HAL open science

Methods Combining Genomic and Epidemiological Data in the Reconstruction of Transmission Trees: A Systematic Review

Hélène Duault, Benoit Durand, Laetitia Canini

To cite this version:

Hélène Duault, Benoit Durand, Laetitia Canini. Methods Combining Genomic and Epidemiological Data in the Reconstruction of Transmission Trees: A Systematic Review. Pathogens, 2022, 11 (2), pp.252. 10.3390/pathogens11020252 . anses-03926779

HAL Id: anses-03926779

https://anses.hal.science/anses-03926779v1

Submitted on 2 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

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

check for updates

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

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

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).

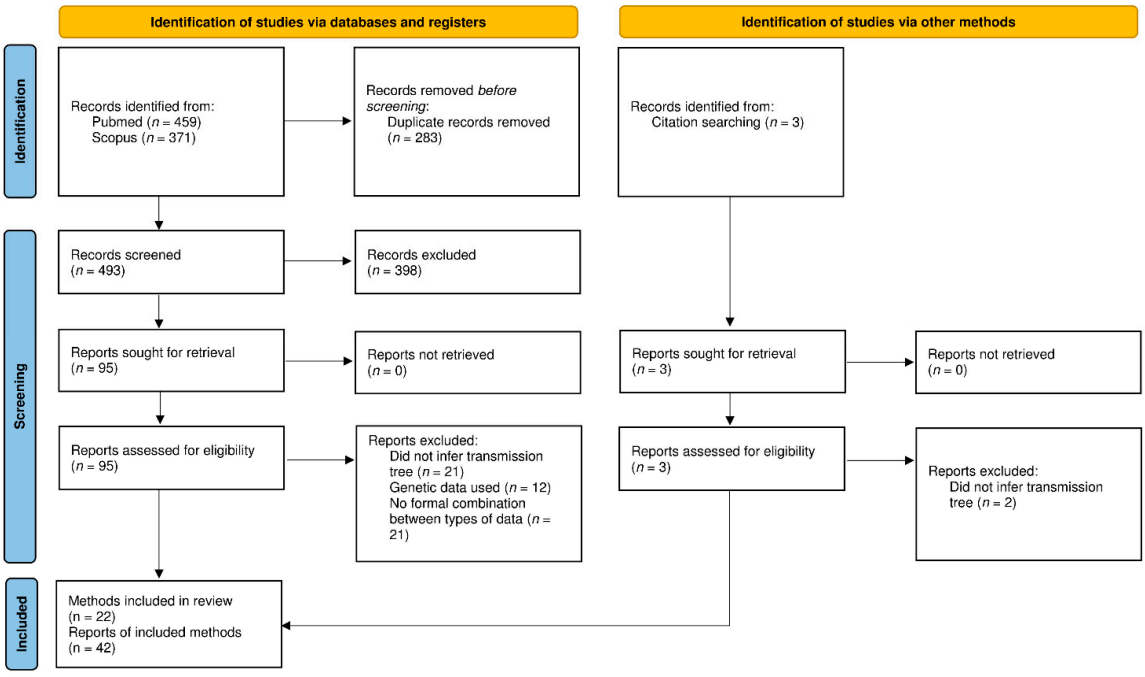

Figure 1. PRISMA flow diagram representing the article selection process (from [27]).

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].

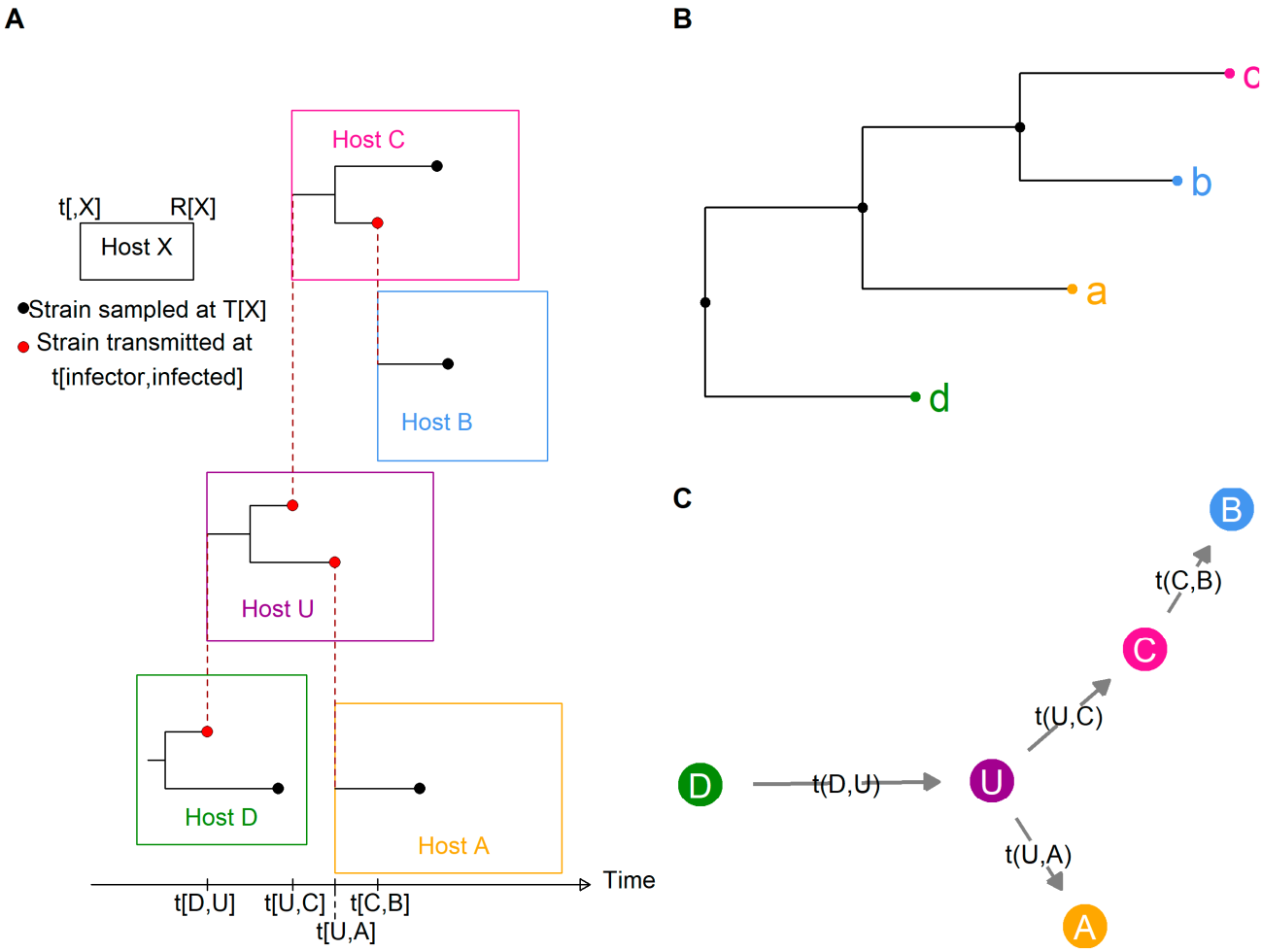

Figure 2. A simple transmission scenario (**A**), the reconstructed phylogenetic tree (**B**), and the transmission tree (**C**). Rectangles represent hosts, and black lines within a rectangle represent within-host evolution of the pathogen. Black circles correspond to sampled strains, red circles to 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			

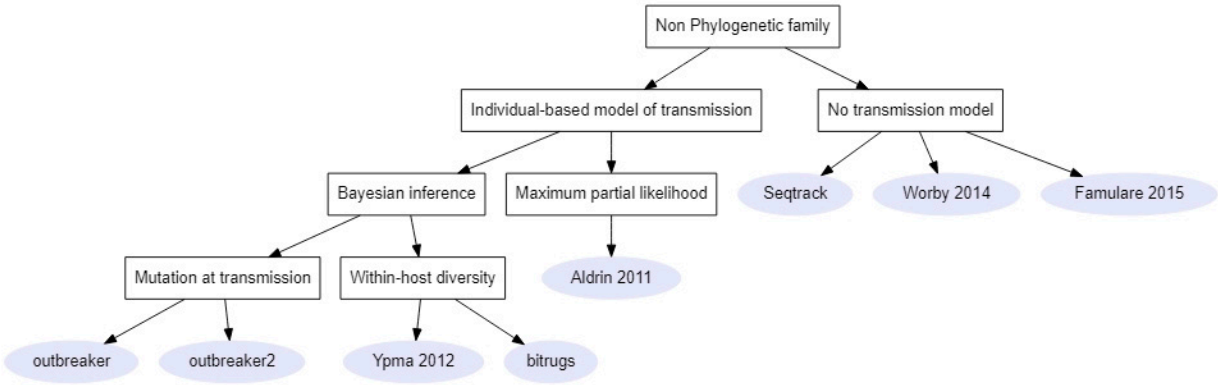

Figure 3. Links between methods of the non-phylogenetic family. Rectangles represent criteria on which to choose a method and the grey circles represent either the name of the method's package or the first author and article date [28,32,37,38].

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).

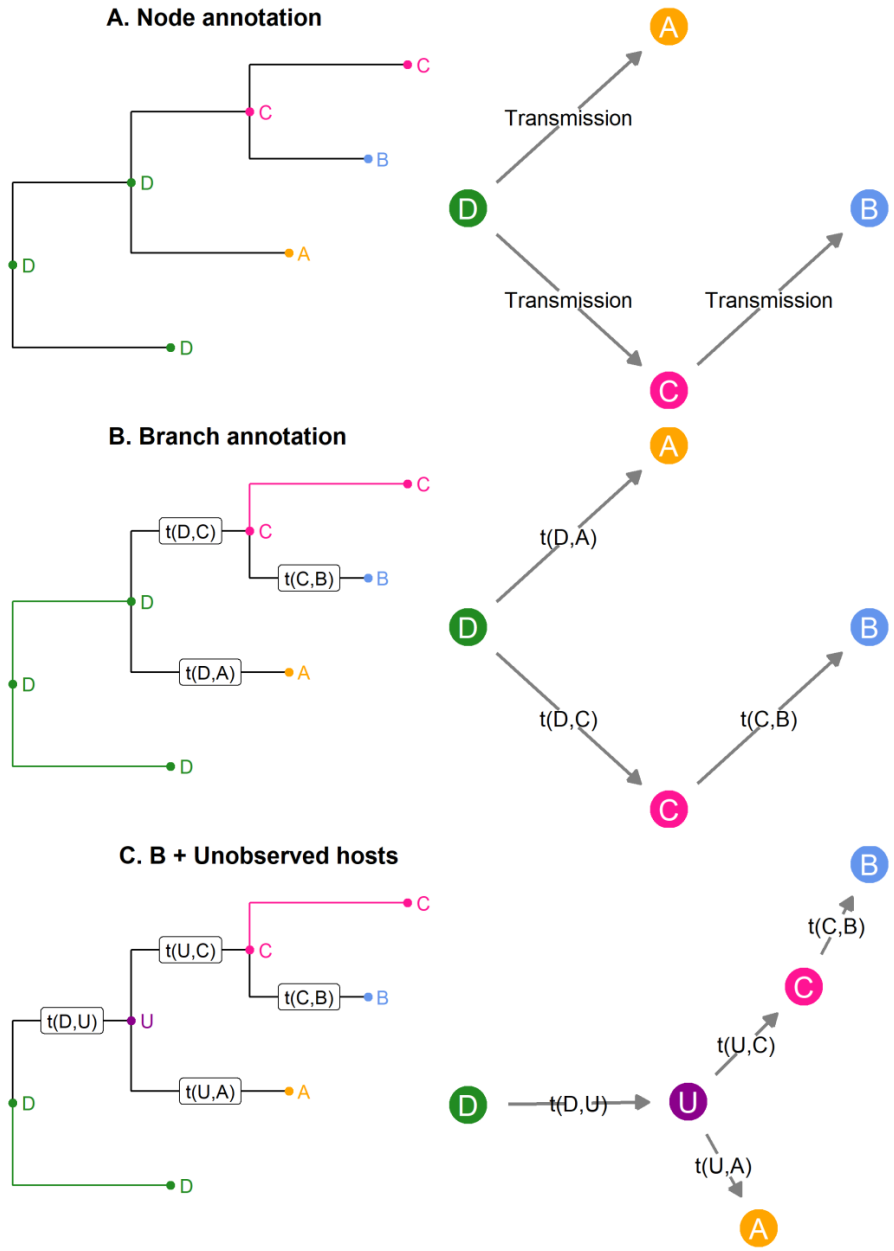

Figure 4. Three links between phylogenetic (on the **left**) and transmission trees (on the **right**). Node annotation with observed hosts (**A**) leads to the identification of transmission links. Annotating the branches (**B**) adds on the time of transmission t. Annotating branches with observed 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].

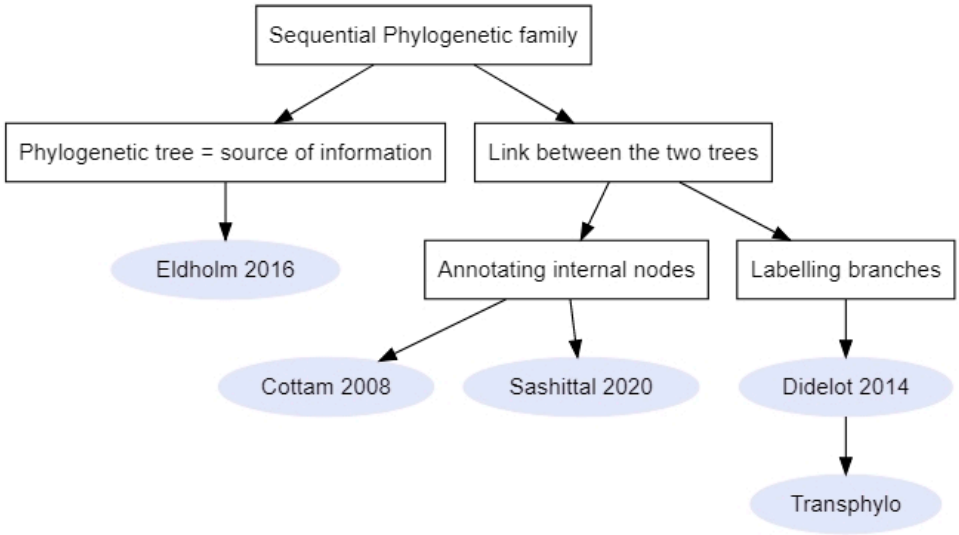

Figure 5. Links between methods of the sequential phylogenetic family. Rectangles represent criteria on which to choose a method and the grey circles represent either the name of the method's package or the first author and article date [2,17,31,39].

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].

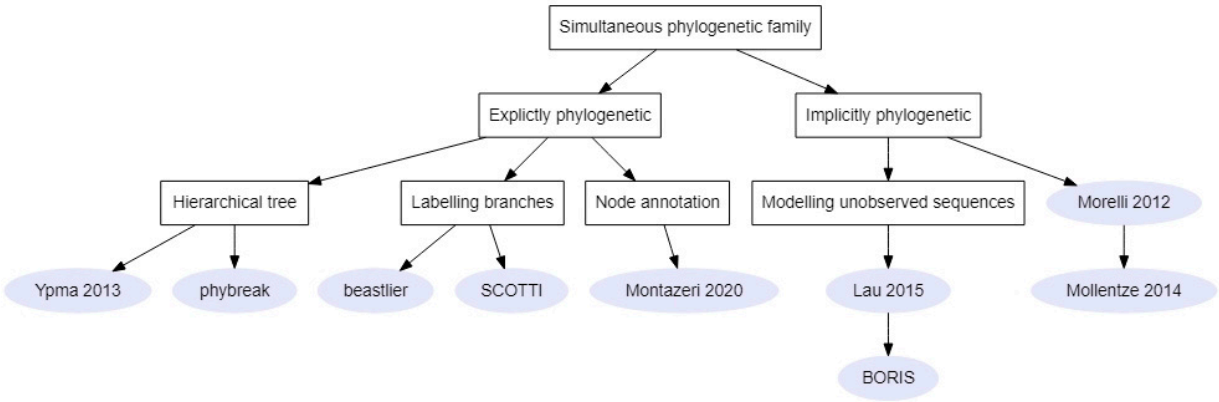

Figure 6. Links between methods of the simultaneous phylogenetic family. Rectangles represent criteria on which to choose a method and the grey circles represent either the name of the method's package or the first author and article date [1,5,23,42,43].

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]