsi utiliser différents modèles d'évolution pour reconstruire les

arbres phylogénétiques avant l'application de méthodes SeqPF sur l'arbre consensus, ou encore sur un sous-échantillon d'arbres postérieurs comme cela a été rendu possible avec *TransPhylo* par Xu *et al.* (149).

5.6 PERSPECTIVES

Pour aller plus loin dans l'étude de systèmes multi-hôtes de bTB en France, la prochaine étape serait d'utiliser des données de terrain différentes (autres régions de France) ou actualisées (depuis 2014 pour la Côte-d'Or et depuis 2017 pour les PA et Landes et la Dordogne/Haute-Vienne) de *M. bovis*.

Premièrement, les données actualisées pourraient être utilisées pour reconstruire des arbres de transmission. Puis, ces arbres de transmission pourraient être comparés avec les résultats obtenus lors de la reconstruction des états ancestraux déjà effectuée dans les PA et Landes, Côte-d'Or et Dordogne/Haute-Vienne. Ensuite de nouveaux systèmes multi-hôtes pourraient être étudiés avec les deux types d'approches. Les cerfs élaphes, sangliers et bovins de la forêt de Brotonne ont constitué le premier système multi-hôtes de bTB impliquant la faune sauvage découvert en France. Contrairement aux autres systèmes multi-hôtes en France, c'est un système épidémiologique clos, bordé par l'autoroute et la Seine qui empêchent les mouvements d'animaux sauvages au-delà de cette zone. Les résultats de travaux de modélisation réalisés par Zanella *et al.* dans le système cerfs-sangliers sont en faveur d'une dynamique de transmission très différente des trois autres systèmes multi-hôtes de bTB en France. En effet, le cerf élaphe semblerait plus contagieux et le sanglier plus sensible (162). Enfin, les données que nous avons simulées dans le système de bTB à trois espèces-hôtes dans les PA et Landes pourraient être utilisées afin d'évaluer d'autres méthodes de reconstruction d'arbres de transmission. Ces données simulées de bTB pourraient également être utilisées pour évaluer la performance des méthodes de reconstruction des états ancestraux en présence ou non de biais d'observation.

5.7 CONCLUSION GENERALE

Nous avons donc testé deux approches permettant l'estimation de la contribution de chaque espèce dans un système multi-hôtes, l'une sur des données réelles, l'autre sur des données simulées. La reconstruction d'états ancestraux et celle d'arbres de transmission sont deux approches complémentaires. L'une permet d'estimer des paramètres de population et l'autre des paramètres épidémiologiques généraux ou spécifiques d'un système multi-hôtes. Cependant, leur application à des systèmes multi-hôtes

complexes soumis à des biais d'observation n'est pas toujours évidente. Il reste à explorer l'impact de la complexité du système épidémiologique et des biais d'observation sur des méthodes de reconstruction d'états ancestraux et développer des méthodes de reconstruction d'arbres de transmission plus adaptées aux systèmes multi-hôtes.

6 BIBLIOGRAPHIE

1. Cleaveland S, Laurenson MK, Taylor LH. Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. Philos Trans R Soc Lond B Biol Sci. 29 juill 2001;356(1411):991-9.

2. Taylor LH, Latham SM, Woolhouse ME. Risk factors for human disease emergence. Philos Trans R Soc Lond B Biol Sci. 29 juill 2001;356(1411):983-9.

3. Cross AR, Baldwin VM, Roy S, Essex-Lopresti AE, Prior JL, Harmer NJ. Zoonoses under our noses. Microbes Infect. 2019;21(1):10-9.

4. Viana M, Cleaveland S, Matthiopoulos J, Halliday J, Packer C, Craft ME, et al. Dynamics of a morbillivirus at the domestic–wildlife interface: Canine distemper virus in domestic dogs and lions. Proc Natl Acad Sci U S A. 3 févr 2015;112(5):1464-9.

5. Gortázar C, Ferroglio E, Höfle U, Frölich K, Vicente J. Diseases shared between wildlife and livestock: a European perspective. Eur J Wildl Res. 1 nov 2007;53(4):241-56.

6. Haydon DT, Cleaveland S, Taylor LH, Laurenson MK. Identifying reservoirs of infection: a conceptual and practical challenge. Emerg Infect Dis. déc 2002;8(12):1468-73.

7. Tuells J, Henao-Martínez AF, Franco-Paredes C. Yellow Fever: A Perennial Threat. Arch Med Res. nov 2022;53(7):649-57.

8. Frierson JG. The Yellow Fever Vaccine: A History. Yale J Biol Med. juin 2010;83(2):77-85.

9. Rhodes CJ, Atkinson RP, Anderson RM, Macdonald DW. Rabies in Zimbabwe: reservoir dogs and the implications for disease control. Philos Trans R Soc Lond B Biol Sci. 29 juin 1998;353(1371):999-1010.

10. Fenton A, Pedersen AB. Community Epidemiology Framework for

Classifying Disease Threats. Emerg Infect Dis. déc 2005;11(12):1815-21.

11. Portier J, Ryser-Degiorgis MP, Hutchings MR, Monchâtre-Leroy E, Richomme C, Larrat S, et al. Multi-host disease management: the why and the how to include wildlife. BMC Vet Res. 14 août 2019;15:295.

12. Viana M, Mancy R, Biek R, Cleaveland S, Cross PC, Lloyd-Smith JO, et al. Assembling evidence for identifying reservoirs of infection. Trends Ecol Evol. mai 2014;29(5):270-9.

13. Delavenne C, Desvaux S, Boschiroli ML, Carles S, Durand B, Forfait C, et al. Surveillance de la tuberculose due à Mycobacterium bovis en France métropolitaine en 2019 : Résultats et indicateurs de fonctionnement. Bulletin épidémiologique, santé animale et alimentation. 2021;94(13):1-23.

14. Pham TT, Meng S, Sun Y, Lv W, Bahl J. Inference of Japanese encephalitis virus ecological and evolutionary dynamics from passive and active virus surveillance. Virus Evol. janv 2016;2(1):vew009.

15. Gervasi V, Marcon A, Bellini S, Guberti V. Evaluation of the Efficiency of Active and Passive Surveillance in the Detection of African Swine Fever in Wild Boar. Vet Sci. 30 déc 2019;7(1):5.

16. Kao RR, Haydon DT, Lycett SJ, Murcia PR. Supersize me: how whole-genome sequencing and big data are transforming epidemiology. Trends in microbiology. 2014/03/26 éd. mai 2014;22(5):282-91.

17. Alizon S. Phylogénies d'infections et phylodynamique. In: Etude de l'approche mathématique et informatique [Internet]. Gilles Didier and Stéphane Guindon. London, UK: ISTE; 2022. Disponible sur: https://hal.archives-ouvertes.fr/hal-02884408/file/hal.pdf

18. Dudas G, Carvalho LM, Rambaut A, Bedford T. MERS-CoV spillover at the camel-human interface. Ferguson NM, éditeur. eLife. 16 janv 2018;7:e31257.

19. Hamming R. Error-detecting and error-correcting codes. The Bell System Technical Journal. 1950;29(2):147-60.

20. Tavaré S. Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on Mathematics in the Life Sciences. 1 janv 1986;17:57-86.

21. Hasegawa M, Kishino H, Yano T aki. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution. oct 1985;22(2):160-74.

22. Freese E. The difference between spontaneous and base-analogue induced mutation of phage T4. Proc Natl Acad Sci U S A. avr 1959;45(4):622-33.

23. Jukes TH, Cantor CR. Evolution of protein molecules. In: Munro HN (ed) Mammalian Protein Metabolism. New York: Academic Press; 1969. p. 21-132.

24. Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. In: Bryson, V and Vogel, HJ, Eds, Evolving Genes and Proteins. New York: Academic Press; 1965. p. 97-166.

25. dos Reis M, Donoghue PCJ, Yang Z. Bayesian molecular clock dating of species divergences in the genomics era. Nat Rev Genet. févr 2016;17(2):71-80.

26. Li WH. So, what about the molecular clock hypothesis? Curr Opin Genet Dev. déc 1993;3(6):896-901.

27. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed Phylogenetics and Dating with Confidence. PLoS Biol. mai 2006;4(5).

28. Kingman JFC. The coalescent. Stochastic Processes and their Applications. 1 sept 1982;13(3):235-48.

29. Kuhner MK, Yamato J, Felsenstein J. Maximum likelihood estimation of population growth rates based on the coalescent. Genetics. mai 1998;149(1):429-34.

30. Pybus OG, Rambaut A, Harvey PH. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics. juill 2000;155(3):1429-37.

31. Nordborg M. Coalescent theory. In: Handbook of statistical genetics. John Wiley & Sons; 2001.

32. Nee S, May RM, Harvey PH. The reconstructed evolutionary process. Philos Trans R Soc Lond B Biol Sci. 28 mai 1994;344(1309):305-11.

33. Novozhilov AS, Karev GP, Koonin EV. Biological applications of the theory of birth-and-death processes. Brief Bioinform. mars 2006;7(1):70-85.

34. He W, Auclert LZ, Zhai X, Wong G, Zhang C, Zhu H, et al. Interspecies Transmission, Genetic Diversity, and Evolutionary Dynamics of Pseudorabies Virus. J Infect Dis. 5 mai 2019;219(11):1705-15.

35. Swofford D, Maddison W. Parsimony, character-state reconstructions and evolutionary inferences. In: Systematics, Historical Ecology, and North American Freshwater Fishes. Stanford University Press; 1992. p. 186-223.

36. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian Phylogeography Finds Its Roots. PLoS Comput Biol. 25 sept 2009;5(9).

37. Müller NF, Rasmussen DA, Stadler T. The Structured Coalescent and Its Approximations. Mol Biol Evol. nov 2017;34(11):2970-81.

38. De Maio N, Wu CH, O'Reilly KM, Wilson D. New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. PLoS Genetics. 12 août 2015;11(8).

39. Firestone SM, Hayama Y, Lau MSY, Yamamoto T, Nishi T, Bradhurst RA, et al. Transmission network reconstruction for foot-and-mouth disease outbreaks incorporating farm-level covariates. PloS one. 2020/07/16 éd. 2020;15(7):e0235660.

40. Lau MSY, Marion G, Streftaris G, Gibson G. A systematic Bayesian integration of epidemiological and genetic data. PLoS Computational Biology. 2015;11(11).

41. Ypma RJ, van Ballegooijen WM, Wallinga J. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. Genetics. 2013/09/17 éd. nov 2013;195(3):1055-62.

42. Redd AD, Quinn TC, Tobian AAR. Frequency and Implications of HIV Superinfection. Lancet Infect Dis. juill 2013;13(7):622-8.

43. Blackard JT. HCV superinfection and reinfection. Antivir Ther. 2012;17(7 Pt B):1443-8.

44. Ypma RJ, Jonges M, Bataille A, Stegeman A, Koch G, van Boven M, et al. Genetic data provide evidence for wind-mediated transmission of highly pathogenic avian influenza. The Journal of infectious diseases. 2012/12/12 éd. 1 mars 2013;207(5):730-5.

45. Amundsen EJ, Stigum H, Røttingen JA, Aalen OO. Definition and estimation of an actual reproduction number describing past infectious disease transmission: application to HIV epidemics among homosexual men in Denmark, Norway and Sweden. Epidemiol Infect. déc 2004;132(6):1139-49.

46. Faye O, Boëlle PY, Heleze E, Faye O, Loucoubar C, Magassouba N, et al. Chains of transmission and control of Ebola Virus Disease in Conakry, Guinea in 2014. The Lancet Infectious diseases. mars 2015;15(3):320-6.

47. Podsiadło Ł, Polz-Dacewicz M. Molecular evolution and phylogenetic implications in clinical research. Annals of Agricultural and Environmental Medicine. 20 sept 2013;20(3):455-9.

48. Hillis DM, Huelsenbeck JP. Support for dental HIV transmission. Nature. 5 mai 1994;369(6475):24-5.

49. Blanchard A, Ferris S, Chamaret S, Guétard D, Montagnier L. Molecular evidence for nosocomial transmission of human

immunodeficiency virus from a surgeon to one of his patients. J Virol. mai 1998;72(5):4537-40.

50. Hassan AS, Pybus OG, Sanders EJ, Albert J, Esbjörnsson J. Defining HIV-1 transmission clusters based on sequence data. AIDS. 1 juin 2017;31(9):1211-22.

51. Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SDW. Phylodynamics of infectious disease epidemics. Genetics. déc 2009;183(4):1421-30.

52. Varia M, Wilson S, Sarwal S, McGeer A, Gournis E, Galanis E, et al. Investigation of a nosocomial outbreak of severe acute respiratory syndrome (SARS) in Toronto, Canada. CMAJ: Canadian Medical Association Journal. 19 août 2003;169(4):285-92.

53. Garry M, Hope L, Zajac R, Verrall AJ, Robertson JM. Contact tracing: a memory task with consequences for public health. Perspectives on Psychological Science. janv 2021;16(1):175-87.

54. Crozet G, Dufour B, Rivière J. Investigation of field intradermal tuberculosis test practices performed by veterinarians in France and factors that influence testing. Research in Veterinary Science. 1 juin 2019;124:406-16.

55. Campbell F, Strang C, Ferguson N, Cori A, Jombart T. When are pathogen genome sequences informative of transmission events? PLoS Pathog. févr 2018;14(2):e1006885.

56. Didelot X, Gardy J, Colijn C. Bayesian inference of infectious disease transmission from whole-genome sequence data. Molecular biology and evolution. 2014/04/10 éd. juill 2014;31(7):1869-79.

57. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis. févr 2013;13(2):137-46.

58. Worby CJ, Lipsitch M, Hanage WP. Within-host bacterial diversity

hinders accurate reconstruction of transmission networks from genomic distance data. PLoS computational biology. 2014/03/29 éd. mars 2014;10(3):e1003549.

59. Michelet L, Durand B, Boschiroli ML. Tuberculose bovine: Bilan génotypique de M. bovis à l'origine des foyers bovins entre 2015 et 2017 en France Métropolitaine. Bulletin épidémiologique. 2019;9.

60. O'Reilly LM, Daborn CJ. The epidemiology of Mycobacterium bovis infections in animals and man: a review. Tuber Lung Dis. août 1995;76 Suppl 1:1-46.

61. Kean JM, Barlow ND, Hickling GJ. Evaluating potential sources of bovine tuberculosis infection in a New Zealand cattle herd. New Zealand Journal of Agricultural Research. janv 1999;42(1):101-6.

62. O'Brien DJ, Schmitt SM, Fierke JS, Hogle SA, Winterstein SR, Cooley TM, et al. Epidemiology of Mycobacterium bovis in free-ranging white-tailed deer, Michigan, USA, 1995–2000. Preventive Veterinary Medicine. 30 mai 2002;54(1):47-63.

63. Simpson VR. Wild Animals as Reservoirs of Infectious Diseases in the UK. The Veterinary Journal. 1 mars 2002;163(2):128-46.

64. Naranjo V, Gortazar C, Vicente J, de la Fuente J. Evidence of the role of European wild boar as a reservoir of Mycobacterium tuberculosis complex. Veterinary Microbiology. févr 2008;127(1-2):1-9.

65. Zanella G, Durand B, Hars J, Moutou F, Garin-Bastuji B, Duvauchelle A, et al. Mycobacterium bovis in wildlife in France. J Wildl Dis. janv 2008;44(1):99-108.

66. Réveillaud É, Desvaux S, Boschiroli ML, Hars J, Faure É, Fediaevsky A, et al. Infection of Wildlife by Mycobacterium bovis in France Assessment Through a National Surveillance System, Sylvatub. Front Vet Sci. 2018;5:262.

67. Michelet L, De Cruz K, Hénault S, Tambosco J, Richomme C,

Réveillaud É, et al. Mycobacterium bovis Infection of Red Fox, France. Emerg Infect Dis. juin 2018;24(6):1151-3.

68. Richomme C, Réveillaud E, Moyen JL, Sabatier P, De Cruz K, Michelet L, et al. Mycobacterium bovis Infection in Red Foxes in Four Animal Tuberculosis Endemic Areas in France. Microorganisms. 17 juill 2020;8(7):1070.

69. Keck N, Moyen JL, Gueneau E, Boschiroli ML. Particular features of screening and diagnosis of bovine tuberculosis. Epidémiologie et santé animale. 1 janv 2014;65:5-19.

70. Rivière J, Le Strat Y, Dufour B, Hendrikx P. Sensitivity of Bovine Tuberculosis Surveillance in Wildlife in France: A Scenario Tree Approach. PLoS One. 30 oct 2015;10(10).

71. Bekara MEA, Courcoul A, Bénet JJ, Durand B. Modeling Tuberculosis Dynamics, Detection and Control in Cattle Herds. PLOS ONE. 25 sept 2014;9(9):e108584.

72. Bénet JJ, Boschiroli ML, Dufour B, Garin-Bastuji B. Lutte contre la tuberculose bovine en France de 1954 à 2004: analyse de la pertinence épidémiologique de l'évolution de la réglementation. Epidémiologie et Santé Animale. (50):127-43.

73. Robert J, Boulahbal F, Trystram D, Truffot-Pernot C, de Benoist AC, Vincent V, et al. A national survey of human Mycobacterium bovis infection in France. Network of Microbiology Laboratories in France. Int J Tuberc Lung Dis. août 1999;3(8):711-4.

74. Lepesqueux G, Mailles A, Aubry A, Veziris N, Jaffré J, Jarlier V, et al. Epidémiologie des cas de tuberculose à Mycobacterium bovis diagnostiqués en France. Médecine et Maladies Infectieuses. juin 2018;48(4):115-6.

75. Reviriego Gordejo FJ, Vermeersch JP. Towards eradication of bovine tuberculosis in the European Union. Veterinary Microbiology. 25 févr 2006;112(2):101-9.

76. Delavenne C, Pandolfi F, Girard S, Réveillaud É, Boschiroli ML,

Dommergues L, et al. Tuberculose bovine: Bilan et évolution de la situation épidémiologique entre 2015 et 2017 en France Métropolitaine. Bulletin épidémiologique. 2019;22.

77. Guétin-Poirier V, Rivière J, Dufour B. Cost-effectiveness of two different protocols for animal tracing investigations of bovine tuberculosis outbreaks in France. Prev Vet Med. févr 2020;175:104868.

78. Ladreyt H, Saccareau M, Courcoul A, Durand B. In silico Comparison of Test-and-Cull Protocols for Bovine Tuberculosis Control in France. Front Vet Sci. 23 oct 2018;5:265.

79. Canini L, Durand B. Resilience of French cattle farms to bovine tuberculosis detection between 2004 and 2017. Prev Vet Med. mars 2020;176:104902.

80. Garnier T, Eiglmeier K, Camus JC, Medina N, Mansoor H, Pryor M, et al. The complete genome sequence of Mycobacterium bovis. Proc Natl Acad Sci U S A. 24 juin 2003;100(13):7877-82.

81. Kao RR, Price-Carter M, Robbe-Austerman S. Use of genomics to track bovine tuberculosis transmission. Rev - Off Int Epizoot. avr 2016;35(1):241-58.

82. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, et al. Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. J Clin Microbiol. avr 1997;35(4):907-14.

83. Roring S, Scott A, Brittain D, Walker I, Hewinson G, Neill S, et al. Development of Variable-Number Tandem Repeat Typing of Mycobacterium bovis: Comparison of Results with Those Obtained by Using Existing Exact Tandem Repeats and Spoligotyping. J Clin Microbiol. juin 2002;40(6):2126-33.

84. Hauer A, Cruz KD, Cochard T, Godreuil S, Karoui C, Henault S, et al. Genetic Evolution of Mycobacterium bovis Causing Tuberculosis in Livestock and Wildlife in France since 1978. PLOS ONE. 6 févr 2015;10(2):e0117103.

85. Desvaux S, Réveillaud É, Richomme C, Boschiroli ML, Delavenne C, Calavas D, et al. Sylvatub: Bilan 2015-2017 de la surveillance de la tuberculose dans la faune sauvage. Bulletin épidémiologique. 2019;91(14):10.

86. Hauer A, Michelet L, Cochard T, Branger M, Nunez J, Boschiroli ML, et al. Accurate Phylogenetic Relationships Among Mycobacterium bovis Strains Circulating in France Based on Whole Genome Sequencing and Single Nucleotide Polymorphism Analysis. Front Microbiol. 2019;10.

87. Price-Carter M, Brauning R, de Lisle GW, Livingstone P, Neill M, Sinclair J, et al. Whole Genome Sequencing for Determining the Source of Mycobacterium bovis Infections in Livestock Herds and Wildlife in New Zealand. Front Vet Sci. 30 oct 2018;5.

88. Crispell J, Benton CH, Balaz D, De Maio N, Ahkmetova A, Allen A, et al. Combining genomics and epidemiology to analyse bi-directional transmission of Mycobacterium bovis in a multi-host system. eLife. 17 déc 2019;8.

89. Bouchez-Zacria M, Courcoul A, Jabert P, Richomme C, Durand B. Environmental determinants of the Mycobacterium bovis concomitant infection in cattle and badgers in France. Eur J Wildl Res. oct 2017;63(5):74.

90. Bouchez-Zacria M, Courcoul A, Durand B. The Distribution of Bovine Tuberculosis in Cattle Farms Is Linked to Cattle Trade and Badger-Mediated Contact Networks in South-Western France, 2007–2015. Front Vet Sci. 26 juill 2018;5:173.

91. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 1 févr 2019;35(3):526-8.

92. Felsenstein J, Churchill G. A Hidden Markov Model approach to variation among sites in rate of evolution. Mol Biol Evol 13: 93-104. Molecular biology and evolution. 1 févr 1996;13:93-104.

93. Crispell J, Zadoks RN, Harris SR, Paterson B, Collins DM, de-Lisle

GW, et al. Using whole genome sequencing to investigate transmission in a multi-host system: bovine tuberculosis in New Zealand. BMC Genomics. 16 2017;18(1):180.

94. Salvador LCM, O'Brien DJ, Cosgrove MK, Stuber TP, Schooley AM, Crispell J, et al. Disease management at the wildlife-livestock interface: Using whole-genome sequencing to study the role of elk in Mycobacterium bovis transmission in Michigan, USA. Molecular Ecology. 2019;28(9):2192-205.

95. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. Pertea M, éditeur. PLOS Computational Biology. 8 avr 2019;15(4):e1006650.

96. Müller NF, Rasmussen D, Stadler T. MASCOT: parameter and state inference under the marginal structured coalescent approximation. Bioinformatics. 15 nov 2018;34(22):3843-8.

97. Maturana P, Brewer BJ, Klaere S, Bouckaert R. Model selection and parameter inference in phylogenetics using Nested Sampling. Systematic Biology. 1 mars 2019;68(2):219-33.

98. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst Biol. 1 sept 2018;67(5):901-4.

99. Heled J, Bouckaert RR. Looking for trees in the forest: summary tree from posterior samples. BMC Evol Biol. 4 oct 2013;13:221.

100. Wang LG, Lam TTY, Xu S, Dai Z, Zhou L, Feng T, et al. Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. Molecular Biology and Evolution. 1 févr 2020;37(2):599-603.

101. Yu G. Using ggtree to Visualize Data on Tree-Like Structures. Current Protocols in Bioinformatics. 2020;69(1):e96.

102. Louca S, Doebeli M. Efficient comparative phylogenetics on large trees. Bioinformatics. 15 mars 2018;34(6):1053-5.

103. Biek R, O'Hare A, Wright D, Mallon T, McCormick C, Orton RJ, et al. Whole Genome Sequencing Reveals Local Transmission Patterns of Mycobacterium bovis in Sympatric Cattle and Badger Populations. PLOS Pathogens. 29 nov 2012;8(11):e1003008.

104. Trewby H, Wright D, Breadon EL, Lycett SJ, Mallon TR, McCormick C, et al. Use of bacterial whole-genome sequencing to investigate local persistence and spread in bovine tuberculosis. Epidemics. mars 2016;14:26-35.

105. Mestre O, Luo T, Vultos TD, Kremer K, Murray A, Namouchi A, et al. Phylogeny of Mycobacterium tuberculosis Beijing Strains Constructed from Polymorphisms in Genes Involved in DNA Replication, Recombination and Repair. PLOS ONE. 20 janv 2011;6(1):e16020.

106. Smith NH, Berg S, Dale J, Allen A, Rodriguez S, Romero B, et al. European 1: A globally important clonal complex of Mycobacterium bovis. Infection, Genetics and Evolution. 1 août 2011;11(6):1340-51.

107. Rossi G, Crispell J, Brough T, Lycett SJ, White PCL, Allen A, et al. Phylodynamic analysis of an emergent Mycobacterium bovis outbreak in an area with no previously known wildlife infections. Journal of Applied Ecology. 2022;59(1):210-22.

108. Jacquier M, Vandel JM, Léger F, Duhayer J, Pardonnet S, Say L, et al. Breaking down population density into different components to better understand its spatial variation. 12 mai 2021;21.

109. Donnelly CA, Nouvellet P. The contribution of badgers to confirmed tuberculosis in cattle in high-incidence areas in England. PLoS Curr. 10 oct 2013;5:ecurrents.outbreaks.097a904d3f3619db2fe78d24bc776098.

110. Bouchez-Zacria M. Rôles de l'environnement et des contacts intra et interspécifiques dans la transmission de Mycobacterium bovis dans le système bovins-blaireaux en Pyrénées-Atlantiques – Landes [These de doctorat]. Université Paris-Saclay (ComUE);

2018.

111. Byrne AW, Quinn JL, O'Keeffe JJ, Green S, Sleeman DP, Martin SW, et al. Large-scale movements in European badgers: has the tail of the movement kernel been underestimated? J Anim Ecol. juill 2014;83(4):991-1001.

112. Podgórski T, Baś G, Jędrzejewska B, Sönnichsen L, Śnieżko S, Jędrzejewski W, et al. Spatiotemporal behavioral plasticity of wild boar (Sus scrofa) under contrasting conditions of human pressure: primeval forest and metropolitan area. Journal of Mammalogy. 15 févr 2013;94(1):109-19.

113. Cayley A. A theorem on trees. Quarterly Journal of Pure and Applied Mathematics. 1889;23:376-8.

114. Klinkenberg D, Backer JA, Didelot X, Colijn C, Wallinga J. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. PLoS computational biology. 2017/05/26 éd. mai 2017;13(5):e1005495.

115. Alexandersen S, Zhang Z, Donaldson AI, Garland AJM. The pathogenesis and diagnosis of Foot-and-Mouth Disease. Journal of Comparative Pathology. 1 juill 2003;129(1):1-36.

116. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution. juin 1980;16(2):111-20.

117. Woolhouse M. Quantifying Transmission. Microbiol Spectr. juill 2017;5(4).

118. van Seventer JM, Hochberg NS. Principles of infectious diseases: transmission, diagnosis, prevention, and control. International Encyclopedia of Public Health. 2017;22-39.

119. Svensson Å. A note on generation times in epidemic models. Mathematical Biosciences. 1 juill 2007;208(1):300-11.

120. Aldrin M, Lyngstad TM, Kristoffersen AB, Storvik B, Borgan Ø, Jansen PA. Modelling the spread of infectious salmon anaemia

among salmon farms based on seaway distances between farms and genetic relationships between infectious salmon anaemia virus isolates. Journal of the Royal Society Interface. 7 sept 2011;8(62):1346-56.

121. Campbell F, Cori A, Ferguson N, Jombart T. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. PLoS computational biology. 2019/03/30 éd. mars 2019;15(3):e1006930.

122. Sashittal P, El-Kebir M. Sampling and summarizing transmission trees with multi-strain infections. Bioinformatics (Oxford, England). 2020/07/14 éd. 1 juill 2020;36(Supplement_1):i362-70.

123. Drummond AJ, Bouckaert RR. Bayesian Evolutionary Analysis with BEAST. Cambridge University Press; 2015. 263 p.

124. Hall M, Woolhouse M, Rambaut A. Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. PLoS Comput Biol. déc 2015;11(12):e1004613.

125. Ypma RJ, Bataille AM, Stegeman A, Koch G, Wallinga J, van Ballegooijen WM. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proceedings Biological sciences. 2011/07/08 éd. 7 févr 2012;279(1728):444-50.

126. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS computational biology. 2014/01/28 éd. janv 2014;10(1):e1003457.

127. Cerutti F, Luzzago C, Lauzi S, Ebranati E, Caruso C, Masoero L, et al. Phylogeography, phylodynamics and transmission chains of bovine viral diarrhea virus subtype 1f in Northern Italy. Infect Genet Evol. nov 2016;45:262-7.

128. Stoesser N, Giess A, Batty EM, Sheppard AE, Walker AS, Wilson DJ, et al. Genome Sequencing of an Extended Series of NDM-Producing Klebsiella pneumoniae Isolates from Neonatal

Infections in a Nepali Hospital Characterizes the Extent of Community- versus Hospital-Associated Transmission in an Endemic Setting. Antimicrob Agents Chemother. déc 2014;58(12):7347-57.

129. Kanamori H, Parobek CM, Weber DJ, van Duin D, Rutala WA, Cairns BA, et al. Next-generation sequencing and comparative analysis of sequential outbreaks caused by multidrug-resistant Acinetobacter baumannii at a large academic burn center. Antimicrobial agents and chemotherapy. 2015/12/09 éd. 7 déc 2015;60(3):1249-57.

130. Makke G, Bitar I, Salloum T, Panossian B, Alousi S, Arabaghian H, et al. Whole-genome-sequence-based characterization of extensively drug-resistant Acinetobacter baumannii hospital outbreak. mSphere. 2020/01/17 éd. 15 janv 2020;5(1).

131. Worby CJ, Chang HH, Hanage WP, Lipsitch M. The distribution of pairwise genetic distances: a tool for investigating disease transmission. Genetics. 2014/10/15 éd. déc 2014;198(4):1395-404.

132. Worby CJ, O'Neill PD, Kypraios T, Robotham JV, De Angelis D, Cartwright EJ, et al. Reconstructing transmission trees for communicable diseases using densely sampled genetic data. The annals of applied statistics. 2016/04/05 éd. mars 2016;10(1):395-417.

133. Jombart T, Eggo RM, Dodd PJ, Balloux F. Reconstructing disease outbreaks from genetic data: a graph approach. Heredity. 2010/06/17 éd. févr 2011;106(2):383-90.

134. Edmonds J. Optimum branchings. Journal of Research of the National Bureau of Standards Section B Mathematics and Mathematical Physics. oct 1967;71B(4):233.

135. Hughes J, Allen RC, Baguelin M, Hampson K, Baillie GJ, Elton D, et al. Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks. PLoS pathogens. 2013/01/12 éd. déc 2012;8(12):e1003081.

136. Guerra-Assuncao JA, Crampin AC, Houben RM, Mzembe T, Mallard K, Coll F, et al. Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. Elife. 2015/03/04 éd. 3 mars 2015;4(4).

137. Spencer MD, Winglee K, Passaretti C, Earl AM, Manson AL, Mulder HP, et al. Whole genome sequencing detects inter-facility transmission of carbapenem-resistant Klebsiella pneumoniae. The Journal of infection. 2018/12/07 éd. mars 2019;78(3):187-99.

138. Famulare M, Hu H. Extracting transmission networks from phylogeographic data for epidemic and endemic diseases: Ebola virus in Sierra Leone, 2009 H1N1 pandemic influenza and polio in Nigeria. Int Health. mars 2015;7(2):130-8.

139. Eldholm V, Rieux A, Monteserin J, Lopez JM, Palmero D, Lopez B, et al. Impact of HIV co-infection on the evolution and transmission of multidrug-resistant tuberculosis. eLife. 2016/08/10 éd. 9 août 2016;5.

140. Cottam EM, Thebaud G, Wadsworth J, Gloster J, Mansley L, Paton DJ, et al. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. Proceedings Biological sciences. 2008/01/31 éd. 22 avr 2008;275(1637):887-95.

141. Didelot X, Fraser C, Gardy J, Colijn C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. Molecular biology and evolution. 2017/01/20 éd. 1 avr 2017;34(4):997-1007.

142. Séraphin MN, Didelot X, Nolan DJ, May JR, Khan MSR, Murray ER, et al. Genomic investigation of a Mycobacterium tuberculosis outbreak involving prison and community cases in Florida, United States. American Journal of Tropical Medicine and Hygiene. 2018;99(4):867-74.

143. Xu Y, Cancino-Munoz I, Torres-Puente M, Villamayor LM, Borras R, Borras-Manez M, et al. High-resolution mapping of tuberculosis transmission: Whole genome sequencing and

phylogenetic modelling of a cohort from Valencia Region, Spain. PLoS medicine. 2019/11/02 éd. oct 2019;16(10):e1002961.

144. Sobkowiak B, Banda L, Mzembe T, Crampin AC, Glynn JR, Clark TG. Bayesian reconstruction of Mycobacterium tuberculosis transmission networks in a high incidence area over two decades in Malawi reveals associated risk factors and genomic variants. Microb Genom. avr 2020;6(4).

145. Kwong JC, Lane CR, Romanes F, Goncalves da Silva A, Easton M, Cronin K, et al. Translating genomics into practice for real-time surveillance and response to carbapenemase-producing Enterobacteriaceae: evidence from a complex multi-institutional KPC outbreak. PeerJ. 2018/01/10 éd. 2018;6:e4210.

146. Van Dorp L, Wang Q, Shaw LP, Acman M, Brynildsrud OB, Eldholm V, et al. Rapid phenotypic evolution in multidrug-resistant Klebsiella pneumoniae hospital outbreak strains. Microbial Genomics. 2019;5(4):1-11.

147. Wang L, Didelot X, Yang J, Wong G, Shi Y, Liu W, et al. Inference of person-to-person transmission of COVID-19 reveals hidden super-spreading events during the early outbreak phase. Nature communications. 2020/10/08 éd. 6 oct 2020;11(1):5006.

148. Stapleton PJ, Eshaghi A, Seo CY, Wilson S, Harris T, Deeks SL, et al. Evaluating the use of whole genome sequencing for the investigation of a large mumps outbreak in Ontario, Canada. Scientific reports. 2019/09/01 éd. 30 août 2019;9(1):12615.

149. Xu Y, Stockdale JE, Naidu V, Hatherell H, Stimson J, Stagg HR, et al. Transmission analysis of a large tuberculosis outbreak in London: a mathematical modelling study using genomic data. Microbial genomics. 11 nov 2020;

150. Morelli MJ, Thebaud G, Chadoeuf J, King DP, Haydon DT, Soubeyrand S. A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. PLoS computational biology. 2012/11/21 éd. 2012;8(11):e1002768.

151. Mollentze N, Nel LH, Townsend S, le Roux K, Hampson K, Haydon DT, et al. A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. Proceedings Biological sciences. 2014/03/13 éd. 7 mai 2014;281(1782):20133251.

152. Hayama Y, Firestone SM, Stevenson MA, Yamamoto T, Nishi T, Shimizu Y, et al. Reconstructing a transmission network and identifying risk factors of secondary transmissions in the 2010 foot-and-mouth disease outbreak in Japan. Transboundary and emerging diseases. 2019/05/28 éd. sept 2019;66(5):2074-86.

153. Montazeri H, Little S, Mozaffarilegha M, Beerenwinkel N, DeGruttola V. Bayesian reconstruction of transmission trees from genetic sequences and uncertain infection times. Stat Appl Genet Mol Biol. 21 oct 2020;/j/sagmb.ahead-of-print/sagmb-2019-0026/sagmb-2019-0026.xml.

154. Choi SC. Inferring transmission routes of avian influenza during the H5N8 outbreak of South Korea in 2014 using epidemiological and genetic data. Korean Journal of Microbiology. 2018;54(3):254-65.

155. De Maio N, Wu CH, Wilson DJ. SCOTTI: Efficient Reconstruction of Transmission within Outbreaks with the Structured Coalescent. PLoS computational biology. 2016/09/30 éd. sept 2016;12(9):e1005130.

156. Azarian T, Maraqa NF, Cook RL, Johnson JA, Bailey C, Wheeler S, et al. Genomic Epidemiology of Methicillin-Resistant Staphylococcus aureus in a Neonatal Intensive Care Unit. PloS one. 2016/10/13 éd. 2016;11(10):e0164397.

157. De Maio N, Worby CJ, Wilson DJ, Stoesser N. Bayesian reconstruction of transmission within outbreaks using genomic variants. Koelle K, éditeur. PLoS Comput Biol. 18 avr 2018;14(4):e1006117.

158. Alamil M, Hughes J, Berthier K, Desbiez C, Thébaud G, Soubeyrand S. Inferring epidemiological links from deep sequencing data: a

statistical learning approach for human, animal and plant diseases. Philosophical Transactions of the Royal Society B: Biological Sciences. 24 juin 2019;374(1775):20180258.

159. Willgert K, Didelot X, Surendran-Nair M, Kuchipudi SV, Ruden RM, Yon M, et al. Transmission history of SARS-CoV-2 in humans and white-tailed deer. Sci Rep. 15 juill 2022;12:12094.

160. Bouchez-Zacria M, Ruette S, Richomme C, Lesellier S, Payne A, Boschiroli ML, et al. Analysis of a multi-type resurgence of Mycobacterium bovis in cattle and badgers in Southwest France, 2007-2019. Veterinary Research. 3 mai 2023;54(1):41.

161. Baubet E, Servanty S, Brandt S, Toïgo C, Klein F. Améliorer la connaissance du fonctionnement démographique des populations de sangliers : vers une meilleure gestion de l'espèce Sus scrofa. ONCFS Rapport scientifique 2004. 2004;30-3.

162. Zanella G, Bar-Hen A, Boschiroli ML, Hars J, Moutou F, Garin-Bastuji B, et al. Modelling transmission of bovine tuberculosis in red deer and wild boar in Normandy, France. Zoonoses Public Health. sept 2012;59 Suppl 2:170-8.

163. Duault H, Michelet L, Boschiroli ML, Durand B, Canini L. A Bayesian evolutionary model towards understanding wildlife contribution to F4-family Mycobacterium bovis transmission in the South-West of France. Veterinary Research. 2 avr 2022;53(1):28.

164. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. J Phys Chem. 1 déc 1977;81(25):2340-61.

165. De Maio N, Boulton W, Weilguny L, Walker CR, Turakhia Y, Corbett-Detig R, et al. phastSim: efficient simulation of sequence evolution for pandemic-scale datasets. bioRxiv. 23 sept 2021;2021.03.15.435416.

166. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. janv 2016;2(1):vew007.

167. Duault H, Durand B, Canini L. Methods Combining Genomic and Epidemiological Data in the Reconstruction of Transmission Trees: A Systematic Review. Pathogens. 15 févr 2022;11(2):252.

168. Plummer M, Best N, Cowles K, Vines K. CODA: convergence diagnosis and output analysis for MCMC. R News. mars 2006;6(1):7-11.

169. Jombart T, Cori A, Finger F. Small Helpers and Tricks for Epidemics Analysis [Internet]. [cité 10 mars 2022]. Disponible sur: http://www.repidemicsconsortium.org/epitrix/

170. Didelot X, Kendall M, Xu Y, White PJ, McCarthy N. Genomic Epidemiology Analysis of Infectious Disease Outbreaks Using TransPhylo. Curr Protoc. févr 2021;1(2):e60.

171. Kendall M, Ayabina D, Xu Y, Stimson J, Colijn C. Estimating Transmission from Genetic and Epidemiological Data: A Metric to Compare Transmission Trees. Statistical Science. 2018;33(1):70-85.

172. Firestone SM, Hayama Y, Bradhurst R, Yamamoto T, Tsutsui T, Stevenson MA. Reconstructing foot-and-mouth disease outbreaks: a methods comparison of transmission network models. Scientific reports. 2019/03/20 éd. 18 mars 2019;9(1):4809.

173. Sobkowiak B, Romanowski K, Sekirov I, Gardy JL, Johnston J. Comparing transmission reconstruction models with Mycobacterium tuberculosis whole genome sequence data. bioRxiv; 2022. p. 2022.01.07.475333.

174. Nigsch A, Robbe-Austerman S, Stuber TP, Pavinski Bitar PD, Gröhn YT, Schukken YH. Who infects whom?-Reconstructing infection chains of Mycobacterium avium ssp. paratuberculosis in an endemically infected dairy herd by use of genomic data. PLoS One. 2021;16(5):e0246983.

175. van Tonder AJ, Thornton MJ, Conlan AJK, Jolley KA, Goolding L, Mitchell AP, et al. Inferring Mycobacterium bovis transmission between cattle and badgers using isolates from the Randomised

Badger Culling Trial. PLoS Pathog. nov 2021;17(11):e1010075.

176. Akhmetova A, Guerrero J, McAdam P, Salvador LCM, Crispell J, Lavery J, et al. Genomic epidemiology of Mycobacterium bovis infection in sympatric badger and cattle populations in Northern Ireland. Microbial Genomics. 2023;9(5):001023.

177. Canini L, Modenesi G, Courcoul A, Boschiroli ML, Durand B, Michelet L. Deciphering the role of host species for two Mycobacterium bovis genotypes from the European 3 clonal complex circulation within a cattle-badger-wild boar multihost system. MicrobiologyOpen. 2023;12(1):e1331.

178. Xue Y, Chen D, Smith SR, Ruan X, Tang S. Coupling the Within-Host Process and Between-Host Transmission of COVID-19 Suggests Vaccination and School Closures are Critical. Bull Math Biol. 2023;85(1):6.

179. Colangeli R, Arcus VL, Cursons RT, Ruthe A, Karalus N, Coley K, et al. Whole Genome Sequencing of Mycobacterium tuberculosis Reveals Slow Growth and Low Mutation Rates during Latent Infections in Humans. PLoS One. 11 mars 2014;9(3):e91024.

180. Layan M, Müller NF, Dellicour S, De Maio N, Bourhy H, Cauchemez S, et al. Impact and mitigation of sampling bias to determine viral spread: Evaluating discrete phylogeography through CTMC modeling and structured coalescent model approximations. Virus Evol. 2023;9(1):vead010.

181. Marsot M, Bernard C, Payne A, Rossi S, Ruette S, Desvaux S, et al. "BACACIX", a spatial index combining proxies of bovine and badger space use associated with extended Mycobacterium bovis circulation in France. Preventive Veterinary Medicine. 1 févr 2023;211:105817.

182. Akhmetova A, Guerrero J, McAdam P, Salvador LCM, Crispell J, Lavery J, et al. Genomic epidemiology of Mycobacterium bovis infection in sympatric badger and cattle populations in Northern Ireland. bioRxiv; 2021. p. 2021.03.12.435101.

183. Müller NF, Dudas G, Stadler T. Inferring time-dependent migration and coalescence patterns from genetic sequence and predictor data in structured populations. Virus Evol. 14 août 2019;5(2):vez030.

184. Scaduto DI, Brown JM, Haaland WC, Zwickl DJ, Hillis DM, Metzker ML. Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. Proceedings of the National Academy of Sciences of the United States of America. 2010/11/17 éd. 14 déc 2010;107(50):21242-7.

185. Schurch AC, Kremer K, Daviena O, Kiers A, Boeree MJ, Siezen RJ, et al. High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. Journal of clinical microbiology. 2010/07/02 éd. sept 2010;48(9):3403-6.

186. Shiino T. Phylodynamic analysis of a viral infection network. Front Microbiol. 31 juill 2012;3:278.

187. Zarrabi N, Prosperi M, Belleman RG, Colafigli M, De Luca A, Sloot PM. Combining epidemiological and genetic networks signifies the importance of early treatment in HIV-1 transmission. PloS one. 2012/10/03 éd. 2012;7(9):e46156.

188. Alam SJ, Zhang X, Romero-Severson EO, Henry C, Zhong L, Volz EM, et al. Detectable signals of episodic risk effects on acute HIV transmission: strategies for analyzing transmission systems using genetic data. Epidemics. 2013/02/27 éd. mars 2013;5(1):44-55.

189. Stack JC, Murcia PR, Grenfell BT, Wood JLN, Holmes EC. Inferring the inter-host transmission of influenza A virus using patterns of intra-host genetic variation. Poceedings - Royal Society of London, B. 2013;280(1750):20122173.

190. Gavryushkina A, Welch D, Stadler T, Drummond AJ. Bayesian Inference of Sampled Ancestor Trees for Epidemiology and Fossil Calibration. PLOS Computational Biology. 4 déc 2014;10(12):e1003919.

191. Mehaffy C, Guthrie JL, Alexander DC, Stuart R, Rea E, Jamieson FB.

Marked microevolution of a unique Mycobacterium tuberculosis strain in 17 years of ongoing transmission in a high risk population. PloS one. 2014/11/19 éd. 2014;9(11):e112928.

192. Numminen E, Chewapreecha C, Siren J, Turner C, Turner P, Bentley SD, et al. Two-phase importance sampling for inference about transmission trees. Proceedings Biological sciences. 2014/09/26 éd. 7 nov 2014;281(1794):20141324.

193. Croucher NJ, Didelot X. The application of genomics to tracing bacterial pathogen transmission. Current opinion in microbiology. 2014/12/03 éd. févr 2015;23:62-7.

194. Janies DA, Pomeroy LW, Krueger C, Zhang Y, Senturk IF, Kaya K, et al. Phylogenetic visualization of the spread of H7 influenza A viruses. Cladistics. 2015;31(6):679-91.

195. Valdazo-Gonzalez B, Kim JT, Soubeyrand S, Wadsworth J, Knowles NJ, Haydon DT, et al. The impact of within-herd genetic variation upon inferred transmission trees for foot-and-mouth disease virus. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases. 2015/04/12 éd. juin 2015;32:440-8.

196. Folarin OA, Ehichioya D, Schaffner SF, Winnicki SM, Wohl S, Eromon P, et al. Ebola Virus Epidemiology and Evolution in Nigeria. The Journal of infectious diseases. 2016/07/06 éd. 15 oct 2016;214(suppl 3):S102-9.

197. Hall MD, Woolhouse ME, Rambaut A. Using genomics data to reconstruct transmission trees during disease outbreaks. Revue scientifique et technique (International Office of Epizootics). 2016/05/25 éd. avr 2016;35(1):287-96.

198. Hoffmann M, Luo Y, Monday SR, Gonzalez-Escalona N, Ottesen AR, Muruvanda T, et al. Tracing Origins of the Salmonella Bareilly Strain Causing a Food-borne Outbreak in the United States. The Journal of infectious diseases. 2015/05/23 éd. 15 févr 2016;213(4):502-8.

199. Kenah E, Britton T, Halloran ME, Longini IM Jr. Molecular Infectious Disease Epidemiology: Survival Analysis and Algorithms Linking Phylogenies to Transmission Trees. PLoS computational biology. 2016/04/14 éd. avr 2016;12(4):e1004869.

200. Parratt SR, Numminen E, Laine AL. Infectious Disease Dynamics in Heterogeneous Landscapes. 2016. 283-306 p. (Annual Review of Ecology, Evolution, and Systematics; vol. 47).

201. Ray B, Ghedin E, Chunara R. Network inference from multimodal data: A review of approaches from infectious disease transmission. Journal of biomedical informatics. 2016/09/11 éd. déc 2016;64:44-54.

202. Ren H, Jin Y, Hu M, Zhou J, Song T, Huang Z, et al. Ecological dynamics of influenza A viruses: cross-species transmission and global migration. Sci Rep. 2016/11/09 éd. 9 nov 2016;6:36839.

203. Wohl S, Schaffner SF, Sabeti PC. Genomic Analysis of Viral Outbreaks. Annual review of virology. 2016/08/09 éd. 29 sept 2016;3(1):173-95.

204. Xu W, Berhane Y, Dube C, Liang B, Pasick J, VanDomselaar G, et al. Epidemiological and Evolutionary Inference of the Transmission Network of the 2014 Highly Pathogenic Avian Influenza H5N2 Outbreak in British Columbia, Canada. Scientific reports. 2016/08/05 éd. 4 août 2016;6:30858.

205. Agoti CN, Munywoki PK, Phan MVT, Otieno JR, Kamau E, Bett A, et al. Transmission patterns and evolution of respiratory syncytial virus in a community outbreak identified by genomic analysis. Virus evolution. 2017/05/02 éd. janv 2017;3(1):vex006.

206. Baele G, Suchard MA, Rambaut A, Lemey P. Emerging Concepts of Data Integration in Pathogen Phylodynamics. Systematic biology. 2017/02/09 éd. 1 janv 2017;66(1):e47-65.

207. Glebova O, Knyazev S, Melnyk A, Artyomenko A, Khudyakov Y, Zelikovsky A, et al. Inference of genetic relatedness between viral quasispecies from sequencing data. BMC Genomics. 2017/12/16

éd. 6 déc 2017;18(Suppl 10):918.

208. Snitkin ES, Won S, Pirani A, Lapp Z, Weinstein RA, Lolans K, et al. Integrated genomic and interfacility patient-transfer data reveal the transmission pathways of multidrug-resistant Klebsiella pneumoniae in a regional outbreak. Science translational medicine. 2017/11/24 éd. 22 nov 2017;9(417).

209. Worby CJ, Lipsitch M, Hanage WP. Shared Genomic Variants: Identification of Transmission Routes Using Pathogen Deep-Sequence Data. Am J Epidemiol. 2017/11/18 éd. 15 nov 2017;186(10):1209-16.

210. Campbell F, Didelot X, Fitzjohn R, Ferguson N, Cori A, Jombart T. outbreaker2: a modular platform for outbreak reconstruction. BMC bioinformatics. 2018/10/23 éd. 22 oct 2018;19(Suppl 11):363.

211. Choi SC. Genomic epidemiology for microbial evolutionary studies and the use of Oxford Nanopore sequencing technology. Korean Journal of Microbiology. 2018;54(3):188-99.

212. Ezeoke I, Galac MR, Lin Y, Liem AT, Roth PA, Kilianski A, et al. Tracking a serial killer: Integrating phylogenetic relationships, epidemiology, and geography for two invasive meningococcal disease outbreaks. PLoS One. 2018/11/30 éd. 2018;13(11):e0202615.

213. Gotoh Y, Taniguchi T, Yoshimura D, Katsura K, Saeki Y, Hirabara Y, et al. Multi-step genomic dissection of a suspected intra-hospital Helicobacter cinaedi outbreak. Microbial genomics. 2019/01/11 éd. déc 2018;4(12).

214. Leitner T, Romero-Severson E. Phylogenetic patterns recover known HIV epidemiological relationships and reveal common transmission of multiple variants. Nature microbiology. 2018/08/01 éd. sept 2018;3(9):983-8.

215. Meehan CJ, Moris P, Kohl TA, Pecerska J, Akter S, Merker M, et al. The relationship between transmission time and clustering

methods in Mycobacterium tuberculosis epidemiology. EBioMedicine. 2018/10/21 éd. nov 2018;37:410-6.

216. Payne DC, Biggs HM, Al-Abdallat MM, Alqasrawi S, Lu X, Abedi GR, et al. Multihospital Outbreak of a Middle East Respiratory Syndrome Coronavirus Deletion Variant, Jordan: A Molecular, Serologic, and Epidemiologic Investigation. Open forum infectious diseases. 2018/10/09 éd. mai 2018;5(5):ofy095.

217. Blackburn RM, Frampton D, Smith CM, Fragaszy EB, Watson SJ, Ferns RB, et al. Nosocomial transmission of influenza: A retrospective cross-sectional study using next generation sequencing at a hospital in England (2012-2014). Influenza and other respiratory viruses. 2019/09/20 éd. nov 2019;13(6):556-63.

218. DeSilva MB, Styles T, Basler C, Moses FL, Husain F, Reichler M, et al. Introduction of Ebola virus into a remote border district of Sierra Leone, 2014: use of field epidemiology and RNA sequencing to describe chains of transmission. Epidemiology and infection. 2019/03/15 éd. janv 2019;147:e88.

219. Guthrie JL, Strudwick L, Roberts B, Allen M, McFadzen J, Roth D, et al. Whole genome sequencing for improved understanding of Mycobacterium tuberculosis transmission in a remote circumpolar region. Epidemiol Infect. 2019/08/01 éd. janv 2019;147:e188.

220. Hall MD, Colijn C. Transmission Trees on a Known Pathogen Phylogeny: Enumeration and Sampling. Molecular biology and evolution. 2019/03/16 éd. 1 juin 2019;36(6):1333-43.

221. Hall MD, Holden MT, Srisomang P, Mahavanakul W, Wuthiekanun V, Limmathurotsakul D, et al. Improved characterisation of MRSA transmission using within-host bacterial sequence diversity. eLife. 2019/10/09 éd. 8 oct 2019;8.

222. Ratmann O, Grabowski MK, Hall M, Golubchik T, Wymant C, Abeler-Dörner L, et al. Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis. Nat Commun. 29 mars 2019;10(1):1411.

223. Sashittal P, El-Kebir M. SharpTNI: Counting and Sampling Parsimonious Transmission Networks under a Weak Bottleneck. 2019;12.

224. van Beek J, Raisanen K, Broas M, Kauranen J, Kahkola A, Laine J, et al. Tracing local and regional clusters of carbapenemase-producing Klebsiella pneumoniae ST512 with whole genome sequencing, Finland, 2013 to 2018. Euro Surveill. 2019/09/26 éd. sept 2019;24(38).

225. Vaughan TG, Leventhal GE, Rasmussen DA, Drummond AJ, Welch D, Stadler T. Estimating Epidemic Incidence and Prevalence from Genomic Data. Molecular biology and evolution. 2019/05/07 éd. 1 août 2019;36(8):1804-16.

226. Bbosa N, Ssemwanga D, Ssekagiri A, Xi X, Mayanja Y, Bahemuka U, et al. Phylogenetic and Demographic Characterization of Directed HIV-1 Transmission Using Deep Sequences from High-Risk and General Population Cohorts/Groups in Uganda. Viruses. 18 mars 2020;12(3).

227. de Bernardi Schneider A, Ford CT, Hostager R, Williams J, Cioce M, Catalyurek UV, et al. StrainHub: a phylogenetic tool to construct pathogen transmission networks. Bioinformatics (Oxford, England). 2019/08/17 éd. 1 févr 2020;36(3):945-7.

228. Dhar S, Zhang C, Mandoiu I, Bansal MS. TNet: Transmission Network Inference Using Within-Host Strain Diversity and its Application to Geographical Tracking of COVID-19 Spread. IEEE/ACM Trans Comput Biol Bioinform. 13 juill 2021;PP.

229. Nelson KN, Gandhi NR, Mathema B, Lopman BA, Brust JCM, Auld SC, et al. Modeling Missing Cases and Transmission Links in Networks of Extensively Drug-Resistant Tuberculosis in KwaZulu-Natal, South Africa. American journal of epidemiology. 2020/04/04 éd. 1 juill 2020;189(7):735-45.

230. Nelson KN, Jenness SM, Mathema B, Lopman BA, Auld SC, Shah NS, et al. Social Mixing and Clinical Features Linked With Transmission in a Network of Extensively Drug-resistant

Tuberculosis Cases in KwaZulu-Natal, South Africa. Clin Infect Dis. 2019/07/26 éd. 23 mai 2020;70(11):2396-402.

231. Wang X, Zhou Q, He Y, Liu L, Ma X, Wei X, et al. Nosocomial outbreak of COVID-19 pneumonia in Wuhan, China. The European respiratory journal. 2020/05/06 éd. juin 2020;55(6).

232. Worobey M, Pekar J, Larsen BB, Nelson MI, Hill V, Joy JB, et al. The emergence of SARS-CoV-2 in Europe and North America. Science. 30 oct 2020;370(6516):564-70.

233. Romero-Severson E, Nasir A, Leitner T. What Should Health Departments Do with HIV Sequence Data? Viruses. 2020/09/17 éd. 12 sept 2020;12(9).

234. Bataille A, van der Meer F, Stegeman A, Koch G. Evolutionary analysis of inter-farm transmission dynamics in a highly pathogenic avian influenza epidemic. PLoS pathogens. 2011/07/07 éd. juin 2011;7(6):e1002094.

235. Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature. juin 2009;459(7250):1122-5.

236. Briand FX, Niqueux E, Schmitz A, Martenot C, Cherbonnel M, Massin P, et al. Highly Pathogenic Avian Influenza A(H5N8) Virus Spread by Short- and Long-Range Transmission, France, 2016–17. Emerging Infectious Diseases journal - CDC. févr 2021;27(2).

237. Murcia PR, Wood JLN, Holmes EC. Genome-scale evolution and phylodynamics of equine H3N8 influenza A virus. J Virol. juin 2011;85(11):5312-22.

238. Biek R, Henderson JC, Waller LA, Rupprecht CE, Real LA. A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. Proc Natl Acad Sci U S A. 8 mai 2007;104(19):7993-8.

239. Vega VB, Ruan Y, Liu J, Lee WH, Wei CL, Se-Thoe SY, et al.

Mutational dynamics of the SARS coronavirus in cell culture and human populations isolated in 2003. BMC Infect Dis. déc 2004;4(1):32.

240. Nie Q, Li X, Chen W, Liu D, Chen Y, Li H, et al. Phylogenetic and phylodynamic analyses of SARS-CoV-2. Virus Res. 2 oct 2020;287:198098.

241. Vrancken B, Rambaut A, Suchard MA, Drummond A, Baele G, Derdelinckx I, et al. The genealogical population dynamics of HIV-1 in a large transmission chain: bridging within and among host evolutionary rates. PLoS Comput Biol. avr 2014;10(4):e1003505.

242. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 12 sept 2014;345(6202):1369-72.

243. Burns CC, Shaw J, Jorba J, Bukbuk D, Adu F, Gumede N, et al. Multiple Independent Emergences of Type 2 Vaccine-Derived Polioviruses during a Large Outbreak in Northern Nigeria. J Virol. mai 2013;87(9):4907-22.

244. Devold M, Karlsen M, Nylund A. Sequence analysis of the fusion protein gene from infectious salmon anemia virus isolates: evidence of recombination and reassortment. J Gen Virol. juill 2006;87(Pt 7):2031-40.

245. Kibenge FSB, Kibenge MJT, Wang Y, Qian B, Hariharan S, McGeachy S. Mapping of putative virulence motifs on infectious salmon anemia virus surface glycoprotein genes. J Gen Virol. nov 2007;88(Pt 11):3100-11.

246. Karami-Zarandi M, Douraghi M, Vaziri B, Adibhesami H, Rahbar M, Yaseri M. Variable spontaneous mutation rate in clinical strains of multidrug-resistant Acinetobacter baumannii and differentially expressed proteins in a hypermutator strain. Mutat Res. août 2017;800-802:37-45.

247. Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK,

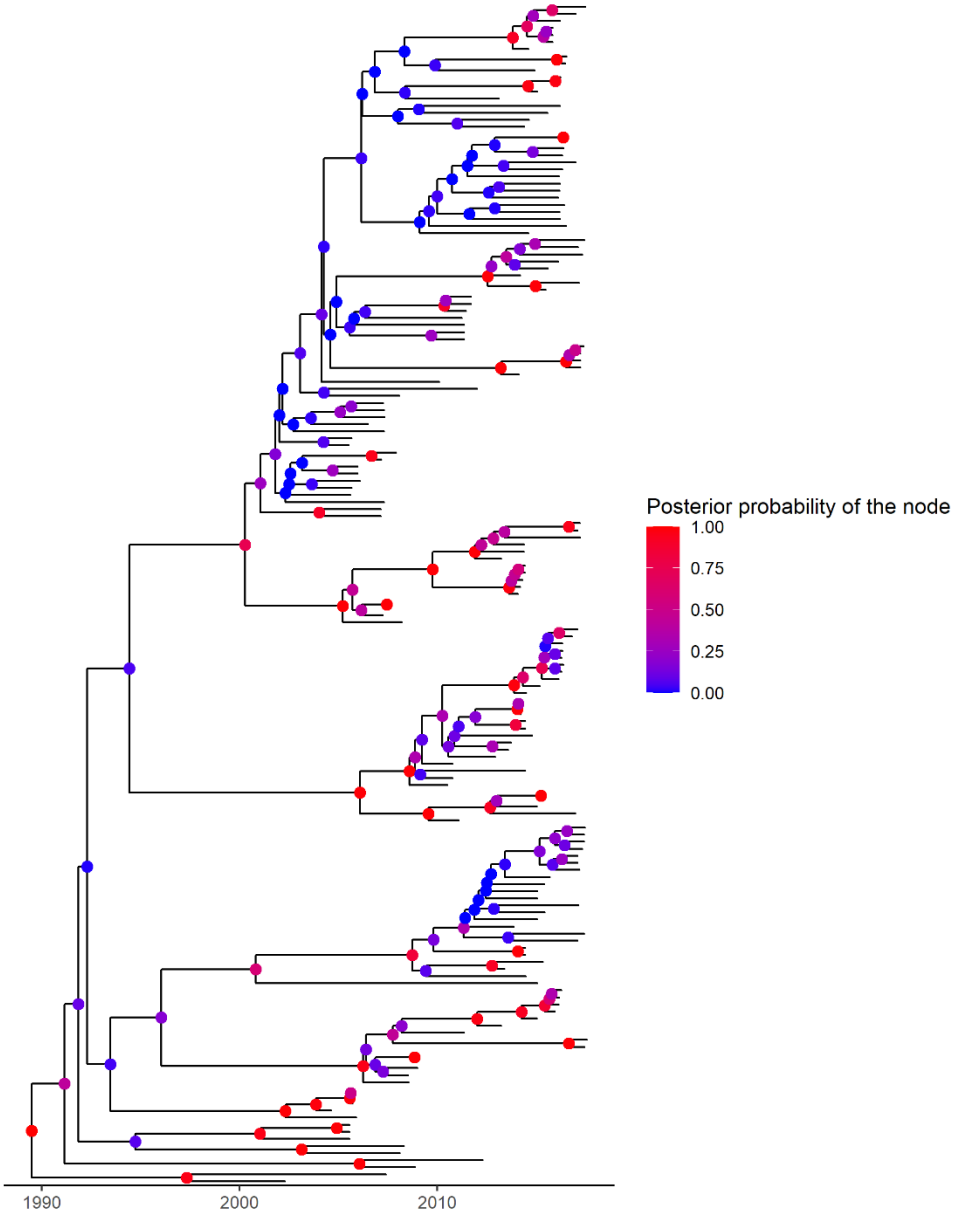
Chantratita N, et al. Evolution of MRSA During Hospital Transmission and Intercontinental Spread. Science. 22 janv 2010;327(5964):469-74.

7 ANNEXES

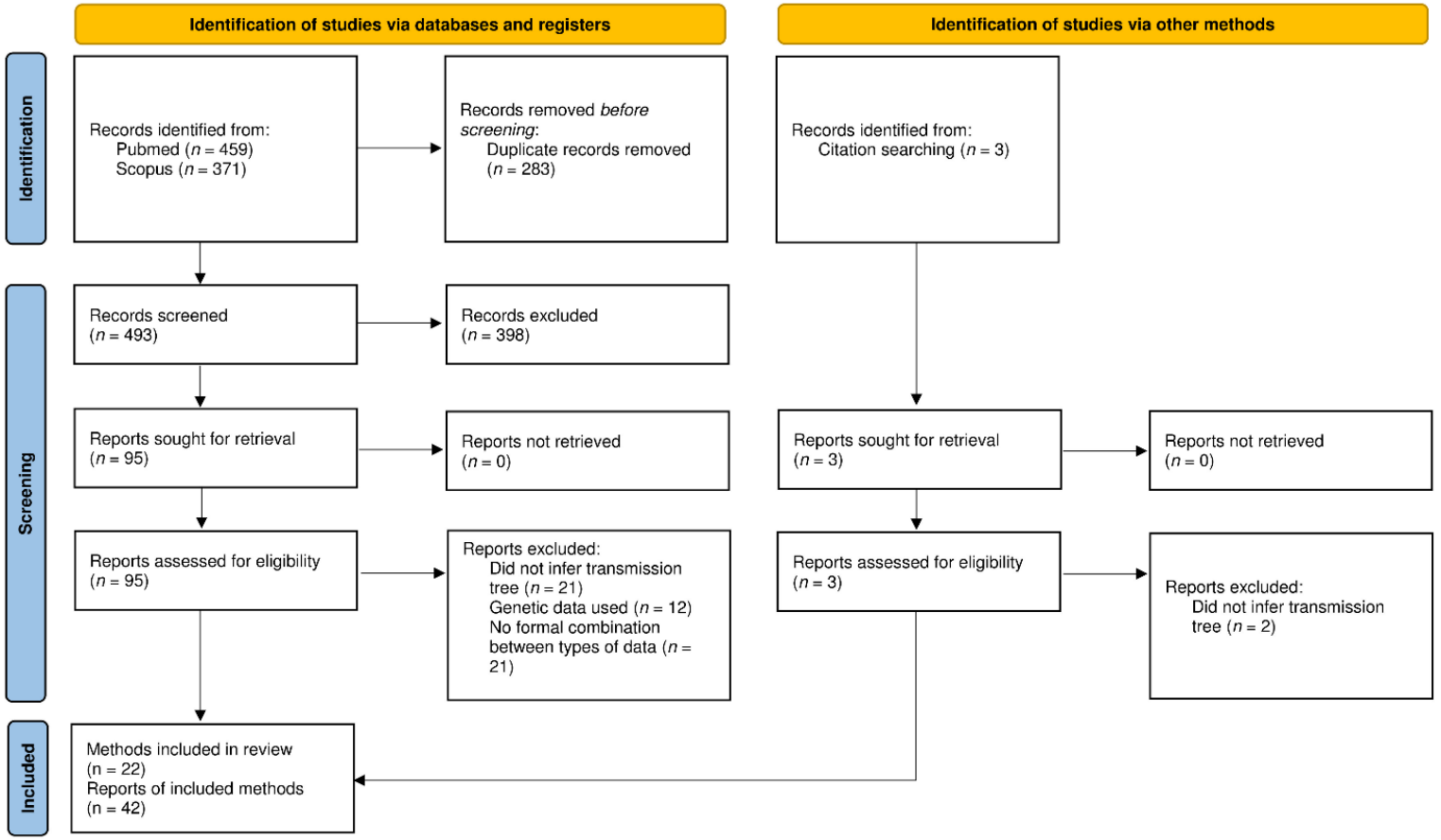
Annexe 1 : Résultats de la comparaison deux à deux par le BF des différents modèles d'évolution. *Chaque modèle est décrit par le modèle de substitution puis le modèle d'horloge et enfin le modèle de population. N est le nombre de particules, VM, la vraisemblance marginale et SD, la déviation standard.*

Modèle 1	Modèle 2	N	VM1	VM2	SD1	SD2	Log(BF)	Modèle favorisé (choisi)
JC Stricte Constant	HKY Stricte Constant	10	-1995,24	-1933,18	6,84	6,96	-62,06	**2**
HKY Stricte Constant	GTR Stricte Constant	10	-	-	-	-	-	**(2)**
HKY Stricte Constant	HKY Exponentielle Constant	1	-1894,05	-39504,85	21,57	6,29	37610,8	**1**
HKY Stricte Constant	HKY Lognormale Constant	1	-1894,05	-2203,36	21,57	22,97	309,31	**1**
HKY Stricte Constant	HKY Stricte Exponentiel	10	-1933,18	-2004,58	6,96	7,4	71,4	**1**
HKY Stricte Constant	HKY Stricte Skyline	10	-1933,18	-1958,8	6,96	7,03	25,62	**1**

Annexe 2 : Arbre MCC avec la probabilité postérieure de chaque nœud interne (couleur).



Annexe 3 : Diagramme PRISMA du processus de sélection des articles



Annexe 4 : Articles exclus de la revue systématique

	Title	Reference	Reasons for exclusion
1	Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences	Scaduto *et al.*, 2010 (184)	Only uses genetic data to infer transmission
2	High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster	Schurch *et al.*, 2010 (185)	No formal combination of genetic and epidemiological data
3	Phylodynamic analysis of a viral infection network	Schiino, 2012 (186)	Review on transmission clusters
4	Combining epidemiological and genetic networks signifies the importance of early treatment in HIV-1 transmission	Zarrabi *et al.*, 2012 (187)	Graphical representation of contact info with a SNP cut-off
5	Detectable signals of episodic risk effects on acute HIV transmission systems using genetic data	Alam *et al.*, 2013 (188)	Only uses a transmission model and not genetic data
6	Inferring the inter-host transmission of influenza A virus using patterns of intra-host genetic variation	Stack *et al.*, 2013 (189)	Cannot use consensus sequence
7	Inferring the source of transmission with phylogenetic data	Volz *et al.*, 2013 (51)	Uses simulated data
8	Bayesian Inference of Sampled Ancestor Trees for Epidemiology and Fossil Calibration	Gavryushkina *et al.*, 2014 (190)	Reconstructs a phylogenetic tree with sampled ancestors and not a transmission tree
9	Marked microevolution of a unique Mycobacterium tuberculosis strain in 17 years of ongoing transmission in a high risk population	Mehaffy *et al.*, 2014 (191)	No formal combination of genetic and epidemiological data
10	Two-phase importance sampling for inference about transmission trees	Numminen *et al.*, 2014 (192)	Each data cluster was analyzed independently from others

11	Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data	Worby *et al.*, 2014 (58)	Uses only (simulated) genetic data
12	The application of genomics to tracing bacterial pathogen transmission	Croucher *et* Didelot, 2015 (193)	General overview of use of WGS
13	Phylogenetic visualization of the spread of H7 influenza A viruses	Janies *et al.*, 2015 (194)	Phylogeographic reconstruction represented by a transmission network
14	The impact of within-herd genetic variation upon inferred transmission trees for foot-and-mouth disease virus	Valdazo-Gonzalez *et al.*, 2015 (195)	Only uses genetic data to infer transmission
15	Ebola Virus Epidemiology and Evolution in Nigeria	Folarin *et al.*, 2016 (196)	No formal combination of genetic and epidemiological data
16	Using genomics data to reconstruct transmission trees during disease outbreaks	Hall *et al.*, 2016 (197)	Review
17	Tracing Origins of the Salmonella Bareilly Strain Causing a Food-borne Outbreak in the United States	Hoffman *et al.*, 2016 (198)	Geographic mapping of a phylogeny
18	Molecular Infectious Disease Epidemiology: Survival Analysis and Algorithms Linking Phylogenies to Transmission Trees	Kenah *et al.*, 2016 (199)	Enumerates transmission trees compatible with phylogeny and epidemiological data, doesn't select
19	Infectious Disease Dynamics in Heterogeneous Landscapes	Parratt *et al.*, 2016 (200)	Review
20	Network inference from multimodal data: A review of approaches from infectious disease	Ray *et al.*, 2016 (201)	Review

	transmission		
21	Ecological dynamics of influenza A viruses: cross-species transmission and global migration	Ren *et al.*, 2016 (202)	Phylogeographic reconstruction represented by a transmission network
22	Genomic Analysis of Viral Outbreaks	Wohl *et al.*, 2016 (203)	Review
23	Epidemiological and Evolutionary Inference of the Transmission Network of the 2014 Highly Pathogenic Avian Influenza H5N2 Outbreak in British Columbia, Canada	Xu *et al.*, 2016 (204)	Compares two network one from genetic data and the other from epidemiological data
24	Transmission patterns and evolution of respiratory syncytial virus in a community outbreak identified by genomic analysis	Agoti *et al.*, 2017 (205)	Only uses genetic data to infer transmission
25	Emerging Concepts of Data Integration in Pathogen Phylodynamics	Baele *et al.*, 2017 (206)	Review
26	Inference of genetic relatedness between viral quasispecies from sequencing data	Glebova *et al.*, 2017 (207)	Only uses genetic data to infer transmission and cannot use consensus sequence
27	Integrated genomic and interfacility patient-transfer data reveal the transmission pathways of multidrug-resistant Klebsiella pneumoniae in a regional outbreak	Snitkin *et al.*, 2017 (208)	Phylogeographic reconstruction represented by a transmission network
28	Shared Genomic Variants: Identification of Transmission Routes Using Pathogen Deep-Sequence Data	Worby *et al.*, 2017 (209)	Only uses deep sequencing data data to infer transmission
29	outbreaker2: a modular platform for outbreak reconstruction	Campbell *et al.*, 2018 (210)	Implements Transphylo method into outbreaker2 using simulated data (phybreak package)

30	Genomic epidemiology for microbial evolutionary studies and the use of Oxford Nanopore sequencing technology	Choi *et al.*, 2018 (211)	Review
31	Bayesian reconstruction of transmission within outbreaks using genomic variants	De Maio *et al.*, 2018 (157)	Only at a single locus resolution (deep sequencing data), cannot use consensus sequence
32	Tracking a serial killer: Integrating phylogenetic relationships, epidemiology, and geography for two invasive meningococcal disease outbreaks	Ezeoke *et al.*, 2018 (212)	Phylogeographic reconstruction represented by a transmission network
33	Multi-step genomic dissection of a suspected intra-hospital Helicobacter cinaedi outbreak	Gotoh *et al.*, 2018 (213)	No formal combination of genetic and epidemiological data
34	Estimating Transmission from Genetic and Epidemiological Data: A Metric to Compare Transmission Trees	Kendall *et al.*, 2018 (171)	Transmission tree comparaison
35	Phylogenetic patterns recover known HIV epidemiological relationships and reveal common transmission of multiple variants	Leitner *et* Romero-Severson, 2018 (214)	Uses topology of phylogenetic trees to determine whether direct transmission occured
36	The relationship between transmission time and clustering methods in Mycobacterium tuberculosis epidemiology	Meehan *et al.*, 2018 (215)	Estimates transmission events with the age difference between MRCA node and youngest node in a cluster
37	When are pathogen genome sequences informative of transmission events?	Campbell *et al.*, 2018 (55)	Uses simulated data to compare two methods
38	Multihospital Outbreak of a Middle East Respiratory	Payne *et al.*, 2018 (216)	Only uses epidemiological data to

	Syndrome Coronavirus Deletion Variant, Jordan: A Molecular, Serologic, and Epidemiologic Investigation		infer transmission
39	Nosocomial transmission of influenza: A retrospective cross-sectional study using next generation sequencing at a hospital in England (2012-2014)	Blackburn *et al.*, 2019 (217)	No formal combination of genetic and epidemiological data
40	Inferring epidemiological links from deep sequencing data: a statistical learning approach for human, animal and plant diseases	Alamil *et al.*, 2019 (158)	Deep sequencing data
41	Introduction of Ebola virus into a remote border district of Sierra Leone, 2014: use of field epidemiology and RNA sequencing to describe chains of transmission	DeSilva *et al.*, 2019 (218)	Only uses epidemiological data to infer transmission
42	Reconstructing foot-and-mouth disease outbreaks: a methods comparison of transmission network models	Firestone *et al.*, 2019 (172)	Transmission tree comparaison on simulated data
43	Whole genome sequencing for improved understanding of Mycobacterium tuberculosis transmission in a remote circumpolar region	Guthrie *et al.*, 2019 (219)	No formal combination of genetic and epidemiological data
44	Transmission Trees on a Known Pathogen Phylogeny: Enumeration and Sampling	Hall *et* Colijn, 2019 (220)	Aim is not to reconstruct a transmission tree
45	Improved characterisation of MRSA transmission using within-host bacterial sequence	Hall *et al.*, 2019 (221)	Only uses genetic data to infer transmission

	diversity		
46	Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis	Ratmann *et al.*, 2019 (222)	Deep sequencing data
47	SharpTNI: Counting and Sampling Parsimonious Transmission Networks under a Weak Bottleneck	Sashittal *et al.*, 2019 (223)	Similar to Hall *et* Colijn, 2019
48	Tracing local and regional clusters of carbapenemase-producing Klebsiella pneumoniae ST512 with whole genome sequencing, Finland, 2013 to 2018	van Beek *et al.*, 2019 (224)	No formal combination of genetic and epidemiological data + uses allele differences
49	Estimating Epidemic Incidence and Prevalence from Genomic Data	Vaughan *et al.*, 2019 (225)	No difference between phylogeny and transmission tree
50	Phylogenetic and Demographic Characterization of Directed HIV-1 Transmission Using Deep Sequences from High-Risk and General Population Cohorts/Groups in Uganda	Bbosa *et al.*, 2020 (226)	Deep sequencing data
51	StrainHub: a phylogenetic tool to construct pathogen transmission networks	de Bernardi Schneider *et al.*, 2020 (227)	Phylogeographic reconstruction represented by a transmission network
52	TNet: Phylogeny-Based Inference of Disease Transmission Networks Using Within-Host Strain Diversity	Dhar *et al.*, 2020 (228)	Deep sequencing data
53	Modeling Missing Cases and Transmission Links in Networks of Extensively Drug-Resistant	Nelson *et al.*, 2020 (229) a	Compares an empirical network (> or = to 5 SNPs difference) with network reconstructed

	Tuberculosis in KwaZulu-Natal, South Africa		using ergm with different missing cases scenarios
54	Social Mixing and Clinical Features Linked With Transmission in a Network of Extensively Drug-resistant Tuberculosis Cases in KwaZulu-Natal, South Africa	Nelson *et al.*, 2020 (230) b	Graphical representation of contact info with a SNP cut-off
55	Nosocomial outbreak of COVID-19 pneumonia in Wuhan, China	Wang *et al.*, 2020 (231)	Only uses epidemiological data to infer transmission
56	The emergence of SARS-CoV-2 in Europe and North America	Worobey *et al.*, 2020 (232)	Compares simulated data to obtained sequence data then took a phylogeographic approach

Annexe 5 : Séquences de pathogènes étudiées par des méthodes de la NPF. *D'abord, nous avons enregistré la période d'étude (J signifie jours, M mois et A années) puis le nombre de séquences et enfin la taille du génome étudié (en nombre de nucléotides N ou SNP). TC signifie temps de calcul et NT, non trouvé.*

Méthodes	**Pathogène**	**Période d'étude**	**Nombre de séquences**	**Taille du génome**	**Références**	**TC**	**Unité épidémiologique**
Aldrin 2011	*Salmon isavirus*	6 A	61	891 N	(120)	NT	Elevage
seqTrack	H1N1	105 J	433	3,025 N	(133)	NT	Individu
	H3N8	56 J	48	903 N	(135)		Individu
	Mycobacterium tuberculosis	15 A	1687	~4.1 x 10^6 N	(136)		Individu
	Klebsiella pneumoniae	3 A	71	~5.5 x 10^6 N	(137)		Individu
Ypma 2012	H7N7	51 J	185	5,354 N	(125)	NT	Elevage
Outbreaker	SARS-CoV1	48 J	13	29,731 N	(126)	NT	Individu
	Acinetobacter baumanii	9 M	19	1207 SNP	(130)		Individu
		3 A	42	4,461,520 N	(129)		Individu
	Klebsiella pneumoniae	305 J	34	51 SNP	(128)		Individu
	BVDV	~2 A	12	381 N	(127)		
Worby 2014	*MRSA*	300 J	35	~2.8 x 10^6 N	(131)	NT	Individu
Famulare 2015	Ebola	25 J	78	~19,000 N	(138)	NT	Individu
	H1N1	105 J	433	3025 N			
	poliovirus	~6 A	358	~2600 N			
Bitrugs	*MRSA*	300 J	35	~2.8 x 10^6 N	(132)	NT	Individu
		~3 A	40	345 SNP	(156)		
			16	209 SNP	(156)		
Outbreaker2	SARS-CoV1	48 J	13	29,731 N	(121)		Individu
	Données simulées		20		(210)	4,5 sec/1000 iterations	/

Annexe 6 : Séquences de pathogènes étudiées par des méthodes de la SeqPF. *D'abord, nous avons enregistré la période d'étude (J signifie jours, M mois et A années) puis le nombre de séquences et enfin la taille du génome étudié (en nombre de nucléotides N ou SNP). TC signifie temps de calcul et NT, non trouvé.*

Méthodes	**Pathogène**	**Période d'étude**	**Nombre de séquences**	**Taille du génome**	**Références**	**TC**	**Unité épidémiologique**
Cottam 2008	FMDV	96 J	22	8,176 N	(140)	NT	Elevage
Didelot 2014	*Mycobacterium tuberculosis*	~3 A	33	20 SNP	(56)	100 000 iterations en 5 min	Individu
Eldoholm 2016	*Mycobacterium tuberculosis*	13 A	252	509 SNP	(139)	NT	Individu
Transphylo	SARS-CoV2	~60 J	208	29,718 N	(147)		Individu
	Myxovirus parotidis	~1 A	24	51 SNP	(148)		Individu
	Mycobacterium tuberculosis	13 A	86	85 SNP	(141)		Individu
		7 A	21	134 SNP	(142)		
		19 A	1857	~4.1 x 10^6 N	(144)		Individu
		3 A	117		(143)		Individu
		~15 A	329	269 SNP	(149)		Individu
	Klebsiella pneumoniae	~3 A	73	~5.5 x 10^6 N	(145)		Individu
		14 M	82		(146)		
	Données simulées		20		(210)	4,6 s/1000 iterations	/
TiTUS	HIV	18 A	212	1620 N	(122)	NT	Individu

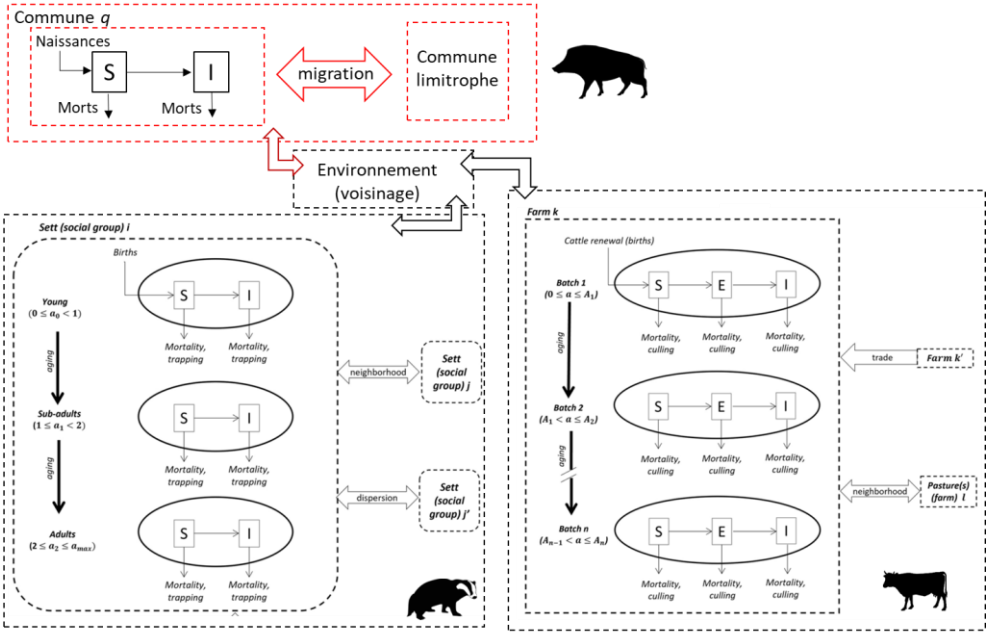
Annexe 7 : Séquences de pathogènes étudiées par des méthodes de la SimPF. *D'abord, nous avons enregistré la période d'étude (J signifie jours, M mois et A années) puis le nombre de séquences et enfin la taille du génome étudié (en nombre de nucléotides N ou SNP). TC signifie temps de calcul et NT, non trouvé.*

Méthodes	**Pathogène**	**Période d'étude**	**Nombre de séquences**	**Taille du génome**	**Références**	**TC**	**Unité épidémiologique**
Ypma 2013	FMDV	64 J	12	8176 N	(41)	NT	Elevage
Beastlier	H7N7	51 J	185	5354 N	(124)	NT	Elevage
	H5N8		37	83 SNP	(154)		Elevage
SCOTTI	FMDV	57 J	11	8176 N	(155)		Elevage
	HIV				(233)		Individu
	Klebsiella pneumoniae	305 J	34	51 SNP	(155)		Individu
	Données simulées		100		(155)	1-2 h	/
Phybreak	FMDV	57 J	11	30 SNP	(114)		Elevage
		65 J	15	58 SNP			
	H7N7	71 J	241	293 N		7 h/ 25000 iterations	Elevage
	Mycobacterium tuberculosis	~3 A	33	21 SNP		30 min/ 25000 iterations	Individu
	MRSA	204 J	36	27 SNP		30 min/ 25000 iterations	Individu
	Données simulées		50	10,000 N		1 min/1000 iterations	/
Morelli 2012	FMDV	57 J	10	8176 N	(150)	NT	Elevage
		65 J	15				
Mollentze 2014	RABV	464 J	176	760 N	(151)	NT	Individu
Lau 2015	FMDV	64 J	12	8176 N	(40)		Elevage
	Données simulées	~3 M	104	7667 N	(152)		
			150	8000 N	(40)	400 000 iterations 17,7 h	/
BORIS	FMDV	2.5 M	104	7667 N	(39)	NT	Elevage
Montazeri 2020	HIV	~15 A	19	~3000 N	(153)		Individu
	Ebola	~5 M	78	~19,000 N			
	Données simulées		50	3000		3 h	

Annexe 8 : Agents pathogènes étudiés dans les articles sélectionnés et l'estimation de leur taux de mutation.

Pathogène	**Taux de mutation**
FMDV	$2{,}1 \times 10^{-5}$/site/jour (140)
H7N7	Gènes HA et NA ~1×10^{-2} ET gène PB2 0.54×10^{-2}/site/an (234)
H1N1	Gènes HA et NA $3{,}6 \times 10^{-3}$/site/an (235)
H5N8	$6{,}68 \times 10^{-3}$/site/an (236)
H3N8	Gène HA $1{,}74 \times 10^{-3}$/site/an (237)
RABV	$2{,}9 \times 10^{-4}$/site/an (238)
SARS-CoV1	$5{,}7 \times 10^{-6}$/site/jour (239)
SARS-CoV2	$9{,}90 \times 10^{-4}$/site/an (240)
HIV	gène *env* $7{,}4 \times 10^{-3}$/site/an (241)
Ebola virus	~2×10^{-3}/site/an (242)
poliovirus	$1{,}1 \times 10^{-2}$ /site/an (243)
BVDV	$3{,}9 \times 10^{-3}$ /site/an (127)
Salmon isavirus	Gène HE [$6{,}1 \times 10^{-6}$; $1{,}13 \times 10^{-3}$]/site/an (244,245)
Myxovirus parotidis	$2{,}24 \times 10^{-3}$/site/an (148)
Mycobacterium tuberculosis	$1{,}15 \times 10^{-7}$/site/an (56)
Acinetobacter baumanii	10^{-6} mutation par division cellulaire (246)
Klebsiella pneumoniae	$3{,}65 \times 10^{-6}$/site/an (128)
Staphylococcus areus	3×10^{-6}/site/an (247)

Annexe 9 : Représentation schématique du modèle de simulation d'après Bouchez-Zacria *et al.*, 2023. *Ce qui a été ajouté dans ce travail de thèse est représenté en rouge. L'unité épidémiologique considérée pour les sangliers n'a pas de valeur écologique (commune et non horde) du fait du manque de données disponibles pour cette population. La force d'infection exercée sur un individu de chaque groupe est la somme de la force exercée au sein du groupe, de la force exercée par les groupes voisins (fraction ε de celle exercée au sein des groupes voisins ou nulle pour les sangliers) et par l'environnement.*



D'après Bouchez-Zacria *et al.*, 2023

$\lambda_{i,sp}(t) = \lambda_{i,sp}^{intra}(t) + \lambda_{i,sp}^{voisins}(t) + \lambda_{i,sp}^{env}(t)$		
$\lambda_{i,sp}^{intra}(t) = \beta_{sp}^{intra} * \frac{I}{N}$		
$\lambda_{i,sp}^{voisins}(\mathrm{t}) = \sum_{j \in V(sp,i)} \varepsilon_{sp} * \lambda_{j,sp}^{intra}(t)$		$\lambda_{i,sp}^{voisins}(\mathrm{t}) = 0$
$\lambda_{i,sp}^{env}(\mathrm{t}) = \beta_{sp}^{env} * 1_{i,sp}^{env}(t)$	$1_{i,sp}^{env}(t) = 0$ *si aucun animal infectieux d'une autre espèce a partagé le même environnement il y a moins de 3 mois.* $1_{i,sp}^{env}(t) = 1$ *sinon.*	

Annexe 10 : Nom, définition, classe et valeur des paramètres MCMC dans *outbreaker2* et *TransPhylo*.

Méthode	Nom	Définition	Classe	Valeur
outbreaker2	n_iter	Nombre d'itérations	Integer	n_iter = 1e5 (or 2e5)
	sample_every	Fréquence d'échantillonnage	Integer	sample_every = 50
	init_tree	Arbre initial	Character	init_tree = SeqTrack (uses seqTrack algorithm to generate the initial tree) (default)
TransPhylo	mcmcIterations	Nombre d'itérations	Numeric	mcmcIterations = 5e5
	thinning	Fréquence d'échantillonnage	Numeric	thinning = 50

Annexe 11 : Nom, définition, classe et valeur des données implémentées dans *outbreaker2* et *TransPhylo*.

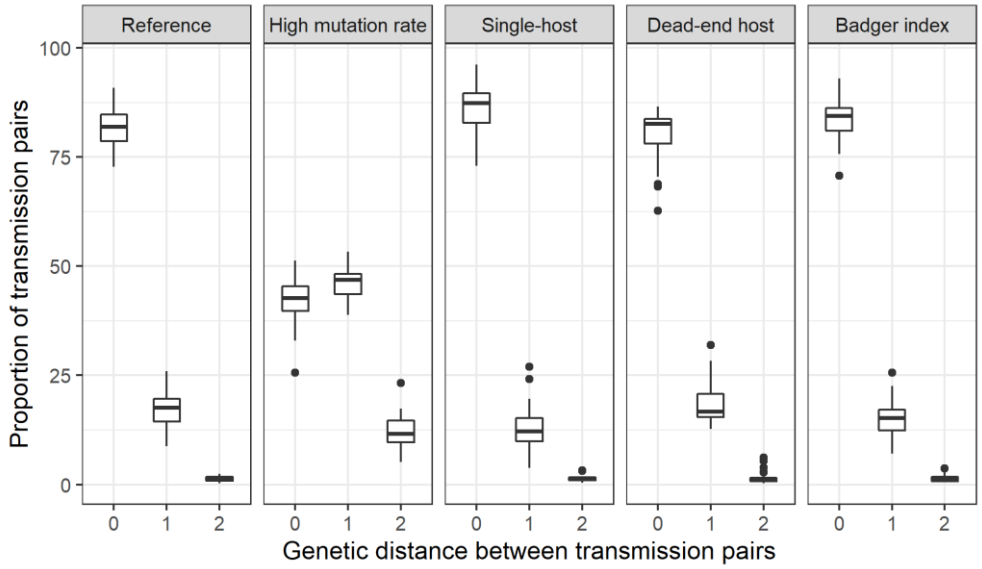
Méthode	Nom	Définition	Classe	Valeur
outbreaker2	dates	Dates d'échantillonnage	Vector of integers (numbers or dates)	Selon la simulation (en mois)
	dna	Séquences ADN	DNAbin	Selon la simulation
	ctd	Données de contact	Matrix/dataframe of two columns	NULL
TransPhylo	p	Phylogénie datée	ptree	Selon la simulation
	dateT	Date de fin d'observation	Numeric	dateT = 2020.1

Annexe 12 : Nom, définition, classe, type (fixe or estimé) et valeur des paramètres du modèle dans *outbreaker2* et *TransPhylo*.

Méthode	Nom	Définition	Prior OR classe	Type	Valeur
outbreaker2	w_dens	Distribution du temps de génération	Vector of numeric values	Fixe	Selon la simulation
	f_dens	Distribution de l'intervalle infection-échantillonnage	Vector of numeric values	Fixe	Selon la simulation
	t_inf	Date d'infection	Vector of integers	Estimé	init_t_inf = NULL, (défaut) move_t_inf = TRUE (défaut) prop_t_inf_move = 0.2 (défaut)
	alpha	Ancêtres	Vector of integers	Estimé	init_alpha = NULL (défaut) move_alpha = TRUE (défaut) prop_alpha_move = 0.25 (défaut)
	kappa	Nombre de générations séparant deux cas	Vector of integers	Estimé	init_kappa = 1 (défaut)
	max_kappa	Nombre maximum de générations entre deux cas	Integer	Fixe	max_kappa = 5 (défaut) move_kappa = TRUE (défaut)
	min_date	Date minimale d'infection	Double	Fixe	Selon la simulation
	mu	Taux de mutation (par site par génération d'infection)	Exp(prior_mu)	Estimé	prior_mu = 1 (défaut) init_mu = 1 e-04 (défaut) move_mu = TRUE (défaut) sd_mu = 1 e-04 (défaut)
	pi	Probabilité d'échantillonnage	Beta(prior_pi1, prior_pi2)	Estimé	prior_pi1 = 1 prior_pi2 = 1 (défaut) init_pi = 0.9 (défaut)

					move_pi = TRUE (défaut) sd_pi = 0.1 (défaut)
	eps	Taux de reportage	Beta(prior_eps1, prior_eps2)	Estimé	Aucune données de contact
	lambda	Taux de contact non-infectieux	Beta(prior_lambda1, prior_lambda2)	Estimé	Aucune données de contact
TransPhylo	w	Temps de génération	Gamma(w.shape, w.scale) OR Gamma(w.mean, w.std)	Fixe	Selon la simulation
	ws	Temps d'échantillonnage	Gamma(ws.shape, ws.scale) OR Gamma(ws.mean, ws.std)	Fixe	Selon la simulation
	Neg	Temps de coalescence moyen	Exp(1)	Estimé	Neg = 100/365 (défaut) updateNeg = TRUE (défaut)
	Off	Distribution des cas secondaires	NegBinom(Off.r, Off.p) Off.r → Exp(1) Off.p → Unif([0;1])	Off.r estimé Off.p fixe (défaut)	startOff.r = 1 updateOff.r = TRUE (défaut) startOff.p = 0.5 updateOff.p = FALSE (défaut)
	pi	Probabilité d'échantillonnage	Unif([0;1])	Estimé	startPi = 0.9 updatePi = TRUE (défaut)

Annexe 13 : Proportion de paires de transmission séparées par 0, 1 et 2 SNP selon le scénario de transmission. *La référence correspond au système multi-hôtes où les bovins sont des cas index et les sangliers contribuent à la transmission. High mutation rate correspond au même scénario simulé avec un taux de mutation plus élevé. Single-host correspond au système avec uniquement des bovins. Dead-end host correspond au scénario où les sangliers ne contribuent pas et badger index, au scénario de référence avec des blaireaux en cas index.*



Annexe 14 : Proportion d'évènements de transmission présents dans les arbres de référence en fonction de la méthode et du schéma d'échantillonnage. *Les () représentent les plages de valeurs. T signifie "biais temporel", S_B "biais blaireau" et S_W "biais sanglier". T+S_B (T+S_W) combine le biais temporel avec le biais blaireau (sanglier).*

Schéma d'échantillonnage	*outbreaker2*	seqTrack	*TransPhylo*
Référence	8,0 (2,2-11,3)	3,4 (1,3-12,1)	8,9 (6,0-16,8)
T	8,0 (2,8-13,3)	3,2 (0,9-12,3)	8,4 (5,8-15,0)
S_B	6,9 (3,2-16,0)	3,0 (1,3-10,7)	10,2 (6,5-15,7)
T+S_B	7,0 (3,2-16,0)	3,0 (1,3-10,7)	10,7 (4,9-16,7)
S_W	8,2 (4,3-12,6)	5,0 (1,1-14,4)	10,2 (7,4-13,0)
T+S_W	8,0 (3,8-13,4)	4,6 (0,8-18,3)	9,4 (6,6-16,3)

Annexe 15 : Nombre maximum d'évènements de transmission dus à un seul super-spreader dans un arbre reconstruit par seqTrack et l'espèce-hôte de ce super-spreader, selon le schéma d'échantillonnage. *Les () représentent les plages de valeurs. T signifie "biais temporel", S_B "biais blaireau" et S_W "biais sanglier". T+S_B (T+S_W) combine le biais temporel avec le biais blaireau (sanglier).*

Schéma d'échantillonnage	Nombre d'évènements de transmission reconstruits	Nombre maximum d'évènements de transmission dus à un seul super-spreader	Nombre de bovins (%)	Nombre de blaireaux (%)	Nombre de sangliers (%)
Référence	244 (116-449)	108 (52-275)	12 (57 %)	6 (29 %)	3 (14 %)
T	240 (114-443)	108 (52-273)	15 (71 %)	3 (14 %)	3 (14 %)
S_B	200 (75-406)	90 (40-253)	16 (76 %)	0	5 (24 %)
T+S_B	200 (75-406)	90 (40-252)	17 (81 %)	0	4 (19 %)
S_W	210 (98-404)	93 (50-257)	14 (67 %)	7 (33 %)	0
T+S_W	204 (96-398)	93 (45-256)	18 (86 %)	3 (14 %)	0

Annexe 16 : Nombre d'hôtes infectés présents dans l'épidémie reconstructible et dans les arbres reconstruits selon la méthode et le schéma d'échantillonnage. *T signifie "biais temporel", S_B "biais blaireau" et S_W "biais sanglier". T+S_B (T+S_W) combine le biais temporel avec le biais blaireau (sanglier).*

Schéma d'échantillonnage	Taille de l'épidémie reconstructible	*outbreaker2*		*TransPhylo*	
		Nombre d'hôtes échantillonnés divisé par π	Nombre d'hôtes dans l'arbre reconstruit	Nombre d'hôtes échantillonnés divisé par π	Nombre d'hôtes dans l'arbre reconstruit
Référence	245 (118-458)	250 (125-462)	245 (118-458)	407 (190-15666)	393 (180-1219)
T	242 (116-452)	246 (123-456)	241 (116-452)	368 (166-16490)	380 (180-1222)
S_B	212 (87-428)	204 (84-418)	201 (77-415)	257 (118-8669)	312 (109-945)
T+S_B	212 (87-428)	204 (84-418)	201 (77-415)	255 (120-8438)	313 (112-947)
S_W	222 (103-432)	217 (107-417)	211 (100-413)	269 (127-4965)	307 (129-1063)
T+S_W	219 (101-426)	210 (105-411)	205 (98-407)	270 (134-5885)	310 (131-1001)

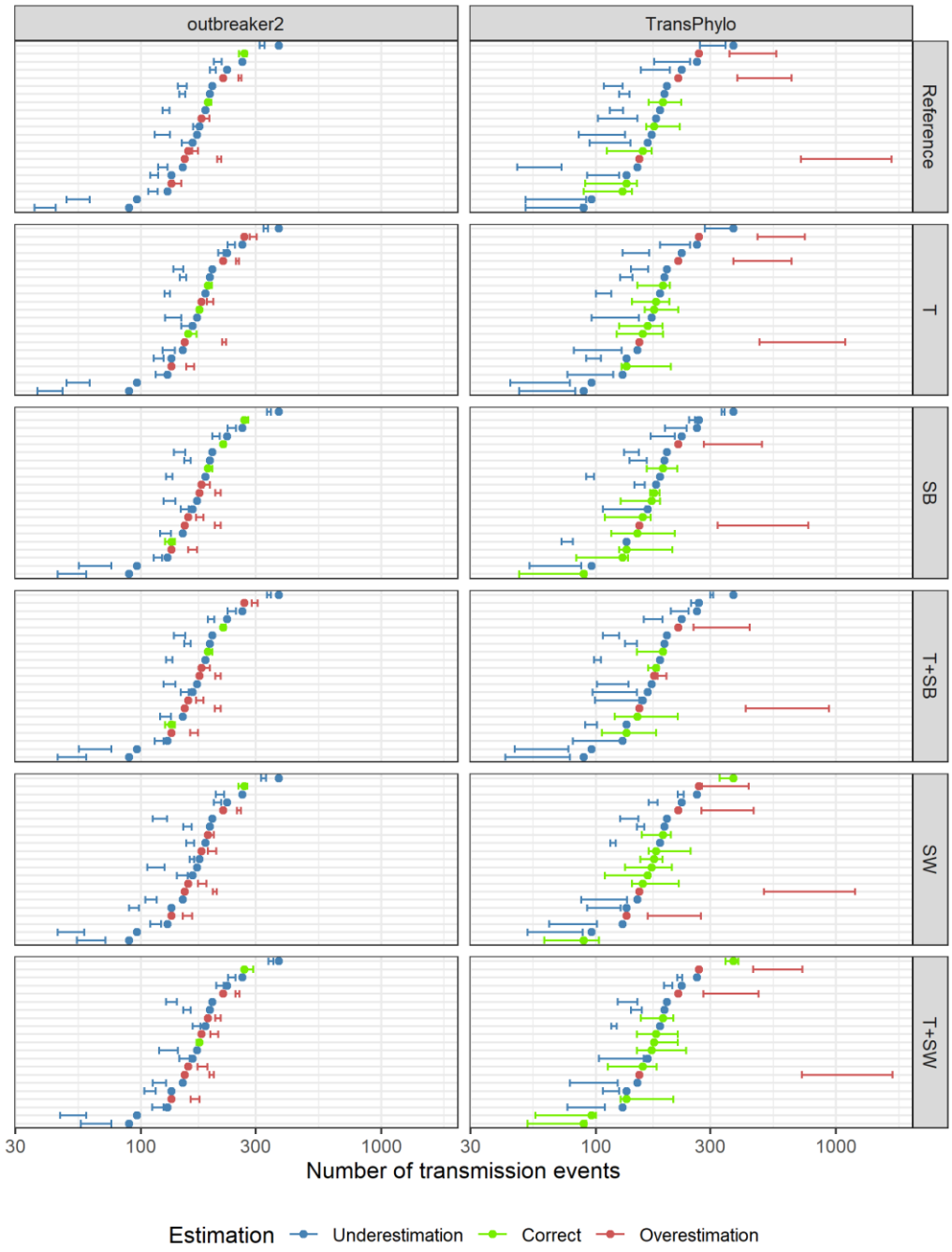
Annexe 17 : Estimation de la taille de l'épidémie testée par un GLMM Négatif Binomial avec la méthode et l'interaction entre la méthode et le schéma d'échantillonnage en effets fixes. *IRR signifie ratio des taux d'incidence. Les nombres en gras signalent des IRRs significativement différent de 1 (p<0,05). La taille de l'épidémie reconstructible était l'offset et le schema de référence, la référence. T signifie "biais temporel", S_B "biais blaireau" et S_W "biais sanglier". T+S_B (T+S_W) combine le biais temporel avec le biais blaireau (sanglier).*

Effets fixes	*outbreaker2*		*TransPhylo*	
	IRR	p	IRR	p
Méthode	1,01	0,85	**1,83**	**<0,001**
Méthode :T	0,99	0,93	0,98	0,75
Méthode :S_B	0,94	0,32	**0,77**	**<0,001**
Méthode :T+S_B	0,94	0,31	**0,76**	**<0,001**
Méthode :S_W	0,94	0,31	**0,85**	**0,008**
Méthode :T+S_W	0,93	0,27	**0,83**	**0,004**

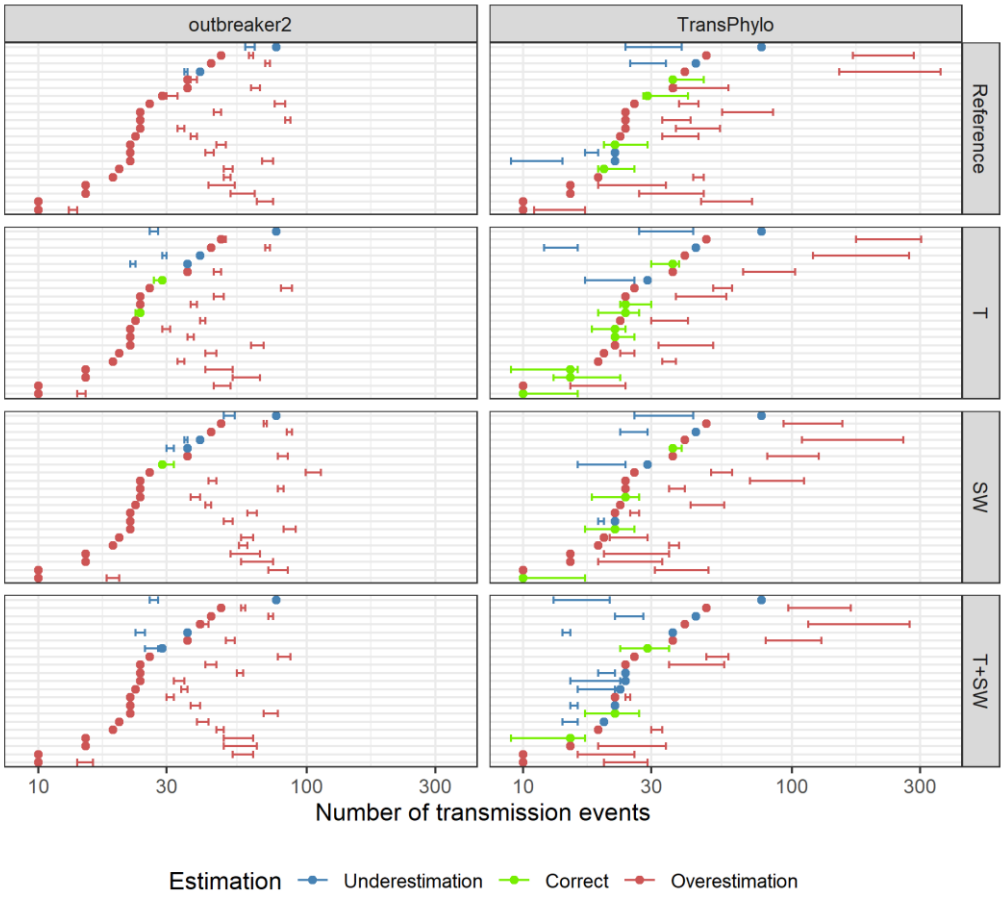
Annexe 18 : Nombre médian d'évènements de transmission dus à chaque espèce-hôte dans l'épidémie reconstructible et dans les arbres reconstruits, selon la méthode et le schéma d'échantillonnage. *Les () représentent les plages de valeurs. T signifie "biais temporel", S_B "biais blaireau" et S_W "biais sanglier". T+S_B (T+S_W) combine le biais temporel avec le biais blaireau (sanglier).*

Schéma d'échantillonnage	Espèce-hôte	Nombre d'évènements de transmission dans l'épidémie reconstructible	*outbreaker2*		*TransPhylo*	
			Nombre d'évènements de transmission entre hôtes échantillonnés divisé par π	Nombre d'évènements de transmission dus aux hôtes échantillonnés	Nombre d'évènements de transmission entre hôtes échantillonnés divisé par π	Nombre d'évènements de transmission dus aux hôtes échantillonnés
Référence	Bovin	175 (89-374)	151 (38-314)	148 (36-311)	124 (58-1102)	87 (18-279)
	Blaireau	24 (10-77)	50 (13-84)	49 (13-83)	37 (12-232)	22 (2-45)
	Sanglier	40 (4-123)	37 (7-86)	37 (7-85)	25 (0-464)	18 (1-48)
T	Bovin	175 (84-369)	156 (40-326)	154 (37-323)	151 (58-738)	98 (16-283)
	Blaireau	23 (9-76)	40 (14-81)	40 (14-80)	26 (12-229)	21 (3-56)
	Sanglier	40 (4-123)	30 (5-88)	30 (5-87)	30 (7-738)	22 (3-62)
S_B	Bovin	158 (50-354)	160 (49-337)	157 (45-335)	149 (65-492)	135 (31-348)
	Sanglier	40 (4-109)	38 (9-90)	38 (9-89)	29 (5-151)	27 (7-48)
T+S_B	Bovin	158 (50-354)	162 (49-337)	160 (45-335)	136 (56-626)	119 (35-326)
	Sanglier	40 (4-109)	37 (9-90)	36 (9-89)	28 (11-111)	25 (4-60)
S_W	Bovin	160 (87-351)	157 (48-319)	154 (45-316)	165 (65-765)	98 (39-345)
	Blaireau	21 (8-74)	57 (18-102)	56 (18-99)	33 (13-165)	22 (8-57)
T+S_W	Bovin	159 (82-346)	167 (49-342)	164 (46-339)	162 (68-1088)	114 (48-361)
	Blaireau	21 (8-73)	43 (14-80)	42 (14-78)	23 (12-173)	17 (7-53)

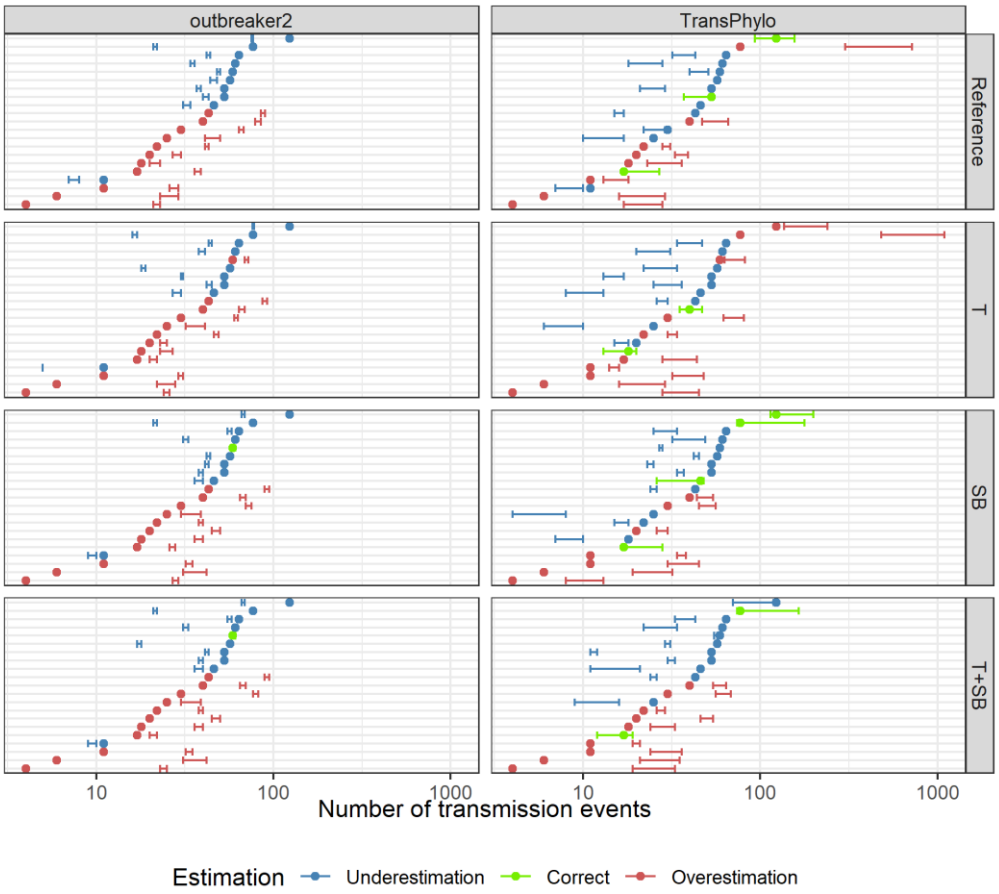
Annexe 19 : Intervalle de crédibilité du nombre d'évènements de transmission dus aux bovins estimé par *outbreaker2* et *TransPhylo* comparé (couleur) au nombre dans l'épidémie simulée (point), selon le schéma d'échantillonnage. *T signifie "biais temporel", S_B "biais blaireau" et S_W "biais sanglier". T+S_B (T+S_W) combine le biais temporel avec le biais blaireau (sanglier).* Chaque ligne représente un arbre reconstruit.



Annexe 20 : Intervalle de crédibilité du nombre d'évènements de transmission dus aux blaireaux estimé par *outbreaker2* et *TransPhylo* comparé (couleur) au nombre dans l'épidémie simulée (point), selon le schéma d'échantillonnage. *T signifie "biais temporel", S_B "biais blaireau" et S_W "biais sanglier". T+S_B (T+S_W) combine le biais temporel avec le biais blaireau (sanglier).* Chaque ligne représente un arbre reconstruit.



Annexe 21 : Intervalle de crédibilité du nombre d'évènements de transmission dus aux sangliers estimé par *outbreaker2* et *TransPhylo* comparé (couleur) au nombre dans l'épidémie simulée (point), selon le schéma d'échantillonnage. *T signifie "biais temporel", S_B "biais blaireau" et S_W "biais sanglier". T+S_B (T+S_W) combine le biais temporel avec le biais blaireau (sanglier).* Chaque ligne représente un arbre reconstruit.



Annexe 22 : Nombre d'évènements de transmission dus à chaque espèce-hôte testé par un GLMM Négatif Binomial avec la méthode et l'interaction entre la méthode et le schéma d'échantillonnage en effets fixes. *IRR signifie ratio des taux d'incidence. Les nombres en gras signalent des IRRs significativement différent de 1 (p<0,05). Le nombre d'évènements dans l'épidémie reconstructible était l'offset et le schema de référence, la référence. T signifie "biais temporel", S_B "biais blaireau" et S_W "biais sanglier". T+S_B (T+S_W) combine le biais temporel avec le biais blaireau (sanglier).*

Effets fixes	*outbreaker2*		*TransPhylo*	
	IRR	p	IRR	p
1. Contribution des bovins				
Méthode	**0,85**	**0,02**	**0,55**	**<0,001**
Méthode :T	1,05	0,54	1,02	0,81
Méthode :S_B	**1,19**	**0,04**	**1,45**	**<0,001**
Méthode :T+S_B	**1,20**	**0,04**	**1,42**	**<0,001**
Méthode :S_W	1,08	0,40	**1,22**	**0,02**
Méthode :T+S_W	1,14	0,13	**1,30**	**0,003**
2. Contribution des blaireaux				
Méthode	**1,96**	**<0,001**	0,92	0,54
Méthode :T	0,82	0,06	0,87	0,23
Méthode :S_W	**1,27**	**0,03**	1,19	0,13
Méthode :T+S_W	1,04	0,71	0,95	0,69
3. Contribution des sangliers				
Méthode	1,24	0,26	**0,61**	**0,01**
Méthode :T	0,92	0,46	1,16	0,24
Méthode :S_B	1,11	0,35	**1,40**	**0,006**
Méthode :T+S_B	1,06	0,60	**1,42**	**0,004**

Annexe 23 : Comparaison entre les arbres de référence ayant convergé dans BEAST2 et *TransPhylo* et ceux qui n'ont pas convergé, selon le scénario de transmission.

Paramètre	Convergence	Résultats précédents	Taux de mutation élevé	Système à un seul type d'hôte	Cul-de-sac épidémio-logique	Index blaireau
Nombre d'arbres	Oui	21	21	26	17	18
	Non	9	9	4	13	12
Taille de l'épidémie	Oui	245 (118-458)	243 (114-392)	101.5 (47-495)	219 (98-468)	183.5 (120-369)
	Non	336 (114-376)	327 (251-458)	464.5 (238-486)	390 (130-490)	371 (113-489)
Proportion de séquences uniques	Oui	6.1 (1.8-8.7)	33.4 (24.9-48.1)	3.6 (0.9-9.4)	8.0 (3.8-13.7)	5.4 (1.5-4.4)
	Non	5.8 (3.4-8.1)	30.6 (27.6-40.5)	3.5 (1.7-3.8)	6.3 (5.2-16.7)	5.2 (1.5-9.8)
Ecart moyen après transmission	Oui	0.19 (0.11-0.28)	0.69 (0.54-0.98)	0.14 (0.04-0.28)	0.19 (0.14-0.38)	0.16 (0.07-0.33)
	Non	0.18 (0.09-0.25)	0.67 (0.57-0.84)	0.12 (0.07-0.17)	0.18 (0.14-0.43)	0.16 (0.11-0.26)

Annexe 24 : Pourcentage (%) d'évènements de transmission présents dans les arbres de référence selon la méthode et le scénario de transmission. *() correspond aux plages de valeurs. n est le nombre d'arbres par scénario pour lesquels la reconstruction a été possible.*

Méthode	Résultats précédents (n=21)	Taux de mutation élevé (n=21)	Système à un seul type d'hôte (n=26)	Cul-de-sac épidémio-logique (n=17)
outbreaker2	8,0 (2,2-11,3)	25,7 (15,9-33,3)	5,5 (0-11,4)	6,8 (2,9-14,6)
seqTrack	3,4 (1,3-12,1)	15,3 (8,2-33,3)	2 (0-11,4)	3,8 (1,1-9,7)
TransPhylo	8,9 (6,0-16,8)	21,2 (13,2-29,3)	6,5 (1,9-14,3)	8,2 (4,1-14,4)

Annexe 25 : Exactitude testée par un GLM Binomial avec la méthode en variable explicative, selon le scénario de transmission. *Les nombres en gras signalent des ORs significativement différent de 1 (p<0,05).*

Scénario de transmission	*outbreaker2*		seqTrack		*TransPhylo*	
	OR	p	OR	p	OR	p
Taux de mutation élevé (n=21)	-	-	**0,54**	**<0,001**	**0,82**	**<0,001**
Système à un seul type d'hôte (n=26)	-	-	**0,48**	**<0,001**	**1,23**	**0,03**

Annexe 26 : Nombre d'arbres reconstruits avec des super-spreaders et le nombre maximum d'évènements de transmission dus à un seul super-spreader, selon la méthode et le scénario de transmission. *() correspond aux plages de valeurs. n est le nombre d'arbres par scénario pour lesquels la reconstruction a été possible.*

Méthode	Scénario	Nombre d'évènements de transmission reconstruits	Nombre maximum d'évènements de transmission dus à un seul super-spreader	Nombre de bovins (%)	Nombre de blaireaux (%)	Nombre de sangliers (%)
outbreaker2	Taux de mutation élevé (n=21)	0	*NA*	*NA*	*NA*	*NA*
	Système à un seul type d'hôte (n=26)	5	39 (22-69)	26 (100 %)	*NA*	*NA*
	Cul-de-sac épidémio-logique (n=17)	0	*NA*	*NA*	*NA*	*NA*
seqTrack	Taux de mutation élevé (n=21)	20	35 (22-61)	12 (57 %)	4 (19 %)	4 (19 %)
	Système à un seul type d'hôte (n=26)	26	79 (33-371)	26 (100 %)	*NA*	*NA*
	Cul-de-sac épidémio-logique (n=17)	17	108 (42-237)	14 (67 %)	3 (14 %)	1 (5 %)
TransPhylo	Taux de mutation élevé (n=21)	0	*NA*	*NA*	*NA*	*NA*
	Système à un seul type d'hôte (n=26)	10	39.5 (19-85)	26 (100 %)	*NA*	*NA*
	Cul-de-sac	0	*NA*	*NA*	*NA*	*NA*

	épidémio-logique (n=17)					

Annexe 27 : Nombre médian d'hôtes infectés présents dans l'épidémie reconstructible et dans les arbres reconstruits, selon la méthode et le schéma d'échantillonnage. *() correspond aux plages de valeurs. n est le nombre d'arbres par scénario pour lesquels la reconstruction a été possible.*

Scénario	Taille de l'épidémie reconstructible	*outbreaker2*		*TransPhylo*	
		Nombre d'hôtes échantillonnés divisé par π	Nombre d'hôtes dans l'arbre reconstruit	Nombre d'hôtes échantillonnés divisé par π	Nombre d'hôtes dans l'arbre reconstruit
Taux de mutation élevé (n=21)	243 (114-392)	247 (121-397)	243 (115-392)	280 (120-552)	323 (151-691)
Système à un seul type d'hôte (n=26)	101.5 (47-495)	103 (48-497)	101.5 (47-495)	139.5 (48-1157)	118.5 (47-747)
Cul-de-sac épidémio-logique (n=17)	219 (98-468)	225 (108-471)	219 (98-468)	484 (103-1946)	351 (112-656)

Annexe 28 : Taille de l'épidémie testée par un GLM Négatif Binomial avec la méthode en variable explicative, selon le scénario de transmission. *Les nombres en gras signalent des IRRs significativement différent de 1 (p<0,05).*

Scénario de transmission	Référence				Reconstructible			
	outbreaker2		*TransPhylo*		*outbreaker2*		*TransPhylo*	
	IRR	p	IRR	p	IRR	p	IRR	p
Taux de mutation élevé (n=21)	1,04	0,48	**1,25**	**<0,001**	1,01	0,84	**1,46**	**<0,001**
Système à un seul type d'hôte (n=26)	1,02	0,75	**1,26**	**<0,001**	1,00	0,99	**1,10**	**<0,001**

Annexe 29 : Nombre médian d'évènements de transmission dus à chaque espèce-hôte dans l'épidémie reconstructible et dans les arbres reconstruits, selon la méthode et scénario de transmission. *() correspond aux plages de valeurs. n est le nombre d'arbres par scénario pour lesquels la reconstruction a été possible.*

Scénario	Espèce-hôte	Nombre d'évènements de transmission dans l'épidémie reconstructible	*outbreaker2*		*TransPhylo*	
			Nombre d'évènements de transmission entre hôtes échantillonnés divisé par π	Nombre d'évènements de transmission dus aux hôtes échantillonnés	Nombre d'évènements de transmission entre hôtes échantillonnés divisé par π	Nombre d'évènements de transmission dus aux hôtes échantillonnés
Taux de mutation élevé (n=21)	Bovin	171 (89-289)	180 (76-310)	178 (68-307)	140 (59-294)	144 (22-294)
	Blaireau	24 (10-77)	31 (10-56)	30 (9-55)	32 (7-59)	32 (11-63)
	Sanglier	25 (4-123)	30 (7-65)	30 (7-64)	25 (8-59)	24 (5-65)
Cul-de-sac épidémio-logique (n=17)	Bovin	187 (66-363)	128 (38-248)	124 (37-244)	159 (36-323)	104 (6-236)
	Blaireau	32 (6-104)	55 (6-151)	54 (6-147)	31 (10-117)	32 (2-101)
	Sanglier	0	35 (13-112)	34 (13-111)	34 (11-97)	16 (3-76)

Annexe 30 : Contribution des espèces-hôtes testée par un GLM Négatif Binomial avec la méthode en variable explicative, selon le scénario de transmission. *Les nombres en gras signalent des IRRs significativement différent de 1 (p<0,05).*

Scénario	Référence				Reconstructible			
	outbreaker2		*TransPhylo*		*outbreaker2*		*TransPhylo*	
	IRR		p		IRR		p	
a) Contribution des bovins								
Taux de mutation élevé (n=21)	1,02	0,61	**0,80**	**<0,001**	0,99	0,77	**0,82**	**<0,001**
Dead-end host (n=17)	**0,65**	**<0,001**	**0,77**	**<0,001**	**0,64**	**<0,001**	**0,52**	**<0,001**
b) Contribution des blaireaux								
Taux de mutation élevé (n=21)	**1,34**	**0,006**	1,20	0,09	**1,29**	**0,01**	**1,26**	**0,02**
Dead-end host (n=17)	**1,89**	**<0,001**	**1,32**	**0,03**	**1,84**	**<0,001**	0,89	0,44
c) Contribution des sangliers								
Taux de mutation élevé (n=21)	**1,46**	**0,03**	1,38	0,07	**1,41**	**0,05**	1,40	0,05

Annexe 31 : Cas index (blaireau) testé par un GLM Binomial avec la méthode en variable explicative. *Les nombres en gras signalent des ORs significativement différent de 1 (p<0,05).*

outbreaker2		seqTrack		*TransPhylo*	
OR	p-value	OR	p	OR	p
-	-	1,00	1,00	0,33	0,22

VETERINARY RESEARCH

RESEARCH ARTICLE

Open Access

A Bayesian evolutionary model towards understanding wildlife contribution to F4-family *Mycobacterium bovis* transmission in the South-West of France

Hélène Duault[1,2], Lorraine Michelet[3], Maria-Laura Boschiroli[3], Benoit Durand[1] and Laetitia Canini[1*]

Abstract

In two "départements" in the South-West of France, bovine tuberculosis (bTB) outbreaks due to *Mycobacterium bovis* spoligotype SB0821 have been identified in cattle since 2002 and in wildlife since 2013. Using whole genome sequencing, the aim of our study was to clarify badger contribution to bTB transmission in this area. We used a Bayesian evolutionary model, to infer phylogenetic trees and migration rates between two pathogen populations defined by their host-species. In order to account for sampling bias, sub-population structure was inferred using the marginal approximation of the structured coalescent (Mascot) implemented in BEAST2. We included 167 SB0821 strains (21 isolated from badgers and 146 from cattle) and identified 171 single nucleotide polymorphisms. We selected a HKY model and a strict molecular clock. We estimated a badger-to-cattle transition rate (median: 2.2 transitions/lineage/year) 52 times superior to the cattle-to-badger rate (median: 0.042 transitions/lineage/year). Using the maximum clade credibility tree, we identified that over 75% of the lineages from 1989 to 2000 were present in badgers. In addition, we calculated a median of 64 transition events from badger-to-cattle (IQR: 10–91) and a median of zero transition event from cattle-to-badger (IQR: 0–3). Our model enabled us to infer inter-species transitions but not intra-population transmission as in previous epidemiological studies, where relevant units were farms and badger social groups. Thus, while we could not confirm badgers as possible intermediaries in farm-to-farm transmission, badger-to-cattle transition rate was high and we confirmed long-term presence of *M. bovis* in the badger population in the South-West of France.

Keywords: Multi-host system, bovine tuberculosis, genomic epidemiology

Introduction

Bovine tuberculosis (bTB) mainly affects cattle, however bTB's most frequent etiological agent, *Mycobacterium bovis*, can also infect other domestic species as well as wildlife species [1]. *M. bovis* host-species depend on the studied area and the role played by wildlife in these various multi-host systems can sometimes prove to be substantial. Indeed, different wildlife species have been implicated as reservoirs of *M. bovis* around the world; e.g. brush-tailed possums (*Trichosurus vulpecula*) in New Zealand [2] and white-tailed deer (*Odocoileus virginianus*) in Michigan, USA [3]. In Europe, evidence supports badgers (*Meles meles*) in Ireland and Britain [4] and wild boars (*Sus scrofa*) in Spain [5] as bTB reservoirs. In France, wildlife *M. bovis* infection was first detected in red deer (*Cervus elaphus*) and wild boars in Normandy in 2001 [6]. Since then, a national wildlife surveillance

*Correspondence: laetitia.canini@anses.fr
[1] Epidemiology Unit, Laboratory for Animal Health, Paris-Est University, Anses, 94700 Maisons-Alfort, France
Full list of author information is available at the end of the article

program, "Sylvatub" has reported infected badgers and boars in persistent clusters of infection such as the Pyrénées-Atlantiques, Landes and Dordogne "départements" (a French administrative subdivision) as well as infected red deer and roe deer (*Capreolus capreolus*) in Dordogne [7]. In addition, *M. bovis* infection has recently been investigated in red foxes (*Vulpes vulpes*) and infection rates comparable to those in badgers and wild boars were found in Dordogne, Charente and Landes [8].

In the European Union (EU), bTB has been until now subject to control programs (EU directive 64/432/EEC). In France, the program for eradication of bTB in cattle, which started in 1954 was quickly followed by a decrease of bTB herd incidence from 13% in 1965 to <0.1% in 2000 [9]. Following a 6 year period with herd prevalence <0.1%, the officially free of bTB (OTF) status was obtained in 2001. The OTF status mainly presents an economic interest since it facilitates live cattle trade in the EU and with other countries (EU directive 64/432/EEC, [10]). However, this status is currently endangered by persistent clusters of infection, especially in the South-West of France [11]. These past 2 years, the majority of infected herds (68/92 (74%) in 2019 and 84/104 (81%) in 2020) were detected in Nouvelle-Aquitaine (according to the Animal Health Epidemiological platform ESA), a "région", which contains the Pyrénées-Atlantiques, Landes, Charente and Dordogne, amongst other "départements".

Systematic *post mortem* inspection of bovine carcasses for lesions compatible with bTB in French abattoirs constitutes the first component of cattle surveillance and periodical herd skin-testing, the second and main component, which currently detects around 70% of bTB clusters (according to the Animal Health Epidemiological platform ESA). Single intradermal tuberculin tests (SITT) or single intradermal comparative tuberculin tests (SICTT) are performed in the cervical region and results are read 72 h post-injection. Herd skin-testing regularity, ranging from annual testing of all animals older than 6 weeks to no testing, is decided at the "département" level. In the Pyrénées-Atlantiques and Landes, herd skin-testing was reinforced in 2012 to an annual regularity in "communes" (i.e. the smallest French administrative subdivision) where bTB outbreaks were detected the previous year. Before the generalization of this annual regularity to all "communes" in 2018, herd skin-testing regularity for the "communes" where bTB outbreaks had not been detected the previous year, was every 2 years in the Landes and every 3 years in the Pyrénées-Atlantiques. Skin-testing is also performed before introduction of all cattle in transit for more than 6 days, coming from at-risk herd, transiting through a high-risk herd with high turnover or coming from a "département" with a 5 year cumulative incidence higher than the national average incidence and when investigating an epidemiological link to a confirmed outbreak.

After culling following a positive skin-testing or after a *post mortem* lesion detection, subsequent PCR testing and bacterial culture are conducted in order to detect mycobacteria. Strains are then sent to the National Reference Laboratory (NRL) where genotypes are determined using spoligotyping and VNTR (Variable Number Tandem Repeat) typing methods, this has led to the identification of regional genotypes [12]. Positive identification of a bTB case in a farm leads to an official declaration of infection and control measures are then implemented; depending on the control strategy, this official declaration of infection could cause long-term depopulation of vulnerable cattle farms [13].

Following the report of *Mycobacterium bovis* infection in red deer and wild boars in 2001, with genotyping linking this wildlife outbreak to cases in nearby cattle [6] and the discovery of infected wildlife in other affected regions elsewhere in France, "Sylvatub", was started in 2011 to investigate bTB infection in badgers, boars, red deer and roe deer [7]. "Sylvatub" submits road-kill, hunting carcasses and animals captured in annual campaigns designed at the "département" level, to a protocol similar to cattle surveillance, i.e. PCR testing, bacterial culture and genotyping [7].

The majority of *M. bovis* detection in wildlife are located in the vicinity of cattle outbreaks and present the same genotypes [14]. A current example of this can be found in the Pyrénées-Atlantiques and Landes "départements", where "Sylvatub" was started in 2012 and surveillance data reported two spoligotypes belonging to the F4-family/cluster A [15, 16] shared by cattle, badgers and wild boars: SB0821 and SB0832 [14]. In the Pyrénées-Atlantiques and Landes, the number of newly infected herds declared each year ranged from 16 to 31 between 2012 and 2017, without any obvious trend (Boschiroli, personal communication). Moreover, the apparent prevalence of bTB in badgers in the region was estimated at 5.9% [3.9–6.8%] 95% CI in 2013–2014 (by culture) and 7.9% [5.2–11.2%] 95% CI in 2016–2017 (by PCR testing) [7].

Since the same genotype profile is shared by both cattle and wildlife, a more discriminating method to differentiate strains is necessary in order to understand transmission dynamics. Whole genome sequencing data has been previously selected for its higher resolution to investigate bTB transmission [17, 18].

When studying transmission dynamics between different populations, a Bayesian evolutionary model applied to *M. bovis* transmission between cattle and wildlife, while not always conclusive on the direction of transmission [17], has recently brought insights to badger

intervention in bTB transmission in the UK [19]. In a Bayesian evolutionary model, genetic sequences are annotated by a state e.g. geographical locations [20, 21] or host-species [22, 23]. Reconstruction of ancestral node states in the phylogeny enables estimation of migration processes between populations. In this study, our aim was to analyze whole genome sequencing data using a Bayesian evolutionary model in order to better understand badger contribution to transmission in a SB0821 bTB multi-host system, in the Pyrénées-Atlantiques and Landes French "départements".

Materials and methods

Study area and data collection

Our study area consisted of the "communes" selected in previous works on the badger-cattle bTB system in the South-West of France [24, 25]. This study area was restricted to a 3754 km^2 area of 335 "communes" (Figure 1), straddling the border of Pyrénées-Atlantiques and Landes.

A maximum of three SB0821 strains per official declaration of infection per farm, collected between 2002 and 2017, were included in the study. All SB0821 badger strains collected during our study period by "Sylvatub" were included.

The sampling date considered was either the date of slaughter for cattle strains or the date of capture recorded by "Sylvatub" for badger strains. Cattle information was provided by the "Base de données nationale d'identification" (BDNI), in which every bovine is registered in France. BDNI records date of birth, date of death, cattle movements and their cause (e.g. trade, slaughter).

Genomic data

Upon reception at the NRL, liquid culture media (7H9 + ADC) was employed to grow the *M. bovis* strains. After a heating step, the lysate obtained was sent for purification and Illumina sequencing (paired-end 2*150 bp) to the Paris Brain Institute (ICM). At the ICM, sequencing quality was controlled using FASTQC with an acceptability Phred score threshold of 30. Sequence alignment and Single Nucleotide Polymorphism (SNP) calling were computed at the NRL using the AF2122/97 reference strain on Bionumerics software, version 7.6 (AppliedMath, Belgium). SNPs identified were selected according to strict criteria of wgSNP module: (i) they had to be present on at least 5 reads in both forward and reverse direction, (ii) 12 base pairs had to separate

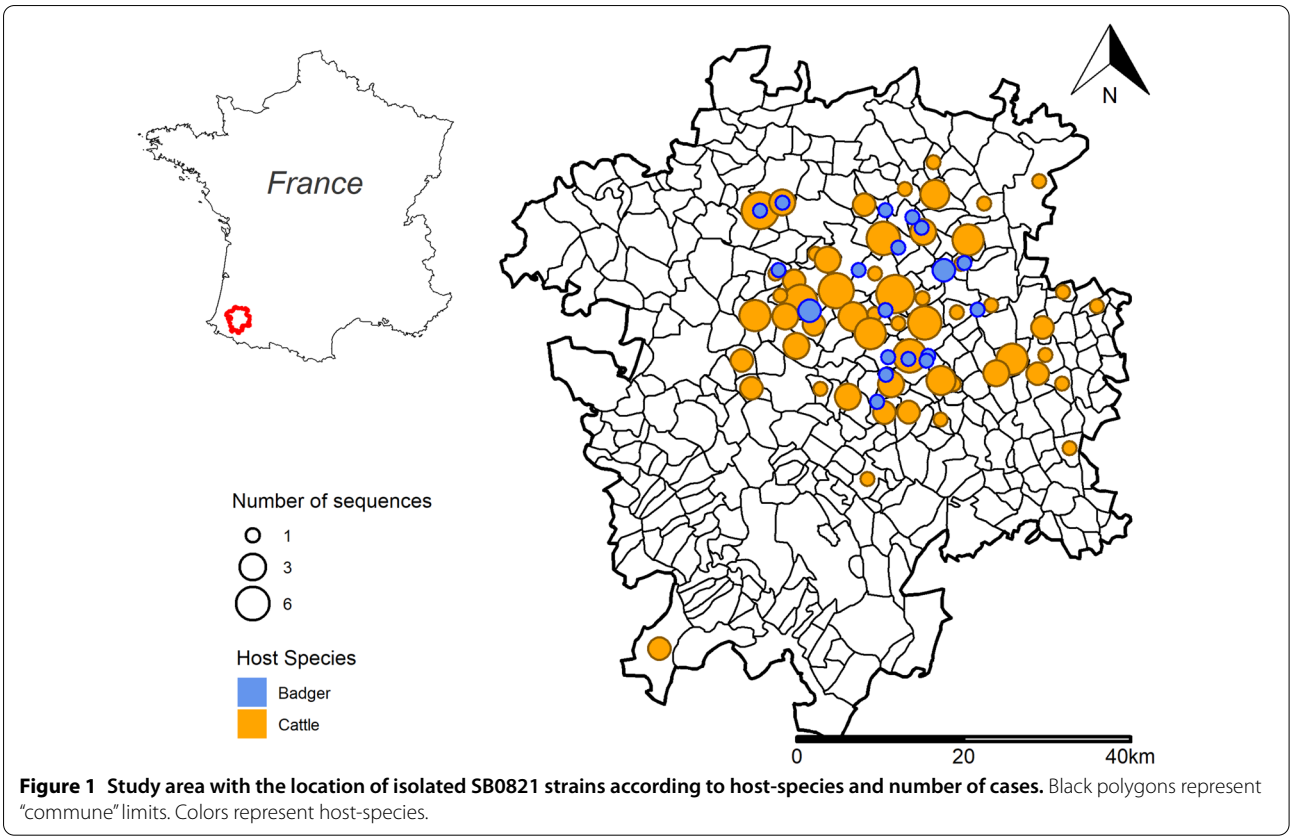


Figure 1 Study area with the location of isolated SB0821 strains according to host-species and number of cases. Black polygons represent "commune" limits. Colors represent host-species.

them, (iii) they were not present in repetitive regions of the genome and (iv) ambiguous SNPs (at least one unreliable (N) base, ambiguous (non ATCG) base or gap) were not included. SNPs were then used to reconstruct a maximum parsimony tree on Bionumerics in order to identify genetic outliers. We estimated pairwise distances between sequences using the *dist.dna* function available in the *ape* package version 5.0 [26] on R4.0.3. The model considered was the F84 model [27] since it closely resembles the HKY model [28] that was previously used on *M. bovis* strains [17, 29]. Distances were estimated between all, badger and then cattle sequences.

Bayesian evolutionary model

We used a Bayesian evolutionary model consisting of a structured coalescent population model, in order to capture the transition between two pathogen populations defined by their host-species, using BEAST2 (Bayesian Evolutionary Analysis by Sampling Trees) v2.6.3 [30]. The probability of nucleotide substitutions was described by the substitution model and the molecular clock modeled the evolution of substitution rates across branches. Moreover, the pathogen sub-population structure was inferred using the marginal approximation implemented in the Mascot package v2.1.2 [21].

We needed to select the most appropriate substitution and molecular clock models. Models were finally compared using the Bayes Factor (BF) after estimating their marginal likelihoods with the "Nested Sampling" algorithm implemented in the NS package v1.1.0 [31]. In the NS estimation, the number of particles was $N=1$ (or 10 if results were inconclusive with $N=1$) and subchain length was fixed to 100 000. We first tested three substitution models: the Jukes-Cantor (JC) model [32], in which all substitutions are equally likely and base frequencies are equal, the Hasegawa-Kishino-Yano (HKY) [28], in which substitution probabilities depend on the nature of bases and all base frequencies differ, and Generalized-Time-Reversible (GTR) [33] model, where all substitution probabilities and base frequencies are independent. We then tested three molecular clock models: the strict clock with constant substitution rates across branches [34], and the relaxed uncorrelated lognormal and exponential clocks with substitution rates varying over branches [35]. All models were tested in BEAST2 software after annotation on BEAUti interface [30]. We set the site model frequencies parameter to empirical and the constant effective population prior to a lognormal distribution (mean: 0, standard deviation: 1); other parameters kept their default settings. We selected a chain length of 300 million iterations, a burn-in period of 10% and a sampling frequency of 1 in 30 000. Four replicates were performed and combined in LogCombiner v.2.6.3 with a lower sampling frequency of 1 in 120 000.

We checked for convergence i.e. stationary distribution of the MCMC (Markov chain Monte Carlo), and independence of sampling (Effective Sample Size (ESS) above 200 for each parameter), on TRACER v1.7.1 [36]. To summarize the posterior sampled trees, a maximum clade credibility (MCC) tree was built via Tree Annotator using the common ancestor heights option [37]. In the MCC tree, host-species of internal nodes were considered unknown if the posterior probability of their "host" (i.e. "Badger" or "Cattle") was lower than 0.70, otherwise three host-species probability categories were represented:]0.7; 0.8],]0.8; 0.9] and]0.9; 1]. We then inferred the lineages' host-species through time by considering that state transition between two nodes occurred at the parental node. The MCC tree was visualized on R4.0.3 with treeio [38] and ggtree packages [39].

In addition, we resampled the posterior trees at a frequency of 1 in 1 200 000 using LogCombiner and imported the resulting 1004 trees in R. In a phylogenetic tree, we have information on the host-species (badger or cattle) of each node. We can therefore count the number of times a parental node and a descendant node do not belong to the same host-species. This number corresponds to inter-species lineage transitions (badger-to-cattle and cattle-to-badger). However, when two consecutive nodes belong to the same host-species, we cannot infer with our method whether the lineage remains in the same animal, in the same group of animals (social group for badgers, farm for cattle), or if there is one or multiple within-species transmission events, within and/or between groups of animals. Similarly, between two nodes hosted by different species, at least one transmission event took place (between a badger and cattle) however, other transmission events could have taken place. Therefore for each tree, we counted the number of inter-species lineage transitions, the number of times lineages remained in the same host-species between two nodes (which we called intra-species persistence) as well as the number of unknown transitions, i.e. the number of times one or both consecutive nodes are considered unknown (host-species probability lower than 0.70). We then calculated the proportion of lineage transitions through time by summing the number of each transition type per year divided by the number of transitions occurring in that year. We considered that state transition between two nodes occurred at the parental node and dated these transitions using the *castor* package version 1.7.0 [40]. Thus, the number of transitions through time corresponds to the sum of internal nodes dated from each year multiplied by two (since one node diverges into

two lineages). So transition from an internal node to a tip (representing the isolates) will not be represented at the time of sampling of isolates but at the time of the internal node, that immediately precedes the isolate in the tree. Transition from an internal node to a tip representing a cattle (badger) isolate could correspond to either a badger-to-cattle (cattle-to-badger) transition, an unknown transition or cattle (badger) persistence.

Similarly to the MCC tree, three probability thresholds were used to determine the host-species of internal nodes: 0.7, 0.8 and 0.9. However, since all 1004 resampled trees are studied rather than the consensus tree, the most recent common ancestor (MRCA) of some trees can be dated from before the MRCA of the MCC tree. Therefore, the time range considered is wider than for the MCC tree.

Results

Genomic data

From 167 SB0821 strains, 171 SNPs were identified. Pairwise distances between sequences ranged from 0 to 0.145 for all and cattle strains (median: 0.042) and from 0 to 0.108 for badger strains (median: 0.036). Among these 167 SB0821 strains, 146 were isolated from cattle and 21 from badgers (Figure 2). In 2002, SB0821 strains were first detected in cattle, which preceded our first SB0821 badger strain (2013).

Bayesian evolutionary model

We selected a strict molecular clock based on the BF comparisons (Additional file 1). However, we could not differentiate between the HKY and GTR substitution models; HKY was chosen based on previous works [17, 29].

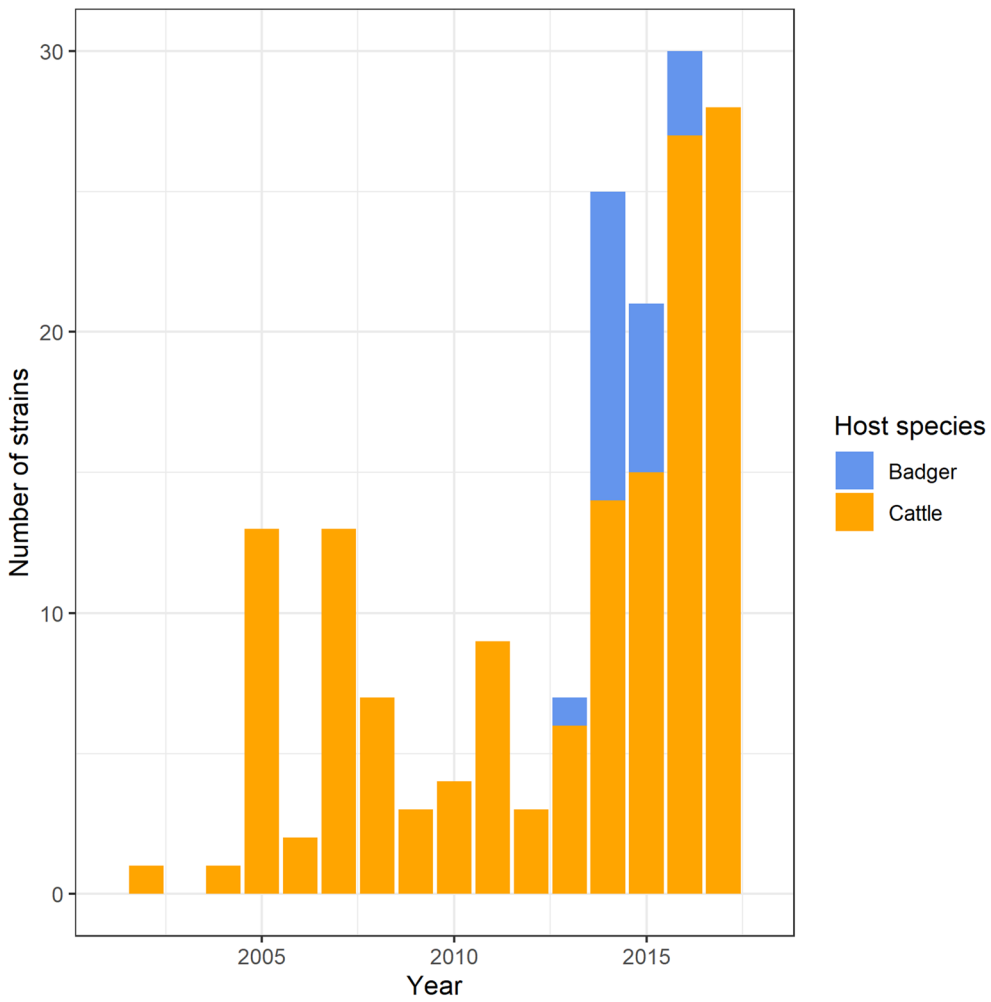


Figure 2 Number of SB0821 strains isolated by year of death. Colors represent host-species.

The median transition/transversion ratio (kappa) parameter of the HKY model was estimated at 5.9 (95% HPD: "High Posterior Density": [4.2; 8.2]) (Additional file 2). Estimations of median substitution rate and tree height were respectively 0.41 substitutions/genome/year (95% HPD: [0.29; 0.55]) and 27.5 years (95% HPD: [21.0; 36.6]). Therefore, the MRCA was estimated to have been circulating in 1990 (95% HPD: [1980; 1996]). The model estimated a badger-to-cattle transition rate (median of 2.2 transitions/lineage/year, 95% HPD: [0.74; 4.5]) 52 times superior to the cattle-to-badger transition rate (median 0.042 transitions/lineage/year, 95% HPD: $[3.5 \times 10^{-5}$; 0.24]). Estimation of effective population sizes Ne was higher for the badger population (median of 34, 95% HPD: [20; 51]) than for the cattle population (median of 1.2, 95% HPD: [0.27; 2.7]) (Additional file 2).

In the MCC tree (Figure 3 and Additional files 3, 4), 81 out of 166 internal nodes including the root (with a host probability equal to 0.94) and the nodes closest to the root were identified as hosted by badgers (55 with a posterior probability > 0.9, 15 with a posterior probability between 0.8 and 0.9 and 11 between 0.7 and 0.8). Among the remaining 85 internal nodes, host-species were identified as cattle for 57 nodes (42 with a posterior probability > 0.9, 7 with a posterior probability between 0.8 and 0.9 and 8 between 0.7 and 0.8) and

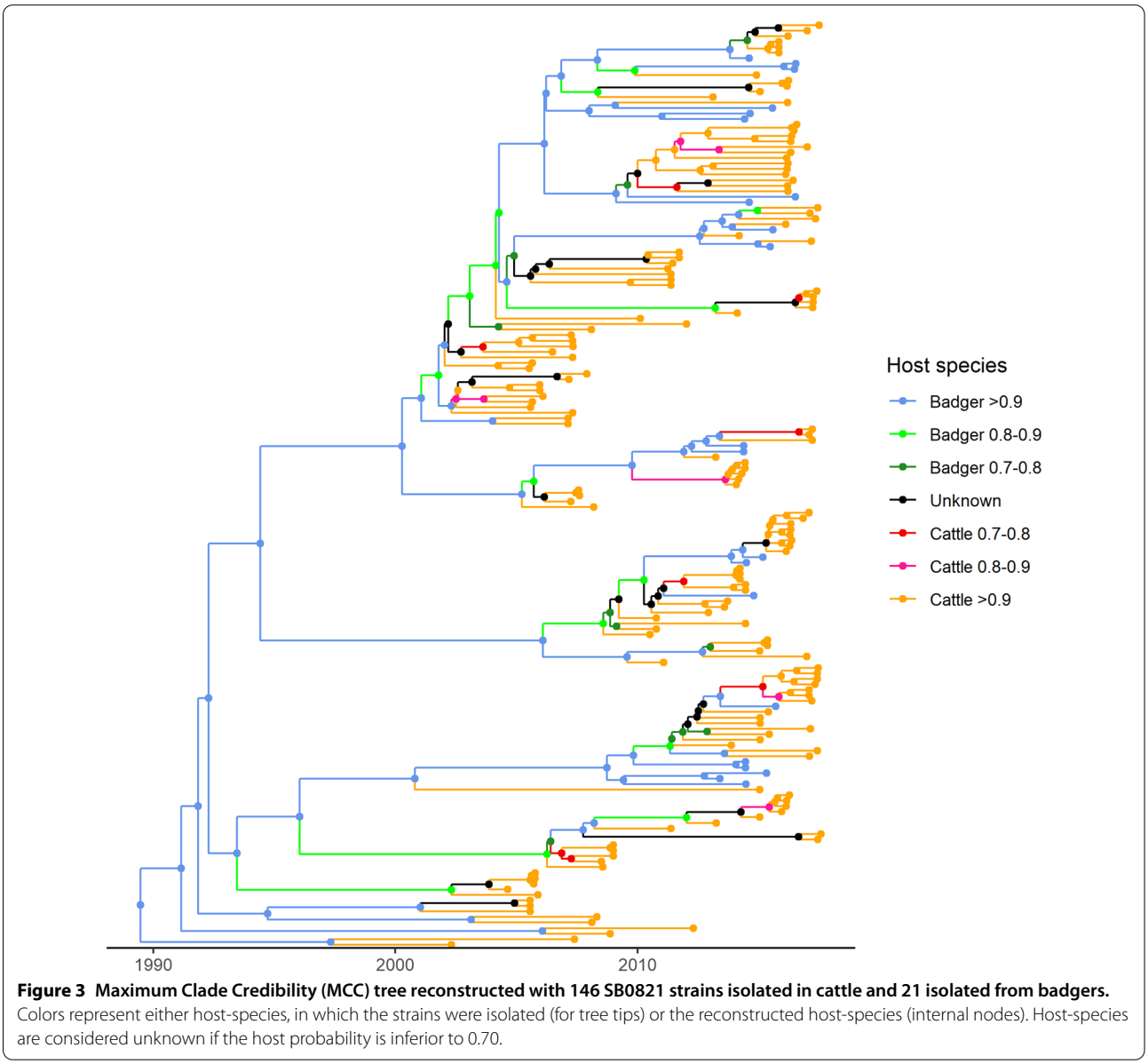


Figure 3 Maximum Clade Credibility (MCC) tree reconstructed with 146 SB0821 strains isolated in cattle and 21 isolated from badgers. Colors represent either host-species, in which the strains were isolated (for tree tips) or the reconstructed host-species (internal nodes). Host-species are considered unknown if the host probability is inferior to 0.70.

unknown for 28. Figure 4 depicts the host-species of lineages in the MCC tree through time, lineages were estimated to have been circulating solely in badgers until 1996 and over 75% of lineages per year were present in badgers until 2000. Moreover, we predicted two peaks of cattle lineages, one occurring in the mid-2000s and another in the mid-2010s, following a badger peak.

Among the 1004 resampled trees, the median tree height was estimated and corresponded to a MRCA circulating in 1990 (1st quartile: 1978, 3rd quartile: 1996). Moreover, in these trees and when considering the 0.9 probability threshold, we calculated a median of 64 badger-to-cattle transition events (1st quartile: 10, 3rd quartile: 91) and zero cattle-to-badger transition (1st quartile: 0, 3rd quartile: 3). However, the number of times a lineage persists in the same host-species are similar when considering the badger (median: 109, 1st quartile: 14, 3rd quartile: 137) and the cattle population (median: 112, 1st quartile: 78, 3rd quartile: 158). This asymmetry between the number of inter-species transitions as well as the similarity between the intra-species persistence were observed for the three different thresholds (Additional file 5).

Figure 5 shows that the type of lineage transitions observed in the 1004 trees varies over time. The proportion of badger persistence constituted over 50% of lineage transitions from 1964 to 2001 (excepted in 1974, where 50% of lineages were unknown with the 0.9 probability threshold) while cattle persistence started in 1990 at the earliest. Cattle persistence represented over 50% of transitions in 2005 and again in 2016 and 2017, which corresponds to the dates of the two cattle lineage peaks in the MCC tree. In addition, the proportion of cattle-to-badger transitions never exceeded 1.3% of lineage transitions.

Discussion

In this work, we used whole genome sequencing data in order to investigate the role played by badgers in SB0821 *M. bovis* strains transmission in the Pyrénées-Atlantiques and Landes. This region situated in the South-West of France is of major interest concerning bTB control in the country since it has been continuously harboring persistent clusters of infection, especially in the past decade (according to surveillance data available on the Animal Health Epidemiological platform ESA).

For the Bayesian evolutionary model, we selected a strict molecular clock. We also chose a HKY

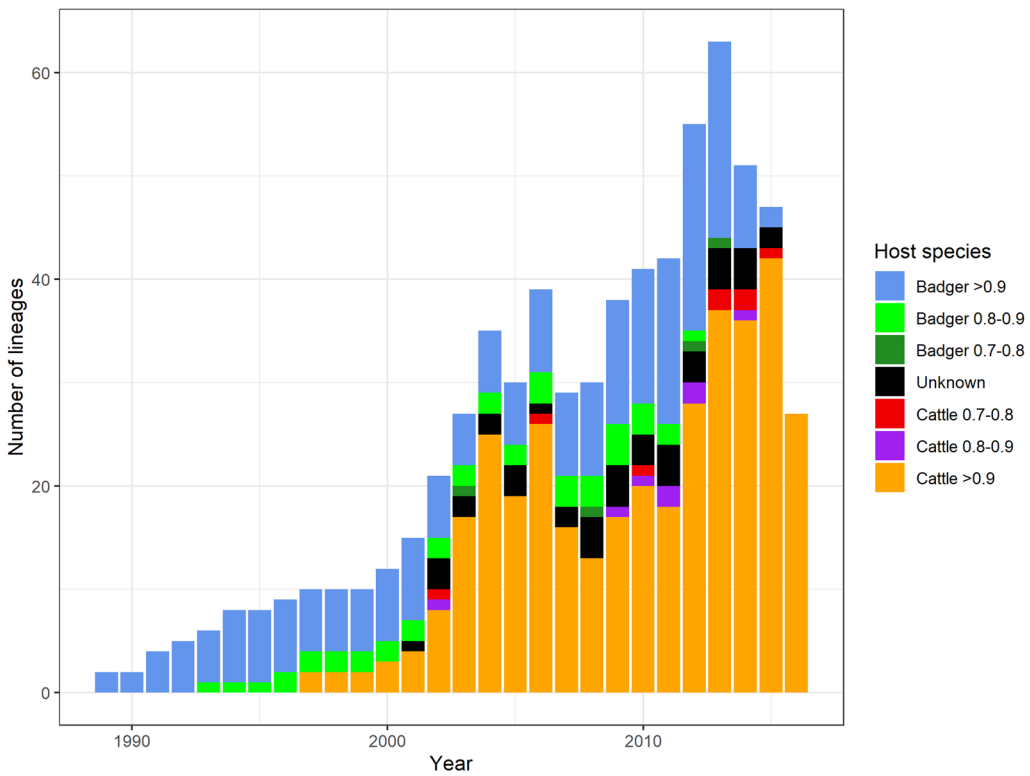


Figure 4 Host-species of lineages through time estimated in the Maximum Clade Credibility (MCC) tree. Colors represent host-species and posterior probability. Host-species are considered unknown if the host probability is inferior to 0.70.

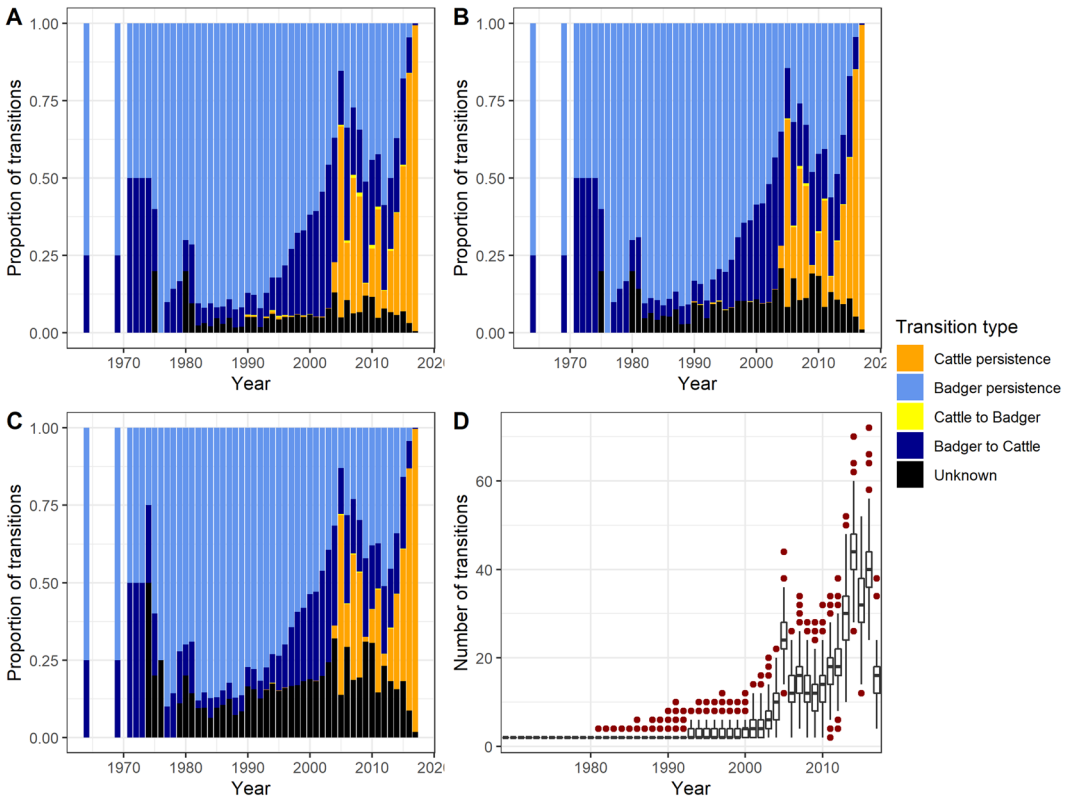


Figure 5 Proportion of lineage transitions estimated in the 1004 resampled trees. We used a 0.7 (**A**), 0.8 (**B**) and a 0.9 (**C**) probability threshold. In (**D**), the number of transitions per tree from the 1004 trees is represented over time. Colors represent transition type.

substitution model over GTR according to past Bayesian studies [17, 29]. Estimated substitution rate of 0.41 substitutions/genome/year (95% HPD: [0.29; 0.55]) was higher than estimations from past studies in Northern Ireland (0.15, 95% HPD: [0.04–0.26] [41] and 0.2, 95% HPD: [0.1–0.3] [42]) and in Michigan, USA (0.2, 95% HPD: [0.1–0.3] [29]). However, Crispell et al. in 2017 [17] estimated a higher rate of 0.53 substitutions/genome/year, 95% HPD: [0.22–0.94]. Observed differences between these studies could be attributed to the *M. bovis* lineage; specific lineage characteristics have been highlighted in *M. tuberculosis* [43]. *M. bovis* strains studied by Biek et al. and Trewby et al. [41, 42] are part of the Eu1 clonal complex [44]. In France, *M. bovis* lineages differ according to the area studied [15], our estimations were based on SB0821 strains, which belong to the F4-family/cluster A [16]. Lastly, studies based on *M. bovis* strains, which did not share the same spoligotype nor VNTR profile [17] could estimate a higher substitution rate. Variation between estimations could also depend on the sampled host-species. Wildlife species studied vary, e.g. the white-tailed deer and the elk in Michigan [29] or even the badger in Northern Ireland [41, 42].

Bayesian inference methods are used to reconstruct ancestral node states (e.g. in our case, host-species) and to estimate parameters such as inter-species transition rates. In order to mitigate the bias due to sampling process in the migration rate estimation, we chose to use a structured coalescent method. More specifically, we used the Marginal Approximation of the Structured COalescenT (MASCOT) to model the evolution of SB0821 strains isolated from cattle and badgers [21]. This method assumes a constant effective population size over time. The past demographics until 2017 are difficult to estimate in the region due to the late implementation of the wildlife surveillance system in 2012. The discovery of bTB in wildlife caused an increase in cattle surveillance and thus an increase in bTB detection in cattle. However, the apparent prevalence in badgers did not seem to vary significantly between the two estimations (2013–2014 and 2016–2017) by Réveillaud et al. [7] and the number of newly infected farms did not follow an obvious trend over the same period of time (Boschiroli, personal communication).

Similar genetic distances were estimated between cattle and badger sequences. However, considering the fact that this was based on 21 badger sequences and 146 cattle

sequences, this means that badger strains presented a higher genetic diversity. This was consistent with the higher effective population size estimated in the badger population compared to the cattle population.

The average badger density in 13 study sites (including a 50 km^2 site situated in our study area) in France was estimated at 3.8 badgers per km^2 (range: 1.7–7.9), which is "relatively lower than those found in the UK and concordant with global estimates from Ireland" [45]. We inferred a badger-to-cattle transition rate 52 times greater than the cattle-to-badger rate. Our results are consistent with a previous study by Crispell et al. showing a badger-to-cattle transition rate (0.045 transitions/lineage/year) 10 times superior to the cattle-to-badger transition rate (0.0044 per lineage per year) on a subset of cattle ($n=83$) and badger ($n=97$) strains isolated in the UK [19]. Conversely, Rossi et al. estimated a higher cattle-to-badger transition rate in a newly infected region in the North-West of England and concluded on the possible requirement of a "build-up in badger infections […] before badger-to-cattle infections become probable" [46].

Crispell et al. estimated similar results than in our work concerning inter-species transmission events comprised mainly of badger-to-cattle transmission and a median of zero cattle-to-badger transmission. Moreover, the variation of lineages' host-species through time in our consensus tree showed an increase in cattle lineages following an increase in badger lineages. These results suggest that badger-to-cattle transmission may be amplified by onward cattle-to-cattle transmission, a hypothesis proposed by Donnelly and Nouvellet in the UK [47]. Similarly, a study that analyzed an empirical contact network of cattle farms in the same region, concluded on the importance of badger-mediated contacts in bTB spread [24].

Crispell et al. had an interesting approach and estimated the minimum number of intra-species and inter-species transmission events [19]. To this end, the authors assumed that a coalescent event corresponded to at least one transmission event. Therefore, the existence of a single pathogen lineage within an infected animal was implied. We did not make any assumptions on the within-host evolution nor on the timing of transmission. While we considered an inter-species transition to correspond to at least one transmission event between cattle and badger, we did not estimate the number of within-species transmission events. However, the majority of transitions being identified as persistence suggests the importance of intra-species transmission events highlighted in Crispell et al.'s work [19]. Indeed, the majority of the lineage transitions until 2001 were identified as badger persistence, which either correspond to the evolution of *M. bovis* lineages in the same badger or bTB transmission between badgers belonging to the same social group (i.e. sett) or to neighboring setts. Conversely, Bouchez-Zacria et al. used a stochastic model of *M. bovis* transmission within the badger-cattle system in the Pyrénées-Atlantiques and Landes and determined limited inter-social group bTB transmission [48]. Therefore, the majority of badger-to-badger persistence until 2001 suggest either long-term carriage of bTB in a badger and/or a long transmission history within a social group.

Similarly to badger persistence, cattle persistence could either correspond to the evolution of *M. bovis* lineages in the same animal, intra-farm or between-farm bTB transmission. Trade data provided by the BDNI determined the significant role of cattle movements in bTB transmission in the Pyrénées-Atlantiques and Landes [24], which could explain between-farm bTB transmission. However, transmission modeling of this badger-cattle system determined that 49.3% of farm infections were due to proximity to pastures belonging to an infected farm [48].

Nonetheless, badger-to-cattle and cattle-to-cattle transmissions are not the only possible source of farm infection. In these two models [24, 48] as well as in our work, contribution of other wildlife species were not included. Infected wild boars have been detected in the Pyrénées-Atlantiques and Landes since the implementation of "Sylvatub" in the area. While a badger movement study in Europe estimated a mean distance of 1.7 km traveled by badgers with some rare long distance travels of up to 22 km [49], the mean daily distances traveled by wild boars is estimated to be around 7–13 km [50]. Therefore, wild boars could have contributed to *M. bovis* spread in a way badgers, typically traveling shorter distances, could not have.

We used genomic data to study the role of badgers in bTB transmission in the Pyrénées-Atlantiques and Landes. However, the sampling process differs between cattle and wildlife strains. According to expert opinions, possible environmental contamination and deterioration of wildlife carcasses could lower the culture sensitivity by 35% and since wildlife samples are pooled, PCR test sensitivity could decrease by 15% [51].

Moreover, in practice, while herd skin-testing rhythm varies between "communes", cattle surveillance concerns all animals over 24 months. In our study area, while the testing of road-killed badgers did not depend on the presence of bTB in cattle, badger capture protocol changed over the years and varied from one place to another according to the detection of nearby cases in cattle. Contrary to the registered and easily accessible cattle population, the entirety of the badger population cannot be surveilled for practical (free-ranging population) and financial reasons, which contributes to an unavoidable underestimation of cases. However, our

choice in the MASCOT method was motivated by the fact that it does not treat the number of each host-species as data and thus helps reduce the impact of sampling bias [52].

In conclusion, our Bayesian evolutionary model enabled us to infer inter-species (badger-to-cattle and cattle-to-badger) transitions but not intra-species transmission as in previous epidemiological studies, where relevant units were farms and badger social groups. Therefore, we could not confirm badger social groups as possible intermediaries in farm-to-farm transmission. However, our results highlighted long-term *M. bovis* presence in the badger population and a high badger-to-cattle transition rate in the Pyrénées-Atlantiques and Landes, which justifies control measures implemented to prevent contacts between cattle and badgers. Further research including transmission tree reconstruction of this multi-host system could help us better understand intra-species bTB transmission and integrate the contribution of other wildlife species in this bTB multi-host system. Including further genomic data isolated from cattle, badgers and especially wild boars would improve our work.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13567-022-01044-x.

Additional file 1: Model selection with the Bayes Factor. N corresponds to the number of particles, ML1 (2) to the log maximum likelihood of model 1 (2) and SD to the standard deviation. Log(BF) is the difference between ML1 and ML2. If (BF) > 0 (< 0) than model 1 (2) is favored. "-" means that the results were inconclusive. Subpopulations defined by host-species are not taken into account.

Additional file 2: Parameters estimation (median and 95% High Posterior Density (HPD) interval) and effective sample size (ESS). Ne is the effective population size and ESS stands for effective sample size.

Additional file 3: Maximum Clade Credibility (MCC) tree reconstructed with strain names. Colors represent either host-species, in which the strains were isolated (for tree tips) or the reconstructed host-species (internal nodes). Host-species are considered unknown if the host probability is inferior to 0.70.

Additional file 4: Maximum Clade Credibility (MCC) tree reconstructed with posterior probabilities. Colors represent the posterior probability of the nodes.

Additional file 5: Number of inter-species transitions per tree calculated over 1004 sampled trees. Number of transitions are represented according to transition type and various probability thresholds (0.7, 0.8 and 0.9).

Additional file 6: List of sample names, host-species and their accession numbers.

Authors' contributions

LC, LM and MLB designed the work. HD, LC and LM contributed to the analysis of data. All authors contributed to the interpretation of data. HD drafted the work and all authors revised it. All authors read and approved the final manuscript.

Funding

The French ministry of Agriculture financed the sampling and the analyses in the framework of the RFSA call on TB projects (Anses-DGAl credit agreement no 170131). This work was also financially supported by the Université Paris-Saclay, which funded H.D.'s PhD grant.

Availability of data and materials

All WGS data used for these analyses have been uploaded to the European Nucleotide Archive (PRJE45853). The individual isolates can be accessed under the following Biosample accession numbers: SAMEA8939070-SAMEA8939236 (see Additional file 6).

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Epidemiology Unit, Laboratory for Animal Health, Paris-Est University, Anses, 94700 Maisons-Alfort, France. [2]Université Paris-Saclay, Faculté de Médecine, 94270 Le Kremlin-Bicêtre, France. [3]Tuberculosis Reference Laboratory, Bacterial Zoonosis Unit, Laboratory for Animal Health, Paris-Est University, Anses, 94700 Maisons-Alfort, France.

Received: 13 July 2021 Accepted: 6 March 2022
Published online: 02 April 2022

References

1. O'Reilly LM, Daborn CJ (1995) The epidemiology of *Mycobacterium bovis* infections in animals and man: a review. Tuber Lung Dis 76:1–46. https://doi.org/10.1016/0962-8479(95)90591-x
2. Kean JM, Barlow ND, Hickling GJ (1999) Evaluating potential sources of bovine tuberculosis infection in a New Zealand cattle herd. N Z J Agric Res 42:101–106. https://doi.org/10.1080/00288233.1999.9513358
3. O'Brien DJ, Schmitt SM, Fierke JS, Hogle SA, Winterstein SR, Cooley TM, Moritz WE, Diegel KL, Fitzgerald SD, Berry DE, Kaneene JB (2002) Epidemiology of *Mycobacterium bovis* in free-ranging white-tailed deer, Michigan, USA, 1995–2000. Prev Vet Med 54:47–63. https://doi.org/10.1016/S0167-5877(02)00010-7
4. Simpson VR (2002) Wild animals as reservoirs of infectious diseases in the UK. Vet J 163:128–146. https://doi.org/10.1053/tvjl.2001.0662
5. Naranjo V, Gortazar C, Vicente J, de la Fuente J (2008) Evidence of the role of European wild boar as a reservoir of *Mycobacterium tuberculosis* complex. Vet Microbiol 127:1–9. https://doi.org/10.1016/j.vetmic.2007.10.002
6. Zanella G, Durand B, Hars J, Moutou F, Garin-Bastuji B, Duvauchelle A, Fermé M, Karoui C, Boschiroli ML (2008) *Mycobacterium bovis* in wildlife in France. J Wildl Dis 44:99–108. https://doi.org/10.7589/0090-3558-44.1.99
7. Réveillaud É, Desvaux S, Boschiroli M-L, Hars J, Faure É, Fediaevsky A, Cavalerie L, Chevalier F, Jabert P, Poliak S, Tourette I, Hendrikx P, Richomme C (2018) Infection of wildlife by *Mycobacterium bovis* in France assessment through a national surveillance system. Sylvatub Front Vet Sci 5:562. https://doi.org/10.3389/fvets.2018.00262
8. Richomme C, Réveillaud E, Moyen J-L, Sabatier P, De Cruz K, Michelet L, Boschiroli ML (2020) *Mycobacterium bovis* infection in red foxes in four animal tuberculosis endemic areas in France. Microorganisms 8:1070. https://doi.org/10.3390/microorganisms8071070
9. Bekara MEA, Courcoul A, Bénet J-J, Durand B (2014) Modeling tuberculosis dynamics, detection and control in cattle herds. PLoS One 9:e108584. https://doi.org/10.1371/journal.pone.0108584
10. Reviriego Gordejo FJ, Vermeersch JP (2006) Towards eradication of bovine tuberculosis in the European union. Vet Microbiol 112:101–109. https://doi.org/10.1016/j.vetmic.2005.11.034
11. Delavenne C, Pandolfi F, Girard S, Réveillaud É, Boschiroli M-L, Dommergues L, Garapin F, Keck N, Martin F, Moussu M, Philizot S, Rivière J, Tourette I, Dupuy C, Dufour B, Chevalier F (2020) Tuberculose bovine: bilan et évolution de la situation épidémiologique entre 2015 et 2017 en France

Métropolitaine. Bull Epidemiol 91:12 (in French). Available online at: https://be.anses.fr/sites/default/files/O-034_2021-08-04_Tub-Bilan_Delavenne_MaqF.pdf
12. Michelet L, Durand B, Boschiroli ML (2020) Tuberculose bovine: Bilan génotypique de M. bovis à l'origine des foyers bovins entre 2015 et 2017 en France Métropolitaine. Bull Epidemiol 91:13 (in French) Available online at: https://be.anses.fr/sites/default/files/O-035_2021-08-04_Tub-genotyp-Michelet_MaqF.pdf
13. Canini L, Durand B (2020) Resilience of French cattle farms to bovine tuberculosis detection between 2004 and 2017. Prev Vet Med 176:104902. https://doi.org/10.1016/j.prevetmed.2020.104902
14. Desvaux S, Réveillaud É, Richomme C, Boschiroli M-L, Delavenne C, Calavas D, Chevalier F, Jabert P (2020) Sylvatub: Bilan 2015–2017 de la surveillance de la tuberculose dans la faune sauvage. Bull Epidemiol 91:14 (in French). Available online at: https://be.anses.fr/sites/default/files/O-036_2021-08-04_Sylvatub_Desvaux_MaqF.pdf
15. Hauer A, Cruz KD, Cochard T, Godreuil S, Karoui C, Henault S, Bulach T, Bañuls A-L, Biet F, Boschiroli ML (2015) Genetic evolution of *Mycobacterium bovis* causing tuberculosis in livestock and wildlife in France since 1978. PLoS One 10:e0117103
16. Hauer A, Michelet L, Cochard T, Branger M, Nunez J, Boschiroli M-L, Biet F (2019) Accurate phylogenetic relationships among *Mycobacterium bovis* strains circulating in France based on whole genome sequencing and single nucleotide polymorphism analysis. Front Microbiol 10:955. https://doi.org/10.3389/fmicb.2019.00955
17. Crispell J, Zadoks RN, Harris SR, Paterson B, Collins DM, de Lisle GW, Livingstone P, Neill MA, Biek R, Lycett SJ, Kao RR, Price-Carter M (2017) Using whole genome sequencing to investigate transmission in a multi-host system: bovine tuberculosis in New Zealand. BMC Genom 18:180. https://doi.org/10.1186/s12864-017-3569-x
18. Price-Carter M, Brauning R, de Lisle GW, Livingstone P, Neill M, Sinclair J, Paterson B, Atkinson G, Knowles G, Crews K, Crispell J, Kao R, Robbe-Austerman S, Stuber T, Parkhill J, Wood J, Harris S, Collins DM (2018) Whole genome sequencing for determining the source of *Mycobacterium bovis* infections in livestock herds and wildlife in New Zealand. Front Vet Sci 5:272. https://doi.org/10.3389/fvets.2018.00272
19. Crispell J, Benton CH, Balaz D, De Maio N, Ahkmetova A, Allen A, Biek R, Presho EL, Dale J, Hewinson G, Lycett SJ, Nunez-Garcia J, Skuce RA, Trewby H, Wilson DJ, Zadoks RN, Delahay RJ, Kao RR (2019) Combining genomics and epidemiology to analyse bi-directional transmission of *Mycobacterium bovis* in a multi-host system. eLife 8:e45833. https://doi.org/10.7554/eLife.45833
20. Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009) Bayesian phylogeography finds its roots. PLoS Comp Biol 5:1000520. https://doi.org/10.1371/journal.pcbi.1000520
21. Müller NF, Rasmussen D, Stadler T (2018) MASCOT: parameter and state inference under the marginal structured coalescent approximation. Bioinformatics 34:3843–3848
22. De Maio N, Wu C-H, O'Reilly KM, Wilson D (2015) New routes to phylogeography: a bayesian structured coalescent approximation. PLoS Genet 11:e1005421. https://doi.org/10.1371/journal.pgen.1005421
23. Dudas G, Carvalho LM, Rambaut A, Bedford T (2018) MERS-CoV spillover at the camel-human interface. eLife 7:e31257. https://doi.org/10.7554/eLife.31257
24. Bouchez-Zacria M, Courcoul A, Durand B (2018) The distribution of bovine tuberculosis in cattle farms is linked to cattle trade and badger-mediated contact networks in South-Western France, 2007–2015. Front Vet Sci 5:173
25. Bouchez-Zacria M, Courcoul A, Jabert P, Richomme C, Durand B (2017) Environmental determinants of the *Mycobacterium bovis* concomitant infection in cattle and badgers in France. Eur J Wildl Res 63:74. https://doi.org/10.1007/s10344-017-1131-4
26. Paradis E, Schliep K (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35:526–528. https://doi.org/10.1093/bioinformatics/bty633
27. Felsenstein J, Churchill G (1996) A Hidden Markov model approach to variation among sites in rate of evolution. Mol Biol Evol 13:93–104. https://doi.org/10.1093/oxfordjournals.molbev.a025575
28. Hasegawa M, Kishino H, Yano T-A (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174. https://doi.org/10.1007/BF02101694
29. Salvador LCM, O'Brien DJ, Cosgrove MK, Stuber TP, Schooley AM, Crispell J, Church SV, Gröhn YT, Robbe-Austerman S, Kao RR (2019) Disease management at the wildlife-livestock interface: Using whole-genome sequencing to study the role of elk in *Mycobacterium bovis* transmission in Michigan, USA. Mol Ecol 28:2192–2205. https://doi.org/10.1111/mec.15061
30. Bouckaert R, Vaughan TG, Sottani BJ, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N, Matschiner M, Mendes FK, Müller NF, Ogilvie HA, du Plessis L, Popinga A, Rambaut A, Rasmussen D, Siveroni I, Suchard MA, Wu CH, Xie D, Zhang C, Stadler T, Drummond AJ (2019) BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comp Biol 15:6650. https://doi.org/10.1371/journal.pcbi.1006650
31. Maturana P, Brewer BJ, Klaere S, Bouckaert R (2019) Model selection and parameter inference in phylogenetics using nested sampling. Syst Biol 68:219–233
32. Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism, vol 3. Academic Press, New York, pp 21–132
33. Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. Lect Math Life Sci 17:57–86
34. Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ (eds) Evolving Genes and Proteins. Academic Press, New York, pp 97–166
35. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. PLoS Biol 4:e88. https://doi.org/10.1371/journal.pbio.0040088
36. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA (2018) Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst Biol 67:901–904. https://doi.org/10.1093/sysbio/syy032
37. Heled J, Bouckaert RR (2013) Looking for trees in the forest: summary tree from posterior samples. BMC Evol Biol 13:221. https://doi.org/10.1186/1471-2148-13-221
38. Wang L-G, Lam TT-Y, Xu S, Dai Z, Zhou L, Feng T, Guo P, Dunn CW, Jones BR, Bradley T, Zhu H, Guan Y, Jiang Y, Yu G (2020) Treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. Mol Biol Evol 37:599–603. https://doi.org/10.1093/molbev/msz240
39. Yu G (2020) Using ggtree to visualize data on tree-like structures. Curr Protoc Bioinform 69:e96. https://doi.org/10.1002/cpbi.96
40. Louca S, Doebeli M (2018) Efficient comparative phylogenetics on large trees. Bioinformatics 34:1053–1055. https://doi.org/10.1093/bioinformatics/btx701
41. Biek R, O'Hare A, Wright D, Mallon T, McCormick C, Orton RJ, McDowell S, Trewby H, Skuce RA, Kao RR (2012) Whole genome sequencing reveals local transmission patterns of *Mycobacterium bovis* in sympatric cattle and badger populations. PLoS Pathog 8:e1003008. https://doi.org/10.1371/journal.ppat.1003008
42. Trewby H, Wright D, Breadon EL, Lycett SJ, Mallon TR, McCormick C, Johnson P, Orton RJ, Allen AR, Galbraith J, Herzyk P, Skuce RA, Biek R, Kao RR (2016) Use of bacterial whole-genome sequencing to investigate local persistence and spread in bovine tuberculosis. Epidemics 14:26–35. https://doi.org/10.1016/j.epidem.2015.08.003
43. Mestre O, Luo T, Vultos TD, Kremer K, Murray A, Namouchi A, Jackson C, Rauzier J, Bifani P, Warren R, Rasolofo V, Mei J, Gao Q, Gicquel B (2011) Phylogeny of *Mycobacterium tuberculosis* Beijing strains constructed from polymorphisms in genes involved in DNA replication, recombination and repair. PLoS One 6:e16020. https://doi.org/10.1371/journal.pone.0016020
44. Smith NH, Berg S, Dale J, Allen A, Rodriguez S, Romero B, Matos F, Ghebremichael S, Karoui C, Donati C, Machado AdC, Mucavele C, Kazwala RR, Hilty M, Cadmus S, Ngandolo BNR, Habtamu M, Oloya J, Muller A, Milian-Suazo F, Andrievskaia O, Projahn M, Barandiarán S, Macías A, Müller B, Zanini MS, Ikuta CY, Rodriguez CAR, Pinheiro SR, Figueroa A et al (2011) European 1: a globally important clonal complex of *Mycobacterium bovis*. Infect Genet Evol 11:1340–1351. https://doi.org/10.1016/j.meegid.2011.04.027
45. Jacquier M, Vandel J-M, Léger F, Duhayer J, Pardonnet S, Say L, Devillard S, Ruette S (2021) Breaking down population density into different components to better understand its spatial variation. BMC Evol Ecol 21:82. https://doi.org/10.1186/s12862-021-01809-6
46. Rossi G, Crispell J, Brough T, Lycett SJ, White PCL, Allen A, Ellis RJ, Gordon SV, Harwood R, Palkopoulou E, Presho EL, Skuce R, Smith GC, Kao RR

(2022) Phylodynamic analysis of an emergent *Mycobacterium bovis* outbreak in an area with no previously known wildlife infections. J Appl Ecol 59:210–222. https://doi.org/10.1111/1365-2664.14046
47. Donnelly CA, Nouvellet P (2013) The contribution of badgers to confirmed tuberculosis in cattle in high-incidence areas in England. PLoS Curr. https://doi.org/10.1371/currents.outbreaks.097a904d3f3619db2fe78d24bc776098
48. Bouchez-Zacria M (2018) Rôles de l'environnement et des contacts intra et interspécifiques dans la transmission de *Mycobacterium bovis* dans le système bovins-blaireaux en Pyrénées-Atlantiques—Landes. These de doctorat, Université Paris-Saclay (ComUE) (in French)
49. Byrne AW, Quinn JL, O'Keeffe JJ, Green S, Sleeman DP, Martin SW, Davenport J (2014) Large-scale movements in European badgers: has the tail of the movement kernel been underestimated? J Anim Ecol 83:991–1001. https://doi.org/10.1111/1365-2656.12197
50. Podgórski T, Baś G, Jędrzejewska B, Sönnichsen L, Śnieżko S, Jędrzejewski W, Okarma H (2013) Spatiotemporal behavioral plasticity of wild boar (*Sus scrofa*) under contrasting conditions of human pressure: primeval forest and metropolitan area. J Mammal 94:109–119. https://doi.org/10.1644/12-MAMM-A-038.1
51. Rivière J, Le Strat Y, Dufour B, Hendrikx P (2015) Sensitivity of bovine tuberculosis surveillance in wildlife in France: a scenario tree approach. PLoS One 10:e0141884. https://doi.org/10.1371/journal.pone.0141884
52. Müller NF, Rasmussen DA, Stadler T (2017) The structured coalescent and its approximations. Mol Biol Evol 34:2970–2981. https://doi.org/10.1093/molbev/msx186

Publisher's Note

pathogens

MDPI

Systematic Review

Methods Combining Genomic and Epidemiological Data in the Reconstruction of Transmission Trees: A Systematic Review

Hélène Duault [1,2], Benoit Durand [1] and Laetitia Canini [1,*]

[1] Epidemiology Unit, Paris-Est University, Laboratory for Animal Health, French Agency for Food, Environment and Occupational Health and Safety (ANSES), 94700 Maisons-Alfort, France; helene.duault@anses.fr (H.D.); benoit.durand@anses.fr (B.D.)
[2] Faculté de Médecine, Université Paris-Saclay, 94270 Le Kremlin-Bicêtre, France
* Correspondence: laetitia.canini@anses.fr

Abstract: In order to better understand transmission dynamics and appropriately target control and preventive measures, studies have aimed to identify who-infected-whom in actual outbreaks. Numerous reconstruction methods exist, each with their own assumptions, types of data, and inference strategy. Thus, selecting a method can be difficult. Following PRISMA guidelines, we systematically reviewed the literature for methods combing epidemiological and genomic data in transmission tree reconstruction. We identified 22 methods from the 41 selected articles. We defined three families according to how genomic data was handled: a non-phylogenetic family, a sequential phylogenetic family, and a simultaneous phylogenetic family. We discussed methods according to the data needed as well as the underlying sequence mutation, within-host evolution, transmission, and case observation. In the non-phylogenetic family consisting of eight methods, pairwise genetic distances were estimated. In the phylogenetic families, transmission trees were inferred from phylogenetic trees either simultaneously (nine methods) or sequentially (five methods). While a majority of methods (17/22) modeled the transmission process, few (8/22) took into account imperfect case detection. Within-host evolution was generally (7/8) modeled as a coalescent process. These practical and theoretical considerations were highlighted in order to help select the appropriate method for an outbreak.

Keywords: transmission tree; genomic epidemiology; who-infected-whom

Citation: Duault, H.; Durand, B.; Canini, L. Methods Combining Genomic and Epidemiological Data in the Reconstruction of Transmission Trees: A Systematic Review. *Pathogens* **2022**, *11*, 252. https://doi.org/10.3390/pathogens11020252

Received: 9 December 2021
Accepted: 11 February 2022
Published: 15 February 2022

1. Introduction

Understanding transmission dynamics is pivotal in controlling and preventing infectious diseases. Studies have aimed to reconstruct transmission trees depicting transmission histories of actual outbreaks in order to draw conclusions on how the disease spread [1,2]. For instance, transmission trees have been used to explore hypotheses on mechanisms of transmission [3] and evaluate key transmission parameters, such as the reproduction number R, that is, the number of secondary cases caused by one infected individual [4]. In a transmission tree, nodes represent infected hosts (i.e., entities that the pathogen can infect, e.g., individuals or groups of individuals like farms in a foot-and-mouth disease (FMD) outbreak [5]), connected by transmission events represented by directed edges [6]. Transmission events in a transmission tree generally correspond to the first infection of each host, as superinfections (infection with an additional strain) or reinfections (second infection after clearance) are usually disregarded.

One way to infer transmission history in an outbreak has been the use of contact tracing methods, in which infected individuals are interrogated regarding time of symptom onset, duration of disease, and potential exposures to pathogen [7]. However, data collected by epidemiological investigations are not always available, reliable, or detailed enough for accurate reconstructions [8]. In addition, the fact that not all infected individuals are

known hinders the reconstruction of an observed outbreak. Indeed, asymptomatic infected individuals are less likely to be detected unless a testing strategy has been implemented, and even then, test sensitivity (Se) and field conditions are sometimes mediocre. For instance, on-the-field implementation of the intradermal tuberculin skin test for bovine tuberculosis can differ from the official guidelines (e.g., not respecting the recommended injection area, qualitative reading of results), which in turn lowers the Se [9].

Complementary to epidemiological data, pathogen isolation and subsequent partial or total sequencing of pathogen genomes can inform on the relative closeness of strains. The increasing availability and affordability of sequencing contributes to its mounting popularity and its frequent use in molecular epidemiology. Genomic data have been frequently applied to the reconstruction of phylogenetic trees, which describe the evolutionary relationships between sequences [10]. Indeed, numerous methods and tools exist to reconstruct phylogenetic trees (e.g., [11–14]). Some studies have considered phylogenetic trees to be partially observed transmission trees [15]. However, others have highlighted the differences between these two notions [5,16–18]. Contrary to a transmission tree, internal nodes in a phylogenetic tree represent hypothetical common ancestors and tips correspond to sampled sequences, therefore ancestries between sampled sequences cannot be recovered from a phylogenetic tree on its own [16]. Moreover, nodes are linked by branches, which represent genetic distances, and the timing of nodes correspond to within-host diversification events (reconstructed as coalescent events), which precede transmission when considering a complete bottleneck [5,18]. A complete bottleneck means that during infection, only one strain can be transmitted, as opposed to a weak transmission bottleneck that allows multiple strains to be transmitted. Thus, without explicitly representing the hosts in which each pathogen lineage was present, we cannot identify and time transmission events from a phylogenetic tree. Phylogenetic tree reconstruction has been used to identify "transmission clusters", that is, clusters of sequences more closely related in the evolutionary process and therefore considered epidemiologically linked. For instance, a review on HIV "transmission clusters" definitions showed that a majority were based on statistical criteria defining how likely the existence of the node was (phylogenetic node support) or a combination of phylogenetic node support and a genetic distance threshold [19].

However, inferring actual transmission trees solely from genetic data proves challenging. Indeed, genetic diversity between sampled sequences hinges on the evolutionary rate of the pathogen as well as time-to-sampling, and when diversity is limited, the ability to infer correct transmission histories is affected [20]. For example, in a *Mycobacterium bovis* outbreak, a high proportion of sampled sequences isolated from different hosts can be identical [21] due to the low evolutionary rate. While sequenced strains from pathogens that tend to have a high evolutionary rate show greater dissimilarity, a non-negligible within-host diversity and/or a weak bottleneck complicates the inference of the transmission tree solely from genetic data [22]. Thus, methods were developed to combine epidemiological and genomic data, whether in a simultaneous [5,23,24] or sequential (integrating one type of data then the other, e.g., [2,17]) manner to infer possible transmission trees.

According to graph theory, the number of spanning trees that can be constructed from a complete graph of n nodes is given by Cayley's tree formula: n^{n-2} [25]. When applied to transmission trees, this number corresponds to the number of unrooted transmission trees compatible with n hosts. For instance, when considering 10 hosts, 10^8 transmissions trees are compatible. Therefore, simply enumerating all possible oriented transmission trees is not a viable option when n is high and other strategies need to be applied. Methods that combine both epidemiological and genomic data can model four unobserved processes mentioned by Klinkenberg et al. (2017) [26] that can be defined as follows:

- Mutation: includes nucleotide "indel" (either a deletion or an insertion, i.e., a nucleotide disappears from or is added to the sequence) and substitution (a nucleotide in the sequence changes into another).
- Within-host evolution: represents how the pathogen genome changes within an individual or a group of individuals, which leads to genome diversification.

- Transmission: passage of a pathogen from an infected host to a susceptible host and the subsequent infection in the newly infected host. In transmission models, assumptions are thus made regarding how the disease was introduced in the host population then spread from host to host, as well as regarding the natural history of the disease.
- Case observation: process of identifying and sampling infected hosts in the host population.

We systematically reviewed the literature for methods combining genomic and epidemiological data to reconstruct transmission trees. A problem arises from the existence of numerous methods: how to select the appropriate method for the studied outbreak. Therefore, our goal was to discuss methods according to the epidemiological and genomic data necessary to implement them, as well as the underlying sequence mutation, within-host evolution, transmission, and case observation models.

2. Results

After removal of duplicates, 496 articles were imported to EndNote and screened for their relevance to transmission tree reconstruction methodology. Among these 496 articles, the full texts of 98 articles were screened for eligibility (Figure 1). The reasons for exclusion of full-text articles are detailed in Supplementary Table S1. The main reasons were as follows: the article did not actually aim to infer a transmission tree according to our definition (n = 23), the kind of genetic data considered (n = 12), or the lack of a formal combination of the two types of data (n = 21).

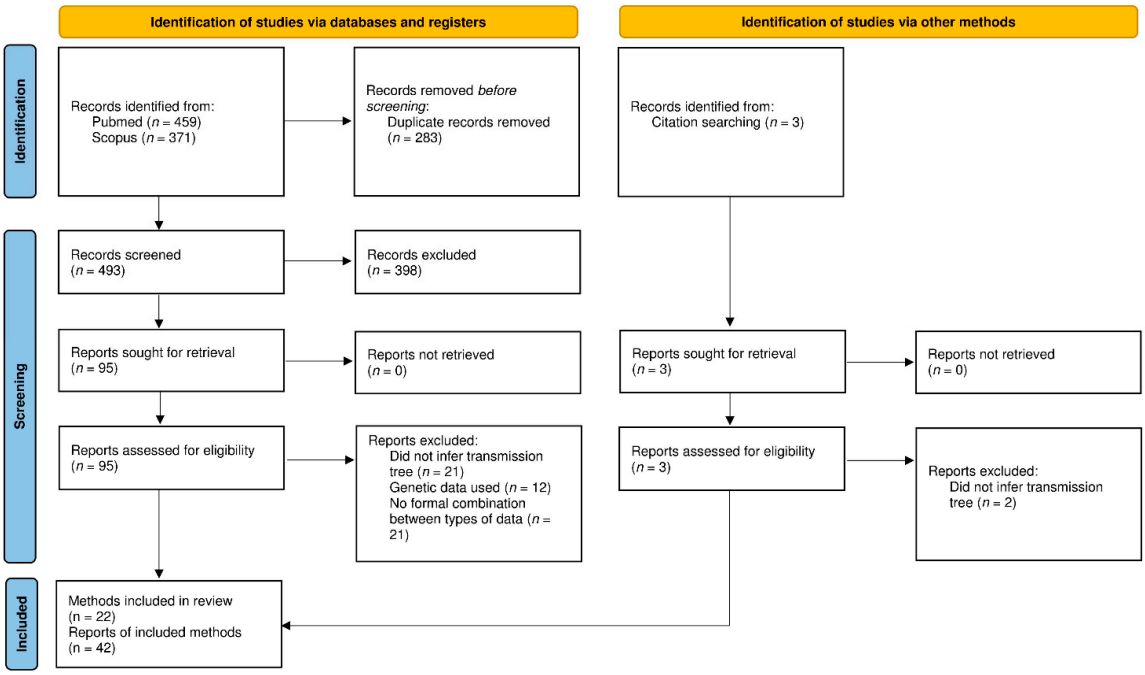


Figure 1. PRISMA flow diagram representing the article selection process (from [27]).

Twenty-two different methods were used in the remaining 41 articles, and we defined three families: a non-phylogenetic family, a sequential phylogenetic family, and a simultaneous phylogenetic family. In the non-phylogenetic family (NPF), phylogenetic trees were not considered in the transmission tree reconstruction, and instead, pairwise genetic distances were estimated. In the phylogenetic families (PF), transmission trees were inferred from phylogenetic trees either by using the phylogenetic tree as a source of information or by establishing a method to link the two types of trees, that is, a transmission tree was obtained by inferring the host of each node or branch in the phylogenetic tree. In the sequential phylogenetic family (SeqPF), phylogenetic trees had to be reconstructed prior to the implementation of the transmission tree reconstruction methods. However, in

the simultaneous phylogenetic family (SimPF), phylogenetic and transmission trees were simultaneously inferred. We decided to distinguish between the two since the two-step approach in the sequential phylogenetic family means the users need to choose an appropriate method to reconstruct the phylogenetic tree and implement it, prior to the transmission tree reconstruction method. Thus, the sequential phylogenetic family assumes that the phylogenetic tree does not depend on the transmission process.

To illustrate the problem these three families tried to address, Figure 2A shows a simplistic transmission and within-host evolution scenario: D transmits to U (an unobserved individual), who in turn transmits to C and A, and finally, C transmits to B. In this figure, the length of each host rectangle represents the time from infection to removal (either recovery or death). From this small outbreak, we consider the sequences a, b, c, and d collected respectively from hosts A, B, C, and D at times T_A, T_B, T_C, and T_D. The removal times of known hosts are also included in the data: R_A, R_B, R_C, and R_D. From the known epidemiological data and either pairwise genetic distances (NPF) or the phylogenetic tree (Figure 2B, PF), each family of methods aimed to reconstruct the transmission tree (Figure 2C), with or without inferring the unknown transmission times t[infector, infected].

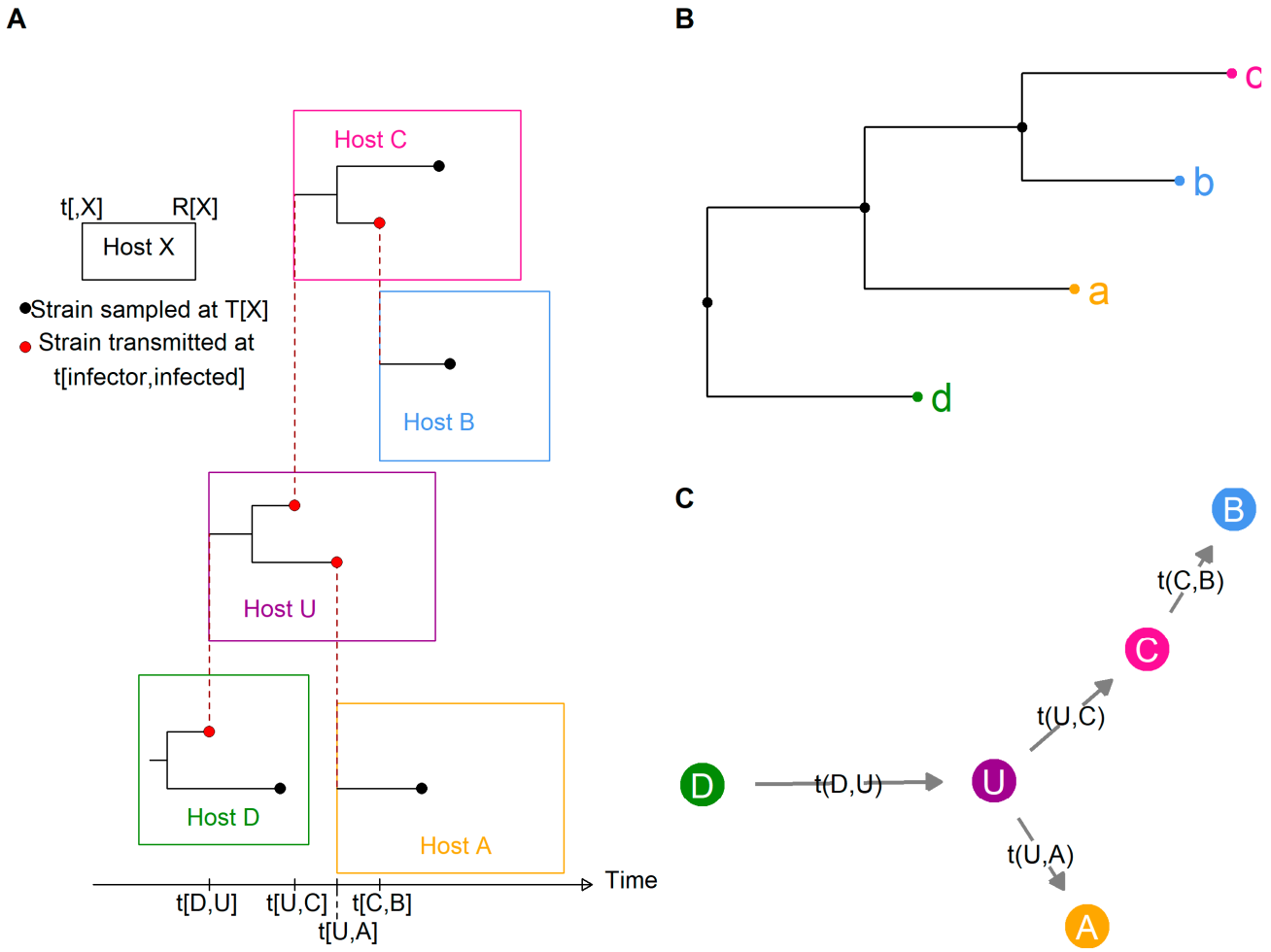


Figure 2. A simple transmission scenario (**A**), the reconstructed phylogenetic tree (**B**), and the transmission tree (**C**). Rectangles represent hosts, and black lines within a rectangle represent within-host evolution of the pathogen. Black circles correspond to sampled strains, red circles to transmitted strains, and red dotted lines to a transmission event. Length of host rectangles represent time from infection (t) to removal (R). The phylogenetic tree is reconstructed from sequences (a, b, c, and d) sampled at time T. The transmission tree considered the unobserved host U.

Table 1 presents the epidemiological and genomic data needed to implement each method. A majority (20/22) of methods used at least sampling times (Table 1). Eleven methods considered removal times, seven the onset of infectiousness, while few (3/22) considered the start of exposure (Table 1). Moreover, intrinsic characteristics (predominant species, number of animals, production period) are only considered in two methods, belonging to the NPF and SimPF, respectively: Aldrin 2011 [28] and BORIS (Bayesian Outbreak Reconstruction Inference and Simulation) [29] (Table 1). Similarly, only two other methods included contact data in their transmission model: one in the NPF, outbreaker2 [30], and one in the SeqPF, TiTUS [31] (Table 1). Ten out of the twenty-two methods (Table 1) were implemented in packages, 7 available as R packages (however, since their implementation, two have been removed from the CRAN repository; for details see Table S6), one code on github (Transmission Tree Uniform Sampler, TiTUS), and the remaining two on BEAST [13] (beastlier) or BEAST2 [14] (Structured COalescent Transmission Tree Inference, SCOTTI).

Within-host evolution was explicitly modeled in fewer than half of the methods (8/22), and most methods made restricting assumptions on the outbreak: all cases are observed and sampled, the transmission bottleneck is complete, or a single introduction event took place (Tables 2–4). Observation is the detection of an infected host, and a host is sampled when a pathogen sequence was isolated.

2.1. Non-Phylogenetic Family

The non-phylogenetic family (Figure 3) contained eight methods. The majority of these methods (5/8) attached a genetic model that described the pairwise genetic distance between two individuals according to their relationship in the transmission tree to an explicit model of disease transmission. In the Bayesian methods (4/5), these models were combined in a likelihood function, which was used to sample from the transmission tree space.

2.1.1. Methods That Consider Mutations to Occur at Transmission

The Bayesian method proposed by Ypma et al. (2012) used three types of data (temporal, geographical, and genetic) from an H7N7 outbreak in poultry farms in the Netherlands and considered them all independent of each other. The likelihood function was therefore a product of contributions given by the three types of data [32]. Similarly, Jombart et al. (2014) decomposed the likelihood into a genetic likelihood and a temporal likelihood in the outbreaker package [24]. Campbell et al. (2019) then extended the transmission model in this method to include contact data in a reporting likelihood in outbreaker2 [30]. Probability of transmission between two sampled individuals was inferred from known generation time Tg and time-to-sampling distributions. In addition, outbreaker and outbreaker2 considered two parameters to model unobserved cases: π, the proportion of sampled cases in the outbreak, and κ, the maximum number of generations separating a sampled infected individual and his sampled ancestor in the transmission tree [24,30]. SARS-CoV-1 [24,30], bovine viral diarrhea virus [33], *Klebsiella pneumoniae* [34], and *Acinetobacter baumannii* [35,36] outbreaks (Table S3) have been studied using outbreaker and outbreaker2, available in R.

In these three methods, mutation was considered to occur during transmission, and thus, the genetic likelihood depended solely on the number of transmission events separating two individuals and not on time [24].

Table 1. Epidemiological and genomic data necessary for each method. S stands for sequences, and P for phylogenetic trees. Packages are available for methods in bold. Removal time corresponds to time at which an individual becomes non-infectious, generally the culling time or end of hospitalization, and intrinsic characteristics are either number of individuals present on site or predominant animal species. Didelot et al.'s (2014) [17] method, while not based on a spatial kernel, penalized transmission trees after reconstruction if they did not respect geographical data, hence the parentheses surrounding the geographical data. Hall et al.'s (2015) [18] method could include contact data, but geographical data was used instead.

Family	Method (*Name*) [Reference]	Start of Exposure	Onset of Infectiousness	Sampling Time	Removal Time	Contact Data	Geographical Data	Intrinsic Characteristics	Phylogenetic Tree or Sequences
Non-phylogenetic	Aldrin et al., 2011 [28]		X		X		X	X	S
	Jombart et al., 2011 (*Seqtrack*) [16]			X					S
	Ypma et al., 2012 [32]		X		X		X		S
	Jombart et al., 2014 (*outbreaker*) [24]			X					S
	Worby et al., 2014 [37]			X					S
	Famulare et al. 2015 [38]			X					S
	Worby et al., 2016 (*bitrugs*) [6]	X		X	X				S
	Campbell et al., 2019 (*outbreaker2*) [30]			X		X			S
Sequential phylogenetic	Cottam et al., 2008 [2]		X	X	X				P
	Didelot et al., 2014 [17]			X			(X)		P
	Eldholm et al., 2016 [39]			X					P
	Didelot et al., 2017 (*Transphylo*) [40]			X					P
	Sashittal et al., 2020 (*TiTUS*) [31]	X		X	X	X			P

Table 1. *Cont.*

Family		Method (*Name*) [Reference]	Start of Exposure	Onset of Infectiousness	Sampling Time	Removal Time	Contact Data	Geographical Data	Intrinsic Characteristics	Phylogenetic Tree or Sequences
Simultaneous phylogenetic	Explicitly phylogenetic	Ypma et al., 2013 [5]		X	X	X		X		S
		Hall et al., 2015 (*beastlier*) [18]			X	X	(X)	X		S
		De Maio et al., 2016 (*SCOTTI*) [41]	X		X	X				S
		Klinkenberg et al., 2017 (*phybreak*) [26]			X					S
	Implicitly phylogenetic	Morelli et al., 2012 [23]		X	X	X		X		S
		Mollentze et al., 2014 [1]			X			X		S
		Lau et al., 2015 [42]		X	X	X		X		S
		Firestone et al., 2020 (*BORIS*) [29]		X	X	X		X	X	S
		Montazeri et al., 2020 [43]			X					S

Table 2. Modeling of unobserved processes in the non-phylogenetic family. Within-host evolution (modeled or not) includes whether the transmission bottleneck is complete or weak. When transmission is modeled, we mention the states hosts can find themselves in (S: susceptible, E: latent, I: infectious, R: removed). In addition, either geographical distance (spatial kernel), contact data, or random mixing are considered. Finally, the transmission model mentions whether there is only one index case possible (single introduction) or multiple.

Method (Name) [Reference]	Sequence Mutation	Within-Host Evolution	Transmission	Case Observation	Inference Method
Aldrin et al., 2011 [28]	Kimura model	No explicit model	SIR (infectious period)	All cases are observed but not always sampled	Partial Maximum Likelihood
		Complete	Distance kernel		
			Multiple		
Jombart et al., 2011 (Seqtrack) [16]	User's choice	No explicit model	No explicit model	All cases are observed and sampled	Edmonds algorithm
		Complete			
Ypma et al., 2012 [32]	Deletion + Transition + Transversion	No explicit model	SEIR (latency/infectious period)	All cases are observed but not always sampled	Bayesian
		Complete	Spatial kernel		
			Single		
Jombart et al., 2014 (outbreaker) [24]	Mutation rate	No explicit model	SI (generation times)	Proportion of sampled cases	Bayesian
		Complete	Random mixing		
			Multiple		
Worby et al., 2014 [37]	Mutation rate	Pathogen population size	No explicit model	All cases are observed and sampled	Observed genetic distance vs. theoretical distribution
		Weak			
Famulare et al., 2015 [38]	Mutation rate	No explicit model	No explicit model	No assumption	Likelihood ratio test + Pruning algorithm
Worby et al., 2016 (bitrugs) [6]	No explicit model	No explicit model	SEIR (latency/infectious period)	Test sensitivity < 1	Bayesian
		No assumption	Random mixing		
			Multiple		
Campbell et al., 2019 (outbreaker2) [30]	Mutation rate	No explicit model	SI (generation times)	Proportion of sampled cases	Bayesian
		Complete	Contact data		
			Multiple		

Table 3. Modeling of unobserved processes in the sequential phylogenetic family. For the sequence mutation process, NA stands for not applicable. Within-host evolution (modeled or not) includes whether the transmission bottleneck is complete or weak. When transmission is modeled, we mention the states hosts can find themselves in (S: susceptible, E: latent, I: infectious, R: removed). In addition, either geographical distance (spatial kernel), contact data, or random mixing are considered. Finally, the transmission model mentions whether there is only one index case possible (single introduction) or multiple. In the inference method, we mention how phylogenetic trees are used to infer transmission trees (either internal nodes or branches are labelled with the host or phylogenetic trees are used as a source of information). * means multiple sequences can be considered per epidemiological unit.

Method (Name) [Reference]	Sequence Mutation	Within-Host Evolution	Transmission	Case Observation	Inference Method
Cottam et al., 2008 [2]	NA	No explicit model	SEIR (latency/infectious period)	All cases are observed and sampled	Label internal nodes
		Complete	Random mixing		Maximum Likelihood
			Single		
Didelot et al., 2014 [17]	NA	Coalescent process	SIR (infectious period)	All cases are observed and sampled	Label branches
		Complete	Random mixing		Bayesian
			Single		
Eldholm et al., 2016 [39]	NA	Coalescent process	SEIR (latency/infectious period)	Probability threshold	Information source
		Complete	Random mixing		Edmonds' algorithm
			Single		
Didelot et al., 2017 (Transphylo) [40]	NA	Coalescent process	SI (generation times)	Proportion of sampled cases	Label branches
		Complete	Random mixing		Bayesian
			Single		
Sashittal et al., 2020 (TiTUS) [31]	NA	No explicit model	No explicit model	All cases are observed and sampled	Label internal nodes
		Weak *			Logical problem

2.1.2. Methods That Allow Within-Host Diversity

Worby et al. (2014) noted that previous methods based their genetic model on strong assumptions, such as a complete transmission bottleneck [24,30] or mutations occurring at time of transmission [24,30,32], thus disregarding within-host diversity. First, they constructed an approximation of the genetic distance distribution and compared it to observed genetic distances in order to determine the probability of direct methicillin-resistant *Staphylococcus aureus* (MRSA) transmission between individuals in a hospital [37]. Then, Worby et al. (2016) incorporated a genetic distance distribution approximation with an explicit transmission model tailored to a nosocomial outbreak, in a Bayesian inference framework, available in a bitrugs package in R [6]. This approach allowed for the within-host diversity previously lacking in other methods while avoiding having to make any assumptions about the within-host evolution process [6], as was necessary in their first work [37]. The transmission model considered a hospital setting, where patients were either susceptible (S) or infectious (I) one day after infection, and transmission rate per infected patient was constant until their discharge. Homogeneous mixing was assumed, meaning that each infected patient had equivalent contact with each susceptible individual. In addition, imperfect case detection was modeled by incorporating test sensibility as a model parameter [6].

Table 4. Modeling of unobserved processes in the simultaneous phylogenetic family. For the sequence mutation process, the user could either use a single substitution model or choose. Within-host evolution (modeled or not) includes whether the transmission bottleneck is complete or weak. When transmission is modeled, we mention the states hosts can find themselves in (S: susceptible, E: latent, I: infectious, R: removed). In addition, either geographical distance (spatial kernel), contact data, or random mixing are considered. Finally, the transmission model mentions whether there is only one index case possible (single introduction) or multiple. * means multiple sequences can be considered per epidemiological unit.

Method (Name) [Reference]	Sequence Mutation	Within-Host Evolution	Transmission	Case Observation	Inference Method
Ypma et al., 2013 [5]	Mutation rate	Coalescent process	SEIR (latency/infectious period)	All cases are observed and sampled	Bayesian
		Complete	Spatial kernel		
			Single		
Hall et al., 2015 (beastlier) [18]	User's choice	Coalescent process	SEIR (latency/infectious period)	All cases are observed but not always sampled	Bayesian
		Complete *	Spatial kernel		
			Single		
De Maio et al., 2016 (SCOTTI) [41]	User's choice	Coalescent process	Migration model	Maximum number of hosts	Bayesian
		Weak *			
Klinkenberg et al., 2017 (phybreak) [26]	Mutation rate	Coalescent process	SI (generation times)	All cases are observed but not always sampled	Bayesian
		Complete	Random mixing		
			Single		
Morelli et al., 2012 [23]	Jukes Cantor model	No explicit model	SEIR (latency/infectious period)	All cases are observed and sampled	Bayesian
		Complete	Spatial kernel		
			Single		
Mollentze et al., 2014 [1]	Kimura model	No explicit model	SEIR (latency/infectious period)	Observed cases contribute to transmission after removal time	Bayesian
		Complete	Spatial kernel		
			Multiple		
Lau et al., 2015 [42]	Kimura model	No explicit model	SEIR (latency/infectious period)	All cases are observed but not always sampled	Bayesian
		Complete	Spatial kernel		
			Multiple		
Firestone et al., 2020 (BORIS) [29]	Kimura model	No explicit model	SEIR (latency/infectious period)	All cases are observed but not always sampled	Bayesian
		Complete	Spatial kernel		
			Multiple		
Montazeri et al., 2020 [43]	Jukes Cantor model	No explicit model	No explicit model	All cases are observed and sampled	Bayesian
		Complete			

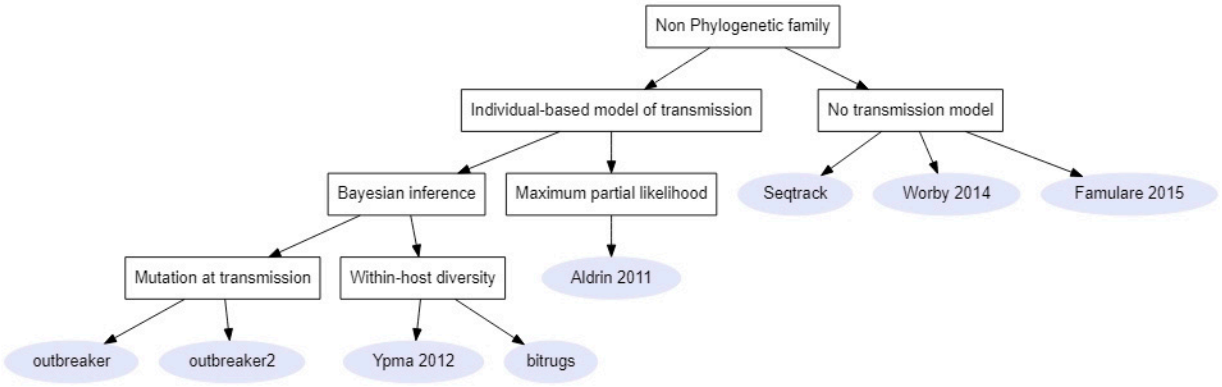


Figure 3. Links between methods of the non-phylogenetic family. Rectangles represent criteria on which to choose a method and the grey circles represent either the name of the method's package or the first author and article date [28,32,37,38].

2.1.3. Other Methods

Conversely, in their work on infectious salmon anemia, Aldrin et al. (2011) did not use the same Bayesian approach. While they did establish a transmission model, a maximum partial likelihood approach was then used to estimate model parameters. From these estimated parameters, they calculated the probability that one salmon farm infected another. In their model, transmission probability exponentially decreased with increasing sea and genetic distances between farms and depended on farm-level characteristics such as the maximum number of fish in a cohort during the production period and when that production period was (spring vs. autumn) [28]. When genetic data was unavailable for a farm, the unknown genetic distance was imputed with the value of a parameter computed from the known genetic data [28].

Finally, the Seqtrack method [16] and Famulare et al. (2015) differ from all the others and only explicitly modeled the mutation process. Indeed, Jombart et al. (2011) computed the transmission tree for which "ancestors always precede [d] their descendants in time" (assuming sampling times follow the same chronological order as infection times) and the total genetic distance between linked nodes was minimal (i.e., the optimum branching, also named minimum spanning tree, of the graph in which all the possible links between infector and infected host are represented) using Edmonds' algorithm [44]. While this method can be used solely with sampling times and genetic data (Table 1), other epidemiological data (e.g., locations) can also be considered to resolve equally likely ancestries. The Seqtrack algorithm was implemented in the adegenet package and has been applied to H1N1 2009 swine-origin pandemic [16], H3N8 equine influenza [45], *M. tuberculosis* [46], and *K. pneumonia* [47] outbreaks (Table S3). Conversely, Famulare et al. identified pairs linked by direct transmission by performing a likelihood ratio test to determine whether the time of the most recent common ancestor of the considered pair (tMRCA) was equal to the earliest sampling time [38]. In order to compute the likelihood for the tMRCA, Famulare et al. assumed the mutation process followed a Poisson model with a known constant mutation rate. Competing ancestries were resolved using a pruning algorithm that the user could specify, for example, by keeping the link minimizing the time between tMCRA and sampling. This method was applied to study the Ebola virus outbreak in Sierra Leone, the 2001 H1N1 influenza pandemic, and the 2005–2008 polio outbreak in Nigeria [38] (Table S3).

2.2. Phylogenetic Families

In phylogenetic families, links were established between phylogenetic and transmission trees. From the small imaginary outbreak (Figure 2), Figure 4 depicts three ways to modify the basic phylogenetic tree (Figure 2A) in order to obtain a transmission tree. Figure 4A shows a phylogenetic tree in which internal nodes are annotated with a sampled host. The transmission tree reconstructed (on the right) from this annotated phylogenetic

tree contains the order of transmission but does not estimate the unknown transmission times t (unless we assume that coalescence and transmission occur at the same time). In Figure 4B, the internal nodes are annotated in the phylogenetic tree (on the left); however, the branch between two nodes hosted by different individuals is considered to be an "infection branch," and transmission occurs along this infection branch. Therefore, we obtain a timed transmission tree (on the right) that does not assume coalescence and transmission to coincide. Finally, in 4C the possibility of annotating unobserved hosts in the phylogenetic tree is added (on the left), thus the unobserved host U can be inferred in the transmission tree (on the right).

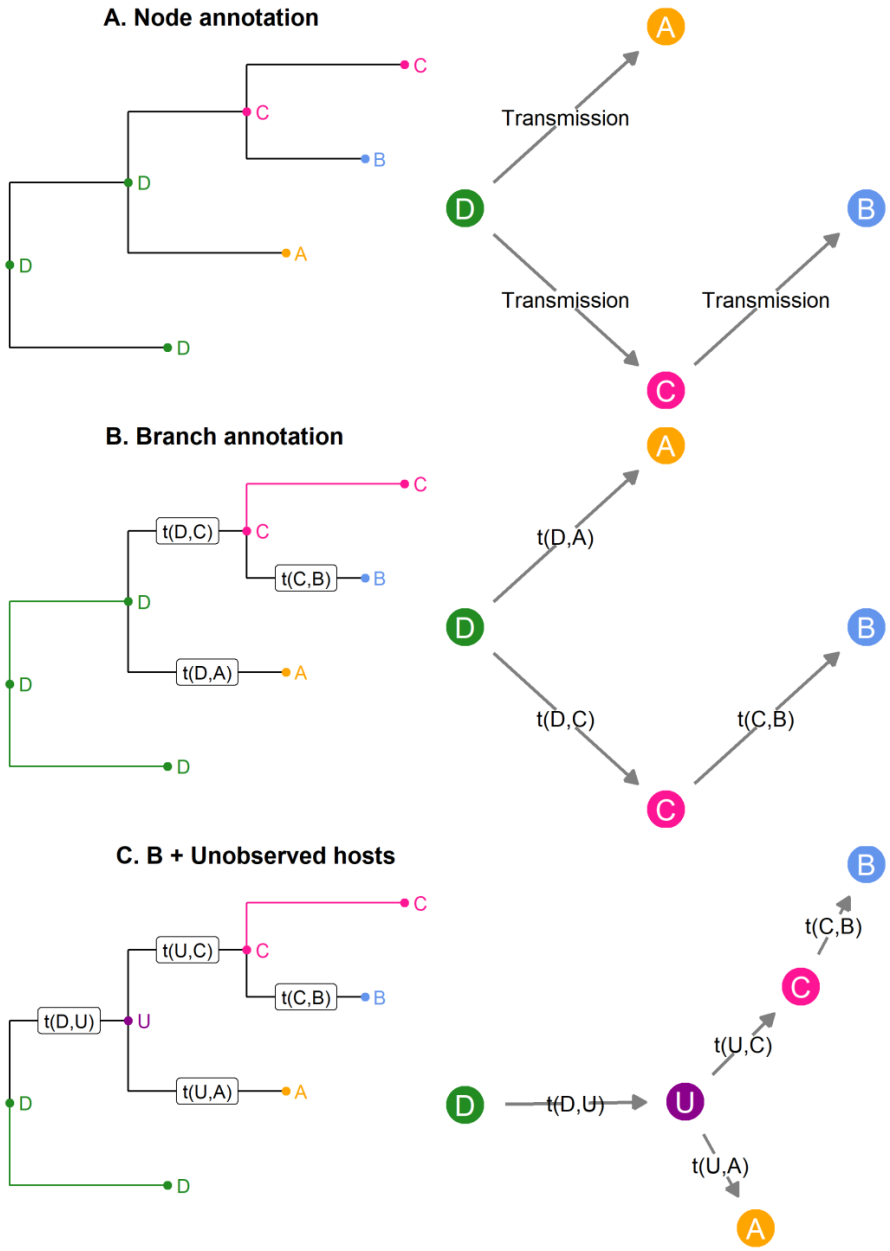


Figure 4. Three links between phylogenetic (on the **left**) and transmission trees (on the **right**). Node annotation with observed hosts (**A**) leads to the identification of transmission links. Annotating the branches (**B**) adds on the time of transmission t. Annotating branches with observed and unobserved hosts (**C**) means the identification of host U is possible.

2.2.1. Sequential Phylogenetic Family

These methods (Figure 5, $n = 5$) required a phylogenetic tree to be reconstructed prior to their implementation. In one method, the phylogenetic tree was used as a source of information on the time of coalescence between two lineages [39]. Indeed, Eldholm et al. (2016) used this information in association with a SEIR model to calculate the likelihood of direct and oriented transmissions between sampled individuals of an *M. tuberculosis* outbreak [39]. When the likelihoods of transmission between every pair of individuals were calculated, the direction of transmission corresponding to the lowest likelihood was removed. Finally, the optimum branching graph was computed using Edmonds' algorithm [44] as in Seqtrack. In order to account for unobserved cases, this method used various thresholds of direct transmission likelihoods to plot the transmission trees [39].

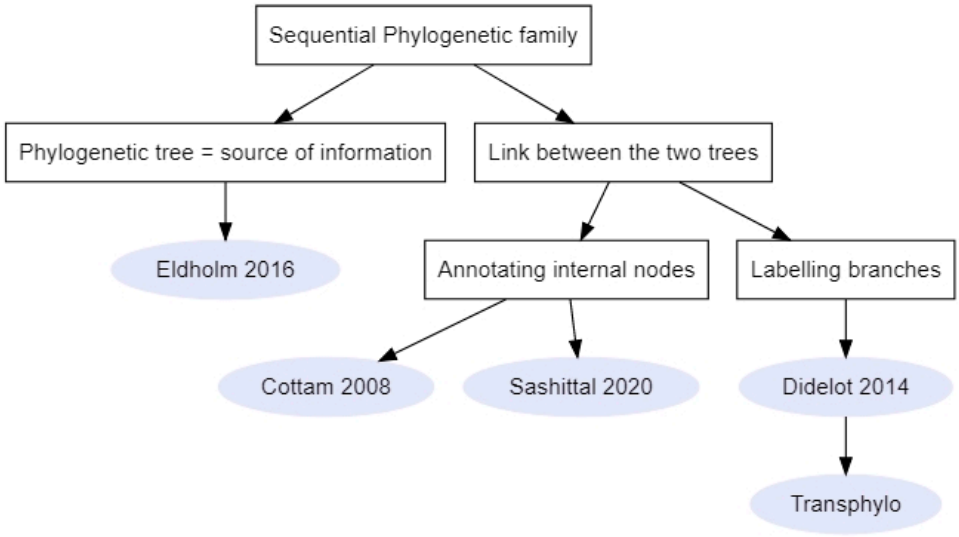


Figure 5. Links between methods of the sequential phylogenetic family. Rectangles represent criteria on which to choose a method and the grey circles represent either the name of the method's package or the first author and article date [2,17,31,39].

The four remaining methods annotated the phylogenetic tree in order to reconstruct the transmission tree using different sampling strategies of the tree space. In Cottam et al.'s (2008) method, no sampling strategy per se was implemented since the number of transmission trees compatible with their data and previous knowledge on transmission events between five farms (identified via animal movements) was relatively small (1728 trees). Every possible transmission tree was enumerated by assigning to every ancestral node one of its two descendants, while moving backwards in time on the phylogeny [2] (node annotation similar to Figure 4A, with added constraints). Then, the likelihood of a transmission tree was computed from the joint likelihood of each transmission pair, which was based on the probability of the epidemiological data (removal dates and onset of infectiousness, Table 1) according to the SEIR transmission model (Table 3). This method was applied to the 2001 FMD outbreak in the United Kingdom (Table S4).

Similarly, Sashittal and El-Kabir (2020) aimed to label the internal nodes in a phylogenetic tree reconstructed from HIV sequences [31] (node annotation similar to Figure 4A). However, in this method, a weak transmission bottleneck was considered. Moreover, the labelling was not restricted to the two descendants of each node. While the transmission process was not explicitly modeled (Table 3), the labelling had to satisfy a number of constraints derived from the known transmission windows (i.e., from exposure time to removal time) and contact information (Table 1). The transmission tree reconstruction was treated as a logical problem and a parsimonious consensus tree was then selected from uniformly sampled transmission trees that satisfied the temporal and contact constraints [31].

Methods that identify transmission events as branching events in a phylogeny and assume a complete bottleneck do not consider within-host evolution [17]. Thus, Didelot et al. (2014) inferred the transmission tree by affecting hosts along branches in the phylogenetic tree [17,40] (branch annotation similar to Figure 4B). Since hosts could change along branches and not only at the nodes, transmission events were no longer restrained to the timing of coalescent events. In their method, Bayesian inference was used to infer the epidemiological parameters of their SIR (Susceptible-Infected-Removed) model, the within-host evolutionary parameters (for which they considered a neutral coalescent process with constant population size Ne and average population generation time g, i.e., duration of the replication cycle), and the transmission tree. Thus, contrary to the complete enumeration in Cottam et al.'s (2008) method and the uniform sampling used in Sashittal and El-Kabir (2020), MCMC (Markov Chain Monte Carlo) sampling was used to explore the transmission tree space. In addition, they used geographical data as well as diagnostic test results to penalize transmission trees [17].

The main limitation of previous methods is the assumption that the outbreak is finished and that all cases were sampled [40]. Didelot et al. (2017) therefore implemented another Bayesian method in an R package called Transphylo, where the user could define the probability for an individual to be sampled and either select the completed or the ongoing outbreak scenario (branch annotation and unobserved hosts similar to Figure 4C). Contrasting with their previous work, the transmission model considered was a branching process [40]. The branching process was defined by a number of offspring distribution (i.e., number of individuals one individual can infect) and a generation time distribution [40]. The Transphylo package was chosen to study (Table S4) bacterial transmission (such as *M. tuberculosis* [48–50] and *K. pneumoniae* outbreaks [51,52]), as well as viral transmission (e.g., part of the recent SARS-CoV-2 pandemic [53] and a large mumps outbreak in Canada [54]). Recently, the Transphylo package [40] was extended to infer transmission trees from multiple phylogenetic trees [55].

None of these transmission tree reconstruction methods explicitly modeled sequence mutation since the method is applied to an already fully reconstructed phylogenetic tree (hence the "not applicable" in Table 3). However, some articles [39,40,48,51,53–55] have used substitution models to reconstruct the phylogenetic tree prior to the implementation of their method.

2.2.2. Simultaneous Phylogenetic Family

Five methods from this family (Figure 6) implicitly considered a phylogenetic tree where internal nodes corresponded to transmission events (node annotation similar to Figure 4A). Morelli et al. (2012) built a likelihood function taking into account correlations between genetic and epidemiological data to study the 2001 and 2007 FMD outbreaks in the United Kingdom [23]. Indeed, the genetic pseudo-likelihood depended on the time from infection to observation and therefore indirectly permitted mutations to occur within the host without explicitly modeling within-host evolution. The transmission model was then extended by Mollentze et al. (2014) to allow multiple introductions of the disease instead of a single index case, which is more suited to endemic situations, and was applied to a canine rabies outbreak in South Africa [1] (Table S5). Both works otherwise used a similar SEIR transmission model (Table 4) and estimated parameters including time-to-infectiousness (or latency period) and time-to-sampling distributions [1,23]. However, Mollentze et al. (2014) indirectly modeled unobserved cases by allowing observed cases to transmit after their removal time and considered two categories of individuals, those that could transmit the virus (dogs) and those that could not [1].

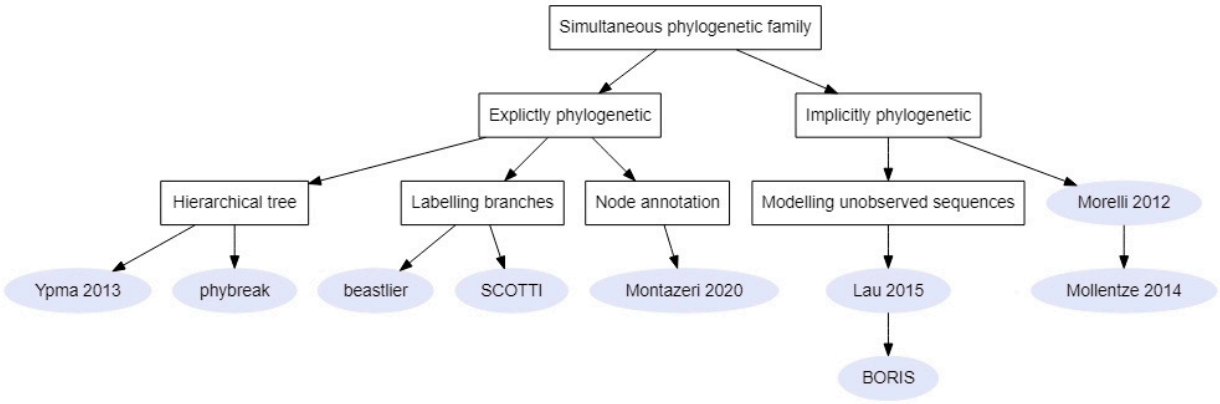


Figure 6. Links between methods of the simultaneous phylogenetic family. Rectangles represent criteria on which to choose a method and the grey circles represent either the name of the method's package or the first author and article date [1,5,23,42,43].

Lau et al. (2015) noted that these previous methods lacked a way to explicitly infer the unobserved transmitted sequences. Indeed, Morelli et al. (2012) and Mollentze et al. (2014) considered a genetic pseudo-likelihood computed for only observed sequences [1,23]. Therefore, Lau et al. (2015) proposed a genuine joint inference by modeling missing genetic data and inferring the unobserved sequences alongside the transmission tree [42]. In their transmission model, two types of infections were considered: primary infections corresponding to imported cases whose sequences were derived from a universal sequence G_M and secondary infections. Secondary infections were modeled according to a SEIR model [42]. Hayama et al. (2019) applied this method to the 2010 FMD outbreak in Japan [56] (Table S5). BORIS is an extension of Lau et al.'s model that incorporates farm-level covariates, such as the number of animals and predominant species (Table 1), which are considered to influence susceptibility and infectiousness of farms in the transmission model [29].

The most recent method from this sub-category did not take into account an explicit transmission model [43]. Montazeri et al. (2020) provided two algorithms that reconstructed the phylogenetic tree from a possible transmission tree by considering estimates of infection times and the absence of within-host diversity. Montazeri et al. applied this method to an HIV transmission cluster in San Diego, California, and the 2014 Ebola virus outbreak in Sierra Leone.

Contrary to these five previous methods, four methods aimed to simultaneously infer phylogenetic and transmission trees. In these four Bayesian methods, a formal link is established between phylogenies and transmission trees and in each MCMC step, both trees are updated in a way that guarantees they remain compatible. Ypma et al. (2013) considered a hierarchical tree where every within-host phylogeny was connected through transmission [26]. They focused on a previously studied 2001 FMD outbreak [2,23] and assumed all infected individuals were known [5] (Table 4). Similarly, Klinkenberg et al.'s (2017) method [26] considered a hierarchical tree. This method was implemented in the R package phybreak and was applied to five published datasets: *M. tuberculosis* [17], MRSA, two FMD outbreaks [2,5,23], and H7N7 [18] (Table S5).

Instead of individually modifying within-host phylogenies as in the hierarchical tree approach, Hall et al.'s (2015) method partitioned the phylogeny by annotating internal nodes with hosts then estimating a parameter for each host to determine their time of infection along the branch (branch annotation similar to Figure 4B) [18]. Hall et al. (2015) studied a 2003 H7N7 outbreak at a farm-level and divided avian farms into two categories ("high-risk" vs. "low-risk"), which differed in the distribution of their infectious period due to the implementation of control measures [18]. Case observation was not modeled, while missing genetic data was replaced by non-informative sequences (repetition of nucleotide "N") [18]. This method implemented in the beastlier package in BEAST [18] was then

applied to a H5N8 avian influenza outbreak (Table S5) with birds as epidemiological units [57].

In these three methods, the transmission process was modeled by epidemiological models previously mentioned in the literature, such as a homogeneous branching process [26] or an individual-based SEIR model, which included a function describing host characteristics affecting transmission, such as geographical distances (via a spatial kernel) [5,18] or risk group [18]. However, De Maio et al. (2016) [41] had an original approach and used the Bayesian structured coalescent approximation (BASTA, [58]). They considered hosts as separate populations characterized by their exposure interval (time from start of exposure to removal, Table 1) and between which pathogens can migrate. This transmission model allowed multiple infections of the same host and transmission of multiple strains during an infection. Therefore, this method implemented in the SCOTTI package [41] in BEAST2 [14] was more suited to outbreaks with frequent mixed infections and large transmission inocula and was applied to FMDV and *K. pneumoniae* outbreaks (Table S5). In addition, this method is the only one in this family (Table 4) that modeled the case observation process (the user could specify a maximum number of hosts in the outbreak) (branch annotation and unobserved hosts similar to Figure 4C).

All these methods explicitly modeled the sequence mutation process with a substitution model, and four out of nine modeled the within-host evolution with a coalescent process (Table 4).

2.3. Application to M. tuberculosis, FMDV, and MRSA Outbreaks

M. tuberculosis is characterized by a low mutation rate (Table S2) coupled with a high proportion of identical sampled sequences [21]. Infection by *M. tuberculosis* can lead to a long latency period, and the majority of cases are asymptomatic. Thus, we should not assume that all cases are observed, and not accounting for the possible long latency could lead to incorrect transmission tree inference. However, the within-host evolution could be disregarded considering the low mutation rate. Methods (included in a package) that allow imperfect case detection are outbreaker and outbreaker2, bitrugs, Transphylo, and SCOTTI (Tables 2 and 4). Among these five methods, Transphylo, outbreaker, and outbreaker2 could allow for a long latency period by selecting an appropriate generation time distribution (Table S6), as has previously been demonstrated with the Gamma generation time density in Transphylo [40]). *M. tuberculosis* outbreaks have been reconstructed using five methods from phylogenetic and non-phylogenetic families: Seqtrack (NPF) [46], Didelot et al. (2014) [17], Eldholm et al. (2016) [39], and Transphylo [40,48–50,55] from the SeqPF and phybreak (SimPF) [26] (Tables S3–S5).

Conversely, FMDV has a high mutation rate (Table S2), and farms are generally the most relevant epidemiological units in an FMDV outbreak. In addition, wind-mediated transmission can play a role in disease spread [3], and pigs shed more than ruminants, who are more susceptible to FMDV [59]. Thus, disregarding within-host evolution seems difficult to justify when the "host" is a farm and the pathogen has a high mutation rate. Moreover, considering the fact that farms have fixed locations and the role played by indirect transmission, it seems unwise to assume random mixing of hosts as well as disregard the information provided by geographical data. Finally, considering the predominant species in the transmission model could help exploit the dissymmetry in roles played by pig and cattle farms. The methods (included in a package) that have an individual-based transmission model with a spatial kernel are BORIS and beastlier (Table 4). However, while BORIS takes into account farm characteristics, beastlier models within-host evolution (Table 4). Seven methods have been applied to FMDV outbreaks: Cottam et al. (2008) (SeqPF) [2], Ypma et al. (2013) [5], SCOTTI [41], phybreak [26], Morelli et al. (2012) [23], Lau et al. (2015) [42,56], and BORIS [29] from the SimPF (Tables S4 and S5).

MRSA has a low mutation rate (Table S2); however, within-host diversity is important to consider when studying *S. aureus* [40]. Studied outbreaks have taken place in neonatal ICUs [6,30,60]. A hospital setting implies a higher proportion of sampled or at least de-

tected cases and multiple possible introductions. Detailed contact data could be available. Therefore, methods used to reconstruct a MRSA outbreak could assume that all cases are observed. However, depending on the outbreak, assuming a single disease introduction could be inappropriate. In addition, contact data would be interesting to consider. Methods (included in a package) that do not assume a single disease introduction are Seqtrack, outbreaker and outbreaker2, bitrugs, and BORIS (Tables 2 and 4). Among these, only bitrugs allows within-host diversity and was specifically designed to study a nosocomial outbreak, while outbreaker2 considers contact data (Table 1). Therefore, the choice between the two methods depends on the type of data available and whether accounting for within-host evolution is necessary to answer our question about the studied outbreak. Bitrugs [6,60] and phybreak [26] were chosen to study MRSA outbreaks in neonatal ICUs (Tables S3 and S5).

3. Discussion

We systematically reviewed the literature for methods combining genomic and epidemiological data to reconstruct transmission trees. The epidemiological data necessary to implement each method was first used to differentiate them. Methods were then divided into three families according to the way genetic data was integrated in the transmission tree inference. We thus differentiated the methods in order to offer practical considerations to examine when selecting transmission tree reconstruction methods.

We were interested in the integration of epidemiological and genetic data in transmission tree inference; however, two methods (Cottam et al. 2008 and Seqtrack) [2,16] were criticized by others for not fully integrating the information provided by both types of data. Even though the possible transmission trees were based on the phylogenetic tree, Cottam et al. (2008) [2] calculated transmission tree likelihood solely from epidemiological data, disregarding any further information that could have been derived from the genetic data [32]. Similarly, Seqtrack [16] only considered additional epidemiological data to distinguish multiple cases when their genetic sequences were identical [32].

The non-phylogenetic family estimated transmission probability from calculated pairwise genetic distances. However, two families used phylogenetic trees to reconstruct transmission trees, either by inferring the host of each node or branch in the phylogenetic tree [2,17,18,31,40,41], considering within-host phylogenetic trees as part of a hierarchical tree [5,26], or by using the phylogenetic tree as a source of information [39]. In the sequential phylogenetic family, phylogenetic trees were reconstructed prior to the implementation of the method and thus called for an additional choice, the phylogenetic tree reconstruction method. Moreover, the phylogenetic tree needs to be correctly reconstructed, or it will lead to errors in the transmission tree. At first, all sequential phylogenetic methods used a single fixed tree generated beforehand by a standard phylogenetic method as an input. As such, these methods ignored any uncertainty in the estimation of the phylogeny [18] and therefore did not take the full uncertainty in the evolutionary process into account [26]. Thus, the Transphylo package was extended to reconstruct transmission trees from multiple phylogenetic trees [55]. However, another strategy was to infer transmission trees and phylogenetic trees simultaneously; we grouped these methods in the simultaneous phylogenetic family.

As mentioned by Klinkenberg et al. (2017), four unobserved processes could be taken into account or ignored [26]: sequence mutation, within-host evolution, transmission, and case observation. Substitution models explicitly model sequence mutation, while genetic distances calculated without a substitution model do not consider intermediary or back mutations and can therefore lead to incorrect estimates. Sequential phylogenetic methods either modeled the mutation process indirectly or did not model it, depending on the method used to pre-generate the phylogenetic tree. Cottam et al. (2008) used a parsimony method [2], while the others [17,39,40] generally opted for Bayesian methods, which supported a number of substitution models. In the two remaining families (non-phylogenetic family and simultaneous phylogenetic family), all methods had the similar option to take into account an explicit substitution model.

Since we expect a non-negligible within-host evolution in infections by pathogens with long generation times [17] combined with a high evolutionary rate, ignoring the fact that mutations occur within-host (e.g., by considering mutations that occur at transmission, such as in outbreaker [18,26]) is inappropriate in this case. In addition, some methods (Morelli et al. 2012, Mollentze et al. 2014, and Lau et al. 2015), while allowing for within-host mutation, only allowed a single pathogen lineage to exist within each host at any given time [18], therefore disregarding any within-host diversity. However, when dealing with a highly sampled outbreak, Ypma et al. stated in 2013 that ignoring within-host diversity's contribution to the observed differences between sampled sequences could lead to incorrect inference of the transmission tree [5]. Methods that modeled within-host evolution generally assimilated it into a coalescent process [5,17,18,26,39–41], which requires the assumption of a low sampling fraction within the host [18]. While this condition is usually verified at an individual scale, it should be kept in mind when reconstructing an outbreak between farms, where the "host" is actually a group of individuals.

Furthermore, farms as epidemiological units could also make it more difficult to disregard within-host population diversity and assume a single infection (multiple introductions are likely to occur), as well as a single within-host pathogen lineage. The reconstructed transmission trees generally considered only the first transmission event, or when it was necessary to account for these secondary transmission events, hosts could simply be duplicated in the transmission tree and infection events were considered independent [24]. Aldrin et al. disregarded completely the possibility of multiple infections of the same farm and chose the least distant genetic data when multiple sequences were available for one farm [28]. The possibility of transmitting genetically diverse strains was overlooked in most methods due to a strong assumption, that is, a transmission bottleneck size of one transmitted sequence [37]. This assumption was relaxed in three methods, Worby et al. (2014), for whom transmission bottleneck size varied [37], and De Maio et al. (2016) and Sashittal et al. (2020), who disregarded transmission bottlenecks completely, allowing the transmission of multiple strains [31,41] and even multiple infections in SCOTTI [41].

While epidemiological models contribute to estimating the most probable transmission tree, a number of underlying assumptions are made on the natural history of the disease and how the disease spread, which need to be considered before choosing a method. For instance, assuming random mixing between hosts means that every infected host is equally likely to infect any susceptible host (used in Didelot et al. 2014, Eldholm et al. 2016, and bitrugs) [6,17,39]. This could be problematic, for example, when considering an FMD outbreak between farms where wind-mediated transmission can play a role in disease spread [3], and thus transmission between farms is no longer equally likely but depends on wind direction and geographical distances. Therefore, some methods have used an individual-based model with a spatial kernel [1,5,18,23,42] or even included farm characteristics influencing infectivity and susceptibility, such as predominant species or herd size (BORIS) [29]. Lastly, considering that the outbreak has a single introduction event is not suited to an endemic situation [1] or even the spread of nosocomial infections in a hospital setting, where multiple introductions can occur [6]. Therefore, some methods did not assume a single disease introduction and either identified genetic outliers [1,24,30] or included disease introduction in the transmission model [6,29,42]. Five methods (Seqtrack, Worby et al. 2014, Famulare et al. 2015, Montazeri et al. 2020, and TiTUS) did not explicitly model transmission.

The final unobserved process to be considered is case observation. According to Didelot et al. (2017), the main limitation of some works preceding the development of Transphylo was the assumption that the outbreak was over and that all cases had been sampled [40]. Indeed, assuming all cases to be linked by direct transmission leads to incorrect estimates on the natural history of the disease or false transmission links. Thus, some methods explicitly modeled case observation by estimating a proportion of observed cases [24,30,40], test sensitivity [6], or the maximum number of hosts in the outbreak [41]. Mollentze et al. (2014) indirectly accounted for unobserved cases by allowing hosts to transmit the pathogen after their removal [1]. Whether the case observation process needs

to be modeled in a transmission tree reconstruction method depends on the possibility of missing infected individuals in the studied outbreak. Therefore, natural disease history, testing strategies, and their effectiveness should be considered.

Moreover, the choice of a method also depends on its availability, as well as its applicability to a wide range of datasets. This can be attested by the number of studies found in our search that reconstructed transmission trees with methods available in packages (e.g., Seqtrack algorithm, $n = 4$ [16,45–47], outbreaker, $n = 4$ [24,34–36], and especially Transphylo, $n = 9$ [40,48–55]) compared to methods like Ypma et al. (2013) and Morelli et al. (2012) [5,23], which are designed for specific datasets and rarely used for other purposes. Unfortunately, computational time was not always available in the selected articles, which makes it difficult to estimate the size of the dataset that can be studied.

Finally, we decided to exclude methods that needed deep-sequencing data. For instance, a Bayesian inference method called BadTrIP (BAyesian epiDemiological TRansmission Inference from Polymorphisms) using genetic and epidemiological data considered a genetic data format (in the form of nucleotide counts for each position in the genome) [61] that greatly differed from the other methods. Another method called SLAFEEL (Statistical Learning Approach For Estimating Epidemiological Links) considered a set of sequences for each host, and epidemiological data was used to calibrate a penalization of the pseudo-likelihood (describing the probability of obtaining the set of sequences in the infected host from the set of sequences present in the infector) [62]. These methods (which do not constitute an exhaustive list) could be interesting to use when multiple sequences are available for a host, when usual model assumptions are unsuitable (SLAFEEL), or when we cannot assume the absence of recombination (BadTrIP).

The choice of a transmission tree reconstruction method thus depends on the characteristics of the pathogen such as mutation rate and natural history of the disease, the epidemiological and genetic data available from the outbreak, as well as the questions we wish to see answered. The impact that violating underlying assumptions of the evolutionary and epidemiological models has on the reconstructed transmission tree, as well as the use of biased data, would be interesting to further investigate.

4. Materials and Methods

4.1. Search Strategy

We searched two electronic databases, Pubmed and Scopus, from 13 October to 17 November 2020. The list of references from the selected studies were screened in order to find further studies to be included. We selected keywords revolving around transmission trees ("transmission chain", "transmission tree", "transmission reconstruction", "transmission network", "who infected whom") and those pertaining to the use of genomic data ("genome", "SNP", "genetic data", "phylogenetic data"). We formulated the following search query: ("transmission chain" OR "transmission tree" OR "transmission reconstruction" OR "transmission network" OR "who infected whom") AND ("genome" OR "genomic" OR "sequence data" OR "genetic data" OR "phylo* data"). Depending on search databases, the search query was entered in "all fields" (Pubmed) or in "Title, abstract, or author-specified keywords" (Scopus). In the database that did not support wild cards (Scopus), "phylo* data" was replaced by "phylogenetic data".

4.2. Eligibility Criteria

Studies were included when they inferred a transmission tree for an infectious disease outbreak using non-simulated epidemiological and genomic data. The genomic data considered was single-nucleotide polymorphisms identified from consensus sequences or the consensus sequences of entire genes themselves and not deep sequencing data, where multiple nucleotides are available for a single locus. We defined a transmission tree as a rooted graph consisting of nodes (representing cases, i.e., infected individuals or groups of individuals) connected by edges (representing transmission events). Transmission trees reconstructed using solely one type of data were excluded. Methods that estimated

possible transmission events compatible with the epidemiological data separately from those compatible with the genomic data were excluded. Even if they graphically combined these transmission events or compared the results obtained by each type of data, in the absence of an algorithm linking the two types of data to reconstruct a transmission tree, we considered them to not formally combine epidemiological and genomic data.

4.3. Data Management

Citations were exported from the two electronic databases to EndNote X9 (2018), where we proceeded to remove duplicates and screened the title, abstract, and when necessary to reach a decision, the material and methods section of the remaining articles. The full texts of selected articles were then assessed for eligibility in chronological order to better understand how the methods relate to one another and their interdependency.

4.4. Data Collection Process

We recorded the inference method (e.g., Bayesian, maximum-likelihood) and the limits of a reconstruction method when they were discussed in an article.

Since genetic diversity affects the ability to reconstruct transmission histories [20], we systematically sought the following information concerning the genomic data. We documented the pathogen, the mutation rate, the number of genetic sequences, and the number of single-nucleotide polymorphisms or the sequence length used to reconstruct the trees, as well as the time period covered. When pathogen mutation rate was not estimated in the article, we searched the literature for this information.

We recorded the epidemiological unit studied, for example, individual or group of individuals. We sought this information because depending on the epidemiological unit, within-host evolution can mean either intra-individual pathogen evolution or intra-group, and therefore incorporate transmission dynamics between individuals within the group considered as a host. Moreover, we identified the type of epidemiological data needed and recorded computational time when available, in order to give practical reasons for method selection. Types of epidemiological data included start of exposure, onset of infectiousness, sampling time, removal time, contact and geographical data, as well as intrinsic characteristics that could influence either infectiousness or susceptibility. For instance, predominant species are intrinsic characteristics of a farm that could be interesting to include in the transmission model of an FMD outbreak [29]. Indeed, pigs shed more virus than ruminants, who are more susceptible; therefore, the most likely pattern of airborne FMDV spread is from pig to cattle and sheep [59].

Finally, we were interested in whether unobserved processes (e.g., mutation, within-host evolution, transmission, and case observation) were explicitly modeled.

1. Substitution models (e.g., Kimura [63] and Jukes Cantor [64] models) are often used to describe sequence mutation. We recorded the type of substitution model used for the sequence mutation.
2. Within-host evolution can be modeled by population models (e.g., the coalescent [65]) that are commonly used in phylogenetic tree reconstruction to describe the ancestry between sampled pathogens. When possible, we recorded the population model describing the within-host evolution.
3. Three sub-categories were considered to describe the transmission model. Since an individual's infectiousness varies over time depending on pathogen shedding [66], transmission models consider different stages of an infectious disease according to transmission potential. Parameters such as latency period and generation time can be fixed beforehand or estimated in the inference. The latency period corresponds to the time from infection by a pathogen to onset of infectiousness and is followed by an infectious period during which the individual can transmit the pathogen to others [67]. Generation times (Tg) represent the time interval between the infection of an index case and the time of transmission from that index case to secondary cases; Tg are related to the latency and infectious periods but also to the variation of an

individual's infectiousness over time [68]. Thus, we identified the different states considered for a host (for instance, S: susceptible, E: exposed, I: infectious, R: removed) and whether latency and infectious periods or generation times were considered to model the natural history of the disease. Moreover, since a transmission event is the result of direct or indirect contact between an infectious individual and a susceptible individual, this contact can be modeled by assuming a random mixing of individuals, considering transmission probability as a function of geographical distances (i.e., a spatial transmission kernel) or taking into account explicit contact data. In our second subcategory, we were interested in how contacts between hosts were modeled (random mixing, spatial kernel, or contact data). Finally, we recorded whether the method assumed that a single introduction of the disease was responsible for the outbreak or if multiple introductions into the host population were possible.

4. For case observation, we were interested in how the methods accounted for imperfect case detection and whether all observed cases were sampled or if the method had a way to handle missing genomic data.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/pathogens11020252/s1. Table S1: Excluded articles and reasons for exclusion. Table S2: Pathogens studied in the selected articles and their estimated mutation rates. Table S3: Pathogen sequences studied by methods in the non-phylogenetic family. Table S4: Pathogen sequences studied by methods in the sequential phylogenetic family. Table S5: Pathogen sequences studied by methods in the simultaneous phylogenetic family. Table S6: Details found in online instruction manuals. References [69–134] are cited in the supplementary materials.

Author Contributions: Conceptualization, L.C., B.D. and H.D.; methodology, L.C., B.D. and H.D.; formal analysis, H.D.; investigation, H.D.; writing—original draft preparation, H.D.; writing—review and editing, L.C. and B.D.; visualization, H.D.; supervision, L.C. and B.D.; funding acquisition, H.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported by Université Paris-Saclay, which funded H.D.'s PhD grant.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mollentze, N.; Nel, L.; Townsend, S.; le Roux, K.; Hampson, K.; Haydon, D.T.; Soubeyrand, S. A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proc. R. Soc. B Boil. Sci.* **2014**, *281*, 20133251. [CrossRef] [PubMed]
2. Cottam, E.M.; Thébaud, G.; Wadsworth, J.; Gloster, J.; Mansley, L.; Paton, D.J.; King, D.P.; Haydon, D.T. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc. R. Soc. B Boil. Sci.* **2008**, *275*, 887–895. [CrossRef] [PubMed]
3. Ypma, R.J.; Jonges, M.; Bataille, A.; Stegeman, A.; Koch, G.; van Boven, M.; Koopmans, M.; van Ballegooijen, W.M.; Wallinga, J. Genetic data provide evidence for wind-mediated transmission of highly pathogenic avian influenza. *J. Infect. Dis.* **2012**, *207*, 730–735. [CrossRef] [PubMed]
4. Faye, O.; Boëlle, P.-Y.; Heleze, E.; Faye, O.; Loucoubar, C.; Magassouba, N.; Soropogui, B.; Keita, S.; Gakou, T.; Bah, E.H.I.; et al. Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: An observational study. *Lancet Infect. Dis.* **2015**, *15*, 320–326. [CrossRef]
5. Ypma, R.J.F.; van Ballegooijen, W.M.; Wallinga, J. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics* **2013**, *195*, 1055–1062. [CrossRef] [PubMed]
6. Worby, C.; O'Neill, P.D.; Kypraios, T.; Robotham, J.; de Angelis, D.; Cartwright, E.J.P.; Peacock, S.J.; Cooper, B. Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *Ann. Appl. Stat.* **2016**, *10*, 395–417. [CrossRef]
7. Varia, M.; Wilson, S.; Sarwal, S.; McGeer, A.; Gournis, E.; Galanis, E.; Henry, B.; Team, H.O.I. Investigation of a nosocomial outbreak of severe acute respiratory syndrome (SARS) in Toronto, Canada. *Can. Med. Assoc. J.* **2003**, *169*, 285–292.
8. Garry, M.; Hope, L.; Zajac, R.; Verrall, A.J.; Robertson, J.M. Contact tracing: A memory task with consequences for public health. *Perspect. Psychol. Sci.* **2021**, *16*, 175–187. [CrossRef]
9. Crozet, G.; Dufour, B.; Rivière, J. Investigation of field intradermal tuberculosis test practices performed by veterinarians in France and factors that influence testing. *Res. Veter. Sci.* **2019**, *124*, 406–416. [CrossRef]

10. Podsiadło, Ł.; Polz-Dacewicz, M. Molecular evolution and phylogenetic implications in clinical research. *Ann. Agric. Environ. Med.* **2013**, *20*, 455–459.
11. Vaz, C.; Nascimento, M.; Carriço, J.A.; Rocher, T.; Francisco, A.P. Distance-based phylogenetic inference from typing data: A unifying view. *Brief. Bioinform.* **2021**, *22*, 147. [CrossRef] [PubMed]
12. Francisco, A.P.; Vaz, C.; Monteiro, P.T.; Melo-Cristino, J.; Ramirez, M.; Carriço, J.A. PHYLOViZ: Phylogenetic inference and data visualization for sequence based typing methods. *BMC Bioinform.* **2012**, *13*, 87. [CrossRef] [PubMed]
13. Suchard, M.A.; Lemey, P.; Baele, G.; Ayres, D.L.; Drummond, A.J.; Rambaut, A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **2018**, *4*, vey016. [CrossRef] [PubMed]
14. Bouckaert, R.; Vaughan, T.G.; Barido-Sottani, J.; Duchêne, S.; Fourment, M.; Gavryushkina, A.; Heled, J.; Jones, G.; Kühnert, D.; de Maio, N.; et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **2019**, *15*, e1006650. [CrossRef] [PubMed]
15. Volz, E.M.; Pond, S.; Ward, M.J.; Brown, A.L.; Frost, S. Phylodynamics of infectious disease epidemics. *Genetics* **2009**, *183*, 1421–1430. [CrossRef]
16. Jombart, T.; Eggo, R.M.; Dodd, P.; Balloux, F. Reconstructing disease outbreaks from genetic data: A graph approach. *Heredity* **2010**, *106*, 383–390. [CrossRef]
17. Didelot, X.; Gardy, J.; Colijn, C. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol. Biol. Evol.* **2014**, *31*, 1869–1879. [CrossRef]
18. Hall, M.; Woolhouse, M.; Rambaut, A. Epidemic reconstruction in a phylogenetics framework: Transmission trees as partitions of the node set. *PLoS Comput. Biol.* **2015**, *11*, e1004613. [CrossRef]
19. Hassan, A.S.; Pybus, O.; Sanders, E.J.; Albert, J.; Esbjörnsson, J. Defining HIV-1 transmission clusters based on sequence data. *AIDS* **2017**, *31*, 1211–1222. [CrossRef]
20. Campbell, F.; Strang, C.; Ferguson, N.; Cori, A.; Jombart, T. When are pathogen genome sequences informative of transmission events? *PLoS Pathog.* **2018**, *14*, e1006885. [CrossRef]
21. Walker, T.M.; Ip, C.L.; Harrell, R.H.; Evans, J.T.; Kapatai, G.; Dedicoat, M.J.; Eyre, D.; Wilson, D.; Hawkey, P.M.; Crook, D.W.; et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: A retrospective observational study. *Lancet Infect. Dis.* **2013**, *13*, 137–146. [CrossRef]
22. Worby, C.J.; Lipsitch, M.; Hanage, W.P. Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. *PLoS Comput. Biol.* **2014**, *10*, e1003549. [CrossRef] [PubMed]
23. Morelli, M.J.; Thébaud, G.; Chadœuf, J.; King, D.P.; Haydon, D.T.; Soubeyrand, S. A Bayesian Inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Comput. Biol.* **2012**, *8*, e1002768. [CrossRef]
24. Jombart, T.; Cori, A.; Didelot, X.; Cauchemez, S.; Fraser, C.; Ferguson, N. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput. Biol.* **2014**, *10*, e1003457. [CrossRef] [PubMed]
25. Cayley, A. A theorem on trees. *Collect. Math. Pap.* **2011**, *23*, 26–28. [CrossRef]
26. Klinkenberg, D.; Backer, J.A.; Didelot, X.; Colijn, C.; Wallinga, J. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput. Biol.* **2017**, *13*, e1005495. [CrossRef]
27. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef]
28. Aldrin, M.; Lyngstad, T.M.; Kristoffersen, A.B.; Storvik, B.; Borgan, Ø.; Jansen, P.A. Modelling the spread of infectious salmon anaemia among salmon farms based on seaway distances between farms and genetic relationships between infectious salmon anaemia virus isolates. *J. R. Soc. Interface* **2011**, *8*, 1346–1356. [CrossRef]
29. Firestone, S.M.; Hayama, Y.; Lau, M.S.Y.; Yamamoto, T.; Nishi, T.; Bradhurst, R.A.; Demirhan, H.; Stevenson, M.A.; Tsutsui, T. Transmission network reconstruction for foot-and-mouth disease outbreaks incorporating farm-level covariates. *PLoS ONE* **2020**, *15*, e0235660. [CrossRef]
30. Campbell, F.; Cori, A.; Ferguson, N.; Jombart, T. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLoS Comput. Biol.* **2019**, *15*, e1006930. [CrossRef]
31. Sashittal, P.; El-Kebir, M. Sampling and summarizing transmission trees with multi-strain infections. *Bioinformatics* **2020**, *36* (Suppl. S1), i362–i370. [CrossRef] [PubMed]
32. Ypma, R.J.F.; Bataille, A.; Stegeman, A.; Koch, G.; Wallinga, J.; van Ballegooijen, W.M. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proc. R. Soc. B Boil. Sci.* **2011**, *279*, 444–450. [CrossRef] [PubMed]
33. Cerutti, F.; Luzzago, C.; Lauzi, S.; Ebranati, E.; Caruso, C.; Masoero, L.; Moreno, A.; Acutis, P.L.; Zehender, G.; Peletto, S. Phylogeography, phylodynamics and transmission chains of bovine viral diarrhea virus subtype 1f in Northern Italy. *Infect. Genet. Evol.* **2016**, *45*, 262–267. [CrossRef] [PubMed]
34. Stoesser, N.; Giess, A.; Batty, L.; Sheppard, A.; Walker, A.S.; Wilson, D.; Didelot, X.; Bashir, A.; Sebra, R.; Kasarskis, A.; et al. Genome sequencing of an extended series of NDM-producing *Klebsiella pneumoniae* isolates from neonatal infections in a Nepali hospital characterizes the extent of community- versus hospital-associated transmission in an endemic setting. *Antimicrob. Agents Chemother.* **2014**, *58*, 7347–7357. [CrossRef]

35. Kanamori, H.; Parobek, C.; Weber, D.J.; van Duin, D.; Rutala, W.A.; Cairns, B.A.; Juliano, J.J. Next-generation sequencing and comparative analysis of sequential outbreaks caused by multidrug-resistant Acinetobacter baumannii at a large academic burn center. *Antimicrob. Agents Chemother.* **2016**, *60*, 1249–1257. [CrossRef] [PubMed]
36. Makke, G.; Bitar, I.; Salloum, T.; Panossian, B.; Alousi, S.; Arabaghian, H.; Medvecky, M.; Hrabak, J.; Merheb-Ghoussoub, S.; Tokajian, S. Whole-genome-sequence-based characterization of extensively drug-resistant *Acinetobacter baumannii* hospital outbreak. *mSphere* **2020**, *5*, e00934-19. [CrossRef]
37. Worby, C.J.; Chang, H.-H.; Hanage, W.P.; Lipsitch, M. The distribution of pairwise genetic distances: A tool for investigating disease transmission. *Genetics* **2014**, *198*, 1395–1404. [CrossRef]
38. Famulare, M.; Hu, H. Extracting transmission networks from phylogeographic data for epidemic and endemic diseases: Ebola virus in Sierra Leone, 2009 H1N1 pandemic influenza and polio in Nigeria. *Int. Health* **2015**, *7*, 130–138. [CrossRef]
39. Eldholm, V.; Rieux, A.; Monteserin, J.; Lopez, J.M.; Palmero, D.; Lopez, B.; Ritacco, V.; Didelot, X.; Balloux, F. Impact of HIV co-infection on the evolution and transmission of multidrug-resistant tuberculosis. *eLife* **2016**, *5*, e16644. [CrossRef]
40. Didelot, X.; Fraser, C.; Gardy, J.; Colijn, C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol. Biol. Evol.* **2017**, *34*, 997–1007. [CrossRef]
41. De Maio, N.; Wu, C.-H.; Wilson, D. SCOTTI: Efficient reconstruction of transmission within outbreaks with the structured coalescent. *PLoS Comput. Biol.* **2016**, *12*, e1005130. [CrossRef] [PubMed]
42. Lau, M.S.Y.; Marion, G.; Streftaris, G.; Gibson, G. A systematic Bayesian integration of epidemiological and genetic data. *PLoS Comput. Biol.* **2015**, *11*, e1004633. [CrossRef] [PubMed]
43. Montazeri, H.; Little, S.; Mozaffarilegha, M.; Beerenwinkel, N.; DeGruttola, V. Bayesian reconstruction of transmission trees from genetic sequences and uncertain infection times. *Stat. Appl. Genet. Mol. Biol.* **2020**, *19*, 1–13. [CrossRef] [PubMed]
44. Edmonds, J. Optimum branchings. *J. Res. Natl. Bur. Stand. Sect. B Math. Math. Phys.* **1967**, *71B*, 233. [CrossRef]
45. Hughes, J.; Allen, R.C.; Baguelin, M.; Hampson, K.; Baillie, G.J.; Elton, D.; Newton, J.R.; Kellam, P.; Wood, J.L.N.; Holmes, E.C.; et al. Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks. *PLoS Pathog.* **2012**, *8*, e1003081. [CrossRef] [PubMed]
46. Guerra-Assunção, J.A.; Crampin, A.; Houben, R.M.G.J.; Mzembe, T.; Mallard, K.; Coll, F.; Khan, P.; Banda, L.; Chiwaya, A.; Pereira, R.P.A.; et al. Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. *eLife* **2015**, *4*, e05166. [CrossRef]
47. Spencer, M.D.; Winglee, K.; Passaretti, C.; Earl, A.M.; Manson, A.L.; Mulder, H.P.; Sautter, R.L.; Fodor, A.A. Whole genome sequencing detects inter-facility transmission of carbapenem-resistant Klebsiella pneumoniae. *J. Infect.* **2019**, *78*, 187–199. [CrossRef]
48. Séraphin, M.N.; Didelot, X.; Nolan, D.J.; May, J.R.; Khan, S.R.; Murray, E.R.; Salemi, M.; Morris, J.G., Jr.; Lauzardo, M. Genomic investigation of a *Mycobacterium tuberculosis* outbreak involving prison and community cases in Florida, United States. *Am. J. Trop. Med. Hyg.* **2018**, *99*, 867–874. [CrossRef]
49. Xu, Y.; Cancino-Muñoz, I.; Torres-Puente, M.; Villamayor, L.M.; Borrás, R.; Borrás-Máñez, M.; Bosque, M.; Camarena, J.J.; Colomer-Roig, E.; Colomina, J.; et al. High-resolution mapping of tuberculosis transmission: Whole genome sequencing and phylogenetic modelling of a cohort from Valencia Region, Spain. *PLoS Med.* **2019**, *16*, e1002961. [CrossRef]
50. Sobkowiak, B.; Banda, L.; Mzembe, T.; Crampin, A.C.; Glynn, J.R.; Clark, T. Bayesian reconstruction of Mycobacterium tuberculosis transmission networks in a high incidence area over two decades in Malawi reveals associated risk factors and genomic variants. *Microb. Genom.* **2020**, *6*, mgen000361. [CrossRef]
51. Kwong, J.; Lane, C.R.; Romanes, F.; da Silva, A.G.; Easton, M.; Cronin, K.; Waters, M.J.; Tomita, T.; Stevens, K.; Schultz, M.; et al. Translating genomics into practice for real-time surveillance and response to carbapenemase-producing Enterobacteriaceae: Evidence from a complex multi-institutional KPC outbreak. *PeerJ* **2018**, *6*, e4210. [CrossRef]
52. Van Dorp, L.; Wang, Q.; Shaw, L.P.; Acman, M.; Brynildsrud, O.; Eldholm, V.; Wang, R.; Gao, H.; Yin, Y.; Chen, H.; et al. Rapid phenotypic evolution in multidrug-resistant *Klebsiella pneumoniae* hospital outbreak strains. *Microb. Genom.* **2019**, *5*, e000263. [CrossRef] [PubMed]
53. Wang, L.; Didelot, X.; Yang, J.; Wong, G.; Shi, Y.; Liu, W.; Gao, G.F.; Bi, Y. Inference of person-to-person transmission of COVID-19 reveals hidden super-spreading events during the early outbreak phase. *Nat. Commun.* **2020**, *11*, 5006. [CrossRef] [PubMed]
54. Stapleton, P.J.; Eshaghi, A.; Seo, C.Y.; Wilson, S.; Harris, T.; Deeks, S.L.; Bolotin, S.; Goneau, L.W.; Gubbay, J.B.; Patel, S.N. Evaluating the use of whole genome sequencing for the investigation of a large mumps outbreak in Ontario, Canada. *Sci. Rep.* **2019**, *9*, 1–11. [CrossRef] [PubMed]
55. Xu, Y.; Stockdale, J.E.; Naidu, V.; Hatherell, H.; Stimson, J.; Stagg, H.R.; Abubakar, I.; Colijn, C. Transmission analysis of a large tuberculosis outbreak in London: A mathematical modelling study using genomic data. *Microb. Genom.* **2020**, *6*, e000450. [CrossRef]
56. Hayama, Y.; Firestone, S.M.; Stevenson, M.A.; Yamamoto, T.; Nishi, T.; Shimizu, Y.; Tsutsui, T. Reconstructing a transmission network and identifying risk factors of secondary transmissions in the 2010 foot-and-mouth disease outbreak in Japan. *Transbound. Emerg. Dis.* **2019**, *66*, 2074–2086. [CrossRef] [PubMed]
57. Choi, S.C. Inferring transmission routes of avian influenza during the H5N8 outbreak of South Korea in 2014 using epidemiological and genetic data. *Korean J. Microbiol.* **2018**, *54*, 254–265. [CrossRef]

58. De Maio, N.; Wu, C.-H.; O'Reilly, K.; Wilson, D.J. New routes to phylogeography: A Bayesian structured coalescent approximation. *PLoS Genet.* **2015**, *11*, e1005421. [CrossRef]
59. Alexandersen, S.; Zhang, Z.; Donaldson, A.; Garland, A. The pathogenesis and diagnosis of foot-and-mouth disease. *J. Comp. Pathol.* **2003**, *129*, 1–36. [CrossRef]
60. Azarian, T.; Maraqa, N.F.; Cook, R.L.; Johnson, J.A.; Bailey, C.; Wheeler, S.; Nolan, D.; Rathore, M.H.; Morris, J.G., Jr.; Salemi, M. Genomic epidemiology of methicillin-resistant *Staphylococcus aureus* in a neonatal intensive care unit. *PLoS ONE* **2016**, *11*, e0164397. [CrossRef]
61. De Maio, N.; Worby, C.; Wilson, D.; Stoesser, N. Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLoS Comput. Biol.* **2018**, *14*, e1006117. [CrossRef]
62. Alamil, M.; Hughes, J.; Berthier, K.; Desbiez, C.; Thébaud, G.; Soubeyrand, S. Inferring epidemiological links from deep sequencing data: A statistical learning approach for human, animal and plant diseases. *Philos. Trans. R. Soc. B Biol. Sci.* **2019**, *374*, 20180258. [CrossRef]
63. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **1980**, *16*, 111–120. [CrossRef] [PubMed]
64. Jukes, T.H.; Cantor, C.R. Evolution of protein molecules. In *Mammalian Protein Metabolism*; Munro, H.N., Ed.; Academic Press: New York, NY, USA, 1969; Volume 3, pp. 21–132.
65. Kingman, J. The coalescent. *Stoch. Process. Appl.* **1982**, *13*, 235–248. [CrossRef]
66. Woolhouse, M. Quantifying transmission. *Microbiol. Spectr.* **2017**, *5*, 279–289. [CrossRef]
67. Van Seventer, J.M.; Hochberg, N.S. Principles of infectious diseases: Transmission, diagnosis, prevention, and control. *Int. Encycl. Public Health* **2017**, *6*, 22–39. [CrossRef]
68. Svensson, Å. A note on generation times in epidemic models. *Math. Biosci.* **2007**, *208*, 300–311. [CrossRef] [PubMed]
69. Scaduto, D.I.; Brown, J.; Haaland, W.C.; Zwickl, D.J.; Hillis, D.M.; Metzker, M.L. Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 21242–21247. [CrossRef]
70. Schürch, A.; Kremer, K.; Daviena, O.; Kiers, A.; Boeree, M.J.; Siezen, R.J.; van Soolingen, D. High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. *J. Clin. Microbiol.* **2010**, *48*, 3403–3406. [CrossRef]
71. Shiino, T. Phylodynamic analysis of a viral infection network. *Front. Microbiol.* **2012**, *3*, 278. [CrossRef]
72. Zarrabi, N.; Prosperi, M.; Belleman, R.G.; Colafigli, M.; de Luca, A.; Sloot, P. Combining epidemiological and genetic networks signifies the importance of early treatment in HIV-1 transmission. *PLoS ONE* **2012**, *7*, e46156. [CrossRef] [PubMed]
73. Alam, S.J.; Zhang, X.; Romero-Severson, E.O.; Henry, C.; Zhong, L.; Volz, E.; Brenner, B.G.; Koopman, J.S. Detectable signals of episodic risk effects on acute HIV transmission: Strategies for analyzing transmission systems using genetic data. *Epidemics* **2013**, *5*, 44–55. [CrossRef] [PubMed]
74. Stack, J.C.; Murcia, P.R.; Grenfell, B.T.; Wood, J.L.N.; Holmes, E.C. Inferring the inter-host transmission of influenza A virus using patterns of intra-host genetic variation. *Proc. R. Soc. B Boil. Sci.* **2013**, *280*, 20122173. [CrossRef] [PubMed]
75. Gavryushkina, A.; Welch, D.; Stadler, T.; Drummond, A.J. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Comput. Biol.* **2014**, *10*, e1003919. [CrossRef]
76. Mehaffy, C.; Guthrie, J.; Alexander, D.C.; Stuart, R.; Rea, E.; Jamieson, F.B. Marked microevolution of a unique mycobacterium tuberculosis strain in 17 years of ongoing transmission in a high risk population. *PLoS ONE* **2014**, *9*, e112928. [CrossRef]
77. Numminen, E.; Chewapreecha, C.; Sirén, J.; Turner, C.; Turner, P.; Bentley, S.D.; Corander, J. Two-phase importance sampling for inference about transmission trees. *Proc. R. Soc. B Boil. Sci.* **2014**, *281*, 20141324. [CrossRef] [PubMed]
78. Croucher, N.; Didelot, X. The application of genomics to tracing bacterial pathogen transmission. *Curr. Opin. Microbiol.* **2015**, *23*, 62–67. [CrossRef]
79. Janies, D.A.; Pomeroy, L.W.; Krueger, C.; Zhang, Y.; Senturk, I.; Kaya, K.; Çatalyürek, Ü.V. Phylogenetic visualization of the spread of H7 influenza A viruses. *Cladistics* **2015**, *31*, 679–691. [CrossRef]
80. Valdazo-González, B.; Kim, J.T.; Soubeyrand, S.; Wadsworth, J.; Knowles, N.J.; Haydon, D.T.; King, D.P. The impact of within-herd genetic variation upon inferred transmission trees for foot-and-mouth disease virus. *Infect. Genet. Evol.* **2015**, *32*, 440–448. [CrossRef]
81. Folarin, O.A.; Ehichioya, D.; Schaffner, S.F.; Winnicki, S.M.; Wohl, S.; Eromon, P.; West, K.L.; Gladden-Young, A.; Oyejide, N.E.; Matranga, C.; et al. Ebola virus epidemiology and evolution in Nigeria. *J. Infect. Dis.* **2016**, *214*, S102–S109. [CrossRef]
82. Hall, M.; Woolhouse, M.; Rambaut, A. Using genomics data to reconstruct transmission trees during disease outbreaks. *Rev. Sci. Tech. Int. Off. Epizoot.* **2016**, *35*, 287–296. [CrossRef] [PubMed]
83. Hoffmann, M.; Luo, Y.; Monday, S.R.; Gonzalez-Escalona, N.; Ottesen, A.R.; Muruvanda, T.; Wang, C.; Kastanis, G.; Keys, C.; Janies, D.; et al. Tracing origins of the *Salmonella bareilly* strain causing a food-borne outbreak in the United States. *J. Infect. Dis.* **2015**, *213*, 502–508. [CrossRef] [PubMed]
84. Kenah, E.; Britton, T.; Halloran, M.E.; Longini, I.M. Molecular infectious disease epidemiology: Survival analysis and algorithms linking phylogenies to transmission trees. *PLoS Comput. Biol.* **2016**, *12*, e1004869. [CrossRef]
85. Parratt, S.R.; Numminen, E.; Laine, A.-L. Infectious disease dynamics in heterogeneous landscapes. *Annu. Rev. Ecol. Evol. Syst.* **2016**, *47*, 283–306. [CrossRef]
86. Ray, B.; Ghedin, E.; Chunara, R. Network inference from multimodal data: A review of approaches from infectious disease transmission. *J. Biomed. Inform.* **2016**, *64*, 44–54. [CrossRef] [PubMed]

87. Ren, H.; Jin, Y.; Hu, M.; Zhou, J.; Song, T.; Huang, Z.; Li, B.; Li, K.; Zhou, W.; Dai, H.; et al. Ecological dynamics of influenza A viruses: Cross-species transmission and global migration. *Sci. Rep.* **2016**, *6*, 36839. [CrossRef]
88. Wohl, S.; Schaffner, S.F.; Sabeti, P.C. Genomic analysis of viral outbreaks. *Annu. Rev. Virol.* **2016**, *3*, 173–195. [CrossRef]
89. Xu, W.; Berhane, Y.; Dubé, C.; Liang, B.; Pasick, J.; van Domselaar, G.; Alexandersen, S. Epidemiological and evolutionary inference of the transmission network of the 2014 highly pathogenic avian influenza H5N2 outbreak in British Columbia, Canada. *Sci. Rep.* **2016**, *6*, 30858. [CrossRef]
90. Agoti, C.N.; Munywoki, P.K.; Phan, M.V.T.; Otieno, J.R.; Kamau, E.; Bett, A.; Kombe, I.; Githinji, G.; Medley, G.F.; Cane, P.A.; et al. Transmission patterns and evolution of respiratory syncytial virus in a community outbreak identified by genomic analysis. *Virus Evol.* **2017**, *3*, vex006. [CrossRef]
91. Baele, G.; Suchard, M.A.; Rambaut, A.; Lemey, P. Emerging concepts of data integration in pathogen phylodynamics. *Syst. Biol.* **2016**, *66*, e47–e65. [CrossRef]
92. Glebova, O.; Knyazev, S.; Melnyk, A.; Artyomenko, A.; Khudyakov, Y.; Zelikovsky, A.; Skums, P. Inference of genetic relatedness between viral quasispecies from sequencing data. *BMC Genom.* **2017**, *18*, 918. [CrossRef]
93. Snitkin, E.S.; Won, S.; Pirani, A.; Lapp, Z.; Weinstein, R.A.; Lolans, K.; Hayden, M.K. Integrated genomic and interfacility patient-transfer data reveal the transmission pathways of multidrug-resistant *Klebsiella pneumoniae* in a regional outbreak. *Sci. Transl. Med.* **2017**, *9*, eaan0093. [CrossRef]
94. Worby, C.J.; Lipsitch, M.; Hanage, W.P. Shared genomic variants: Identification of transmission routes using pathogen deep-sequence data. *Am. J. Epidemiol.* **2017**, *186*, 1209–1216. [CrossRef] [PubMed]
95. Campbell, F.; Didelot, X.; Fitzjohn, R.; Ferguson, N.; Cori, A.; Jombart, T. Outbreaker2: A modular platform for outbreak reconstruction. *BMC Bioinform.* **2018**, *19*, 17–24. [CrossRef]
96. Choi, S.C. Genomic epidemiology for microbial evolutionary studies and the use of Oxford Nanopore sequencing technology. *Korean J. Microbiol.* **2018**, *54*, 188–199. [CrossRef]
97. Ezeoke, I.; Galac, M.R.; Lin, Y.; Liem, A.T.; Roth, P.A.; Kilianski, A.; Gibbons, H.S.; Bloch, D.; Kornblum, J.; del Rosso, P.; et al. Tracking a serial killer: Integrating phylogenetic relationships, epidemiology, and geography for two invasive meningococcal disease outbreaks. *PLoS ONE* **2018**, *13*, e0202615. [CrossRef] [PubMed]
98. Gotoh, Y.; Taniguchi, T.; Yoshimura, D.; Katsura, K.; Saeki, Y.; Hirabara, Y.; Fukuda, M.; Takajo, I.; Tomida, J.; Kawamura, Y.; et al. Multi-step genomic dissection of a suspected intra-hospital *Helicobacter cinaedi* outbreak. *Microb. Genom.* **2019**, *5*, 1–11. [CrossRef]
99. Kendall, M.; Ayabina, D.; Xu, Y.; Stimson, J.; Colijn, C. Estimating transmission from genetic and epidemiological data: A metric to compare transmission trees. *Stat. Sci.* **2018**, *33*, 70–85. [CrossRef]
100. Leitner, T.; Romero-Severson, E. Phylogenetic patterns recover known HIV epidemiological relationships and reveal common transmission of multiple variants. *Nat. Microbiol.* **2018**, *3*, 983–988. [CrossRef]
101. Meehan, C.J.; Moris, P.; Kohl, T.A.; Pečerska, J.; Akter, S.; Merker, M.; Utpatel, C.; Beckert, P.; Gehre, F.; Lempens, P.; et al. The relationship between transmission time and clustering methods in *Mycobacterium tuberculosis* epidemiology. *EBioMedicine* **2018**, *37*, 410–416. [CrossRef]
102. Payne, D.C.; Biggs, H.M.; Al-Abdallat, M.M.; Alqasrawi, S.; Lu, X.; Abedi, G.R.; Haddadin, A.; Iblan, I.; Alsanouri, T.; Al Nsour, M.; et al. Multihospital outbreak of a middle east respiratory syndrome coronavirus deletion variant, Jordan: A molecular, serologic, and epidemiologic investigation. *Open Forum Infect. Dis.* **2018**, *5*, ofy095. [CrossRef] [PubMed]
103. Blackburn, R.M.; Frampton, D.; Smith, C.M.; Fragaszy, E.B.; Watson, S.J.; Ferns, R.B.; Binter, S.; Coen, P.G.; Grant, P.; Shallcross, L.; et al. Nosocomial transmission of influenza: A retrospective cross-sectional study using next generation sequencing at a hospital in England (2012–2014). *Influ. Other Respir. Viruses* **2019**, *13*, 556–563. [CrossRef]
104. DeSilva, M.B.; Styles, T.; Basler, C.; Moses, F.L.; Husain, F.; Reichler, M.; Whitmer, S.; McAuley, J.; Belay, E.; Friedman, M.; et al. Introduction of Ebola virus into a remote border district of Sierra Leone, 2014: Use of field epidemiology and RNA sequencing to describe chains of transmission. *Epidemiol. Infect.* **2019**, *147*, e88. [CrossRef] [PubMed]
105. Firestone, S.M.; Hayama, Y.; Bradhurst, R.; Yamamoto, T.; Tsutsui, T.; Stevenson, M.A. Reconstructing foot-and-mouth disease outbreaks: A methods comparison of transmission network models. *Sci. Rep.* **2019**, *9*, 2074–2086. [CrossRef] [PubMed]
106. Guthrie, J.L.; Strudwick, L.; Roberts, B.; Allen, M.; McFadzen, J.; Roth, D.; Jorgensen, D.; Rodrigues, M.; Tang, P.; Hanley, B.; et al. Whole genome sequencing for improved understanding of *Mycobacterium tuberculosis* transmission in a remote circumpolar region. *Epidemiol. Infect.* **2019**, *147*, e188. [CrossRef]
107. Hall, M.D.; Colijn, C. Transmission trees on a known pathogen phylogeny: Enumeration and sampling. *Mol. Biol. Evol.* **2019**, *36*, 1333–1343. [CrossRef]
108. Hall, M.D.; Holden, M.T.; Srisomang, P.; Mahavanakul, W.; Wuthiekanun, V.; Limmathurotsakul, D.; Fountain, K.; Parkhill, J.; Nickerson, E.K.; Peacock, S.J.; et al. Improved characterisation of MRSA transmission using within-host bacterial sequence diversity. *eLife* **2019**, *8*, e46402. [CrossRef]
109. Ratmann, O.; PANGEA Consortium and Rakai Health Sciences Program; Grabowski, M.K.; Hall, M.; Golubchik, T.; Wymant, C.; Abeler-Dörner, L.; Bonsall, D.; Hoppe, A.; Brown, A.L.; et al. Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis. *Nat. Commun.* **2019**, *10*, 1–13. [CrossRef]
110. Sashittal, P.; El-Kebir, M. SharpTNI: Counting and Sampling Parsimonious Transmission Networks under a Weak Bottleneck. Article Number (bioRxiv: 842237). Available online: https://www.biorxiv.org/content/10.1101/842237v1 (accessed on 4 February 2022).

111. Van Beek, J.; Räisänen, K.; Broas, M.; Kauranen, J.; Kähkölä, A.; Laine, J.; Mustonen, E.; Nurkkala, T.; Puhto, T.; Sinkkonen, J.; et al. Tracing local and regional clusters of carbapenemase-producing *Klebsiella pneumoniae* ST512 with whole genome sequencing, Finland, 2013 to 2018. *Eurosurveillance* **2019**, *24*, 1800522. [CrossRef]
112. Vaughan, T.G.; Leventhal, G.E.; Rasmussen, D.A.; Drummond, A.J.; Welch, D.; Stadler, T. Estimating epidemic incidence and prevalence from genomic data. *Mol. Biol. Evol.* **2019**, *36*, 1804–1816. [CrossRef]
113. Bbosa, N.; Ssemwanga, D.; Ssekagiri, A.; Xi, X.; Mayanja, Y.; Bahemuka, U.; Seeley, J.; Pillay, D.; Abeler-Dörner, L.; Golubchik, T.; et al. Phylogenetic and demographic characterization of directed HIV-1 transmission using deep sequences from high-risk and general population cohorts/groups in Uganda. *Viruses* **2020**, *12*, 331. [CrossRef] [PubMed]
114. Schneider, A.D.B.; Ford, C.T.; Hostager, R.; Williams, J.; Cioce, M.; Çatalyürek, Ü.V.; Wertheim, J.O.; Janies, D. StrainHub: A phylogenetic tool to construct pathogen transmission networks. *Bioinformatics* **2020**, *36*, 945–947. [CrossRef] [PubMed]
115. Dhar, S.; Zhang, C.; Mandoiu, I.I.; Bansal, M.S. TNet: Transmission network inference using within-host strain diversity and its application to geographical tracking of COVID-19 spread. *IEEE ACM Trans. Comput. Biol. Bioinform.* **2022**, *19*, 230–242. [CrossRef]
116. Nelson, K.N.; Gandhi, N.R.; Mathema, B.; Lopman, B.A.; Brust, J.C.M.; Auld, S.C.; Ismail, N.; Omar, S.V.; Brown, T.S.; Allana, S.; et al. Modeling missing cases and transmission links in networks of extensively drug-resistant tuberculosis in KwaZulu-Natal, South Africa. *Am. J. Epidemiol.* **2020**, *189*, 735–745. [CrossRef] [PubMed]
117. Nelson, K.N.; Jenness, S.M.; Mathema, B.; Lopman, B.A.; Auld, S.C.; Shah, N.S.; Brust, J.C.M.; Ismail, N.; Omar, S.V.; Brown, T.S.; et al. Social mixing and clinical features linked with transmission in a network of extensively drug-resistant tuberculosis cases in KwaZulu-Natal, South Africa. *Clin. Infect. Dis.* **2019**, *70*, 2396–2402. [CrossRef] [PubMed]
118. Wang, X.; Zhou, Q.; He, Y.; Liu, L.; Ma, X.; Wei, X.; Jiang, N.; Liang, L.; Zheng, Y.; Ma, L.; et al. Nosocomial outbreak of COVID-19 pneumonia in Wuhan, China. *Eur. Respir. J.* **2020**, *55*, 2000544. [CrossRef] [PubMed]
119. Worobey, M.; Pekar, J.; Larsen, B.B.; Nelson, M.I.; Hill, V.; Joy, J.B.; Rambaut, A.; Suchard, M.A.; Wertheim, J.O.; Lemey, P. The emergence of SARS-CoV-2 in Europe and North America. *Science* **2020**, *370*, 564–570. [CrossRef]
120. Bataille, A.; van der Meer, F.; Stegeman, A.; Koch, G. Evolutionary analysis of inter-farm transmission dynamics in a highly pathogenic avian influenza epidemic. *PLoS Pathog.* **2011**, *7*, e1002094. [CrossRef]
121. Smith, G.J.D.; Vijaykrishna, D.; Bahl, J.; Lycett, S.J.; Worobey, M.; Pybus, O.G.; Ma, S.K.; Cheung, C.L.; Raghwani, J.; Bhatt, S.; et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **2009**, *459*, 1122–1125. [CrossRef]
122. Briand, F.-X.; Niqueux, E.; Schmitz, A.; Martenot, C.; Cherbonnel, M.; Massin, P.; Kerbrat, F.; Chatel, M.; Guillemoto, C.; Guillou-Cloarec, C.; et al. Highly pathogenic avian influenza A(H5N8) virus spread by short- and long-range transmission, France, 2016–17. *Emerg. Infect. Dis.* **2021**, *27*, 508–516. [CrossRef]
123. Murcia, P.R.; Wood, J.L.N.; Holmes, E. Genome-scale evolution and phylodynamics of equine H3N8 influenza A virus. *J. Virol.* **2011**, *85*, 5312–5322. [CrossRef] [PubMed]
124. Biek, R.; Henderson, J.C.; Waller, L.A.; Rupprecht, C.E.; Real, L.A. A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 7993–7998. [CrossRef] [PubMed]
125. Vega, V.B.; Ruan, Y.; Liu, J.; Lee, W.H.; Wei, C.L.; Se-Thoe, S.Y.; Tang, K.F.; Zhang, T.; Kolatkar, P.R.; Ooi, E.E.; et al. Mutational dynamics of the SARS coronavirus in cell culture and human populations isolated in 2003. *BMC Infect. Dis.* **2004**, *4*, 32. [CrossRef] [PubMed]
126. Nie, Q.; Li, X.; Chen, W.; Liu, D.; Chen, Y.; Li, H.; Li, D.; Tian, M.; Tan, W.; Zai, J. Phylogenetic and phylodynamic analyses of SARS-CoV-2. *Virus Res.* **2020**, *287*, 198098. [CrossRef]
127. Vrancken, B.; Rambaut, A.; Suchard, M.A.; Drummond, A.; Baele, G.; Derdelinckx, I.; van Wijngaerden, E.; Vandamme, A.-M.; van Laethem, K.; Lemey, P. The genealogical population dynamics of HIV-1 in a large transmission chain: Bridging within and among host evolutionary rates. *PLoS Comput. Biol.* **2014**, *10*, e1003505. [CrossRef]
128. Gire, S.K.; Goba, A.; Andersen, K.G.; Sealfon, R.S.G.; Park, D.J.; Kanneh, L.; Jalloh, S.; Momoh, M.; Fullah, M.; Dudas, G.; et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **2014**, *345*, 1369–1372. [CrossRef]
129. Burns, C.C.; Shaw, J.; Jorba, J.; Bukbuk, D.N.; Adu, F.; Gumede, N.; Pate, M.A.; Abanida, E.A.; Gasasira, A.; Iber, J.; et al. Multiple independent emergences of type 2 vaccine-derived polioviruses during a large outbreak in Northern Nigeria. *J. Virol.* **2013**, *87*, 4907–4922. [CrossRef]
130. Devold, M.; Karlsen, M.; Nylund, A. Sequence analysis of the fusion protein gene from infectious salmon anemia virus isolates: Evidence of recombination and reassortment. *J. Gen. Virol.* **2006**, *87*, 2031–2040. [CrossRef]
131. Kibenge, F.S.B.; Kibenge, M.J.T.; Wang, Y.; Qian, B.; Hariharan, S.; McGeachy, S. Mapping of putative virulence motifs on infectious salmon anemia virus surface glycoprotein genes. *J. Gen. Virol.* **2007**, *88*, 3100–3111. [CrossRef]
132. Karami-Zarandi, M.; Douraghi, M.; Vaziri, B.; Adibhesami, H.; Rahbar, M.; Yaseri, M. Variable spontaneous mutation rate in clinical strains of multidrug-resistant *Acinetobacter baumannii* and differentially expressed proteins in a hypermutator strain. *Mutat. Res. Mol. Mech. Mutagen.* **2017**, *800-802*, 37–45. [CrossRef]
133. Harris, S.R.; Feil, E.J.; Holden, M.T.G.; Quail, M.A.; Nickerson, E.K.; Chantratita, N.; Gardete, S.; Tavares, A.; Day, N.; Lindsay, J.A.; et al. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **2010**, *327*, 469–474. [CrossRef] [PubMed]
134. Romero-Severson, E.; Nasir, A.; Leitner, T. What should health departments do with HIV sequence data? *Viruses* **2020**, *12*, 1018. [CrossRef] [PubMed]