THE UNIVERSITY
of EDINBURGH

Virus phylogeography at the wild/domestic animal interface

Duchatel Florian

THE UNIVERSITY *of* EDINBURGH

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy to the University of Edinburgh

2019

DECLARATION

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or processional qualification except as specified.

Parts of this work has been published on Frontiers in Ecology and Evolution:

Duchatel Florian

13.07.2019

ACKNOWLEDGEMENTS

I thank my supervisors, Samantha Lycett and Mark Bronsvoort, for all their support and input throughout this whole thesis, I would not have been able to finish it without them.

I thank the other researchers from the EERA group and Roslin institute for providing an intellectually stimulating environment.

Je remercie ma maman, pour avoir toujours cru en moi et pour tous les efforts et sacrifices qu'elle a consenti tout au long de ces années pour me pousser jusqu'où j'en suis actuellement. Je remercie aussi mon papa pour m'avoir appris à me surpasser et à toujours m'améliorer. Merci à Baptiste et à Coralie pour être les meilleurs frères et soeur qu'on puisse rêver avoir et pour m'avoir toujours donné des bons conseils.

I finally thank Mado, for being part of my world and for all her support over this whole thesis.

ABSTRACT

With the recent advances in sequencing technology it is increasingly common to find freely available large genomic datasets composed of thousands of genetic samples of viral or bacterial origin. Because Bayesian approaches possess multiple advantages against more traditional phylogenetic methods, such methods are considered as a standard tool to study the evolution and circulation of fast-evolving pathogens such as viruses. With those Bayesian phylogenetic approaches, the evolutionary and transmission parameters of fast evolving infectious diseases can be estimated to inform their control of in both epidemic and endemic contexts. Zoonoses are diseases that have passed from a non-human animal to a human population. The emergence of such diseases in human populations can be caused by environmental changes bringing wild and domestic animals closer to each other, facilitating the transmission to humans. Phylogenetic approaches can help us to understand how diseases can be transmitted between wild and domestic animal populations.

Using recently developed Bayesian phylogenetic methods, the first aim of this thesis was to understand and study the transmission of infectious disease amongst wild and domestic animal populations. Therefore, I performed those epidemiological analysis in an endemic context, using both Eurasian avian influenza sequences and African foot and mouth disease virus (FMDV). The second aim was to develop a software capable of reducing potential sampling bias between multiple populations while analysing large genetic datasets in a short running time compared to currently available phylogenetic methods.

I first studied the transmission and reassortment pattern of avian influenza within Europe and Asia using internal segments sequences (PB2) originated from wild and

domestic birds. Using both a non-structured and structured coalescent approach, I determined that the two continents constitute distinct demes that are sporadically connected. Most of the reassortment pattern observed occurred within western Europe and Eastern China. I also determined that while wild Anseriformes are responsible for most of the of the virus circulation in Europe, domestic Anseriformes birds are responsible for the virus movements in Asia. The circulation of the virus between Asia and Europe is mostly done by both domestic and wild Anseriformes birds.

Secondly, to understand the patterns of FMDV, I compared the transmission patterns of four FMDV (FMDO, FMDA, FMD SAT1, FMD SAT2) serotypes and estimated the factors influencing the circulation of these viruses in Africa using a discrete and a continuous phylogeographic approach. One conclusion of this chapter is that FMDV strains currently circulating in African livestock were probably introduced in the early 18^{th} century trough livestock movements for the serotype A/O and reintroduced from wild Buffalo populations after the African rinderpest epidemic for the SAT serotypes. I also show that movements of domestic cattle were responsible of the FMDV propagation and circulation in Africa, with a small role played by wild animal populations. Thirdly, using advanced Bayesian structured coalescent model approximations, I studied the role played by antelopes in the transmission of FMDV SAT1 and SAT2 in Africa. I found that for both serotypes, antelopes seem to act as an intermediate host between buffalo and cattle.

In the last part of the thesis, I present a new software **"Epitree-sim"** that allows the fast estimation of phylogenetic trees and transmission patterns between demes using a fast-dating algorithm and repeated subsampling of the sequence analysed.

LAY SUMMARY

With the recent advances in sequencing technology, it is increasingly common to find large datasets composed of thousands of genetic viral or bacterial samples publicly available. Multiple statistical tools exist that can make use of such datasets to understand the evolution and circulation of fast-evolving pathogens to inform their control in both epidemic and endemic contexts.

The study of the transmission of human diseases originating from animal populations is of first importance to prevent future public health emergencies. The emergence of such disease in human populations can be caused by environmental changes bringing domestic and wild animals closer to each other facilitating the transmission to humans. Phylogenetic approaches can help us to understand how diseases can be transmitted between wild and domestic animal populations.

The first aim of this thesis is to use recently developed techniques to understand and study the transmission between wild and domestic animal populations in an endemic context using Eurasian avian influenza sequences and African foot and mouth disease virus (FMDV). The second aim is to develop a software capable of analysing large genetic datasets in a short running time compared to currently available methods, while tackling some issues caused by potential sampling bias.

I studied the transmission and reassortment of avian influenza within Europe and Asia using sequences from wild and domestic birds. I determined that the two continents constitute distinct demes that are sporadically connected. Most of the reassortment pattern observed occurred within western Europe and Eastern China. I also determined that while wild birds are responsible for most of the of the virus circulation in Europe, domestic birds are responsible for virus movements in Asia. The circulation of the virus between Asia and Europe is mostly done by both domestic and wild birds.

While analysing transmission of FMDV, I determined that currently circulating African FMDV strains were either introduced by livestock movements in the early 18th century or reintroduced from wild Buffalo populations following the African rinderpest epidemic. I also observed that domestic cattle were responsible of the FMDV circulation in Africa with no or little role played by wild animal populations. I also determined that antelopes seem to act as an intermediate host between buffalo and cattle.

In the last part of the thesis, I present a new software “Epitree-sim” that allows the fast estimation of evolutionary trees and transmission pattern between discrete trait using a new procedure and a repeated sampling of the sequence analysed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

A version of chapter 3 has now been published in Frontiers in Ecology and Evolution: Duchatel, F., Bronsvoort, B M de C. & Lycett, S. “Phylogeographic Analysis and Identification of Factors Impacting the Diffusion of Foot and Mouth Disease Virus in Africa” Front. Ecol. Evol., 09 October 2019 https://doi.org/10.3389/fevo.2019.00371

1 GENERAL INTRODUCTION

1.1 ABSTRACT

In this chapter, the general background of the currently used phylogenetic analysis methods such as the neighbour joining tree, maximum likelihood and Bayesian inference methods is given. More recently developed methods, such as the structural coalescent approach and how it can be used to reconstruct potential transmission networks using epidemiological and sequence data, are also introduced.

Also, a general overview on foot and mouth disease virus and avian influenza is provided. Their viral structure, pathogenicity, transmissibility, presence in wild and domestic population and global circulation pattern are presented.

1.2 GENERAL OVERVIEW

The phylodynamic study of how epidemiological, immunological, and evolutionary processes affect and potentially interact to shape the observed phylogenies of viral disease is the subject of this thesis[1]. Although the use of genetic data have increased our understanding of the transmission and evolution of infectious diseases, multiple challenges still remain (mostly on how sampling pattern, host population structure and viral reassortment affect the outcome of the phylogenetic analysis) [2]. Furthermore, most of the current scientific effort aiming to understand the propagation of contagious diseases and to answer, "who infected whom?" (e.g. location, host) is done by reconstructing transmission networks. A transmission network is a graphical representation of an epidemic which encompasses all the known parameters of the disease propagation amongst given hosts, locations or populations[3]. Additionally, with the increasing availability of high-resolution temporal and spatial pathogen sequence datasets and the continuous development of new computational and analytical tools, we now have the ability to reconstruct such transmission networks using pathogen genetic information[4,5]. Using those approaches, such methods have an improved resolution compared to earlier methods such as the observation of contact patterns between infected animals[6] or infected patients[7–10].

With currently 847,748 whole influenza genomes and 239,018 coming exclusively from avian hosts (accessed on 26th June 2019) publicly available on Genbank / NCBI Influenza Virus resource[11], viral genetic sequences are easily available and allow new opportunities to arise in order to study the transmission and evolution of infectious diseases.

The neutral theory of molecular evolution suggests that the majority of genetic variations in populations is not caused by Darwinian natural selection, but is the result

of random mutations that do not give competitive advantage[12]. However, because of natural selection pressure unfavourable to genetic change would be eliminated from a population, whilst favourable ones will be kept. Once one such modification gets fixed in the population, it becomes a genetic mutation (mutation rate). During replication cycles of pathogens (or cells) within a host, errors can occur when copying the genomic DNA or RNA. Those changes may involve a single nucleotide, with the substitution of a nucleotide for another (substitution rate), or can involve larger segments, with the insertion or deletion of multiple nucleotides, or even changes in chromosome numbers.

Thanks to their fast and error prone replication cycles[13], the observed mutation rates in RNA genome (in the range of 10^{-3}– 10^{-5} mutations per nucleotide copied) are estimated to be 6 orders of magnitude higher than those observed in vertebrates [14]. This high observed mutation rate explains why RNA viruses populations are thought of as being composed of a distribution of closely related mutants, known as a quasispecies, instead of organisms sharing an identical genome[15].

Due to their high realised genome mutation[13], we can estimate the evolutionary relationships that exist between multiple RNA viral strains by reconstructing phylogenetic trees[16]. Because of the continuous emergence of genome variants in RNA populations[14], the appearance of new mutations could coincide with the appearance of new conditions (environmental or not), influenced by population dynamic[5] or random events. Consequently, for RNA viruses, by looking at viral neutral genetic mutations, we can observe the variation caused by environmental or population changes that were not directly observed by the scientific community[5]. In contrast, for organisms with a slower observed mutation rate, the evolutionary and

ecological processes occur on different time scales since mutation occur ahead of the environmental changes[4].

Through the use of a nucleotide substitution model, a statistical model of evolution[17], the genetic distance existing amongst several sequences can be calculated[18]. The simplest available phylogenetic approaches (such as the neighbour-joining[19] or UPGMA[20]) make use of it to infer the evolutionary relationship existing between genetic sequences[4]. Other phylogeny methods, such as the maximum likelihood approach, make use of a statistical approach to estimate the most probable phylogenetic tree given the available genetic sequences[21].

More advanced phylogenetic approaches make use of Bayesian statistics to estimate the relationship between multiple sequences[17,22,23]. Additionally, by using sequences sampled at multiple points in time (heterochronous sequences), a 'molecular clock' can be used to represent the correlation between the observed genetic distances and the time elapsed used to determine the timescale of the studied evolutionary process[2,4]. By using a Bayesian approach to reconstruct a phylogenetic tree, additional epidemiological parameters can be estimated, such as the transmission rate and transition events between sampled populations or locations[17], as well as the influence that multiple factors (environmental and anthropological) might have on the circulation and transmission of the studied pathogen. Furthermore, with the recent advances in structural coalescent model approximation, we are now one step closer to account for the structure present in real world populations considering the existence of multiple subpopulations while reconstructing a phylogeny[23–25]. Additionally, De Maio and al.[22] estimated the possible transmission network of a specific epidemic using epidemiological and genetic information to estimate who infected whom over the whole course of an outbreak[22]. All those cited phylogenetic methods have already

been used on several occasions and have allowed us to gain insight into the factors influencing the transmission of multiple pathogens, such as Ebola virus in Africa[26] or Zika virus in South-America[27].

Since infectious disease transmission comes from the interaction between infective and susceptible hosts, the environmental characteristics where those interactions take place are important features that we have to take into account to properly understand the dynamics of infectious diseases[28]. Landscape epidemiology aims to show the links between spatial context and evolutionary processes by using genetic material to determine the causal forces leading the evolutionary and ecological trajectories[29]. The composition and connectivity between types of landscape patches (such as a forest and an agricultural area) influence the environment structure and therefore impact the disease dynamic, spread and prevalence by changing the biotic or abiotic conditions encountered by the virus or its host [30–32]. Therefore, a better understanding of the influence of landscape patterns on risk factors of viral circulation is a prerequisite for efficient control of diseases[33]. With new approaches combining phylogenetic and real-word landscape measures, we are now able to understand the effect that landscape heterogeneity has on the circulation and evolution of infectious diseases[34]. However, because the environmental conditions that may affect a particular disease's dynamic are highly heterogeneous and might influence the disease propagation over a broad range of potential scales (from a single host to a whole community or continent) it is important to take into account those multiple potential scales at which ecological systems might affect epidemiological processes[35]. However, by adopting a graphical theory framework, it is possible to explicitly evaluate the effect of multiple environmental features while estimating the genetic connectivity across large geographic areas and accounting for small-scale, or local,

heterogeneity[36–39]. Therefore, this approach seems able to efficiently integrate natural and anthropogenic features of landscapes in epidemiological models.

This thesis is focussed on two zoonotic RNA virus diseases, avian influenza[40,41], and foot and mouth disease virus (FMDV)[42]. Those two diseases are both of wild animal origin but have the potential to be transmitted to domestic animals causing massive economic losses with the underlying risk to be transmitted to human populations, starting a potential massive pandemic[42,43]. With environmental and anthropogenic changes bringing wild and domestic animal closer to each other, there is the risk for such factor to impact the virus transmission between the two populations[44]. Therefore, it is particularly important to understand for those diseases which set of factors (animal, human or environmental) might most influence the transmission of zoonotic diseases.

Taken together these two RNA viruses enable the application of recent phylogenetic approaches to gain a better understanding of the causes leading to their propagation and transmission. This thesis also aims to address some of the challenges encountered when analysing genetic data, especially those regarding the effect of sampling patterns and host population structure on phylogenetic analysis results. Avian influenza and foot and mouth disease viruses were selected as examples because both are transmitted to the wild and domestic animal interface leading to a potentially uneven sampling between the two involved populations, depending on the sampling possibilities, with the wild populations often under-represented. Specifically, the aim was to gain knowledge about the circulation of avian influenza amongst wild and domestic bird populations in Eurasia and the effect it has on the virus reassortment patterns. Also, knowledge about the circulation of foot and mouth disease between wild and domestic animals and the effect that those host types have on the circulation of the disease in sub-Saharan Africa was improved. Finally, a new

fast phylogenetic approach able to deal with biased samples was introduced in the last section.

1.3 VIRAL PHYLODYNAMIC TECHNIQUES

The rediscovery of Mendelian genetics and the development of DNA/RNA sequencing technology such as next-generation sequencing paved the way for the development of the new field of molecular evolution. Over the last century, those theoretical and technological advancement provided a framework for molecular evolution studies to estimates phylogenetics trees based on genetic sequences comparison[45]. With the constant increase in computational resources, the development of new mathematical tools and the improvement of sequencing techniques, phylogenetic approaches are now used in disease transmission studies, sometimes even down to the individual host-to-host scale[9]. Such phylogenetic studies aim to use genetic material to track transmission patterns, with the assumption that the parts of the genetic sequences being used to construct the trees are not under strong selection pressure (for example, sequence positions associated with drug resistance mutations are usually removed before tree construction as they can have a biasing effect on the inferred transmission patterns)[46]. This section introduces some of the concepts important to understand how modern phylogenetic approaches work.

1.3.1 Model of molecular evolution

The genetic distance between two sequences is a measure of how divergent the sequences are, and in its most simple form is just the number of bases different between the two. This divergence comes from random nucleotide substitution events

that can be statistically modelled using a model of evolution. Phylogenetic methods are based on few assumptions over those models; therefore, the use of different models may change the outcome of phylogenetic analysis. Phylogenetic methods may be less accurate or inconsistent if the model of evolution chosen is wrong[17].

The substitution process is usually modelled as a continuous-time Markov chain (CTMC) where the whole process can be described by a matrix of infinitesimal exchange rates. CMTC assumes that the process is memoryless (the future state is only influenced by the current state and not the past ones), homogeneous (substitution rates do not change through time) and stationary (relative frequency of bases do not change)[47]. Different parameters are needed to describe nucleotide substitution patterns. Four parameters describe the substitution between purines and pyrimidines (transversion), two describe substitutions within purines or pyrimidines (transitions) and four parameters describe the relative base frequencies[48].

Different models of substitution with increasing levels of complexity exist. The simplest model was created by Jukes and Cantor in 1969 (JC69). It assumes that all bases are present at equal frequencies and that each base can change with the same frequency. In this model, only one parameter (the instantaneous rate of change between two nucleotides) is needed to describe the whole substitution process. The Kimura 2 parameter assumes that all base frequencies are equal but that transitions and transversions have different rates. The Hasegawa-Kishino-Yano (HKY) model assumes an unequal base frequency and a different rate of different rate between transitions and transversions. The general time reversible model (GTR) assumes a unequal base frequency and a different rate between all substitutions[48].

We can observe different rates of substitution amongst sites present on the same genetic sequence[12]. Those variations arise from the interactions of functional and

structural pressures that each site experiences[49]. For example, protein surface sites usually evolve faster than interior sites[50] but surface sites involved in protein–protein interactions are often more conserved and therefore evolve more slowly than other surface proteins [49,51]. This heterogeneity may have a huge influence in the inference of genetic distance, and a discrete gamma distribution with 4-8 discrete categories is commonly used to simulate those sequence variations[52–54]. The use of such gamma variation is not only limited to amino-acid substitution, but also between nucleotides where rate variation also exists among sites[55].

1.3.2 Molecular clock model

Viruses evolve at a high enough rate that it is possible to detect measurable genetic differences in samples taken only few months apart within the same epidemic or individual[13]. Because those genetic differences seem to accumulate over time, the existence of a molecular clock defining the relationship between the rate of evolution of a particular gene with the time is implied[56]. If this molecular clock rate of evolution can be estimated, then the divergence time between two sequences can be easily determined. Therefore, the length of the branches linking two genetic sequences within a phylogenetic tree can be expressed in time units and correspond to the divergence time between two organisms (see Figure 1-1)[4].

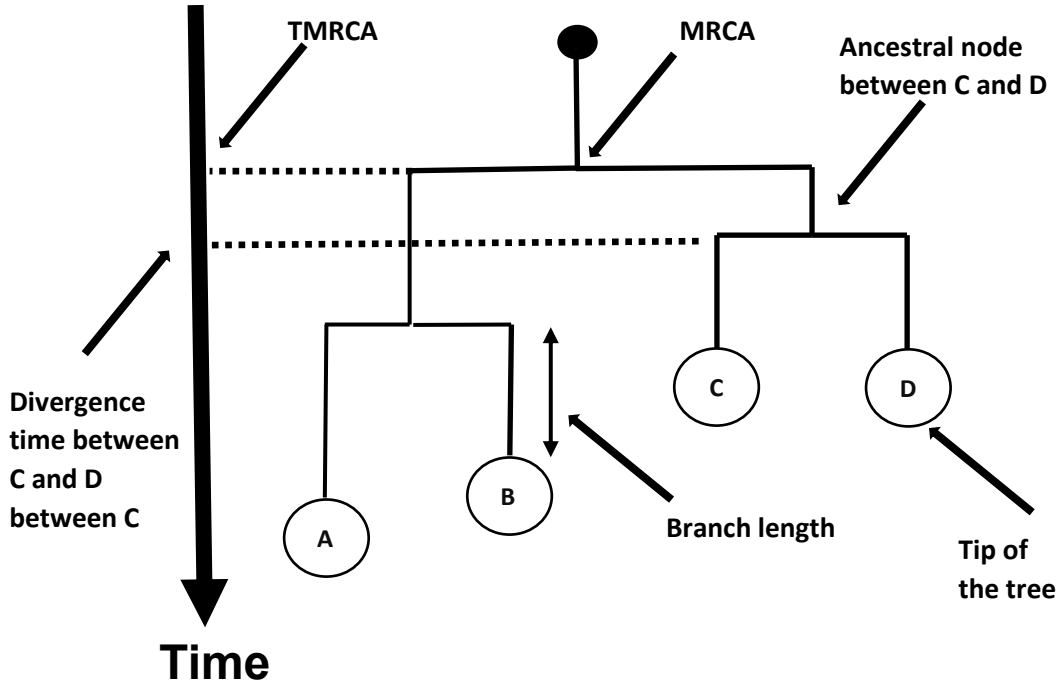


Figure 1-1: Time scaled phylogenetic tree between four sequences showing the MRCA, TMRCA, ancestral node, divergence time and branch length. The most recent common ancestor or MRCA is the hypothetic most recent sequence from which all the sampled sequences used in the phylogenetic analysis descend. The time to the most recent common ancestor is an estimate of the time at which the MRCA would have existed. An ancestral node is a node that connect multiple tree tips representing the sampled sequences. The time where the divergence between multiple sequences is supposed to have occurred is called the divergence time. The different tips and nodes of the tree are connected through branches. In a time scaled phylogenetic tree, the branch length represents the time elapsed between two consecutive nodes.

The best type of data that can be used when doing such analyses is data sampled years apart (heterochronous data) because their temporal structure provides an additional source of information to calibrate the clock[57]. In contrast, data sampled in the same period does not allow the mutation rate and time of divergence to be estimated at the same time [58], and typically an average substitution rate obtained from other studies is applied.

A simple method for estimating the rate of a molecular clock using an heterochronous dataset is a linear regression of the root-to-tip genetic distance against the sampling times from a rooted phylogeny (see Figure 1-2)[5]. Nevertheless, this method has some

shortcomings such as the assumption of independency between each genetic distance (or root to tip) when all sampled sequences share a common evolutionary history. There are improvements to this basic method coded within the Tempest software or the treedater R package[59,60], and linear regression is still used to get an early idea about the molecular clock when studying a new dataset[13] or to remove diverging sequences before using a more sophisticated method such as Bayesian inference.

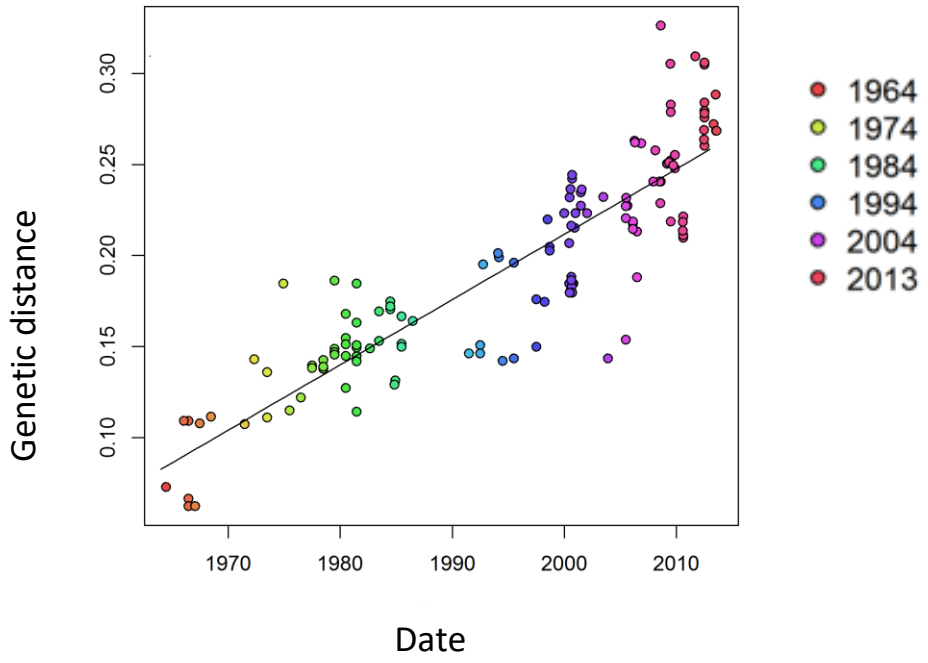


Figure 1-2: Example of root-to-tip divergence plot. This graphic shows the correlation that exists between the genetic distance and time of sampling for a set of heterochronous sequences. The identified correlation can be used as molecular clock estimates. Such regression approach is still used to identify problematic or diverging sequences and as preliminary analysis before performing more computationally demanding phylogenetic analysis.

For a long time, a strict molecular clock assuming a linear relation between time and number of substitutions was used [61].

However, because neutral evolution is arguable in most viral RNA[62], the fluctuations in adaptive environments and population size will affect the constancy of their

substitution rates[63]. Consequently, the idea of a molecular clock was relaxed by allowing the clock rate to vary along the phylogeny[64] by using local molecular clocks, where different rates are estimated for different clades. In relaxed local molecular clocks, the rate in each branch is determined from a parametric distribution which depends on the rate found in the parent branch. This phenomenon, called autocorrelation, means that adjacent branches have more similar molecular clocks than more distant branches[65]. The idea of autocorrelation comes from two key assumptions[66]. The first is that the mutation rate is influenced by life-history characteristics such as the generation time or the DNA repair efficiency[67,68]. The second is that the mutation rate and the substitution rate are correlated, which is possible if two closely related species (or individual) experience the same selection intensity[66].

In contrast, uncorrelated relaxed clocks are defined when the rate of each of the tree branches is drawn independently and identically from an underlying distribution of rates of substitution for the whole tree. Uncorrelated relaxed clocks, based upon an exponential or a log normal distribution, are widely used for modelling infectious diseases such as influenza or FMDV[69].

Additionally, other models of clocks exist, such as the local molecular clock where different regions of a phylogenetic tree have different rates[70]. Extensions of this model exist such as the Bayesian random local clock, where all possible local clock configurations are nested using a Bayesian apporach[71], or the host-specific local clock that allows different phylogeny states or hosts to have different molecular clocks[72].

1.3.3 Coalescent model

The genealogy of a population contains information about the processes that shaped it. For example, individuals from an isolated small village probably have a most recent common ancestor (MCRA) than two persons living in a big city. The time to the most recent common ancestor (TMRCA) depends on parameters such as the size of the population, the rate of migration and changes in population size. Coalescent theory makes inference on those historical population processes and demography and was first developed by Kingman in 1982.

The simplest way to present a coalescent model is to use a Wright–Fisher population model, a random mating haploid population that has a constant size over time and which is composed of N individuals. In this population, if each new generation is solely composed of the descendants of the preceding generation and if there are no selective forces acting on the population and that all individuals have an equal chance of producing offspring, the probability that two individuals have a common ancestor in the preceding generation is $P(t_1) = \frac{1}{N}$. Likely, the probability that they share a common ancestor two generations away is $P(t_2) = \frac{1}{N}\ (1-\frac{1}{N})$. By generalising this approach, we can assume that after k generations, the probability that n sampled individuals have a common ancestor is:

$$P(t_k) = (\frac{n(n-1)}{2N})exp(\frac{n(n-1)}{2N})dt$$

However, natural populations do not necessarily have constant population sizes. Therefore, instead of N we often use an effective population number (Ne) which can be defined as the size of a fraction of the studied population which has the same genetic diversity and the same coalescence rate as the global population[74,75]. The size of this population is often assumed by phylodynamic studies to be proportional

to the effective number of infected individuals, however this relation does not seem to be true in general[74]. The effective population size is an important determining factor of the change in allele frequencies over time[75].

We can easily adopt the concept of substitution rate when speaking about phylogeny. In that case, each generation leads to μ substitutions, N generations to Nμ substitutions and t' generations to b= μtk substitutions. Therefore, the time is not measured in generations anymore, but in number of substitution t'. The probability to have the ancestor after t' substitutions is:

$$P(t'_b) = (\frac{n(n-1)}{2N\mu})exp(\frac{n(n-1)}{2N\mu})dt'$$

In 1999, Rodrigo and Felsenstein established the serial coalescent theory which takes into account heterochronous samples. This approach allows a direct estimation of the mutation rate by considering the sampling interval. This approach offers a better estimation of the number of past lineages, therefore mutation rate and time can be separated when it was not the case before[17].

1.3.4 Effective population size

A ‘‘coalescent model" is a mathematical representation of the changes in effective population size observed through time and is used by coalescent method to infer a population genealogy. Each coalescent model has one or more "demographic parameters". Some of the most commonly used demographic models with a simple parametric function are (classified by increasing parametrization): the constant size model, the exponential growth model (constant growth rate), the logistic growth model (decreasing growth rate) and the expansion growth model (increasing growth)[77].

However, simple demographic models can sometimes not describe the population history in an adequate way. The use of an incorrect coalescent model usually leads to false estimates of demographic history and bias in the estimation of evolutionary model parameters.

The choice of model can be simplified by using non-parametric coalescent models such as the Bayesian skyline plot method. Those models estimate the population size directly from the sequence and can be used to simplify the parametric model selection[5,77].

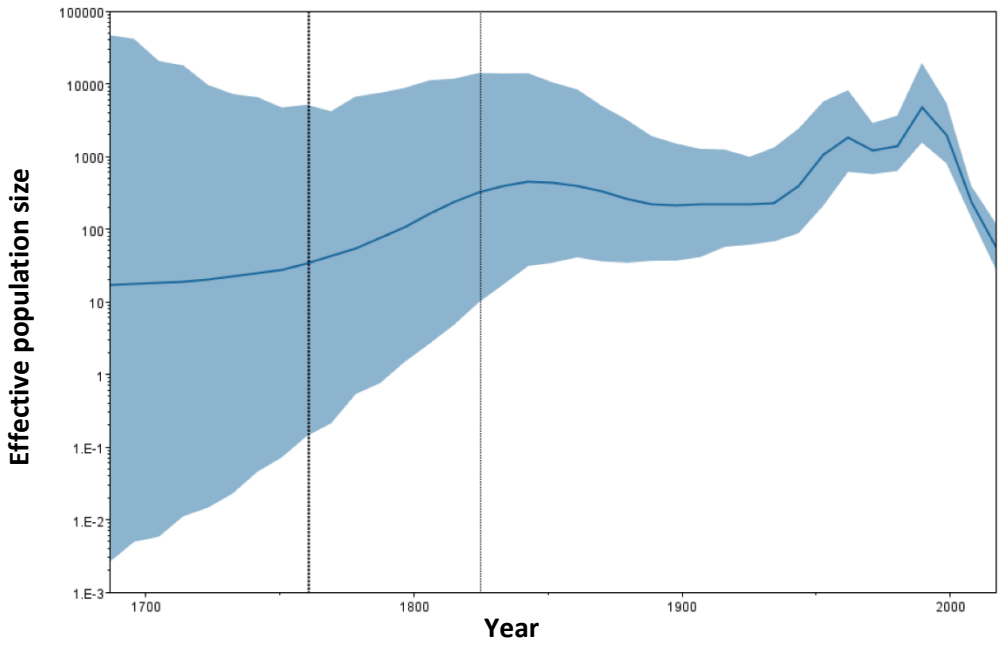


Figure 1-3: A Bayesian skygrid plot obtained using an alignment 134 foot and mouth disease virus serotype SAT1 sequences sampled from 1961 to 2015. The graphic represents the demographic history of the virus by showing the evolution of the viral effective population size over time. The x axis represents the years before 2015 whilst the y axis is equal to the estimated effective population size. The median estimate is represented by a thick solid line and the 95% highest posterior density. We can observe that the viral effective population size drastically dropped after 2000 following a sharp increase during the second half of the 20th century.

Another non-parametric coalescent model is the Bayesian Skyride model. This model uses a Gaussian Markov random field (GMRF) to estimate trajectory changes at coalescent times, and allows the estimation of the effective population size through

time[78]. The Bayesian Skygrid model is an extension of the Bayesian Skyride model that allows the estimated trajectory to change at pre-specified fixed points in real time (see Figure 1-3). This model gives the user extra flexibility and enables the Skygrid's GMRF prior to be independent of the genealogy[79].

1.3.5 Statistical phylogeny and coalescence

1.3.5.1 Neighbour joining tree and Maximum likelihood phylogenetics

The neighbour joining (NJ) reconstruct phylogenetic trees by performing a hierarchical clustering of the genetic distance between each pair of available DNA or protein sequence analysed[19]. This approach uses a greedy algorithm to estimate the reconstructed tree according to a 'balanced minimum evolution' (BME) measure[80]. This BME estimates that the optimum tree typology is the one that minimizes the total tree length. Because this approach is fast, it is well suited to analyse large datasets and to perform bootstrap analysis[81]. However, the biggest drawback of the NJ approach is that we need closely related good quality sequences to reliably estimate the genetic distances[82,83].

The maximum likelihood (ML) approach infers the phylogenies by using a statistical model of molecular evolution[21]. This approach aims to find the tree (with a given topology, branch lengths and substitution model parameters) which gives the highest probability considering the available sequences:

$$p(D/M) = \int_{\theta \epsilon \Theta} p(D/\theta, M) p(\theta/M) d\theta$$

With D being the observed data (the sequences), M the model to evaluate and θ being the parameters of the model[84]. Because this method assumes that the evolution between the sites and lineages is statistically independent, the maximum likelihood approach is appropriate to analyse non-closely related sequences. A drawback of this approach is that it is computationally demanding and might get trapped in local optima[85,86]. Therefore, to avoid getting trapped in one of those local optima, subtree pruning and regrafting (SPR) moves can be performed. This approach relies on moving parts of the trees at a different position in order to perform a better and more exhaustive search of the possible tree typologies[86].

1.3.5.2 Bayesian inference

Phylogeny reconstruction using Bayesian inference is a commonly used technique because of its advantages above other tree reconstruction approaches, such as the NJ and the maximum likelihood approaches. Bayesian phylogenetic inference has the ability to take into account phylogenetic uncertainty in combination with other substitution, molecular clock and effective population size models[87]. In Bayesian inference, the posterior probability of a specific tree is calculated as a function of a prior probability belief on the tree and the likelihood of the observed sequence over the tree tested. For a phylogenetic tree τ using a set aligned genetic sequence X the posterior probability $P\left(\frac{\tau}{X}\right)$ of the tree τ using the set of sequences X is:

$$P(\tau/X) = \frac{P(\tau)P(X/\tau)}{P(X)}$$

With:

$$P(X) = \sum_{1}^{n} P(X/\tau_n)P(\tau_n)$$

Therefore, to calculate P(X) we must sum the probability across all the tree topology space which therefore makes the maximum posterior probability impractical to calculate analytically. The solution is to estimate the posterior distribution using a Markov chain Monte Carlo sampling (MCMC) [88].

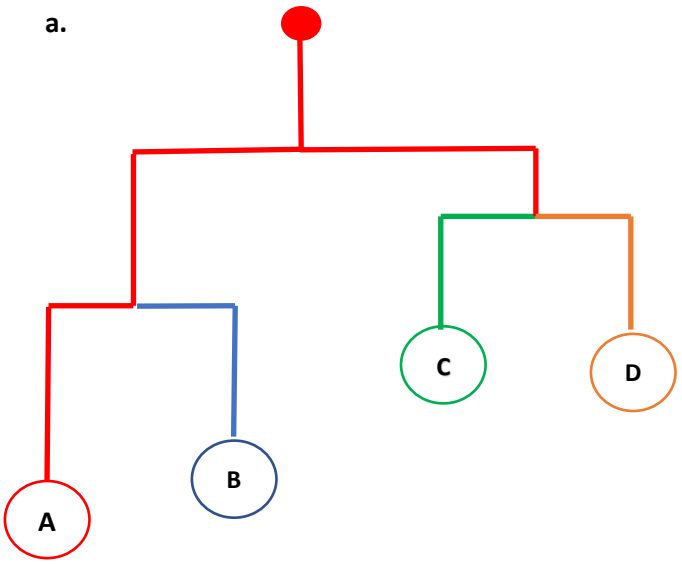
The central idea is to randomly change a current parameter state to form a conceptual chain. For time-scaled phylogenies, the possible changes include rearrangements of the tree topology and changes to the parameters of the molecular clock model [87]. For each chain iteration, a new value of the parameter is proposed as next step. If the new parameter has a higher posterior probability than the present parameter in the chain, the move is accepted, and the new parameter is now the current step in the chain, and the cycle is repeated. If the new parameter has a lower probability, the new value will be accepted with a probability p (equals to the ratio of the posterior of the proposed state compared with the posterior of the current state). If the new step is rejected, the current parameter is used as next step in the chain. By repeating this process millions of times, the chain will attempt to find the complete distribution of parameters probability[88].

1.3.5.3 Mugration model

CTMC models are not only restricted to nucleotides but are used in a broad range of data analyses. In particular, they can be used with many different discrete states data types, such as spatial locations [17,89] or hosts [90].

Discrete spatial diffusion and population transmission can be modelled as evolutionary processes where the hosts or locations are considered as discrete virus traits and distinct populations. Discrete reconstruction of the location history is particularly suitable when sequences are from a limited set of location or populations (single cities or country of origin)[91,92] and are particularly suitable to represent human mobility between specific locations (airports), but less suitable to represent the diffusion of animals or plants pathogens in a continuous space[93].

Using a CTMC approach, a discrete model estimates the set of transition rates along each of the branches of the reconstructed phylogenetic tree. The resulting rate matrix can be symmetric or asymmetric depending of the model used. The Figure 1-4 shows an example of a discrete phylogenetic tree and corresponding rate matrix using an asymmetric model.[17,94]



b.

	A	B	C	D
A	0	λab	λac	λad
B	λba	0	λbc	λbd
C	λca	λcb	0	λcd
D	λda	λdb	λdc	0

Figure 1-4: Example of a discrete phylogenetic tree and corresponding rate matrix. Using a CTMC the discrete model aims to estimate a possible transition rate matrix (b) between the different states (A, B, C and D) observed over the whole phylogeny (a). An asymmetric model will consider antagonist rates (λ_{ab} and λ_{ba}) to be different while a symmetric model estimates those rates to be similar. Using this approach, we can infer ancestral states of all the internal nodes.

Based on previous work[89,95], the model for discrete phylogeny reconstruction called “mugration” approach was introduced by Philippe Lemey et al. in 2009 and is now the most widely used discrete Bayesian approach due to its computational efficiency and ease of use. However, this method suffers from diverse limitations, especially regarding how it models the viral migrations. Indeed, one of the consequences of the different approximations this model take is that the migration events between the modelized population (or trait) do not affect the overall tree shape, and instead the model assumes that the population changes according to a standard coalescent process[25]. Because of those assumptions the model makes and regarding how it models the migration between the different populations, the “mugration” approach might suffer from statistical bias and approximation when performed on a biased dataset[24,25]. More specifically, the “mugration” approach assumes that the sampling intensity is proportional to population size, therefore if the sampling is not even across the studied trait, the model inference might be biased with difficulties to understand the role of the under-sampled population[9,96,97].

1.3.5.4 Bayesian stochastic variable selection - BSSVS

Lemey et al. 2009 extended the discrete Bayesian approach by implementing a mixture model where there is some probability that exchange rates in the Markov model might be null. For GTR models of nucleotide substitutions, all transitions have a non-negligible probability of occurring. However, in the case of locations (or hosts), where only few transitions are possible, it is unlikely that all transitions occur and where most of the transmission rates are suspected to be equal to zero. To circumvent this, a Bayesian stochastic variable selection (BSSVS) is performed to determine the

most parsimonious explanation of the viral diffusion process within the different trait states. The use of a BSSVS allows the creation of a Bayes Factor (BF) which describes the significance of an individual rate. The BF of particular rate k is the posterior odds that the rate is non-zero divided by the equivalent prior odds[91].

$$BF = \frac{\frac{p_k}{1-p_k}}{\frac{q_k}{1-q_k}}$$

With p_k = posterior probability that rate k is non-zero and q_k = prior probability.

1.3.5.5 Markov Jumps and Markov Rewards

Recent improvements in explicit calculation enables the estimation of the expected number and timing of all the transition events (Markov jumps) between the different discrete states of a trait. Moreover, the timing of each jump along the phylogeny can be traced so that the time spent (Markov reward) in each of the states can be determined [78,91].

1.3.5.6 Continuous phylogeographic approach

When the different sequences used for phylogeographic inference are continuously distributed over a geographic area, the continuous phylogeographic approach may be more appropriate than the discrete approach. The discrete phylogeny approach is not able to explicitly model the diffusion process on a continuous space and is only able to infer ancestor locations from the sampled locations[98]. Continuous phylogeographic methods are based on a diffusion process (Brownian motion) over a

two dimensional space and allow ancestral viruses to reside anywhere on a continuous geographical landscape and not only on a previously observed location [91].

The Brownian diffusion model assumes that the diffusion process remains homogeneous over the entire phylogeny[99]. Therefore, for each branch of a phylogeny F, the diffusion process starts at the tip of the branch at the time s and location $X_{(s)}$ to end at the location $X_{(t)}$ at time t. Under a time-homogeneous spatial diffusion process, moving from the location $X_{(s)}$ to the location X(t) would depend on an infinitesimal precision matrix P. With this approach, the whole process likelihood is multivariate normally distributed with a variance $P_{-1} \times (t - s)$ which depends on time differences (t-s) and not the actual location values. This is a similar restriction to the strict molecular clock assumption in the molecular sequence substitution process [98].

This homogeneous diffusion process assumption was relaxed by allowing the diffusion rate to vary according to an underlying discrete rate, approach previously used in molecular clock theory[100]. Therefore, for each branch b of F a rate scalar φ_b is assigned to produce a relaxed random walk model. The scalar φ_b takes the diffusion variance (rate) matrix P_{-1} and rescales it to $P_{-1}\varphi_b$. As a consequence, the underlying diffusion process can now vary from branch to branch along the phylogeny [98].

The continuous approach has been used to model the diffusion of multiple pathogens including the dengue virus in Vietnam[101], the rabies virus in Tanzania[102] and H5N1 avian influenza in Asia [103].

1.3.6 Resistance distance and "isolation-by-resistance" approach

With the emergence or re-emergence of landscape and spatial epidemiology [30], there is a growing interest for tools explicitly able to incorporate landscape heterogeneity into gene flow and genetic differentiation analyses[34]. Landscape characteristics, such as landscape patches layout, can influence organism circulation between localities. Therefore, an effective distance is often used instead of Euclidian distances to represent the connectivity and the aptitude of organisms to move between two locations by including the effects of landscape on organism movements [32,104,105].

However, because the environmental conditions that may affect a particular disease's dynamic are highly heterogeneous and might influence the disease propagation over a broad range of potential scales (from a single host to a whole community or continent), it is important to be able to take into account the multiple potential scales at which ecological systems might affect epidemiological processes[35].

Isolation-by-resistance (IBR) model can take resistance surfaces as input to estimate the effective distances between locations[106]. Resistance surfaces are well suited to understand which landscape and ecological feature influence the effective distances between locations[32]. A resistance surface can be calculated using a graphical theory framework based on circuit theory[107] which treats environmental raster as grids of electric resistance or conductance[108]. When a raster is considered as being a resistance grid, high raster values will impede the circulation. At the opposite, if the raster is considered a being a conductance grid, high raster value will be more permeable to the circulation. In the case of a resistance grid, if two locations are connected by multiple low resistance cells, the circulation between the two will be

facilitated and the effective distance between the two locations will be lowered[106,107]. By adopting a graphical theory framework, we can explicitly estimate the effect of multiple environmental features over large geographic areas while still accounting for small-scale, or local, heterogeneity[36–39].

To study the impact that environmental variables have on a pathogen transmission, at least two types of rasters must be considered: the environmental rasters and their corresponding 'null' rasters acting as negative controls. Such null rasters can take the form of a copy of the environmental rasters with all the cell values equal to one[108]. When performing the analysis on the null raster, there is no environmental heterogeneity impacting the effective distance between locations and only the spatial distance is considered.

Consequently, this circuit theory approach seems to be able to efficiently integrate natural and anthropogenic features of landscapes in epidemiological models.

Recently, resistance surfaces have been used in phylogeny reconstruction to identify and test the impact of environmental features on disease transmission. This approach has been used in both discrete and continuous phylogeny reconstruction approaches [93,109].

1.3.6.1 GLM to test environmental predictors

The discrete phylogeny can be extended by parameterising each rate of among-location movement as a function of various environmental predictors. The GLM considers every instantaneous movement rate Δij for $i \neq j$ in as Δ a log linear function of the set **P** of predictors $x = (x_1, x_2, \ldots, x_P)$, such that:

$$log(\Delta_{ij}) = 1 + \beta_1\delta_1 x_{i,j,1} + \beta_2\delta_2 x_{i,j,2} + \cdots . + \beta_p\delta_p x_{i,j,p}$$

Where $\beta = (\beta_1, \beta_2, \ldots, \beta_P)$ quantifies the contribution of the predictor (in log space), and $\delta = (\delta_1, \delta_2, \ldots, \delta_P)$ is a binary indicator variable that allows the predictor to be included or excluded from the model. The variable δ can be estimated by using a BSSVS to estimate the support for each of the predictors (including as Bayes Factors) and expresses the importance of the predictor in the model[110,111]. We can estimate the impact that the environment or human activities have on the virus circulation by parametrising the GLM analysis with values obtained from a resistance surface analysis between two locations[34]. It is interesting to notice that this approach identifies the best combination of predictors that explains the rates of transition amongst the discrete locations and therefore estimates the impact that environmental factors have on the dispersal 'frequency' of the virus[108].

The GLM approach has already been used in multiple occasions and different settings. For example, this approach has been used to understand the effect that live swine trade has on the circulation of influenza A in swine population[112], or to understand the factors affecting the intracontinental spread of dengue virus in human population, or avian influenza in bird populations[113,114]. Additionally, the use of resistance surfaces in discrete analysis has been used to understand the circulation of the rice yellow mottle virus in Africa[109], or to estimate the impact of the environment on the circulation of rabies in Africa[102].

1.3.6.2 Test of environmental predictors impact in a continuous approach

Such as the discrete approach, we can use the output from a continuous analysis to test and estimate the effect that environmental and anthropological features might have on a virus circulation[93,98,115].

To perform this approach, the spatio-temporal information contained in the posterior set of trees is extracted and phylogenetic branch is treated as a distinct movement vector with an associated start/end location and duration[116]. For each branch, environmental factor and environmental distance can be computed which allows the correlation between the time spent on each branch and the associated environmental distance value to be evaluated. The statistical significance of this correlation is tested using a randomised phylogeny and expressed in the form of a BF[115]. This analysis estimates the impact of tested environmental factors on the dispersal 'velocity' of the virus[108].

This approach has already been used to estimate the effect that the altitude, direction of the wind and cattle density have on the Bluetongue virus circulation in Europe[117].

1.3.7 Model comparison

Comparing potential models is an essential part of phylogenetic hypothesis testing[84,118]. A standard approach to compare models in Bayesian phylogenetic is to evaluate a Bayes factors equal to the ratio of the two model marginal likelihoods[119]:

$$BF = \frac{p(D/M_1)}{p(D/M_2)}$$

With D being the observed data (the sequences) and M the evolutionary model evaluated (in this case it would be a phylogenetic tree and all its parameters). A BF

value superior to one would be in favour for the model M_1 while a BF value above 3 would give a strong favour to the model M_1[116]. However, a problem of this approach is that the evaluation of BF values tends to favour complex models[120].

Another approach to compare model maximum-likelihood context is to use an Akaike information criteria[121]:

$$AIC = 2k - 2l_{max}$$

With l_{max} being the maximum log likelihood and k the effective number of parameters of the model evaluated. In this case, I would favour models with lower values of AIC. It is possible to estimate a posterior simulation–based equivalent of the AIC criteria, the AICM, by using posterior samples generated by a MCMC[122]:

$$AICM = 2s_l^2 - 2\bar{l}$$

With $\bar{l}$ and s_l^2 being the sample mean and variance of the posterior log likelihoods.

Another method to compare evolutionary hypotheses is to use non-posterior-sampling methods such as the path sampling (PS)[123] and stepping stone sampling (SS)[84,124]. Those methods estimate the marginal likelihood in a series (or path) that links the posterior and prior distribution of the evaluated model. The PS models estimate this path by estimating a succession of power posteriors that are only differing by their power values [69]:

$$q_\beta(\theta) = p(D/\theta, M)^\beta p(\theta/M)$$

With $p(D/\theta, M)$ being the likelihood function and $p(\theta/M)$ the prior. With the power posterior being equal to the posterior distribution when β=1.0 and to the prior

distribution when β=0.0. The PS method estimates the marginal likelihood of the model in K+1steps by evenly selecting the β value[120]:

$$lnp(D/M) = \frac{1}{2K}\sum_{k=o}^{k-1}(lnp\left(D/\frac{\theta_k}{M}\right) + lnp(D/\theta_{k+1}, M))$$

With $\beta_0 = 0$ and $\beta_k = 1$.

The SS methods improve the PS approach by the β values according to a Beta (α, 1.0) distribution instead of selecting the β values regularly between 0.0 and 1.0. Therefore, we can calculate the marginal likelihood using n samples from a series of K+1 power posterior[124,125]:

$$\mathrm{p(D/M)} = \prod_{\mathrm{k=1}}^{\mathrm{K}}\frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}}\mathrm{p(D/\theta_i, M)}^{\beta_k-\beta_{k-1}}$$

Even though the PS and SS methods require additional computational time, it has been shown that they outperform posterior-sampling methods to accurately estimate marginal likelihoods [69].

1.3.7.1 Structured coalescent model

Models used by population geneticists and epidemiologists often possess some kind of structure that might, for example, represent two subpopulations (or demes) that are spatially separated due to a physical barrier (such as the sea or a mountain)[24]. Similarly to the “mugration” model, ignoring the influence that population structure has on the tree shape of the inferred genealogies may lead to statistical bias if ignored [25,126]. A way to incorporate some structure in phylogenetic analyses is to use a

structured coalescent model. This model is an extension of Kingman's coalescent model where the unstructured Wright-Fisher population model is replaced by a structured equivalent composed of a number of discrete sub-populations[24,25,127–130].

Such as the Kingman's coalescent model, the structural coalescent model has few assumptions: the different unstructured subpopulations are assumed to have a stable population size trough time, defined by the vector θ. It is also assumed that there is a constant migration between the different populations and there are no differences in fitness between the randomly sampled individuals.

In the structural coalescent model, the backward migration rate matrix m, $m_{b,a}$ is the total rate of migration from the population a to b, divided by the effective number of individuals present in the population b. The aligned sequences are represented by the set $S = \{\frac{s_i}{i} \epsilon\, I\}$, the sampling dates by the set $t_1 = \{\frac{t_i}{i} \epsilon\, I\}$ and the sampling location as $L = \{\frac{l_i}{i} \epsilon\, I\}$. The parameter T represents the genealogy, μ represents the nucleotide substitution rate matrix and M represents the migration history of lineages in the tree (the timing, source, sink and lineage involved in every migration event). Using this notation, to estimate the parameters of the structured coalescent by Bayesian inference, we have to calculate[24,25]:

$$P(T, M, \mu, m, \theta / S, t_1, L) \propto P(S/T, t_1, \mu) P(T; M/t_1, L, m, \theta) P(\mu, m, \theta)$$

In the structural coalescent approach, the coalescent process depends on the effective size and rate of migration between the studied subpopulations[25]. When based on such models, phylogenetic analyses have better model specification in their genealogy inference and parameters estimation[24]. However, because this method infers all the migration history M and all the parameters of primary interest (such as

the migration rates, population sizes and phylogeny), the structured coalescent model is highly computationally demanding and is impractical to use on many occasions.

Therefore, new techniques are developed to approximate the structural coalescent model and overcome the limitations of the phylogeny reconstruction methods currently in use, such as the "mugration" model[25,131]. For example, the BASTA model (BAyesian STructured coalescent Approximation), instead of calculating all possible migration events for the whole history, the model splits it in sub-intervals corresponding to the time between successive coalescent events. Therefore, at the opposite of the structural coalescent model, BASTA assumes that the rate of coalescence between lineages depends on the probability that they are in the same deme at the same time[25]. Therefore, BASTA approximates the same posterior than the structured coalescent model, but integrates it over all the possible migration histories:

$$P(T, \mu, m, \theta / S, t_1, L) \propto P(S/T, t_1, \mu) P(T/t_1, L, m, \theta) P(\mu, m, \theta)$$

BASTA approximates $P(T/t_1, L, m, \theta)$ by using the probability density of each interval between the successive coalescent and sampling events.

BASTA has already been used in some occasions, such as to reconstruct the circulation of the ZIKA virus in Central and South America[27], or to understand the circulation of H7N3 in between the US and Mexico [25,132].

1.3.7.1.1 Combining Epidemiologic and Genomic Data

The transmission and the phylogenetic tree of an outbreak may be thought of as representing the same thing; however, their tree nodes conceptually represent different events. In a transmission tree, node timing corresponds to transmission

events, whereas node timing corresponds to coalescent events in a phylogenetic tree. Consequently, to have a similar transmission and phylogenetic tree, transmission and coalescent timing, and therefore trees typologies, must correspond. This is often the case in a sparse sampling situation, because coalescent events are led by inter host infection dynamic, meaning that transmission and coalescent events happen at the same time. However, as the sampling rate of the hosts increase, the branch lengths in the phylogenetic tree decrease. Consequently, the difference between coalescent time and transmission timing increase at the same rate as the host sampling grow[133].

The combined analysis of epidemiological and sequence data is a challenging task that would help to reconstruct transmission trees for densely sampled outbreaks. Some methods combining the two types of data have been published over the past few years, but typically make some phylogenetic approximations that could lead to bias in the estimation of infection timing[134]. One of those main sources of bias is the existence of unsampled host in phylogenetic studies. There is indeed the possibility that two closely related sampled hosts i and j are not always directly linked, and therefore when calculating probability that i infected j, we should account for the possibility that an intermediate unsampled third host k exists and infected both i and j [135].

However, a recent Bayesian approach named SCOTTI can consider unsampled hosts but also other sources of error (such as multiple infections of the same host). The idea behind this method is to model each host as a separate pathogen population and define transmission events between them as being equal to migration events between those populations. By using available epidemiological information, this model simplifies the number of parameters calculated by rejecting all impossible direct transmission events between populations due to non-overlapping exposure times.

SCOTTI is an extension of the structural coalescent model approximation BASTA, but assumes that population sizes and migration rates between populations are both equal. As a consequence, by combining epidemiological and genetic data and integrating them in a structural coalescent approximation, SCOTTI is able to estimate transmission trees [22].

Similarly to BASTA, Scotti estimates the parameters of the structured coalescent using:

$$P(T, \mu, m, \theta / S, t_1, L) \propto P(S/T, t_1, \mu) P(T/t_1, L, m, \theta) P(\mu, m, \theta)$$

But this time SCOTTI approximates $P(T/t_1, L, m, \theta)$ by using the probability density of each interval between successive coalescent, sampling, population introduction or population removal events[22].

1.3.8 Transmission network analysis from sequence data

A transmission network is a way of depicting infection processes that contain information about the interactions that resulted in the propagation of the disease amongst the studied hosts, locations or populations[3]. Networks are a graphical representation of interacting components composed of nodes (or vertices) joined by links (or edges)[136,137]. Each node may represent an individual, group of individuals or even locations. Edges represent an interaction between nodes that could lead to the transmission of an infection. With a hypothetical perfect knowledge of a disease epidemic, we could fully represent its transmission course by linking each infected host to its infector. This “transmission tree” would therefore perfectly represent the circulation of a disease within an entire population. The use of this network modelling

approach has been first used to reconstruct sexually transmitted diseases[138] and is now complemented with other epidemiological information to study the transmission of other diseases, such as the tuberculosis outbreak in Canada[139,140]. It is important to note that the observed transmission network (or tree) is usually a subset of an underlying contact network that contains all the information about the contacts (their strength, duration, and time of occurrence[140]) existing amongst the potential hosts, whether or not they have resulted in a transmission event[140,141]. Consequently, it is important to note that a single contact network can lead to multiple potential transmission networks and that an observed transmission network depends on its underlying contact network. Thus, a transmission network cannot predict the disease spread in a different outbreak, but it can reveal part of the underlying contact network[142]. Because they are less challenging to estimate (with the use of tracking devices for example[143]), ecologists often use contact networks as proxy for the transmission networks[141].

Combined with other epidemiological data, the use of genetic material could potentially help to reconstruct a transmission network[22,144]. However, a key issue regarding their use is the completeness of the data collected. It will be indeed difficult to reconstruct a transmission network and draw conclusions about its analysis if only a small fraction of infections are sampled[145]. Nonetheless, for most disease outbreaks only a certain percentage of the total infected individuals are sampled[26]. For example, it is possible that two individuals (i and j) with closely related viruses were infected by a shared unsampled infector (the individual k). Thus, the probability that the individual i infected the individual j must account for the possibility that the unsampled individual k infected both i and j. Likewise, there is a probability that the individual i infected an unsampled individual k that then infected an individual j[135]. Therefore, when transmission networks are used to predict an infectious disease spread, analysing an

uncomplete network with missing nodes and edges may have an impact on the quality of the results[146].

Diverse network properties that are relevant for epidemiologically analysis can be estimated[147]. Within a network graph composed of diverse nodes (representing a population or individual) linked by edges (representing any sort of contact or transmission), a community is a set of nodes that has a higher probability of being linked together than any other node[148]. Communities are an important characteristic of complex networks, because they help to understand how a network works by categorising the nodes based on the community they are part of[137]. Those tightly connected nodes may for example represent individuals belonging to a scientific community[149], or represent biological pathways in metabolic networks[150]. Additionally, identifying a node with a central position within their community might mean that it plays an important role or function within their community and might represent the existence of a super spreader[149,151]. Community detection methods aim to identify those structures by only using the information encoded in the graph topology[149] and were first used to identify working group within a government agency[152]. The technique used in this first paper, cutting the edges between the different groups, is still used in several modern community detection approaches[149]. Random walk dynamics, such as the walktrap method, are by far the most used method in community detection. A random walk is defined as a route were a node reached at a time t is a random neighbour of the node reached at the time t-1. Additionally, with this approach, the distance between two nodes is the average number of edges that a random walker must pass to link them[153]. Close nodes have a high probably of being part of the same community[154]. Once within a community, a random walker would spend a long time within it due to the high density of edges found within this kind of structure and the few edges leaving this group of nodes[148,155].

In the walktrap approach developed by Latapy and Pons[156], the distance between the nodes i and j is defined as the probability that a random walker moves from i to j in a fixed number of steps t. This distance must be large enough to explore a significant portion of the graph, but not too big as to avoid a stationary limit where the transition probabilities only depend on the degrees of the node[153]. Therefore, the edges present in the same communities will be more tightly connected by a random walk process than nodes that are parts of different communities. Using this approach, the nodes similarity is expected to be considerably higher within groups than between groups[148,153], and communities can be identified using standard hierarchical or partitional clustering techniques[148,154,157]. Walktrap approaches are highly computationally demanding and therefore cannot be used on large networks[25,3].

1.4 INFLUENZA VIRUS

1.4.1 Challenge

Over the last century, four different human influenza pandemics have occurred, causing millions of deaths among humans[158]. Amongst the strains previously involving those human plagues, such as the 2009 flu pandemic, some had genetic elements of avian origin [44,159,160]. It was formerly hypothesised that in order to infect humans, avian influenza strains had to pass through a swine population. However, it is now well understood that direct transmission from birds to humans is a recurrent event[161,162] raising concerns about the possible emergence of new high-mortality human influenza strains of avian origin with the ability to be easily transmitted amongst humans populations[163]. Due to multiple underlying socio-economic, environmental and ecological factors (such as the intensification of the livestock industry or the

climate change), the emergence of such zoonoses, easily transmitted amongst animal and human populations, is expected to increase in the upcoming decades[158,164]. This statement is especially true for developing regions such as China, India, or other parts of Southeast Asia[165]. Because those zoonotic pathogens represent a threat for human and wildlife populations, it is therefore of primary importance for the scientific community to study the processes leading to their emergence[166].

1.4.2 Structure of the influenza A virus

Influenza viruses are in the family of Orthomyxoviridae. Four different types of influenza viruses exist: A, B, C and D. Amongst them the influenza type A virus has the largest pool of hosts (amongst other human, birds, pigs, horses and crocodiles), whereas the other types only have a limited number of potential hosts (humans for influenza B, humans and pigs for influenza C and pigs and cattle for influenza D) [167,168]. In this thesis, I will focus exclusively on influenza type A virus strains.

The genome of the type A influenza virus comprises eight negative-sense, single-stranded viral RNA (vRNA) segments (see Figure 1-5)[169]. Its genome encodes for multiple proteins encoded on its different segments: three polymerases (PA on segment 3, PB1 on segment 2 and PB2 on segment 1), a nucleoprotein (NP on segment 5), two matrix proteins (M1 and M2 on segment 7), two non-structural proteins (NS1 and NEP/NS2 on segment 8) and two surface antigen proteins, hemagglutinin (HA on segment 4) and neuraminidase (NA on segment 6)[170,171]. The RNA replication and transcription are done by the ribonucleoprotein (RNP) complex which is composed of the nucleoprotein and the three polymerases [159].

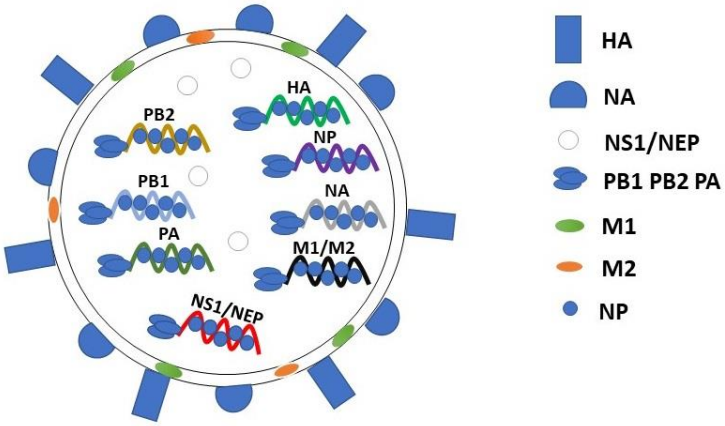


Figure 1-5: Schematic representation of the avian influenza A virus. Influenza virus is enveloped with HA and NA surface proteins. The genome is composed of 8 negative sense RNA segments, each one coding for a different set of proteins.

Each influenza A strain has one type of HA and NA glycoproteins at their surface. There are 18 subtypes of HA (H1-H17) and nine NA subtypes (N1-N9) which differ by at least 30% in their amino acid sequence[170]. Different combinations of those surface glycoproteins are possible[172] and define the influenza A subtypes. Apart from the newly discovered bat-specific H17, H18, N10 and N11 proteins[173,174], almost all of the possible HA/NA combinations have been found in birds, while only a few have been reported in other host populations[175].

Two mechanisms can change the nature of the HA and NA proteins found at the surface of the virus: antigenic shift and antigenic drift. Antigenic drift simply refers to the single amino acid changes found at the antigenic sites of the NA and HA proteins. Antigenic shift refers to exchange of genome segment amongst strains coinfecting the same cells[169] caused by the segmented nature of the virus genome. This mechanism can lead to the acquisition of new antigens by viral reassortment, allowing the virus to efficiently evade the immune system of the previously infected individuals.

This phenomenon will ultimately facilitate the infection to be established and the following transmission events to occur.

1.4.3 Type of influenza virus

Two types of influenza strains can be distinguished based on their virulence and mortality in chickens[170]. Low pathogenic influenza (LPAI) strains may cause a reduction in weight gain or a decline in egg production in poultry[40], whereas highly pathogenic influenza (HPAI) can have 100% mortality within 48 hours in poultry[176]. However, some duck species may show no or limited symptoms while infected by HPAI strains[177]. The difference in pathogenicity between strains comes from the insertion of multiple basic amino-acids in the cleavage site of the precursor of the mature hemagglutinin[170]. In LPAI strains, the type of proteases able to cleave this maturation site are “trypsine-like” and are present in host respiratory and intestinal organ cells. However, the “furine-like” protease that cleave HPAI HA precursors may be found in numerous cells type throughout the entire host organism. This difference in HA cleavability affects in which tissue the proteins can be activated and is considered as the main outcome determinant of a viral infection[178].

HPAIV strains have only been found amongst the H5 and H7 subtypes[40]. New HPAI strains are considered to emerge from LPAI strains after their transmission from wild birds to domestic bird populations where it can slowly gain in pathogenicity through successive infection cycles[179]. This transformation might take from a few days[180] to months, depending on the outbreak[181].

1.4.4 Reservoirs of influenza type A viruses

Within each one of Influenza A virus potential hosts, we can observe a continuous virus circulation and only a sporadic transmission between them[170,182]. Additionally, by showing signs of a potential co-evolution with the avian influenza A, water birds, especially Anseriformes (ducks, geese, and swans) and Charadriiforms (gulls, terns, sandpipers) are believed to be the natural reservoir of the influenza A virus[44,170].

1.4.5 Avian influenza

1.4.5.1 Influenza in wild birds

Until now, influenza A viruses have been isolated from at least 105 wild bird species from 26 families[175]. Almost all HA and NA antigenic subtypes can be found in the Anseriformes order. Because we can observe higher levels of infection in mallard ducks (Anseriform) than in any other wild bird species[183], they potentially constitute an important reservoir of influenza virus strains[44]. Moreover, by being asymptomatic Eurasian-teals and whooper swan birds could play a role in the long-range avian influenza propagation through their seasonal migrations[184]. In North America and Eurasia, mallard infection levels are subject to a seasonal variation ranging from 60% in autumn, right after the breeding period, down to 10% in spring[183,185]. This variation is thought to be caused by the arrival of juvenile birds, acting as naïve population for the virus, but also by environmental conditions, that could influence the virus survivability outside the host [44].

Charadriiforms (or Shorebirds) birds, such as gulls and waders, demonstrate lower levels of infection of avian influenza A than Anseriformes birds, but are also healthy

carriers of the virus[186] and long range migrant birds meaning that their movements could strongly affect influenza distribution[44,187,188]. At the opposite of what we observe in North America for Anseriformes birds, the highest infection for Charadriiforms birds can be found in the late-spring and early-summer. This suggests that Charadriiforms birds could play a role in the overwinter maintenance of avian influenza prior to spreading the virus to the northern breeding areas in spring where Charadriiforms birds could infect other bird populations[44]. Because there is no significant genetic difference between the influenza A strains found in Anseriformes and Charadriiform populations, the two species groups might not be separated, and the virus could circulate between them[175].

1.4.5.2 Influenza A viruses in domestic birds

There is a continuous co-circulation and reassortment of influenza A strains within domestic bird populations. Some of those strains are phylogenetically and antigenically related to previous human-derived influenza strains and can therefore represent a potential source for human infection[161]. Not all domestic bird species have the same susceptibility to influenza A virus. Experimental infections have shown that quails, pigeons, geese and ducks show few clinical signs of infection[176,189,190]. At the opposite end of the spectrum, gallinaceous birds are usually dead within 48 hours following the infection with a highly pathogenic influenza strain[40,177]. During a recent H7N9 LPAI outbreak, domestic ducks were identified as key intermediate hosts between wild birds and chickens[191]. Furthermore, it has been shown that domestic ducks could play an important role in HPAI strains persistence and emergence[189,192,193]. Domestic ducks share the same habitat, water and food than wild

waterfowl birds and are thus easily infected by those, hence the importance of domestic ducks in the epidemiology of the disease[193].

1.4.5.3 Transmission and Host switching

Because LPAI and HPAI strains can be found in different parts of the infected host's body, they have different mechanism of excretion. In the case of a LPAI strain, the virus excretion is mainly done through cloacal shedding. In contrast, HPAI virus' shedding is predominantly done via respiratory tract[194]. This indicates a difference in transmission potential between the two kinds of strains. The transmission of a HPAI strain such as H5N1 could involve close-contact route of transmission as opposed to an indirect faecal-oral transmission through the environment for LPAI strains.

Once in the environment, the virus may survive for some time depending on parameters such as temperature, salinity, pH and type of substrate (such as faeces, sediments, meat). At low temperature, infected faeces and meat may remain infectious for an extended period, up to a few months, and may act as source of infection for susceptible birds[195]. Therefore, different ways of indirect transmission are possible: through ingestion of infected water, inhalation or contact with untreated poultry faeces used as fertiliser[196].

Important risk-factors of influenza transmission have been identified such as the proportion of free grazing ducks and rice paddy fields (where wild and domestic birds share the same environment), the number and connectivity of live bird markets and the presence of water. All those factors might influence the persistence and spread of the virus in domestic bird population[158,197,198]. Other factors such as the breed susceptibility, the virulence of the strain and the number of young individuals within a

bird population might also influence the propagation of the virus[199]. There is some strong evidence that primary infections in domestic bird populations emerge from the introduction of new influenza strains by wild birds[200]. However, it is still assumed that the main source of secondary infections in domestic populations comes from human actions. It has indeed been shown that in various occasions and settings, the movements farm owners and staff, trucks and drivers moving birds or farm material and level of biosecurity encountered in the infected areas have been shown to influence the spread of the virus between and within farms[172,201].

1.4.5.4 Global spread of avian influenza

The worldwide spread of HPAI probably resulted from the synergy between trade of infected domestic birds and wild bird movements[202]. The influence of wild bird migratory flyways on the virus circulation has been shown on multiple occasions[203,204] using remote sensing analysis and phylogenetical analysis. For example, it has been shown that the circulation of H5N1 in Eastern Asia follows the wild birds migratory flyways[205] and that wild birds following the North American flyways were responsible for the introduction of H7N3 introduction into Mexico[132].

We can see the effect that those migratory flyways have on the circulation of the disease by looking at the worldwide phylogeny of all avian influenza. If we do so, we can easily distinguish two major clades, corresponding to the American (North, Central and Southern American continent) and the Eurasian (Asia, Europe, Africa and Oceania) avian influenza clades[170]. However, cross-flyway and inter-continental spread of avian influenza still occurs[39]. For example, the East Asian-Australasian flyway connecting South Asia to the Pacific Americas has already been shown to be

involved in the inter-continental transmission of LPAI and HPAI [187,206,207]. Similarly, transmission of avian influenza from Americas to Eurasia has been shown to happen through movements of long range migratory wild ducks[208,209].

However, even if migratory birds might be good vectors, transmission patterns indicate that it is at least partially maintained through trade of infected Galliforms (chickens) [210,211]. In this regard, the 2004 HPAI H5N1 long distance expansion was found to be caused by human movements of domestic poultry[211], while the disease circulation was found to be driven by human activities in China [114].

1.5 FOOT AND MOUTH DISEASE VIRUS

1.5.1 Background

Foot and mouth disease (FMD) is one of the most important viral diseases of cattle globally. It has been eradicated in many wealthy nations, but it is still endemic in many countries, especially lower and middle income countries (LMICs), that all together, contain three-quarters of the world's human population[212]. However, disease-free countries are still under the threat of sporadic incursions of FMD virus, which may have severe economic impact such as during the 2001 FMD epidemic in the United Kingdom [213–215]. Multiple spatial and environmental attributes (e.g. type of land use or host density, environmental and climatic conditions) indirectly influence the dynamics of the infectious disease by impacting the distribution and movement of its potential hosts[30,33]. Therefore, understanding how the virus circulates and which factors influence its spread are not only a prerequisite for an efficient control in

endemic countries, but also to prevent sporadic introductions into disease free countries.

1.5.2 Virion Structure

Foot and mouth disease (FMD) is a vesicular disease which can infect more than 70 species of cloven-hoofed animals, including domestic ruminants and pigs [216]. The causal agent of FMD is a positive-sense, single-stranded RNA virus called foot and mouth disease virus (FMDV). This virus is classified as an Aphthovirus and is a member of the Picornaviridae family [217]. Its genome encodes the information for 4 structural (VP1-4) and 8 non-structural proteins (7 proteases and one RNA polymerase) (see Figure 1-6). Its genome is divided in three regions: a non-coding regulator region, a protein-coding region and a non-coding regulator region. The viral particles are composed of 60 copies of each of the four structural proteins, with the VP1-3 proteins being external structural proteins and the VP4 protein being internal, in contact with the viral genome [218]. The exact function and interaction with the proteins encoded within the FMDV genome are areas of active research [219]. Due to the absence of a proofreading-repair activity, the FMDV genome is subject to a high substitution rate estimated at 3-7x10-3 per site per year [218].

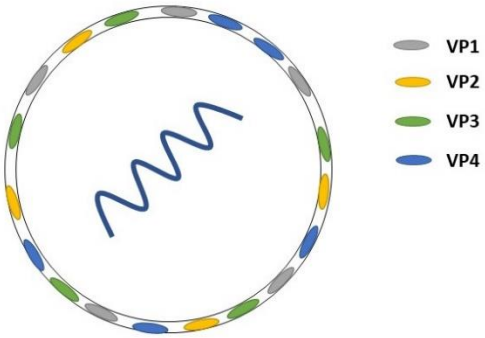


Figure 1-6: Schematic representation of foot and mouth disease virus. FMDV is non-enveloped and has a capsid structure composed of the VP1,2,3 and 4 proteins. It has a positive sense single stranded RNA genome.

Based on the level of cross protection between each strain, the viral population is divided into seven distinct viral serotypes: O, A, C, Southern African Territories [SAT] 1, SAT 2, SAT 3 and Asia 1[218,219]. Overall, FMDVs are contained in three distinct epidemiological clusters: Africa, Asia and South America. They are further divided into seven inter-connected major virus pools[219,220]. Africa has three of these pools: East Africa constituting 'pool 4' with the presence of serotypes A, O, SAT-1, SAT-2 and SAT-3; West Africa is covered by 'pool 5' with the serotypes A, O , SAT-1 and SAT-2; and Southern Africa has 'pool 6' with the presence of the serotypes SAT-1, SAT-2 and SAT-3[221]. These pools only relate to countries currently infected with FMDV. Moreover, based on VP1 structural differences between geographical localities, each of the serotypes can be subsequently subdivided into 'topotypes' [220,222]. In Africa, WRL/OIE define two topotypes for serotype A, six for serotype O, three for serotype C (not reported in Africa since 2004) and 9, 14 and 5 topotypes for serotypes SAT-1, SAT-2 and SAT-3[220,221].

1.5.3 Host range

FMD is a severe, clinically acute, vesicular disease that can infect more than 70 different species of cloven-hoofed and Camelidae animals [219,223,224]. Three categories of hosts can be considered in the epidemiology of FMDV: those who play a role in the natural epidemiology of the virus (cattle, buffaloes, pigs, sheep and goats), those that may play a role under certain conditions (deer, camels, llamas and alpacas) and hosts that are susceptible to infection but without importance under field conditions (horses and carnivores) [225].

1.5.4 Pathogenicity of foot and mouth disease

FMD susceptibility varies with the host species and the strain involved. But the severity of the infection depends on the amount of virus inoculated, its serotype as well as the host species and individual immunity[216,226]. In this regard, it has been proven that acute clinical FMD is more severe in pigs than in ruminants. However, pigs are more likely to clear the infection[225] and are therefore less likely to develop a "carrier" state than cows[227,228].

The disease is characterised by the development of vesicles in and around the mouth, on the feet and other sites of the skin, loss of appetite and fever. Lesions are usually observed as white areas that develop into vesicles. The mortality following the infection is usually low but might be higher in young animals and calves, due to high level of myocarditis[216]. In such cases, calves often suffer from a "tiger heart", a state characterised by the presence of a soft, flaccid heart with white or greyish stripes[216]. The relation between those myocarditis in young calves and the spread of the disease is still not well understood [216].

Within the first 24 hours of infection, the virus can be present in the respiratory epithelium, subepithelium, and interstitial areas of the lung of infected cattle. After 72 hours, the virus can be detected in other tissues such as the tongue, soft palate, feet, tonsils, and tracheobronchial lymph nodes[229,230]. However, conflicting observations suggest the pharynx instead of the lung as initial site of viral replication, with multiple parameters such as the aerosol particle size, strain of virus, or how the aerosol was generated which influence the primary site of viral infection[216]. Overall, we still have a parse knowledge of the specific mechanisms responsible for acute clinical signs[216].

1.5.5 Impact of foot and mouth disease

Since the animal species that are most significant in the natural epidemiology of FMDV are of major importance in the production of food (cattle, sheep, pigs, goats), the disease usually leads to important direct and indirect economic impacts in countries with a developed agricultural industry [225,231,232]. Direct economic losses are caused by mortality and decreased animal production whereas indirect economic losses are caused by trade restrictions and the cost of control measures (e.g. vaccination, cleaning, movement control)[219]. Furthermore, due to the market exclusion of countries where FMD is detected, the disease can have an impact on food security by impairing livestock movement and trade[220]. Therefore, even if FMD is a low mortality disease, the number and the species of the infected hosts can have a high and continuous impact for the infected countries [232].

Even though the impact of the disease in industrialised countries is well recognised, there is still a lack of data to quantify the consequences of FMD in developing countries where the virus is often endemic. Although work has been done to

understand the impact of FMDV in large scale dairy farms in low and middle income countries[233], there is still a lack of data to quantify its broad impact on the economy of endemically infected countries[234]. However, it is almost certain that the disease has important consequences in traditional extensive systems as a result of the direct and indirect effects on the economy of infected countries[234].

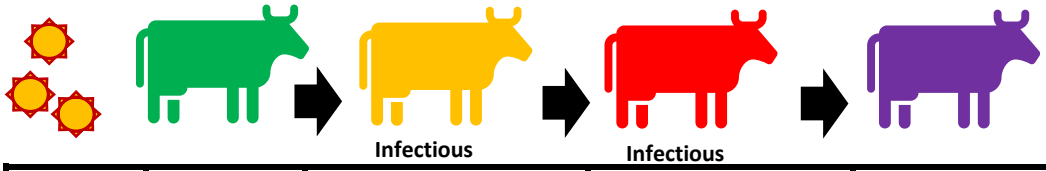
1.5.6 Transmission

1.5.6.1 Routes of transmission

Most of our knowledge on FMDV transmission routes comes from studies involving simulated natural methods or animals infected by artificial methods[225]. Most of them involve the inoculation of animals with FMDV and the study of the transmission to susceptible animals in a controlled environment.

Generally, the transmission of the disease between susceptible hosts is done mechanically, through viral entry via skin cuts or mucosae following physical contact with infected secretions or excretions[216,235]. Following the infection by FMDV serotype O strain, the virus can be detected in all infected cattle excretions and secretions, such as saliva, nasal and lachrymal fluid, milk, expired breath and, to a smaller extent, urine and faeces[216]. Following a short latent phase of one day, the viral excretion follows three distinct phases: a subclinical infectious phase lasting between 1 and 3 days where the host does not show signs of infections, a clinical infectious phase lasting from 6 to 11 days where the host shows signs of infection and, in some cases, a carrier state where the virus is excreted at low levels without signs of infection[216,236]. In total for cattle infected by a serotype O, the incubation phase lasts on average 3.6

days and the clinical phase lasts on average 8.5 days, making the total infectious phase lasts for 10.7 days overall[237] (see figure 1.7).



<table>
<tr><td rowspan="3">Exposure</td><td rowspan="2">Latent
(1.6 days)</td><td>Subclinical infectious
(2.2 days)</td><td>Clinical infectious
(8.5 days)</td><td>Carrier</td></tr>
<tr><td colspan="2">Total infectious (10.7 days)</td><td rowspan="2">Convalescent</td></tr>
<tr><td colspan="2">Incubation (3.6 days)</td><td>Clinical disease
(8.5 days)</td></tr>
</table>

Figure 1-7: FMDV progression in cattle with its three distinct phases of infection. Following the exposition to the virus, infected cattle enter an incubation and a clinical disease phase. During the first phase, where no signs of infection can be seen, cattle go through a latent and subclinical infectious phase. The total infectious phase includes the subclinical and clinical infectious phase. Figure adapted from Yadav and al. 2019 [237].

FMDVs are able to persist in the environment in low temperatures and high humidity, and contact with secretions or excretions from an infected animal indirectly through environmental contamination or mechanical transfer by people, animals or vehicles can lead to indirect short distance transmission of viral particles[216,225,235]. However, the most common transmission mechanism between ruminants and pigs is by the circulation of infected animals and the aerogenous transmission of infectious droplets they exhale[238]. The major determinants of such small scale airborne transmission are the strain of virus and species of infected animals involved[238]. In the field, such transmission mostly occur from pigs to cattle due to the high concentration of virus liberated by pigs and the large amount of air inhaled by cattle[216,239]. Theoretically, long-distance spread, through airborne propagation, is possible under well-known specific conditions[238,240,241] that depends on wind direction and strength, temperature, relative humidity and geographical topography. Predictive models have shown that way of propagation to be quite uncommon in temperate conditions and extremely

unlikely in arid conditions. However, this phenomenon is still plausible and has been shown to play an important primary role during the 2001 UK foot and mouth disease outbreak[242–244]. It is important to notice that pigs are also less susceptible to airborne infection compared to ruminants[227]. For example, when exposed to naturally generated aerosols of the O1 Lausanne and O SKR 2000 strains, doses of at least 6000 $TCID_{50}$ for O1 Lausanne and 1000 $TCID_{50}$ for the O SKR 2000 were needed to infect pigs[245]. At the opposite, cattle and sheep can be infected by a dose of 10 $TCID_{50}$ [246,247]. The pathogenicity of the disease differs also between topotypes. Pacheco and Mason[248] show that in inoculated pigs (through the heel bulb of each major digit of each foot with 10^2, 10^3, 10^4 or 10^5 PFU of virus/5 µL), strains coming from the Cathay topotype are more virulent than those from the PanAsia topotype.

It is considered that movements of animals and animal products are the main factors of long-range FMDV transmission[220]. In endemic areas such as in Africa, movement of animals and animal products can be an important cause of FMDV spread. In contrast, introduction of FMDV in previously disease-free countries has often be related to food waste when the primary source of infection cannot be linked to the movement of animals. When this situation occurs, such as what happened during the 2001 FMD UK epidemic, it is likely that pigs constitute the primary source of infection caused by the ingestion of infected animal products [249–251].

For cattle, the peak of viral production coincides with the presence of clinical signs of infection, whereas in sheep it occurs mainly in the pre-clinical period[252]. Depending on the host species and the viral strain, there might be large variations in the amount of virus produced by infected animals [245,253]: pigs are the largest producers of aerosol virus, whereas cattle and sheep usually shed 3000 times less virus than pigs and are therefore considered as less important sources of aerosol virus[241]. Finally, after

approximately 4 to 5 days, the viral load declines, as the antibodies and other host immune responses lead to control of the infection within the host [252].

1.5.6.2 Carrier state transmission

Animals in the carrier state are defined as asymptomatic continuously infected animals that have recovered from acute FMD; have been vaccinated, or have been exposed, and in which FMDV can be isolated in the oropharynx for more than 28 days after the infection by the disease[254]. Since carrier animals are asymptomatic, they may remain undetected for a long period of time and represent a major challenge for FMDV control, since they theoretically have the potential to trigger new outbreaks months after the apparent control of the disease[252,255]. Both the host species and viral strain have an influence in the development and maintenance of this state[216]. Depending on the species and other factors such as the immune status of the host population, up to 50% of infected ruminants may become carriers[216,256]. In free-living African buffalo populations, the prevalence of carrier animals can be up to 60-70% [257], and in endemic areas the prevalence of carrier animals in domestic cattle can be around 20%[258]. In the carrier state, the virus can be isolated for up to 3.5 years in cattle, 9 months in sheep, 4 months in goats and 5 years in African buffaloes[216]. Although the presence of the carrier state is an integrated part of FMDV pathogenies, little is known about the mechanisms leading to the apparition of this state[259]. However, in an experiment involving naïve and vaccinated cattle, Stenfeldt and al. determined that the carrier state is associated with the suppression of the host antiviral response and the persistence of the virus in specific nasopharyngeal epithelial sites[259,260].

The potential for carrier animals to provoke new outbreaks of FMD is controversial and has been debated for years [224,261]. Experimental studies involving cattle and swine have shown that FMDV obtained from carrier animals might be less virulent than samples from acute infected hosts[262–264]. Indeed, it seems difficult for carrier animals to transmit the virus to susceptible animals, and only transmission of SAT serotypes from carrier buffaloes to cattle has been demonstrated under both experimental and natural conditions [265,266]. Apart from that, transmission from other persistently infected livestock or wildlife to susceptible animals has never been proven[226].

1.5.7 FMD situation in Africa

In Africa, the epidemiology of FMD is considered to be more complex than in any other region of the world. On this continent, six of the seven possible serotypes are found, with high regional variances in both their distribution and serotype prevalence [267]. In total 9, 14 and 5 topotypes have been defined for SAT 1, SAT 2 and SAT 3 respectively in Africa[268]. In Africa, SAT 2 seems to be the more widely distributed serotype, whilst serotype O is the least prevalent and has not been observed since 2004. The serotype prevalence in Africa can be broadly described by SAT 2 > O > A > SAT 1 > SAT 3 > C [267].

It is mostly accepted that the origin of FMDV is in Africa due to observed long-term subclinical infections of African buffaloes and the greater genetic diversity of the SAT serotypes compared to all other FMDV serotypes[234,267]. Furthermore, all FMDV serotypes, apart from Asia1, have been recorded on the African continent [221,267]. The African continent is divided into three of the FMDV pools: East Africa (serotypes O,

A, SAT-1, SAT-2 and SAT-3), West Africa (serotypes O, A, SAT-1 and SAT-2) and Southern Africa (serotypes SAT-1, SAT-2 and SAT- 3)[234,269].

Even if most FMD outbreaks remain unrecorded, FDMV is considered endemic in almost all African countries. This situation might be partially explained by the extensive livestock-raising system which is practised in several African regions, leading to a low direct impact from the disease. Therefore, many African livestock owners and authorities do not consider FMD as a serious disease and do not invest in its report or control[267]. However, the economic impact of the disease in such pastoralist and agropastoralist settings is still present, with a proven association between FMDV presence and lower milk yield for rural smallholders[270]. Moreover, sub-Saharan Africa is generally poorly equipped to face transboundary disease spread due to the lack of financial resources, military conflicts and need to answer other problems (e.g. education, human health)[267].

1.5.7.1 FMD distribution in Africa

In North African countries (Morocco, Western Sahara, Algeria, Tunisia, Libya and Egypt), the spread of the disease is suspected to be mainly through livestock trade of domestic animals[221,271]. For countries in the western part of this region (Morocco, Algeria and Tunisia), FMD occurrence is sporadic and well controlled by vaccination and animal movement restrictions. Even in those countries with limited trade, movements of small infected populations may have played a role in the observation of individual infection cases[221]. During the period 2003 to 2012, FMD became endemic in Libya and Egypt, possibly caused by importations of infected livestock from FMD endemic countries.

West Africa is composed of Mauritania, Mali, Niger, Cape Verde, Senegal, Gambia, Guinea-Bissau, Guinea, Sierra Leone, Liberia, Côte d'Ivoire, Burkina Faso, Ghana, Togo, Benin and Nigeria. Central African countries are Chad, Cameroon, Central African Republic (CAR), Equatorial Guinea, Gabon, Republic of Congo and the Democratic Republic of Congo (DRC). East Africa countries are Sudan, Eritrea, South Sudan, Ethiopia, Djibouti, Somalia, Kenya, Uganda, Tanzania, Rwanda and Burundi. Historically, six out of the seven serotypes (O, A, C, SAT-1, SAT-2 and SAT-3) can be found in these three regions and currently five out of the seven serotypes can be found in those regions (O, A, SAT-1, SAT-2 and SAT-3). Moreover, two out of the three FMDV virus pools present in Africa (four and five) can be found in this part of the continent, with a considerable co-occurrence in virus serotypes and topotypes.

Serotypes O, A and SAT-2 can be found throughout those regions, with a limited presence of SAT-1 in East and West Africa. Serotype SAT-3 was isolated on one occasion in African buffaloes in Uganda and has never been isolated in livestock. It seems that the only region were FMDV serotype C was isolated in recent times was in Kenya in 2004 in cattle population [221,267].

In the Southern Africa region (Angola, Zambia, Malawi, Namibia, Botswana, Zimbabwe, Comoros, Mozambique, Madagascar, Swaziland, Lesotho, South Africa, Mauritius and the Seychelles), FMD circulation in livestock is clearly influenced by the livestock–wildlife interface, and more particularly the presence of African buffaloes.

From 1991 onwards, four serotypes (O, SAT-1, SAT-2 and SAT-3) have been found in the region, with a high presence of the SAT 1–3 serotypes, but with a distribution of SAT-3 limited to South Africa[221]. In South Africa, serotypes O strains were shown to be imported from Europe/South America and the Far/Middle East. However, almost all the FMD outbreaks in Southern Africa were caused by FMD SAT serotypes [272].

1.5.7.2 Agricultural system and livestock trades in Africa

Five different livestock production systems involving increasing levels of animal movements can be found in Africa[212]: total nomadism, semi-nomadism, transhumance, partial nomadism and sedentary animal husbandry. Along the Mediterranean coast of North Africa, a transhumance system with some sedentary small holder farming can be found. In the Sahel region, nomadic, semi-nomadic and transhumance systems predominate. In sub-Saharan Africa, all five systems can be found, with nomadism, semi-nomadism and transhumance systems being the most widely practiced. In Southern Africa, the livestock production system is mainly sedentary. In the nomadic, semi-nomadic and transhumance systems, domestic animals can move freely during long international border travels. During the driest parts of the year, those moving animals are more exposed to wildlife around rivers and watering points, where grazing is still possible, increasing the risk of FMDV transmission between wild and domestic animals[221].

Livestock trade is one of the main economic activities on which the poorest parts of the African population depend. However, statistics on those trades are still limited (FAOSTAT.fao.org). Furthermore, available data do not often take into account informal trades from the nomadic and transhumance herds[221]. Therefore, most of the trade is unrecorded and it is indeed estimated that only 10% of all cross-border trades are done through official channels[220]. However, we know that there are well established livestock trade routes across the sub-Saharan region, with Nigeria as the biggest African importer of meat due to an increasing demand of protein boosted by a growing human population. The greater horn of Africa acts as an important animal supplier for the Arabian Peninsula and the Gulf States, and represents the main destination for livestock trading routes within East-Africa, with Sudan acting as a key

commercial intermediate between the Western and Eastern parts of the continent [220,221]. Somalia, Sudan and Ethiopia are the continent's largest sheep meat exporters, whereas Southern Africa is the main beef exporter. Not all exported animals are originating from those exporting countries and are often transported from surrounding countries. Therefore, all those trades lead to massive national and international animal movements within the continent and are of major importance for the FMDV dissemination in Africa [221,273].

1.5.7.3 Livestock/Wildlife interface in Africa

Amongst all wildlife species in Africa that FMDV can infect, only buffaloes and impalas were implicated in the transmission of the virus toward cattle [226,274].

In Ethiopia, the cattle populations with the highest FMDV antibody prevalence were those in close contact with wild animals, and located near wildlife sanctuaries where large populations of African buffaloes can be found[275]. In Southern Africa, buffaloes are suspected to be the source for livestock outbreaks, and particularly the SAT-type FMD viruses, and its true maintenance host[224,265]. However, there is still a lot of uncertainty regarding the role of buffaloes elsewhere in Africa as a source of livestock outbreaks, and how they are able to sustain endemic cycles of infection in livestock[234]. Additionally, there is still no proof that African buffaloes have a role in the epidemiology of the FMDV serotypes O, A and C [221].

Impala populations, in contrast to African buffaloes, are not thought to exhibit a carrier state and are therefore suspected to be accidental hosts that act as intermediates between buffaloes and cattle populations[276]. It is justified by the fact that in impala populations, the infection prevalence can climb up to 90% between June and

November, when buffalo calves lose their maternal antibodies and become infected with FMDV, meaning that impalas potentially get the spill over from the infections in naïve buffalo calves [277].

The existence of large wild free ranging populations of cloven-hoofed species in Africa has led to the establishment of important wild reservoirs of FMDV in this part of the world. For several years now, African impalas and buffaloes have been recognised as major wild reservoirs of FMDV and carriers[224,278]. When these animals get in close contact with domestic livestock, they provide a possibility for viral transmission that might lead to an FMD outbreak[225].

1.5.7.4 Landscape influence on the transmission of FMDV in Africa

Sub-Saharan countries possess common characteristics associated with FMD virus incursion and maintenance, all mostly related to wild animal proximity and movements[273,279] .

The landscape epidemiology of endemic FDMV between cattle and buffaloes has been studied in the Kruger National Park situated in South Africa by Dion, VanSchalkwyk and Lambin [33]. They developed a spatially explicit agent-based model to understand the effect of the environment on the spatial and seasonal distribution of transmission risk between the cattle and buffalo populations. The aim was to understand the direct and indirect impact of the environment on animal movements, used as proxy for transmission risk. Therefore, the impact of different factors related to climate, location of water points, growth of human settlements, controlled burning, location of fence breaks, and landscape configurations were assessed.

In this study, Dion and al. showed that an increase in the number of cattle–buffalo contacts was mostly dependent on the number of holes in the park fences, the presence of human settlements and the movement extent of the cattle and buffaloes, influenced by landscape configuration. One factor, spatial homogeneity, was associated with a decrease in the number of contacts between the two populations. Most of the observed contacts were taking place close to water points and grazing areas, far from human settlements.

Even if the transmission from wild buffaloes to domestic animals is generally supposed to be the dominant pathway of disease propagation [216,274], other important ways of transmission exist [279]. Using data from passive surveillance for 2001-2006, a risk factor analysis was set to assess the factors associated with FMDV occurrence in Tanzania[279]. Here, increased distances to the main roads were associated with the highest effect in FMDV occurrence, followed with increased distances to the main railways and international borders. Moreover, this study also showed that cattle density was globally associated with an increase in FMD occurrence. Therefore, it was supposed that in Tanzania, FMD incidence was mostly related to domestic animal movements and human activities and seems not to be caused by transboundary movements of infected animals or contact with infected wild animals, the factors that are usually associated with FMDV incidence in domestic animal populations.

A similar study was done in Zambia for the 1981-2012 period[280]. In this case, a decrease in FMD-outbreak occurrence was linked to distances to international borders, nearest major road, wetland areas, as well as with index of wetness and elevation. Moreover, a link between distances to the major railways and FMDV occurrence was found and assumed to be caused by its spatial correlation with another, potentially more direct, determinant of disease, such as human density along the railways.

1.6 RESEARCH QUESTIONS

The two studied viruses in this thesis are both zoonotic diseases that are transmitted at the wild-domestic animal interface. Both viruses may lead to important economic losses when circulating in domestic populations with the potential to be transmitted to humans. In a constantly changing world, land-use and environmental changes bring wild and domestic animals closer to each other. Therefore, understanding the circulation of zoonotic diseases at the interface between wild and domestic animals is important to mitigate potential economic losses and reduce the risk that such diseases represent for human health.

Although the transmission patterns between wild and domestic birds have been widely studied, most approaches used external proteins (HA or NA) to study the circulation of the virus. However, by analysing reassortment and transmission patterns of LPAI and HPAI strains, using internal segments has the potential to gain new insights in the virus epidemiology. New phylogenetic approaches enable us to reconstruct potential viral transmission networks, allowing us to gain insight in the timing and routes of transmission that occurred in the studies' timeframe[281]. Therefore, in Chapter 2, by applying such approaches, I tried to determine the circulation and reassortment patterns of the Avian Influenza between Asia and Europe. I especially focused on the timing and role played by different bird populations on the transmission events between the two continents.

Using phylogenetic approaches, there are ways of quantifying the influence that different environmental and anthropological features have on the circulation of infectious diseases using real world data[34,115,282]. Such approaches have been partially used to study the influence of multiple environmental factors of avian influenza circulation[114], but have never been applied to study the influence of

environmental factors on the circulation of FMDV. Therefore, in Chapter 3, I estimated the impact that various environmental and human features have on the circulation of FMDV in sub-Saharan Africa using viral sequences coming from domestic animals. Moreover, there are still numerous questions regarding the impact that wild animals such as buffaloes and impalas have on the circulation of FMDV in livestock. This phenomenon is exacerbated by difficulties to gather virus samples from wild animals. Therefore, new phylogenetic approaches able to cope with sampling imbalance were applied to study the circulation of FMDV between wild and domestic animals in Southern Africa in Chapter 4.

Finally, although more complex methods are available to perform phylogenetic studies, such approaches are often more computationally intensive with more parameters to set or estimate to study the circulation of diseases amongst multiple populations. Therefore, there is a growing need for simple methods able to deal with the growing number of available sequences, whilst being statistically robust to deal with potential sampling imbalance. I tried to answer those different issues in Chapter 5 of this thesis with a new approach able to cope with those issues.

t

2 SPATIAL, SPECIES AND SUBTYPE NETWORK RECONSTRUCTIONS FOR AVIAN INFLUENZA VIRUSES BETWEEN EUROPE AND ASIA

2.1 ABSTRACT

Avian influenza is a worldwide disease, causing significant economic losses in domestic bird populations with the underlying threat of direct transmission to humans. Multiple subtypes of avian influenza virus co-circulate in avian populations, and the viruses acquire different subtypes of the surface proteins, hemagglutinin (HA) and neuraminidase (NA) through reassortment. Using viral sequences and phylogenetic techniques, migratory birds have previously been implicated in the spread of influenza between Asia and Europe, with Eastern Asia, and especially China, acting as a source of infection for Europe. Using phylodynamic techniques, I reconstructed the general transmission network of avian influenza viruses (AIV) within Europe and Asia, and studied the role of migratory and domestic birds in the spread and reassortment of the AIV.

To infer transmission routes and reassortment, I used a representative sample of 282 polymerase basic 2 (PB2) sequences from avian influenza viruses of any subtype sampled within Asia and Europe over the period 2001-2017. Considering the different sampled locations, host-type, HA-subtype and NA-subtype as discrete traits or populations, I inferred Bayesian time-resolved phylogeographic trees using the "mugration" model. By analysing the changes in subtype over transmission routes as defined by the phylogeographic trees, I calculated measures of reassortment rates within each host type and sampled location. I also reconstructed a general transmission network of avian influenza between locations and host-types using an

extension of the structural coalescent model approximation. This analysis indicated an intensive continuous circulation within Europe and Eastern Asia separately, but also sporadic connections through different wild and domestic bird transmission routes. High reassortment rates were observed in European wild Anseriformes and Asian domestic Anseriformes, indicating the importance of Anseriformes as generators of new influenza variants, but highlighting the different epidemiologies in the two regions.

2.2 INTRODUCTION

The emergence of zoonotic pathogens is expected to increase in the upcoming decades[158,164], and because such diseases represent a threat for human and wildlife populations, it is of primary importance for the scientific community to study the process leading to their appearance[166].

Over the last century, four different human influenza pandemics have occurred, causing millions of deaths[158], with all of these having genomic elements of avian origin that were directly or indirectly transmitted to humans through intermediate swine hosts [44,159,160]. It is now well known that direct transmission of influenza viruses from birds to humans is a recurrent event, with birds representing a likely viral source for humans[161,162]. Although it has not happened yet, there are ongoing concerns about the possible emergence of a new high mortality rate influenza strain with an avian origin and the ability to be easily transmitted amongst humans[163].

The influenza virus is a member of the family Orthomyxoviridae. Because the influenza virus type A has the largest pool of hosts (which include humans, pigs, horses, sea mammals, wild and domestic birds), amongst the three influenza types, I focused my study on this type[167,168]. The genome of the type influenza virus type A comprises eight negative-sense, single-stranded viral RNA (vRNA) segments[169] that encode at least 10 different proteins. The two surface antigen proteins (hemagglutinin, HA, and neuraminidase, A) are encoded by one segment each[170]. The remaining six segments encode the “internal proteins”: three polymerases (PA, PB1, PB2), a nucleoprotein (NP), two matrix proteins (M1 and M2), two non-structural proteins (NS1 and NS2/NEP). In some strains, a few additional smaller proteins are expressed from PB1 and PA [283,284].

2.2. Introduction

Influenza A virions present one type of HA and NA glycoproteins on their surface. There are 18 subtypes of HA (H1-H17) and 11 NA subtypes (N1-N9) which differ by at least 30% in their amino acid sequence[170]. Different combinations of those surface glycoproteins are possible[172] and define the influenza A subtypes. The segmented nature of the virus genome can lead to reassortment - the exchange of genome segments amongst strains co-infecting the same cells. This phenomenon can change the nature of the proteins found at the surface of the recombined virus (antigenic shift) that can therefore efficiently evade the immune system of previously infected cells and allow an effective transmission[169].

Although influenza A virus has a wide range of potential hosts[170], by showing signs of a potential co-evolution with the avian influenza A, water birds, and especially Anseriformes (ducks, geese, and swans) and Charadriiforms (gulls, terns, sandpipers), are thought to be the natural reservoir of influenza A virus[44,170]. Moreover, by being asymptomatic carriers, both Anseriformes and Charadriiforms birds could play a role in the overwinter maintenance and long-range propagation of LPAI and HPAI strains through their seasonal migrations[44,175,184,187,285]. Multiple subtypes of influenza A viruses co-circulate in avian populations, providing ample opportunity for reassortment.

There is evidence that the primary infections in domestic bird populations usually come from the introduction of a new influenza strain by wild birds[200]. Because they share the same habitat, water and food as wild waterfowl[193], domestic ducks are a key intermediate host between wild birds and poultry, and play an important role in the HPAI strains persistence and emergence[189,191–193]. However, the source of secondary infections in domestic birds populations mostly comes from human actions through movements of infected farm material and animals[172].

Low pathogenic influenza (LPAI) strains and highly pathogenic influenza (HPAI) strains can be distinguished based on their virulence and mortality in chicken[170]. To date, HPAI virus strains have only been found amongst the H5 and H7 subtypes [40]. New HPAI strains are considered to emerge from LPAI strains following their transmission from wild birds to domestic birds where they progressively gain in pathogenicity through successive infection cycles[179]. This transformation might take from a few days[180] to months, depending on the outbreak[181]. This gain in pathogenicity comes from the insertion of multiple basic amino-acids in the cleavage site of the precursor of the mature hemagglutinin[170]. This difference in HA cleavability affects in which tissue the proteins can be activated and is considered as the main determinant of the pathogenicity of a particular strain[178].

Even though the isolation of HPAI strains within living wild birds is uncommon, HPAI outbreaks are shown to follow wild bird migrations[205]. However, it has been observed that the disease circulation is partially maintained through trade of infected Galliformes (chickens)[210,211]. Therefore, the worldwide spread of HPAI probably comes from the synergy between trade of infected domestic birds and wild bird movements[202].

Using Bayesian phylogenetic methods on viral sequence data, it is possible to infer ecological processes and the evolutionary events that shaped the evolution of influenza strains[5]. Furthermore, by considering discrete locations (e.g. countries) as a discrete trait, and modelling the transitions from one location to another with a continuous-time Markov chain (CTMC) upon phylogenetic trees, phylogenetic methods can be used to perform spatiotemporal analysis over the evolutionary history of a strain [17,91,286]. These models of diffusion are not limited to spatiotemporal analysis and have already been applied to understand viral host switching [90,103] and genotype

evolution of the avian influenza virus[206]. Additionally, with the recent advances in structural coalescent model approximation and by combining epidemiological and genetic data, we are now better able to reconstruct transmission networks, putting us closer to determining “who infected whom” over the entire course of an outbreak[22,25].

Many previous works studying the circulation and transmission of infection between multiple hosts or strains have used HA or NA avian influenza , and are often focused on H5, H7 and H9 subtypes [103,191,206]. However, internal protein coding segments have their own reassortment history and have the possibility to group with numerous external segment partners[287,288], while LPAI strains have the potential to cause zoonotic outbreaks by contributing to the genesis and spread of new reassortant viruses[114,289]. Consequently, analysing the reassortment pattern of low and high pathogenicity subtypes using the internal segments could lead to new insights for avian influenza epidemiology.

The aim of this study was to understand the role that Asian and European wild and domestic bird populations play in the spatial diffusion of avian influenza and how their movements affects the genetic exchange among viral subtypes in circulation at the time. I also wanted to estimate the differences and similarities in influenza circulation patterns and timing within and between Asia and Europe. Therefore, using a Bayesian phylogeographic framework, I performed a set of phylogenetic analysis using an internal protein coding segment (PB2) of avian influenza A strains circulating in bird populations in Asia and Europe for the period 2001 to 2017 in LPAI and HPAI together. By using an internal avian influenza sequence such as PB2[290], with a low natural genetic variation and little evolutionary pressure, instead of external protein such as HA and NA, subject to balancing selection [287,291], I aim to better estimate the ecological and evolutionary dynamic events that shaped the evolution of the virus

compared to previous phylogenetic studies that used external proteins. Furthermore, to gain a more detailed view of the circulation between the sampled locations and bird populations, I also applied recent Bayesian structural coalescent approximation reconstruction in order to reconstruct transmission networks.

2.3 MATERIAL AND METHODS

To assemble comprehensive genetic data sets representative of the avian influenza A circulation, I retrieved all available full length European and Asian sequences of Avian influenza PB2 (13596 sequences) sampled over the period 2001 and 2017 available in December 2017 on the GISAID platform[292]. I selected all the sequences with known strain, sampling location, sampling year and host. The sequences were aligned with the Multiple Alignment using Fast Fourier Transform (MAFFT)[293] program version 6.864b, and a manual edition of the alignment was carried out to remove low-quality sequences. I grouped the sequences according to their region of sampling, neuraminidase and hemagglutinin type and type of hosts (Wild or domestic Anseriformes, domestic galliform, wild charadriiform and other wild birds). Prior to the sequence selection and to obtain a good representation of the virus strains circulating between the two continents, I constructed a neighbour joining (NJ) tree using the MEGA software[294]. Because most of the clades composing the phylogenetic tree were made of sequences coming from a single population (domestic or wild) or from a single outbreak or area, I selected a clade with a high diversity in the type of strain and location present in the NJ tree.

In total, the selected clade was composed of 825 sequences. To reduce the effect of potential sampling bias, I performed a stratified subsampling procedure to allow a maximum of three sequences per country of origin, per subtype and per year. Following the subsampling procedure, the sequences were then grouped according to their geographical origin to reduce the number of discrete states to analyse.

The final PB2 avian influenza dataset comprised 282 sequences coming from nine different geographical regions: four European regions (Western, Eastern, Southern and Northern Europe), two regions located in Central Asia (Central Asia and North

Central Asia) and three regions situated in Asia (Southern and Eastern Asia, and Eastern China) (see Table 2-1). All the sequences came from five distinct bird populations: wild-Anseriformes, Domestic-Anseriformes, Domestic-Galliformes, Wild-Charadriiforms and Wild-others (see Table 2-3 to observe the number of sequences per host type and the Supplementary table 8-2 for a number of sequences per location and host type).

Table 2-1: Number of sequences per discrete locations for the 282 avian influenza PB2 sequences dataset.

Location	Number of sequences	Date range
Central Asia	20	2007.6-2017.3
Eastern Asia	32	2004-2017.1
North-Central Asia	20	2001.6-2016.4
South Asia	26	2008.97-2016.9
East China	48	2002-2016
Eastern Europe	36	2005.7-2017.3
Northern Europe	18	2003.8-2016.95
Southern Europe	29	2001-2017.6
Western Europe	53	2002-2017

Table 2-2: Most precise sampling location composing each one of the 9 different location traits used in the analysis.

Location	Sampling area
Central Asia	Iran, Altai, Republic of Georgia, Georgia, Chany, Siberia, Tatarstan, Astrakhan
Eastern Asia	Hokkaido, Korea, Heilongjiang, Hebei, Xianghai, Shimane, Kagoshima, Taiwan
North-Central Asia	Primorie, Mongolia, Yunnan, Guizhou, Sichuan, Tyva, Uvs-Nuur Lake, Qinghai
South Asia	Bangladesh, Thailand, Vietnam, Quang Ninh, India
East China	Hunan, Eastern China, Shantou, SanJiang, Jiangxi, Jiang Xi, Henan, Guangxi, Hong Kong, Hubei, Fujian, Shanghai, Jiangsu, Zhejiang, Dongting, Wuhan, Wuxi
Eastern Europe	Ukraine, Hungary, Poland, Czech Republic, Moscow, Kalmykia, Rostov, Sergiyev Posad
Northern Europe	Denmark, Sweden, England, Iceland
Southern Europe	Italy, Spain, Portugal, Slovenia, Croatia
Western Europe	Netherlands, France, Belgium, Germany, Bavaria, Germany-NI, NL-Zeewolde, NL-Marker Wadden, Germany-MV, NL-Durgerdam, NL-Abbega, Germany-SH, NL-Stolwijk

Table 2-3: Number of sequences per host for the 282 avian influenza PB2 sequences dataset.

Host	Number of sequences	Date range
Domestic Anseriformes	101	2001.64-2017.29
Domestic Galliform	49	2002-2017.6
Wild Anseriformes	106	2001-2017.64
Wild Charadriiform	15	2003.29-2016.9
Wild other	11	2002-2017.2

Subsequently, I performed a joint analysis of the host type and location using a discrete trait representing both the host type (wild or domestic) and the location (Europe, Central Asia, Southern Asia and China-Eastern Asia) of the samples (see Table 2-4). In total, I performed my analysis on 7 states for this joint trait analysis (Table 2-5).

Table 2-4: Most precise sampling location composing each one of the 4 different location trait used in the analysis with the host and location trait combined.

Location	Sampling area
Central Asia	Iran, Altai, Republic of Georgia, Georgia, Chany, Siberia, Tatarstan, Astrakhan, Primorie, Mongolia, Yunnan, Guizhou, Sichuan, Tyva, Uvs-Nuur Lake, Qinghai
Eastern Asia	Hokkaido, Korea, Heilongjiang, Hebei, Xianghai, Shimane, Kagoshima, Taiwan, Hunan, Eastern China, Shantou, SanJiang, Jiangxi, Jiang Xi, Henan, Guangxi, Hong Kong, Hubei, Fujian, Shanghai, Jiangsu, Zhejiang, Dongting, Wuhan, Wuxi
South Asia	Bangladesh, Thailand, Vietnam, Quang Ninh, India
Eastern Europe	Ukraine, Hungary, Poland, Czech Republic, Moscow, Kalmykia, Rostov, Sergiyev Posad, Denmark, Sweden, England, Iceland, Italy, Spain, Portugal, Slovenia, Croatia, Netherlands, France, belgium, Germany, Bavaria, Germany-NI, NL-Zeewolde, NL-Marker Wadden, Germany-MV, NL-Durgerdam, NL-Abbega, Germany-SH, NL-Stolwijk

Table 2-5: Number of sequences per join trait for the 282 avian influenza PB2 sequences dataset.

Host	Number of sequences
Domestic South Asia	26
Domestic Central Asia	13
Wild Central Asia	27
Domestic East China	58
Wild East China	22
Domestic Europe	54
Wild Europe	82

Using a Bayesian discrete phylogeographic approach model, I estimated the virus evolution and circulation among the different regions, hosts and subtypes involved[17]. Each Markov chain Monte Carlo (MCMC) analysis was performed in BEAST 1.8 using the BEAGLE library[295,296]. Different substitution, clock and effective population size (tree priors) evolution models were evaluated by estimating their marginal likelihoods using the Akaike's Information Criterion for MCMC samples (AICM) in Tracer[297]. A Bayesian skygrid population model with a relaxed uncorrelated exponential molecular clock model and a general-time-reversible (GTR) model with site to site rate variation across two nucleotide partitions (codon positions 1 and 2; and codon position 3) were selected as the most appropriate model[61,79].

The transition history among the different regions, hosts and subtypes was inferred using a discrete traits model. Even though the Bayesian framework I chose to work with could cope with the joint inference of sequences evolution and the virus circulation between different populations, it would have remained a computationally challenging task. Therefore, by keeping the circulation and the evolution processes independently modelled, I was able to split my analysis in two independent steps. In the first step, I modelled the sequence evolution to generate an empirical tree distribution and then in the second step, I fitted my different discrete traits analysis over the set of posterior trees. Therefore, a subset of 1000 trees from the reconstructed posterior tree distribution was selected and was used as an empirical distribution in the further discrete state analysis.

For all the discrete states analysis, I used an asymmetric CTMC model and incorporated Bayesian Stochastic Search Variable Selection (BSSVS) analysis to identify the smallest sets of transitions rates that could summarise the epidemiological connectivity between the locations, hosts or strain[17]. From those BSSVS analysis, Bayes factor (BF) values were calculated to determine the significant non-zero

transition rates (BF = 10) between the discrete locations, host or strains. Additionally, for each of the discrete analysis I incorporated a posterior inference on all the possible state transitions (Markov jump) through the whole phylogeny[78,110] to count the number of state transitions that occurred. To identify the existence of any potential subpopulations or clusters where the virus tend to circulate more, I performed a walktrap community detection analysis [298] on the BSSVS and MJ results. Community detection methods aim to identify closely connected subnetworks of nodes (in this case, between the discrete states of location or host populations), and random walk dynamics, such as the walktrap method, are a commonly used method for community detection. A random walk is defined as a route where a node reached at step i is a random neighbor of the node reached at step i-1, and once within a community a random walker would spend a long time within it due to the high density of within-community edges and only a few edges leaving this group of nodes.

Also using the mappings of the discrete traits on the posterior trees, in order to determine in which host species or regions the reassortment events occurred, I calculated a reassortment measure using the number of branches in which the subtype changed but the host species or region traits remained identical (normalised by the sum of branch lengths, see also [206]).

Finally, I used the package SCOTTI, an extension of the BEAST software, to identify the existence of possible unsampled hosts as well as to reconstruct the transmission tree and transmission network between all sampled locations and host types[22]. This approach uses the recent advances in structured coalescent model approximation[22] to model each host as a separate pathogen population. By using available epidemiological information, this model simplifies the number of parameters calculated by rejecting all impossible direct transmissions events between populations due to non-overlapping exposure times.

Therefore, two transmission network analyses between Asia and Europe were performed, one using a discrete state encompassing the information about the region and time of sampling, and a second analysis using a discrete state encompassing the information about bird type (wild or domestic), region (Europe, Central Asia, China-East Asia and South Asia) and time of sampling (for the definition of the different regions used in this analysis, see table 2-5). At the end, I used 99 discrete states for the first analysis and 84 for the second analysis. Each one of those discrete states were then used as proxy to represent the bird population existing in each one of those location during that specific year. For this part of my work, because the evolutionary process and the circulation between the discrete states are not modelled independently by SCOTTI, the sequences evolution and the discrete traits were analysed together.

Table 2-6: Different locations composing the different regions used in the network reconstitution analysis between the regions/hosts/year using SCOTTI. The locations had to be aggregated in accordance with their geographical extent in order to reduce the number of discrete states to analyse. This was a trade-off between keeping as much geographical definition as possible while reducing the complexity of the analysis.

Region	**Location**
Central Asia	Central Asia, North-Central Asia
East China	East China, Eastern Asia
Europe	Eastern Europe, Southern Europe, Western Europe, Northern Europe
South Asia	South Asia

I used TreeAnnotator to summarise the maximum clade credibility (MCC) trees and FigTree version 1.4.1 to visualise the annotated trees[299,300]. The software SPREAD3 and cytoscape were used to identify and visualise the well supported rates of transmission[301,302]. Custom R scripts using the R package ape 5.0[303] were used to perform the stratified sampling and the Markov jump visualisation.

2.4 RESULTS

For this section, I first present the results for the geographical dispersal and the host contribution and then present the estimations for the virus reassortment patterns and population changes. I finish by presenting the results for the temporal transmission network analysis.

2.4.1 "Mugration" model results for the distinct analysis of host and location trait

2.4.1.1 Geographical dispersal and host contribution

Using a relaxed exponential clock and a skygrid coalescent population model, I determined multiple evolutionary parameters and reconstructed the evolutionary tree between the different sequences analysed. I observed a mutation rate of $3.9e^{-3} \pm 3e^{-4}$ [95% HPD:$3.35e^{-3}$-$4.53e^{-3}$] mutations per site per year with a TMRCA of 22.86 [95% HPD:19.69-26.63] year (1995.28).

To capture the circulation of Avian influenza within Eurasia, I performed a discrete trait analysis between the 9 discrete locations (Northern Europe, Western Europe, Eastern Europe, Southern Europe, Central Asia, North-Central Asia, Southern Asia, Eastern China and Eastern Asia). Overall and by looking at the reconstructed phylogenetic tree, I estimated that the virus mostly circulated within Southern Europe, Western Europe and East China with multiple introductions into the other sampled regions (see Figure 2-1a). An analysis of the root state probabilities pointed Southern Europe (with a 76% probability) above Western Europe (11% probability) as the location of origin for the selected clade, and a time-to-most-recent common ancestor of 22.8 years [HPD: 19.69-26.63]. However, the observation of Southern Europe as the origin might be caused by the existence of few early sequences originating from

this region (see Figure 2-1). By estimating the time spent in each of the locations using a Markov reward analysis, I observed that the virus spent significantly more time in East China than in Western Europe and Southern Europe. Interestingly, the virus did not seem to spend much time in Northern Europe, North-Central Asia and Central Asia (see Table 2-7).

2.4. Results

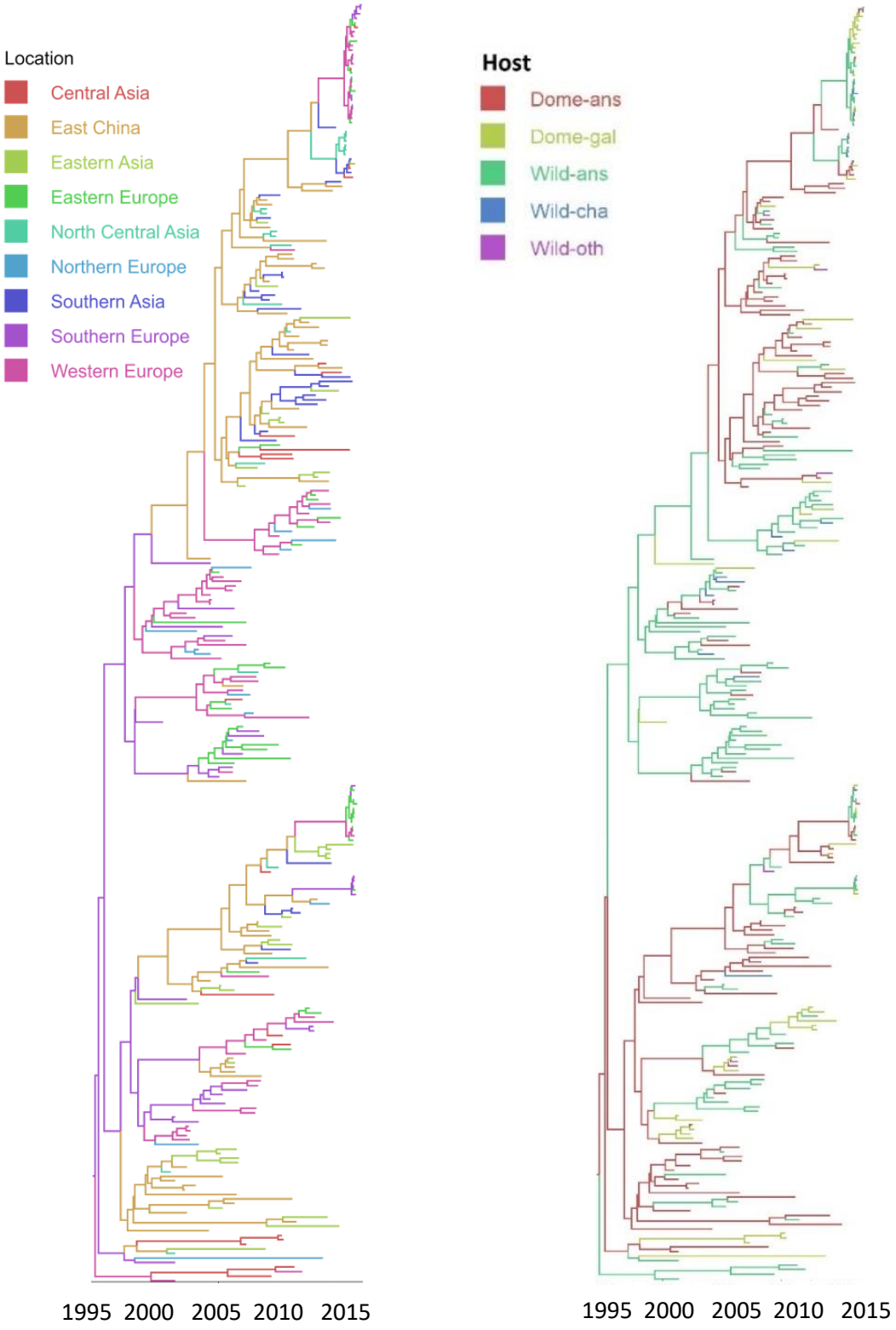


Figure 2-1: Bayesian MCC time scaled discrete phylogeographic tree of 282 PB2 avian influenza sequences. The phylogeny branches are coloured according: location a. To their descendent nodes location. b. To their descendent nodes host. The key for colours for each of the subgraph is shown on the left.

Table 2-7: Markov-reward analysis results for the discrete location analysis using 282 avian influenza PB2 sequences.

Location	Time spend in each location (years)	SD (years)	95% HPD interval (years)
East China	233	43.7	143.4-303.6
Western Europe	160	22.2	119.8-207
Southern Europe	103	28.2	42-149.2
South Asia	75	39.2	29.8-155.3
Eastern Asia	65	12.4	42.1-89.2
Eastern Europe	57	11.6	33.8-78.5
Central Asia	48	10.3	27.2-67.6
North-Central Asia	30	10	12.76-48.7
Northern Europe	29	7.4	15.9-43.7

Table 2-8: Markov-reward analysis results for the host analysis using 282 avian influenza PB2 sequences.

Host	Time spend in each host (years)	SD (years)	95% HPD interval (years)
Domestic Anseriformes	377.376	37.8	300.2-450.9
Wild Anseriformes	334.883	34.4	269.4-404.9
Domestic Galliformes	71.098	11	51.6-93.9
Wild Charadriiforms	11.724	3.6	5.5-19
Wild other	4.239	2.4	1-8.7

Using a Bayesian stochastic search variable selection (BSSVS) procedure, I identified a set of well supported rates of transition between the sampled countries that I characterised using a BF value. Most of the well supported rates of transition were observed within the European and Asian continents, with a connecting role of Central Asia between Southern Asia and Eastern Europe. I also observed many well supported rates starting in East China toward all the other Asian regions (see Figure 2-2a and Supplementary table 8-4).

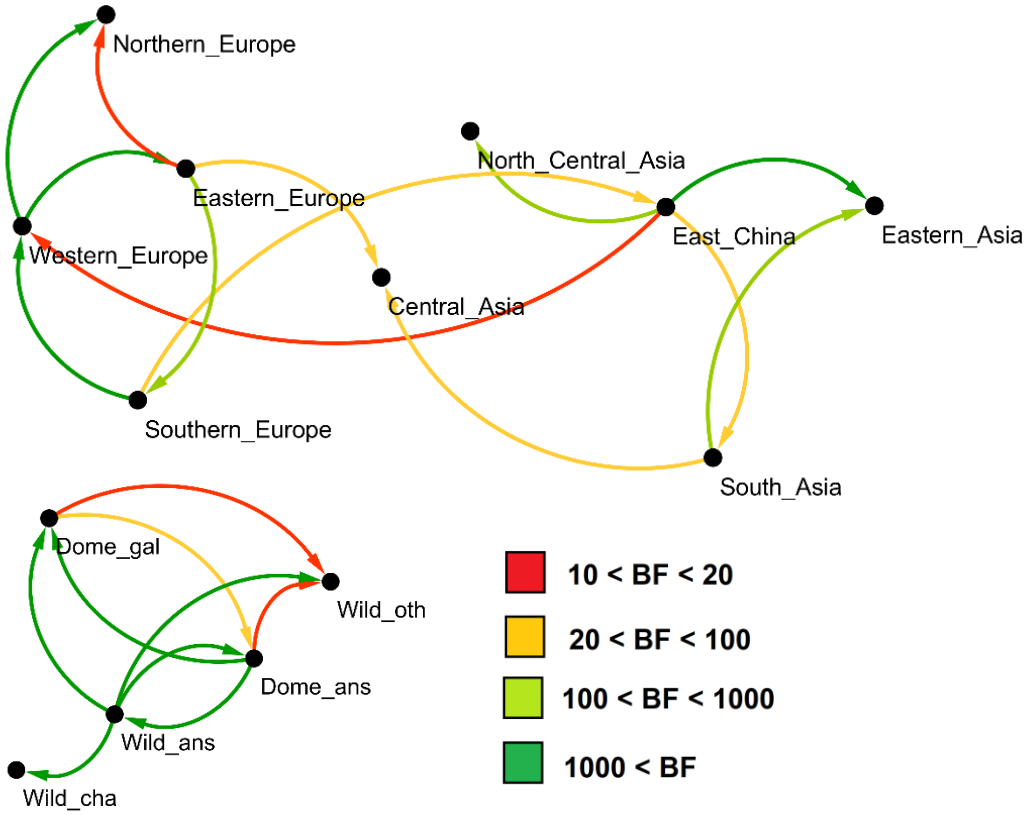


Figure 2-2: Output of the BSSVS analysis for avia influenza showing the best supported rates of transition between the sampled locations and hosts. The edge colours represent the relative strength by which the rates are supported BSSVS analysis result for a. The location analysis b. The host analysis.

I obtained a quantitative measure of gene flow between the sampled regions using a 'Markov jump' counts evaluation (see Figure 2-3 and Supplementary table 8-6). This value represents the number of expected state transitions between each location over the whole phylogeny. I observed a high number of transitions within Asia and Europe with a few transitions between the two. Overall East China and Southern Asia seemed to be the two main sources of viral material for the rest of the Asian continent. I observed a tendency for Eastern-China to infect more easily North-Central Asia than Central Asia, whilst Southern Asia had a higher number of transitions toward Central

Asia. I observed that North-Central Asia had a few transmission events toward Europe and Asia but slightly more toward Central Asia. Only a few transmission events from Central Asia to Western/Eastern Europe and Southern Asia were observed. I observed many jumps starting in Western Europe toward Northern and Eastern Europe while Southern Europe seemed to be an important viral source for Western Europe. I noticed many transitions starting in Eastern and Western Europe and ending in Central but only a few toward North-Central Asia. We can observe a transition from Southern Europe and Western Europe toward East-China. From those results, it seems that there are three main routes of transmission between Asia and Europe, even though such event seems rare. The first is through a direct transmission event from East China to Europe, a second indirect route following the sequence Southern Asia, Central Asia and Europe, and a third less common indirect route following the sequence East-China, North-Central Asia, Central Asia and Europe.

From the bird population analysis, I observed that the domestic and wild Anseriformes birds were composing most of the backbone on reconstructed phylogenetic tree (see Figure 2-1b). Using the Markov reward analysis, I observed that the virus tended to spend most of its time in those two populations and was barely present in the wild Charadriiform and other wild bird populations (see Table 2-8). Additionally, the BBSVS and MJ analysis showed multiple transmission links and events starting in Wild and Domestic Anseriformes bird toward the three other bird populations (see Figure 2-2b, Figure 2-4, Supplementary table 8-5 and Supplementary table 8-7). Interestingly, by looking at the BBSVS analysis, I observed that the domestic Galliformes population was the only population (apart the wild Anseriformes) able to infect the domestic Anseriformes birds (see Figure 2-2b).

2.4. Results

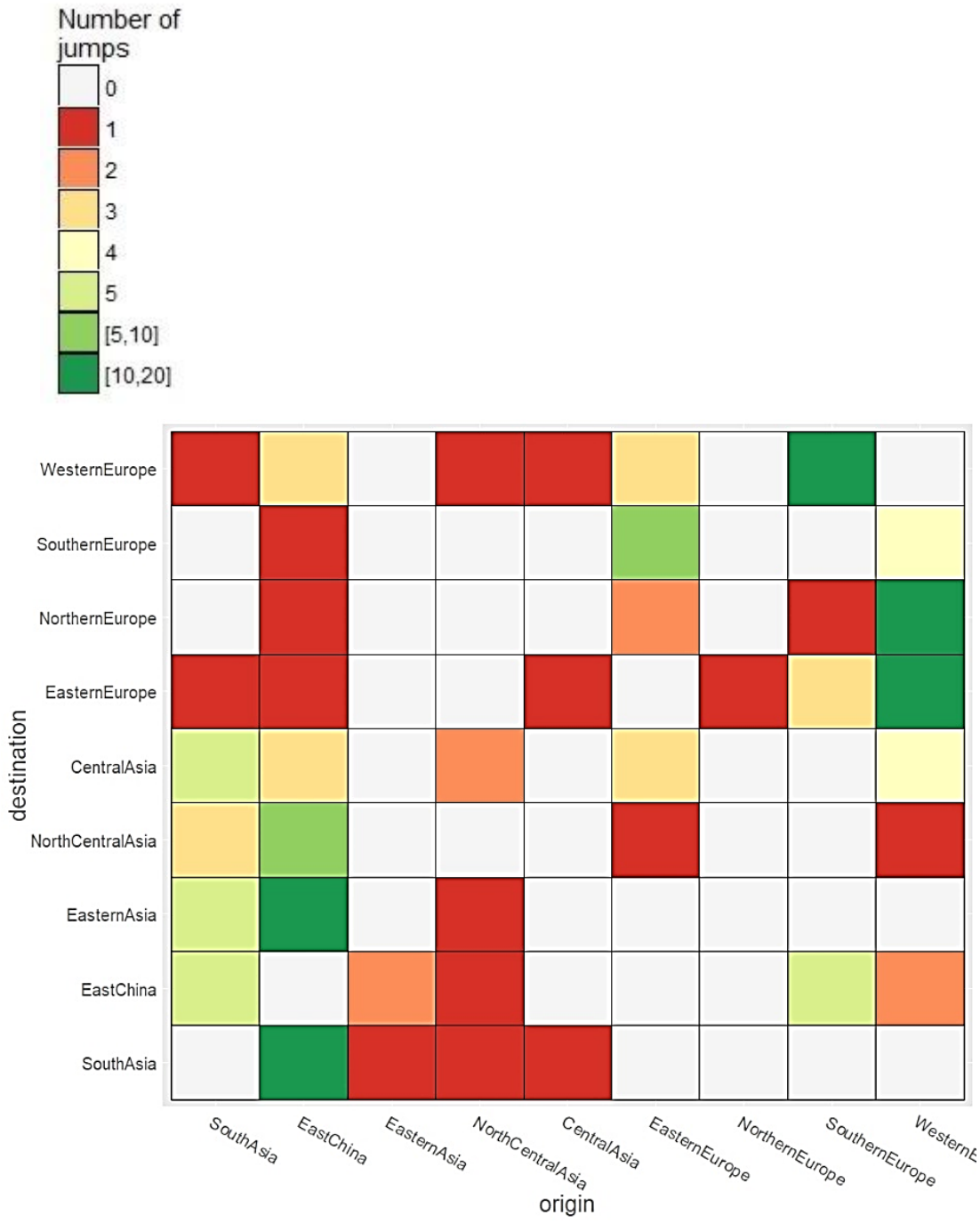


Figure 2-3: Heatmap showing the number of transitions between the sampled locations for avian influenza. The heatmap is coloured according to the number of observed transitions between locations.

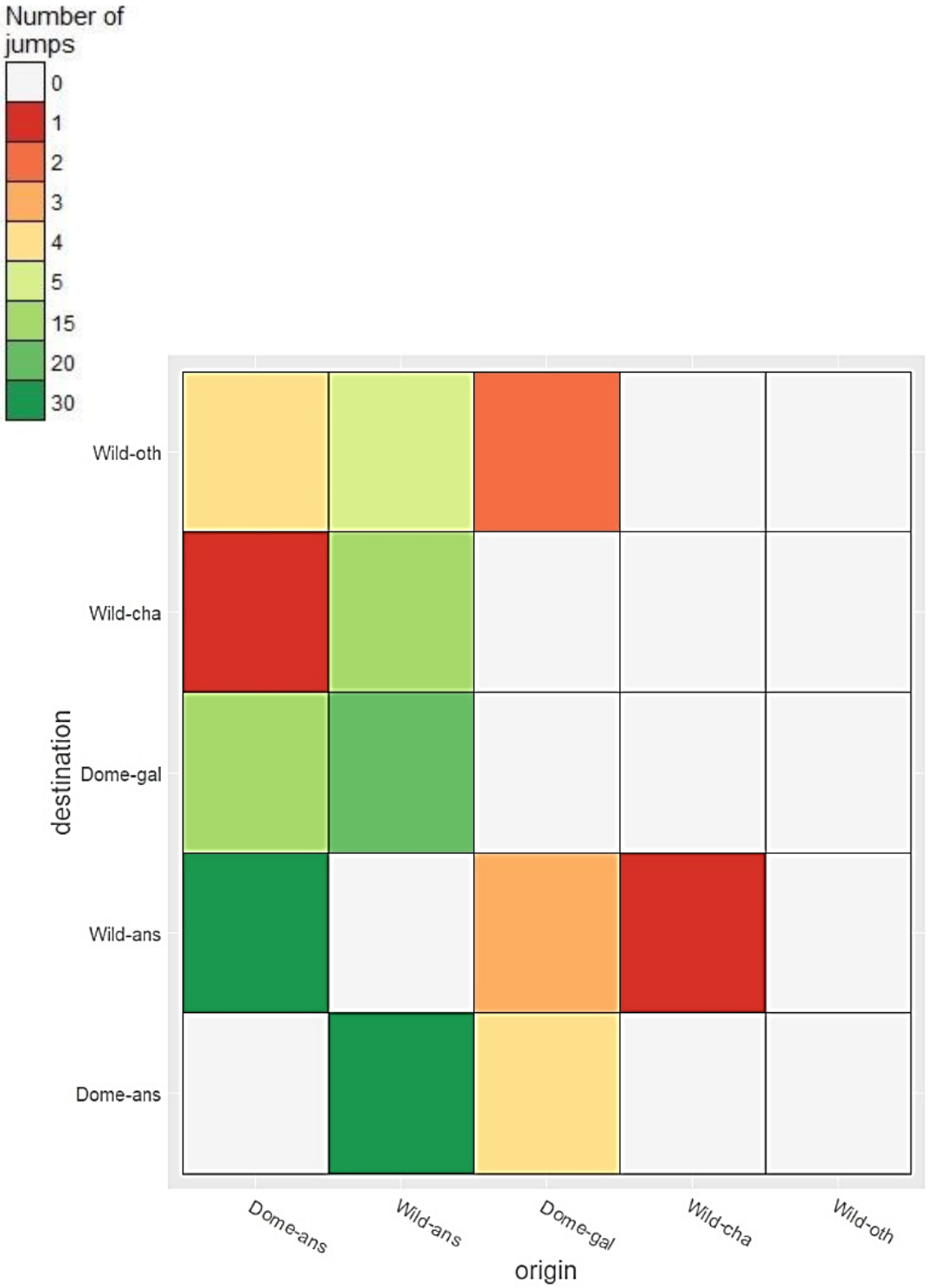


Figure 2-4: Heatmap showing the number of transitions between the sampled hosts for avian influenza. The heatmap is coloured according to the number of observed transitions between hosts.

To determine the existence of different communities within the sampled regions, I performed a walktrap community detection on the BSSVS and Markov jumps results. Using the BSSVS analysis results, I observed four communities (with a modularity of 0.37), one comprising all the European regions, a second one comprising the East-China and Eastern-Asia regions, a third one comprising the South Asia and Central Asia regions and the last one comprising the North-Central Asia region. Using the Markov-jump analysis results, I detected two clear communities (with a modularity of 0.33), one comprising the Asian and Central-Asian regions and the other one comprising the European regions.

2.4.1.2 Reassortment in different populations

In the previous section, the circulation of avian influenza between different but connected geographic regions and host-types was inferred using the PB2 gene. However, this dataset is comprised of viruses of many subtypes, i.e. the PB2 gene is paired with different HA and NA subtypes during circulation, due to the reassortment process. In this section, I now consider whether reassortment seems to occur in some regions or host types more than others.

To do this, a reassortment measure was calculated using the number of branches in the posterior set of phylogenetic trees for which the HA and NA subtype changed whilst the host or region traits remained the same (see Methods) (See Table 2-9 and Table 2-10).

Table 2-9: Reassortment measure estimation for HA changes per sampled regions. The measure is the proportion of all reassortment events taking place within a location.

Location	Mean (HA)	95% HPD interval (HA)
East China	0.1	0.08-0.12
Western Europe	0.05	0.06-0.04
Southern Europe	0.02	0.01-0.03
South Asia	0.02	0.0-0.04
Eastern Asia	0.02	0.01—0.03
Central Asia	0.01	0.0 – 0.2
Eastern Europe	0.01	0.0 – 0.2
Northern Europe	0	0
North-Central Asia	0	0

Table 2-10: Reassortment measure estimation for HA changes per sampled host populations. This measure is the proportion of reassortment changes taking place within a host type.

Host	Mean (HA)	95% HPD interval (HA)
Domestic Anseriformes	0.19	0.17-0.21
Wild Anseriformes	0.14	0.12-0.16
Domestic Galliform	0.01	0.0 – 0.02
Wild Charadriiform	0	0
Wild others	0	0

2.4.1.3 Viral host circulation

Using a similar approach, I looked at the circulation of the virus between the different bird populations and locations sampled. By looking at the overall population transfer, I observed that most of the population changes occurred in East China (see Table 2-11). However, by looking in more detail, I observed that most of the transfers from wild to domestic hosts occurred in East China, Western Europe and Southern Europe, while most of the transfer from domestic to wild birds occurred in Eastern China (see

Supplementary table 8-8 and Supplementary table 8-9). Additionally, I isolated the host responsible for the transfers between the diverse regions and noticed that most of the direct virus transfers from Asia (East China, Southern Asia and Eastern Asia) to Europe were done by domestic and wild Anseriformes birds, while the transfers from Europe to Asia were mostly done by domestic Anseriformes birds (see tables Table 2-12 and Table 2-13). Interestingly, by looking at the indirect transfers between Asia and Europe, through Central Asia and North-Central Asia, I observed that most of the circulation between Asia and Central Asia (composed of Central Asia and North-Central Asia) was done through the circulation of domestic Anseriformes, whilst most of the circulation between Europe and Central Asia was done through wild Anseriformes birds (see Supplementary table 8-10, Supplementary table 8-11, Supplementary table 8-12 and Supplementary table 8-13). Furthermore, it seems that wild Anseriformes birds were responsible for the circulation within Europe whilst domestic Anseriformes birds were responsible for most of the virus circulation within Asia (See Supplementary table 8-14 and Supplementary table 8-15).

Table 2-11: Population change measure estimation between the different sampled regions. The population change measure is the proportion of all bird population changes taking place within a location.

Location	Mean	95% HPD interval
East China	0.05	0.04-0.06
Western Europe	0.03	0.02-0.04
Southern Europe	0.02	0.01-0.03
Eastern Asia	0.01	0.0-0.02
Central Asia	0.01	0.0 – 0.02
Northern Europe	0	0
South Asia	0	0.0 – 0.01
North-Central Asia	0	0
Eastern Europe	0	0

Table 2-12: Main host responsible for the direct circulation between Asia and Europe. This measure is the proportion of edges going from Asia to Europe where the bird population did not change.

Bird population	Mean	Sd	Median
Domestic Anseriformes	0.01	0.01	0.01
Wild Anseriformes	0.01	0.00	0.01
Domestic Galliform	0.00	0.00	0.00
Wild charadriiform	0.00	0.00	0.00
Wild others	0.00	0.00	0.00

Table 2-13: Main host responsible for the direct circulation between Europe and Asia. This measure is the proportion of edges going from Europe to Asia where the bird population did not change.

Bird population	Mean	Sd	Median
Domestic Anseriformes	0.02	0.01	0.02
Wild Anseriformes	0.01	0.00	0.01
Domestic Galliform	0.00	0.00	0.00
Wild charadriiform	0.00	0.00	0.00
Wild others	0.00	0.00	0.00

2.4.2 "Mugration" model results for the combined analysis of host and location traits

To get a better understanding of the relation between the host status (wild or domestic) and the sampling location, I performed the same set analysis (BSSVS, MJ, community detection, reassortment and transmission network estimation) using a discrete trait representing both the host type (wild or domestic) and the location (Europe, Central Asia, Southern Asia and China-Eastern Asia) of the samples (see Table 2-5).

The phylogenetic tree obtained through this approach shows that avian influenza seems to have mostly circulated within domestic birds in East China and wild birds in Europe (see Supplementary figure 8-2).

2.4. Results

From this BSSVS analysis, I observed that the domestic bird population in China and Eastern Asia was the major source of the virus for all other bird populations present in Asia, and that European wild birds act as a source for domestic birds in China-Eastern Asia. I also observed few links between domestic birds in Europe and domestic and wild birds present in Central Asia (See Supplementary table 8-16). From the Markov jumps analysis, I observed multiple transitions amongst the Asian bird populations and European bird populations, but only observed few transitions between the European and Asian bird populations (See Supplementary table 8-17). Interestingly, only the Central-Asian and European wild birds were able to infect all other bird populations. At the opposite, wild European birds were only infected by Central Asian wild birds and European domestic birds. Additionally, I observed that the domestic birds in Central Asia were the only bird population that could be infected by all other bird populations. Using a reassortment measure, I observed that most of the reassortment occurred in wild birds in Europe and domestic birds in China-East Asia (See Supplementary table 8-18).

Using the join trait BSSVS analysis results, I detected three communities (with a modularity of 0.2). The first community was composed of domestic European birds, wild European birds and wild Central Asia birds. The second community was composed of domestic birds from China-East Asia, domestic birds from South Asia and domestic birds from Central Asia. The last community was composed of wild China-Eastern Asia birds. Using the Markov jump analysis, I detected two communities (with a modularity of 0.29), the first comprising all the European traits and the second comprising all the Asian regions' traits.

2.4.3 Transmission network reconstitution using SCOTTI

Using a constant clock and an HKY substitution model, I used the BEAST extension SCOTTI to identify the existence of possible unsampled hosts, as well as to reconstruct a probable transmission tree and transmission network between 99 different possible host states encompassing the information about region and time of sampling.

Using this approach, I estimated mutation rate of $2.19e^{-3}$ [95% HPD:$2.07e^{-3}$-$2.3e^{-3}$] mutations per site per year with a TMRCA of 25.5 (95% HPD: 24.18-27.18) years. Because most of the backbone of the tree was in an unsampled location state, most of the virus circulation probably occurred within a state (a particular combination of region and year) that was not present in the 99 analysed states.

To reconstruct the transmission network, I selected the transmission events with a posterior probability above 50%. In total, two distinct subgraphs were isolated, both composed of Asian and European hosts (see Figure 2-5 and Figure 2-6). The first subgraph is composed of an Asian and a European cluster linked together by Northern Europe. In this graph, I observed that the virus circulated for some time in East China (2012 to 2014) before being exported to Europe. I observed that East-China was subject to reinfection cycles, this area being its own viral source over multiple years. It also seemed that Eastern Europe and Western Europe were closely linked with the observation of multiple transmission routes over several years between the two areas (see Figure 2-5). In the second subgraph, I observed that the virus was introduced in Europe in 2017 from Southern Asia or Eastern China by passing through Central Asia (see Figure 2-6). From there, the newly introduced strain circulated intensively in Europe in 2017.

2.4. Results

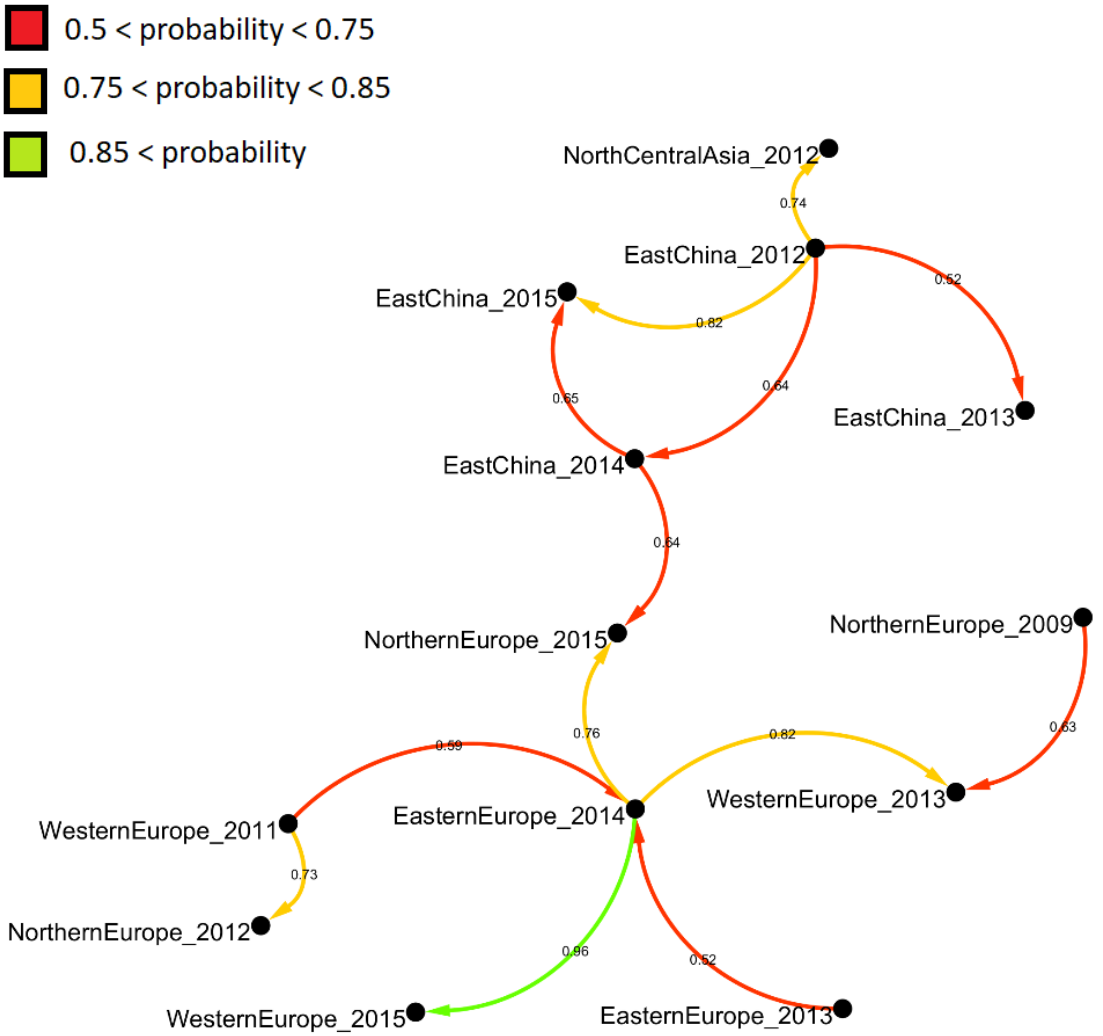


Figure 2-5: First subgraph estimated by SCOTTI using 282 PB2 avian influenza sequences and the information about the region and time of sampling. The edge colours represent the posterior probability of the edge. The key for colours can be found above the network.

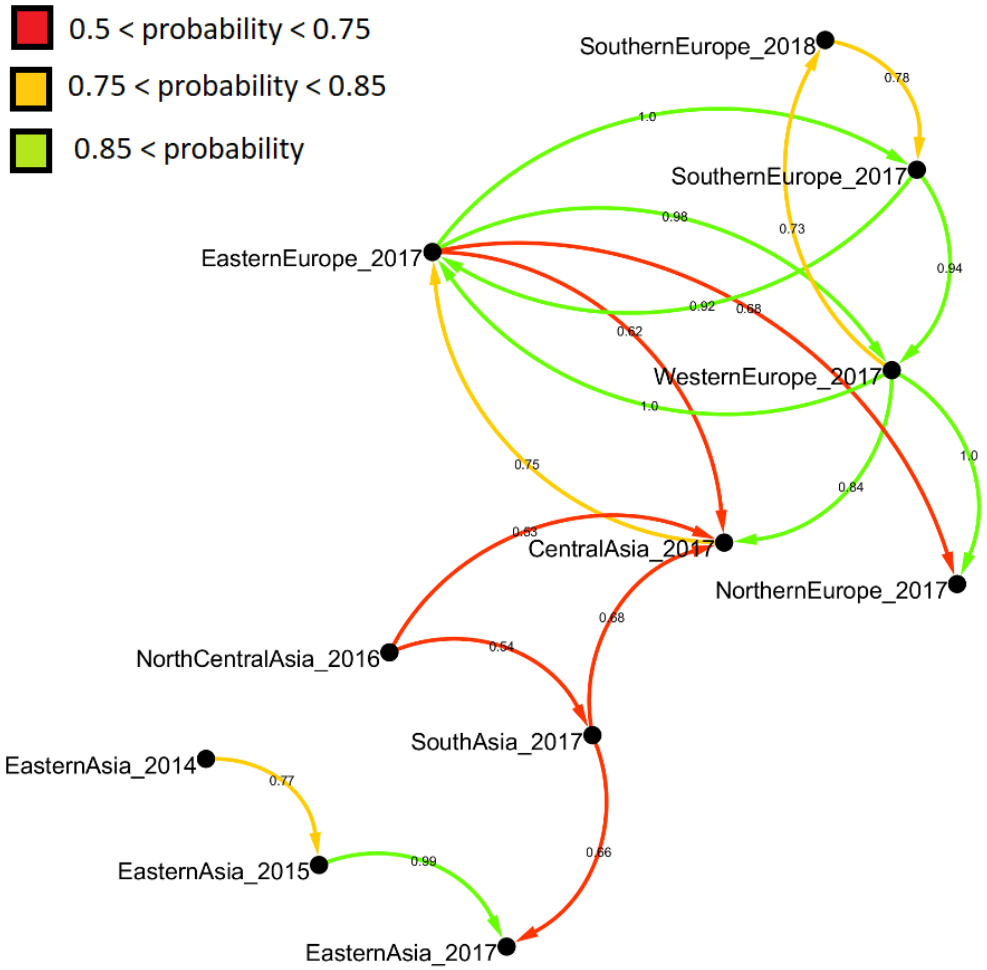


Figure 2-6: Second subgraph estimated by SCOTTI using 282 PB2 avian influenza sequences and information about the region and time of sampling. The edge colours represent the posterior probability of the edge. The key for colours can be found above the network.

Finally, using a constant clock and an HKY substitution model, I used the package SCOTTI to reconstruct a probable transmission tree and transmission network between 84 possible states encompassing information about the host type, location and time of sampling. With this approach, the estimated mutation rate was $2.12e^{-3}$ [95% HPD:$2.00e^{-3}$-$2.23e^{-3}$] mutations per site per year with a TMRCA of 26.1 [95% HPD: 24.63-27.76] years. The reconstructed phylogenetical showed signs that most

of the virus circulation was done within an unsampled host because most of the backbone of the tree was in an unsampled state.

By selecting the estimated transmission routes with a posterior probability above 50%, I reconstructed two transmission networks, both composed of 18 European or Asian hosts (see *Figure 2-7* and *Figure 2-8*). In each of those networks, I observed a single link between the Asian and European continents, one involving two domestic hosts (Domestic Europe 2017 to Domestic Central-Asia 2017) and the other one involving two wild hosts (Wild China-East 2014 to Wild Europe 2015). In the first graph, I observed a continuous circulation of the virus within wild European birds with few links toward European domestic birds. In the second graph, I observed a continuous circulation of the virus within Asian domestic birds with some links toward wild Asian birds. By looking at the two networks, we can observe that most of the links are done between hosts of the same type (wild-wild or domestic-domestic).

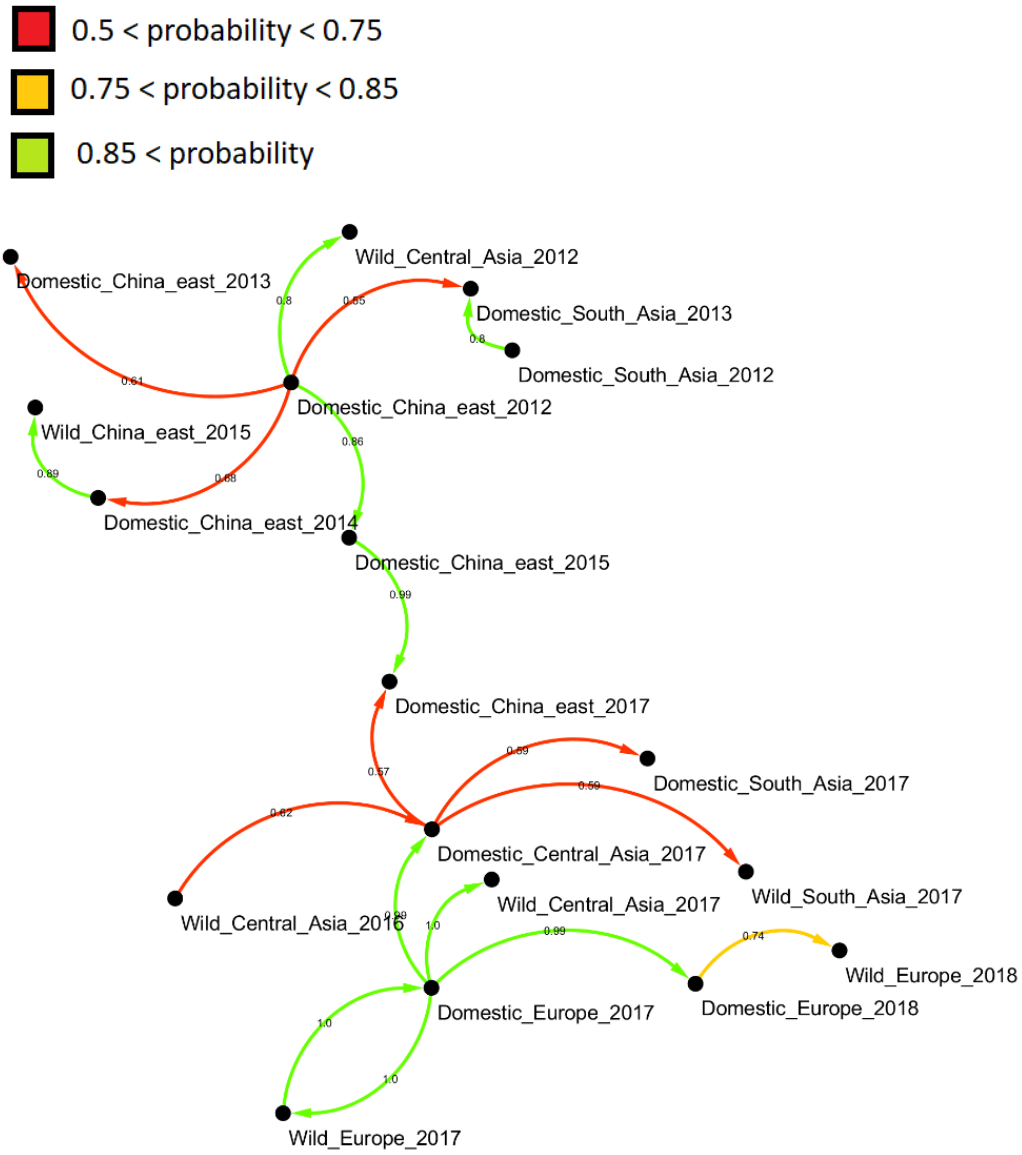


Figure 2-7: First subgraph estimated by SCOTTI using 282 PB2 avian influenza sequences and using the information about the region, host and time of sampling. The edge colours represent the posterior probability of the edge. The key for colours can be found above the network.

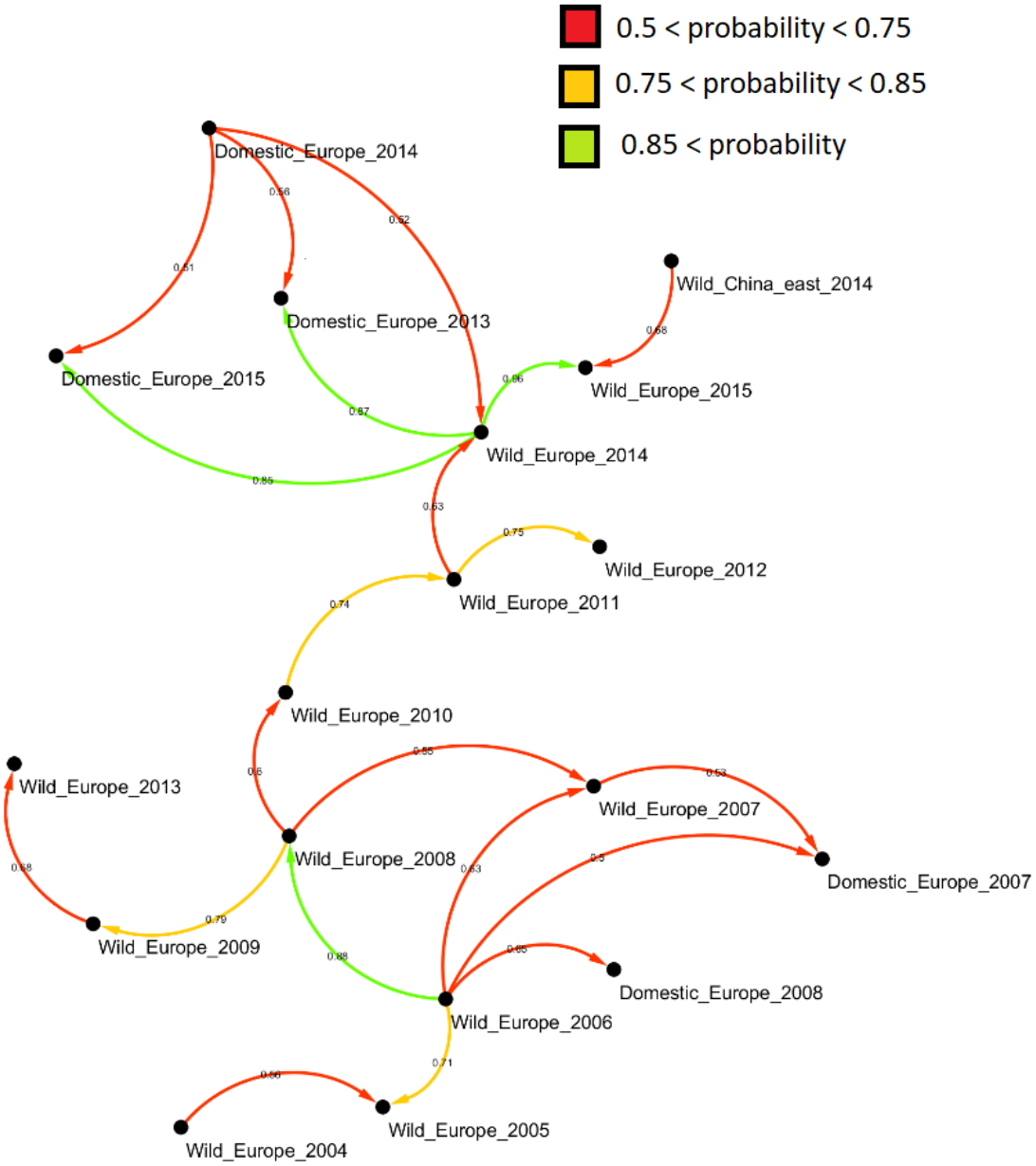


Figure 2-8: Second subgraph estimated by SCOTTI using 282 PB2 avian influenza sequences and using the information about the region, host and time of sampling. The edge colours represent the posterior probability of the edge. The key for colours can be found above the network.

2.5 DISCUSSION

In this paper, I have applied advanced phylogenetic methods on an Avian influenza PB2 sequences dataset representative of the avian influenza A circulation between Europe and Asia.

To generate an understanding of the circulation of the virus between Europe and Asia, I first used a phylogeographic discrete trait approach ("mugration")[17]. This approach, although useful and intuitive, is susceptible to bias in the transmission rate estimation from one state to another when applied to datasets with biased number of sequences per state. Therefore, to reduce this potential source of bias, I performed my analysis on a stratified subsample of all the available sequences and grouped them according to location and/or host type. This first approach enabled us to estimate the phylogenetic tree of the virus, the number of transition events between different sampled regions and hosts, as well as the potential existence of separated communities amongst the different regions and hosts sampled. This method also allowed us to estimate drivers measure of viral reassortment and estimates of the role played by the different potential hosts in the virus circulation.

Secondly, assuming each subpopulation has at least some representation, I estimated a potential transmission network by combining genetic and epidemiological information using a structured coalescent approach, a method less susceptible to sampling bias compared to the "mugration" approach. In order to perform this analysis, I used the R package SCOTTI[281] to evaluate the transmission network statistics amongst numerous discrete states (up to 100) representing the location of sampling, type of host, or a combination of both traits.

2.5. Discussion

With those two phylogenetic approaches, I was able to highlight three main key points in this analysis.

First, the results obtained through the “mugration” approach and supported by walktrap community detection results suggest that avian influenza virus tend to mostly circulate within Asia or Europe. When complemented with the transmission network reconstitution results, where only few links between the two continents can be observed, those observations suggest that the two continents constitute distinct demes that sporadically connected, where the virus tends to circulate intensively and where re-infection cycles are common. When looking at the circulation between Asia and Europe, the BSSVS and MJ analysis results suggest that two main transmission routes between Asia and Europe exist. The first is a direct transmission route where the virus moves from Eastern China to Europe without an intermediate location. The second is through an indirect transmission route that can take two forms, one that is passing by South-Asia/Central-Asia/Europe and another one, less common, passing through East-Asia or East-China/North-Central Asia/Central Asia/Europe. My findings are a good complement to multiple prior studies of H5 viruses[103,206,211,304] suggesting that in Eurasia, the continental-scale dynamics of avian influenza is structured between a few demes delimited by the presence of domestic birds and connected by relatively rare transmission events by migratory[103,206,304] and domestic bird movements[211].

Secondly, the population change measures obtained through the “mugration” approach highlighted the central role that Asian domestic Anseriformes birds and European wild Anseriformes birds have in the virus circulation and transmission in their respective continent. In Asia, the predominant role that domestic birds play in the circulation of avian influenza is probably coming from the existence of numerous

live bird markets acting as large transmission hub enhancing the virus dissemination[305]. Because of more strongly enforced biosecurity measures in Europe, there is a lower risk for domestic birds to circulate while being infected by the virus[306], making wild birds the main influenza vector in Europe. It was also determined that, when done directly, the transmission between Asia and Europe seems to mostly involve domestic animals and partially wild animals through the circulation of long-range migratory birds. Moreover, I observed a critical role that wild Anseriformes from Central Asia play in the transmission of avian influenza acting as a link between wild European birds and domestic Asian birds. However, due to the lack of sequences originating from those regions and the loss of definition due to their merge in most analyses performed, it remains difficult to estimate the different roles that North-Central Asia and Central Asia birds might have in the circulation of the virus. This predominant role that Central Asian wild birds have on virus circulation is probably caused by the intersection of multiple bird migratory flyways in this region (East-Asia/Australian flyway, Central-Asian flyway and Black Sea/Mediterranean flyway)[307], resulting in bird populations originating from different regions to mix and spread the virus circulation.

Finally, with the reconstitution of the transmission network, I observed that avian influenza tends to mostly circulate within domestic and wild birds whilst the circulation between the two populations seems to be an infrequent event. However, estimated population changes showed that bi-directional viral transfer between wild and domestic birds were present in East China while lower viral transfer rates from wild to domestic birds were present in Western Europe and Southern Europe. This observation is probably the result of the close proximity observed between wild and domestic birds in Asia, and especially between wild birds and domestic ducks[304,308], leading frequent transmission events between the two populations. At the opposite,

in Europe, my findings nicely complement previous genetic studies[309] by suggesting the existence of infrequent mono-directional transmission events from wild to domestic birds in Europe, which are highly dependent on host range restriction factors.

Thirdly, I studied the reassortment pattern of avian influenza and observed that most of the reassortment occurred in China-Eastern Asian domestic birds, but also in Western-European wild birds. Continuous circulation of multiple influenza strains in the same bird population can lead to reassortment of parts of viral genomes to produce various new influenza genotypes[179] which may lead to the emergence of new HPAI strains[200,310]. My results suggest that not only a well-known risk for the emergence of new HPAI in China[193] exists, but also high chances to observe the emergence of new HPAI strains in Europe due to the continuous reassortment on LPAI strains in wild European birds.

This work represents a quantitative analysis at the inter-continental scale studying the general viral dynamics and reassortment of low and high pathogenetic avian influenza using internal segment sequences. It is also the first time that a structural coalescent model approach is used to understand the circulation of Avian influenza over multiple locations and hosts. Overall, this analysis links multiple theories on the transmission of avian influenza between Europe and Asia that were suggested separately and it connects them in a global analysis using the most recent advances in phylogeographic techniques.

My work has some limitations, especially regarding the limited availability of sequences. Although I tried to limit its impact, the sampling is obviously unbalanced as it is based on submissions of individual countries or ad hoc research projects, and the effect it has on the results quality is uncertain. An example of possible bias caused

by a potential sampling issue is the fact that southern Europe was detected as being the source of the infection for avian influenza, which could be seen as uncommon. Multiple explanations could explain those results caused by the existence of early samples originated from southern Europe. Removing those samples would probably change the outcome of the analysis and can be discarded. First, we could estimate that those results are wrong and caused by the application of an inappropriate model applied on a biased dataset. Therefore, more advanced model such as the structural coalescent model, less prone to the effect of sampling bias, could have led to a different result. However, this observation could also be explained by the circular nature of the virus circulation (approximated by the reconstructed BSSVS network (see Figure *2-2*)) which is itself strongly influenced by the shape and seasonality of existing wild bird migratory routes (in our case more particularly the Black sea/Mediterranean flyway linking the southern part of Europe to central Asia)[307]. Consequently, the timing of sampling and location of the oldest sample would influence location of the root of the tree and could put it anywhere on this flyway, and in our case southern Europe. This same effect of the Black sea/Mediterranean flyway on the virus circulation between central Asia and southern Europe has been partially observed for subsequent years (2016/2017) in other studies where the same virus movements can be observed for reassortment sequences of HPAI and LPAI[311].

Ultimately, increasing the number of available sequences from diverse locations and hosts would help to develop models that better represent the diffusion of avian influenza and lead to better understanding of its epidemiology.

2.5. Discussion

3 PHYLOGEOGRAPHIC ANALYSIS AND IDENTIFICATION OF FACTORS IMPACTING THE DIFFUSION OF FOOT AND MOUTH DISEASE VIRUS IN AFRICA

3.1 ABSTRACT

Foot and mouth disease (FMD) is endemic in sub-Saharan Africa. Due to the complexity of the disease epidemiology and the lack of data available, there is a need to use modelling approaches to fill the gaps in our understanding of the virus circulation on this continent. Using a phylogeographic approach, I reconstructed the circulation of FMD virus serotypes A, O, SAT1 and SAT2 in Africa and evaluated the influence of potential environmental and anthropological predictors of virus diffusion. My results show that the serotypes O and A were introduced in Africa over the last century while the SAT1 and SAT2 serotypes have been circulating for at least 400 years in wildlife. My results also suggest that outside Southern-Africa, wildlife does not play a role in the maintenance and circulation of the disease within domestic animals. Further, the circulation of serotype O in Eastern Africa appears to be facilitated by both indirect transmission through persistence in the environment and anthropological activities such as cattle movements. Evidence for the different epidemiologies of serotypes has been lacking but is essential in developing a modern approach to control FMD viruses in Africa.

3.2 INTRODUCTION

Foot and mouth disease (FMD) affects more than 70 species of cloven-hoofed animals[216]. The disease is characterised by the development of vesicles in and around the mouth, on the feet and possibly in places on the skin[216]. The causal agent is a positive-sense, single-stranded RNA virus of the Picornaviridae family[312] called foot and mouth disease virus (FMDV). Its genome encodes the information for 4 structural proteins (VP1-4) and 8 non-structural proteins (7 proteases and one RNA polymerase). Based on the level of cross protection between strains, the virus can be divided into seven serotypes: O, A, C, Southern African Territories [SAT] 1, 2, 3 and Asia 1 [218,219], which are clinically indistinguishable from each other but which have different epidemiologies. The hosts that are considered to play an active role in these epidemiologies are cattle, buffaloes, pigs, sheep and goats[225].

FMD susceptibility varies according to the host and strain of FMDV involved. The severity of the infection depends on the amount of virus inoculated, the serotype, the host species and the individual immunity[226]. The commonest route of infection for a new host is by direct contact with an infected animal[216,235]. The infection may also occur indirectly by contact with contaminated surfaces or products, such as personnel, vehicle or fomites[225]. Movement of animals and animal products are considered to play an important role in the disease circulation in endemic areas and are considered the main factors for FMDV transboundary spread[220].

FMD has been eradicated in many high income countries but is still endemic in numerous low and middle income countries (LMICs)[313], particularly in Africa and South and East Asia. Although FMD has a low mortality rate in adult animals, it causes significant productivity losses that may lead to important and continuous economic losses for farmers, and impact countries' trading ability at a national level [220]. Although work has been done to understand the impact of FMDV in large scale dairy farms in

LMICs[233], there is still a lack of data to quantify its impact more broadly on the economy of endemically infected countries[234].

FMDV is endemic in most of sub-Saharan Africa with an epidemiology considered to be more complex than in other regions of the world due to multiple serotypes and wildlife reservoirs[267]. However, due to a general lack of surveillance and animal traceability, very few statistics on disease incidence and circulation exist for Africa. There are already a few studies on animal trade and seasonal migration of nomadic and pastoralist herds in sub-Saharan Africa[220,221,267,314]. However, to develop more modern control approaches, there is a need for analytical methods that use existing data to improve our understanding of the FMDV circulation and epidemiologies in this part of the world. Even the standard on the endemicity of FMD in Africa, based on clinical observation, ignore the possibility that the disease might arise from epidemic waves of different serotypes, a theory first proposed in 2006 by Bronsvoort et al.[315] and more recently again by Casey-Bryars et al.[270].

Many wildlife species can be infected by FMDVs in Africa[226] but amongst all these potential hosts, only the Cape buffalo (*Syncerus caffer*) and impala (*Aepyceros melampus*) have been implicated in the transmission of FMDV to domestic cattle[274]. Even though the Cape buffalo is suspected to be the primary reservoir and the main source of SAT serotypes in Southern Africa[278], its role as a viral source for livestock epidemics for the FMDV O, A and C serotypes elsewhere in Africa is still unclear[221,224,234,265], and might be unimportant[270].

It has already been observed that the spatiotemporal occurrence and circulation of FMDV in Africa is mainly affected by human activities through domestic animal movements[279,280,316,317]. However, several environmental characteristics and attributes such as the landscape, vegetation, and natural barriers to animal

movements such as roads, rivers or mountains have the potential to influence the dynamics and circulation of FMD [33,318,319].

Since FMDV are single stranded RNA viruses and lack a proof-reading mechanism for their genome replication, and replicate to high rates within a host, they display a high realised substitution rate at the between-host level[218]. The history of these mutations can provide information on the ecological processes and population events that shaped the virus evolution, even if not directly observed. These processes, along with other evolutionary parameters, can be modelled while reconstructing the phylogenetic trees[5,13]. Furthermore, by combining genetic data and spatial information, phylogeographic tree reconstruction can be used to estimate the unobserved geographical circulation of a pathogen[4]. For example, Cottam et al.[320] exposed that it was possible to determine transmission routes using FMD sequences for the UK 2001 outbreak. Virus movements can be modelled as discrete transmission events between the sampled locations[17], or as a continuous process using different random walk diffusion models[98]. Recently, both discrete and continuous approaches have been extended to test and quantify the contribution of potential environmental and anthropological parameters (predictors of viral diffusion) that might influence the spread and circulation of the studied pathogen[110,321].

The aim of this chapter is to gain a better understanding of the circulation of FMDV in Africa, comparing discrete and continuous approaches[110,115]. A detailed discrete phylogeographical analysis of serotypes A, O, SAT1 and SAT2 sequences was performed, and the influence of 13 potential environmental and anthropological predictors of virus diffusion were quantified and tested using both discrete and continuous approaches.

3.3 MATERIALS AND METHODS

3.3.1 Data Collection

To obtain a comprehensive genetic dataset, I first retrieved all available African FMDV A, FMDV O, FMDV SAT1 and FMDV SAT2 genetic sequences in Genbank (accessed on the 15/12/2016 for the serotypes A, O and SAT2 and on the 11/09/2018 for SAT1). From these datasets, I selected all VP1 sequences with at least information on the country of sampling and the year of sampling. In total, I gathered 191 FMDV A, 351 FMDV O, 214 FMDV SAT1 and 477 FMDV SAT2 sequences. The sequences were aligned using Multiple Alignment Fast Fourier transformation (MAFFT)[293]. Potential recombinant sequences were detected with RDP4 software and any such sequences were removed[322].

To reduce the effect of potential sampling, I ran a stratified subsampling procedure to thin the number of sequences to a maximum of three sequences per country of origin per month. For countries with less than three sequences available in total (regardless of temporal span), I grouped them with neighbouring countries if possible, or else removed them. The final FMDV A dataset was composed of 107 sequences from eight countries, dates ranging from 1966 to 2016 (see Table 3-1 for more details). The final FMDV O dataset was composed of 192 sequences from 12 countries, dates ranging from 1964 to 2016 (see Table 3-2 for more details). The final FMDV SAT1 dataset was comprised of 117 sequences from 10 countries in total, but grouped into 5 regions for further analysis (because there were only 1 or 2 sequences from some countries), dates ranging from 1961 to 2015 (see Table 3-3 for more details).The final FMDV SAT2 dataset was composed of 135 sequences from 15 countries, dates ranging from 1970 to 2015 (see Table 3-4 for more details). See Supplementary table

8-19, Supplementary table 8-20, Supplementary table 8-21 and Supplementary table 8-22 for more details on the sequences.

Table 3-1: Number of FMD A sequences per country/region sampled

FMDA		
Country	**Number of sequences**	**Year range**
Kenya	34	1966-2013
Cameroon	16	1975-2001
Nigeria	13	2009-2013
Ethiopia	12	1974-2008
Egypt	11	1972-2016
Tanzania	10	1967-2009
Sudan	7	1981-2006
Eritrea	4	1997-2006

Table 3-2: Number of FMD O sequences per country/region sampled

FMDO		
Country	**Number of sequences**	**Year range**
Kenya	60	1964-2011
Ethiopia	36	1973-2007
Sudan	23	1974-2011
Uganda	23	1974-2011
Tanzania	12	1980-2009
Nigeria	10	2007-2014
Egypt	9	2009-2016
Niger	5	2001-2014
Cameroon	4	2010-2012
Somalia	4	1977-1983
Libya	3	2013
Togo	3	2004-2005

Table 3-3: Number of FMD SAT1 sequences per country/region sampled

FMD SAT1		
Country	***Number of sequences***	***Year range***
South Africa	*45*	*1961-2010*
Kenya Uganda	*30*	*1970-2012*
Soutern Africa	*16*	*1977-2010*
Zimbabwe	*14*	*1990-2015*
Tanzania	*12*	*1971-2010*

Table 3-4: Number of FMD SAT2 sequences per country/region sampled

FMD SAT1		
Country	**Number of sequences**	**Year range**
South Africa	23	1986-2012
Zimbabwe	23	1979-2015
Kenya	12	1982-2013
Ethiopia	11	1990-2015
Namibia	10	1989-2011
Botswana	9	1978-2006
Uganda	9	1970-2013
Zambia	7	1982-1996
Libya	5	2003-2012
Tanzania	5	1975-2010
Nigeria	5	1975-2012
Egypt	5	2012-2015
Cameroon	5	2000-2005
Sudan	5	1977-2010

3.3.2 Bayesian Evolutionary Inference

3.3.2.1 Discrete phylogeographic tree inference

Time-scaled phylogenetic trees were inferred using BEAST 1.8 with the BEAGLE library[323], and different substitution clock and population evolution models were evaluated by estimating their marginal likelihoods using the Akaike's Information Criterion for MCMC samples (AICM) in Tracer 1.6. Ultimately, a general-time-reversible (GTR) model with site to site rate variation between two categories was selected as nucleotide substitution model[324], with a Bayesian skygrid population model and a relaxed uncorrelated log-normal molecular clock model were chosen to model the evolution of the FMDV A , FMDV O and FMDV SAT1 serotypes. The simpler HKY nucleotide substitution model with a constant clock model and a Bayesian skygrid population model were chosen to model the evolution of the FMDV SAT2 serotype[79,100] since this combination was favoured by the AICM scores and was

appropriate for this diverse dataset. Posterior sets of trees were generated for each serotype by combining at least 2 independent Markov Chain Monte Carlo runs of 40 million steps sampling every thousand with 10% burn-in with an ESS value above 200.

I first reconstructed the time-scaled phylogenetic trees for the four studied serotypes. Thereafter, to reduce the computation time of the GLM and the spatial diffusion analyses, I estimated the spatial model components using subsets of 1000 trees from the original posterior distributions of trees as input empirical tree distributions. I used TreeAnnotator to summarise maximum clade credibility (MCC) trees and FigTree version 1.4.1 to visualise the annotated trees[299,300]. The software SPREAD3 and Cytoscape were used to identify and visualise the well supported rates of transmission through a Bayes factor test[301].

For the four serotypes, I reconstituted the discrete transition events between the different sampled countries (or five African regions for SAT1) through the whole phylogeny using the “mugration model”. Therefore, an asymmetric continuous-time Markov chain (CTMC) model with an incorporated Bayesian stochastic search variable selection (BSSVS) was used to determine which set of transition rates sufficiently summarises the epidemiological connectivity between the countries[17]. A posterior inference of the complete Markov jump history through the whole genealogy was also performed, in order to quantify state transitions and infer the time spent in each state by the virus.

3.3.2.2 Environmental and anthropological effect estimation

3.3.2.2.1 Monophyletic clade selection

Using the previously reconstructed discrete phylogreographic tree of the FMDV O serotype, I selected a monophyletic clade with a MRCA under 25 years and a posterior probability over 50% on the location for all its nodes (see Supplementary figure 8-22). To avoid uncertainty in the predictor effect estimation analysis, I removed all sequences connected to branches with lengths more than 10 years. At the end of the process, the dataset was composed of 46 FMDV O sequences coming from 31 locations across Kenya, Uganda and Tanzania (see Supplementary table 8-37). The spatial coordinates of sampling for each sequence were retrieved using the GGMAP package in R and the most precise sampling localisation name available for each sequence[325].

3.3.2.2.2 Generation of predictive factors of FMDV diffusion

A Generalised Linear Model (GLM) extension of the discrete approach was used to test and quantify the enhancing (positive) or impeding (negative) effects of potential predictors on the viral diffusion process[110]. This model parametrises the transmission rate matrix between discrete locations as a log linear function of the potential predictive factor matrices. While reconstructing the phylogeographic history, the model performs Bayesian model averaging to determine which combinations of predictor matrices are best to explain the spatial diffusion process. For each predictor, a Bayes factor (BF) value is calculated based on the ratio of posterior to prior probabilities of inclusion[109].

The different predictors of FMD diffusion considered were: the accessibility of the sampled location (travel time to the nearest city of 50,000 people in 2000)[326], cattle density, crop density, elevation of the location, forest density, human density, average yearly precipitations, shrubland area density, average daily temperature (see Table 3-5 and for the provenance, see Supplementary table 8-23). Each potential predictor was retrieved as a raster matrix, representing the predictor spatial localisation, and aggregated to a resolution of 0.08 by 0.08, corresponding to pixels of approximately 8km by 8km (for an illustration of the rasters used in this analysis, see Supplementary figure 8-3 to Supplementary figure 8-15).

Table 3-5: Environmental and anthropological predictors tested for an effect on the FMDV serotype O diffusion in Eastern Africa. Those predictors were chosen following discussions with experts.

Environmental predictors	**Anthropological predictor**
Distance	Accessibility
Elevation	Cattle density
Precipitation	Presence of crop
Temperature	Presence of fragmented crop
Presence of forest	Human density
Presence of herbaceous vegetation	Logarithm of cattle density
	Logarithm of human density

The “circuitscape” software was used to determine the predictors’ values used in my GLM analysis[34]. For each predictor, two predictor values were generated, one using the raster as resistance values (impeding the viral diffusion) and the other using the raster as conductance value (enhancing the viral diffusion). To obtain those values, I used a circuit theory approach to estimate modified distances, used as predictor values, between each pair of locations using the raster values as heterogeneity factors[109]. Consequently, if a raster was used as a resistance surface, I would then estimate large predictor values between the locations separated by high raster values and small predictor values between the locations separated by small raster values.

Prior to their inclusion in the GLM analyses, the predictor values were log transformed and standardised. Each analysis was run by comparing the effect of a predictor with a null predictor, corresponding to a random raster.

Complementary to the discrete GLM approach, I tested and quantified the effect of potential predictors using a continuous coordinate approach. Therefore, I inferred the diffusion of the virus using a random walk model of diffusion and used the SERAPHIM package to test and estimate the effect of the predictors on the virus diffusion[98,115]. Like the discrete approach, SERAPHIM estimates a modified distance for each pair of locations found at the start and end of the phylogeny branches. The correlation between the time spent on each branch and the estimated distance value is then estimated. The statistical significance of this correlation is tested using a randomised phylogeny and expressed in the form of a BF[115].

3.4 RESULTS

3.4.1 Discrete phylogenetic analysis

3.4.1.1 Evolutionary parameters estimation

Overall, I observed a mean substitution rate of $4.67x10^{-3}$ substitutions per site per year and $3.69x10^{-3}$ for the serotypes A and O respectively. I also estimated a significantly slower substitution rate of $1.8x10^{-3}$ and $1.1x10^{-3}$ for the serotypes SAT1 and SAT2 (see Supplementary table 8-28).

3.4.1.2 Phylogeographic tree reconstruction for serotype A

The reconstructed phylogeographic tree of the African serotype A virus has a time to most recent common ancestor (TMRCA) of around 1926 (1890-1950 95% HPD), with geographic origin in the eastern part of Africa and high posterior probabilities for Kenya (49.83%) and Ethiopia (35.95%) (see
Figure 3-1a). For serotype A, there is no clear clade separation between the western and eastern sides of Africa, as the first isolated clade combines all the western African sequences as well as sequences from Sudan, Ethiopia and Egypt. Considering the lineages in serotype A, a few transmission events are observable between the two sides of Africa, and all of them involve Sudan as a link between them.

3.4.1.3 Phylogeographic tree reconstruction for serotype O

The TMRCA of the African serotype O is estimated to be 1937 (1921-1952 95% HPD) and located in the eastern part of Africa with high posterior probabilities for Kenya (61.49%), Sudan (17.15%) and Uganda (11.42%) (see

Figure 3-1b). The reconstructed phylogeographical tree is composed of four large clades. The first clade is almost entirely composed of Kenyan, Tanzanian and Ugandan sequences with only a few transmissions to other countries. The second clade is mostly situated in Ethiopia with few transitions to Kenya and Somalia. The third clade is centred in Sudan with incursions into Nigeria, Cameroon, Egypt and Ethiopia. The fourth clade is centred in West and Central African countries (Cameroon, Nigeria, Niger and Togo) and seems to originate from Sudan. Overall, we can see that the situation for the serotype O is quite similar to the one for the serotype A with only a few observed transmissions between the Eastern and Western sides of Africa, with Sudan acting as a link between the two sides of Africa.

3.4.1.4 Phylogeographic tree reconstruction for serotype SAT1

The TMRCA for the serotype SAT1 was estimated at 1755 (1665-1833 95% HPD) (see

Figure 3-1c). Due to the long timescale and low posterior probabilities near the root of the tree, it is difficult to estimate precisely the location of origin. Unsurprisingly, considering the composition of the dataset, the inferred origin location was in the southern parts of Africa (23% South-Africa, 37% Zimbabwe and 30% Southern Africa (other countries)).

The

Figure 3-1c shows three major clades with posterior probabilities above 75%. Clade 1 seems to have emerged in the middle of the 19th century and is composed of sequences coming from Kenya, Tanzania, Zimbabwe, Mozambique and Zambia (here indicated as ‘Southern Africa’). Clade 2 emerged at the end of the end of the

19th century and is composed of sequences almost entirely from South Africa (and one from Swaziland, coloured as 'Southern Africa' in

Figure 3-1c), with a single introduction into Zimbabwe. Clade 3 emerged at the start of the 20th century and is composed of sequences from Botswana and Namibia (coloured as 'Southern Africa' in

Figure 3-1c), with an introduction into South Africa and another one in Zimbabwe. It can also be seen that SAT 1 was introduced in the eastern part of Africa (Tanzania and Kenya) from Zimbabwe in a single introduction near the start of the 20th century.

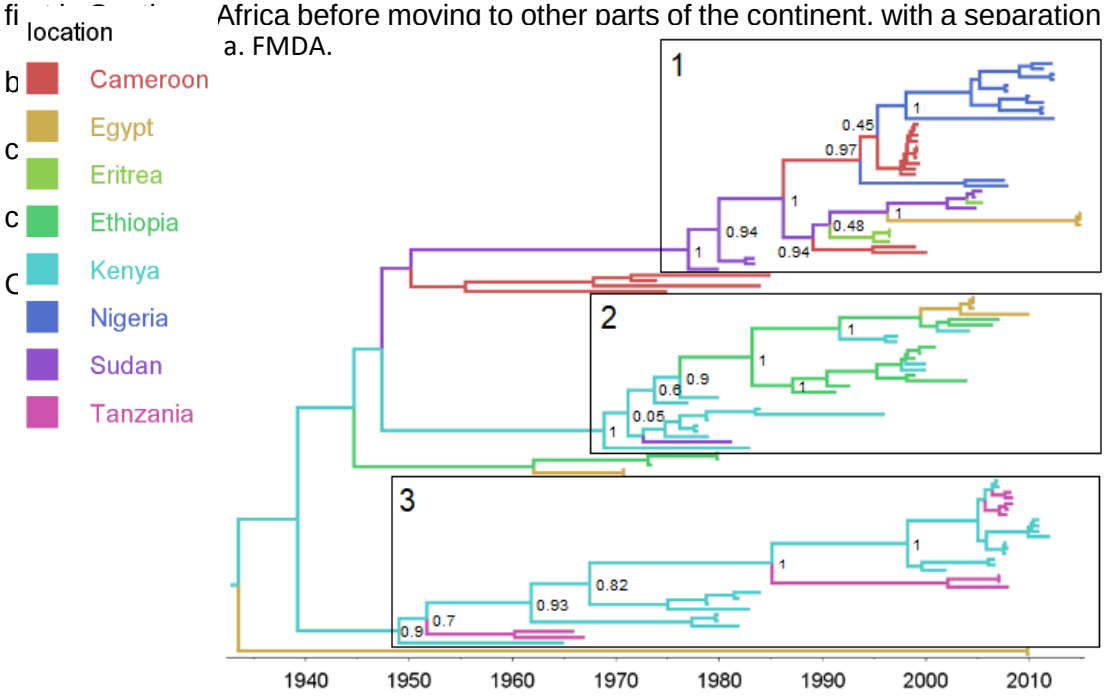
3.4.1.5 Phylogeographic tree reconstruction for serotype SAT2

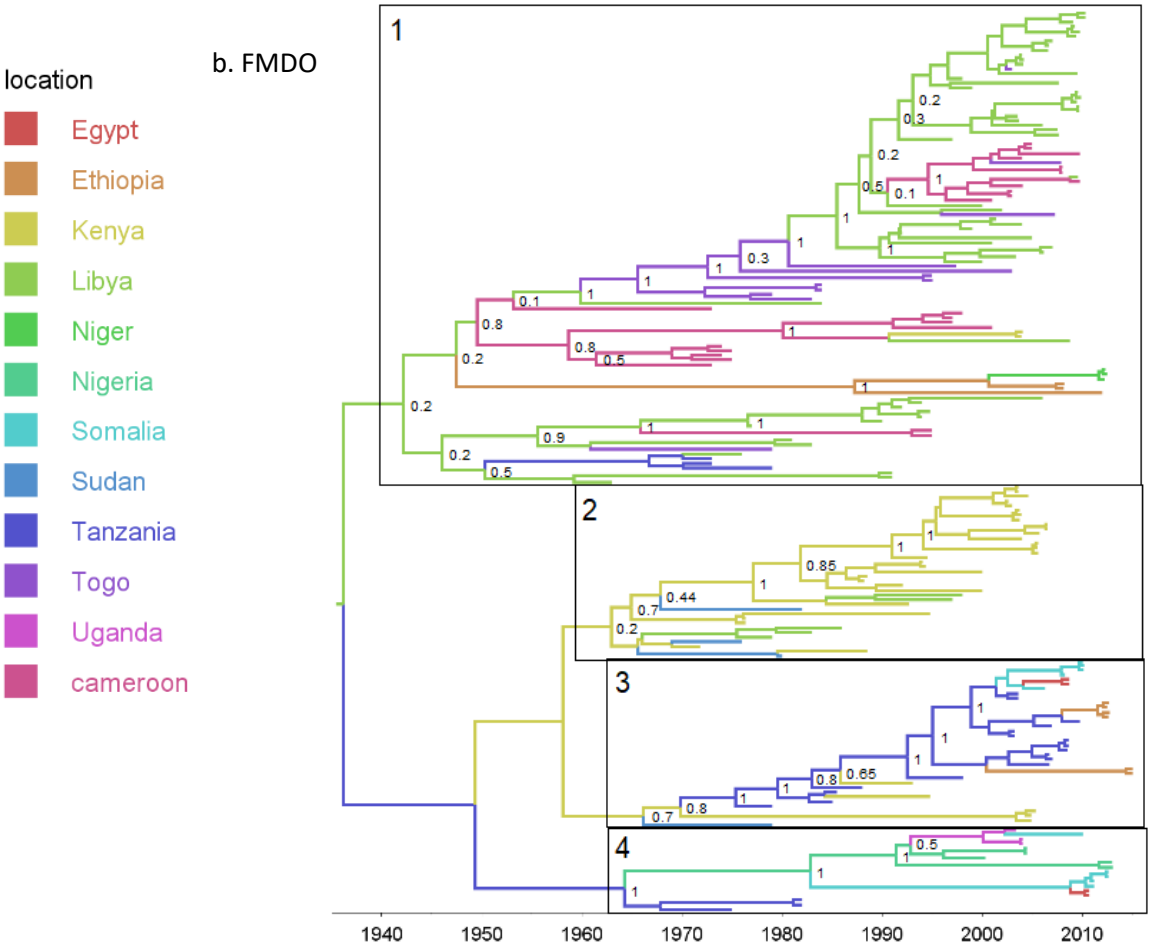
The TMRCA for serotype SAT2 is estimated as 1583 (1440-1722 95% HPD). Similarly to SAT1, due to these long timescales, long branches and low posterior probabilities for the location at the ancestral nodes, it is difficult to estimate an origin location (see

Figure 3-1d).

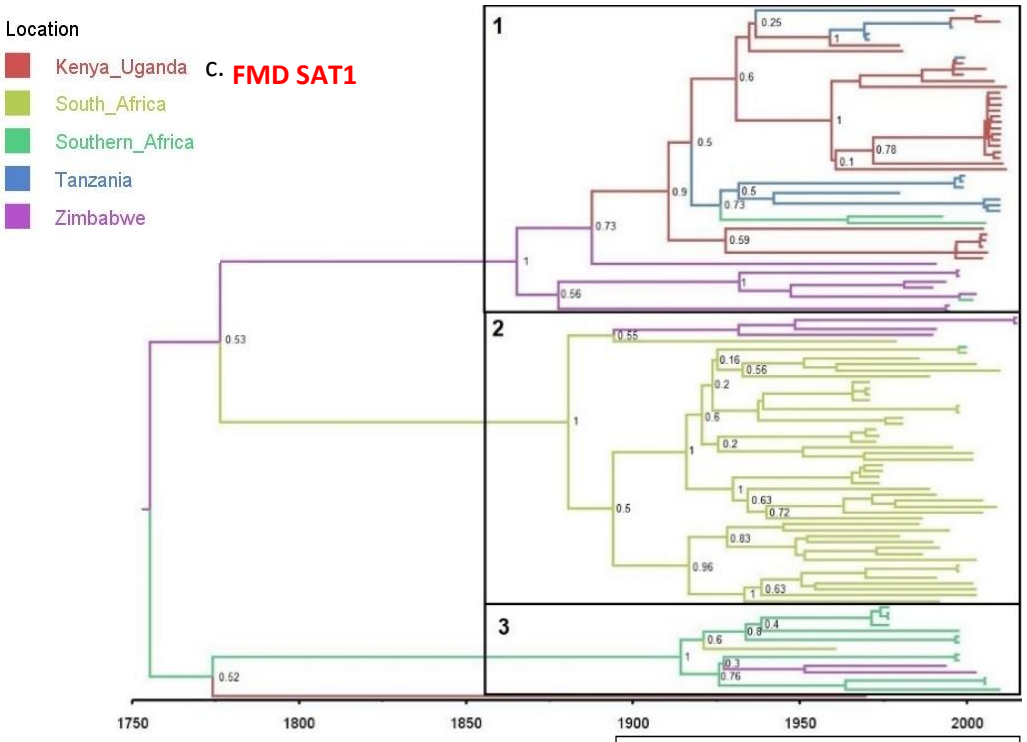
Five geographically defined main clades, with location posterior probabilities above 45%, can be observed. The first clade is exclusively composed of sequences from Botswana, Namibia and Zimbabwe and seems to have its origin in the first half of the 19th century. The second clade is composed of Ethiopian, Kenyan, Ugandan and Tanzanian sequences and seems to originate at the transition between the 19th and 20th centuries. The third clade seems to have emerged at the end of the 18th century and is composed of Zimbabwean and all the South African sequences. The fourth clade has its TMRCA in the first half of the 19th century and is composed of sequences from Botswana, Namibia and Zambia. The last clade emerged over the last century and is the most diverse in the observed locations, with sequences coming from

Eastern, Western and Northern Africa (Cameroon, Egypt, Ethiopia, Libya, Nigeria and Sudan). Similar to the SAT1 serotype, the SAT2 serotype seems to have appeared fi Africa before moving to other parts of the continent, with a separation





3.4. Results


Location
Kenya_Uganda
South_Africa
Southern_Africa
Tanzania
Zimbabwe
c. FMD SAT1
1
2
3
1750
1800
1850
1900
1950
2000

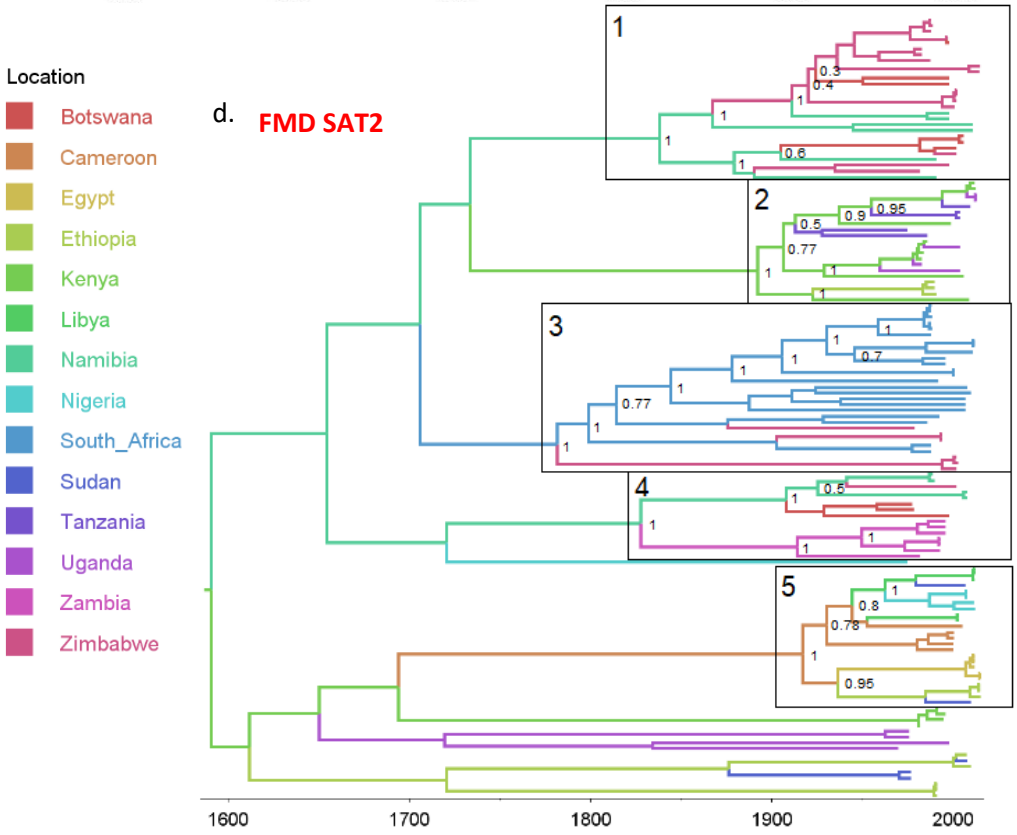

Location
Botswana
Cameroon
Egypt
Ethiopia
Kenya
Libya
Namibia
Nigeria
South_Africa
Sudan
Tanzania
Uganda
Zambia
Zimbabwe
d. FMD SAT2
1
2
3
4
5
1600
1700
1800
1900
2000

Figure 3-1: Bayesian MCC time scaled discrete phylogeographic tree for the four studied serotypes. a. Bayesian phylogeographic tree for serotype A using 107 VP1 sequences. b. Bayesian phylogeographic tree for serotype O using 192 VP1 sequences. c. Bayesian phylogeographic tree for serotype SAT1 using 117 VP1 sequences. d. Bayesian phylogeographic tree for serotype SAT2 using 135 VP1 sequences. The phylogeny branches are coloured according to their descendent node location with the key for colours shown on the right. The main clades for each of the studied serotypes are identified on the phylogeographic trees. The nodes of the isolated clades are annotated with their posterior probabilities.

3.4.1.6 Bayesian stochastic search variable selection analysis

Using a Bayesian stochastic search variable selection (BSSVS) analysis, I identified well-supported rates of transition between the sampled countries. The support for the rates was quantified with Bayes factors (BF), and rates with BF >=3 are shown in Figure 3-2. Globally, the results for the serotypes A and O look quite similar, with Sudan acting as a link between the Eastern, Northern and Western parts of the continent (see Figure 3-2a and Figure 3-2b). For both serotypes, there is a clear transmission route starting from Ethiopia, passing through Kenya to Tanzania.

Although the observed pattern for the SAT1 serotype is slightly different, due to the lack of data from outside Southern Africa, we can still observe strong links between Tanzania and Kenya-Uganda for this serotype (see Figure 3-2c). I observed multiple links between South-Africa, Zimbabwe and other countries in the Southern-African region (Zambia, Botswana, Namibia, Mozambique and Swaziland). Additionally, a well-supported transition was observed between the Zimbabwe and Kenya-Uganda region, however this should not be interpreted as direct and contemporary link between these two regions.

Despite having lower BF values, the situation for the SAT2 serotype is fairly similar to what is observed for serotypes A and O (see Figure 3-2d). For SAT2, multiple transitions rates can be observed within Eastern and Western Africa with Sudan acting as link between the two sides. However, with only two rates linking South Africa to the rest of the continent, in general, Southern African are quite isolated from the

other African countries (see Supplementary table 8-29, Supplementary table 8-30, Supplementary table 8-31 and Supplementary table 8-32).

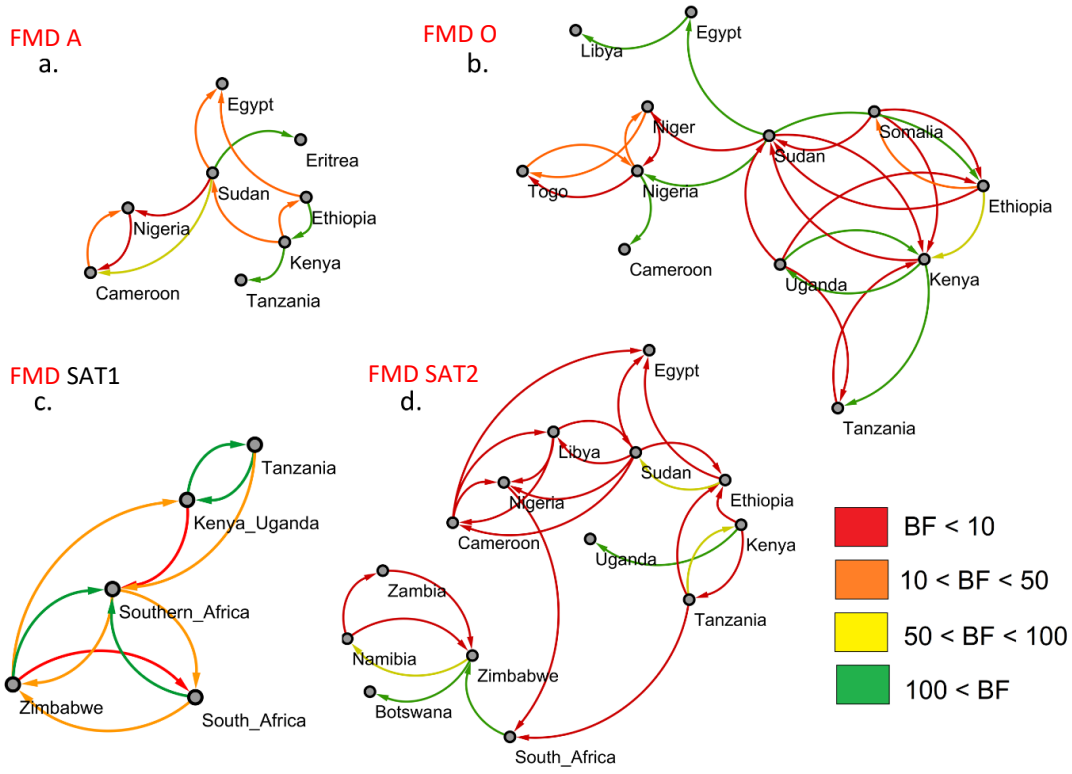


Figure 3-2 : Output of the BSSVS analysis for the four studied FMDV serotypes showing the best supported rates of transition between the sampled countries. The edge colours represent the relative strength by which the rates are supported. a. For FMDV serotype A. b. For FMDV serotype O. c. For FMDV serotype SAT1. d. For FMDV serotype SAT2.

3.4.1.7 Markov jumps analysis

To complement the BSSVS analysis, an estimation of the number of transmission events between the different locations using a Markov jump analysis was performed (see Figure 3-3). For both FMDV serotypes A and O, I observed some transmission events from Kenya to other East African countries such as Tanzania, Ethiopia and Uganda (see Figure 3-3a and Figure *3-3*b). For these two serotypes, I also detected jumps from Sudan in the direction of North-Eastern and Western African countries such as Egypt, Eritrea, Cameroon and Nigeria. Therefore, it seems that for these two serotypes, Kenya and Sudan act as distributors of the virus, but toward different directions (see Supplementary table 8-33 and Supplementary table 8-34). For the SAT1 serotype, I observed very few transitions between the Southern African regions and many transitions between Kenya-Uganda and Tanzania (see Figure 3-3c and Supplementary table 8-35). Only one transition from Zimbabwe to Kenya-Uganda was observed in SAT1 (see Figure 3-3c), and this occurred around the 1900s (see Figure 3-1c). For serotype SAT2, most of the observed transitions occurred within Eastern African and Southern Africa with no clear link between them (see Figure 3-3d and Supplementary table 8-36).

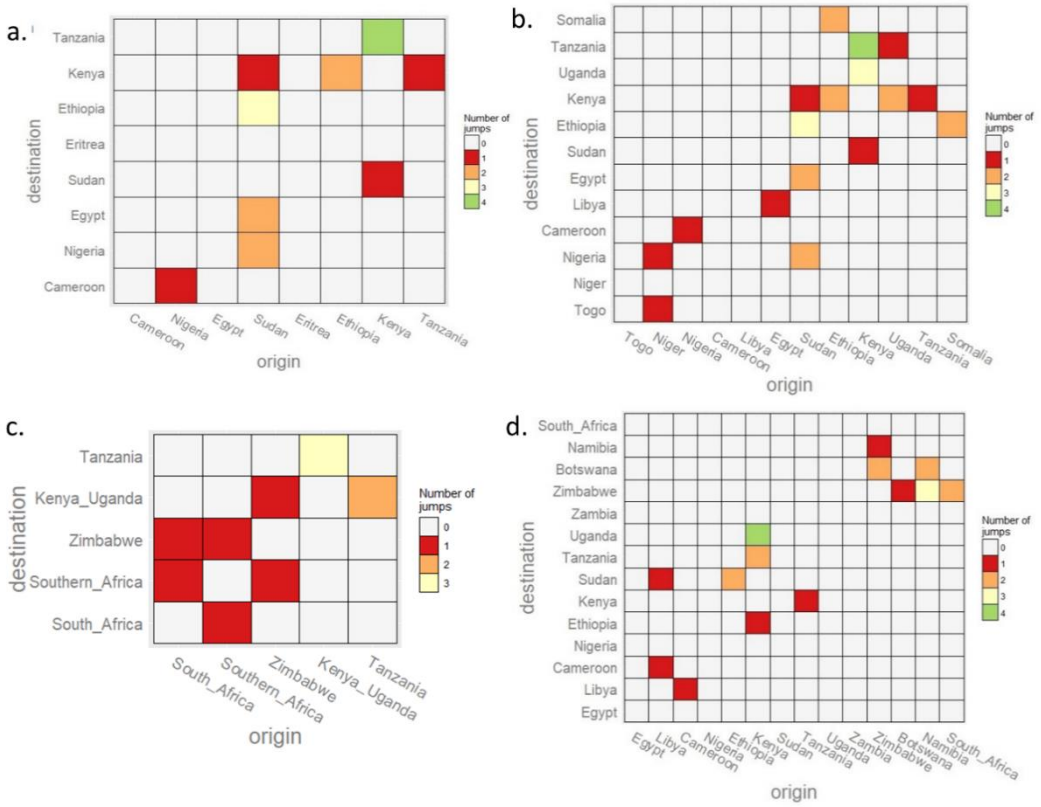


Figure 3-3: Heatmap showing the number of transitions between the sampled countries for the three studies FMDV serotypes. The heatmaps are coloured according to the number of observed transitions between countries. a. FMDV serotype A. b. FMDV serotype O. c. FMDV serotype SAT1. d. FMDV serotype SAT2.

3.4.2 Environmental and anthropological factors affecting FMDV diffusion

Using the output from previous discrete phylogeographical analysis, I isolated a FMDV serotype O monophyletic clade with a time to the most recent common ancestor (TMRCA) below 25 years and a high posterior probability on the location for all its nodes. The selected 46 sequences originated from Kenya, Tanzania and Uganda (See Supplementary figure 8-22 and Supplementary table 8-37). Using a general linear model (GLM) for the discrete location approach and the recently developed SERAPHIM package[115] for the continuous location approach, I tested the impact of 13 different anthropological and environmental factors (predictors) on the FMDV diffusion in Eastern Africa.

Table 3-6: Bayes factor values associated with the effect of each predictor on the connectivity between the sampled locations using a discrete or continuous location approach. Each predictor raster was used as conductance or resistance to evaluate if the predictor have a positive or negative effect on the viral genetic connectivity.

	Discrete locations		**Continuous locations**	
Predictor	**Conductance**	**Resistance**	**Conductance**	**Resistance**
	Bayes factor	**Bayes factor**	**Bayes factor**	**Bayes factor**
Accessibility	0	8	0	0
Cattle density difference	-	0	-	-
Cattle density	2	0	0	3
Presence of crop	0	0	1	0
Presence of crop (combined)	0	0	0	4
Distance	-	8	-	-
Elevation	1	0	0	1
Presence of forest	1	0	1	0
Presence of fragmented crop	1	0	0	6
Human density	1	0	1	0
Precipitation	2	7	0	0
Presence of herbaceous vegetation	1	0	0	0
Temperature	4	7	0	1
Logarithm Cattle density	4	1	1	0
Logarithm Human density	9	2	0	0

3.4.2.1 Predictive factors for FMDV diffusion using a discrete location approach

A generalized linear model (GLM) was used to parametrise the transition rate matrices between the sampled locations as a function of the selected predictors[110] on a posterior set of time-resolved trees. I considered the set of predictors to be 'conductors' – i.e. enhancing viral diffusion, or 'resistors' – i.e. impeding viral diffusion. I observed that the diffusion process was enhanced by the average daily temperature (BF 4), the logarithm of the cattle density (BF 4) and human densities (BF 9) (see Table 3-6). It was impeded by accessibility (BF 8), distance between sampled locations (BF 8), average amount of precipitation (BF 7) per year, and by average daily temperature (BF 7) (for all the results, see Supplementary table 8-38 and Supplementary table 8-39). To gain a better understanding of the impact of the average temperature and precipitation on the viral diffusion, I selected different thresholds of precipitation and temperature to parametrise the GLM analysis (see Supplementary table 8-40 and Supplementary table 8-41). I detected that low precipitation values (<80mm/year) were associated with an impeding (negative) impact on the viral diffusion processes, whereas high precipitation was associated with an enhancing (positive) effect on the diffusion process. I also observed that in the case of low temperature (below 22℃), a positive effect on the diffusion could be observed whereas temperatures around 22℃ had a negative effect on virus diffusion. Temperatures above 24℃ seemed again to have a positive impact on the virus spread. It was difficult to distinguish between the effects of accessibility and human density because the two were strongly negatively correlated, thus confounding the analysis (see Supplementary table 8-42).

3.4.2.2 Predictive factors for FMDV diffusion using a continuous diffusion approach

Using a random walk model, I was able to reconstruct the virus diffusion in a continuous setting for the isolated FMDV serotype O (depicted in Figure 3-4). Using the R package SERAPHIM[115], I evaluated the impact of the predictors on the virus diffusion speed and observed a negative influence of the cattle density (BF 3), the presence of cropland (fragmented cropland and pure cropland areas combined) (BF4) and by the presence of fragmented cropland (BF6). I was not able to detect a predictor with an enhancing (positive) influence on the diffusion process (see Table 3-6).

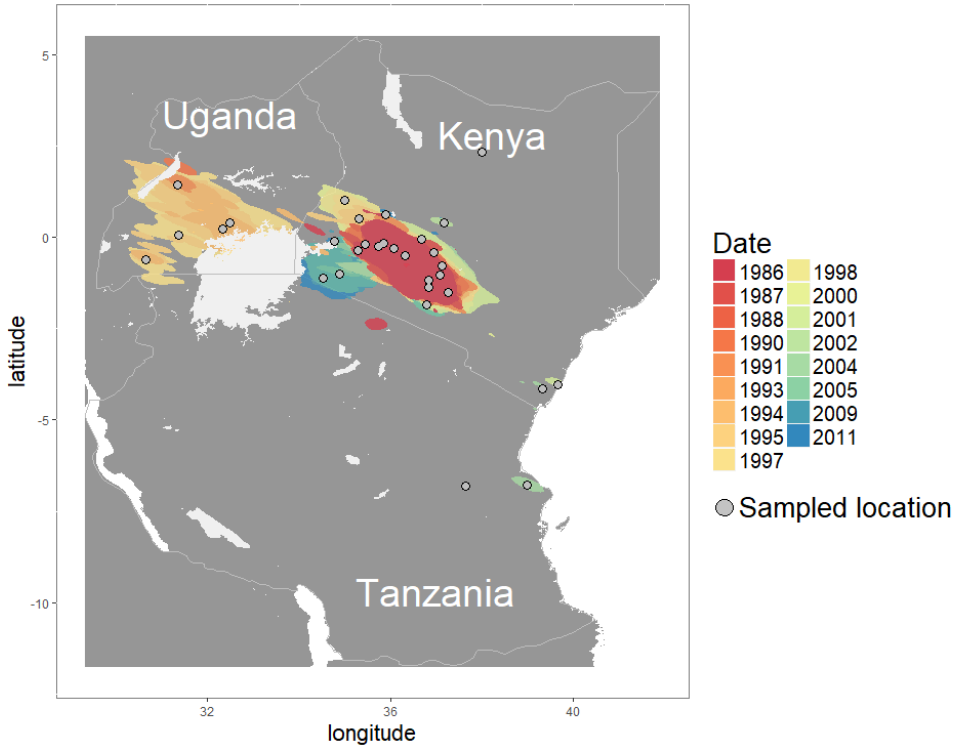


Figure 3-4: Map showing the continuous diffusion of the isolated clade of FMDV serotype O, with the sampled locations as grey circles. The virus movements were reconstructed using a random walk model with an underlying lognormal distribution

To gain a better understanding of the role of the fragmented crop and cattle density, I isolated the areas newly covered over the course of the infection and observed how the presence of the two predictors evolved. Overall, I noticed an opposite trend in how their densities evolved with the elapsed time (see Supplementary figure 7-24). For the fragmented crop density, high values of crop densities became more common over the course of the epidemic, with the disease moving from areas with low crop densities to areas with high crop densities. For the cattle density, the opposite trend was observed, with high values of cattle density more common at the start of the epidemic, the disease starting in an area where cattle density was high and moving toward areas with lower density of cattle. To better understand the effect that the cattle density had on virus diffusion, I looked at selected areas above different thresholds of cattle density and used them as inputs in the SERAPHIM package. By doing so, I was able to observe that densities of cattle above 125 cattle per square kilometre have the biggest negative impact on the virus diffusion (see Supplementary table 8-43).

3.5 DISCUSSION

For this work, I have applied state-of-the-art phylogenetic methods to the available African FMDV VP1 sequences for the serotypes A, O, SAT1 and SAT2.

The estimated substitution rates of $4.67x10^{-3}$, $3.69Ex10-^{3}$,$1.7x10^{-3}$ and $1.1x10^{-3}$ substitutions per site per year for serotypes A, O and SAT2 from my results are similar to previous estimates of 4.26×10^{-3}, 3.14×10^{-3} and 1.07×10^{-3} substitutions per site per year for the same serotypes, as found by Tully et al[327]. The observed rate of $1.7x10^{-3}$ for the SAT1 is significantly lower than the rate found by Tully et al.[327] but similar to the one found by Sangula et al. [328], who already pointed out this difference. Overall, I observed similar evolutionary patterns for both FMDV serotypes A and O. My findings suggest that those serotypes have appeared in Eastern Africa around 1930, which is consistent with what previous findings suggested[234]. Additionally, my results point the possible role of Kenya as a viral source for East African countries and the role of Sudan as a link between East Africa and North-East Africa. The evolution of the SAT1 and SAT2 serotypes seems to be quite different from that of the serotypes A and O since both SAT serotypes were present in Southern-Africa up to 400 years prior to the appearance of serotypes A and O in Africa. Interestingly, both SAT serotypes appear to have spread outside of Southern Africa around the start of the 20th century. For the FMDV serotypes A and O, the observation of well supported rates of viral transmission between Eastern Africa and Western Africa can be explained by the existence of commercial routes between those areas. It is indeed acknowledged that livestock trades play an important role in FMDV circulation in sub-Saharan Africa, through trading routes existing between the Horn of Africa and Eastern Africa, with Sudan acting as key commercial intermediate[220]. Additionally, the existence of a relatively recent common ancestor for the FMDV serotypes A and O further supports

the idea that these serotypes were imported into Africa at the start of the 20th century through livestock trade from Asia and Europe[234].

The SAT1 and SAT2 serotype analysis shows the signs of the impact that the African rinderpest epidemics (that occurred in 1890[329]) had on FMDV circulation in Africa. Although FMDV was first reported in Southern Africa in 1795, it had likely coevolved with buffaloes over millennia, resulting in a large diverse viral pools, but the rinderpest epidemic decimated almost all FMDV potential carriers and probably pushed it through a huge bottleneck[330]. It is thought that FMDV re-emerged from wild buffalo populations that survived the rinderpest epidemic, before being reported again in 1932 in Southern Africa[234]. Consistent with this hypothesis, the introduction of SAT1 in Eastern Africa seems to have occurred after the rinderpest epidemic. Similarly, of all the clades present in the reconstructed phylogeny for SAT2, only those originating from Southern African countries have a TMRCA older than the African rinderpest epidemic (clades 1, 3 and 4). The SAT1 and SAT2 serotypes probably spread outside Southern African countries through infected livestock movements, increasing virus mobility, explaining the more recent emergence of the virus in Eastern and Western Africa and the apparent more common transboundary movements observed in those regions. This idea is supported by observation of the BSSVS analysis output that show similar transmission patterns amongst the serotypes A, O and SAT2. This would suggest that, outside of South Africa and over the last hundred years, the SAT1 and SAT2 circulation was mainly driven by domestic animal movements, and the observed relative isolation of Southern Africa is the result of the different livestock trades control measures in place in this region[331]. However, I cannot exclude with certitude the role that wildlife might play in those regions considering the lack of wildlife samples outside Southern Africa and the frequently observed interactions between domestic and wild animals in those regions[332].

Using both discrete and continuous frameworks, I looked at the effect that diverse environmental and anthropological factors had on the diffusion of an isolated FMDV serotype O clade that circulated in Kenya, Uganda and Tanzania. The results of the discrete approach suggest that the FMDV diffusion is facilitated by low average daily temperature (<22℃), high average precipitations (>80 mm/year) as well as high human and cattle densities. I also saw that the virus diffusion was negatively impacted by accessibility i.e. slow viral diffusion was associated with long travel times, as well as high daily temperatures and low average precipitations.

Since lower temperatures and higher humidity values are usually associated with a longer virus survivability in the environment[333], these results raise the possibility of a more important role than what was previously believed of the indirect transmission through viral persistence in the environment for FMV in this region. Additionally, with the viral diffusion being positively affected by high cattle and human densities, and negatively affected by large accessibility values, anthropological activities seem to have an impact on the virus diffusion. These observations could be the consequence of infected herds of cattle moving from smaller rural localities toward nearby larger cities with cattle markets[314,334].

Regarding the effect of the different selected predictors on virus diffusion in a continuous setting, my results suggest that cattle densities above 125 cattle per km^2 and the presence of cropland (pure cropland or mixed with other types of land) both have a negative impact on virus diffusion. My results suggest that the virus had difficulties to spread beyond the geographic region located at the root of the tree, where high cattle densities and low crop densities were present, and to spread to areas with low cattle densities but high crop densities, presumably due to lack of suitable hosts. However, it is difficult to know exactly whether it is the cattle or crop

density that had the most impact due to high correlation of the two variables at the time and in the region of origin.

The location uncertainty found at the root of the continuous tree could explain the differences between the discrete and continuous methods in estimating the effect of the cattle density on virus diffusion. For my analysis, this uncertainty seems to be translated by the SERAPHIM programme as a period where the virus is almost not moving. This uncertainty seems to drive SERAPHIM to the conclusion that the high cattle densities found near the origin of the epidemic are related to this lack of movement, and therefore estimate that they have a negative influence on the virus diffusion by slowing its wavefront. Although I suspect a link between the cattle density and the location of emergence of the analysed clade, I think that the continuous analysis does not offer the resolution needed to understand that relation (i.e. the spatial HPD confidence interval is too large). By parameterising each rate of among-location movement as a function of predictors, the discrete approach seems therefore more appropriate to characterise the environmental and anthropological effects of the virus diffusion in this endemic situation.

In conclusion, the reconstructed phylogeographical tree pattern for the FMDV serotypes A, O, SAT1 and SAT2 reflects a situation where the recent FMDV circulation between non-Southern-African countries is mainly driven by commercial exchanges, through pastoral herd movements, where wildlife seems to have almost no influence on the circulation of the disease. The observations for A and O suggest that those serotypes were imported in Africa at the start of the 20th century, while the observed patterns for SAT1 and SAT2 reflects a situation where wildlife constitutes the original host of the serotype. I observed that indirect transmission through the environment and direct transmission through anthropological activities had an enhancing effect on the virus diffusion in Eastern Africa.

My work has some limitations, especially regarding the limited availability of sequences. My sampling is obviously unbalanced as it is based on submissions by individual countries or ad hoc research projects, and the effect that it has on the results quality is uncertain. On the other hand, I am combining sequence data from many African countries and over as large a time span as possible; I have applied a subsampling scheme to reduce over-representation and bias as much as practical; and have used both discrete and continuous Bayesian phylogeographic methodologies which are able to infer transmission patterns in sparsely sampled situations. Nevertheless, increasing the number of available FMDV sequences from diverse locations and hosts would help to develop models that better represent the diffusion of FMDV in Africa and lead to better environmental and anthropological effect estimation.

Although previous studies have suggested similar findings on the origins and natural hosts for FMD in Africa[55,56], and on the disease circulation being driven by livestock rather than by wildlife[220,270], my work represents the first comprehensive quantitative analysis of continental scale in support for different epidemiologies between the serotypes, and on different roles played by wildlife and livestock animals on the virus circulation. The use of such analytical methods is important in developing a modern approach to FMD control where different serotypes could be targeted and controlled in different regions, particularly in areas where wildlife may be less important.

3.5. Discussion

4 IMPORTANCE OF WILDLIFE IN THE CIRCULATION AND MAINTENANCE OF FOOT AND MOUTH DISEASE VIRUS SAT1 AND SAT2 IN AFRICA

4.1 ABSTRACT

Foot and mouth disease (FMD) virus is endemic in sub-Saharan Africa. Due to the complexity of the disease epidemiology and the lack of available data, there is a need to use phylogenetic approaches to understand the role of potential hosts involved in the circulation and maintenance of the virus. The uneven host sampling of the available sequences requires us to take advantage of the recent advances in phylogenetic reconstruction. Therefore, using two structural coalescent model approximations, I estimated the circulation of FMD virus serotypes SAT1 and SAT2 between cattle, buffalo and impala populations. The results suggest that in Africa, the impala population seems to act as an intermediate host between the cattle and buffalo populations and plays a more important role in the circulation of the disease than was previously suspected. Until now, the role of the impala population in the circulation of FMDV has been suggested, but never explicitly shown.

4.2 INTRODUCTION

Foot and mouth disease (FMD) is a vesicular disease affecting more than 70 species of cloven-hoofed animals, including domestic ruminants and pigs[216]. The causal agent of foot and mouth disease is a positive-sense, single-stranded RNA virus (FMDV). Since the most significant hosts in the natural epidemiology of FMDV are of major importance in the production of food (cattle, sheep, pigs, goats), the disease can potentially lead to important direct and indirect economic impact in countries with a developed agricultural industry[225,231,232].

The most common transmission route between infected and susceptible hosts is by direct contact. In this situation, the viral transmission is mechanical, with virus entry through skin cuts or mucosae, following physical contact with infected secretions or excretions[216,235]. Based on the level of cross protection between each strain, we can divide the virus population into seven distinct viral serotypes: O, A, C, Southern African Territories (SAT) 1, SAT 2, SAT 3 and Asia 1[218,219].

With five of the seven possible serotypes present in Africa over the last decade and with high regional variances in both their distribution and prevalence[267], the epidemiology of FMD in Africa is considered to be more complex than anywhere else. It is generally accepted that FMDV originated on the African continent due to the long-term subclinical infection status observed in the African buffalo (Syncerus caffer) and the important genetic diversity observed in the SAT serotypes[234,267]. Moreover, although most FMD outbreaks in sub-Saharan Africa go unrecorded, FMDV is considered endemic in almost all sub-Saharan African countries. This situation might be partially explained by the extensive livestock-raising system which is practised in several African regions, leading to a low direct impact from the disease[267].

Amongst all the wildlife species in Africa that are susceptible to FMDV, only the buffalo and the impala (Aepyceros melampus) have been implicated in transmission of the virus to cattle[226,274]. In Southern Africa, buffaloes are suspected to be the source for many livestock outbreaks, in particular those concerning the SAT-type FMD viruses for which they are considered the true maintenance host[224,265]. Hence, it is considered that the transmission from buffaloes to domestic animals is the dominant pathway of disease propagation[216,274]. For example, in Ethiopia, it has been observed that the cattle populations with the highest FMDV antibody prevalence are those in close contact with wild animals and located near wildlife sanctuaries where large populations of African buffaloes can be found[275,336]. However, there is still a lot of uncertainty regarding the role of buffaloes elsewhere in Africa as a source of livestock outbreaks, and how those buffalo populations are able to sustain endemic cycles of infections in livestock[234]. Additionally, although impalas do not seem to become persistently infected with FMDV, they have been suspected to act as intermediate hosts between buffalo and cattle populations in Southern Africa[276].

Due to a lack of a proof-reading mechanism in the FMDV genome, replication of the virus is subject to high mutation rates[218]. The unobserved ecological and population events impacting virus evolution can be estimated while reconstructing the phylogenetic tree of the epidemic[5,13]. By combining genetic and epidemiological information, we can estimate the circulation of fast evolving pathogens between multiple discrete states such as the location and the host involved[4,17].

This approach has already been used on a few occasions to study the role of the cattle and buffalo populations in the circulation and maintenance of FMD viruses in Africa[271,335]. However, the results of such analyses need to be interpreted with caution due to the uneven spatiotemporal sampling across the different discrete states used in these analyses[9,96,97,271]. It has been shown that the phylogeographical method used

in these papers (the “mugration” model[17]) suffers from statistical bias, exacerbated by an uneven sampling between the populations[25,126]. A way to mitigate the impact of these biases is to use a structural coalescent phylogenetic model approximation[23,25]. Such approaches, by taking into account the effect of the migration events on the phylogeny structure, have a better model specification in their phylogenetic and evolutionary parameter estimations, such as the effective population size (Ne) of the viral populations estimated in each of the analysed discrete state[24]. Ne being the size of an idealised population randomly mating and having the same gene frequency changes as the whole population under study[75].

The aim of this paper is to compare the results of two recent structural coalescent model approximations (BASTA[23] and MASCOT[28]) with the previously used “mugration” approach. Therefore, I estimated the evolution and transmission between cattle, impala and buffalo populations of FMDV serotypes SAT1 and SAT2 across Africa.

4.3 MATERIALS AND METHODS

Two datasets of FMDV SAT1 and SAT2 strains compiled for a previous analysis were used[337]. The FMDV SAT1 dataset was composed of 117 sequences with dates ranging from 1961 to 2015 and including 16 sequences from impalas, 35 sequences from buffaloes and 68 sequences from cattle. The FMDV SAT2 dataset was composed of 135 sequences with dates ranging from 1970 to 2015 and including 7 sequences from impalas, 34 sequences from buffaloes and 98 sequences from cattle (see Table 4-1 and Table 4-2 for the number of sequences per host and location, and Supplementary figure 8-24 and Supplementary figure 8-25 for the annotated "mugration" trees with the host and location for each sequence).

Table 4-1:Host and origin of the SAT1 sequences utilised in the phylogenetical analysis. The year of sampling range can be seen for each one of the combinations of host and origin.

	Buffalo (Date range)	**Cattle (Date range)**	**Impala (Date range)**	**Total per country**
Kenya-Uganda	2 (1970-2012)	27 (1980-2012)	1 (2010)	30
South-Africa	21 (1986-2005)	14 (1961-2010)	10 (1971-1998)	45
Southern-Africa	3 (1998-2000)	11 (1977-2010)	2 (1977)	16
Tanzania	3 (2010)	9 (1971-1999)	0	12
Zimbabwe	7 (1990-1998)	7 (1994-2015)	0	14
Total per host	36 (1970-2010)	68 (1971-2015)	13 (1977-2010)	117

Table 4-2: Host and origin of the SAT2 sequences utilised in the phylogenetical analysis. The year of sampling range can be seen for each one of the combinations of host and origin.

	Buffalo	**Cattle**	**Impala**	**Total per country**
Botswana	3 (1998)	6 (1977-2006)	0	9
Cameroon	0	5 (200-2005)	0	5
Egypt	2 (2012-2015)	3 (2012-2015)	0	5
Ethiopia	0	11 (1990-2015)	0	11
Kenya	0	12 (1982-2012)	0	12
Libya	0	5 2003-2012()	0	5
Namibia	6	4 (1989-2008)	0	10
Nigeria	0	5 (1975-2012)	0	5
South-Africa	10 (1998-2010)	6 (2001-2012)	7 (1985-1992)	23
Sudan	0	5 (1977-2010)	0	5
Tanzania	0	5 (1975-2009)	0	5
Uganda	1 (1970)	8 (1976-2013)	0	9
Zambia	4 (1993-1996)	3 (1981-1996)	0	7
Zimbabwe	8 (1988-2002)	15 (1979-2015)	0	23
Total per host	34 (1970-2015)	93 (1975-2015)	7 (1985-1992)	134

The “mugration” phylogenetic trees were reconstructed using BEAST 1.8 with the BEAGLE library[323]. For both serotypes, a Hasegawa-Kishono-Yano (HKY) nucleotide substitution model with a constant clock model and a Bayesian skygrid population model were chosen to model the evolution the virus[79,100].

I first reconstructed the time-scaled phylogenetic trees for the two studied serotypes by combining at least two independent Markov Chain Monte Carlo runs of 40 million steps sampling every thousand with a 10% burn-in. Thereafter, to reduce the computation time needed for the hosts’ diffusion analysis, I used subsets of 1000 trees from the original posterior distributions of trees as empirical tree distributions.

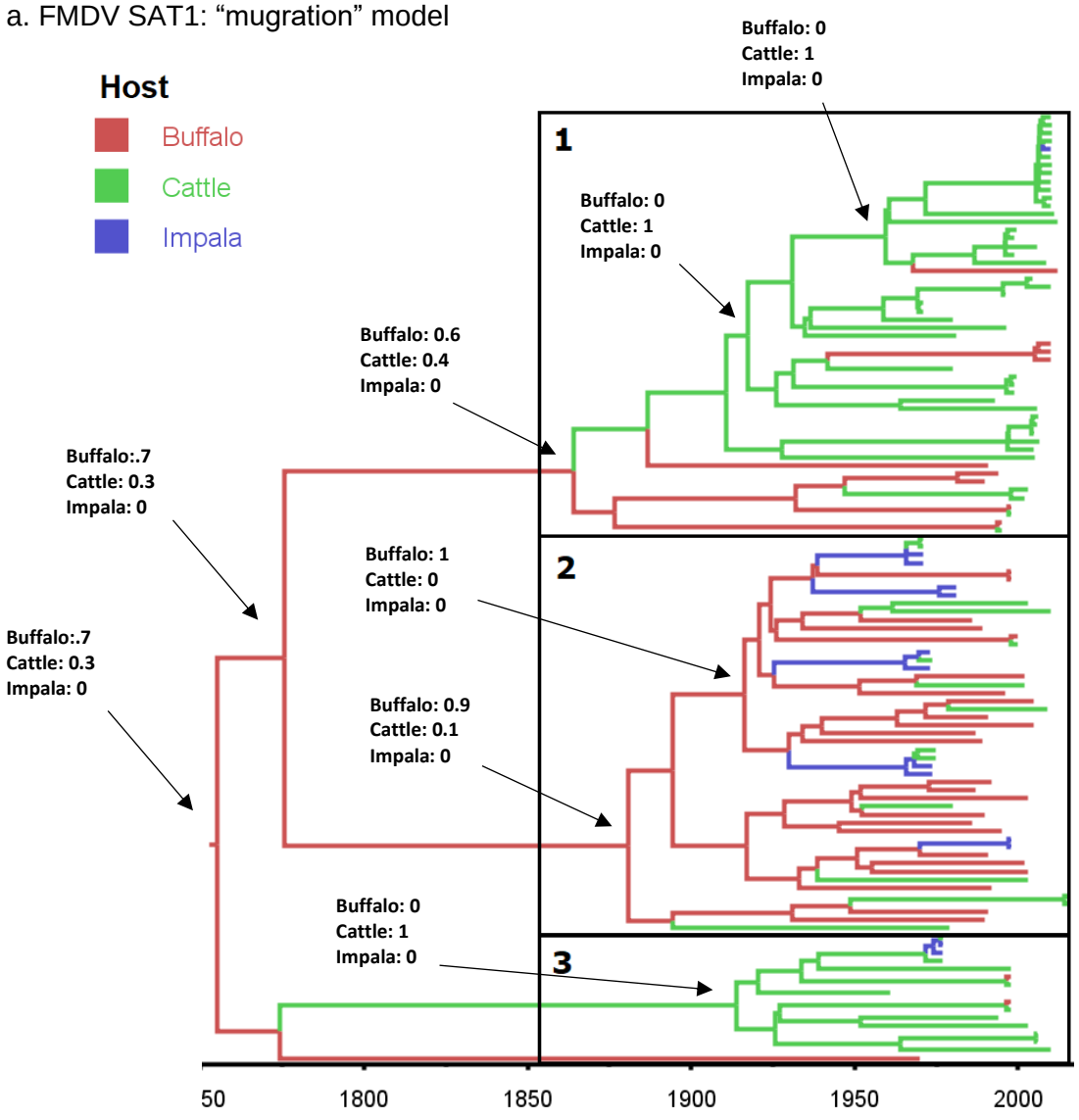
Simultaneously, I performed a phylogenetic tree reconstitution using two structural coalescent model approximations with the BASTA[25] and MASCOT[23] packages available within BEAST2[300]. For both models and serotypes, I used an HKY model of molecular evolution and a constant clock model. For each analysis, I combined three converging runs of at least 50 million steps (BASTA) or 10 million steps (MASCOT).

For all the analyses, I used TreeAnnotator to summarise maximum clade credibility (MCC) trees and FigTree version 1.4.1 to visualise the annotated trees[299,300].

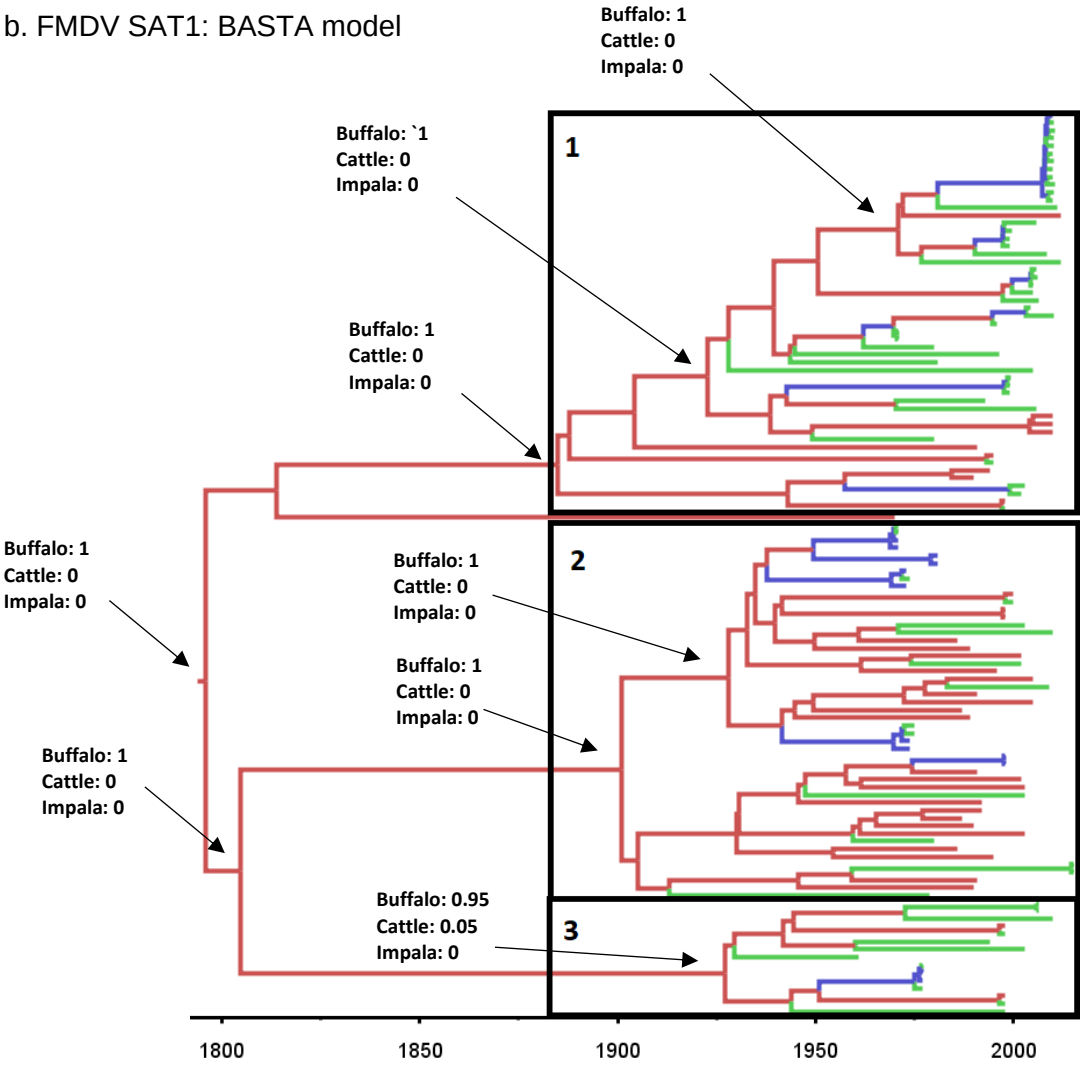
For the “mugration” and BASTA approaches, I was able to perform a Markov jump analysis to determine the number of transmission events that occurred between the three host over the whole phylogeny[17].

4.4 RESULTS

a. FMDV SAT1: “mugration” model



b. FMDV SAT1: BASTA model



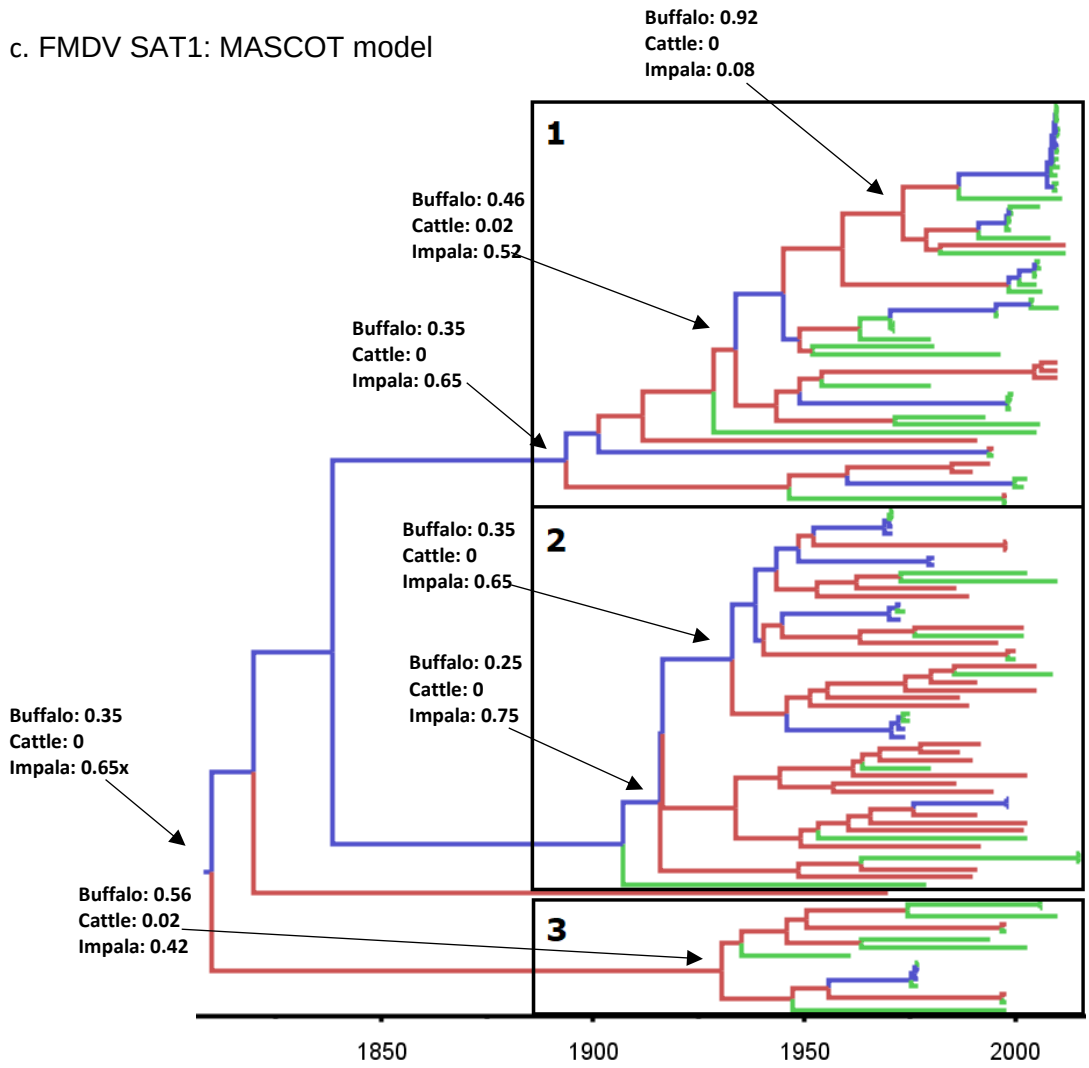


Figure 4-1: Bayesian MCC time scaled discrete phylogeographic tree for the serotype SAT1 using 113 VP1 sequences. a. Phylogenetic tree estimated using the “mugration” approach implemented in BEAST b. Phylogenetic tree estimated using the BASTA approach implemented in BEAST2. c. Phylogenetic tree estimated using the MASCOT approach implemented in BEAST2. The phylogeny branches are coloured according to their descendent nodes host with the key for colours shown on the upper left of the figure. The identified clades were isolated and numerated. Specific nodes of the trees were annotated with hosts posterior probabilities.

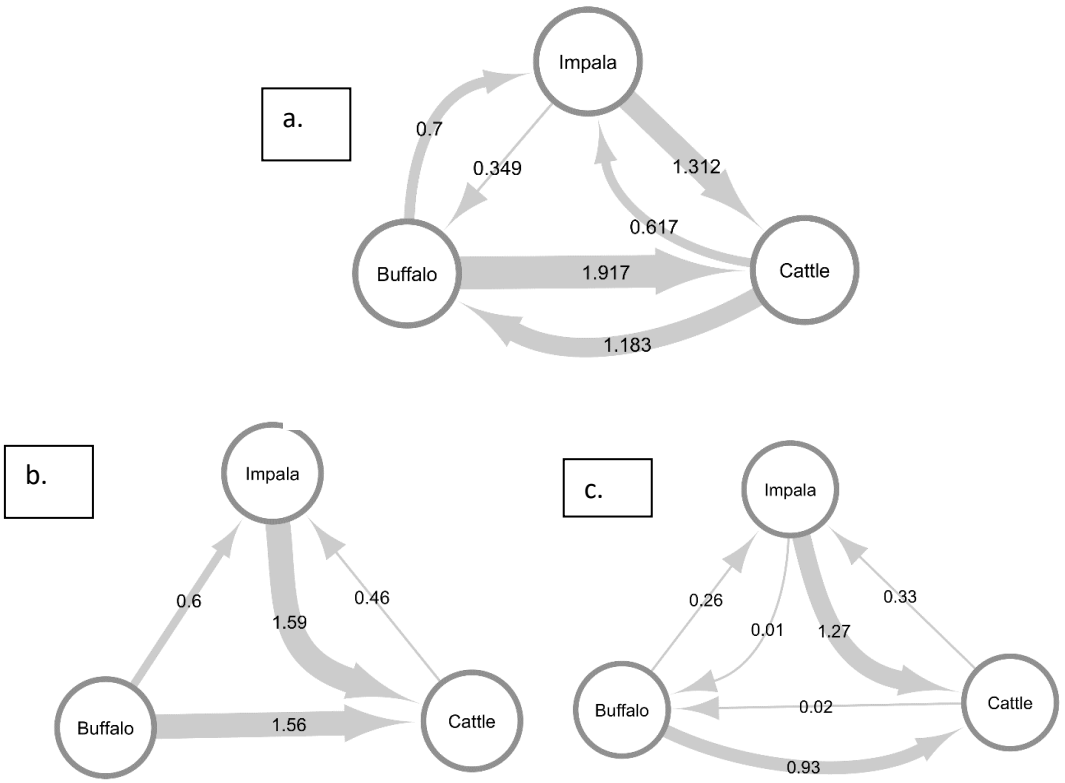


Figure 4-2: Estimated transmission rates between the three potential hosts for the SAT1 FMDV serotype. a. Using the “mugration” approach in BEAST. b. Using the BASTA approach in BEAST2. c. Using the MASCOT approach in BAST2.

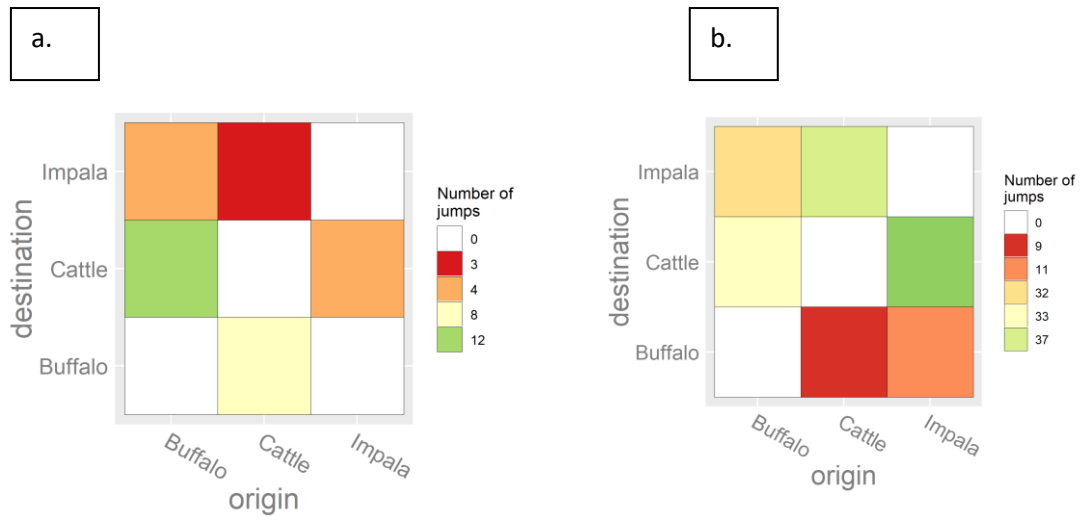


Figure 4-3: Heatmap showing the number of transitions between the sampled hosts for the SAT1 FMDV serotype obtained through a Markov jump analysis. The heatmaps are coloured according to the number of estimated transitions between hosts. a. Host transition using the "mugration" model b. Host transition using the BASTA model.

For the SAT1 serotype, I estimated a similar clock rate of around $2x10^{-3}$ substitutions/site/year for the three phylogenetic methods. I observed a tree height of 260 years for the “mugration” method, and a tree height closer to 200 years for the two structural coalescent approximation methods. Both BASTA and MASCOT models estimated that an important viral population present in buffaloes (736 ± 117 for MASCOT and 903± 187.6 for BASTA) and a medium viral population size present in cattle (7.35 ± 4.6 for BASTA and 4.19 ± 3.14 for MASCOT) had a role in the serotype transmission. The two models estimated that only a small viral population present in impalas was involved in the circulation of the disease (2 ± 1.15 for BASTA and 2.43 ± 1.45 for MASCOT) (see Supplementary table 8-44 and Supplementary figure 8-26).

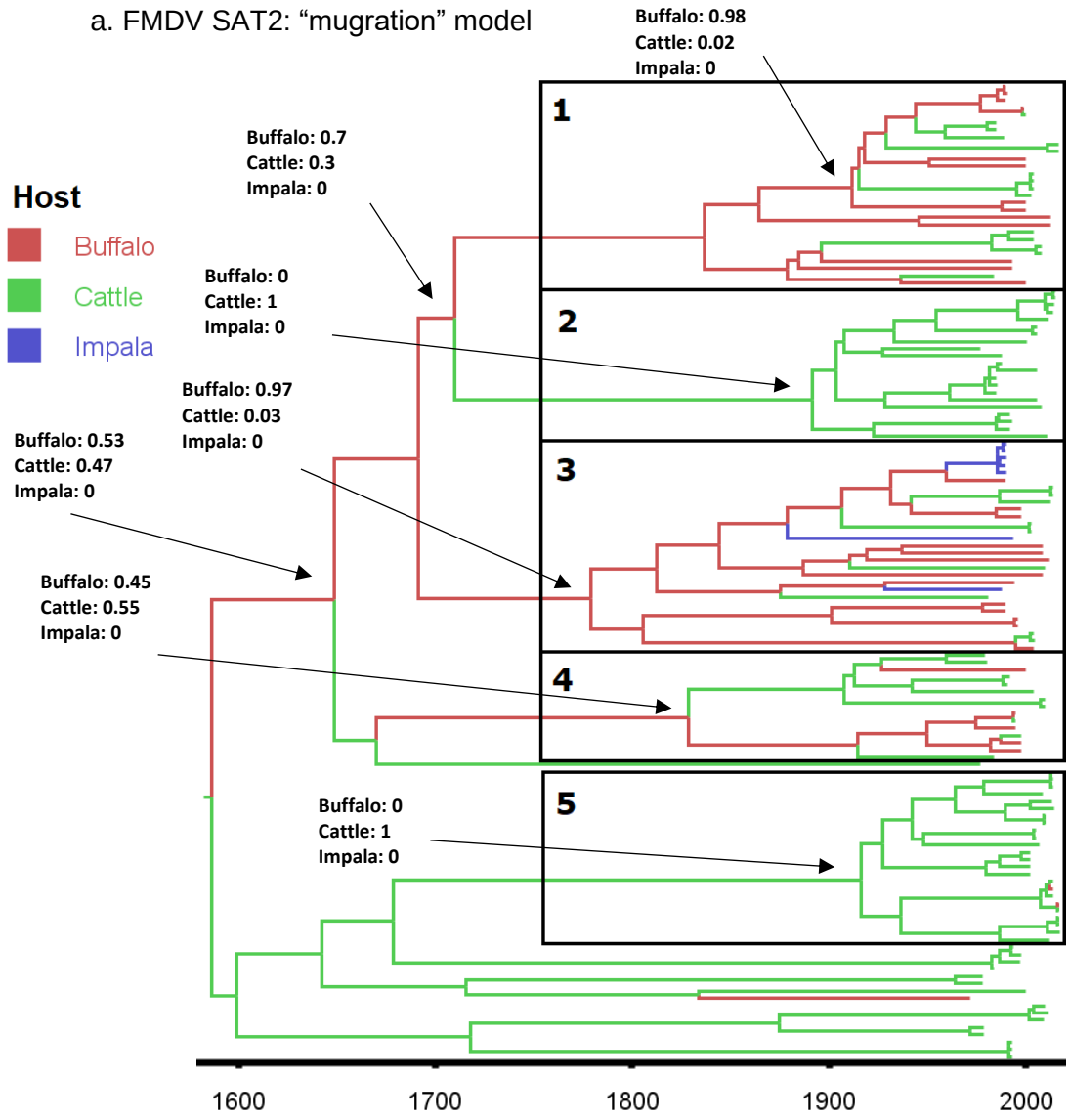
In all three methods, we can observe the same three main clades. With two out of the three clades of the reconstructed tree being almost entirely composed of cattle nodes (clades 1 and 3), the “mugration” approach estimated an important role for cattle in the transmission of the SAT1 virus (see fig. Figure 4-1a). With the “mugration” model, we can observe that in only a few occasions (mostly in clade 2), the impala population acted as an intermediate host between the buffalo and cattle populations. However, for both structural coalescent models, I observed that most of the reconstructed trees were either composed of buffalo nodes (for BASTA) or a mixture of buffalo and impala nodes (for MASCOT), with the cattle population being present only at the tips of the trees (see Figure 4-1b and Figure 4-1c). In both structural approaches, I observed multiple occasions where the impala population acted as the intermediate host between the buffalo and cattle populations.

The Markov jump analysis results and the estimated transmission rates between the different populations for FMDV serotype SAT1 emphasise the observed differences

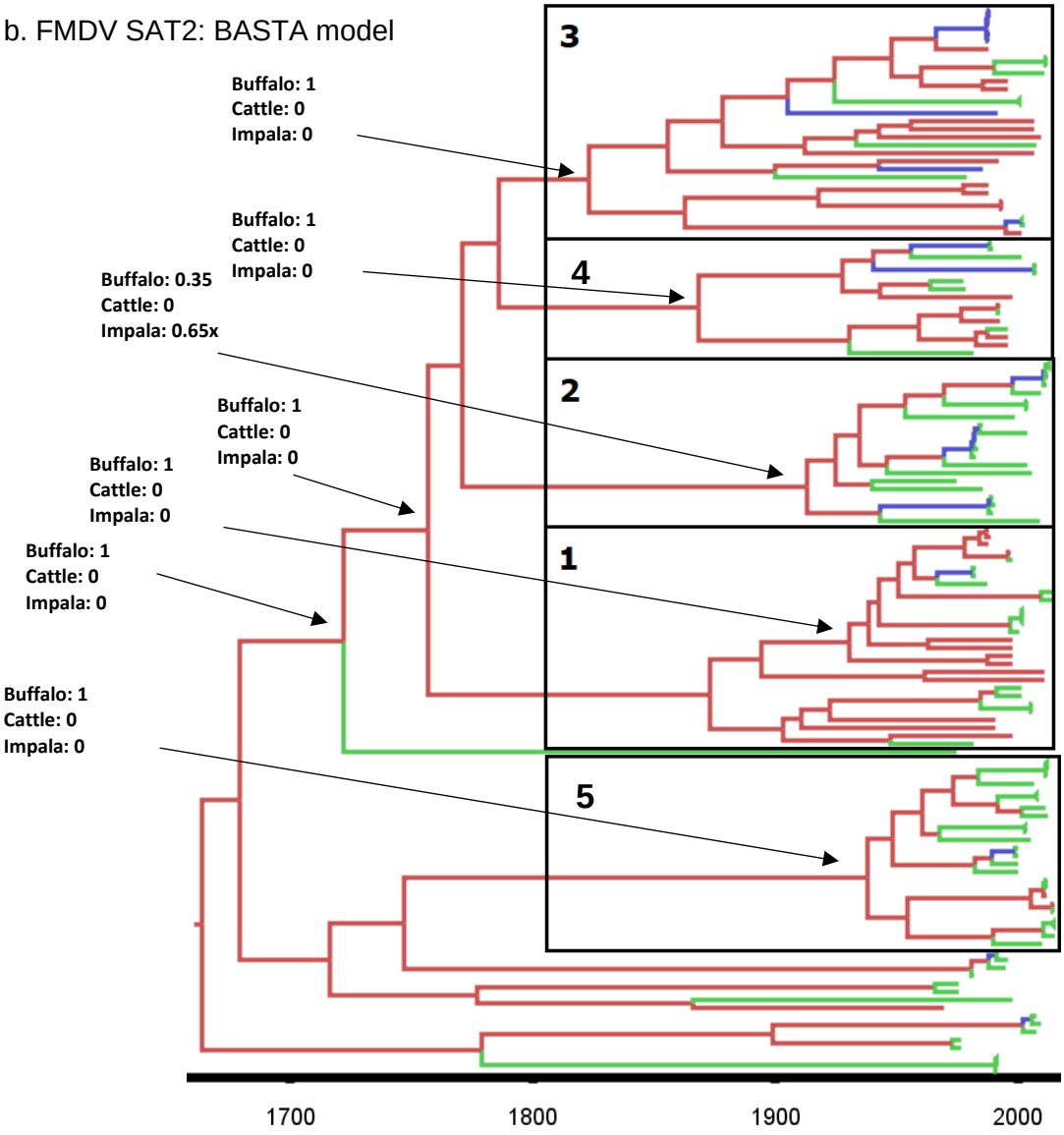
between the “mugration” model and the two structural coalescent approximation approaches.

For both structural coalescent models, I obtained similar transmission rates between the three populations (see Figure 4-2b and Figure 4-2c). I estimated low transmission rates from the impala and cattle populations toward the buffalo population (with a rate lower than 1^{-2} in both approaches). I observed high transmission rates from the impala and buffalo population toward the cattle population (1.59 ± 0.64 and 1.56 ± 0.42 for BASTA and 1.27 ± 0.39 and 0.93 ± 0.44 for MASCOT) and lower transmission rates from the buffalo and cattle populations to the impala population (0.6 ± 0.32 and 0.46 ± 0.43 for BASTA and 0.26 ± 0.18 and 0.33 ± 0.39 for MASTCOT). These results are quite different to those obtained with the “mugration” model (see Figure 4-2a). The main differences with the structural approaches being the high transmission rates estimated between the buffalo and cattle populations (respectively 1.9 ± 0.97 and 1.18 ± 0.69) and the transmission rate of 0.35 ± 0.35 from the impala to the buffalo populations.

Using the “mugration” model, I determined numerous transmission events between buffalo and cattle populations and lower numbers of transition events between the impala and cattle populations (see Figure 4-3a). In comparison, using BASTA, I observed few transmission events from the cattle to the buffalo populations but more transmission events between the impala and cattle populations (see Figure 4-3b). It is important to notice that it is difficult to perform a straight comparison between the Markov jump result of the “mugration” approach and the BASTA approach since the former allows state transitions over single branches leading to a higher jump count.


a. FMDV SAT2: “mugration” model
Host
Buffalo
Cattle
Impala
Buffalo: 0.98
Cattle: 0.02
Impala: 0
Buffalo: 0.7
Cattle: 0.3
Impala: 0
Buffalo: 0
Cattle: 1
Impala: 0
Buffalo: 0.97
Cattle: 0.03
Impala: 0
Buffalo: 0.53
Cattle: 0.47
Impala: 0
Buffalo: 0.45
Cattle: 0.55
Impala: 0
Buffalo: 0
Cattle: 1
Impala: 0
1
2
3
4
5
1600
1700
1800
1900
2000

b. FMDV SAT2: BASTA model



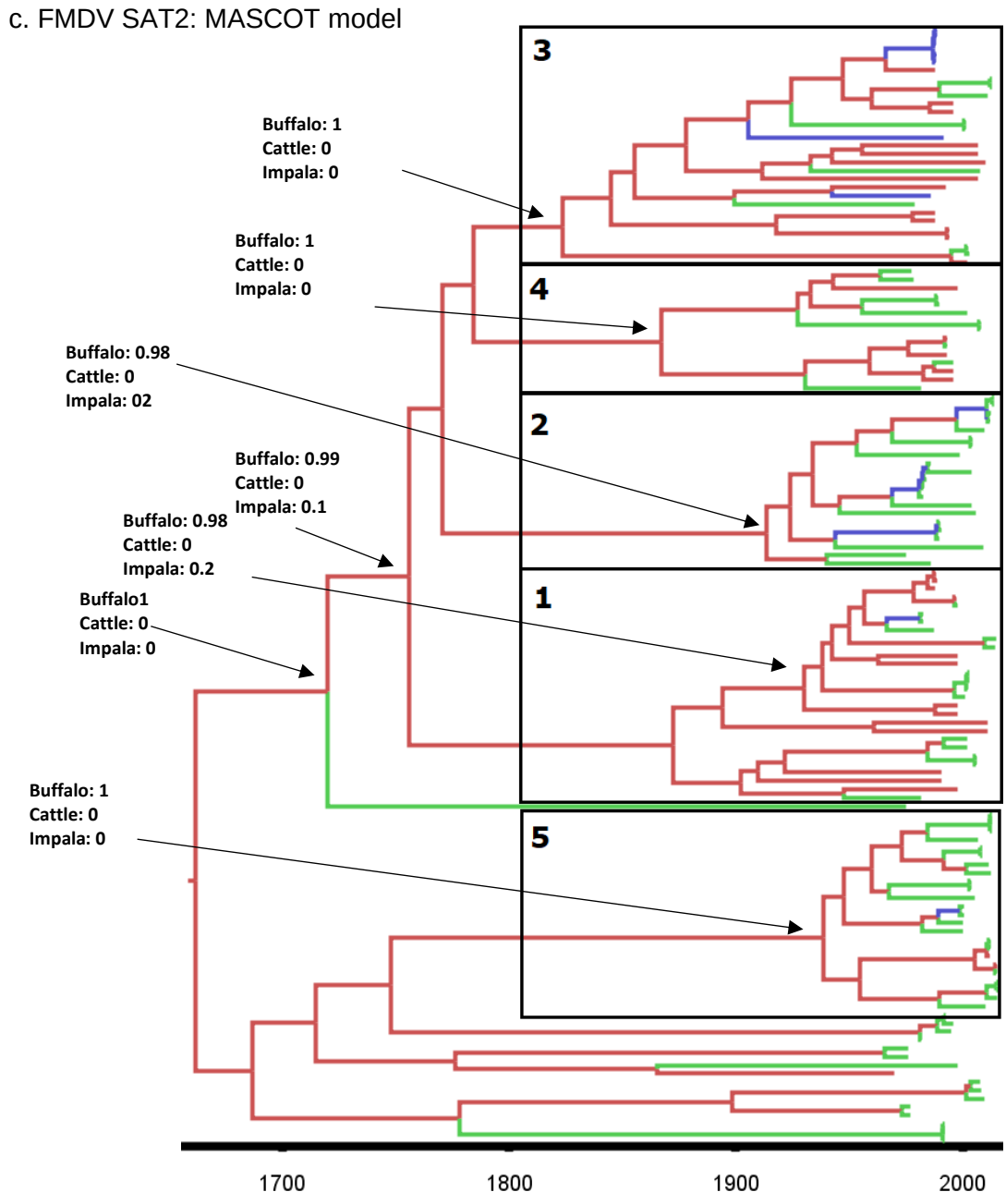


Figure 4-4: Bayesian MCC time scaled discrete phylogeographic tree for the serotype SAT2 using 135 VP1 sequences. a. Phylogenetic tree estimated using the “mugration” approach implemented in BEAST b. Phylogenetic tree estimated using the BASTA approach implemented in BEAST2. c. Phylogenetic tree estimated using the MASCOT approach implemented in BEAST2. The phylogeny branches are coloured according to their descendent nodes host with the key for colours shown on the upper left of the figure. The identified clades were isolated and numerated. Specific nodes of the trees were annotated with hosts posterior probabilities.

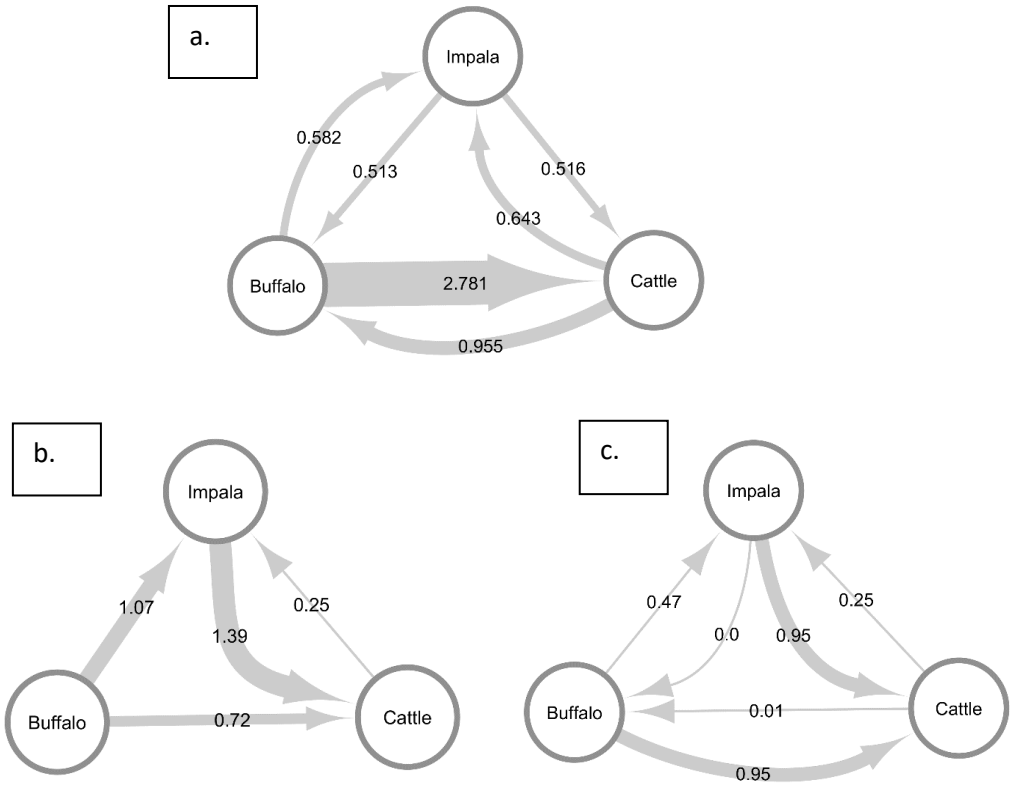


Figure 4-5: Estimated transmission rates between the three potential hosts for the SAT2 FMDV serotype. a. Using the “mugration” approach in BEAST. b. Using the BASTA approach in BEAST2. c. Using the MASCOT approach in BEAST2.

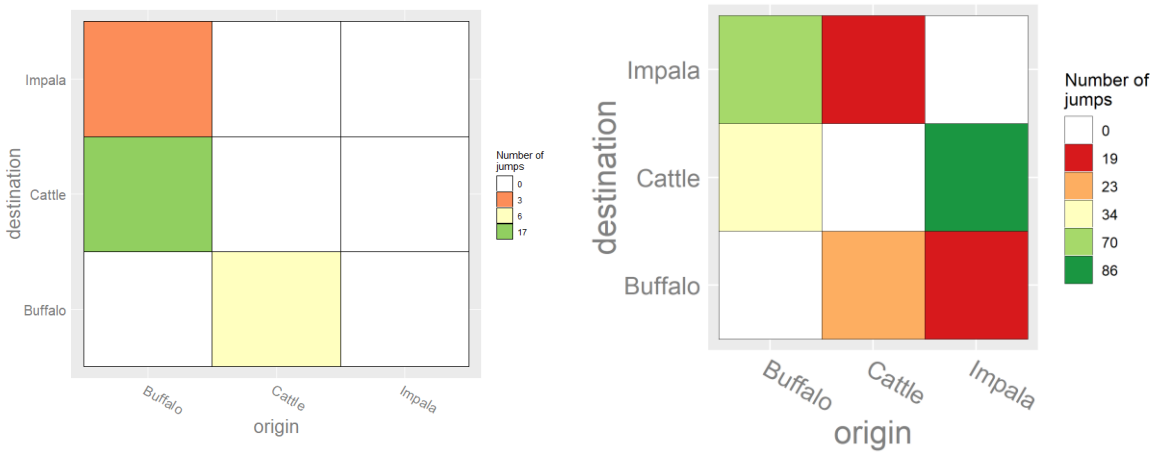


Figure 4-6: Heatmap showing the number of transitions between the sampled hosts for the SAT2 FMDV serotype. The heatmaps are coloured according to the number of estimated transitions between hosts. a. Host transition using the "mugration" model b. Host transition using the BASTA model.

When analysing the FMDV SAT2 serotype dataset, I estimated a similar clock rate of $1.1x10^{-3}$ substitutions/site/year between the three models. However, I estimated a longer tree height of 430 years for the “mugration” method than for the two structural coalescent approximation methods that had a tree height closer to 350 years. In both BASTA and MASCOT models, I estimated a similar buffalo and impala population size involved in the SAT2 serotype spread (around 1200 for the buffalo and the impala). However, the BASTA model estimated a slightly larger cattle population (3.9 ± 1.89) than MASCOT (1.8 ± 1) (see Supplementary table 8-44 and Supplementary table 8-45).

With all three methods, I isolated the same main five clades. Because the “mugration” model estimated that two out of five clades composing the phylogenetic tree of SAT2 were entirely composed of cattle nodes (clades two and five), I estimated an important role of the cattle population in the transmission of the virus (see Figure 4-4a). However, for this serotype, we cannot observe transitions from the impala to the cattle populations. In comparison, with most of the tree composed of buffalo nodes, the BASTA and MASCOT models estimate a less important role for the cattle population in the transmission of FMDV SAT2 serotype (see Figure 4-4b and Figure 4-4c.). Additionally, in these two models, I estimated multiple transmission events of FMDV SAT2 serotypes from buffalo to impala and then to the cattle population. Again, as with the FMDV SAT1 serotype, the estimated number of transitions and transmission rates between the different populations estimated for the SAT2 serotype stresses the differences between the “mugration” model and both structural coalescent approximation approaches.

In both the BASTA and MASCOT models, I estimated low transmission rates (with a rate lower than 1^{-2} in both approaches) from the impala and cattle populations toward

the buffalo population (see Figure 4-5b and Figure 4-5c). In both approaches, I estimated higher rates of transmission from buffaloes to impalas (1 ± 0.44 in BASTA and in 0.466 ± 0.28 in MASCOT) than from cattle to impalas (0.24 ± 0.31 in BASTA and 0.25 ± 0.33 in MASCOT). By looking at the transmission rate, the main difference between the two structural coalescent models is that the BASTA model estimated a higher transmission rate from impalas to cattle than from buffaloes to cattle (1.38 ± 0.77 and 0.7 ± 0.46), while the MASCOT model estimated similar transmission rates (0.95 ± 0.66 and 0.96 ± 0.43). In comparison to the structural coalescent approximation approaches, most of the transmission rates obtained with the "mugration" model were between 0.5 and 1, except for the transmission rate from the buffalo to the cattle population which is equal to 2.8 ± 1.44 (see Figure 4-5a).

The "mugration" model estimated that multiple transmission events occurred between the buffalo and cattle populations and that no transmission events occurred between the impala and cattle populations (see Figure 4-6a). In comparison, the BASTA approach estimated multiple transmission events from the impala and buffalo populations toward the cattle population and a small number of transmissions with the cattle population as origin (see Figure 4-6b).

4.5 DISCUSSION

In this study, I applied three recently developed methods of phylogenetic reconstruction to estimate the transmission and circulation of the VP1 gene sequence between cattle, buffalo and impala populations for two FMDV serotypes (SAT1 and SAT2) in Africa. Two of these methods, the BASTA and MASCOT models, consider the existence of different host populations to reconstruct the evolution of the pathogen. The last approach, the "mugration" model, does not consider the existence

of multiple host populations, making it more prone to statistical errors caused by sampling bias. Considering that both my BASTA and MASCOT models produced similar estimates, it seems that the host structure of the populations plays an important role in the outcome of the analyses. However, the differences observed between the “mugration” and the BASTA/MASCOT models were expected, given the incomplete nature of the sampling for the two serotypes[25].

With the “mugration” model, I estimated an important role of the buffalo and cattle populations in the maintenance of both FMDV SAT1 and SAT2 viruses in Africa, with high value of transmission rate between the two populations, and where the impala population acts as a spill over host for the virus. With this strong connection between cattle and buffalo populations and the apparent absence of a role played by the impala population in the circulation of the viruses, my observations based on the ”mugration” model are consistent with those made in previous studies[271,335].

However, with the BASTA and MASCOT models, I observed a different interaction between the three hosts in the circulation and maintenance of the viruses. The observation of a proportionally higher number of transmission events as well as higher transmission rates between buffalo and impala populations in the BASTA model compared to the “mugration” approach is consistent with the real life observations of regular transmission events between buffaloes and impalas[265,276]. Moreover, with the observed lower transmission rates toward the buffalo population and large population sizes in both structural approaches, my observations enforce the general consensus that buffaloes are the maintenance and original host of the SAT serotypes[267]. Additionally, although the population size in impalas was estimated as small, the large transmission rate from the impala population toward the cattle population estimated by the structural approaches suggest an at least as important role played by the

impala population in the viral transmission toward the cattle population in Southern Africa. Overall, in both structural approaches, I observed that the impala population plays the role of an intermediate host between the buffalo and cattle populations, with the cattle as final host and the buffalo as original host of the viruses.

Overall, my results point out a more important role of the impala population than what was previously thought with buffaloes as being the original host of the virus. However, since most of the analysed samples originate from Southern Africa, my observations could be the result of locally observed relationships between the three populations, and might not be generalised to sub-Saharan Africa. This situation emphasises the need of a more systematic sampling of impala and buffalo populations. Such approach would enable us to detect currently unrecorded/sub-clinical buffalo/impala outbreaks and confirm the results of this analysis more globally. Similar to what the "mugration" approach suggests, our current knowledge sees impalas as intermediate hosts whereas the structural coalescent results suggest a simultaneous buffalo to cattle and buffalo to impala transmission, with a possible final impala to cattle transmission. My study is the first phylogenetic analysis using a structural coalescent approach to reconstruct the circulation and maintenance of the SAT serotypes between multiple hosts. It is also the first time that the importance of the impala population in the transmission and circulation of the virus has been quantified. The fact that the SAT serotypes are mainly maintained within wild buffalo populations does not go against the observation that most of the recent spread of the viruses in Africa seems to be driven by livestock movement [220,270,337] . Moreover, this observation is in concordance with the theory that currently circulating lineage of SAT serotypes re-emerged from a small number of wild buffaloes that survived the African rinderpest epidemic of 1975[338].

Although the use of more complex phylogenetic models allows us to gain more insight into the epidemiology and evolution of FMDVs, greater sampling effort is needed to obtain more diversity in the types of host sampled and the temporal and spatial distributions of isolates in order to observe and explain these results epidemiologically. The use of analytical methods is important in uncovering the drivers of disease maintenance and spread, and thus will help in developing a modern approach to FMD control where different hosts could be targeted and controlled in different regions.

5 EPITREE-SIM: AN APPLICATION FOR EPIDEMIC SIMULATION AND PHYLOGENETIC TREE RECONSTRUCTION

5.1 ABSTRACT

Viruses and bacteria accumulate mutations over time and Bayesian phylogenetic approaches are nowadays a standard tool to study the spread and circulation of those fast-evolving pathogens. Categorising those pathogen sequences into discrete demes or populations enables the transition rates between those multiple discrete states to be estimated and used to infer transmission patterns. Nowadays, it is increasingly common to find large datasets composed of thousands of genetic viral or bacterial samples to analyse. However, currently available Bayesian phylogeographic methods remain computationally demanding when analysing such large datasets composed of multiple discrete demes. Furthermore, such methods can suffer from statistical approximations when analysing biased datasets. Here, I present a new software - Epitree-sim - that allows the fast estimation of phylogenetic tree and transmission patterns using a fast-dating algorithm and repeated subsampling of the sequence datasets. To illustrate this software, I compare the results obtained using an avian influenza and foot and mouth disease datasets using Epitree-sim, the "mugration" model available in BEAST and a structural coalescent model approximation, BASTA, available in BEAST2.

5.2 INTRODUCTION

Phylogenetic techniques are now a standard tool in the study of the spatial and demographic history of organisms[91]. Nowadays, some of the most commonly used techniques to perform such analyses are anchored in a Bayesian framework[300,323,339]. When coupled with epidemiological data, phylogenetic analyses can determine the circulation of fast-evolving pathogens within multiple demes or discrete traits such as a location or host. A few methods can combine genetic and epidemiological data to perform those phylodynamic analysis [340]. Amongst them, the most popular approach is the widely used “mugration” approach, which is embedded in the popular BEAST software[17,323]. This method has already been used on numerous occasions to study the evolution and circulation of multiple organisms and pathogens[341–344]. In this approach, the evolution of the studied trait is modelled the same way as a substitution event[17]. Despite its convenience of use and computational speed, the “mugration” approach makes multiple wrong statistical assumptions when applied on the migration of lineages between multiple geographically defined demes or populations[25]. For example, with the “mugration” approach, subpopulations can be driven extinct and be resurrected. The model also assumes that the sampling intensity is proportional to population size, which might lead to biased migration rate estimates when performed on a biased sample dataset. Finally, ignoring the population structure when estimating the phylogenetic tree can lead to bias or loss of power of the model [25,110,340].

An alternative to the “mugration” approach is the use of a structural coalescent model approximation[127]. Such approach is able to take into account the effects that migrations between different populations have on the tree typology. Therefore, such approach addresses some of the problems of the “mugration” approach since it does not suffer from the same statistical approximation[25]. Over the last few years, multiple

structural coalescent model approximations have been released, each with different assumptions[340]. Although such approaches are efficient to reconstruct the transmission between a few numbers of discrete states, it might be challenging to run those analyses when analysing the transition to large datasets with numerous possible states.

Consequently, with the increasing size of pathogen sequence datasets currently available, it is becoming increasingly difficult to apply these computationally-intensive Bayesian-based methods to real world data[60]. Therefore, there is currently a need for new approaches that would allow the estimation of phylodynamic trees in an efficient and statistically appropriate way.

Here, I present Epitree-sim, a software that combines multiple computational and statistical methods to study the circulation and diffusion of large genetic datasets amongst multiple demes. Epitree-sim is available as a standalone software to (i) generate genetic datasets based on the simulation of epidemics in user-defined populations; (ii) quickly infer the phylogenetic tree between the multiple discrete states composing the simulated or imported dataset; (iii) estimate the significant non-zero transition rates between those discrete states by performing a BBSVS analysis[17].

5.3 FEATURES

Epitree-sim is composed of two distinct parts: an epidemics simulation part and a phylogenetic tree inference part. Each part can be performed independently of the other, and at the end of each part, the user can download the generated data and varied descriptive statistics. Additionally, for each part of the process, the user can find multiple help buttons describing each of the possible actions. The two parts of Epitree-sim are introduced in this section.

I. **Epidemic simulation:** The software is based on the epinet package available within R to simulate epidemics[345]. Using this approach, the spread of the pathogen is modelled as a stochastic compartmental model (a SEIR model where each potential host is one of the four following categories: susceptible, exposed, infectious and recovered) within a static undirected contact network, constituted of nodes representing individual hosts, and edges representing the contacts between two individuals[345]. The whole population of potential hosts can be composed of multiple subpopulations. In this case, each subpopulation is modelled as a random static undirected contact network and all the subpopulations are connected through undirected random contacts. Individuals within a subpopulation have a higher probability of contact among each other than with individuals belonging to another subpopulation (see Figure 5-1 for a schematic representation of the approach used to represent the structure of subpopulations in the epidemic simulation part of Epitree-sim). Because the epidemic is modelled as a stochastic process, the existence of a contact between two individuals **a** and **b** does not necessary mean that if **a** gets infected, **b** would be automatically infected too. The epidemic process depends on few parameters

that the Epitree-sim user can define. The different parameters are the type of contact network and the population size where the epidemic occurs, the waiting time for a transmission across an edge, modelled as an exponential random variable with mean 1/β, the time spent in the exposed and infectious state, modelled as gamma random variable, with parameters (θ_E, k_E) or (θ_I, k_I), a mean $k\theta$ and variance $k\theta^2$ [3,4,5]. The output of this approach is a transmission tree, subgraph of the full contact network, that contains all the infected nodes and their timing of infection and recovery over the whole of the epidemic. From the simulated transmission tree, a time scaled phylogenetic tree is created, and from this, the user can generate the corresponding genetic sequences using a Monte Carlo simulation approach implemented within πBUSS[3,4,6], while specifying the substitution model and evolutionary rates required when simulating the sequences. Using this approach, the user can simulate multiple types of epidemic while generating the corresponding phylogenetic tree and corresponding sequences that can be subsequently used in phylogenetic analysis.

II. **Phylogenetic tree inference:** The user can upload their own genetic dataset and corresponding trait table to set the different demes to analyse. The user can either use real world data or data generated through the epidemic simulation part of Epitree-sim. To speed up the inference of phylogenetic tree and transmission between multiple discrete demes, the software uses a fast dating algorithm R package, treedater[60]. This method is based on recent advances in maximum likelihood and least-squares phylogenetic and molecular clock dating methods to fit a molecular clock to a phylogenetic tree with associated data on sampling times. The user can provide an outgroup sequence that will be used to root the phylogenetic tree, or can let treedater handle the rooting procedure. This approach allows a much faster estimation of the molecular clock of a particular

phylogeny than current Bayesian approaches[60]. At its core, Epitree-sim applies a procedure similar to a random forest analysis[347]: the same model is applied over multiple different subsets of the same dataset in order to boost the performance of the final model. First, my approach uses a two-tier subsample of the whole dataset to create a collection of phylogenetic trees (an initial subsampling followed by a subsampling aiming to balance the samples in the sub-populations more evenly). Second, Epitree-sim uses this collection as a pseudo-empirical tree distribution in discrete state analysis within BEAST using the "mugration" approach. Third, it performs a Bayesian Stochastic Search Variable Selection (BSSVS) analysis to estimate under the form of a Bayes factor (BF) the significant non-zero transition rates between the discrete states[17]. Once the BSSVS analysis is performed, the user can visualise within the software the reconstructed phylogeny and the significant transmission rates between each deme, as well as download all the estimated trees and log files. See Figure 5-2 for a complete explanation of the procedure to perform a discrete phylogenetic analysis, including an explanation of the subsampling procedure applied.

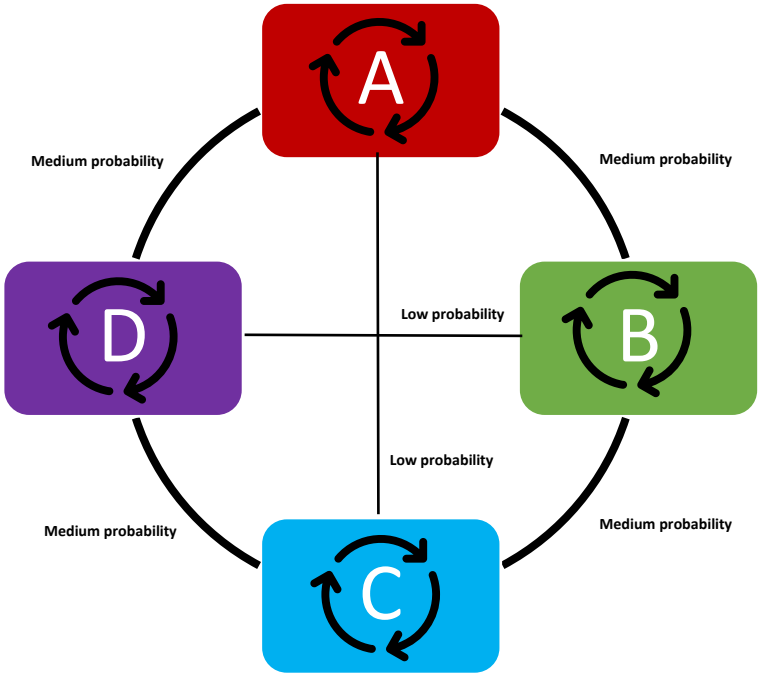


Figure 5-1: Representation of a four-subpopulation relationship and probability of infection within a virus host population. Within each of the subpopulations, the virus can be transmitted with a high probability, amongst adjacent subpopulation the virus can be transmitted with a medium. Finally, amongst non-adjacent subpopulations, the virus can be transmitted with a low probability. The black lines represent the undirected probability of transmission of the virus between the subpopulations.

5.3. Features

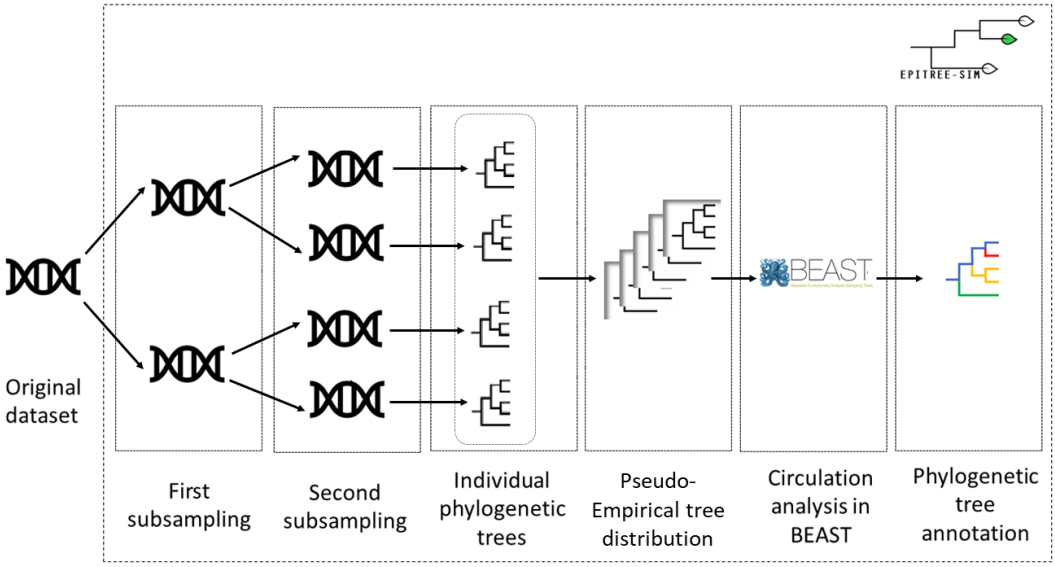


Figure 5-2: Illustration of the use of Epitree-sim to perform a discrete analysis between multiple discrete states. First, Epitree-sim performs a two-tier subsampling procedure using the available sequence information (here, the host type and location). During the first subsampling step, the original dataset is split in multiple subsets based on the combination of host/location of the sequences, the maximum number of samples allowed per set for each combination and the final number of datasets this step has to produce (in this case two). During the second subsampling step, another subsampling procedure will be performed on the newly generated sequence subset based on the location or host of each one of the sequences, the maximum number of sequences allowed per set per category and the final number of datasets this step has to produce per dataset (in this case two). For each one of the new sequence subsets, Epitree-sim will then generate a rooted timescaled NJ tree using treedater[60] and, if available, a provided outgroup sequence. The set of tree generated will then be used as a pseudo-empirical distribution in a discrete trait analysis using the “mugration” approach[17] with a BSSVS analysis set up to estimate the significant non-zero transition rates between the sampled host and therefore estimate potential routes of transmissions.

5.4 APPLICATION AND COMPARISON WITH OTHER PHYLOGENETIC APPROACHES

The aim of this section is to compare the results of my subsampling method with the "mugration" approach available in BEAST[348], and the BASTA[25] method available in BEAST2[349] if possible. The new approach is first compared to the "mugration" approach and BASTA on multiple simulated epidemics in order to apply those approaches in a wide range of situations and compare the different outcomes with the true epidemic. Second, those three approaches were applied on two real world datasets in order to gain more insights on the quality of the Epitree-sim results.

5.4.1 Material and methods

5.4.1.1 Comparison on simulated datasets

To illustrate the application of the method, I first performed the Epitree-sim and "mugration" approaches on six sequence datasets produced by the epidemic simulation feature of Epitree-sim. Because it was difficult to set up the analysis for those large datasets, the structured coalescent model BASTA was performed on the first dataset only. Each simulated epidemic was generated using different epidemic parameters with either 3 or 4 discrete populations to simulate various epidemiological situations.

For an epidemic involving four populations A, B, C and D, the virus can be transmitted with a high probability between individuals belonging to the same population, with a medium probability between individuals belonging to populations close to each other and with a low probability between individuals belonging to the respective non-adjacent subpopulations (for a representation of this see Figure 5-1). For an epidemic

involving 3 populations, there is a medium probability between individuals belonging to different populations to infect each other.

Once the epidemic and the corresponding transmission networks had been generated, I simulated the sequences that each epidemic would generate on 2000 nucleotide long virus with a strict molecular clock of $3e^{-3}$ mutations/site/year following an HKY substitution model. For each set of sequences, the Epitree-sim phylogenetic approach was performed using a different threshold value for the subsampling procedure to mitigate any 'sampling' bias present in the original dataset. While generating the pseudo-empirical tree distribution, I rooted each phylogenetic tree using treedater since I did not provide an outgroup.

For the "mugration" analysis, I selected a constant population size, constant molecular clock model and a HKY model with no site variation as substitution model[350]. An empirical tree distribution of 1000 trees was generated for each set of sequences.

For the BASTA analysis, on the first dataset, I used a constant molecular clock model and a HKY model with no site variation as substitution model[350].

For both the Epitree-sim and "mugration" approaches, a Bayesian Stochastic Search Variable Selection (BSSVS) analysis was set up to identify the smallest set of transition rates that could summarise the epidemiological connectivity between the locations, hosts or strains[17]. From those BSSVS analyses, Bayes factor (BF) values were calculated to determine the significant non-zero transition rates (BF=10) between the discrete locations or hosts.

More details for each dataset can be found on Table 5-1. Performing those phylogenetic approaches on multiple simulated epidemics obtained using different

epidemiological parameters will allow me to compare their output against the known 'true' transmission trees.

Table 5-1: Table describing the number of samples per subpopulations, the epidemiological parameters used to generate the different transmission networks resulting in the different sets of simulated sequences, the overall duration of each one of the simulated epidemics and the maximum number of samples per population allowed in the Epitree-sim phylogenetical analyses. In Epitree-sim, for each set a pseudo-empirical tree distribution of 1000 trees was generated to perform the discrete trait analysis. For the "mugration" analysis, an empirical distribution of 1000 trees was generated. For each epidemic, the waiting time for a transmission across an edge is modelled as an exponential random variable with a mean 1/β, the time spent in the exposed and infectious state is modelled as gamma random variable with parameters (θ_E, k_E) or (θ_I, k_I) and a mean equal to kθ and variance to $k\theta^2$ [345].

	Samples per populations (A, B, C, D)	**β**	**k_e**	**θ**	**k_I**	**Epidemic duration**	**Max samples allowed in Epitree-sim**
Set 1	50, 50, 20, 20	0.3	0.1,	1	0.1	6.19	20
Set 2	198, 200, 79, 30	0.3	0.1	0.7	0.1	2.53	50
Set 3	146, 192, 146, 40	0.5	0.1	0.4	0.1	2.27	50
Set 4	90, 87, 24, 15	0.3	0.1	1	0.1	5	30
Set 5	136, 185, 71, 26	0.3	0.1	0.5	0.1	3.37	50
Set 6	22, 138, 51	1	0.1	0.4	0.1	2.6	35

5.4.1.2 Comparison using Eurasian avian influenza sequences

Secondly, to compare the Epitree-sim "mugration", I performed the two approaches on a real dataset composed of 282 PB2 avian influenza sequences sampled in Eurasia between 2012 and 2017 and grouped in five discrete bird populations (see Table 5-2) and nine discrete regions spread between Europe and Asia (see Table 5-3).

For the host analysis in Epitree-sim, I allowed a maximum of 2 sequences per joint trait (host/location and year) and then 60 sequences per sampled location to generate the empirical trees. For the location analysis, I allowed a maximum of 3 sequences per joint trait (host/location and year) and 30 sequences per sampled location to create the empirical tree distribution (see Table 5-2 and Table 5-3 for the procedure applied and the resulting sequences distribution).

For the “mugration” analysis, I selected a Bayesian skygrid population model and a relaxed uncorrelated exponential molecular clock model to model the evolution of the virus and a general-time-reversible (GTR) model with site-to-site rate variations between two categories as nucleotide substitution model[79,100,324]. Using those parameters, I generated an empirical tree distribution by combining three independent runs of one billion steps each to ultimately perform a BSSVS analysis to estimate significant non-zero transition rates between the different hosts and locations[17].

Table 5-2: Sequences distribution for the avian influenza dataset between the 5 sampled host populations before the subsampling procedure performed within Epitree-sim. The subsampling procedure allowed to generate a pseudo empirical tree distribution to perform a BSSVS analysis on the discrete location to estimate the non-zero transition rate between the hosts. The “mugration” analysis was performed on the unsampled dataset.

Host	**Number of sequences before the subsampling procedure**	**Number of sequences after the subsampling procedure**	**Date range**
Domestic Anseriformes	101	60	2001.64-2017.29
Domestic Galliform	49	40	2002-2017.6
Wild Anseriformes	106	60	2001-2017.64
Wild Charadriiform	15	14	2003.29-2016.9
Wild other	11	11	2002-2017.2

Table 5-3: Sequence distribution for the avian influenza dataset between the 9 sampled locations before and after the subsampling procedure performed with Epitree-sim. The subsampling procedure allowed to generate a pseudo-empirical tree distribution to perform a BSSVS analysis on the discrete location to estimate the non-zero transition rate between the locations. The "mugration" analysis was performed on the unsampled dataset.

Location	Number of sequences before the subsampling procedure	Number of sequences after the subsampling procedure	Date range
Central Asia	20	20	2007.6-2017.3
Eastern Asia	32	30	2004-2017.1
North-Central Asia	20	20	2001.6-2016.4
South Asia	26	24	2008.97-2016.9
East China	48	30	2002-2016
Eastern Europe	36	29	2005.7-2017.3
Northern Europe	18	18	2003.8-2016.95
Southern Europe	29	29	2001-2017.6
Western Europe	53	30	2002-2017

All the "mugration" and unspecified evolution analyses were performed within BEAST 1.8 using the BEAGLE library[323]. For all the analyses, I used TreeAnnotator to summarise maximum clade credibility (MCC) trees and FigTree version 1.4.1 to visualise the annotated trees[299,300].

5.4.1.3 Comparison using FMDV SAT1 sequences

Thirdly, to compare the Epitree-sim approach, "mugration" approach and BASTA approach, I performed the three approaches on a dataset composed of 135 VP1 FMDV SAT1 sequences sampled in Africa between 1961 and 2015 between three potential hosts: impalas, buffaloes and cattle (see Table 5-4).

Table 5-4: Sequences distribution between the host and origin of the SAT1 sequences analysed.

	Buffalo	Cattle	Impala	Total per country
Kenya-Uganda	2	27	1	30
South Africa	21	14	10	45
Southern Africa	3	11	2	16
Tanzania	3	9	0	12
Zimbabwe	7	7	0	14
Total per host	36	68	13	117

For the 'Epitree-sim' analysis, I generated an empirical tree distribution of a thousand trees by allowing a maximum of 3 sequences per joint trait (host/location and year) and 30 sequences per sampled host (see Table 5-5).

Table 5-5: Sequence distribution for the foot and mouth disease dataset between the 3 sampled host populations following the subsampling procedure performed within Epitree-sim.

Host	Number of sequences
Impala	13
Buffalo	36
Cattle	40

For the "mugration" approach, I chose a Hasegawa-Kishono-Yano (HKY) nucleotide substitution model with a constant clock model, and a Bayesian skygrid population model were chosen to model the evolution the virus[79,100].

For the BASTA analysis, I chose a Hasegawa-Kishono-Yano (HKY) nucleotide substitution model with a constant population. To obtain the empirical tree distribution used in the "mugration" approach, I combined two runs of 40 million steps each. For the BASTA analysis, I combined three converging runs of 50 million steps (BASTA).

For both the subsampling procedure and the "mugration" approach, I performed a BSSVS analysis to estimate the well-supported rates of transition between the different hosts. I was not able to perform a BSSVS analysis using BASTA due to model limitations.

5.4.2 Results

In this section, I compare the output obtained by performing the phylogenetic model found in Epitree-sim, the "mugration"[17] approach found in BEAST[348] or the BASTA approach found in BEAST2 when estimating the spread of various simulated epidemics or the transmission of avian influenza amongst Eurasian birds.

5.4.2.1 Simulated datasets

I am presenting in detail the results for the simulated set 1, 2 and 3 since they englobe most possible situations and observations that could be drawn when comparing the two phylogenetic approaches. A quick description of other results can be found at the end of this section. All the generated phylogenetic trees in higher resolution (see Supplementary figure 8-28 to Supplementary figure 8-45) as well as transmission rates (see Supplementary table 8-46 to Supplementary table 8-57) and BSSVS (see Supplementary table 8-58 to Supplementary table 8-63) output can be found in the supplementary material.

Simulated epidemic 1 is the result of a virus which is transmitted amongst the different populations with a low transmission rate for a long time, simulated epidemic 2 is the result of a virus which is easily transmitted amongst the different populations present, starting in a small population, and simulated epidemic 3 is the result of a virus which is easily transmitted amongst the different populations present, starting in a large population.

5.4.2.1.1 Simulation 1

For the simulated epidemic 1, the Epitree-sim approach estimated the closest approximation of the true mean mutation rate of $3e^{-3}$ mutations/site/year with an estimated 3e-3±9e-5 mutations per site per year (see Table 5-6). The “mugration” approach underestimated the mutation rate by estimating a value close to $2.9e^{-3}$ mutations/site/year. In contrast, the “mugration” approach estimated a TMRCA closer to the true value with an estimation of 6.36±0.18 years compared to 5.84±0.4 for the Epitree-sim approach. The results from the BASTA approach were also similar to

those from the “mugration” approach, with a very similar mean mutation rate ($2.87e^{-3}$±$1.^{7e-4}$) and TMRCA (6.4±0.18 years).

Table 5-6: Evolutionary parameters and standard deviation comparison between the Epitree-sim and “mugration” phylogenetic approaches for the simulated epidemic 1.

	True epidemic	**Subsampling approach**	**“mugration” approach**	**BASTA**
Mean mutation rate (mutations/site/year)	$3e^{-3}$	$3e^{-3}$±$9e^{-5}$	$2.88e^{-3}$±2e-4	$2.87e^{-3}$±$1.7e^{-4}$
TMRCA (years)	6.19	5.83±0.4	6.36±0.18	6.4±0.18

When looking at the estimated phylogenetic trees (see Figure 5-3), my subsampling approach was the only model that estimated properly the population D as being the original source of the epidemic. Although my subsampling approach estimated a more important role of the population D in the circulation of the virus than what really happened (see Figure 5-3.1. and Figure 5-3.2), it was the only method which properly estimated the importance of both populations C and D in the circulation of the virus (see Figure 5-3.3 and Figure 5-3.4). Interestingly, both the Epitree-sim approaches and the BASTA approaches properly estimated the role of population D close to the root of the tree, while the “mugration” approach wrongly estimated that only the population B played a role at the start of the epidemic. Overall, the “mugration” approach overestimated the importance of the population A and B in the circulation of the disease. Interestingly, those are the two populations with the highest number of samples in the dataset.

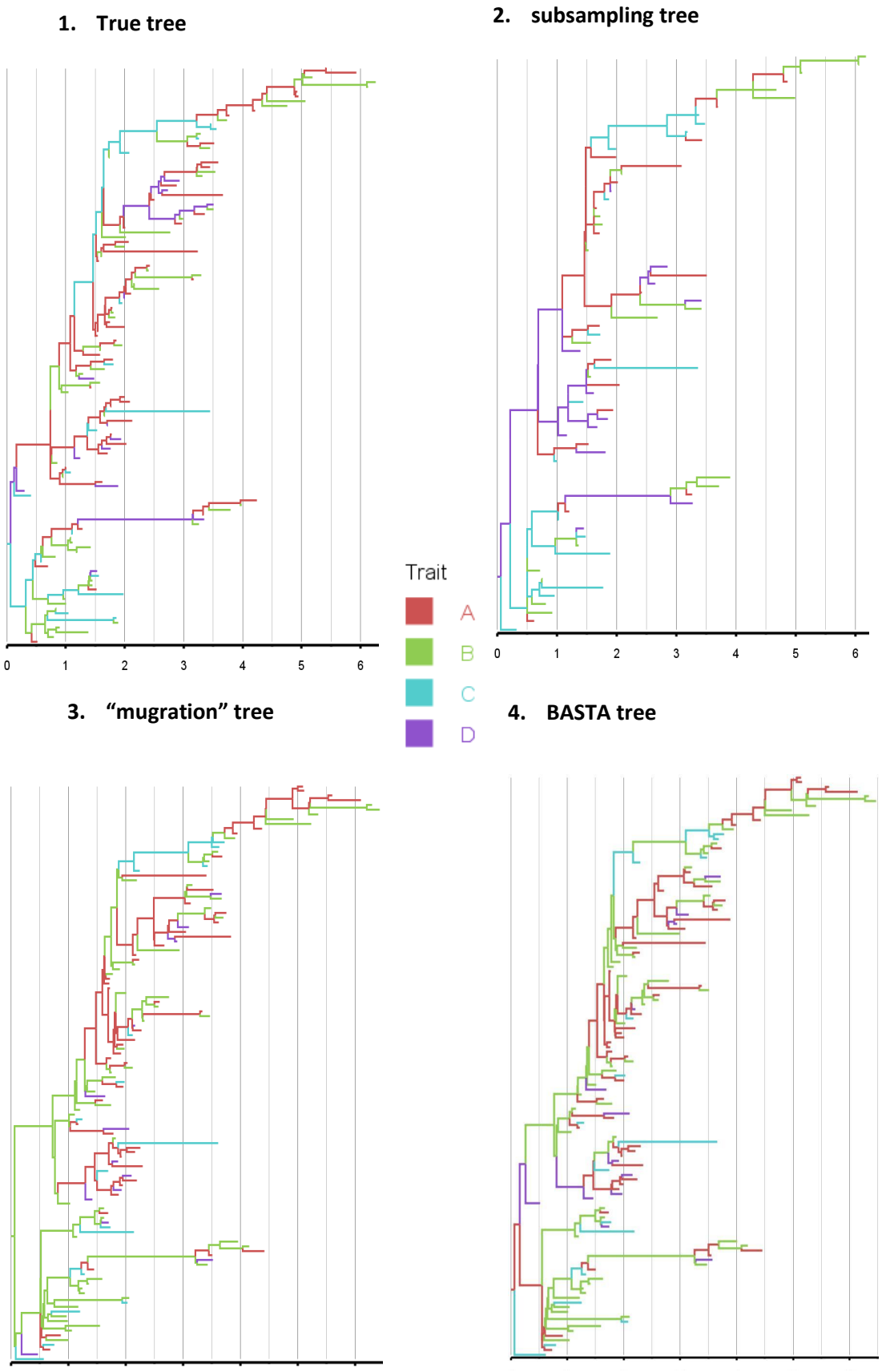


Figure 5-3: Comparison between the different trees obtained for the set of sequence number 1. 1) The true transmission tree results from the simulated epidemic between the different populations in Epitree-sim. 2) Tree obtained through the discrete phylogenetic procedure found in Epitree-sim. 3) Tree obtained through the “mugration” approach found in BEAST. 4) Tree obtained through BASTA. The phylogeny branches are coloured according to their respective population, the phylogeny branches are coloured according to their respective population and the estimated time scale can be found below each tree (see previous page).

When looking at the BSSVS analysis output, the Epitree-sim and “mugration” approaches show clearly different results (see Figure 5-4). At the opposite of the “mugration” results, the Epitree-sim shows multiple strongly supported rates starting from the populations C to B and between the populations A and D with no strongly supported rates between A-C or B-D. At the opposite, the “mugration” model estimated that only the rates starting from the population A and B were well supported.

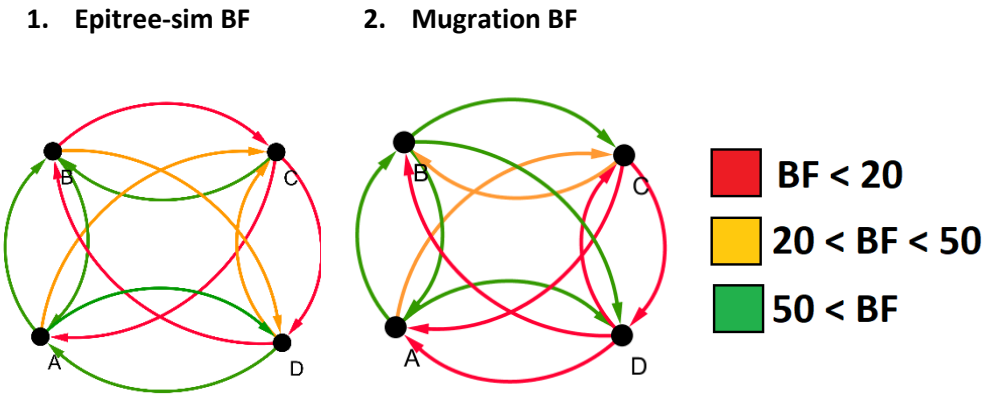


Figure 5-4: BSSVS analysis output for the discrete phylogenetic analysis performed by Epitree-sim or the “mugration” model for the simulated epidemic 1. The edges’ colours represent the relative strength by which the rates are supported by the BSSVS analysis.

The different rates estimated by each model translate what I observed on the different phylogenetical trees. While Epitree-sim estimated high transmission rates from the populations C and D toward the populations A or B and between the populations A and B, the “mugration” model only estimated high transmission rates between the populations A and B and the BASTA model between the populations C to B to A (see Figure 5-5).

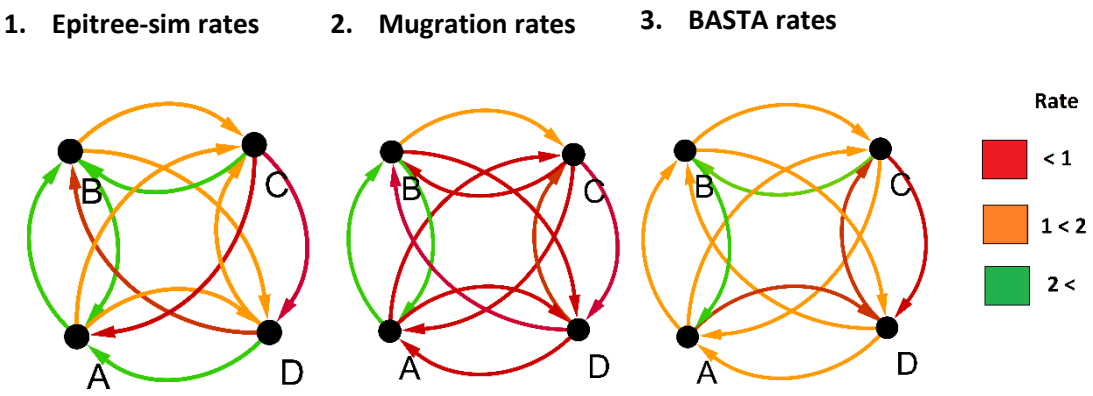


Figure 5-5: Transmission rates between the different population for the first simulated epidemic estimated by 1) Epitree-sim 2) the "mugration" model 3) BASTA. The edges' colours represent the rates of transition between the different locations.

5.4.2.1.2 Simulation 2

Because of the large number of samples and computational limitations, I was not able to perform the BASTA approach on this analysis.

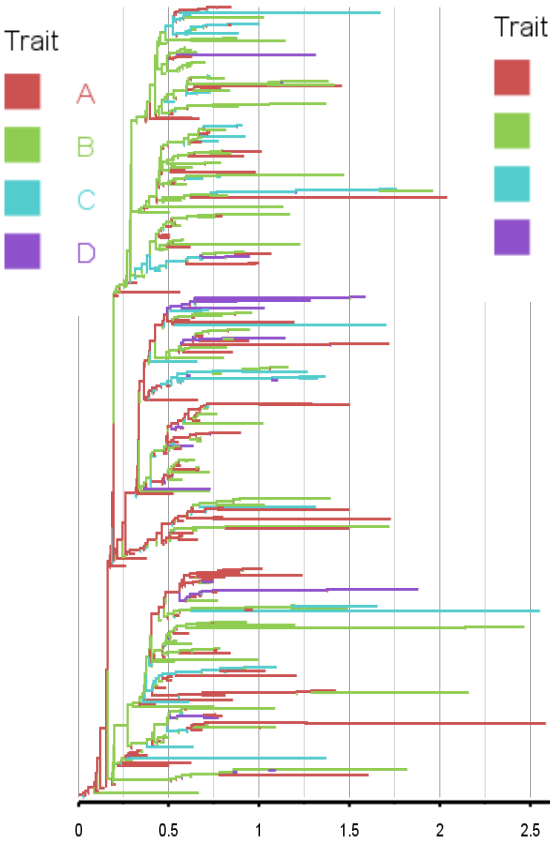
For the simulated epidemic 2, both the Epitree-sim and "mugration" approaches underestimated the real mean mutation rate of $3e^{-3}$ mutations/site/year with an estimate of 2.33e-3±1e-4 and 2.3e-3±1.2e-4 mutations per site per year respectively (see Table 5-7). However, while looking at the true TMRCA value of 2.53 years, the Epitree-sim approach estimated a closer TRMCA value of 2.49±0.21 compared to the 3.3±0.21 years estimated by the "mugration" approach.

Table 5-7: Evolutionary parameters and standard deviation comparison between the Epitree-sim and "mugration" phylogenetic approaches for the simulated epidemic 3.

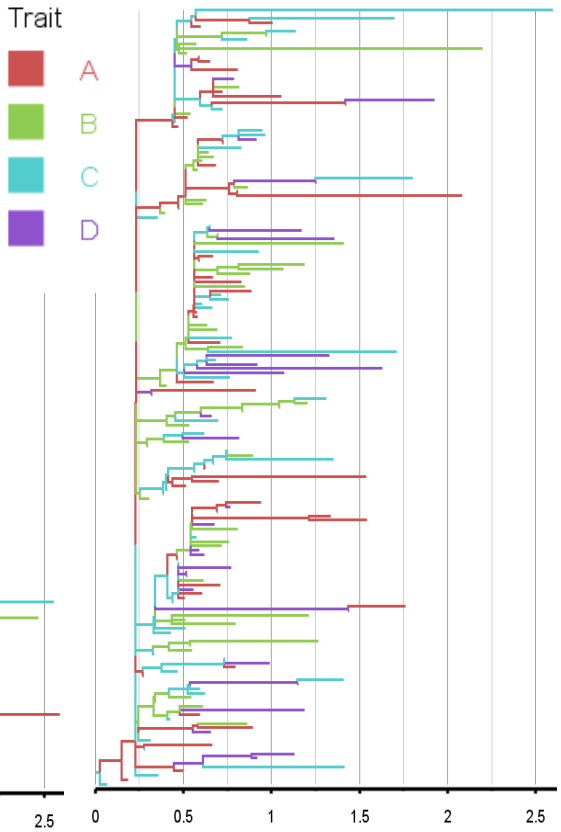
	True epidemic	**Subsampling approach**	**"mugration" approach**
Mean mutation rate (mutations/site/year)	$3e^{-3}$	$2.33e^{-3}$±$1e^{-4}$	2e-3±$1.2e^{-4}$
TMRCA (years)	2.53	2.49±0.21	3.3±0.22

When looking at the estimated phylogenetic trees (see Figure 5-6), we can observe that the overall topology of the Epitree-sim tree was closer to the true topology than the one estimated by the “mugration” approach. Overall, the “mugration” approach estimated that the population A is the main host for most of the epidemic history with the other populations being segregated at the tips of the tree with high uncertainties on the role of each populations in the virus circulation. Closer to the true tree, the Epitree-sim approach properly estimated that the population B and C played an important role in the epidemic history, but still overestimated the importance of the population A in the disease circulation.

1. True tree



2. Subsampling tree



3. "mugration" tree

Figure 5-6: Comparison between the different trees obtained for the set of sequence number 3. 1) The true transmission tree result of the simulated epidemic between the different populations in Epitree-sim. 2) Tree obtained through the discrete phylogenetic procedure found in Epitree-sim. 3) Tree obtained through the "mugration" approach found in BEAST. The phylogeny branches are coloured according to their respective population the key found next to each tree and the estimated time scale can be found below each tree (See previous page).

methods estimated different output (see Figure 5-7). While Epitree-sim estimated that most of the transmission routes reaching and leaving the populations A, B and C had a high BF, the "mugration" model estimated that only the routes starting from A or B were well supported.

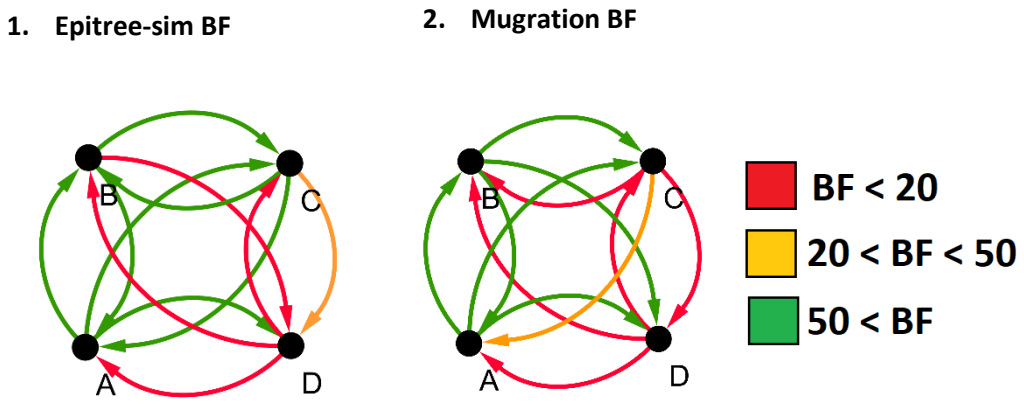


Figure 5-7: BSSVS analysis output for the discrete phylogenetic analysis performed by Epitree-sim or the "mugration" model for the simulated epidemic 3. The edges' colours represent the relative strength by which the rates are supported by the BSSVS analysis.

Finally, when looking at the estimated rates, we can observe that Epitree-sim estimated that most of the possible transmissions had high transmission rates, while the "mugration" approach estimated that only the transmission between A and B had high transmission rates (see Figure 5-8).

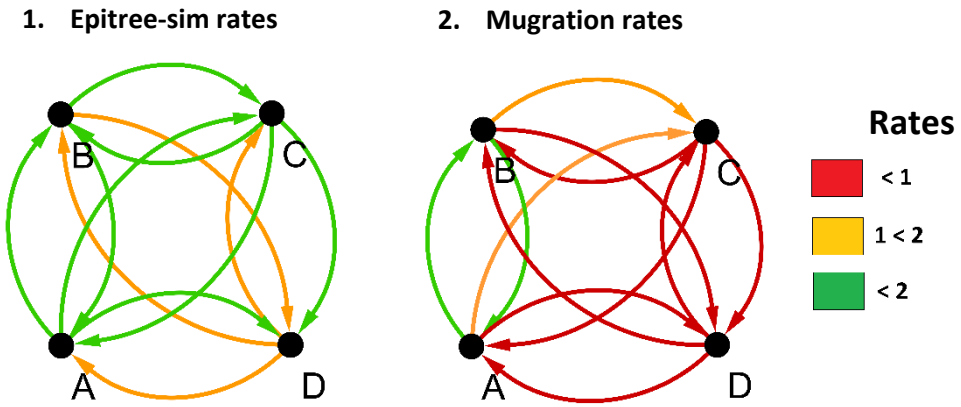


Figure 5-8: Transmission rates between the different populations for the third simulated epidemic estimated by 1) Epitree-sim 2) the "mugration" model 3) BASTA. The edges' colours represent the rates of transition between the different locations.

For the simulated epidemic 3, the "mugration" approach underestimated the real mean mutation rate of $3e^{-3}$ mutations/site/year with an estimate of 4 and 2.46e-3±1.2e-4 mutations per site per year respectively, while the Epitree-sim approach estimated a mutation rate in range with the true value of $2.84e^{-3}$±$3e^{-4}$ (see Table 5-8). However, while looking at the true TMRCA value of 2.27 years, the Epitree-sim approach estimated a closer TRMCA value of 2.49±0.21 compared to the 3.3±0.21 years estimated by the "mugration" approach.

Table 5-8: Evolutionary parameters and standard deviation comparison between the Epitree-sim and "mugration" phylogenetic approaches for the simulated epidemic 3.

	True epidemic	**Subsampling approach**	**"mugration" approach**
Mean mutation rate (mutations/site/year)	$3e^{-3}$	$2.84e^{-3}$±$3e^{-4}$	2e-46±$1.94e^{-4}$
TMRCA (years)	2.27	2.14±0.21	2.83±0.16

When looking at the estimated phylogenetic trees (see Figure 5-9), we can observe that the overall topology of the Epitree-sim tree was closer to the true topology than the one estimated by the "mugration" model. However, the Epitree-sim approach seems to have overestimated the role played by the population B in the circulation of the disease by identifying it as the population at the root of the tree and downplaying the importance of the populations A and C at the start of the epidemic. Overall, the "mugration" approach had lots of difficulties to properly identify the role of each population in the circulation of the disease. For this approach, the population B was estimated to be the most important population in the circulation of the disease.

5.4. Application and comparison with other phylogenetic approaches

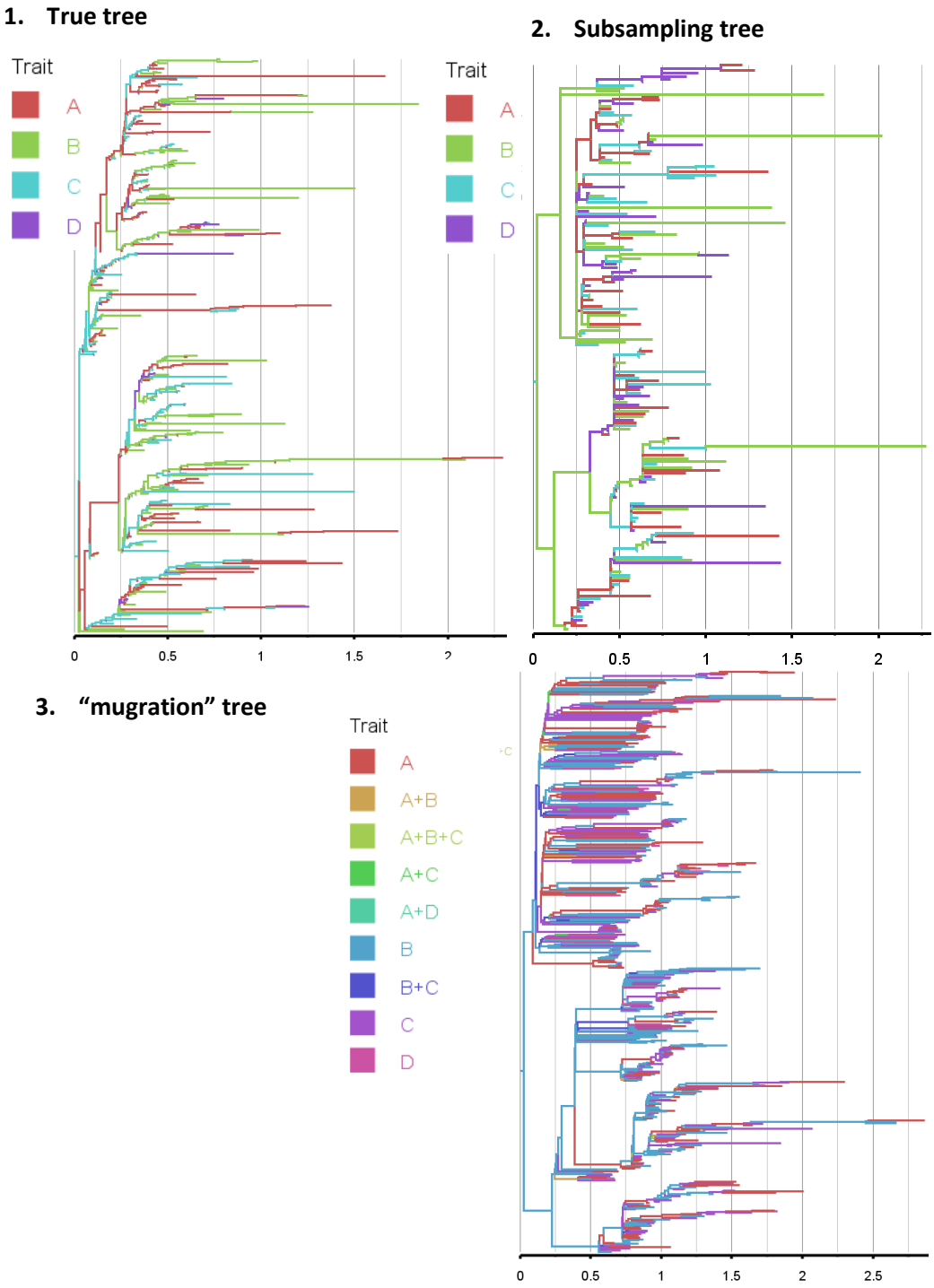


Figure 5-9: Comparison between the different trees obtained for the set of sequence number 5. 1) The true transmission tree, result of the simulated epidemic between the different populations in Epitree-sim. 2) Tree obtained through the discrete phylogenetic procedure found in Epitree-sim. 3) Tree obtained through the “mugration” approach found in BEAST. The phylogeny branches are coloured according to their respective population with the key found next to each tree and the estimated time scale can be found below each tree.

Interestingly, when looking at the BSSVS output for the Epitree-sim model, we can observe that only the transmissions involving the population C had BF values above 50. At the opposite, the "mugration" models estimated that most of the transmission routes, apart from those starting in the population D, were well supported (see Figure 5-10).

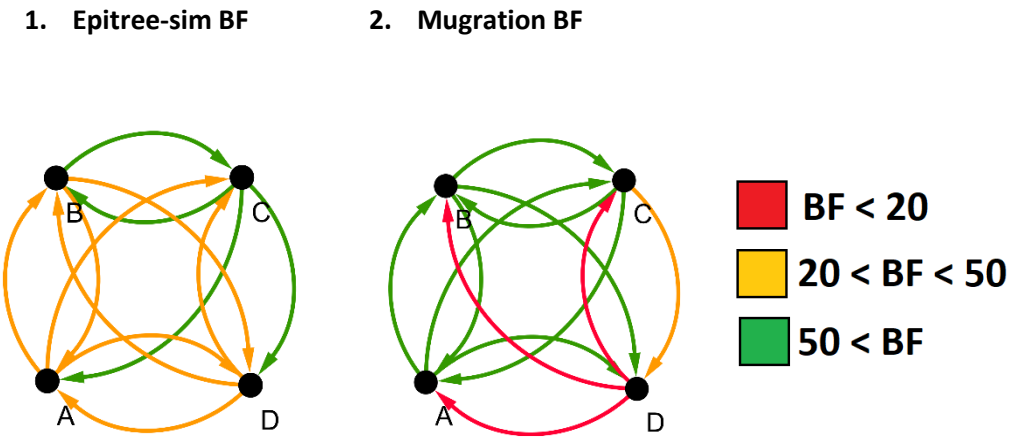


Figure 5-10: BSSVS analysis output for the discrete phylogenetic analysis performed by Epitree-sim or the "mugration" model for the simulated epidemic 3. The edges' colours represent the relative strength by which the rates are supported by the BSSVS analysis.

The Epitree-sim approach estimated that all the possible transmission routes have high transmission rates, while the "mugration" approach estimated that the routes involving population D have low transmission rate values (see Figure 5-11).

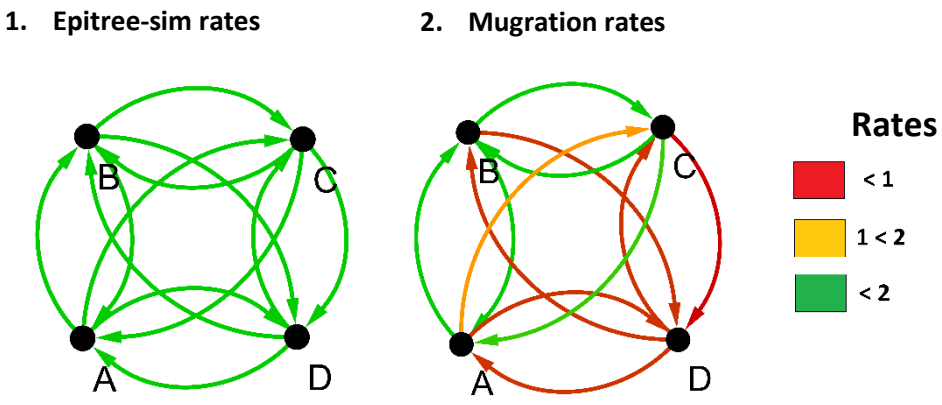


Figure 5-11: Transmission rates between the different populations for the fifth simulated epidemic estimated by 1) Epitree-sim 2) the "mugration" model 3) BASTA. The edges' colours represent the rates of transition between the different locations.

5.4.2.1.4 Other simulations

In this section, the output of the simulated epidemics 4, 5 and 6 are briefly described (for all the results see Supplementary figure 8-38 to Supplementary figure 8-45, Supplementary table 8-52 to Supplementary table 8-57 and Supplementary table 8-61 to Supplementary table 8-63).

Table 5-9: Table comparing the different output obtained by the Epitree-sim approach and the "mugration" approach to estimate the phylogenetic tree and epidemiological parameters for the fourth simulated epidemic.

Epidemic 4					
	Mutation rate (mutations/ site/year)	**TMRCA (Years)**	**Tree structure**	**Transmission rate**	**BSSVS**
True Epidemics	3e-3	5	Epidemic started in population A which compose most of the backbone of the tree with some incursions to population B. Few incursions to the population D at the start of the epidemic		
Epitree-sim	0.26e-3 ±1e-4	5.08±0.06	Epidemic started in population D but population A compose most of the backbone of the tree with some incursions to population B	Most of the rates can be found in the same bracket value of [1-2] apart from the A-D and C-D rates that are below 1	All rates are well supported apart from the route A-D
Mugration	2.54e-3 ±1.8e-4	5.389±0.16	Epidemic started in population B which compose most of the backbone of the tree with some incursions to population A	Most of the rates can be found below 1 apart from the rates A-B and B-A that are above 3	None of the routes starting in population D are well supported. All the routes starting in A and B are well supported

Table 5-10:Table comparing the different output obtained by the Epitree-sim approach and the "mugration" approach to estimate the phylogenetic tree and epidemiological parameters for the fifth simulated epidemic.

Epidemic 5					
	Mutation rate (mutations/site/year)	**TMRCA (Years)**	**Tree structure**	**Transmission rate**	**BSSVS**
True Epidemics	3e-3	3.37	The epidemic started in population B before being transmitted in population A. Two distinct clades can be seen, one with the virus mostly circulating in population A, the other in population B		
Epitree-sim	1.9e-3 ±14e-4	2.74±0.32	The epidemic started in population B. Two distinct clades can be seen, one with the virus circulating mostly in population B, the other one with the virus circulating between the populations A and C	Most of the rates can be found in the bracket [1-2]. Most of the rates involving the population B have a transmission rate above 2	All rates are well supported with the routes involving population D having the lowest BF value
Mugration	2.3e-3 ±1.7e-4	3.76±0.15	The epidemic started in population B with this population composing most of the backbone of the tree	None of the rates starting in population D are above 1. The highest transmission rates that can be found involve both A and B	None of the routes involving the population D are well supported. All of the routes starting in A and B are well supported

5.4. Application and comparison with other phylogenetic approaches

Table 5-11: Table comparing the different output obtained by the Epitree-sim approach and the "mugration" approach to estimate the phylogenetic tree and epidemiological parameters for the sixth simulated epidemic.

Epidemic 6					
	Mutation rate (mutations/site/year)	**TMRCA (Years)**	**Tree structure**	**Transmission rate**	**BSSVS**
True Epidemics	3e-3	2.6	Epidemic started in population C before infecting population B in the second half of the epidemic. Multiple introductions into A from B can be observed		
Epitree-sim	1.16e-3±1.6e-4	2.79±0.1	Epidemic started in population C before infecting population B in the second half of the epidemic. Multiple introductions into A from B can be observed	The transmission routes B-A is the only one with a value above 2. The routes B-C and C-A are the only ones under 1	The routes B-C and C-A are the only ones not well supported
Mugration	2.53e-3±3e-4	2.8±0.12	Epidemic started in population C before infecting population B in the second half of the epidemic. Multiple introductions into A from B can be observed	The rates involving population A are the lowest ones. The exception being the route B-A which is the only one with a value above 2	Apart from the routes between B and A, all routes involving A are not well supported

5.4.2.1.5 Comparison using Eurasian avian influenza sequences

5.4.2.2 Amongst bird populations

5.4.2.2.1 Using the Epitree-sim approach

I estimated the circulation of avian influenza between five types of hosts (domestic Anseriformes, Domestic Galliformes, Wild Anseriformes, Wild Charadriiforms and wild others) using the subsampling approach available in Epitree-sim (see

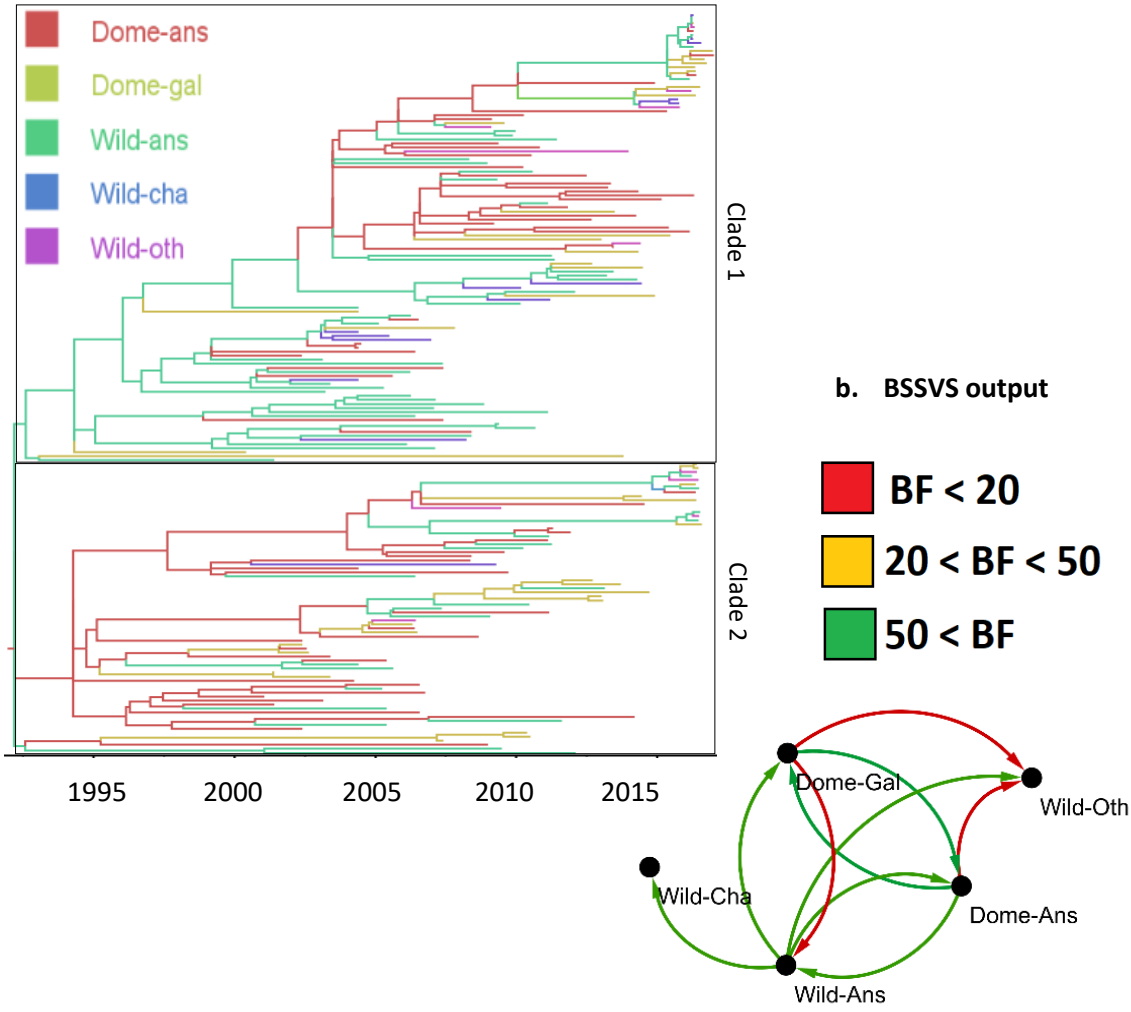


Figure *5-12*).

Figure 5-12: a. Phylogeographic tree representing the circulation of avian influenza amongst 5 bird populations using 282 PB2 avian influenza sequences. The phylogeny branches are coloured according to their respective host. B. BSSVS analysis output for the avian influenza between the sampled hosts using the "mugration" approach. The edges' colours represent the relative strength by which the rates are supported by the BSSVS analysis.

With the Epitree-sim subsampling approach, I estimated a mutation rate of $2e^{-3}$±$3.5e^{-5}$ mutations/per site/per year and a root age of 25.4 ±-0.57 years.

By looking at the annotated edges of the reconstructed tree (

Figure *5-12*a), this method estimated an important role of the wild Anseriformes birds in clade 1 and of domestic Anseriformes birds in clade 2. Additionally, I estimated a fairly important role of the domestics Galliform population in the recent virus circulation, with multiple virus transitions starting from this population near the tips of the tree.

The predominant role of the two Anseriformes and domestic Galliform populations in the circulation of the disease is highlighted by the BSSVS analysis results (

Figure *5-12*b and Supplementary table 8-64). The BSSVS analysis estimated an intense circulation between the three hosts. Interestingly, only the wild Anseriformes population seems to have a well-supported rate toward the wild Charadriiforms population. We can also notice that the wild others population seems to be closer to the domestic Anseriformes and Galliformes populations that to the two other wild populations.

5.4.2.2.2 Comparison with the “mugration” method

I performed the same host analysis using the “mugration” method available in BEAST. For this analysis, I used the whole dataset presented on the Table 5-2 and Table 5-3.

Compared to my subsampling approach, the “mugration” method estimated a faster mutation rate of $3.9e^{-3}\pm3e^{-4}$ mutations per site per year with an earlier TMRCA of 22.86 years (for the comparison between the estimated parameters between the subsampling and the “mugration” approach, see Table 5-12).

Table 5-12: Evolutionary parameters and standard deviation comparison between the subsampling and mugration approaches for the dataset of 282 Eurasian avian influenza PB2 sequences.

	Subsampling approach	**“mugration” approach**
Root height (years)	30±3	22.86±1.9
Mean mutation rate (substitutions/site/year)	$1.6\ e^{-3}\pm3e^{-4}$	$3.9e^{-3}\pm3e^{-4}$

a. Phylogenetic tree **b. BSSVS output**

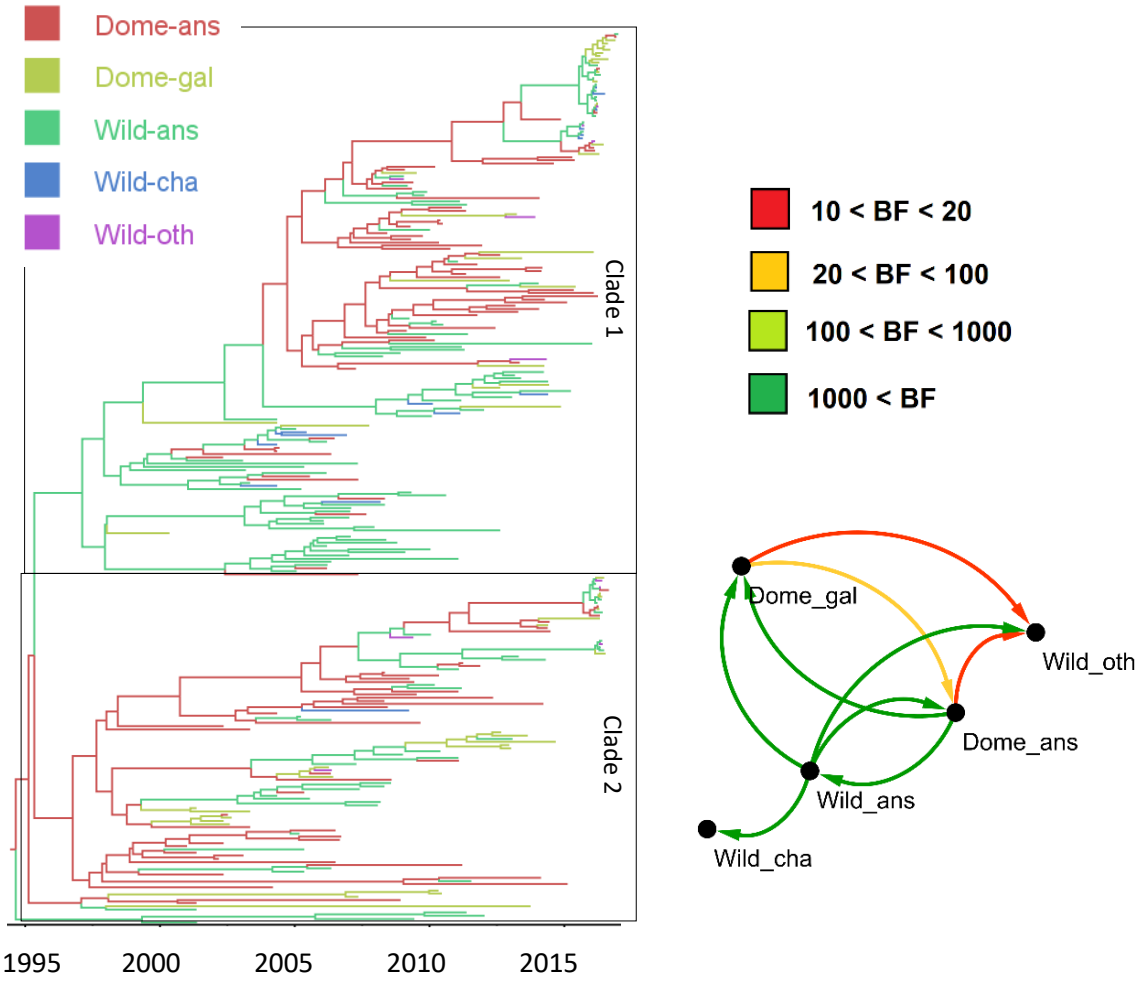


Figure 5-13: a. Phylogeographic tree representing the circulation of avian influenza amongst 5 types of hosts using 282 PB2 avian influenza sequences and obtained using the “mugration” approach. The phylogeny branches are coloured according their respective host. b. BSSVS analysis output for the avian influenza between the sampled hosts using the “mugration” approach. The edges’ colours represent the relative strength by which the rates are supported by the BSSVS analysis.

Such as the subsampling approach, the “mugration” method estimated an important role by the wild Anseriformes birds in clade 1 and the domestic Anseriformes birds in clade 2 (see

Figure *5-13*a).

Similarly, few differences can be observed between the BSSVS results of this subsampling approach and the “mugration” model (fig.

Figure *5-12*b and

Figure *5-13*b., Supplementary table 8-64 and Supplementary table 8-65). The two analyses estimated the same well supported rates, the only difference being that the subsampled method supports the transition between Domestic Galliform and wild Anseriformes. Although the support for the different transition rates was higher with the "mugration" approach, both methods estimated the same rates to be of high or low values. Both models estimated similar migration rates and indicator values.

5.4.2.3 Amongst Eurasian discrete locations

5.4.2.3.1 With the Epitree-sim approach

With the same dataset than the host analysis, I then estimated the circulation of avian influenza amongst 9 discrete Eurasian locations (Western Europe, Northern Europe, Southern Europe, Eastern Europe, Central Asia, North Central Asia, South Asia, East China and Eastern Asia).

Using my subsampling approach, I estimated a mutation rate of 2 e^{-3}±3e^{-5} mutations/site/year with a TMRCA of 25±0.3 years.

5.4. Application and comparison with other phylogenetic approaches

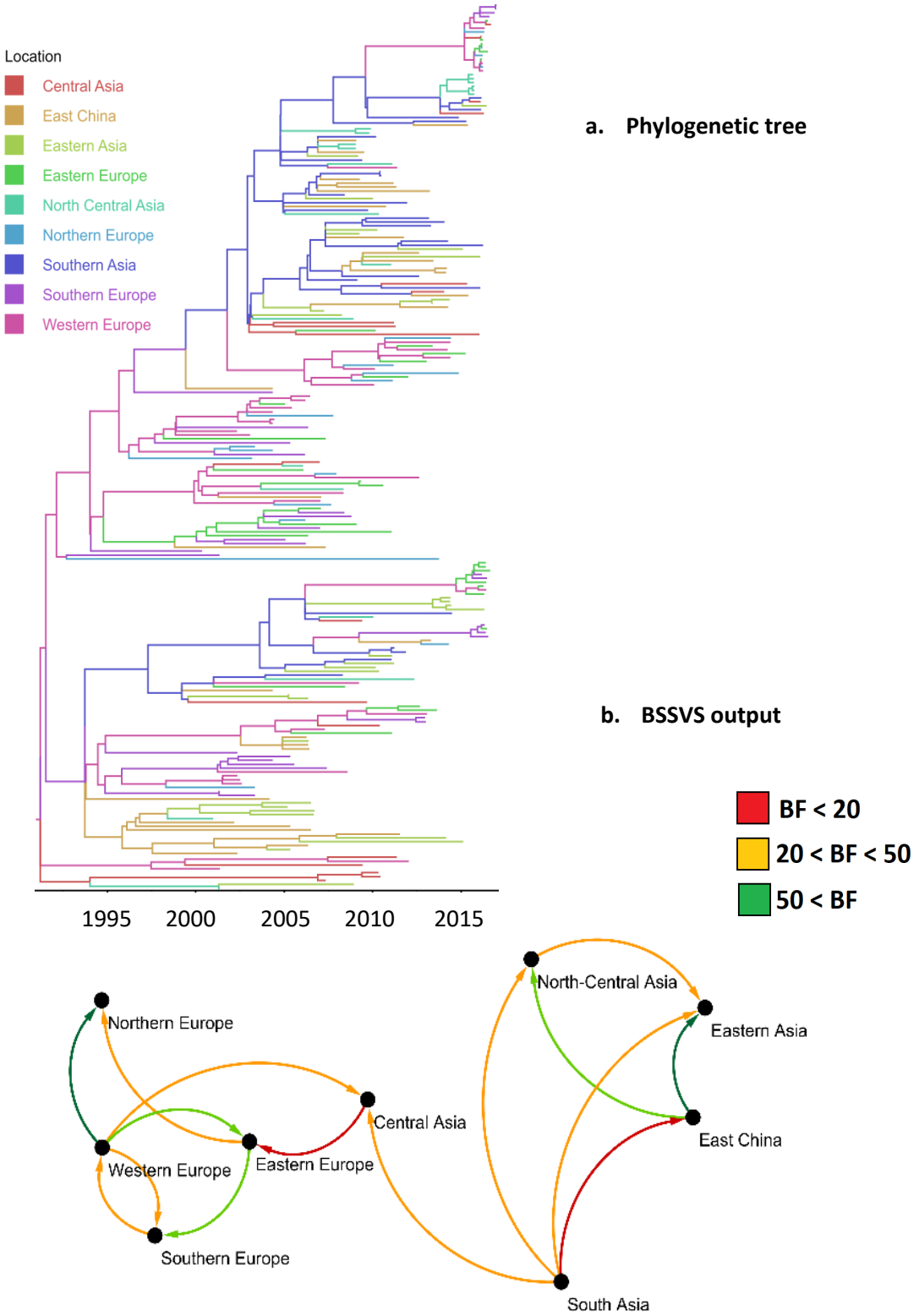


Figure 5-14: a. Phylogeographic tree representing the circulation of avian influenza amongst 9 Eurasian locations using 282 PB2 avian influenza sequences obtained using the subsampling method. The phylogeny branches are coloured according to their host. b. BSSVS analysis output for the avian influenza between the sampled locations using the subsampling approach. The edges' colours represent the relative strength by which the rates are supported by the BSSVS analysis.

By looking at the annotated tree, my approach estimated that most of the evolution of the virus took place in South Asia, and partially in Western Europe (see Figure 5-14a). Over the whole tree, only few connections can be observed between Europe and Asia.

Those preliminary observations of an important role of south Asia in the circulation of avian influenza is confirmed by the BSSVS analysis output (see Figure 5-14b and Supplementary table 8-68). With multiple links starting in Southern Asia toward all the other Asian locations, the central role of Southern Asia is easily observable. Interestingly, my results suggest that Central Asia seems to act as the unique connection between Asia and Europe. The Western part of Europe seems to be of major importance for the circulation of Avian influenza in Europe due to the numerous observations of well supported rates with all other European discrete states.

5.4.2.3.2 Comparison with the “mugration” approach

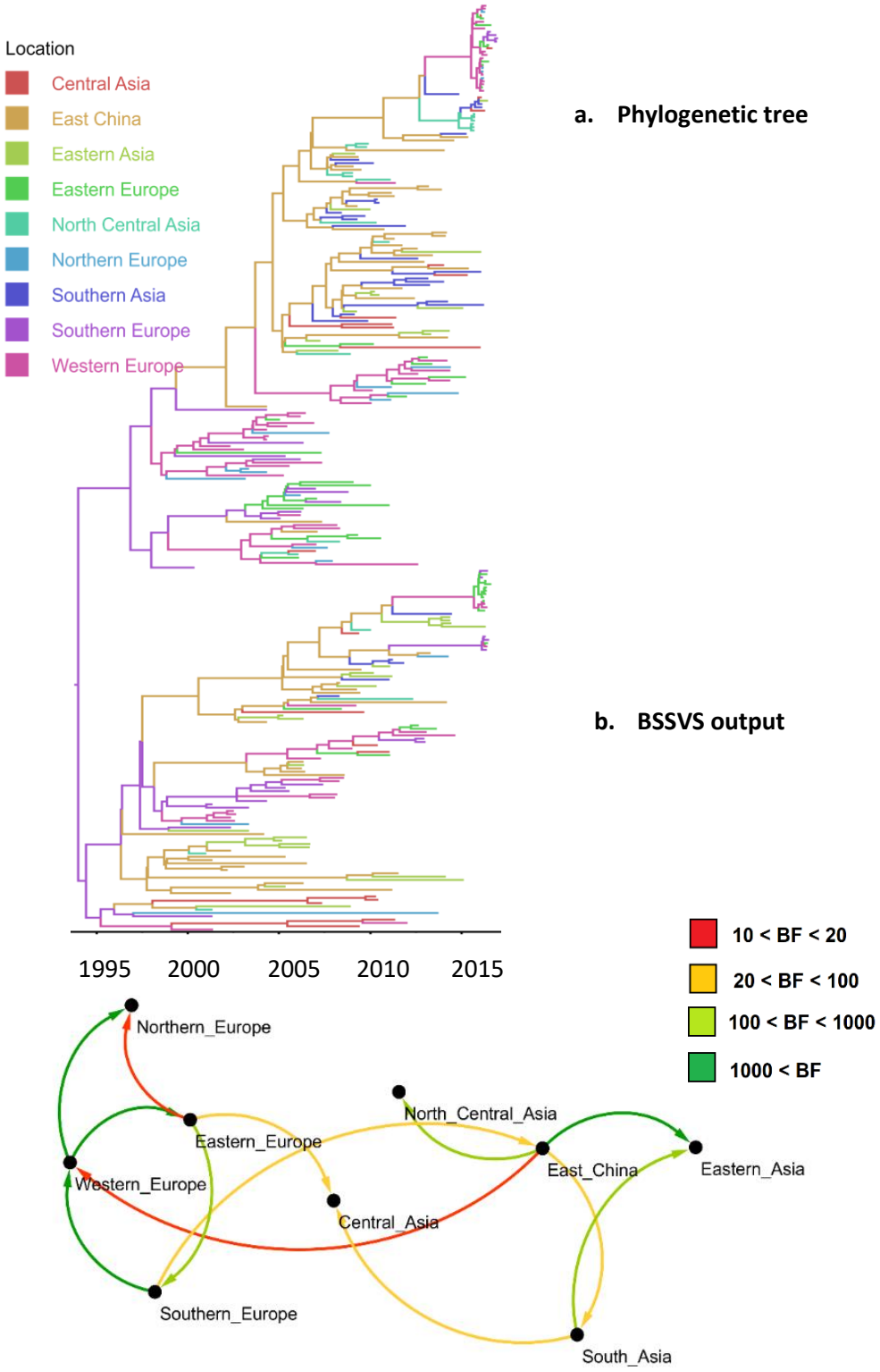


Figure 5-15: a. Phylogeographic tree representing the circulation of avian influenza amongst 9 Eurasian locations using 146 PB2 avian influenza sequences obtained using the “mugration” approach. The phylogeny branches are coloured according to their host. b. BSSVS analysis output for the avian influenza between the sampled locations using the “mugration” approach. The edges’ colours represent the relative strength by which the rates are supported by the BSSVS analysis.

With the “mugration” approach, I estimated a different location at the root of the tree than with the subsampling approach. In fact, the “mugration” approach pointed Southern Europe whereas the subsampling procedure estimated Western Europe as the origin location of the disease (see Figure 5-14a and

Figure 5-15a).

The “mugration” approach estimated that the virus mostly circulated within Southern Europe, Western Europe and East China, with multiple introductions into the other sampled regions, with only a minor role played by South Asia. However, similarly to the subsampling method, the phylogenetic tree obtained with “mugration” approach does not show many transmission events between Asia and Europe.

Similarly to the subsampling approach, the BSSVS output for the “mugration” approach suggest a separation between Europe and Asia with an important role played by Central Asia to connect the two continents (see Figure 5-14b and

Figure 5-15b, Supplementary table 8-68 and Supplementary table 8-69). Additionally, both methods estimated a strong connection between South Asia and Central Asia.

Looking at the BSSVS results, the main difference between the two approaches is the important role that East China plays in the circulation of the disease within Asia for the “mugration” approach, whereas this role is filled by South Asia for the subsampling approach. Both models estimated similar migration rate values and indicator values (see Supplementary table 8-68 and Supplementary table 8-69) for the well supported transmission rates. The only clear differences being the estimated indicator value between South Asia and East China with a higher indicator for the South-Asia-East China transmission and a lower East-China-South Asia within Epitree-sim.

5.4.2.4 Comparison using FMDV SAT1 sequences

5.4.2.4.1 With the subsampling approach

I used 135 VP1 FMDV SAT1 sequences sampled in Africa between 1961 and 2015 to estimate the circulation of the virus between three potential hosts: antelopes, buffaloes and cattle.

With my subsampling approach, I estimated a mutation rate of $1.8e^{-3}\pm1e^{-3}$ mutations/site/year with a TMRCA of around 155±90 years.

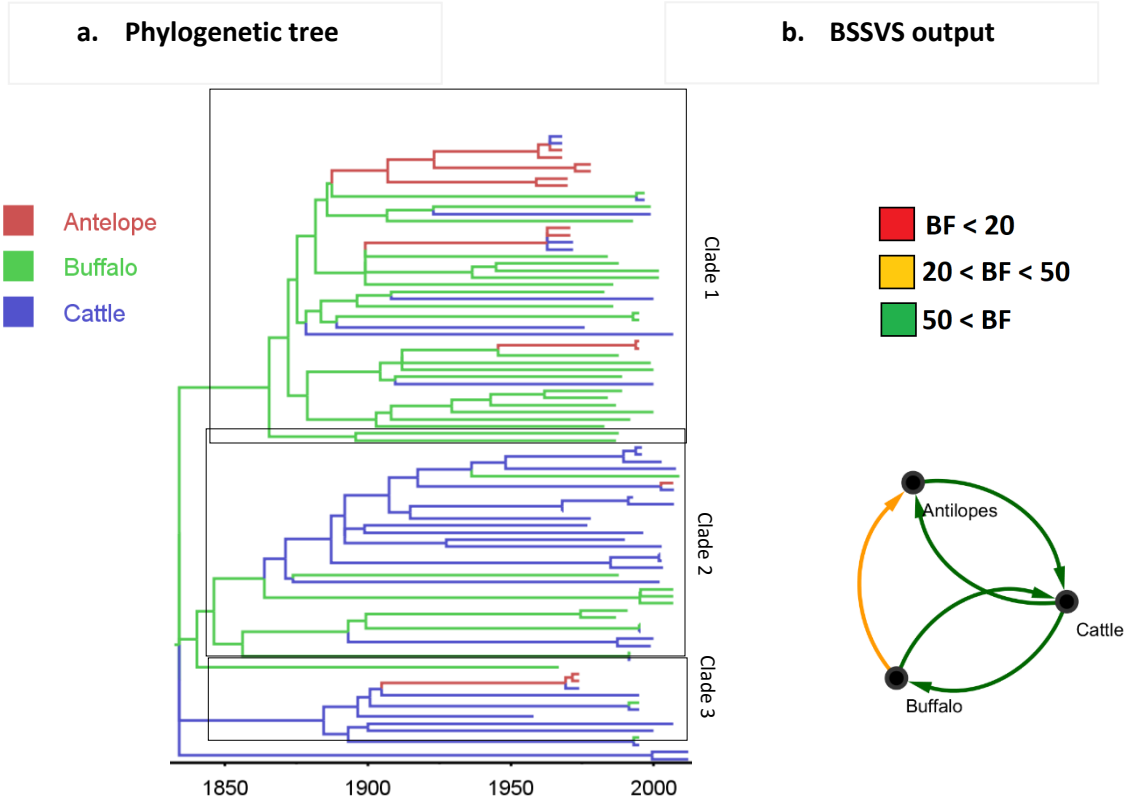


Figure 5-16: a. Phylogeographic tree representing the circulation of foot and mouth disease amongst antelopes, cattle and buffaloes using 134 VP1 sequences. The phylogeny branches are coloured according to their host. B. BSSVS analysis output for the foot and mouth disease analysis. b. BSSVS analysis output for foot and mouth disease analysis between the three hosts using the subsampling approach. The edges' colours represent the relative strength by which the rates are supported by the BSSVS analysis.

By looking at the clades were all three types of hosts can be found, my approach suggests that Buffaloes are the historical hosts of the disease (see Figure 5-16). The annotated tree also suggests that in recent times (since 1950), most of the virus introduced in cattle was of antelope origin. Interestingly, although transmission events from antelopes to cattle seem to be a fairly common event, there is no sign of viral transmission events from antelope to buffalo.

This last observation finds echo with the BSSVS output which does not show signs of supported transmission rates from antelope toward buffalo (Figure 5-16b. and Supplementary table 8-72). However, the BSSVS analysis suggests strong bi-directional transmission rates between cattle and antelopes or cattle and buffaloes.

5.4.2.4.2 Comparison with the “mugration” and BASTA approaches

When using the “mugration” approach, I estimated a mutation rate of $1.79e^{-3}\pm3e^{-4}$ mutations/site/year with TMRCA of 260±44 years. With the BASTA approach, I estimated a slightly faster mutation rate of $1.9e^{-3}\pm1.7e^{-4}$ mutations/site/year with a smaller TMRA of 219±23 years (for a comparison between the models, see Table 5-13).

Table 5-13: Evolutionary parameters and standard deviation comparison between the subsampling, “mugration” and BASTA approaches for the dataset of 135 VP1 FMDV SAT1 sequences.

	Subsampling approach	**“mugration” approach**	**BASTA**
Root height (years)	155±90	260±44	219±23
Mean mutation rate (substitutions/site/year)	$1.8e^{-3}\pm1e^{-3}$	$1.79e^{-3}\pm3e^{-4}$	$1.9e^{-3}\pm1.7e^{-4}$

5.4. Application and comparison with other phylogenetic approaches

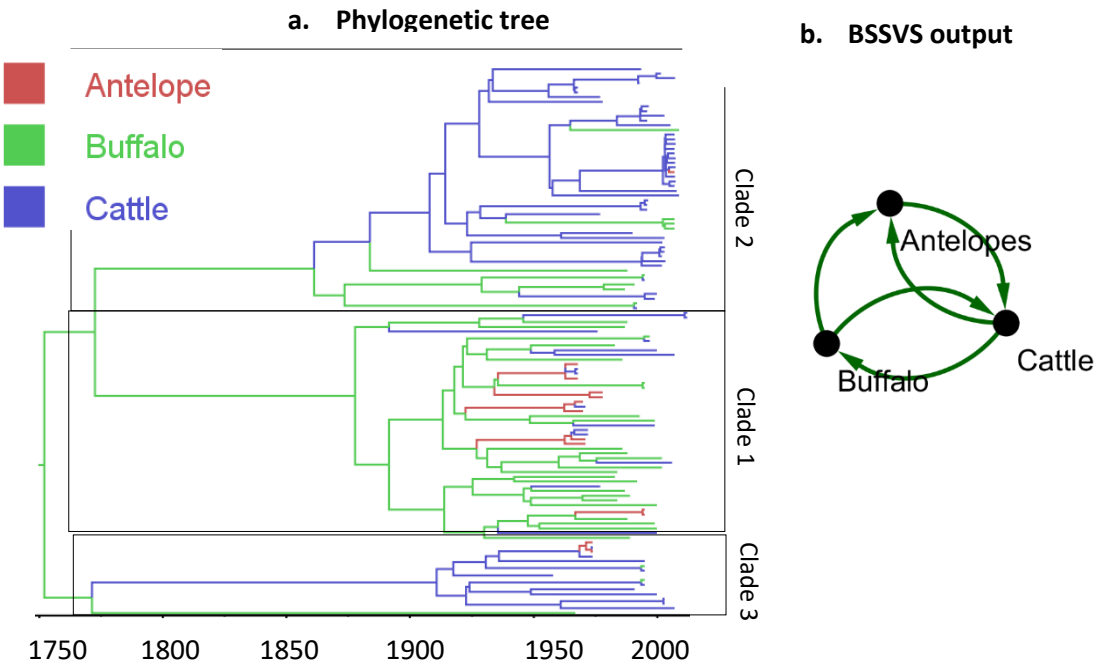


Figure 5-17: a. Phylogeographic tree representing the circulation of foot and mouth disease amongst antelopes, cattle and buffaloes using 134 VP1 sequences and estimated using the "mugration" approach. The phylogeny branches are coloured according to their respective host. The tree is annotated with the three main clades identified. b. BSSVS analysis output for foot and mouth disease analysis between the three hosts using the "mugration" approach. The edges' colours represent the relative strength by which the rates are supported by the BSSVS analysis.

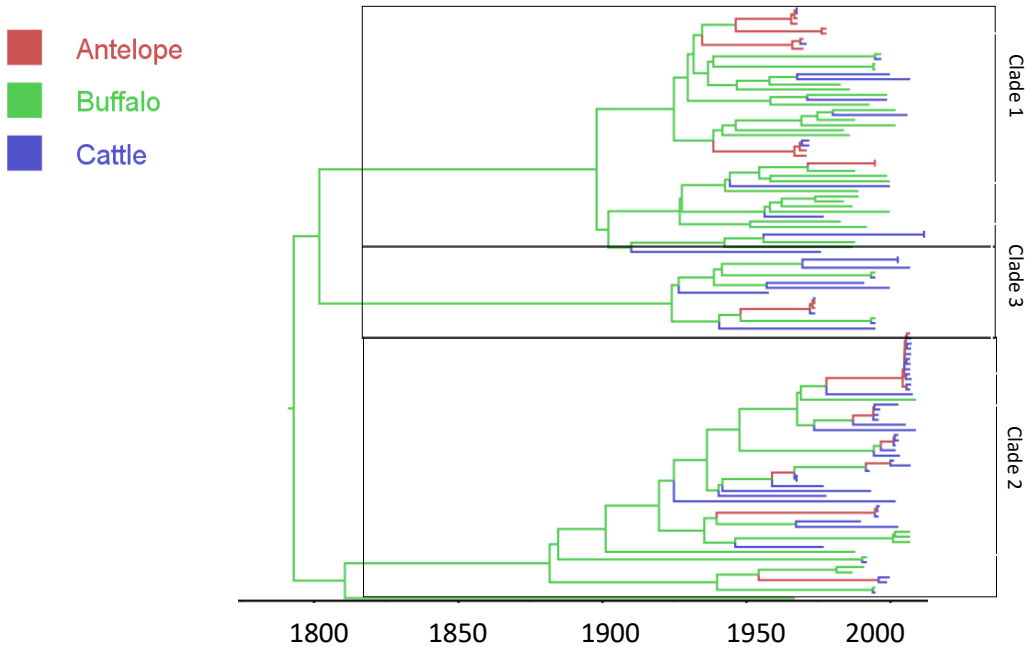


Figure 5-18: Phylogeographic tree representing the circulation of foot and mouth disease amongst antelopes, cattle and buffaloes using 134 VP1 sequences and estimated using the structural coalescent approximation approach BASTA. The phylogeny branches are coloured according to their respective host. The tree is annotated with the three main clades identified.

All the methods tested and identified the same three main clades while reconstructing the phylogeny of the virus (see Figure 5-16, Figure *5-17* and Figure *5-18*). All three analyses suggest that in the first clade the virus mostly circulated amongst the buffalo population with occasional introductions within the impala or cattle populations. For the second clade, both the subsampling and "mugration" approaches estimated that the virus first appeared within the buffalo population before being introduced in the cattle population in a single event at the end of the 19th century. On the other hand, the BASTA analysis suggests that in the second clade, the virus circulated almost entirely within the buffalo population, with only few 'recent' cattle and impala transmissions (see clade 1). Both the subsampling and "mugration" analyses estimated that the third clade circulated mostly within the cattle population with only few introductions within the impala and buffalo populations. Conversely, the BASTA analysis estimated that the buffalo population was the main host in the third clade.

In both the subsampling approach and the "mugration" model, we can observe that in few occasions the impala population acted as an intermediate host between the buffalo and cattle populations (see Figure 5-16a and

Figure *5-17*a). However, this observation seems more common with the BASTA analysis (

Figure *5-18*), where the impala population acted as the intermediate host between the buffalo and cattle populations on multiple occasions.

The two models identified a similar set of rates with a strong support, even though higher support values can be observed for the “mugration” analysis (see Figure 5-16b and

Figure *5-17*b, Supplementary table 8-72 and Supplementary table 8-73). However, the estimated rates antelope-buffalo, antelope-cattle and cattle-antelope were higher in the Epitree-sim approach (see Supplementary table 8-74).

5.5 CONCLUSION AND DISCUSSION

In this study, I presented a subsampling approach aiming to speed up phylodynamic analyses while being easy to use for non-experts in the field. The method is available under the form of a stand-alone software, Epitree-sim. At the core of Epitree-sim is a subsampling approach used to reduce subsampling bias, while keeping the

information available in whole sequence datasets. Epitree-sim can also be used to generate simulated epidemics and their corresponding sequence datasets.

I compared my method with other phylogenetic approaches, namely the “mugration” approach available in BEAST[348] and BASTA available in BEAST2[300]

I first performed those approaches on epidemics generated by Epitree-sim to compare the different final outputs of the three approaches with the true phylogenetic trees from which were generated the different sequence datasets. Secondly, I performed those three phylogenetic approaches on an avian influenza and foot and mouth disease sequence datasets to see how the new approach behaves on real world data.

The impact of the subsampling procedure applied by Epitree-sim can clearly be seen on the results of the simulated epidemic analyses on mitigating the influence of the most sampled populations. Here, at the opposite of the “mugration” approach, Epitree-sim seems to be able to estimate a phylogenetic tree structurally close to the true phylogenetic tree, by properly estimating the influence of the smaller populations on the course of the epidemic. Moreover, when looking at the simulated epidemics for which I was able to set up a BASTA analysis, we can see that the results estimated by Epitree-sim where a close match to this more computationally demanding approach. Finally, the impact of a potential sampling bias on the output of the “mugration” approach can be seen on the structure of the estimated phylogenetic tree with more weight put on the most sampled population, but also on the BSSVS output and estimated transmission rates involving such populations.

For the avian influenza analysis, compared to the “mugration” and BASTA approaches, my model seems to have underestimated the mutation of avian influenza but estimated a comparable mutation rate between the three analyses for the foot and

mouth disease virus. However, I observed that the estimated mutation rate and TMRCA values were highly variable depending on the clock and population models chosen to represent the virus evolution. In this regard, although Epitree-sim estimated the lowest mutation rate value, the TMRCA that Epitree-sim estimated was comparable to some of the values estimated by more complex models available in BEAST.

The results of the BSSVS of my subsampling analysis are only slightly different than those obtained using the “mugration” approach. This was expected since my subsampling approach uses the “mugration” approach to perform the discrete state analysis. However, my approach seems less unambiguous than the “mugration” model when estimating the support for the different transmission rates, probably due to recursive subsampling. I expected this outcome since the recursive subsampling procedure aims to mitigate the influence of potential sampling bias on the circulation analysis. The impact of the sub-sampling procedure on the phylogenetic analysis is particularly true for the avian influenza analyses between the discrete locations and for the simulated dataset analysis.

For the avian influenza circulation within Eurasia analysis, a sampling bias toward East-China and Western Europe was mitigated (from 48 sequences to 30 sequences for East-China and from 53 sequences to 30 sequences for Western Europe). Consequently, compared to the “mugration” approach results, the importance of those two discrete locations over the circulation analysis was reduced in the BSSVS analysis outputs of the subsampling approach. This is due to the “mugration” approach assuming a sampling intensity proportional to population size, leading to biased migration rate estimates when the analysis is performed on a biased sampled dataset.

When applied on the foot and mouth disease dataset, the results obtained by the Epitree-sim approach are closer to the ones obtained through the “mugration” model than from BASTA. In this case, my approach estimated a faster mutation rate (1.8e-3 compared to 1.8 e-4 for the “mugration” approach, and 1.9 e-4 for BASTA) but a similar TMRCA close to 200 years compared to the two other approaches. More, the subsampling and “mugration” approaches estimated that the virus introduction from buffaloes to cattle occurred at the end of the 19th, early 20th century. This is probably caused by the fact that Epitree-sim uses the “mugration” approach to perform the discrete state analysis and therefore is not able to grasp more potentially complex patterns that the BASTA approach can extract. However, because of the recursive subsampling, the Epitree-sim approach seems less unambiguous than the “mugration” model when estimating the support for the different transmission rates. This is mostly the consequence of the recursive subsampling procedure aiming to mitigate the influence of potential sampling bias on the circulation analysis.

Therefore, by performing a recursive subsampling on the analysed trait, Epitree-sim was able to mitigate the impact of the high sampling density on the estimated migration rate. Similarly, compared to the “mugration” approach, by removing any subsampling bias but keeping the information available in whole sequence datasets, my approach was closest to properly reconstruct the true phylogenetic tree of the simulated epidemic. Moreover, my approach was also the best method in terms of estimating the mutation rate of the simulated virus. However, it might have been influenced by the fact that the simulated sequences where generated under the assumption of a strict clock, which is very similar to the underlying assumptions made by treedater to infer the mutation rate.

Therefore, by combining multiple ‘simple’ steps, I was able to develop a robust phylogenetic method which can be seen as a reasonable alternative to some of the most commonly used but computationally intensive phylogenetic approaches.

In general, it is important for phylogenetic approaches to be able to cope with the increasing availability of genetic sequences, including outbreaks with more than a thousand sequences. A solution would be to use simpler, but statistically robust, methods than the Bayesian approaches often used by the popular methods. With its use of the recent advances in fast dating algorithms, my subsampling approach is able to reduce tremendously the time required to performed phylodynamic analyses compared to the “mugration”[17] approach available in BEAST[323], and BASTA[25] available in BEAST2[349] .

6 GENERAL DISCUSSION

Over this whole thesis, I used multiple available phylogenetic tools to study the circulation and transmission of avian influenza virus and foot and mouth disease amongst domestic and wild animals. In the chapters 2, 3 and 4, I used the “mugration” model approach to study the circulation of the viruses. With this approach, I had the possibility to determine multiple statistics (phylogenetic tree, transmission rate and number of transition events amongst multiple demes, etc.) about the circulation of those viruses. In in the chapter 2, I reconstructed the circulation of avian influenza within Eurasia and between multiple wild and domestic bird hosts. Similarly, in the chapter 3, I estimated the transmission routes involved in the circulation of foot and mouth disease in sub-Saharan Africa and between wild and domestic host populations. Using the same approach in the chapter 3, I was able to estimate the effects that multiple environmental and anthropological parameters, such as the density of potential host or humidity, had on the circulation of foot and mouth disease. In this order, I performed two types of analyses, one using a discrete location approach and the other using a continuous coordinate approach. For both methods, I used a resistance surface to represent the relation between each evaluated parameter and the estimated phylogeny.

In chapter 4, I compared the “mugration” approach with two structural coalescent approximation models, BASTA and MASCOT, that, in contrast to the “mugration” approach, consider the existence of multiple populations and the transmission events that might occur between them. The aim of this chapter was to compare the outcome of the “mugration” approach with two approaches supposedly less prone to statistical bias when performed on an uneven sampling dataset. Additionally, in chapter 4,

SCOTTI, another structural coalescent approximation approach that is derived from the BASTA approach, was performed to reconstruct the transmission network of avian influenza amongst multiple demes representing the bird populations present at one location over a specific year.

While the foot and mouth disease virus genome is only composed of a single strand, avian influenza has a segmented genome composed of multiple segments, each one of them coding for different proteins. This gave me the opportunity to study the recombination (reassortment) pattern of the two surface antigens of influenza virus in chapter 2. In this chapter, I more specifically looked at the influence that the population host and location have on the recombination pattern of the virus.

In the last chapter, I presented a new method to estimate the transmission and circulation of a pathogen amongst multiple demes while taking care of potential sampling bias by using a repeated subsampling of the sequence datasets. I demonstrated that this method constitutes a robust alternative to the most common phylogenetic approaches available in BEAST[323] and BEAST2[349].

6.1 DEFINITION OF POPULATION

The assumptions of working with an unstructured population when analysing the circulation of a pathogen is obsolete in the modern world of infectious disease dynamics. Therefore, efforts have been made to incorporate such heterogeneities in the mathematical models used to study the transmission of infectious diseases. We can consider this by acknowledging the existence of multiple populations (representing a type of host or location, for example) possibly involved in the

circulation of a particular pathogen. The aim of such analysis is to identify the original source of infection and the role of each of the identified populations. In Bayesian phylogenetic analysis, there are two main ways to include the existence of multiple populations. The first way to treat those discrete traits in phylogenetic analysis is to use the “mugration” model. This approach will treat the chosen traits as evolving in a similar way than the nucleotides of a genome. Another approach, less prone to statistical error, is the structured coalescent approach. However, for any type of phylogenetic analysis that aims to study the migration of the virus amongst multiple populations, it is important to define what are the populations to be considered.

When speaking about RNA viruses, trying to define what is considered as a single population can be a somewhat difficult task. As said before, RNA virus populations are estimated as being constituted of closely related mutants[15]. Although clear differences can be observed between viruses belonging to the same species (such as the serotypes for FMDV or the American and Eurasian strain for avian influenza), it often does not make sense to analyse them together in the same epidemiological analysis. Therefore, in phylogenetic, viral populations are often defined according to their respective host or location.

The way a population is defined in a phylogenetic analysis will impact its outcome and must be carefully considered to answer the scientific question asked. For example, over this whole thesis and when based on the location, multiple population definitions were used. Those definitions changed depending on the available information and objectives of the analysis. A first parameter considered was the scale of the analysis. In chapter 3, two definitions of population were used to analyse the transmission and circulation of FMDV. The first one, broader, was used when analysing the transmission on a continental scale, the second definition, more precise, was used when analysing the factors influencing the circulation of the disease in Eastern Africa.

For both analyses, the parameters to consider were the maximum number of populations, the analytical approach used and the number of sequences available.

In the chapters 2 and 3 of this thesis, I tried to find a balance between multiple potential populations (represented by the location sampled in this case) and the potential “errors” that might arise when analysing populations with few sequences. For example, in chapter 3, due to the final number of populations (between 8 and 14 depending on the serotype), performing this analysis using a structural coalescent approach would have been computationally difficult. Therefore, I adopted a “mugration” model to study the circulation of the virus. To analyse the factors influencing the circulation of FMDV in Africa, and because fewer sequences were available for this analysis, I used accurate sampling locations (precise up to the local region) to define the populations. Using this approach for this analysis was not problematic thanks to the small number of analysed sequences (less than 30). Additionally, using precise locations was sufficiently detailed when performing the GLM and SERAPHIM analyses. Using approximate locations would have impacted the quality of those analyses. In chapter 2, I used a structural coalescent model approximation (SCOTTI) to understand how the disease was transmitted between wild and domestic birds in Eurasia since 1990. In this regard, I defined a population as being the virus population present in one of the defined geographic areas (country or large area) in a particular year. Although this definition led to analysing numerous populations (around 100), I was able to perform the analysis since SCOTTI assumes that a transmission might only occur between two populations present at the same time.

As previously discussed, the definition of a population changed drastically depending on the method used and the question I was trying to answer. However, the underlying

biological significance of each population was always considered, and whether it made sense to use this definition of a population.

6.2 CONCEPT OF SCALE IN PHYLOGENETIC ANALYSIS

Related to the previous point is the concept of the scale used in phylogenetic analysis. The use of a scale in phylogenetic analysis is a concept of growing interest[351] that impacts any analysis and might influence our understanding of the patterns and processes that shape the observed phylogeny. As stated before, the transmission of infectious disease is influenced by multiple environmental and anthropological factors (such as land use or humidity, for example) that are all present at a certain scale. Therefore, to successfully take into account the effect that those factors might have on the pathogen transmission, it is imperative to use an appropriate scale which effectively represents their influence.

This concept has been more important in chapter 3 where I had to find an appropriate analysis scale to find the right balance between measurement of the genetic connectivity, population definition and computation time, to effectively test the effect of the different factors. In order to test the level where the factors have the strongest impact on the transmission, and the scale which is the most suitable for an effective control of the disease circulation, it would have been better to analyse several spatial scales, as done in a previous study using a similar approach by Jacquot et al (2017)[117]. However, because the presence of one of the main hosts of FMDV, in this case the cattle population, is related to the presence of human populations, it was expected that the human connectivity would intensify the circulation of the disease and overcome most of the natural obstacles to the virus dispersal[352].

Determining the evolutionary pattern of pathogen evolution at both the within- and between-host scales is still a challenging task in phylodynamics[2]. Previously, addressing this issue was limited by the absence of sequence data at both the within- and between-host scales. However, with the decreasing cost of sequencing, the possibility to perform phylogenetic studies at multiple scales is now a reality. Recently, a few studies have started to look at the role of within-host evolution in pathogen phylodynamics[353–356]. However, and because of the sparse nature of the available data, the application of such a complex approach when studying animal diseases would remain challenging and would need some adaptations.

6.3 FINAL CONCLUSION

The study of the transmission of zoonotic disease is as relevant today as it was in the past, and recent events have shown how quickly a competent zoonosis can expand in the human population. The emergence of zoonoses can be driven by changes in landscape and land use, and therefore it is important to understand how such phenomenon can be associated with the emergence of infectious diseases[357]. The concept of landscape epidemiology, the study of the impacts of landscape structure on epidemiological processes, is not new and goes back to the 30's[30]. However, the use of new genetics and phylogenetics tools hold enormous promises for new development in the field of landscape epidemiology and the study of emerging zoonoses[46]. In this regard, phylogenetic analysis is one of the tools that can be used to estimate the impact of environmental changes in the transmission of such diseases. However, such analysis comes with its set of challenges. At each step of phylogenetic analysis, choices can be done that will ultimately impact the quality of the results or the interpretation that can be drawn from such model.

One of the starting points of this thesis was to see if modern phylogenetic approaches could help us understand the effect of the environment on the circulation of an endemic disease, in this case FMDV in sub-Saharan Africa. In this regard, multiple choices of models can be used (direct or indirect) on an infinite number of geographical scales, mostly depending on the available data. In such analysis, the difficulty resides mostly in processing the available information, dealing with missing data, the decision of the resolution for the different raster that will be used, and so on. As said before, each step can affect the outcome of the analysis.

A second point was to understand the effects of the structure of the studied populations on phylogenetic analysis results and estimate if the resulting interpretations are more in line with field observations. Considering all the new parameters required for such structured analysis can be computationally demanding and lead to the existence of multiple models trying to approximate a true structured analysis, each one of them coming with its set of challenges and assumptions. Most of the time, the main difficulty on using such structured phylogenetical model comes from their sensibility on the set of priors used. Therefore, some good prerequisite knowledge of the studied disease is often needed before setting up the different variables used by such models.

The last point was that, although more complex methods are available, there is a growing need for methods able to deal with a growing number of available sequences, whilst remaining statistically robust. Using the more recent advances in phylogenetic analytic methods can be challenging, especially while working on many sequences at once. Therefore, I came up with a new method mimicking the canvas of random-forest analysis in machine learning analysis, a simple model

applied multiple times over a large number of subsets of the whole sequence dataset.

Although this work was performed in the years prior to the emergence of SARS-CoV-2 (COVID-19) in the human population, its relevance in the current situation makes no doubts. In the time between completing the work and finalising this thesis, there has been an enormous sequencing effort globally, with the United Kingdom generating tens of thousands of whole genomes sequences. Because as there have been different responses from different countries (“demes”) with differing travel bans (different migration rates between demes) and different susceptibilities levels to severe infection in different age groups (“host”), phylogenetic analyses of the COVID-19 pandemic need to take the structure of the human population into account. By considering phylodynamic methods in the context of host and spatial structure, and how to approximate such analyses to be able to handle large datasets and reduce bias, this work paves the way for analyses of the current and next pandemics.

7 BIBLIOGRAPHY

1. Grenfell, B. T. *et al.* Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303,** 327–332 (2004).

2. Frost, S. D. W. *et al.* Eight challenges in phylodynamic inference. *Epidemics* **10,** 88–92 (2015).

3. Cauchemez, S. *et al.* Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 2825–2830 (2011).

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

5. Kühnert, D., Wu, C.-H. & Drummond, A. J. Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Infect. Genet. Evol.* **11,** 1825–1841 (2011).

6. Ferguson, N. M., Donnelly, C. A. & Anderson, R. M. The Foot-and-Mouth Epidemic in Great Britain: Pattern of Spread and Impact of Interventions. *Science (80-.).* **292,** 1155–1160 (2001).

7. Harris, S. R. *et al.* Evolution of MRSA During Hospital Transmission and Intercontinental Spread. *Science (80-.).* **327,** 469–474 (2010).

8. Roetzer, A. *et al.* Whole Genome Sequencing versus Traditional Genotyping for Investigation of a Mycobacterium tuberculosis Outbreak: A Longitudinal Molecular Epidemiological Study. *PLoS Med.* **10,** (2013).

9. Kao, R. R., Haydon, D. T., Lycett, S. J. & Murcia, P. R. Supersize me: How whole-genome sequencing and big data are transforming epidemiology. *Trends Microbiol.* **22,** 282–291 (2014).

10. Eshleman, S. H. *et al.* Analysis of genetic linkage of HIV from couples enrolled in the HIV Prevention Trials Network 052 trial. *J. Infect. Dis.* **204,** 1918–26 (2011).

11. Bao, Y. *et al.* The influenza virus resource at the National Center for Biotechnology Information. *J. Virol.* **82,** 596–601 (2008).

12. Kimura, M. & Ohta, T. On some principles governing molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* **71,** 2848–52 (1974).

13. Drummond, A., Pybus, O. G. & Rambaut, A. Inference of Viral Evolutionary Rates from Molecular Sequences. *Adv. Parasitol.* **54,** 331–358 (2003).

14. Andino, R. & Domingo, E. Viral quasispecies. *Virology* **479–480,** 46–51 (2015).

15. Domingo, E., Sheldon, J. & Perales, C. Viral Quasispecies Evolution. *Microbiol. Mol. Biol. Rev.* **76,** 159–216 (2012).

16. Fitch, W. M. & Margoliash, E. Construction of Phylogenetic Trees. *Science*

(80-.). **155,** 279–284 (1967).

17. Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian Phylogeography Finds Its Roots. *PLoS Comput. Biol.* **5,** e1000520 (2009).

18. Felsenstein, J. Phylogenies and Quantitative Characters. *Annu. Rev. Ecol. Syst.* **19,** 445–471 (1988).

19. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–425 (1987).

20. Sokal, R. R. & Michener, C. D. A Statistical Method for Evaluating Systematic Relationships. *Univ. Kansas Sci. Bull.* (1958). doi:citeulike-article-id:1327877

21. Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17,** 368–376 (1981).

22. De Maio, N., Wu, C. H. & Wilson, D. J. SCOTTI: Efficient Reconstruction of Transmission within Outbreaks with the Structured Coalescent. *PLoS Comput. Biol.* **12,** 1–23 (2016).

23. Müller, N. F., Rasmussen, D., Stadler, T. & Kelso, J. MASCOT: parameter and state inference under the marginal structured coalescent approximation. *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty406

24. Vaughan, T. G., Kuhnert, D., Popinga, A., Welch, D. & Drummond, A. J. Efficient Bayesian inference under the structured coalescent. *Bioinformatics* **30,** 2272–2279 (2014).

25. Maio, N. De, Wu, C., Reilly, K. M. O. & Wilson, D. New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. 1–22 (2015). doi:10.1371/journal.pgen.1005421

26. Dudas, G. *et al.* Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* **544,** 309–315 (2017).

27. Faria, N. R. *et al.* Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nat. Publ. Gr.* **546,** 35 (2017).

28. Lambin, E. F., Tran, A., Vanwambeke, S. O., Linard, C. & Soti, V. Pathogenic landscapes: interactions between land, people, disease vectors, and their animal hosts. *Int. J. Health Geogr.* **9,** 54 (2010).

29. Biek, R. & Real, L. A. The landscape genetics of infectious disease emergence and spread. *Mol. Ecol.* **19,** 3515–3531 (2010).

30. Ostfeld, R. S., Glass, G. E. & Keesing, F. Spatial epidemiology: An emerging (or re-emerging) discipline. *Trends Ecol. Evol.* **20,** 328–336 (2005).

31. Viana, M. *et al.* Assembling evidence for identifying reservoirs of infection. *Trends Ecol. Evol.* **29,** 270–279 (2014).

32. Spear, S. F., Balkenhol, N., Fortin, M. J., McRae, B. H. & Scribner, K. Use of resistance surfaces for landscape genetic studies: Considerations for parameterization and analysis. *Mol. Ecol.* **19,** 3576–3591 (2010).

33. Dion, E., VanSchalkwyk, L. & Lambin, E. F. The landscape epidemiology of foot-and-mouth disease in South Africa: A spatially explicit multi-agent

simulation. *Ecol. Modell.* **222,** 2059–2072 (2011).

34. McRae, B. H. Isolation By Resistance. *Evolution (N. Y).* **60,** 1551 (2006).

35. Meentemeyer, R. K., Haas, S. E. & Václavík, T. Landscape Epidemiology of Emerging Infectious Diseases in Natural and Human-Altered Ecosystems. *Annu. Rev. Phytopathol.* **50,** 379–402 (2012).

36. MINOR, E. S. & URBAN, D. L. A Graph-Theory Framework for Evaluating Landscape Connectivity and Conservation Planning. *Conserv. Biol.* **22,** 297–307 (2008).

37. Urban, D. & Keitt, T. Landscape Connectivity: A Graph-Theoretic Perspective. *Ecology* **82,** 1205–1218 (2001).

38. Shah, V. B. & McRae, B. H. Circuitscape : A Tool for Landscape Ecology. *Proc. 7th Python Sci. Conf.* 62–65 (2008). doi:10.1111/j.1523-1739.2008.00942.x

39. Bahl, J. *et al.* Influenza a virus migration and persistence in North American wild birds. *PLoS Pathog.* **9,** e1003570 (2013).

40. Kalthoff, D., Globig, A. & Beer, M. (Highly pathogenic) avian influenza as a zoonotic agent. *Vet. Microbiol.* **140,** 237–245 (2010).

41. Mostafa, A., Abdelwhab, E. M., Mettenleiter, T. C. & Pleschka, S. Zoonotic potential of influenza A viruses: A comprehensive overview. *Viruses* **10,** (2018).

42. Prempeh, H., Smith, R. & Müller, B. Foot and mouth disease: The human consequences: The health consequences are slight, the economic ones huge. *BMJ* **322,** 565–566 (2001).

43. Kandeel, A. *et al.* Zoonotic transmission of avian influenza virus (H5N1), Egypt, 2006-2009. *Emerg. Infect. Dis.* **16,** 1101–1107 (2010).

44. Vandegrift, K. J., Sokolow, S. H., Daszak, P. & Kilpatrick, A. M. Ecology of avian influenza viruses in a changing world. *Ann. N. Y. Acad. Sci.* **1195,** 113–128 (2010).

45. Kimura, M. The neutral theory of molecular evolution: A review of recent evidence. *Japanese J. Genet.* (1991). doi:10.1266/jjg.66.367

46. Archie, E. A., Luikart, G. & Ezenwa, V. O. Infecting epidemiology with genetics: a new frontier in disease ecology. *Trends Ecol. Evol.* **24,** 21–30 (2009).

47. Bielejec, F., Lemey, P., Baele, G., Rambaut, A. & Suchard, M. A. Inferring heterogeneous evolutionary processes through time: From sequence substitution to phylogeography. *Syst. Biol.* **63,** 493–504 (2014).

48. Lio, P. & Goldman, N. Models of Molecular Evolution and Phylogeny. *Genome Res.* **8,** 1233–1244 (1998).

49. Sydykova, D. K. & Wilke, C. O. Calculating site-specific evolutionary rates at the amino-acid or codon level yields similar rate estimates. *PeerJ* **5,** e3391 (2017).

50. Franzosa, E. A. & Xia, Y. Structural Determinants of Protein Evolution Are Context-Sensitive at the Residue Level. *Mol. Biol. Evol.* **26,** 2387–2395 (2009).

51. Jack, B. R., Meyer, A. G., Echave, J. & Wilke, C. O. Functional Sites Induce Long-Range Evolutionary Constraints in Enzymes. *PLOS Biol.* **14,** e1002452 (2016).

52. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11,** 725–36 (1994).

53. Wakeley, J. Substitution-rate variation among sites and the estimation of transition bias. *Mol. Biol. Evol.* **11,** 436–42 (1994).

54. Uzzell, T. & Corbin, K. W. Fitting discrete probability distributions to evolutionary events. *Science* **172,** 1089–96 (1971).

55. Wakeley, J. *Substitution-Rate Variation among Sites and the Estimation of Transition Bias*.

56. Zuckerkandl, E. & Pauling, L. Evolutionary divergence and convergence in proteins. *Evol. genes proteins* 97–166 (1965). doi:10.1209/epl/i1998-00224-x

57. Rambaut, A. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16,** 395–399 (2000).

58. Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R. & Rodrigo, A. G. Measurably evolving populations. *Trends Ecol. Evol.* **18,** 481–488 (2003).

59. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2,** vew007 (2016).

60. Volz, E. M. & Frost, S. D. W. Scalable relaxed clock phylogenetic dating. *Virus Evol.* **3,** (2017).

61. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed Phylogenetics and Dating with Confidence. *PLoS Biol.* **4,** e88 (2006).

62. Bhatt, S., Holmes, E. C. & Pybus, O. G. The Genomic Rate of Molecular Adaptation of the Human Influenza A Virus. *Mol. Biol. Evol.* **28,** 2443–2451 (2011).

63. Jenkins, G. M., Rambaut, A., Pybus, O. G. & Holmes, E. C. Rates of molecular evolution in RNA viruses: A quantitative phylogenetic analysis. *J. Mol. Evol.* (2002). doi:10.1007/s00239-001-0064-3

64. Rambaut, a & Bromham, L. Estimating divergence dates from molecular sequences. *Mol. Biol. Evol.* **15,** 442–448 (1998).

65. Pybus, O. G. Model Selection and the Molecular Clock. *PLoS Biol.* **4,** e151 (2006).

66. Ho, S. Y. W. An examination of phylogenetic models of substitution rate variation among lineages. *Biol. Lett.* **5,** 421–4 (2009).

67. Baer, C. F., Miyamoto, M. M. & Denver, D. R. Mutation rate variation in

multicellular eukaryotes: causes and consequences. *Nat. Rev. Genet.* **8,** 619–631 (2007).

68. Gillespie, J. H. *The causes of molecular evolution*. (Oxford University Press, 1991).

69. Baele, G. *et al.* Accurate Model Selection of Relaxed Molecular Clocks in Bayesian Phylogenetics Letter Fast Track. **30,** 239–243 (2012).

70. Yoder, A. D. & Yang, Z. Estimation of Primate Speciation Dates Using Local Molecular Clocks. *Mol. Biol. Evol.* **17,** 1081–1090 (2000).

71. Drummond, A. J. & Suchard, M. a. Bayesian random local clocks, or one rate to rule them all. *BMC Biol.* **8,** 114 (2010).

72. Worobey, M., Han, G.-Z. & Rambaut, A. A synchronized global sweep of the internal genes of modern avian influenza virus. *Nature* **508,** 254–257 (2014).

73. Kingman, J. F. C. The coalescent. *Stoch. Process. their Appl.* **13,** 235–248 (1982).

74. Frost, S. D. W. & Volz, E. M. Viral phylodynamics and the search for an 'effective number of infections'. *Philos. Trans. R. Soc. B Biol. Sci.* **365,** 1879–1890 (2010).

75. Lemey, P., Salemi, M., Vandamme, A.-M. & (eds). *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. *Recherche* (2009). doi:10.1002/ajhb.20017

76. Rodrigo, A. G. & Felsenstein, J. Coalescent approaches to HIV population genetics. in *The Evolution of HIV* 233–272 (1999).

77. Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22,** 1185–1192 (2005).

78. Minin, V. N. & Suchard, M. a. Fast, accurate and simulation-free stochastic mapping. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **363,** 3985–3995 (2008).

79. Gill, M. S. *et al.* Improving bayesian population dynamics inference: A coalescent-based model for multiple loci. *Mol. Biol. Evol.* **30,** 713–724 (2013).

80. Gascuel, O. & Steel, M. Neighbor-Joining Revealed. *Mol. Biol. Evol.* **23,** 1997–2000 (2006).

81. Kuhner, M. K. & Felsenstein, J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* **11,** 459–468 (1994).

82. Atteson, K. The performance of neighbor-joining algorithms of phylogeny reconstruction. in 101–110 (Springer, Berlin, Heidelberg, 1997). doi:10.1007/BFb0045077

83. St. John, K., Warnow, T., Moret, B. M. . & Vawter, L. Performance study of phylogenetic methods: (unweighted) quartet methods and neighbor-joining. *J. Algorithms* **48,** 173–193 (2003).

84. Baele, G. *et al.* Improving the Accuracy of Demographic and Molecular Clock

Model Comparison While Accommodating Phylogenetic Uncertainty. *Mol. Biol. Evol.* **29,** 2157–2167 (2012).

85. Chor, B. & Tuller, T. Maximum likelihood of evolutionary trees: hardness and approximation. *Bioinformatics* **21,** i97–i106 (2005).

86. Hordijk, W. & Gascuel, O. Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. *Bioinformatics* **21,** 4338–4347 (2005).

87. Ronquist, F. Bayesian inference of character evolution. *Trends Ecol. Evol.* **19,** 475–481 (2004).

88. Holder, M. & Lewis, P. O. Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.* **4,** 275–284 (2003).

89. Sanmartín, I., van der Mark, P. & Ronquist, F. Inferring dispersal: a Bayesian approach to phylogeny-based island biogeography, with special reference to the Canary Islands. *J. Biogeogr.* **35,** 428–449 (2008).

90. Faria, N. R., Suchard, M. a, Rambaut, A., Streicker, D. G. & Lemey, P. Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **368,** 20120196 (2013).

91. Faria, N. R., Suchard, M. a., Rambaut, A. & Lemey, P. Toward a quantitative understanding of viral phylogeography. *Curr. Opin. Virol.* **1,** 423–429 (2011).

92. Ronquist, F. & Sanmartín, I. Phylogenetic Methods in Biogeography. *Annu. Rev. Ecol. Evol. Syst.* **42,** 441–464 (2011).

93. Dellicour, S., Rose, R. & Pybus, O. G. Explaining the geographic spread of emerging viruses: a new framework for comparing viral genetic information and environmental landscape data. *BMC Bioinformatics* **in press,** 1–12 (2016).

94. Minin, V. N. & Suchard, M. a. Counting labeled transitions in continuous-time Markov models of evolution. *J. Math. Biol.* **56,** 391–412 (2007).

95. Pagel, M., Meade, A. & Barker, D. Bayesian estimation of ancestral character states on phylogenies. *Syst. Biol.* **53,** 673–684 (2004).

96. Lemey, P. *et al.* Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2. *PLoS Pathog.* **10,** e1003932 (2014).

97. Faria, N. R. *et al.* Phylogeographical footprint of colonial history in the global dispersal of human immunodeficiency virus type 2 group A. *J. Gen. Virol.* **93,** 889–899 (2012).

98. Lemey, P., Rambaut, A., Welch, J. J. & Suchard, M. A. Phylogeography Takes a Relaxed Random Walk in Continuous Space and Time. *Mol. Biol. Evol.* **27,** 1877–1885 (2010).

99. Edwards, A. W. F. & Cavalli-Sforza, L. L. Reconstruction of evolutionary trees. *Phenetic Phylogenetic Classif.* **6,** 67–76 (1964).

100. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed Phylogenetics and Dating with Confidence. *PLoS Biol.* **4,** e88 (2006).

101. Raghwani, J. *et al.* Endemic Dengue Associated with the Co-Circulation of Multiple Viral Lineages and Localized Density-Dependent Transmission. *PLoS Pathog.* **7,** e1002064 (2011).

102. Brunker, K. *et al.* Landscape attributes governing local transmission of an endemic zoonosis: Rabies virus in domestic dogs. *Mol. Ecol.* **27,** 773–788 (2018).

103. Trovao, N. S., Suchard, M. A., Baele, G., Gilbert, M. & Lemey, P. Bayesian inference reveals host-specific contributions to the epidemic expansion of Influenza A H5N1. 1–12 (2015). doi:10.1093/molbev/msv185

104. Vignieri, S. N. Streams over mountains: Influence of riparian connectivity on gene flow in the Pacific jumping mouse (Zapus trinotatus). *Mol. Ecol.* **14,** 1925–1937 (2005).

105. Crooks, K. R. & Sanjayan, M. Connectivity conservation: Maintaining connections for nature. *Connect. Conserv.* 1–20 (2006). doi:10.1017/CBO9780511754821

106. McRae, B. H., Dickson, B. G., Keitt, T. H. & Shah, V. B. Using circuit theory to model connectivity in ecology, evolution, and conservation. *Ecology* **89,** 2712–2724 (2008).

107. McRae, B. H. & Beier, P. Circuit theory predicts gene flow in plant and animal populations. *Proc. Natl. Acad. Sci.* **104,** 19885–19890 (2007).

108. Dellicour, S., Vrancken, B., Trovão, N. S., Fargette, D. & Lemey, P. On the importance of negative controls in viral landscape phylogeography. *Virus Evol.* **4,** (2018).

109. Trovão, N. S. *et al.* Host ecology determines the dispersal patterns of a plant virus. *Virus Evol.* **1,** vev016 (2015).

110. Lemey, P. *et al.* Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2. *PLoS Pathog.* **10,** e1003932 (2014).

111. Magee, D., Beard, R., Suchard, M. a, Lemey, P. & Scotch, M. Combining phylogeography and spatial epidemiology to uncover predictors of H5N1 influenza A virus diffusion. *Arch. Virol.* **160,** 215–24 (2015).

112. Nelson, M. I. *et al.* Global migration of influenza A viruses in swine. *Nat. Commun.* **6,** 6696 (2015).

113. Nunes, M. R. T. *et al.* Air Travel Is Associated with Intracontinental Spread of Dengue Virus Serotypes 1-3 in Brazil. *PLoS Negl. Trop. Dis.* **8,** (2014).

114. Lu, L., Leigh Brown, A. J. & Lycett, S. J. Quantifying predictors for the spatial diffusion of avian influenza virus in China. *BMC Evol. Biol.* **17,** 16 (2017).

115. Dellicour, S., Rose, R., Faria, N. R., Lemey, P. & Pybus, O. G. SERAPHIM: studying environmental rasters and phylogenetically-informed movements. *Bioinformatics* btw384 (2016). doi:10.1093/BIOINFORMATICS/BTW384

116. Pybus, O. G. *et al.* Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc. Natl. Acad. Sci.* **109,** 15066–15071 (2012).

117. Jacquot, M., Nomikou, K., Palmarini, M., Mertens, P. & Biek, R. Bluetongue virus spread in Europe is a consequence of climatic, landscape and vertebrate host factors as revealed by phylogeographic inference. *Proceedings. Biol. Sci.* **284,** 20170919 (2017).

118. Suchard, M. A., Weiss, R. E. & Sinsheimer, J. S. Bayesian Selection of Continuous-Time Markov Chain Evolutionary Models. *Mol. Biol. Evol.* **18,** 1001–1013 (2001).

119. Jeffreys, H. Some Tests of Significance, Treated by the Theory of Probability. *Math. Proc. Cambridge Philos. Soc.* **31,** 203 (1935).

120. Lartillot, N. & Philippe, H. Computing Bayes Factors Using Thermodynamic Integration. *Syst. Biol.* **55,** 195–207 (2006).

121. Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. in 199–213 (Springer, New York, NY, 1998). doi:10.1007/978-1-4612-1694-0_15

122. Raftery, A. E., Newton, M., Krivitsky, P. N. & Satagopan, J. M. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. in *Bayesian Statistics* **8,** 1–45 (2007).

123. Ogata, Y. A Monte Carlo method for high dimensional integration. *Numer. Math.* **55,** 137–157 (1989).

124. Xie, W., Lewis, P. O., Fan, Y., Kuo, L. & Chen, M.-H. Improving Marginal Likelihood Estimation for Bayesian Phylogenetic Model Selection. *Syst. Biol.* **60,** 150–160 (2011).

125. Baele, G. *et al.* Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty. (2012). doi:10.1093/molbev/mss084

126. Pannell, J. R. Coalescence in a metapopulation with recurrent local extinction and recolonization. *Evolution* **57,** 949–961 (2003).

127. Bodmer, W. F. & Cavalli-Sforza, L. L. A migration matrix model for the study of random genetic drift. *Genetics* **59,** 565–592 (1968).

128. Beerli, P. & Felsenstein, J. Maximum-likelihood estimation of migration rates and effective population numbers. *Genetics* **152,** 763–773 (1999).

129. Wright, S. Isolation by Distance. *Genetics* **28,** 114–138 (1943).

130. Beerli, P. & Felsenstein, J. *Maximum-Likelihood Estimation of Migration Rates and Effective Population Numbers in Two Populations Using a Coalescent Approach migration rate per generation*. (1999).

131. Volz, E. M. Complex population dynamics and the coalescent under neutrality. *Genetics* **190,** 187–201 (2012).

132. Lu, L., Lycett, S. J. & Leigh Brown, A. J. Determining the Phylogenetic and Phylogeographic Origin of Highly Pathogenic Avian Influenza (H7N3) in

Mexico. *PLoS One* **9,** e107330 (2014).

133. Ypma, R. J. F., van Ballegooijen, W. M. & Wallinga, J. Relating Phylogenetic Trees to Transmission Trees of Infectious Disease Outbreaks. *Genetics* **195,** 1055–1062 (2013).

134. Jombart, T. *et al.* Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLoS Comput. Biol.* **10,** e1003457 (2014).

135. Volz, E. M. & Frost, S. D. W. Inferring the Source of Transmission with Phylogenetic Data. **9,** (2013).

136. Newman, M. E. . *Networks. An introduction*. *Oxford University Press* (2010). doi:10.1111/j.1468-5922.2010.01872_2.x

137. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D. Complex networks: Structure and dynamics. *Phys. Rep.* **424,** 175–308 (2006).

138. Klovdahl, A. S. Social networks and the spread of infectious diseases: The AIDS example. *Soc. Sci. Med.* **21,** 1203–1216 (1985).

139. Gardy, J. L. *et al.* Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. *N. Engl. J. Med.* **364,** 730–739 (2011).

140. Eames, K., Bansal, S., Frost, S. & Riley, S. Six challenges in measuring contact networks for use in modelling. *Epidemics* **10,** 72–77 (2015).

141. Craft, M. E. Infectious disease transmission and contact networks in wildlife and livestock. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370,** (2015).

142. Chen, S. & Lanzas, C. Distinction and connection between contact network, social network, and disease transmission network. *Prev. Vet. Med.* **131,** 8–11 (2016).

143. Craft, M. E. & Caillaud, D. Network models: an underutilized tool in wildlife epidemiology? *Interdiscip. Perspect. Infect. Dis.* **2011,** 676949 (2011).

144. Gardy, J. L. *et al.* Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. *N. Engl. J. Med.* **364,** 730–739 (2011).

145. Frost, S. D. W. & Volz, E. M. Modelling tree shape and structure in viral phylodynamics. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **368,** 20120208 (2013).

146. Watts, D. J. & Strogatz, S. H. Collective dynamics of ‘small-world’ networks. *Nature* **393,** 440–2 (1998).

147. Pellis, L. *et al.* Eight challenges for network epidemic models. *Epidemics* **10,** 58–62 (2015).

148. Fortunato, S. & Hric, D. Community detection in networks: A user guide. *Phys. Rep.* **659,** 1–44 (2016).

149. Fortunato, S. Community detection in graphs. *Phys. Rep.* **486,** 75–174 (2010).

150. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A.-L. The large-

scale organization of metabolic networks. *Nature* **407,** 651–654 (2000).

151. Madotto, A. & Liu, J. Super-Spreader Identification Using Meta-Centrality. *Sci. Rep.* **6,** 38994 (2016).

152. Weiss, R. S. & Jacobson, E. A Method for the Analysis of the Structure of Complex Organizations. *Am. Sociol. Rev.* **20,** 661 (1955).

153. Zhou, H. Distance, dissimilarity index, and network community structure. *Phys. Rev. E* **67,** 061901 (2003).

154. Fortunato, S. *Community detection in graphs*.

155. Rosvall, M., Esquivel, A. V., Lancichinetti, A., West, J. D. & Lambiotte, R. Memory in network flows and its effects on spreading dynamics and community detection. *Nat. Commun.* **5,** 4630 (2014).

156. Pons, P. & Latapy, M. *Journal of Graph Algorithms and Applications Computing Communities in Large Networks Using Random Walks*. **10,** (2006).

157. Xu, R. & Wunsch, D. C. *Clustering*. (john Wiley & Son, 2009).

158. Prosser, D. J. *et al.* Mapping Avian Influenza Transmission Risk at the Interface of Domestic Poultry and Wild Birds. *Front. Public Heal.* **1,** 1–11 (2013).

159. De Jong, M. D. & Hien, T. T. Avian influenza A (H5N1). *J. Clin. Virol.* **35,** 2–13 (2006).

160. Smith, G. J. D. *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459,** 1122–5 (2009).

161. Ke, C. *et al.* Circulation of reassortant influenza A(H7N9) viruses in poultry and humans, Guangdong Province, China, 2013. *Emerg. Infect. Dis.* **20,** 2034–2040 (2014).

162. Gerloff, N. A. *et al.* Genetically Diverse Low Pathogenicity Avian Influenza A Virus Subtypes Co-Circulate among Poultry in Bangladesh. 1–29 (2016). doi:10.1371/journal.pone.0152131

163. Shoham, D. Review: Molecular evolution and the feasibility of an avian influenza virus becoming a pandemic strain--a conceptual shift. *Virus Genes* **33,** 127–32 (2006).

164. Jones, K. . *et al.* Global trends in emerging infectious diseases. *Nature* **451,** 990–993 (2008).

165. Jones, K. E. *et al.* Global trends in emerging infectious diseases HHS Public Access. *Nature* **451,** 990–993 (2008).

166. Daszak, P., Cunningham, A. A. & Hyatt, A. D. Emerging infectious diseases of wildlife - threats to biodiversity and human health. *Science (80-.).* **287,** 443–449 (2000).

167. Osterhaus, A. D. *et al.* Influenza B Virus in Seals. *Science (80-.).* **288,** 1051–1053 (2000).

168. Yuanji, G., Fengen, J. & Ping, W. Isolation of influenza C virus from pigs and experimental infection of pigs with influenza C virus. *J. Gen. Virol.* **64,** 177–182 (1983).

169. Bouvier, N. M. & Palese, P. The biology of influenza viruses. *Vaccine* **26,** 49–53 (2008).

170. Webster, R. G., Bean, W. J., Gorman, O. T., Chambers, T. M. & Kawaoka, Y. Evolution and ecology of influenza A viruses. *Microbiol. Rev.* **56,** 152–79 (1992).

171. Bouvier, N. M. & Palese, P. The biology of influenza viruses. *Vaccine* **26 Suppl 4,** D49-53 (2008).

172. Alexander, D. J. An overview of the epidemiology of avian influenza. *Vaccine* **25,** 5637–5644 (2007).

173. Tong, S. *et al.* New World Bats Harbor Diverse Influenza A Viruses. *PLoS Pathog.* (2013). doi:10.1371/journal.ppat.1003657

174. Tong, S. *et al.* A distinct lineage of influenza A virus from bats. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 4269–74 (2012).

175. Olsen, B. *et al.* Global Patterns of Influenza A Virus in Wild Birds. *Science (80-.).* **312,** 384–388 (2006).

176. Swayne, D. E. & Suarez, D. L. Highly pathogenic avian influenza. *Rev. Sci. Tech.* **19,** 463–82 (2000).

177. Perkins, L. E. L. & Swayne, D. E. Pathogenicity of a Hong Kong-origin H5N1 highly pathogenic avian influenza virus for emus, geese, ducks, and pigeons. *Avian Dis.* **46,** 53–63 (2002).

178. Horimoto, T. & Kawaoka, Y. Influenza: lessons from past pandemics, warnings from current incidents. *Nat.Rev.Microbiol.* **3,** 591–600 (2005).

179. Morris, R. S. *et al.* Epidemiology of H5N1 Avian Influenza in Asia and Implications for Regional Control. *Infection* 1–52 (2005).

180. Bowes, V. A. *et al.* Virus characterization, clinical presentation, and pathology associated with H7N3 avian influenza in British Columbia broiler breeder chickens in 2004. *Avian Dis* **48,** 928–934 (2004).

181. Villarreal, C. Avian influenza in Mexico. *Rev. Sci. Tech.* **28,** 261–5 (2009).

182. Sonnberg, S., Webby, R. J. & Webster, R. G. Natural history of highly pathogenic avian influenza H5N1. *Virus Res.* **178,** 63–77 (2013).

183. Runstadler, J., Hill, N., Hussein, I. T. M., Puryear, W. & Keogh, M. Connecting the study of wild influenza with the potential for pandemic disease. *Infect. Genet. Evol.* **17,** 162–187 (2013).

184. Jourdain, E. *et al.* Influenza virus in a natural host, the mallard: Experimental infection data. *PLoS One* **5,** (2010).

185. Munster, V. J. *et al.* Spatial, temporal, and species variation in prevalence of influenza a viruses in wild migratory birds. *PLoS Pathog.* **3,** 0630–0638 (2007).

186. Brown, J. D., Stallknecht, D. E. & Swayne, D. E. Experimental infection of swans and geese with highly pathogenic avian influenza virus (H5N1) of Asian lineage. *Emerg. Infect. Dis.* **14,** 136–42 (2008).

187. Krauss, S. *et al.* Influenza in migratory birds and evidence of limited intercontinental virus exchange. *PLoS Pathog.* **3,** 1684–1693 (2007).

188. Gill, R. E. *et al.* Extreme endurance flights by landbirds crossing the Pacific Ocean: ecological corridor rather than barrier? *Proc. R. Soc. B Biol. Sci.* **276,** 447–457 (2009).

189. Gilbert, M. & Pfeiffer, D. U. Risk factor modelling of the spatio-temporal patterns of highly pathogenic avian influenza (HPAIV) H5N1: A review. *Spat. Spatiotemporal. Epidemiol.* **3,** 173–183 (2012).

190. Pantin-Jackwood, M. J. *et al.* Role of poultry in the spread of novel H7N9 influenza virus in China. *J. Virol.* **88,** 5381–90 (2014).

191. Lam, T. T.-Y. *et al.* The genesis and source of the H7N9 influenza viruses causing human infections in China. *Nature* **502,** 241–4 (2013).

192. Van Boeckel, T. P. *et al.* Improving Risk Models for Avian Influenza: The Role of Intensive Poultry Farming and Flooded Land during the 2004 Thailand Epidemic. *PLoS One* (2012). doi:10.1371/journal.pone.0049528

193. Martin, V. *et al.* Spatial distribution and risk factors of highly pathogenic avian influenza (HPAI) H5N1 in China. *PLoS Pathog.* **7,** e1001308 (2011).

194. Bourouiba, L., Teslya, A. & Wu, J. Highly pathogenic avian influenza outbreak mitigated by seasonal low pathogenic strains: Insights from dynamic modeling. *J. Theor. Biol.* **271,** 181–201 (2011).

195. Nazir, J., Haumacher, R., Ike, A. C. & Marschang, R. E. Persistence of avian influenza viruses in lake sediment, duck feces, and duck meat. *Appl. Environ. Microbiol.* **77,** 4981–4985 (2011).

196. World Health Organization. Potential transmission of avian influenza (H5N1) through water , Sanitation and Hygiene and ways to reduce the risks to human health. 1–20 (2007).

197. Gilbert, M. *et al.* Mapping H5N1 highly pathogenic avian influenza risk in Southeast Asia. *Proc. Natl. Acad. Sci.* **105,** 4769–4774 (2008).

198. Gilbert, M. *et al.* Predicting the risk of avian influenza A H7N9 infection in live-poultry markets across Asia. *Nat. Commun.* **5,** 1–7 (2014).

199. Hénaux, V., Samuel, M. D. & Bunck, C. M. Model-based evaluation of highly and low pathogenic avian influenza dynamics in wild birds. *PLoS One* **5,** 1–7 (2010).

200. Krauss, S. *et al.* Long-term surveillance of H7 influenza viruses in American wild aquatic birds: are the H7N3 influenza viruses in wild birds the precursors of highly pathogenic strains in domestic poultry? *Emerg. Microbes Infect.* **4,** e35 (2015).

201. Henning, J. *et al.* Who Is Spreading Avian Influenza in the Moving Duck Flock Farming Network of Indonesia? *PLoS One* **11,** e0152123 (2016).

202. Kilpatrick, a M. *et al.* Predicting the global spread of H5N1 avian influenza. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 19368–19373 (2006).

203. Lam, T. T.-Y. *et al.* Migratory flyway and geographical distance are barriers to the gene flow of influenza virus among North American birds. *Ecol. Lett.* **15,** 24–33 (2012).

204. Scotch, M. *et al.* Diffusion of influenza viruses among migratory birds with a focus on the Southwest United States. *Infect. Genet. Evol.* **26,** 185–193 (2014).

205. Tian, H. *et al.* Avian influenza H5N1 viral and bird migration networks in Asia. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 172–7 (2015).

206. Lycett, S. J. *et al.* Role for migratory wild birds in the global spread of avian influenza H5N8. *Science (80-.).* **354,** 213–217 (2016).

207. Lee, D.-H. *et al.* Intercontinental Spread of Asian-Origin H5N8 to North America through Beringia by Migratory Birds. *J. Virol.* **89,** 6521–4 (2015).

208. Cheon, S. H. *et al.* Genetic evidence for the intercontinental movement of avian influenza viruses possessing North American-origin nonstructural gene allele B into South Korea. *Infect. Genet. Evol.* (2018). doi:10.1016/j.meegid.2018.09.001

209. Liu, J.-H. *et al.* Interregional Transmission of the Internal Protein Genes of H2 Influenza Virus in Migratory Ducks from North America to Eurasia. *Virus Genes* **29,** 81–86 (2004).

210. Chen, H. *et al.* Establishment of multiple sublineages of H5N1 influenza virus in Asia : Implications for pandemic control. **103,** 2845–2850 (2006).

211. Gauthier-Clerc, M., Lebarbenchon, C. & Thomas, F. Recent expansion of highly pathogenic avian influenza H5N1: A critical review. *Ibis (Lond. 1859).* **149,** 202–214 (2007).

212. Robinson, T. P. *et al. Global livestock production systems*. (2011).

213. THOMPSON, D. K. *et al.* Economic costs of the foot and mouth disease outbreak in the United Kingdom in 2001. *Rev. Sci. Tech. l'OIE* **21,** 675–687 (2002).

214. Valarcher, J.-F. *et al.* Incursions of Foot-and-Mouth Disease Virus into Europe between 1985 and 2006. *Transbound. Emerg. Dis.* **55,** 14–34 (2008).

215. Knowles, N. . & Samuel, A. . Molecular epidemiology of foot-and-mouth disease virus. *Virus Res.* **91,** 65–80 (2003).

216. Alexandersen, S., Zhang, Z., Donaldson, A. I. & Garland, A. J. M. The pathogenesis and diagnosis of foot-and-mouth disease. *J. Comp. Pathol.* **129,** 1–36 (2003).

217. Belsham, G. J. DISTINCTIVE FEATURES OF FOOT-AND-MOUTH DISEASE VIRUS, A MEMBER OF THE PICORNAVIRUS FAMILY; ASPECTS OF VIRUS PROTEIN SYNTHESIS, PROTEIN PROCESSING AND STRUCTURE. *Prog. Biophys. molec. Biol.,* **60,** 241–260 (1993).

218. Domingo, E., Baranowski, E., Escarmís, C. & Sobrino, F. Foot-and-mouth disease virus. *Comp. Immunol. Microbiol. Infect. Dis.* **25,** 297–308 (2002).

219. Fuquay, J. W., Schrijver, R. S. & Vosloo, W. Infectious Diseases: Foot-and-Mouth Disease. *Encycl. Dairy Sci.* 160–167 (2011). doi:http://dx.doi.org/10.1016/B978-0-12-374407-4.00132-1

220. Di Nardo, A., Knowles, N. J. & Paton, D. J. Combining livestock trade patterns with phylogenetics to help understand the spread of foot and mouth disease in sub-Saharan Africa, the East and Southeast Asia.Middle. *Rev. Sci. Tech.* **30,** 63–85 (2011).

221. Tekleghiorghis, T., Moormann, R. J. M., Weerdmeester, K. & Dekker, A. Foot-and-mouth Disease Transmission in Africa: Implications for Control, a Review. *Transbound. Emerg. Dis.* **63,** 136–151 (2016).

222. Samuel, A. & Knowles, N. Foot-and-mouth disease virus type O exhibit genetically and geographically distinct evolutionary lineages (topotypes). *J. Virol. methods* **82,** 609–621 (2001).

223. Coetzer, J. A. W. (ed. ., Thomson, G. R. (ed. . & Tustin, R. C. (ed. . Infectious diseases of livestock with special reference to Southern Africa. (1994).

224. Thomson, G. R., Vosloo, W. & Bastos, A. D. S. Foot and mouth disease in wildlife. *Virus Res.* **91,** 145–161 (2003).

225. Alexandersen, S. & Mowat, N. Foot-and-mouth disease: host range and pathogenesis. *Curr. Top. Microbiol. Immunol.* **288,** 9–42 (2005).

226. Weaver, G. V., Domenech, J., Thiermann, A. R. & Karesh, W. B. Foot and mouth disease: a look from the wild side. *J. Wildl. Dis.* **49,** 759–785 (2013).

227. Stenfeldt, C., Diaz-San Segundo, F., de los Santos, T., Rodriguez, L. L. & Arzt, J. The Pathogenesis of Foot-and-Mouth Disease in Pigs. *Front. Vet. Sci.* **3,** 41 (2016).

228. Stenfeldt, C. *et al.* Detection of Foot-and-mouth Disease Virus RNA and Capsid Protein in Lymphoid Tissues of Convalescent Pigs Does Not Indicate Existence of a Carrier State. *Transbound. Emerg. Dis.* **63,** 152–164 (2016).

229. Grubman, M. J. & Baxt, B. Foot-and-mouth disease. *Clin. Microbiol. Rev.* **17,** 465–93 (2004).

230. Brown, C. C., Piccone, M. E., Mason, P. W., McKenna, T. S. & Grubman, M. J. Pathogenesis of wild-type and leaderless foot-and-mouth disease virus in cattle. *J. Virol.* **70,** 5638–41 (1996).

231. Freimanis, G. L. *et al.* Genomics and outbreaks: foot and mouth disease. *Rev. Sci. Tech.* **35,** 175–89 (2016).

232. Knight-Jones, T. J. D. & Rushton, J. The economic impacts of foot and mouth disease - What are they, how big are they and where do they occur? *Prev. Vet. Med.* **112,** 162–173 (2013).

233. Lyons, N. A. *et al.* Impact of foot-and-mouth disease on milk production on a large-scale dairy farm in Kenya. *Prev. Vet. Med.* **120,** 177–186 (2015).

234. Casey, M. B. *et al. Patterns of Foot-and-Mouth Disease Virus Distribution in Africa: The Role of Livestock and Wildlife in Virus Emergence*. *The Role of Animals in Emerging Viral Diseases* (Elsevier, 2013). doi:10.1016/B978-0-12-405191-1.00002-8

235. Donaldson, A. I., Alexandersen, S., Sorensen, J. H. & Mikkelsen, T. Relative risks of the uncotrollable (airborne) spread of FMD by different species. *Vet. Rec.* **148,** 602–604 (2001).

236. Alexandersen, S., Zhang, Z., Reid, S. M., Hutchings, G. H. & Donaldson, A. I. Quantities of infectious virus and viral RNA recovered from sheep and cattle experimentally infected with foot-and-mouth disease virus O UK 2001. *J. Gen. Virol.* **83,** 1915–1923 (2002).

237. Yadav, S. *et al.* Parameterization of the Durations of Phases of Foot-And-Mouth Disease in Cattle. *Front. Vet. Sci.* **6,** 263 (2019).

238. Donaldson, A. I. & Alexandersen, S. Predicting the spread of foot and mouth disease by airborne virus. *Rev. Sci. Tech.* **21,** 569–75 (2002).

239. Cottam, E. M. *et al.* Transmission pathways of foot-and-mouth disease virus in the United Kingdom in 2007. *PLoS Pathog.* **4,** (2008).

240. Gloster, J., Blackall, R. M., Sellers, R. F. & Donaldson, A. I. Forecasting the airborne spread of foot-and-mouth disease. *Vet. Rec.* **108,** 370–4 (1981).

241. Donaldson, A. I., Alexandersen, S., Sørensen, J. H. & Mikkelsen, T. Relative risks of the uncontrollable (airborne) spread of FMD by different species. *Vet. Rec.* **148,** 602–4 (2001).

242. Paton, D. J., Gubbins, S. & King, D. P. Understanding the transmission of foot-and-mouth disease virus at different scales. *Curr. Opin. Virol.* **28,** 85–91 (2018).

243. Gloster, J. *et al.* The 2001 epidemic of foot-and-mouth disease in the United Kingdom: epidemiological and meteorological case studies. *Vet. Rec.* **156,** 793–803 (2005).

244. König, G. A. *et al.* SHORT COMMUNICATIONS Sequence data and evidence of possible airborne spread in the 2001 foot-and-mouth disease epidemic in the UK. *Vet. Rec.* **3,** 410–412 (2009).

245. Alexandersen, S. & Donaldson, A. I. Further studies to quantify the dose of natural aerosols of foot-and-mouth disease virus for pigs. *Epidemiol. Infect.* **128,** 313–23 (2002).

246. Donaldson, A. I., Gibson, C. F., Oliver, R., Hamblin, C. & Kitching, R. P. Infection of cattle by airborne foot-and-mouth disease virus: minimal doses with O1 and SAT 2 strains. *Res. Vet. Sci.* **43,** 339–46 (1987).

247. Gibson, C. F. & Donaldson, A. I. Exposure of sheep to natural aerosols of foot-and-mouth disease virus. *Res. Vet. Sci.* **41,** 45–9 (1986).

248. Pacheco, J. M. & Mason, P. W. Evaluation of infectivity and transmission of different Asian foot-and-mouth disease viruses in swine. *J. Vet. Sci.* **11,** 133–42 (2010).

249. Cottral, G. E. Persistence of Foot-and-Mouth Disease Virus in Animals, their Products and the Environment. *Bull. Off. int. Epiz* **71,** 3–4 (1969).

250. Sellers, R. F. Quantitative aspects of the spread of foot and mouth disease. *Vet. Bull* **41,** 431–439 (1971).

251. Richard Eales *et al.* The 2001 Outbreak of Foot and Mouth Disease. *Natl. Audit Off.* **939,** (2001).

252. Kitching, R. P. Global epidemiology and prospects for control of foot-and-mouth disease. *Curr. Top. Microbiol. Immunol.* **288,** 133–148 (2005).

253. Kitching, R. P., Hutber, A. M. & Thrusfield, M. V. A review of foot-and-mouth disease with special consideration for the clinical and epidemiological factors relevant to predictive modelling of the disease. *Vet. J.* **169,** 197–209 (2005).

254. FOOT AND MOUTH DISEASE AETIOLOGY Classification of the causative agent. Available at: http://www.oie.int/fileadmin/Home/eng/Animal_Health_in_the_World/docs/pdf/Disease_cards/FOOT_AND_MOUTH_DISEASE.pdf. (Accessed: 27th February 2017)

255. Bronsvoort, B. M. deC *et al.* Redefining the 'carrier' state for foot-and-mouth disease from the dynamics of virus persistence in endemically affected cattle populations. *Sci. Rep.* **6,** 29059 (2016).

256. Arzt, J., Juleff, N., Zhang, Z. & Rodriguez, L. L. The Pathogenesis of Foot-and-Mouth Disease I : Viral Pathways in Cattle. **58,** 291–304 (2011).

257. Condy, J. B., Hedger, R. S., Hamblin, C. & Barnett, I. T. The duration of the foot-and-mouth disease virus carrier state in African buffalo (i) in the individual animal and (ii) in a free-living herd. *Comp. Immunol. Microbiol. Infect. Dis.* **8,** 259–65 (1985).

258. Hedger, R. S. The isolation and characterization of foot-and-mouth disease virus from clinically normal herds of cattle in Botswana. *J. Hyg. (Lond).* **66,** 27–36 (1968).

259. Stenfeldt, C. *et al.* The Foot-and-Mouth Disease Carrier State Divergence in Cattle. *J. Virol.* **90,** 6344–6364 (2016).

260. Grubman, M. J., Moraes, M. P., Diaz-San Segundo, F., Pena, L. & de los Santos, T. Evading the host immune response: how foot-and-mouth disease virus has become an effective pathogen. *FEMS Immunol. Med. Microbiol.* **53,** 8–17 (2008).

261. Sobrino, F. & Domingo, E. *Foot and mouth disease : current perspectives*. (Horizon Bioscience, 2004).

262. Bengis, R. G., Thomson, G. R., Hedger, R. S., De Vos, V. & Pini, A. Foot-and-mouth disease and the African buffalo (Syncerus caffer). 1. Carriers as a source of infection for cattle. *Onderstepoort J. Vet. Res.* **53,** 69–73 (1986).

263. Salt, J. S. The carrier state in foot and mouth disease-an immunological review. *Br. Vet. J.* **149,** 207–223 (1993).

264. Bao, H.-F. *et al.* The infectivity and pathogenicity of a foot-and-mouth disease

virus persistent infection strain from oesophageal-pharyngeal fluid of a Chinese cattle in 2010. *Virol. J.* **8,** 536 (2011).

265. Bastos, A. D., Boshoff, C. I., Keet, D. F., Bengis, R. G. & Thomson, G. R. Natural transmission of foot-and-mouth disease virus between African buffalo (Syncerus caffer) and impala (Aepyceros melampus) in the Kruger National Park, South Africa. *Epidemiol. Infect.* **124,** 591–8 (2000).

266. Bertram, M. R. *et al.* Lack of Transmission of Foot-and-Mouth Disease Virus From Persistently Infected Cattle to Naïve Cattle Under Field Conditions in Vietnam. *Front. Vet. Sci.* **5,** 174 (2018).

267. Vosloo, W., Bastos, A. D., Sangare, O., Hargreaves, S. K. & Thomson, G. R. Review of the status and control of foot and mouth disease in sub-Saharan Africa. *Rev Sci Tech* **21,** 437–449 (2002).

268. Knowles, N. J., Wadsworth, J., Hammond, J. M. & King, D. P. Foot-and-Mouth Disease Virus Genotype Definitions and Nomenclature.

269. Paton, D. J., Sumption, K. J. & Charleston, B. Options for control of foot-and-mouth disease: knowledge, capability and policy. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **364,** 2657–67 (2009).

270. Casey-Bryars, M. *et al.* Waves of endemic foot-and-mouth disease in eastern Africa suggest feasibility of proactive vaccination approaches. *Nat. Ecol. Evol.* 1 (2018). doi:10.1038/s41559-018-0636-x

271. Hall, M. D., Knowles, N. J. & Wadsworth, J. Reconstructing Geographical Movements and Host Species Transitions of Foot-and-Mouth Disease Virus Serotype SAT 2. **4,** 1–10 (2013).

272. Sangare, O., A. D. Bastos, O. Marquardt, E. H. Venter, W. Vosloo, andG. R. T. Molecular epidemiology of serotype O foot-and mouth disease virus with emphasis on West and South Africa. *Virus Genes 22* **22,** 345–351 (2001).

273. Fèvre, E. M., Bronsvoort, B. M. de C., Hamilton, K. A. & Cleaveland, S. Animal movements and the spread of infectious diseases. *Trends Microbiol.* **14,** 125–131 (2006).

274. Vosloo, W., Bastos, A. D. S., Sangare, O., Hargreaves, S. K. & Thomson, G. R. Review of the status and control of foot and mouth disease in sub-Saharan Africa. *Rev. Sci. Tech.* **21,** 437–49 (2002).

275. Molla, B. *et al.* Epidemiological Study on Foot-and-Mouth Disease in Cattle: Seroprevalence and Risk Factor Assessment in South Omo Zone, South-western Ethiopia. *Transbound. Emerg. Dis.* **57,** 340–347 (2010).

276. Vosloo, W., Thompson, P. N., Botha, B., Bengis, R. G. & Thomson, G. R. Longitudinal Study to Investigate the Role of Impala (Aepyceros melampu) in Foot-and-Mouth Disease Maintenance in the Kruger National Park, South Africa. *Transbound. Emerg. Dis.* **56,** 18–30 (2009).

277. Bengis, R. G., Thomson, G. R. & Keet, D. F. Foot and mouth disease in impala (Aepyceros melampus). in *Proceedings of the O.I.E. Scientific Conference on the Control of Foot and Mouth Disease, African Horse Sickness and Contagious Bovine Pleuropneumonia. Gabarone, Botswana,* 13–14 (2014).

278. Miguel, E. *et al.* Contacts and foot and mouth disease transmission from wild to domestic bovines in Africa. *Ecosphere* **4,** 1–32 (2013).

279. Allepuz, A. *et al.* Risk factors for foot-and-mouth disease in Tanzania, 2001-2006. *Transbound. Emerg. Dis.* **62,** 127–136 (2015).

280. Hamoonga, R., Stevenson, M. A., Allepuz, A., Carpenter, T. E. & Sinkala, Y. Risk factors for foot-and-mouth disease in Zambia, 1981-2012. *Prev. Vet. Med.* **114,** 64–71 (2014).

281. De Maio, N., Wu, C.-H. & Wilson, D. J. SCOTTI: Efficient Reconstruction of Transmission within Outbreaks with the Structured Coalescent. *PLoS Comput. Biol.* **12,** e1005130 (2016).

282. Eblé, P., De Koeijer, A., Bouma, A., Stegeman, A. & Dekker, A. Quantification of within- and between-pen transmission of foot-and-mouth disease virus in pigs. *Vet. Res.* **37,** 647–654 (2006).

283. Myaing, M. Z., Jumat, M. R., Huong, T. N., Tan, B. H. & Sugrue, R. J. Truncated forms of the PA protein containing only the C-terminal domains are associated with the ribonucleoprotein complex within H1N1 influenza virus particles. doi:10.1099/jgv.0.000721

284. Varga, Z. T. & Palese, P. The influenza A virus protein PB1-F2: killing two birds with one stone? *Virulence* **2,** 542–6 (2011).

285. Breban, R., Drake, J. M., Stallknecht, D. E. & Rohani, P. The Role of Environmental Transmission in Recurrent Avian Influenza Epidemics. *PLoS Comput. Biol.* **5,** (2009).

286. Bloomquist, E. W., Lemey, P. & Suchard, M. a. Three roads diverged? Routes to phylogeographic inference. *Trends Ecol. Evol.* **25,** 626–632 (2010).

287. Dugan, V. G. *et al.* The evolutionary genetics and emergence of avian influenza viruses in wild birds. *PLoS Pathog.* **4,** (2008).

288. Lu, L., Lycett, S. J. & Leigh Brown, A. J. Reassortment patterns of avian influenza virus internal segments among different subtypes. *BMC Evol. Biol.* **14,** 1–15 (2014).

289. Huang, K. *et al.* Establishment and lineage replacement of H6 influenza viruses in domestic ducks in southern China. *J. Virol.* **86,** 6075–83 (2012).

290. Shirleen Soh, Y., Moncla, L. H., Eguia, R., Bedford, T. & Bloom, J. D. Comprehensive mapping of adaptation of the avian influenza polymerase protein PB2 to humans. *Elife* **8,** (2019).

291. Volz, E. M., Koelle, K. & Bedford, T. Viral phylodynamics. *PLoS Comput. Biol.* **9,** e1002947 (2013).

292. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22,** 30494 (2017).

293. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* (2017). doi:10.1093/bib/bbx108

294. Tamura, K. *et al.* MEGA5 : Molecular Evolutionary Genetics Analysis Using Maximum Likelihood , Evolutionary Distance , and Maximum Parsimony Methods. **28,** 2731–2739 (2011).

295. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29,** 1969–73 (2012).

296. Ayres, D. L. *et al.* BEAGLE: An application programming interface and high-performance computing library for statistical phylogenetics. *Syst. Biol.* **61,** 170–173 (2012).

297. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* **67,** 901–904 (2018).

298. Pons, P. & Latapy, M. Computing Communities in Large Networks Using Random Walks. *J. Graph Algorithms Appl.* **10,** 191–218 (2006).

299. Rambaut, A. FigTree. (2009). Available at: http://tree.bio.ed.ac.uk/software/figtree/. (Accessed: 25th April 2018)

300. Bouckaert, R. *et al.* BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput. Biol.* **10,** e1003537 (2014).

301. Bielejec, F. *et al.* SpreaD3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes. *Mol. Biol. Evol.* **33,** 2167–2169 (2016).

302. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13,** 2498–2504 (2003).

303. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35,** 526–528 (2019).

304. Gilbert, M. *et al.* Flying Over an Infected Landscape: Distribution of Highly Pathogenic Avian Influenza H5N1 Risk in South Asia and Satellite Tracking of Wild Waterfowl. *Ecohealth* **7,** 448–458 (2010).

305. Offeddu, V., Cowling, B. J. & Peiris, J. S. M. Interventions in live poultry markets for the control of avian influenza: A systematic review. *One Health* **2,** 55–64 (2016).

306. Wu, T. & Perrings, C. The live poultry trade and the spread of highly pathogenic avian influenza: Regional differences between Europe, West Africa, and Southeast Asia. *PLoS One* **13,** e0208197 (2018).

307. Wang, G. *et al.* H5N1 avian influenza re-emergence of Lake Qinghai: Phylogenetic and antigenic analyses of the newly isolated viruses and roles of migratory birds in virus circulation. *J. Gen. Virol.* **89,** 697–702 (2008).

308. Cappelle, J. *et al.* Risks of avian influenza transmission in areas of intensive free-ranging duck production with wild waterfowl. *Ecohealth* **11,** 109–119 (2014).

309. Bergervoet, S. A. *et al.* Circulation of low pathogenic avian influenza (LPAI) viruses in wild birds and poultry in the Netherlands, 2006–2016. *Sci. Rep.* **9,** 1–12 (2019).

310. Lycett, S. J., Duchatel, F. & Digard, P. A brief history of bird flu. *Philosophical Transactions of the Royal Society B: Biological Sciences* **374,** 20180257–20180257 (2019).

311. Lycett, S. J. *et al.* Genesis and spread of multiple reassortants during the 2016/2017 H5 avian influenza epidemic in Eurasia. *Proc. Natl. Acad. Sci. U. S. A.* **117,** 20814–20825 (2020).

312. Belsham, G. J. Distinctive features of foot-and-mouth disease virus, a member of the picornavirus family; aspects of virus protein synthesis, protein processing and structure. *Prog. Biophys. Mol. Biol.* **60,** 241–260 (1993).

313. Robinson, T. P., Food and Agriculture Organization of the United Nations., International Livestock Research Institute. & Pro-Poor Livestock Policy Initiative. *Global livestock production systems*. (Food and Agriculture Organization of the United Nations, 2011).

314. Motta, P. *et al.* Implications of the cattle trade network in Cameroon for regional disease prevention and control. *Sci. Rep.* **7,** (2017).

315. Bronsvoort, B. M. D. C. *et al.* Geographical and age-stratified distributions of foot-and-mouth disease virus-seropositive and probang-positive cattle herds in the Adamawa province of Cameroon. *Vet. Rec.* **159,** 299–308 (2006).

316. Wungak, Y. S., Olugasa, B. O., Ishola, O. O., Lazarus, D. D. & Ularamu, G. H. Foot-and-mouth disease (FMD) prevalence and exposure factors associated with seropositivity of cattle in north-central, Nigeria. *African J. Biotechnol.* **15,** 1224–1232 (2016).

317. Dean, A. S. *et al.* Potential Risk of Regional Disease Spread in West Africa through Cross-Border Cattle Trade. *PLoS One* **8,** 1–9 (2013).

318. Bessell, P. R., Shaw, D. J., Savill, N. J. & Woolhouse, M. E. Geographic and topographic determinants of local FMD transmission applied to the 2001 UK FMD epidemic. *BMC Vet. Res.* **4,** 40 (2008).

319. Flood, J. S., Porphyre, T., Tildesley, M. J. & Woolhouse, M. E. The performance of approximations of farm contiguity compared to contiguity defined using detailed geographical information in two sample areas in Scotland: implications for foot-and-mouth disease modelling. *BMC Vet. Res.* **9,** 198 (2013).

320. Cottam, E. M. *et al.* Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc. Biol. Sci.* **275,** 887–895 (2008).

321. Dellicour, S., Rose, R. & Pybus, O. G. Explaining the geographic spread of emerging epidemics: a framework for comparing viral phylogenies and environmental landscape data. *BMC Bioinformatics* **17,** 82 (2016).

322. Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol.* **1,** vev003 (2015).

323. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7,** 214 (2007).

324. Tavare, S. Some probabilistic and statistical problems in the analysis of dna sequences. *Lect. Math. Life Sci. (American Math. Soc.* **17,** 57–86 (1986).

325. Kahle, D. & Wickham, H. ggmap: Spatial Visualization with ggplot2. *R J.* 144–161 (2013).

326. Uchida, H. & Nelson, A. Agglomeration Index: Towards a New Measure of Urban Concentration. in *Urbanization and Development: Multidisciplinary Perspectives* (2011). doi:10.1093/acprof:oso/9780199590148.003.0003

327. Tully, D. C. & Fares, M. A. The tale of a modern animal plague: Tracing the evolutionary history and determining the time-scale for foot and mouth disease virus. *Virology* **382,** 250–256 (2008).

328. Sangula, A. K. *et al.* Evolutionary analysis of foot-and-mouth disease virus serotype SAT 1 isolates from east africa suggests two independent introductions from southern africa. *BMC Evol. Biol.* **10,** 371 (2010).

329. Mack, R. The great African cattle plague epidemic of the 1890's. *Trop. Anim. Health Prod.* (1970). doi:10.1007/BF02356441

330. Knowles, N. J. FMD history. (1990). Available at: http://www.picornaviridae.com/aphthovirus/fmdv/fmd_history.htm. (Accessed: 25th May 2018)

331. Perry, B. D. *et al. The impact and poverty reduction implications of foot and mouth disease control in southern Africa with special reference to Zimbabwe*. (International Livestock Research Institute (ILRI), 2003).

332. Maleko, D. D., Mbassa, G. N., Maanga, W. F. & Sisya, E. S. Impacts of Wildlife-Livestock Interactions in and around Arusha National Park, Tanzania. *Curr. Res. J. Biol. Sci.* **4,** 471–476 (2012).

333. Bartley, L. M., Donnelly, C. A. & Anderson, R. M. Review of foot-and-mouth disease virus survival in animal excretions and on fomites. *Vet. Rec.* **151,** 667–9 (2002).

334. Robinson, S. E. & Christley, R. M. Exploring the role of auction markets in cattle movements within Great Britain. *Prev. Vet. Med.* **81,** 21–37 (2007).

335. Brito, B. P. *et al.* Transmission of foot-and-mouth disease SAT2 viruses at the wildlife-livestock interface of two major transfrontier conservation areas in Southern Africa. *Front. Microbiol.* **7,** 528 (2016).

336. Miguel, E. *et al.* Drivers of foot-and-mouth disease in cattle at wild/domestic interface: Insights from farmers, buffalo and lions. *Divers. Distrib.* **23,** 1018–1030 (2017).

337. Duchatel, F., Bronsvoort, M. & Lycett, S. Phylogeographic analysis and identification of factors impacting the diffusion of Foot-and-Mouth disease virus in Africa. *bioRxiv* 358044 (2018). doi:10.1101/358044

338. Casey, M. B. *et al.* Patterns of Foot-and-Mouth Disease Virus Distribution in Africa: The Role of Livestock and Wildlife in Virus Emergence. *Role Anim. Emerg. Viral Dis.* 21–38 (2014). doi:10.1016/B978-0-12-405191-1.00002-8

339. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of

phylogenetic trees. *Bioinformatics* **17,** 754–755 (2001).

340. Baele, G., Suchard, M. A., Rambaut, A. & Lemey, P. Emerging concepts of data integration in pathogen phylodynamics. *Syst. Biol.* **66,** e47–e65 (2017).

341. Ruchusatsawat, K. *et al.* Long-term circulation of Zika virus in Thailand: an observational study. *Lancet Infect. Dis.* **19,** 439–446 (2019).

342. Lu, L. *et al.* Why is fruit colour so variable? Phylogenetic analyses reveal relationships between fruit-colour evolution, biogeography and diversification. *Glob. Ecol. Biogeogr.* **28,** 891–903 (2019).

343. Ling, J. *et al.* The introduction and dispersal of Sindbis virus from central Africa to Europe. *J. Virol.* JVI.00620-19 (2019). doi:10.1128/JVI.00620-19

344. Hanke, K. *et al.* Reconstruction of the Genetic History and the Current Spread of HIV-1 Subtype A in Germany. *J. Virol.* **93,** e02238-18 (2019).

345. Groendyke, C. & Welch, D. Epinet: An R Package to Analyze Epidemics Spread across Contact Networks. *J. Stat. Softw.* **83,** 1–22 (2018).

346. Bielejec, F. *et al. SOFTWARE Open Access π BUSS: a parallel BEAST/BEAGLE utility for sequence simulation under complex evolutionary scenarios*. *BMC Bioinformatics* **15,** (2014).

347. Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **20,** 832–844 (1998).

348. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7,** 214 (2007).

349. Bouckaert, R. *et al.* BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput. Biol.* **10,** e1003537 (2014).

350. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22,** 160–74 (1985).

351. Graham, C. H., Storch, D. & Machac, A. Phylogenetic scale in ecology and evolution. *Glob. Ecol. Biogeogr.* **27,** 175–187 (2018).

352. Cleaveland, S., Haydon, D. T. & Taylor, L. Overviews of pathogen emergence: which pathogens emerge, when and why? *Curr. Top. Microbiol. Immunol.* **315,** 85–111 (2007).

353. Wakeley, J. & Aliacar, N. Gene genealogies in a metapopulation. *Genetics* **159,** 893–905 (2001).

354. Dearlove, B., Wilson, D. J., B, P. T. R. S., Dearlove, B. & Wilson, D. J. Coalescent inference for infectious disease : meta-analysis of hepatitis C Coalescent inference for infectious disease : meta-analysis of hepatitis C Author for correspondence : (2013).

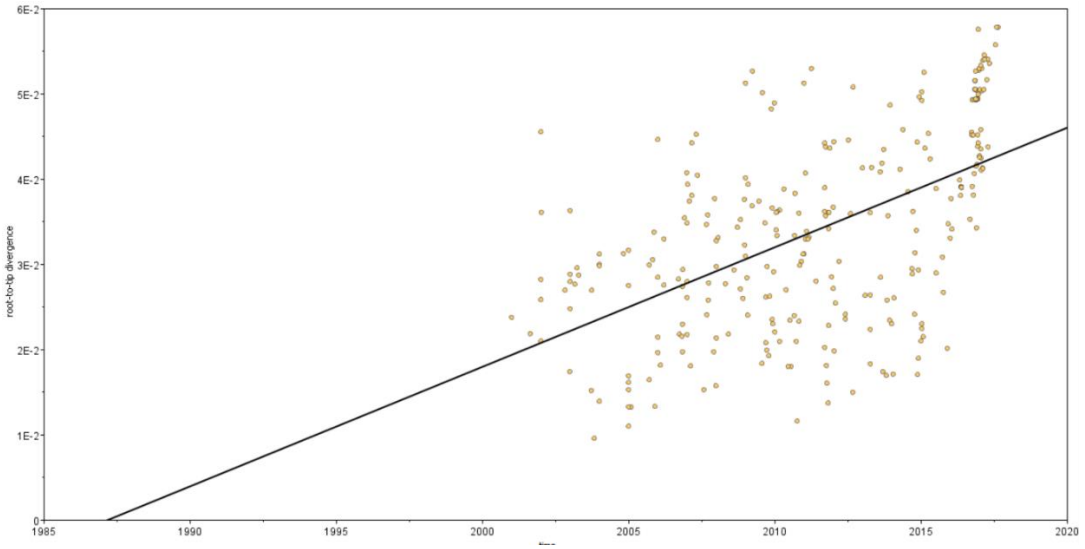
355. Klinkenberg, D., Backer, J. A., Didelot, X., Colijn, C. & Wallinga, J. *Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks*. *PLoS Computational Biology* **13,** (2017).

356. Volz, E. M., Romero-Severson, E. & Leitner, T. Phylodynamic Inference across Epidemic Scales. *Mol. Biol. Evol.* **34,** 1276–1288 (2017).

357. Goodin, D. G., Jonsson, C. B., Allen, L. J. S. & Owen, R. D. Integrating Landscape Hierarchies in the Discovery and Modeling of Ecological Drivers of Zoonotically Transmitted Disease from Wildlife. in 299–317 (Springer, Cham, 2018). doi:10.1007/978-3-319-92373-4_9

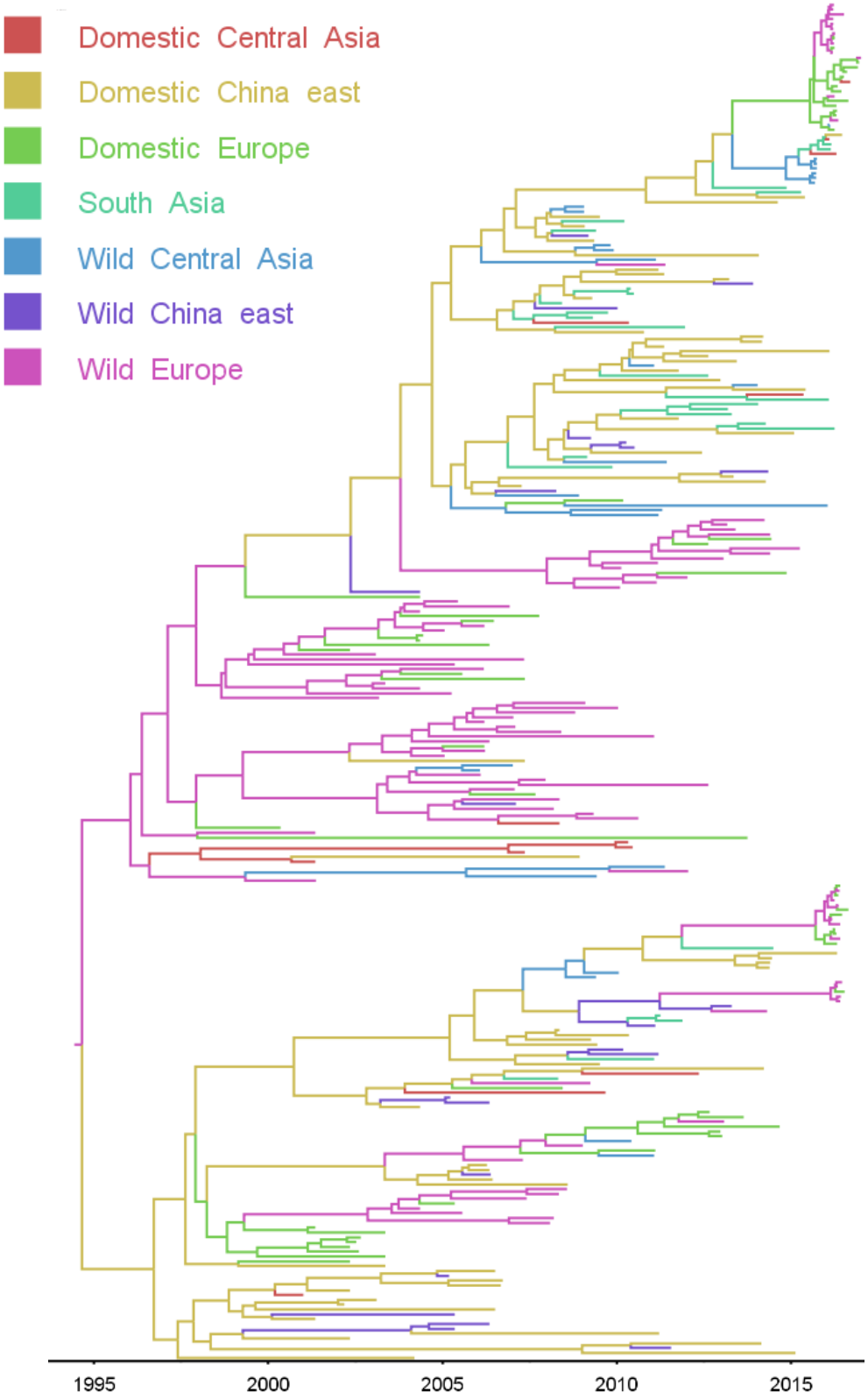
8 SUPPLEMENTARY MATERIAL

8.1 SPATIAL, SPECIES AND SUBTYPE NETWORK RECONSTRUCTIONS FOR AVIAN INFLUENZA VIRUSES BETWEEN EUROPE AND ASIA



Supplementary figure 8-1: Root-to-tip divergence as a function of sampling time for the avian influenza dataset

8.1. Spatial, Species and Subtype network reconstructions for Avian influenza viruses between Europe and Asia



Supplementary figure 8-2:Bayesian MCC time scaled discrete phylogeographic tree of 282 PB2 avian influenza sequences collected in Eurasia from 2001 to 2017. The phylogeny branches are coloured according to a join trait representing the host and location. The key for the graph is shown on the left.

Supplementary table 8-1: Number of sequences used in the analysis splitter by location and host type

Location	Host	Number sequences
Central Asia		**20**
	Domestic Anseriformes	3
	Domestic Galliformes	5
	Wild Anseriformes	11
	Wild others	1
East China		**48**
	Domestic Anseriformes	33
	Domestic Galliformes	8
	Wild Anseriformes	6
	Wild others	1
Eastern Asia		**32**
	Domestic Anseriformes	13
	Domestic Galliformes	4
	Wild Anseriformes	13
	Wild others	2
Eastern Europe		**36**
	Domestic Anseriformes	6
	Domestic Galliformes	7
	Wild Anseriformes	19
	Wild Charadriiforms	2
	Wild others	2
North-Central Asia		**20**
	Domestic Anseriformes	5
	Wild Anseriformes	10
	Wild Charadriiforms	3

	Wild others	2
Northern Europe		**18**
	Domestic Anseriformes	2
	Domestic Galliformes	6
	Wild Anseriformes	8
	Wild Charadriiforms	2
South Asia		**26**
	Domestic Anseriformes	26
Southern Europe		**29**
	Domestic Anseriformes	4
	Domestic Galliformes	10
	Wild Anseriformes	15
Western Europe		**53**
	Domestic Anseriformes	10
	Domestic Galliformes	9
	Wild Anseriformes	24
	Wild Charadriiforms	8
	Wild others	2

Supplementary table 8-2: Root-to-tips parameters for the avian influenza dataset composed of 282 avian influenza PB2 sequences collected in Eurasia from 2001 to 2017

Parameter	**Value**
Date range	16.646
Slope (rate)	1.40E-03
X-Intercept (TMRCA)	1987.2266
Correlation Coefficient	0.5477
R squared	0.2999
Residual Mean Squared	9.17E-05

Supplementary table 8-3: Estimated evolutionary parameters using 282 avian influenza PB2 sequences collected in Eurasia from 2001 to 2017.

Parameter	Estimation
Root height	22.836 ± 1.93
Mean mutation rate (substitutions/site/year)	0.0039 ±3e-4

Supplementary table 8-4: Output of the BSSVS analysis for the avian influenza dataset showing the best supported rates of transition between the sampled locations

Origin	Destination	Bayes factor	Posterior probability
East China	Eastern Asia	7464	1.00
Southern Europe	Western Europe	7464	1.00
Western Europe	Eastern Europe	7464	1.00
Western Europe	Northern Europe	7464	1.00
East China	North Central Asia	491	0.99
Eastern Europe	Southern Europe	318	0.98
South Asia	Eastern Asia	163	0.96
South Asia	Central Asia	96	0.93
Eastern Europe	Central Asia	89	0.92
Southern Europe	East China	89	0.92
East China	South Asia	82	0.92
East China	Western Europe	20	0.73
Eastern Europe	Northern Europe	13	0.64

Supplementary table 8-5: Output of the BSSVS analysis for the avian influenza dataset showing the best supported rates of transition between the sampled hosts

Origin	Destination	Bayes factor	Posterior probability
Domestic Anseriformes	Wild Anseriformes	35225	1.00
Wild Anseriformes	Wild Charadriiforms	35225	1.00
Wild Anseriformes	Domestic Anseriformes	35225	1.00
Wild Anseriformes	Domestic Galliformes	35225	1.00
Wild Anseriformes	Wild others	2707	1.00
Domestic Anseriformes	Domestic Galliformes	1528	1.00
Domestic Galliformes	Domestic Anseriformes	54	0.94
Domestic Anseriformes	Wild others	16	0.83
Domestic Galliformes	Wild others	13	0.80

8.1. Spatial, Species and Subtype network reconstructions for Avian influenza viruses between Europe and Asia

Supplementary table 8-6: Markov jump analysis result for the discrete location analysis using the avian influenza dataset.

Origin	Destination	Number of jumps
Western Europe	Eastern Europe	18
Southern Europe	Western Europe	11
East China	South Asia	13
Western Europe	Northern Europe	13
East China	Eastern Asia	15
East China	North-Central Asia	8
Eastern Europe	Southern Europe	6
South Asia	Central Asia	5
Southern Europe	East China	5
South Asia	East China	5
South Asia	Eastern Asia	5
Western Europe	Central Asia	4
Western Europe	Southern Europe	4
Eastern Europe	Western Europe	3
Eastern Europe	Central Asia	3
East China	Western Europe	3
South Asia	North-Central Asia	3
East China	Central Asia	3
Southern Europe	Eastern Europe	3
North-Central Asia	Central Asia	2
Western Europe	East China	2
Eastern Asia	East China	2
Eastern Europe	Northern Europe	2
South Asia	Eastern Europe	1
Eastern Europe	North-Central Asia	1
Central Asia	South Asia	1
East China	Northern Europe	1
North-Central Asia	Eastern Asia	1
North-Central Asia	East China	1
Central Asia	Western Europe	1
North-Central Asia	Western Europe	1
East China	Eastern Europe	1
Northern Europe	Eastern Europe	1
Southern Europe	Northern Europe	1
Central Asia	Eastern Europe	1
North-Central Asia	South Asia	1
South Asia	Western Europe	1
Eastern Asia	South Asia	1
East China	Southern Europe	1
Western Europe	North-Central Asia	1

Supplementary table 8-7: Markov jump analysis result for the host analysis using the avian influenza dataset

Origin	Destination	Number of jumps
Domestique Anseriformes	Wild Anseriformes	35
Wild Anseriformes	Domestique Anseriformes	27
Wild Anseriformes	Domestique gal	19
Domestique Anseriformes	Domestique gal	15
Wild Anseriformes	Wild Charadriiform	12
Wild Anseriformes	Wild others	5
Domestique gal	Domestique Anseriformes	4
Domestique Anseriformes	Wild others	4
Domestique gal	Wild Anseriformes	3
Domestique gal	Wild others	2
Wild Charadriiform	Wild Anseriformes	1
Domestique Anseriformes	Wild Charadriiform	1
Wild other	Domestique Anseriformes	1

Supplementary table 8-8: population change estimation from wild to domestic birds within each one of the sampled locations.

Location	Mean	Sd	Median
East China	0.015803	0.009379	0.014299
Western Europe	0.012897	0.004676	0.012967
Southern Europe	0.010243	0.005437	0.010534
South Asia	0.002583	0.004403	0.000769
Northern Europe	0.002549	0.002842	0.002443
Eastern Asia	0.002288	0.003655	0.000876
Eastern Europe	0.001724	0.001418	0.00124
Central Asia	0.000596	0.001924	0
North-Central Asia	0.000476	0.00106	0

Supplementary table 8-9: population change estimation from domestic to wild birds within each one of the sampled locations.

Location	Mean	Sd	Median
East China	0.022154	0.007268	0.021163
Southern Europe	0.008017	0.005911	0.006639
Western Europe	0.006857	0.005572	0.005485
Eastern Asia	0.005268	0.00307	0.004612
Central Asia	0.003934	0.005596	0.000825
North-Central Asia	0.001562	0.001974	0.001024
South Asia	0.001395	0.002263	0.000276

Eastern Europe	0.000631	0.001171	0.000284
Northern Europe	3.47E-05	0.000317	0

Supplementary table 8-10: Bird type responsible for the changes from a central Asian location to an Asian location.

Host	**Mean**	**Sd**	**Median**
Domestic Anseriformes	0.01107	0.006702	0.011098
Wild Anseriformes	0.000348	0.00093	0
Domestic Galliformes	5.63E-05	0.000363	0
Wild others	4.06E-07	9.26E-06	0
Wild Charadriiforms	1.72E-07	3.69E-06	0

Supplementary table 8-11: Bird type responsible for the changes from an Asian location to a central Asian location.

Host	**Mean**	**Sd**	**Median**
Domestic Anseriformes	0.022829	0.005036	0.022877
Wild Anseriformes	0.0085	0.007114	0.006341
Domestic Galliformes	0.000204	0.001246	0
Wild Charadriiforms	1.28E-05	0.00014	0
Wild others	1.13E-05	0.000145	0

Supplementary table 8-12: Bird type responsible for the changes from a central Asian location to an European location.

Host	**Mean**	**Sd**	**Median**
Wild Anseriformes	0.004318	0.003376	0.003823
Domestic Anseriformes	0.001094	0.00197	0
Domestic Galliformes	4.58E-05	0.000391	0
Wild Charadriiforms	0	0	0
Wild others	0	0	0

Supplementary table 8-13: Bird type responsible for the changes from a European location to a central Asian location.

Host		**Mean**	**Sd**	**Median**
Wild Anseriformes		0.014617	0.006066	0.014499
Domestic Anseriformes		0.001964	0.002986	0.000646
Domestic Galliformes		0.000595	0.001603	0.000334
Wild Charadriiforms		0	0	0
Wild others		0	0	0

Supplementary table 8-14: Bird type responsible for circulation of the virus within Asia.

Host	Mean	Sd	Median
Domestic Anseriformes	0.023909	0.010514	0.022851
Wild Anseriformes	0.005994	0.006637	0.003474
Domestic Galliformes	0.000273	0.001112	0
Wild Charadriiforms	1.66E-05	0.000171	0
Wild others	3.22E-06	4.53E-05	0

Supplementary table 8-15: Bird type responsible for circulation of the virus within Europe.

Host	Mean	Sd	Median
Wild Anseriformes	0.214868	0.022975	0.217485
Domestic Anseriformes	0.032613	0.01488	0.031851
Domestic Galliformes	0.023165	0.006257	0.022953
Wild Charadriiforms	0.004326	0.002257	0.004474
Wild others	0.000148	0.000266	3.60E-06

Supplementary table 8-16: Output of the BSSVS analysis for avia influenza dataset showing the best supported rates of transition between the joined trait representing both the host and location.

Origin	Destination	Bayes factor	Posterior probability
Domestic China east	Wild China east	4927	1.00
Domestic Europe	Wild Europe	4927	1.00
Wild Europe	Domestic Europe	4927	1.00
Domestic Europe	Wild Central Asia	285	0.98
Domestic China east	Wild Central Asia	165	0.97
Domestic Europe	Domestic Central Asia	144	0.96
Domestic China east	Domestic South Asia	102	0.95
Domestic South Asia	Domestic Central Asia	20	0.79
Wild Central Asia	Wild Europe	18	0.78
Domestic China east	Domestic Central Asia	15	0.74
South Asia	Domestic China east	13	0.72
Wild Europe	Domestic China east	12	0.70

8.1. Spatial, Species and Subtype network reconstructions for Avian influenza viruses between Europe and Asia

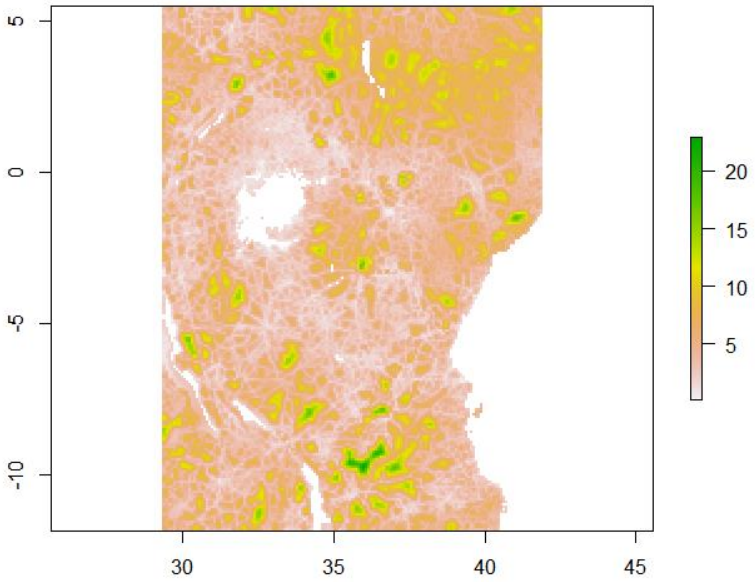
Supplementary table 8-17: Markov jump analysis result for the join trait analysis. The mean is the mean number of jumps observed from the origin trait to the destination trait.

Origin	Destination	Mean
Wild Europe	Domestic Europe	30
Domestic China east	Wild China east	13
Domestic Europe	Wild Europe	11
Domestic China east	South Asia	10
South Asia	Domestic China east	7
Wild Central Asia	Domestic China east	5
South Asia	Domestic Central Asia	5
Wild Europe	Wild Central Asia	5
Wild Europe	Wild China east	3
Domestic China east	Wild Central Asia	3
Domestic China east	Domestic Europe	3
Wild Central Asia	Wild Europe	3
Wild Central Asia	South Asia	2
Wild Europe	Domestic China east	2
Wild China east	Domestic Central Asia	2
Domestic Europe	Wild Central Asia	2
Wild China east	South Asia	1
Wild Europe	South Asia	1
Wild China east	Domestic China east	1
Wild Central Asia	Wild China east	1
Domestic Europe	Wild China east	1
Domestic China east	Domestic Central Asia	1
Wild Central Asia	Domestic Central Asia	1
Domestic Europe	Domestic Central Asia	1
Wild Europe	Domestic Central Asia	1
South Asia	Wild Central Asia	1
South Asia	Domestic Europe	1
Wild Central Asia	Domestic Europe	1

Supplementary table 8-18: Reassortment measure estimation for HA between the different join trait. The reassortment measure is the proportion of all reassortment taking place within a particular trait

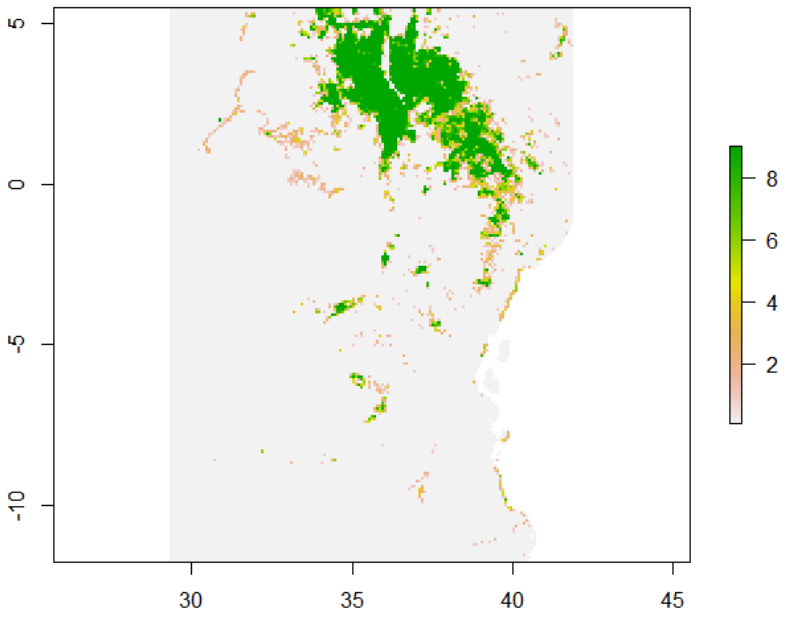
Join trait	Mean	Sd	Median
Domestic China east	0.130	0.025	0.132
Wild Europe	0.103	0.015	0.103
Domestic Europe	0.025	0.014	0.021
Wild Central Asia	0.016	0.009	0.015
South Asia	0.010	0.011	0.007
Domestic Central Asia	0.006	0.006	0.008
Wild China east	0.006	0.004	0.006

8.2 PHYLOGEOGRAPHIC ANALYSIS AND IDENTIFICATION OF FACTORS IMPACTING THE DIFFUSION OF FOOT-AND-MOUTH DISEASE VIRUS IN AFRICA

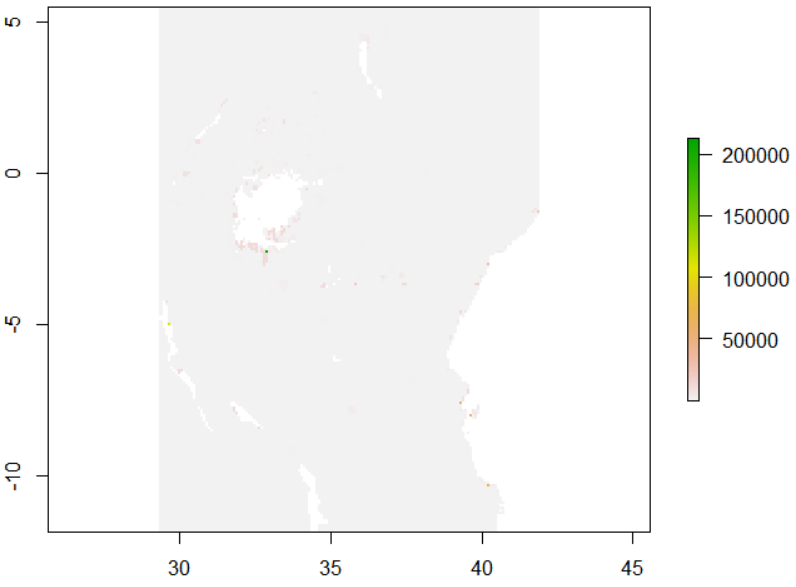


Supplementary figure 8-3: Accessibility raster used in the analysis. The colour code is proportional to the accessibility value of each cell.

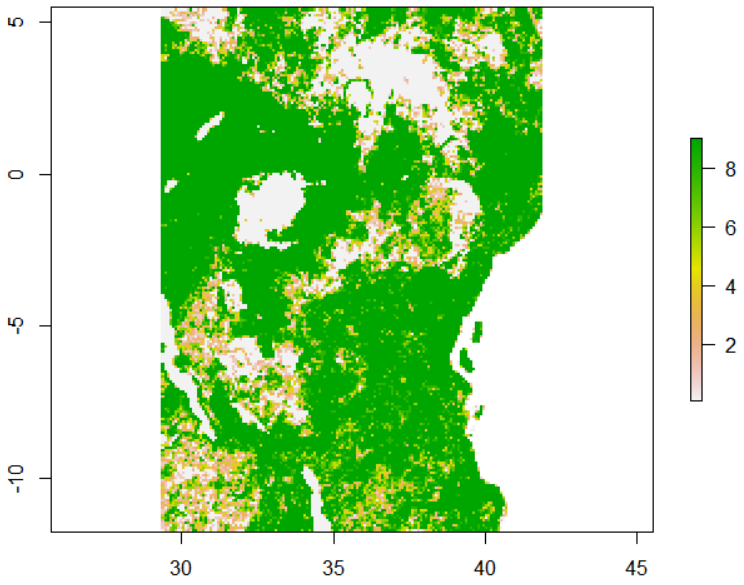
8.2. Phylogeographic analysis and identification of factors impacting the diffusion of Foot-and-Mouth disease virus in Africa



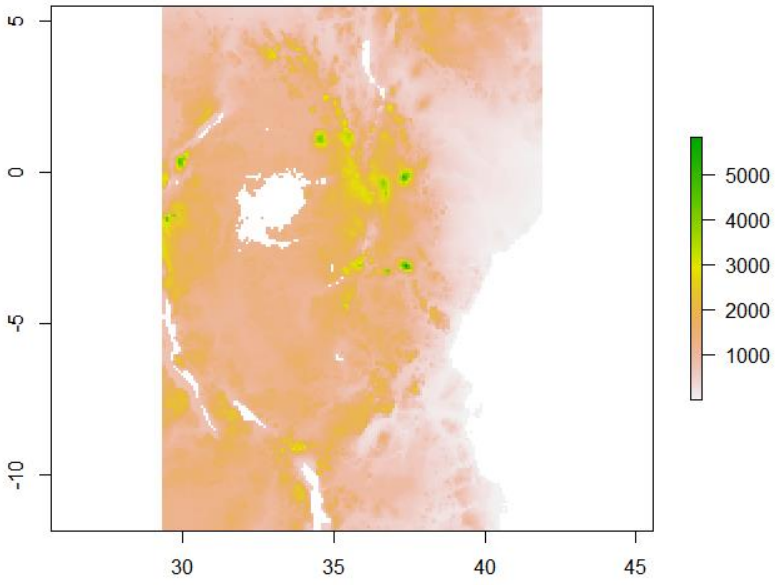
Supplementary figure 8-4: Bare area raster used in the analysis. The colour code is proportional to the number of cells that had a bare area value before the original raster aggregation.



Supplementary figure 8-5: Cattle density raster used in the analysis. The colour code is proportional to the average cattle density value of the cells before the original raster aggregation.

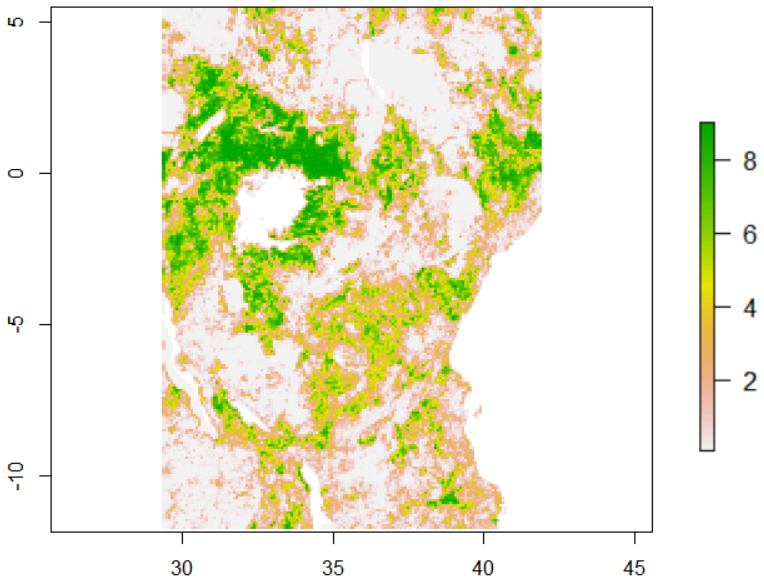


Supplementary figure 8-6 :Crop area raster used in the analysis. The colour code is proportional to the number of cells that had a bare crop value before the original raster aggregation.

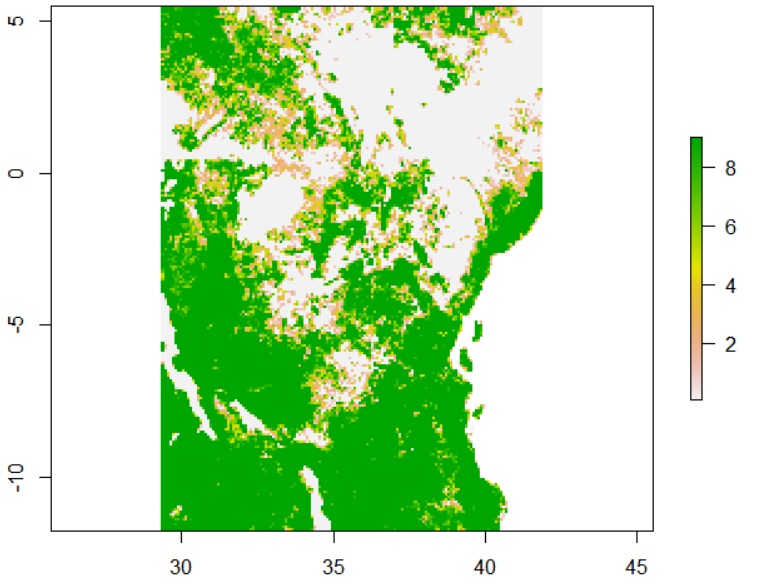


Supplementary figure 8-7: Elevation raster used in the analysis. The colour code is proportional to the average elevation value of the cells before the original raster aggregation.

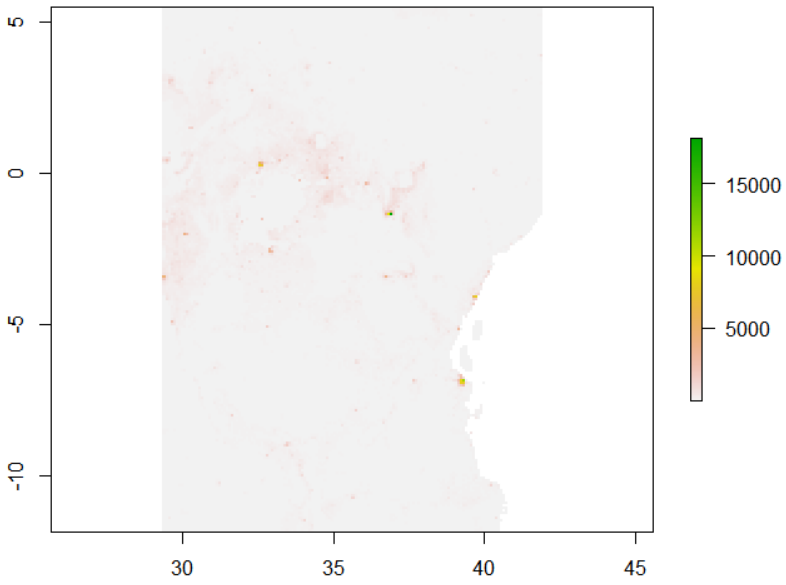
8.2. Phylogeographic analysis and identification of factors impacting the diffusion of Foot-and-Mouth disease virus in Africa



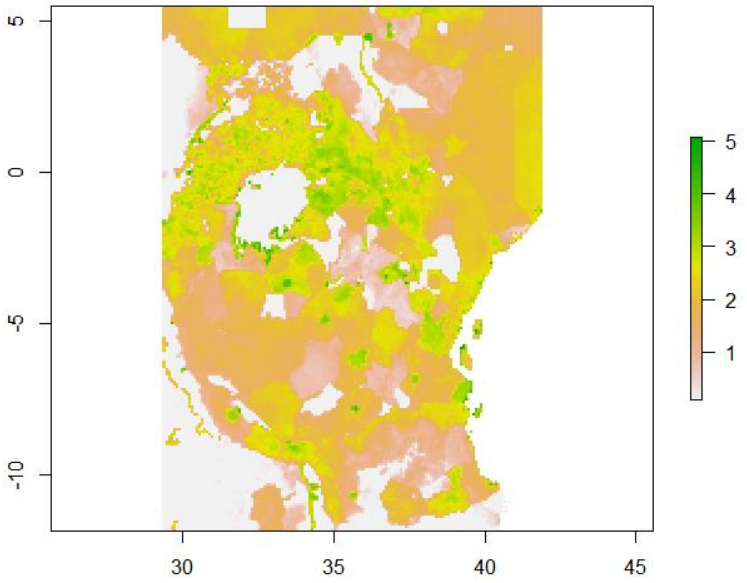
Supplementary figure 8-8: Fragmented crop raster used in the analysis. The colour code is proportional to the number of cells that had a fragmented crop value before the original raster aggregation.



Supplementary figure 8-9: Fragmented forest raster used in the analysis. The colour code is proportional to the number of cells that had a forest value before the original raster aggregation.

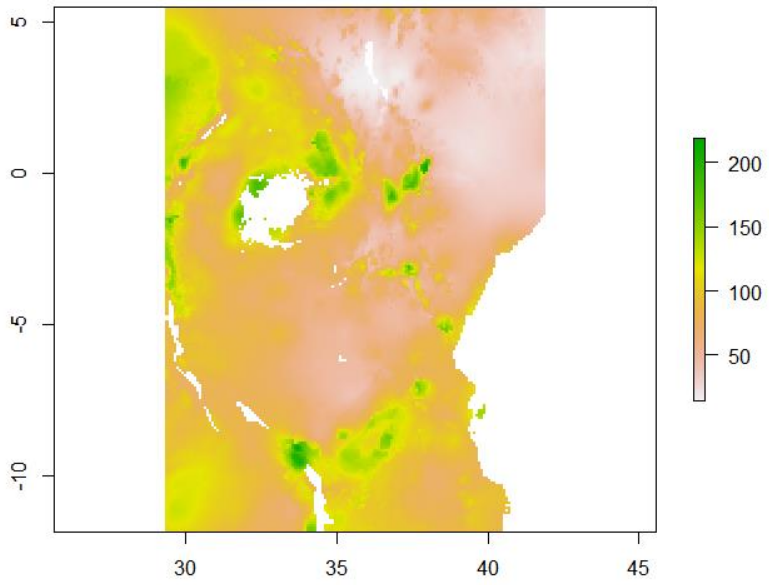


Supplementary figure 8-10 Human density raster used in the analysis. The colour code is proportional to the average human density value of the cells before the original raster aggregation.

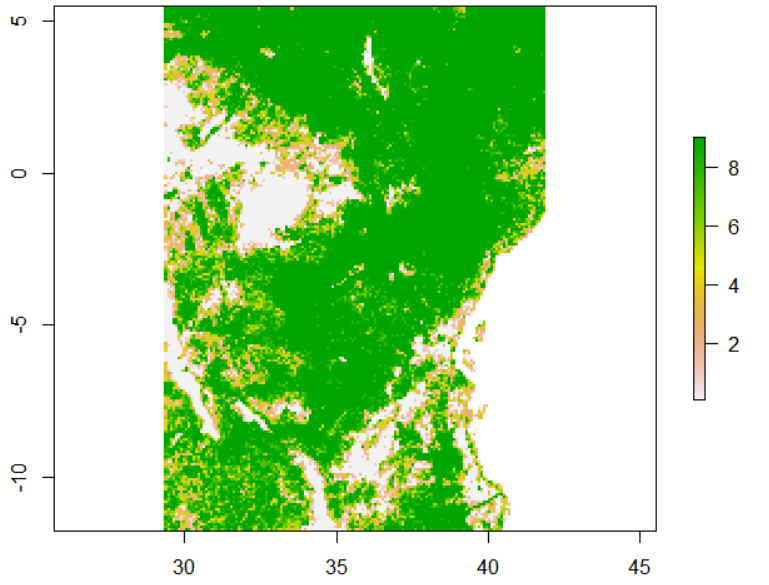


Supplementary figure 8-11: Logarithm of the cattle density raster used in the analysis. The colour code is proportional to the logarithm average cattle density value of the cells before the original raster aggregation.

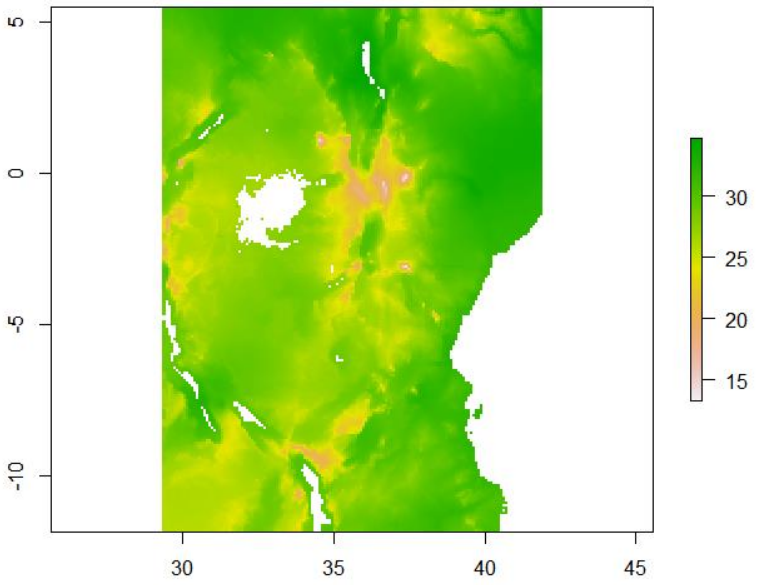
8.2. Phylogeographic analysis and identification of factors impacting the diffusion of Foot-and-Mouth disease virus in Africa



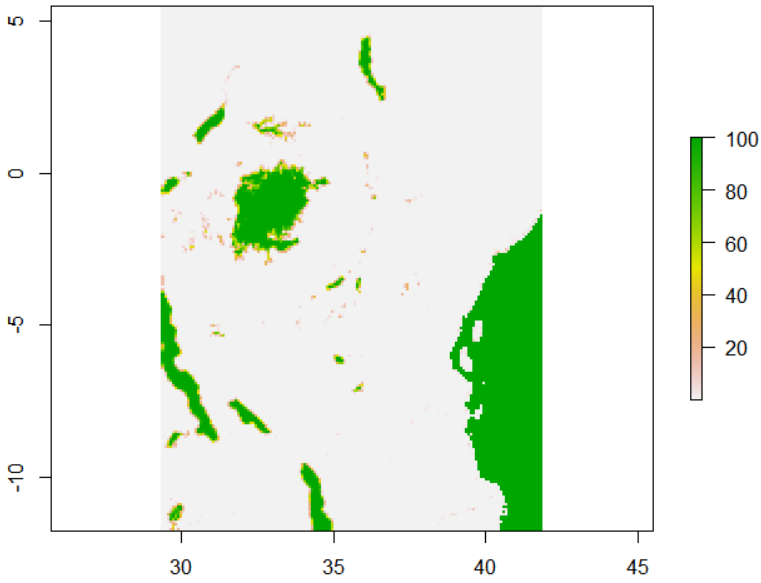
Supplementary figure 8-12: Precipitation raster used in the analysis. The colour code is proportional to the average precipitation value of the cells before the original raster aggregation.



Supplementary figure 8-13: Shrubland raster used in the analysis. The colour code is proportional to the number of cells that had a shrubland value before the original raster aggregation.

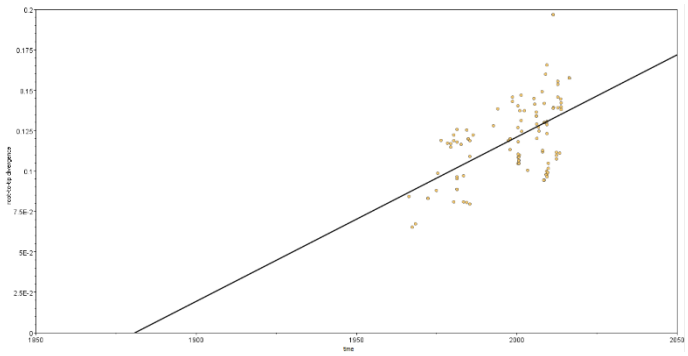


Supplementary figure 8-14: Temperature raster used in the analysis. The colour code is proportional to the average temperature value of the cells before the original raster aggregation.

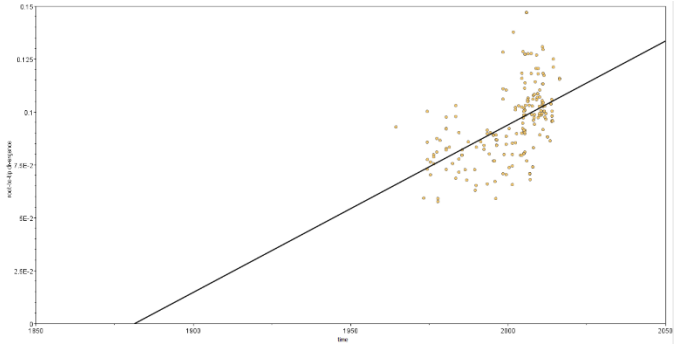


Supplementary figure 8-15: Water raster used in the analysis. The colour code is proportional to the number of cells that had a water value before the original raster aggregation.

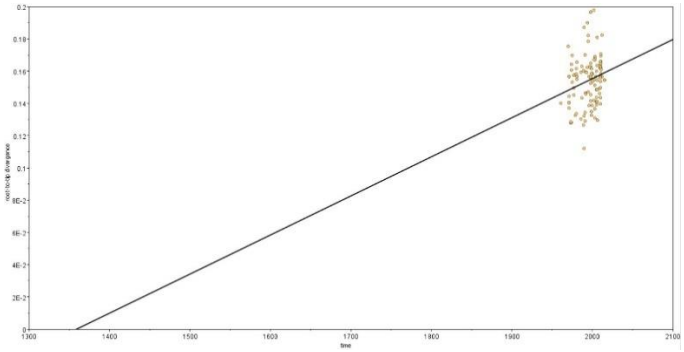
8.2. Phylogeographic analysis and identification of factors impacting the diffusion of Foot-and-Mouth disease virus in Africa



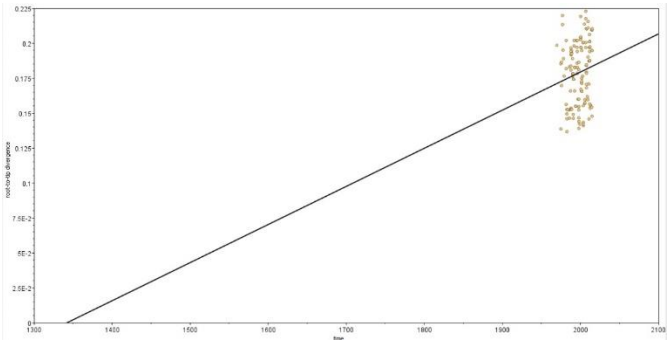
Supplementary figure 8-16: Root-to-tip divergence as a function of sampling time for the FMDA dataset



Supplementary figure 8-17: Root-to-tip divergence as a function of sampling time for the FMDV serotype O

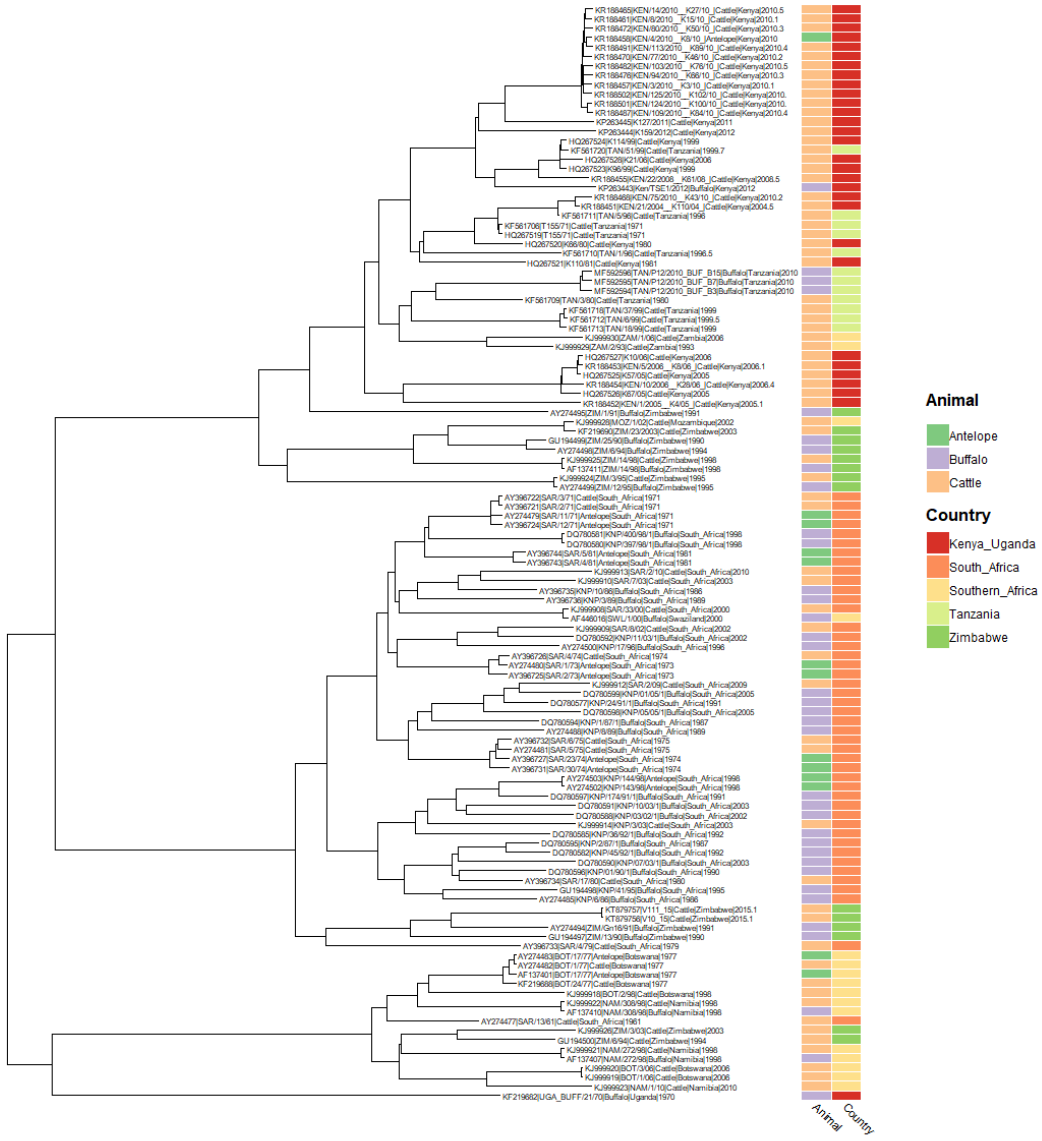


Supplementary figure 8-18: Root-to-tip divergence as a function of sampling time for the FMDV serotype SAT1

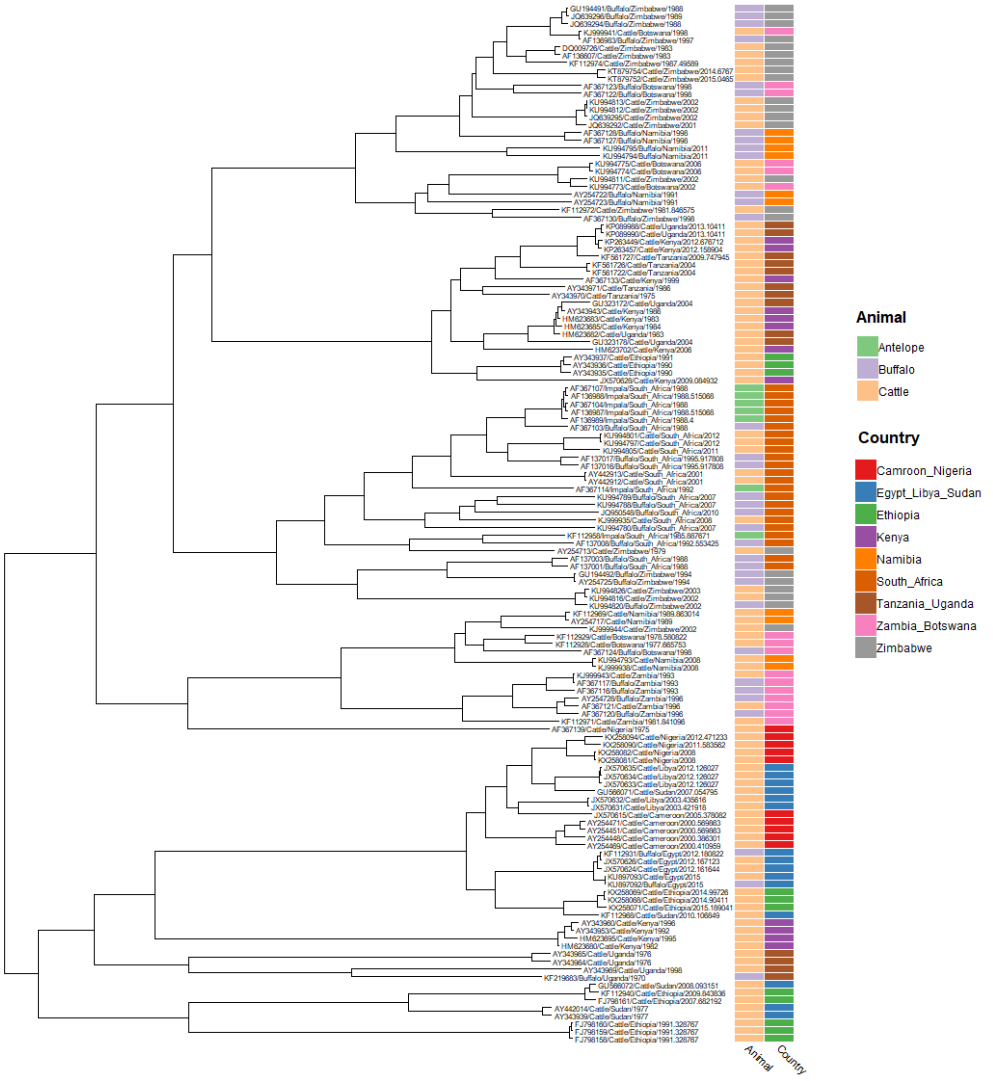


Supplementary figure 8-19: Root-to-tip divergence as a function of sampling time for the FMDV serotype SAT2

8.2. Phylogeographic analysis and identification of factors impacting the diffusion of Foot-and-Mouth disease virus in Africa

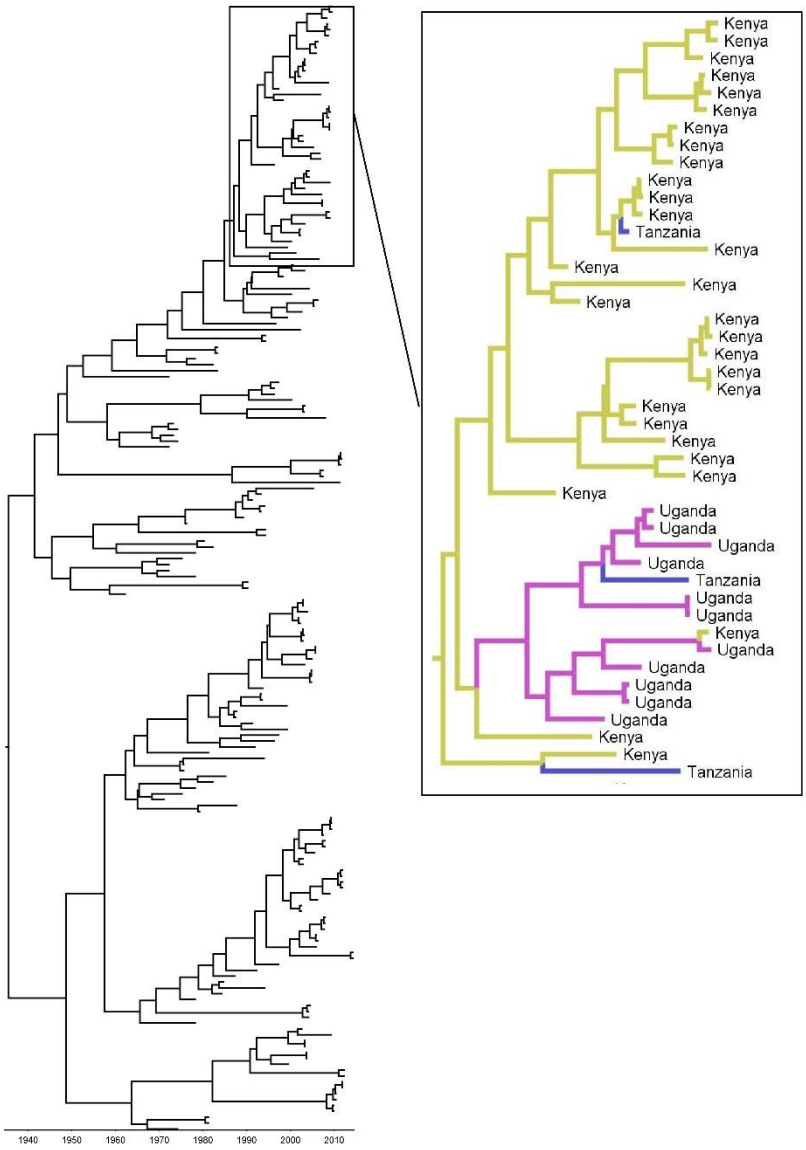


Supplementary figure 8-20: FMDV SAT1 tree annotated with the animal host and the country of sampling.

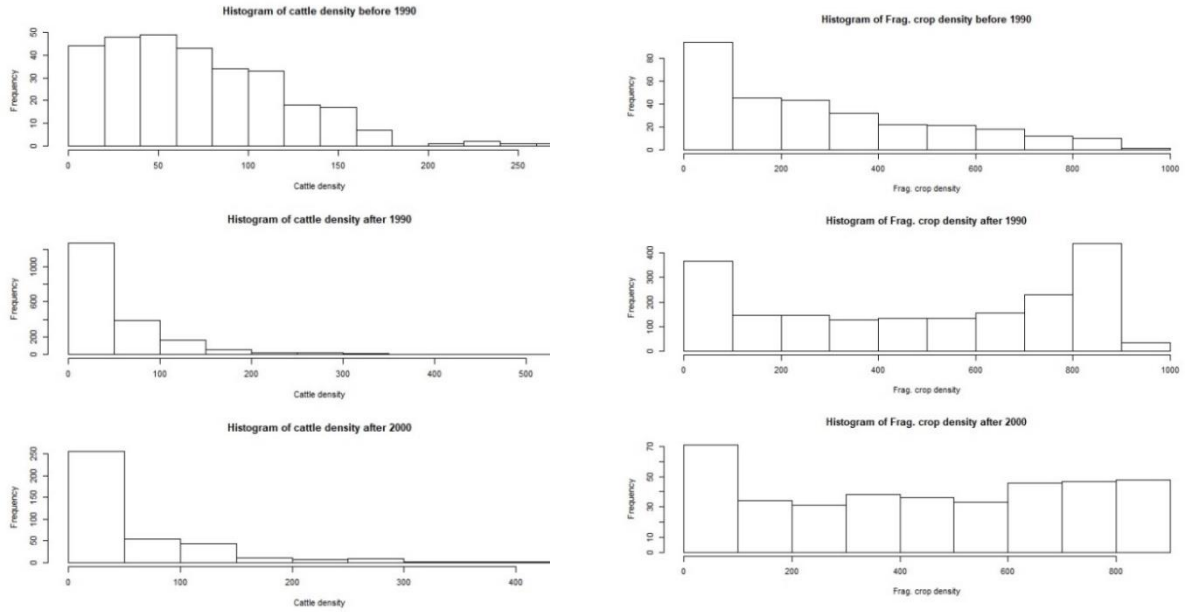


Supplementary figure 8-21: FMDV SAT2 tree annotated with the animal host and the country of sampling

8.2. Phylogeographic analysis and identification of factors impacting the diffusion of Foot-and-Mouth disease virus in Africa



Supplementary figure 8-22: Selected clade for the FMDO dataset



Supplementary figure 8-23: Histogram of different values of cattle densities and crop densities over the course of the virus the infection

Supplementary table 8-19: List of sequences analysed for FMDV A

Accession number	Country	Locality	Date
AY254411	Cameroon	Vina	2000.715
AY254415	Cameroon	Vina	2000.479
AY254419	Cameroon	Mbere	2000.463
AY254424	Cameroon	Djerem	2000.595
AY254425	Cameroon	Djerem	2000.595
AY254428	Cameroon	Vina	2000.715
AY254431	Cameroon	Djerem	2000.655
AY254432	Cameroon	Djerem	2000.595
AY254433	Cameroon	Vina	2000.816
AY254437	Cameroon	Faro et Deo	2000.488
AY254439	Cameroon	Vina	2001.564
AY254440	Cameroon		1975.414
AY254441	Cameroon		1986.414
AY254442	Cameroon		1985.414
AY254443	Cameroon		1976.414
EF208755	Cameroon	Adamawa	2000.414
EF208756	Egypt	Alexandria	1972.362
EF208757	Egypt	Ismailia	2006.107
EF208758	Egypt	Ismailia	2006.107

EF208762	Ethiopia	Tigray	2005.414
EF208763	Ethiopia	Dire Dawa	2000.414
EF208765	Ethiopia	Scena	1992.753
EF208766	Ethiopia	Highland area of Eastern	1994.088
EF208773	Kenya	Embu	2005.644
EF208774	Kenya	Meru	1998.685
EF208775	Kenya	Nakuru	1998.704
FJ798144	Ethiopia	Addis Ababa	1974.937
FJ798145	Ethiopia	Geferssa	1981.414
FJ798146	Ethiopia	Geferssa	1981.414
FJ798147	Ethiopia	Konso	2000.414
FJ798148	Ethiopia	kofele	2000.986
FJ798149	Ethiopia	Gobe	2002.414
FJ798150	Ethiopia	Adaba	2007.932
GU566065	Sudan	Blue Nile	1981.414
GU566066	Sudan		1982.66
GU566067	Sudan	Khartoum	1984.841
GU566068	Sudan	Soba	1984.929
GU566069	Sudan	Kasala State	2006.268
GU566070	Sudan	Gezira state	2006.315
JN680720	Nigeria	Barkin_Ladi	2009.414
KC888937	Egypt	Sharquia	2011.408
KC888938	Egypt	Menoufia	2011.411
KC888939	Egypt	Qaliubia	2011.414
KF112901	Egypt		1972.362
KF112902	Egypt	Ismailia Governorate	2006.107
KF112903	Ethiopia	Yabelo	2008.581
KF112912	Kenya	Olulunga	2009.162
KF112917	Sudan	Amara	2006.855
KF561687	Tanzania		1968.414
KF561688	Tanzania		1967.414
KF561690	Tanzania	Iringa Region	2008.581
KF561691	Tanzania	Iringa Region	2008.581
KF561692	Tanzania	Morogoro Region	2009.329
KF561693	Tanzania	Njombe District	2009.447
KF561694	Tanzania	Kibaha District	2009.468
KF561695	Tanzania	Mpwapwa	2009.682
KF561696	Tanzania	Iringa	2009.847
KF561697	Tanzania	Bagamoyo	2009.833
KF561700	Kenya	Wei	2008.107
KF561701	Kenya	Gathanji	2008.134

KF561702	Kenya	Loitokitok	2008.581
KF561703	Kenya	Olulunga	2009.162
KJ440839	Kenya	Kericho	1966.414
KJ440841	Kenya	Meru Central	1978.414
KJ440843	Kenya	Kiambu	1979.414
KJ440844	Kenya	Thika	1979.414
KJ440845	Kenya	Laikipia	1980.414
KJ440846	Kenya	Embu	1980.414
KJ440847	Kenya	Kajiado	1980.414
KJ440851	Kenya	Teso	1981.414
KJ440852	Kenya	Kilifi	1981.414
KJ440853	Kenya	Kwale	1981.414
KJ440854	Kenya	Kericho	1983.414
KJ440855	Kenya	Taita Taveta	1983.414
KJ440856	Kenya	Meru Central	1984.414
KJ440857	Kenya	Narok	1984.414
KJ440858	Kenya	Isiolo	1985.414
KJ440859	Kenya	Mombasa	1985.414
KJ440865	Kenya	Meru North	1997.414
KJ440866	Kenya	Meru Central	2001.414
KJ440867	Kenya	Nairobi	2001.414
KJ440868	Kenya	Kajiado	2003.414
KJ440870	Kenya	Loitoktok	2008.414
KJ440871	Kenya	Narok South	2009.414
KJ440873	Kenya	Naivasha	2012.414
KJ440874	Kenya	Nakuru North	2012.414
KJ440875	Kenya	Koibatek	2012.414
KJ440876	Kenya	Thika East	2013.414
KX258043	Eritrea	Binbina	1997.934
KX258044	Eritrea	Binbina	1998.038
KX258045	Eritrea	Binbina	1998.038
KX258046	Eritrea		2006.953
KX258049	Nigeria	Shendam	2009.036
KX258051	Nigeria	Yola	2009.414
KX258052	Nigeria	Yola	2009.414
KX258053	Nigeria	Lokoja	2011.482
KX258054	Nigeria	Barakin Ladi	2012.847
KX258055	Nigeria	Barakin Ladi	2012.847
KX258056	Nigeria	Barakin Ladi	2012.847
KX258057	Nigeria	Kaura	2012.866
KX258058	Nigeria	Barakin Ladi	2013.71
KX258059	Nigeria	Toro	2013.888

KX258061	Nigeria	Toro	2013.888
KX258062	Nigeria	Toro	2013.888
KX446997	Egypt	Isamilia	2016.414
KX446998	Egypt	Isamilia	2016.414
KX446999	Egypt	Isamilia	2016.414

Supplementary table 8-20: List of sequences for the FMDV serotype O

Accession number	**Country**	**Locality**	**Date**
AY283376	Ethiopia	Debre Zeit	1977.767
AY283377	Ethiopia	Neguele	1973.244
AY283379	Ethiopia	Nefas Silk Lafto	1989.644
AY283380	Ethiopia	West Hararge	1989.915
AY283382	Ethiopia	Kotebe	1989.918
AY283384	Ethiopia	Highland area	1994.088
AY283387	Ethiopia	Melge	1995.751
AY283388	Ethiopia	Boditi	1995.932
AY283389	Kenya		1996.222
AY283392	Ethiopia	Tigray	1996.222
AY283393	Ethiopia		2001.411
AY283395	Ethiopia		2001.416
AY344591	Kenya		1999.414
AY344593	Kenya		1998.416
AY344595	Somalia		1980.416
AY344596	Somalia		1983.414
AY344597	Somalia		1977.408
AY344598	Somalia		1981.419
AY344599	Sudan		1976.405
AY344600	Sudan		1974.4
AY344601	Sudan		1986.433
AY344602	Sudan		1974.4
AY344603	Sudan		1989.441
AY344604	Sudan		1980.416
AY344605	Kenya		1977.408
AY344606	Sudan		1983.425
AY344607	Sudan		1983.425
AY344608	Sudan		1980.416
AY344609	Sudan		1999.468
AY344611	Tanzania		1996.46
AY344612	Tanzania		1984.427
AY344613	Tanzania		1980.416

AY344614	Tanzania		1996.414
AY344615	Tanzania		1980.414
AY344616	Tanzania		1985.43
AY344618	Uganda		1996.416
AY344619	Uganda		1996.419
AY344620	Uganda		2005.471
AY344621	Uganda		1976.405
AY344622	Uganda		1976.408
AY344623	Uganda		1974.411
AY344624	Uganda		1998.466
AY344625	Uganda		1975.4
AY344627	Uganda		1974.403
AY344628	Uganda		1975.405
AY344629	Uganda		1998.466
AY349950	Uganda	Mbarhra	2002.414
AY349954	Uganda	Jinja	2002.425
DQ165073	Kenya	Nakuru	2002.77
DQ165075	Sudan	Omdurman	1986.877
FJ798106	Ethiopia	Assawa	2005.414
FJ798107	Ethiopia	Debre Birhan	1996.175
FJ798109	Ethiopia	Shashemene	2004
FJ798113	Ethiopia	Guba Lafto	2005.022
FJ798114	Ethiopia	Guba Lafto	2005.022
FJ798117	Ethiopia	Kality	2006
FJ798119	Ethiopia	Robe	2005.126
FJ798121	Ethiopia	Robe	2005.132
FJ798124	Ethiopia	Robe	2005.304
FJ798128	Ethiopia	Yabello	2006.332
FJ798130	Ethiopia	Yabello	2006.337
FJ798133	Ethiopia	Dalocha	2006.748
FJ798134	Ethiopia	Abernosa	2006.989
FJ798135	Ethiopia	Abernosa	2006.992
FJ798136	Ethiopia	Ziway	2006.915
FJ798137	Ethiopia	Ankesha	2007.104
FJ798139	Ethiopia	Fiche	2007.918
FJ798140	Ethiopia	Fiche	2007.921
FJ798143	Ethiopia	Mizan Teferi	2005.304
GU566045	Sudan	Warab	2004.677
GU566052	Sudan	Warab	2004.677
GU566056	Sudan	South Darfur	2005.005
GU566058	Sudan	North Darfur	2005.022
GU566059	Sudan	White Nile State	2008.112

8.2. Phylogeographic analysis and identification of factors impacting the diffusion of Foot-and-Mouth disease virus in Africa

GU566062	Sudan	White Nile State	2008.121
GU566063	Sudan	Gezira state	2008.351
HM211077	Ethiopia		1977.767
HM756587	Kenya	Laikipia	1964.414
HM756588	Kenya	Nakuru	1978.414
HM756589	Kenya	Laikipia	1980.414
HM756590	Kenya	Thika	1982.414
HM756591	Kenya	Kiambu	1984.414
HM756592	Kenya	Kiambu	1984.414
HM756593	Kenya	Kiambu	1985.414
HM756594	Kenya	Kiambu	1987.414
HM756595	Kenya	Kiambu	1991.414
HM756596	Kenya	Nakuru	1992.414
HM756597	Kenya	Kiambu	1992.414
HM756598	Kenya	Kiambu	1993.414
HM756599	Kenya	Laikipia	1993.414
HM756600	Kenya	Kiambu	1995.414
HM756601	Kenya	Kiambu	1995.414
HM756602	Kenya	Kiambu	1998.414
HM756603	Kenya	Nakuru	1999.414
HM756604	Uganda		1999.414
HM756608	Kenya	Trans Nzoia	2000.414
HM756609	Kenya	Nairobi	2000.414
HM756614	Kenya	Nakuru	2001.414
HM756615	Kenya	Mombasa	2001.414
HM756616	Kenya	Nakuru	2002.414
HM756617	Kenya	Nakuru	2003.414
HM756619	Uganda	Hoima	2004.414
HM756621	Uganda	Hoima	2004.414
HM756622	Kenya	Laikipia	2005.414
HM756624	Kenya	Kiambu	2005.414
HM756625	Uganda	Wakiso	2005.414
HM756626	Kenya	Uasin Gishu	2006.414
HM756627	Uganda	Mpigi	2006.414
HM756628	Uganda	Mpigi	2006.414
HM756633	Kenya	Kiambu	2007.414
HM756634	Kenya	Muranga	2007.414
HM756639	Kenya	Kajiado	2008.414
HM756640	Kenya	Thika	2008.414
JN974308	Uganda		2009.414
JN974309	Uganda		2009.414
JQ837834	Egypt		2009.584

JQ837836	Egypt		2009.589
KF135270	Kenya	Machakos Eastern province	2004.803
KF135271	Kenya	Marsabit North-Eastern province	2004.978
KF135272	Kenya	Nyeri Central province	2005.068
KF135273	Kenya	Kiambu Central province	2005.551
KF135274	Kenya	Kiambu Central province	2005.534
KF135275	Kenya	Imenti-North Eastern province	2007.556
KF135276	Kenya	Thika Central province	2008.06
KF135277	Kenya	Nakuru Rift valley province	2008.014
KF135278	Kenya	Molo Rift Valley province	2008.915
KF135279	Kenya	Kipkelion Rift Valley province	2009.085
KF135280	Kenya	Bondo Nyanza province	2009.085
KF135281	Kenya	Eldoret Rift Valley province	2010.159
KF135282	Kenya	Transmara Rift Valley province	2010.51
KF135283	Kenya	Kericho Rift Valley province	2010.797
KF135285	Kenya	Rongai Rift Valley province	2010.877
KF135286	Kenya	Kinango Coast province	2010.89
KF135292	Kenya	Gilgil Rift Valley province	2011.712
KF135293	Kenya	Masaba Rift Valley province	2011.723
KF207881	Kenya	Central regionThika District	2011.093
KF207882	Kenya	Nyanza region Siaya District	2011.115
KF207883	Kenya	Rift Valley region Rongai District	2011.164
KF207884	Kenya	Rift Valley regionBaringo District	2011.266
KF207885	Kenya	Rift Valley regionLaikipia District	2010.649
KF207886	Kenya	Central regionThika West District	2010.959
KF207887	Kenya	Western Kenya Kwanza District	2010.992
KF478938	Uganda	Bukedea	2011.17
KF478941	Uganda	Sembabule	2011.184
KF561676	Tanzania	Iringa Region	1985.414
KF561677	Tanzania	Kyela Mbeya	1998.805
KF561679	Tanzania	Kibaha District Pwani Region	2004.414
KF561682	Tanzania	Songea Ruvuma Region	2004.414
KF561684	Tanzania	Morogoro Region	2008.666
KF561685	Tanzania	Morogoro Region	2009.329
KJ210078	Egypt	EL-Mania	2013.414
KJ831667	Ethiopia		1993.452
KJ831668	Ethiopia		1994.455
KJ831669	Ethiopia		1995.458
KJ831670	Ethiopia		2004.414
KJ831671	Ethiopia		2005.427
KJ831704	Sudan		2008.414
KM921829	Libya	Zliten	2013.666

KM921835	Libya	Zliten	2013.685
KM921841	Libya	Tripoli District	2013.915
KU720572	cameroon	Far North	2010.099
KU720573	cameroon	Far North	2010.099
KU720577	cameroon	Far North	2012.082
KU720580	cameroon	Far North	2012.082
KX258001	Egypt	Dorgram	2013.904
KX258002	Egypt	Karia	2014.041
KX258003	Egypt	Kafr Elsabi	2014.077
KX258004	Egypt	Abo Mostafa	2014.162
KX258016	Niger	in gall	2001.715
KX258017	Niger		2005.915
KX258018	Niger	Dole	2005.915
KX258020	Nigeria	Nabordo	2007.721
KX258021	Nigeria	Federal Lowest Area Joss	2009.647
KX258022	Nigeria	Makurdi	2011.441
KX258023	Nigeria	Oke Buku	2011.482
KX258024	Nigeria	Kachia	2011.553
KX258025	Nigeria	Jos South	2011.584
KX258026	Nigeria	Madagali	2012.586
KX258027	Nigeria	Shuwa	2012.586
KX258029	Nigeria	Kara	2014.005
KX258030	Nigeria	Kara	2014.005
KX258033	Sudan	Omdourman	2009.948
KX258034	Sudan	Hilat Kuku	2010.027
KX258035	Sudan	Keriab	2010.03
KX258036	Sudan	Mwelain	2011.167
KX258038	Togo	Atakpame	2004.786
KX258039	Togo	Gnidjin village	2005.458
KX258041	Togo	Gnidjin village	2005.458
KX424677	Niger	Gaya	2014.416
KX424681	Niger	Kollo	2014.427
KX447125	Egypt	Ismailia	2016.414
KX447138	Egypt	Ismailia	2016.414

Supplementary table 8-21: List of sequences for the FMDV serotype SAT1

Accession number	**Animal**	**Country**	**Region**	**Date**
AF137401	Antelope	Botswana	Southern Africa	1977
AY274483	Antelope	Botswana	Southern Africa	1977
KF219688	Cattle	Botswana	Southern Africa	1977
AY274482	Cattle	Botswana	Southern Africa	1977
KJ999918	Cattle	Botswana	Southern Africa	1998

KJ999919	Cattle	Botswana	Southern Africa	2006
KJ999920	Cattle	Botswana	Southern Africa	2006
KR188458	Antelope	Kenya	Kenya Uganda	2010
KP263443	Buffalo	Kenya	Kenya Uganda	2012
HQ267520	Cattle	Kenya	Kenya Uganda	1980
HQ267521	Cattle	Kenya	Kenya Uganda	1981
HQ267523	Cattle	Kenya	Kenya Uganda	1999
HQ267524	Cattle	Kenya	Kenya Uganda	1999
KR188451	Cattle	Kenya	Kenya Uganda	2004.5
HQ267525	Cattle	Kenya	Kenya Uganda	2005
HQ267526	Cattle	Kenya	Kenya Uganda	2005
KR188452	Cattle	Kenya	Kenya Uganda	2005.1
HQ267527	Cattle	Kenya	Kenya Uganda	2006
HQ267528	Cattle	Kenya	Kenya Uganda	2006
KR188453	Cattle	Kenya	Kenya Uganda	2006.1
KR188454	Cattle	Kenya	Kenya Uganda	2006.4
KR188455	Cattle	Kenya	Kenya Uganda	2008.5
KR188501	Cattle	Kenya	Kenya Uganda	2010
KR188502	Cattle	Kenya	Kenya Uganda	2010
KR188457	Cattle	Kenya	Kenya Uganda	2010.1
KR188461	Cattle	Kenya	Kenya Uganda	2010.1
KR188470	Cattle	Kenya	Kenya Uganda	2010.2
KR188468	Cattle	Kenya	Kenya Uganda	2010.2
KR188476	Cattle	Kenya	Kenya Uganda	2010.3
KR188472	Cattle	Kenya	Kenya Uganda	2010.3
KR188487	Cattle	Kenya	Kenya Uganda	2010.4
KR188491	Cattle	Kenya	Kenya Uganda	2010.4
KR188482	Cattle	Kenya	Kenya Uganda	2010.5
KR188465	Cattle	Kenya	Kenya Uganda	2010.5
KP263445	Cattle	Kenya	Kenya Uganda	2011
KP263444	Cattle	Kenya	Kenya Uganda	2012
KJ999928	Cattle	Mozambique	Southern Africa	2002
AF137410	Buffalo	Namibia	Southern Africa	1998
AF137407	Buffalo	Namibia	Southern Africa	1998
KJ999921	Cattle	Namibia	Southern Africa	1998
KJ999922	Cattle	Namibia	Southern Africa	1998
KJ999923	Cattle	Namibia	Southern Africa	2010
AY396724	Antelope	South_Africa	South Africa	1971
AY274479	Antelope	South_Africa	South Africa	1971
AY274480	Antelope	South_Africa	South Africa	1973
AY396725	Antelope	South_Africa	South Africa	1973
AY396731	Antelope	South_Africa	South Africa	1974
AY396727	Antelope	South_Africa	South Africa	1974
AY396743	Antelope	South_Africa	South Africa	1981

AY396744	Antelope	South_Africa	South Africa	1981
AY274502	Antelope	South_Africa	South Africa	1998
AY274503	Antelope	South_Africa	South Africa	1998
AY274485	Buffalo	South_Africa	South Africa	1986
AY396735	Buffalo	South_Africa	South Africa	1986
DQ780595	Buffalo	South_Africa	South Africa	1987
DQ780594	Buffalo	South_Africa	South Africa	1987
AY396736	Buffalo	South_Africa	South Africa	1989
AY274488	Buffalo	South_Africa	South Africa	1989
DQ780596	Buffalo	South_Africa	South Africa	1990
DQ780577	Buffalo	South_Africa	South Africa	1991
DQ780597	Buffalo	South_Africa	South Africa	1991
DQ780582	Buffalo	South_Africa	South Africa	1992
DQ780585	Buffalo	South_Africa	South Africa	1992
GU194498	Buffalo	South_Africa	South Africa	1995
AY274500	Buffalo	South_Africa	South Africa	1996
DQ780580	Buffalo	South_Africa	South Africa	1998
DQ780581	Buffalo	South_Africa	South Africa	1998
DQ780592	Buffalo	South_Africa	South Africa	2002
DQ780588	Buffalo	South_Africa	South Africa	2002
DQ780591	Buffalo	South_Africa	South Africa	2003
DQ780590	Buffalo	South_Africa	South Africa	2003
DQ780598	Buffalo	South_Africa	South Africa	2005
DQ780599	Buffalo	South_Africa	South Africa	2005
AY274477	Cattle	South_Africa	South Africa	1961
AY396721	Cattle	South_Africa	South Africa	1971
AY396722	Cattle	South_Africa	South Africa	1971
AY396726	Cattle	South_Africa	South Africa	1974
AY274481	Cattle	South_Africa	South Africa	1975
AY396732	Cattle	South_Africa	South Africa	1975
AY396733	Cattle	South_Africa	South Africa	1979
AY396734	Cattle	South_Africa	South Africa	1980
KJ999908	Cattle	South_Africa	South Africa	2000
KJ999909	Cattle	South_Africa	South Africa	2002
KJ999914	Cattle	South_Africa	South Africa	2003
KJ999910	Cattle	South_Africa	South Africa	2003
KJ999912	Cattle	South_Africa	South Africa	2009
KJ999913	Cattle	South_Africa	South Africa	2010
AF446016	Buffalo	Swaziland	Southern Africa	2000
MF592595	Buffalo	Tanzania	Tanzania	2010
MF592594	Buffalo	Tanzania	Tanzania	2010
KF561706	Cattle	Tanzania	Tanzania	1971
HQ267519	Cattle	Tanzania	Tanzania	1971
KF561709	Cattle	Tanzania	Tanzania	1980

KF561711	Cattle	Tanzania	Tanzania	1996
KF561710	Cattle	Tanzania	Tanzania	1996.5
KF561713	Cattle	Tanzania	Tanzania	1999
KF561718	Cattle	Tanzania	Tanzania	1999
KF561712	Cattle	Tanzania	Tanzania	1999.5
KF561720	Cattle	Tanzania	Tanzania	1999.7
KF219682	Buffalo	Uganda	Kenya Uganda	1970
KJ999929	Cattle	Zambia	Southern Africa	1993
KJ999930	Cattle	Zambia	Southern Africa	2006
GU194499	Buffalo	Zimbabwe	Zimbabwe	1990
GU194497	Buffalo	Zimbabwe	Zimbabwe	1990
AY274495	Buffalo	Zimbabwe	Zimbabwe	1991
AY274494	Buffalo	Zimbabwe	Zimbabwe	1991
AY274498	Buffalo	Zimbabwe	Zimbabwe	1994
AY274499	Buffalo	Zimbabwe	Zimbabwe	1995
AF137411	Buffalo	Zimbabwe	Zimbabwe	1998
GU194500	Cattle	Zimbabwe	Zimbabwe	1994
KJ999924	Cattle	Zimbabwe	Zimbabwe	1995
KJ999925	Cattle	Zimbabwe	Zimbabwe	1998
KF219690	Cattle	Zimbabwe	Zimbabwe	2003
KJ999926	Cattle	Zimbabwe	Zimbabwe	2003
KT879756	Cattle	Zimbabwe	Zimbabwe	2015.1
KT879757	Cattle	Zimbabwe	Zimbabwe	2015.1

Supplementary table 8-22: List of sequences for the FMDV serotype SAT2

Accession number	**Host**	**Country**	**Locality**	**Date**
AF136607	Cattle	Zimbabwe		1983
AF136983	Buffalo	Zimbabwe	Mukazi Ranch Chiredzi	1997
AF136987	Impala	South Africa	Kingfisherspruit	1988.515
AF136988	Impala	South Africa	Orpen	1988.515
AF136989	Impala	South Africa	Rabelais Dam	1988.4
AF137001	Buffalo	South Africa	Rietpan	1988
AF137003	Buffalo	South Africa	Shilolweni	1988
AF137008	Buffalo	South Africa	Satara	1992.553
AF137016	Buffalo	South Africa	Monzweni	1995.918
AF137017	Buffalo	South Africa	Monzweni	1995.918
AF367103	Buffalo	South Africa	Kruger NP	1988
AF367104	Impala	South Africa	Kruger NP	1988
AF367107	Impala	South Africa	Kruger NP	1988
AF367114	Impala	South Africa	Kruger NP	1992
AF367116	Buffalo	Zambia	Nanzhila	1993

AF367117	Buffalo	Zambia	Nanzhila	1993
AF367120	Buffalo	Zambia	Mulanga	1996
AF367121	Cattle	Zambia	Mulanga	1996
AF367122	Buffalo	Botswana	Nxaraga	1998
AF367123	Buffalo	Botswana	Nxaraga	1998
AF367124	Buffalo	Botswana	Vumbura	1998
AF367127	Buffalo	Namibia	East Caprivi GR	1998
AF367128	Buffalo	Namibia	East Caprivi GR	1998
AF367130	Buffalo	Zimbabwe	Chizarira	1998
AF367133	Cattle	Kenya	Kaloleni Kilifi	1999
AF367139	Cattle	Nigeria		1975
AY254448	Cattle	Cameroon	Mayo-Banyo	2000.386
AY254451	Cattle	Cameroon	Faro et Deo	2000.57
AY254469	Cattle	Cameroon	Vina	2000.411
AY254471	Cattle	Cameroon	Faro et Deo	2000.57
AY254713	Cattle	Zimbabwe		1979
AY254717	Cattle	Namibia	Sigwe Village East Caprivi	1989
AY254722	Buffalo	Namibia	Matusadona National Park	1991
AY254723	Buffalo	Namibia	Urungwe Safari Area	1991
AY254725	Buffalo	Zimbabwe		1994
AY254728	Buffalo	Zambia		1996
AY343935	Cattle	Ethiopia		1990
AY343936	Cattle	Ethiopia		1990
AY343937	Cattle	Ethiopia		1991
AY343939	Cattle	Sudan		1977
AY343943	Cattle	Kenya		1986
AY343953	Cattle	Kenya		1992
AY343960	Cattle	Kenya		1996
AY343964	Cattle	Uganda		1976
AY343965	Cattle	Uganda		1976
AY343969	Cattle	Uganda		1998
AY343970	Cattle	Tanzania	Sanya Juu Moshi Kilamanjaro Region	1975
AY343971	Cattle	Tanzania	Mtawanya Mtwara Region Tanzania	1986
AY442014	Cattle	Sudan		1977
AY442912	Cattle	South Africa	Craigieburn LP	2001
AY442913	Cattle	South Africa	Craigieburn LP	2001
DQ009726	Cattle	Zimbabwe		1983
FJ798158	Cattle	Ethiopia		1991.329
FJ798159	Cattle	Ethiopia		1991.329
FJ798160	Cattle	Ethiopia		1991.329
FJ798161	Cattle	Ethiopia		2007.682
GU194491	Buffalo	Zimbabwe		1988

GU194492	Buffalo	Zimbabwe		1994
GU323172	Cattle	Uganda	Kiboga district	2004
GU323178	Cattle	Uganda	Kiboga district	2004
GU566071	Cattle	Sudan	Khartoum	2007.055
GU566072	Cattle	Sudan	Gezira state	2008.093
HM623680	Cattle	Kenya	Kiambu	1982
HM623682	Cattle	Uganda		1983
HM623683	Cattle	Kenya	Nakuru	1983
HM623685	Cattle	Kenya	Nakuru	1984
HM623695	Cattle	Kenya	Nairobi	1995
HM623702	Cattle	Kenya	Laikipia	2006
JQ639292	Cattle	Zimbabwe	Lupane area (Jotholo) North	2001
JQ639294	Buffalo	Zimbabwe	Hwang National Park	1988
JQ639295	Cattle	Zimbabwe	Lupane area (Jotholo) North	2002
JQ639296	Buffalo	Zimbabwe	Hwang National Park	1989
JQ950548	Buffalo	South Africa	Pafuri	2010
JX570615	Cattle	Cameroon	Adamawa Region	2005.378
JX570624	Cattle	Egypt	Mnoufia governorate	2012.162
JX570626	Cattle	Egypt	Garbia Governorate	2012.167
JX570628	Cattle	Kenya	Kikuyu Kiambu West Central Province	2009.085
JX570631	Cattle	Libya	Zawiya District	2003.422
JX570632	Cattle	Libya	Zawiya District	2003.436
JX570633	Cattle	Libya	Abu Attni Benghazi East Province	2012.126
JX570634	Cattle	Libya	Abu Attni Benghazi East Province	2012.126
JX570635	Cattle	Libya	Abu Attni Benghazi East Province	2012.126
KF112928	Cattle	Botswana		1977.666
KF112929	Cattle	Botswana		1978.581
KF112931	Buffalo	Egypt	Abo Greda Farasqour Domyat Delta	2012.181
KF112940	Cattle	Ethiopia	Kinbibit North Shewa Oromia Region	2009.844
KF112958	Impala	South Africa	Gudzane Kruger National Park	1985.888
KF112968	Cattle	Sudan	Sheikan Sheikan North Kordafan	2010.107
KF112969	Cattle	Namibia	Sigwe village East Caprivi	1989.863
KF112971	Cattle	Zambia	Kambwa village Monze district	1981.841
KF112972	Cattle	Zimbabwe	Lubu Diptank Manjolo TTL	1981.847
KF112974	Cattle	Zimbabwe	Triangle Chiredzi Masvingo	1987.496
KF219683	Buffalo	Uganda	Queen Elizabeth National Park	1970
KF561722	Cattle	Tanzania	Arusha Region	2004
KF561726	Cattle	Tanzania	Kihonda Morogoro Urban District Morogoro Region	2004
KF561727	Cattle	Tanzania	Makete District Njombe District Iringa Region	2009.748
KJ999935	Cattle	South Africa	Isingiro district	2008
KJ999938	Cattle	Namibia		2008
KJ999941	Cattle	Botswana	Nxaraga	1998

8.2. Phylogeographic analysis and identification of factors impacting the diffusion of Foot-and-Mouth disease virus in Africa

KJ999943	Cattle	Zambia		1993
KJ999944	Cattle	Zimbabwe		2002
KP089988	Cattle	Uganda	Isingiro district	2013.104
KP089990	Cattle	Uganda	Isingiro district	2013.104
KP263449	Cattle	Kenya	Kisii	2012.677
KP263457	Cattle	Kenya	Subukia	2012.159
KT879752	Cattle	Zimbabwe		2015.047
KT879754	Cattle	Zimbabwe		2014.677
KU897092	Buffalo	Egypt	Sharquia	2015
KU897093	Cattle	Egypt	Qaliubia	2015
KU994773	Cattle	Botswana	Matseloje Area	2002
KU994774	Cattle	Botswana	Thabana	2006
KU994775	Cattle	Botswana	Nala	2006
KU994780	Buffalo	South Africa	Shingwedzi Area Bububu	2007
KU994788	Buffalo	South Africa	Shingwedzi Area Boyela	2007
KU994789	Buffalo	South Africa	Shingwedzi Area Dzombo	2007
KU994793	Cattle	Namibia	Katimu Malilo	2008
KU994794	Buffalo	Namibia	Khaiseb/Jaco	2011
KU994795	Buffalo	Namibia	Khaiseb/Jaco	2011
KU994797	Cattle	South Africa	Kruger National Park	2012
KU994801	Cattle	South Africa	Kruger National Park	2012
KU994805	Cattle	South Africa	Kruger National Park	2011
KU994811	Cattle	Zimbabwe	Beitbrug Area	2002
KU994812	Cattle	Zimbabwe	Lupane Area (Jotholo)	2002
KU994813	Cattle	Zimbabwe	Lupane Area (Jotholo)	2002
KU994816	Cattle	Zimbabwe	Harare South	2002
KU994820	Buffalo	Zimbabwe	Bikita	2002
KU994826	Cattle	Zimbabwe	Harare	2003
KX258068	Cattle	Ethiopia	Akaki Kality Akaki Kality Sub City Addis Ababa	2014.904
KX258069	Cattle	Ethiopia	Goro East Wellega Oromia	2014.997
KX258071	Cattle	Ethiopia	Lume East Shawa Oromia	2015.189
KX258081	Cattle	Nigeria	Vom district Plateau state North Central region	2008
KX258082	Cattle	Nigeria	Vom district Plateau state North Central region	2008
KX258090	Cattle	Nigeria	Mickan Plateau State	2011.584
KX258094	Cattle	Nigeria	Igbo Oyo State	2012.471

Supplementary table 8-23: Raster provenance for the different predictors of diffusion used in this analysis

Predictor	Raster provenance
Cattle density	https://harvestchoice.org/maps/cattle-density-headsq-km-2005
Human density	http://www.worldpop.org.uk/
Landcover	http://due.esrin.esa.int/page_globcover.php
Temperature	http://worldclim.org/version2
Precipitation	http://worldclim.org/version2
Accessibility	https://forobs.jrc.ec.europa.eu/products/gam/

Supplementary table 8-24: Root-to-tips parameters for FMDA

Parameter	Value
Date range	50
Slope (rate)	1.0163E-3
X-Intercept (TMRCA)	1880.9882
Correlation Coefficient	0.5655
R squared	0.3198
Residual Mean Squared	3.9926E-4

Supplementary table 8-25: Root-to-tips parameters for FMDO

Parameter	Value
Date range	52
Slope (rate)	7.9215E-4
X-Intercept (TMRCA)	1881.4125
Correlation Coefficient	0.5328
R squared	0.2839
Residual Mean Squared	2.1895E-4

Supplementary table 8-26: Root-to-tips parameters for FMD SAT1

Parameter	Value
Date range	54.1
Slope (rate)	2.43E-04
X-Intercept (TMRCA)	1359.559
Correlation Coefficient	0.2039
R squared	4.16E-02
Residual Mean Squared	2.37E-04

8.2. Phylogeographic analysis and identification of factors impacting the diffusion of Foot-and-Mouth disease virus in Africa

Supplementary table 8-27: Root-to-tips parameters for FMD SAT2

Parameter	Value
Date range	45.189
Slope (rate)	2.7239E-4
X-Intercept (TMRCA)	1341.8646
Correlation Coefficient	0.1387
R squared	1.9247E-2
Residual Mean Squared	4.8211E-4

Supplementary table 8-28: Estimated evolutionary parameters for the three FMDV serotypes

	FMDA	FMDO	FMD SAT1	FMD SAT2
Root height	86,385(+/- 17.04)	79.481 (+/- 8.63)	260.42(+/- 44.94)	431.38(+/- 69.94)
Mean mutation rate (substitutions/site/year)	4.67 e-3(+/- 3.46 e-5)	3.69E-03 (+/- 2.01 e-5)	1.8E-03(+/-2.98 e-4)	1.1E-3(+/- 1,77 e-4)

Supplementary table 8-29: Output of the BSSVS analysis for the FMDV serotypes A showing the best supported rates of transition between the sampled countries

Origin	Destination	Bayes factor	Posterior probability
Kenya	Tanzania	6266.65	1.00
Ethiopia	Kenya	3130.18	1.00
Sudan	Eritrea	3130.18	1.00
Sudan	Cameroon	84.63	0.93
Ethiopia	Egypt	45.56	0.88
Kenya	Ethiopia	23.88	0.79
Kenya	Sudan	18.13	0.74
Cameroon	Nigeria	14.29	0.69
Sudan	Egypt	11.44	0.65

Supplementary table 8-30: Output of the BSSVS analysis for the FMDV serotypes O showing the best supported rates of transition between the sampled countries

Origin	Destination	Bayes factor	Posterior probability
Kenya	Tanzania	7712	1.00
Sudan	Egypt	3011	1.00
Nigeria	Cameroon	1209	0.99
Egypt	Libya	557	0.98

Uganda	Kenya	313	0.97
Sudan	Ethiopia	212	0.95
Sudan	Nigeria	206	0.95
Kenya	Uganda	105	0.91
Ethiopia	Kenya	70	0.87
Ethiopia	Somalia	25	0.71
Togo	Nigeria	22	0.68
Niger	Togo	20	0.66

Supplementary table 8-31: Output of the BSSVS analysis for the FMDV serotypes SAT1 showing the best supported rates of transition between the sampled countries

Origin	**Destination**	**Bayes factor**	**Posterior probability**
Kenya Uganda	Tanzania	13526	1.00
South Africa	Southern Africa	219	0.99
Zimbabwe	Southern Africa	164	0.98
Tanzania	Kenya Uganda	142	0.98
Southern Africa	Zimbabwe	19	0.85
Southern Africa	South Africa	17	0.84
Zimbabwe	Kenya Uganda	11	0.78
South Africa	Zimbabwe	11	0.77
Tanzania	Southern Africa	10	0.76
Zimbabwe	South Africa	4	0.54
Kenya Uganda	Southern Africa	3	0.50

Supplementary table 8-32: Output of the BSSVS analysis for the FMDV serotypes SAT2 showing the best supported rates of transition between the sampled countries

Origin	**Destination**	**Bayes factor**	**Posterior probability**
Kenya	Uganda	11050	1.00
Egypt	Botswana	11050	1.00
Zimbabwe	Namibia	1217	0.99
South Africa	Nigeria	839	0.99
Egypt	Kenya	778	0.98
Zimbabwe	Egypt	725	0.98
South Africa	Zimbabwe	469	0.97
South Africa	Zambia	430	0.97
Egypt	Ethiopia	313	0.96
Ethiopia	Sudan	189	0.94
Kenya	Sudan	155	0.93
Zambia	Botswana	81	0.87
Uganda	Tanzania	58	0.83

Sudan	Libya	37	0.75
Zimbabwe	Botswana	32	0.72
Sudan	Nigeria	27	0.69

Supplementary table 8-33:Result of the FMDA sequences Markov jump analysis

origin	**destination**	**jumps**
Kenya	Tanzania	4
Ethiopia	Kenya	3
Kenya	Ethiopia	2
Sudan	Eritrea	2
Kenya	Sudan	1
Sudan	Cameroon	1
Ethiopia	Egypt	1

Supplementary table 8-34: Result of the FMDO sequences Markov jump analysis

Origin	**Destination**	**Jumps**
Kenya	Tanzania	4
Sudan	Ethiopia	3
Kenya	Uganda	3
Somalia	Ethiopia	2
Uganda	Kenya	2
Sudan	Egypt	2
Ethiopia	Somalia	2
Sudan	Nigeria	2
Ethiopia	Kenya	2
Nigeria	Cameroon	1
Uganda	Tanzania	1
Tanzania	Kenya	1
Kenya	Sudan	1
Sudan	Kenya	1

Supplementary table 8-35: Result of the FMD SAT1 sequences Markov jump analysis

Origin	**Destination**	**Jumps**
Kenya Uganda	Tanzania	3
Tanzania	Kenya Uganda	2
Zimbabwe	Southern Africa	1
South Africa	Southern Africa	1
Southern Africa	Zimbabwe	1
South Africa	Zimbabwe	1
Southern Africa	South Africa	1
Zimbabwe	Kenya Uganda	1

Zimbabwe	South Africa	1
Tanzania	Southern Africa	1

Supplementary table 8-36: Result of the FMD SAT2 sequences Markov jump analysis

Origin	**Destination**	**Number of jumps**
Kenya	Uganda	4
Namibia	Zimbabwe	3
Kenya	Tanzania	2
Ethiopia	Sudan	2
Zimbabwe	Botswana	2
South Africa	Zimbabwe	2
Namibia	Botswana	2
Zimbabwe	Namibia	1
Botswana	Zimbabwe	1
Kenya	Ethiopia	1
Libya	Sudan	1
Libya	Cameroon	1
Cameroon	Libya	1
Tanzania	Kenya	1

Supplementary table 8-37: Selected clade for the FMDO dataset

Accession number	**Country**	**Location**	**latitude**	**longitude**	**Date**
AY349950	Uganda	Mbarhra	-0.60716	30.6545	2002.414
DQ165073	Kenya	Nakuru	-0.3031	36.08003	2002.77
HM756598	Kenya	Kiambu	-1.17481	36.83041	1993.414
HM756602	Kenya	Kiambu	-1.17481	36.83041	1998.414
HM756603	Kenya	Nakuru	-0.3031	36.08003	1999.414
HM756608	Kenya	TransNzoia	1.021852	35.0015	2000.414
HM756609	Kenya	Nairobi	-1.366	36.84	2000.414
HM756614	Kenya	Nakuru	-0.3031	36.08003	2001.414
HM756615	Kenya	Mombasa	-4.04348	39.66821	2001.414
HM756616	Kenya	Nakuru	-0.3031	36.08003	2002.414
HM756619	Uganda	Hoima	1.427355	31.34844	2004.414
HM756621	Uganda	Hoima	1.427355	31.34844	2004.414

HM756622	Kenya	Laikipia	0.396987	37.15878	2005.414
HM756624	Kenya	Kiambu	-1.17481	36.83041	2005.414
HM756625	Uganda	Wakiso	0.398723	32.47927	2005.414
HM756626	Kenya	Uasin	0.5	35.3	2006.414
HM756627	Uganda	Mpigi	0.227353	32.32492	2006.414
HM756628	Uganda	Mpigi	0.227353	32.32492	2006.414
HM756634	Kenya	Muranga	-0.7957	37.1322	2007.414
HM756639	Kenya	Kajiado	-1.84207	36.79186	2008.414
HM756640	Kenya	Thika	-1.03876	37.08338	2008.414
KF135270	Kenya	Machakos	-1.51768	37.26341	2004.803
KF135271	Kenya	Marsabit	2.335497	37.99435	2004.978
KF135272	Kenya	Nyeri	-0.42778	36.94336	2005.068
KF135273	Kenya	Kiambu	-1.17481	36.83041	2005.551
KF135274	Kenya	Kiambu	-1.17481	36.83041	2005.534
KF135275	Kenya	Imenti	-0.06565	36.67486	2007.556
KF135276	Kenya	Thika	-1.03876	37.08338	2008.06
KF135277	Kenya	Nakuru	-0.3031	36.08003	2008.014
KF135278	Kenya	Molo	-0.24884	35.73237	2008.915
KF135279	Kenya	Kipkelion	-0.2077	35.44806	2009.085
KF135282	Kenya	Transmara	-1.00803	34.88358	2010.51
KF135283	Kenya	Kericho	-0.36	35.28	2010.797
KF135285	Kenya	Rongai	-0.17	35.85	2010.877
KF135286	Kenya	Kinango	-4.138	39.315	2010.89
KF135292	Kenya	Gilgil	-0.49227	36.3173	2011.712
KF135293	Kenya	Masaba	-1.12	34.53	2011.723
KF207881	Kenya	Thika	-1.03876	37.08338	2011.093
KF207882	Kenya	Nyanza	-0.09954	34.76248	2011.115
KF207883	Kenya	Rongai	-0.17	35.85	2011.164
KF207884	Kenya	Baringo	0.62	35.9	2011.266
KF207885	Kenya	Laikipia	0.396987	37.15878	2010.649
KF207886	Kenya	Thika	-1.03876	37.08338	2010.959
KF478941	Uganda	Sembabule	0.063772	31.35416	2011.184

KF561679	Tanzania	Kibaha	-6.78132	38.99289	2004.414
KF561685	Tanzania	Morogoro	-6.8	37.65	2009.329

Supplementary table 8-38: Complete result table for the GLM analysis when considering the predictors as "conductance" values

Predictor	**Mean coefficient**	**sd**	**Lower hpd**	**Upper hpd**	**Mean Indicator**	**BF**
Accessibility	-0.25	0.25	-0.74	0.19	0.35	0.24
Cattle density	-0.54	0.26	-1.05	-0.04	0.82	1.97
Presence of crop	-0.14	0.20	-0.52	0.21	0.32	0.21
Presence of crop (combined)	-0.06	1.40	-3.15	2.88	0.50	0.45
Elevation	-0.54	0.38	-1.33	0.13	0.64	0.80
Presence of forest	-0.34	0.25	-0.81	0.12	0.53	0.50
Presence of fragmented crop	-0.35	0.30	-0.85	0.25	0.52	0.49
Human density	-0.15	1.44	-3.55	2.93	0.62	0.72
Precipitation	-0.08	1.46	-3.40	3.39	0.79	1.63
Presence of herbaceous vegetation	-0.06	1.33	-3.08	3.19	0.56	0.56
Temperature	0.05	1.28	-2.69	2.93	0.89	3.67
Logarithm Cattle density	-0.08	1.35	-3.54	2.64	0.95	8.83
Logarithm Human density	-0.05	1.45	-3.65	3.07	0.90	3.94

Supplementary table 8-39: Complete result table for the GLM analysis when considering the predictors as "resistance" values

Predictor	**Mean coefficient**	**sd**	**Lower hpd**	**Upper hpd**	**Mean Indicator**	**BF**
Accessibility	-0.56	0.18	-0.92	-0.22	0.94	7.53
Cattle density difference	-0.11	0.24	-0.58	0.41	0.32	0.20
Cattle density	0.07	0.27	-0.42	0.65	0.26	0.15
Presence of crop	-0.17	0.29	-0.67	0.41	0.34	0.23
Presence of crop (combined)	-0.29	0.18	-0.61	0.08	0.55	0.54
Distance	-0.61	0.21	-1.01	-0.19	0.95	8.35
Elevation	0.10	0.32	-0.36	0.85	0.27	0.17
Presence of forest	0.03	0.28	-0.44	0.63	0.26	0.16

Presence of fragmented crop	-0.27	0.20	-0.57	0.16	0.42	0.32
Human density	-0.07	1.51	-3.08	3.36	0.39	0.28
Precipitation	-0.08	1.31	-2.99	2.94	0.94	6.59
Presence of herbaceous vegetation	-0.14	1.42	-3.00	3.46	0.36	0.24
Temperature	-0.04	1.50	-3.24	3.63	0.94	6.85
Logarithm Cattle density	-0.44	0.20	-0.83	-0.07	0.82	2.05
Logarithm Human density	-0.43	0.23	-0.90	0.00	0.69	0.99

Supplementary table 8-40: Bayes factor values for different values of precipitation and temperature above certain threshold for the isolated FMDO clade in a discrete setting

Predictor	**BF as conductance**	**BF as resistance**
Precipitation above 60	4.2	2.0
Precipitation above 70	2.7	1.2
Precipitation above 80	0.4	2.9
Precipitation above 90	0.3	7.3
Precipitation above 100	0.3	4.8
Precipitation above 110	0.6	10.0
Temperature above 19℃	0.4	4.9
Temperature above 20℃	0.5	5.4
Temperature above 21℃	0.5	4.4
Temperature above 22℃	0.9	10.2
Temperature above 23℃	1.7	7.0
Temperature above 24℃	3.7	4.8

Supplementary table 8-41: Bayes factor values for different values of precipitation and temperature under certain threshold for the isolated FMDO clade in a discrete setting

Predictor	**BF as conductance**	**BF as resistance**
Precipitation below 60	2.6	3.71
Precipitation below 70	1.2	3.44
Precipitation below 80	2.7	0.36
Precipitation below 90	6.6	0.23
Precipitation below 100	5.4	0.30
Precipitation below 110	6.6	0.49
Temperature below 19℃	4	0
Temperature below 20℃	6	1
Temperature below 21℃	4	0
Temperature below 22℃	10	0
Temperature below 23℃	4	0
Temperature below 24℃	3	1

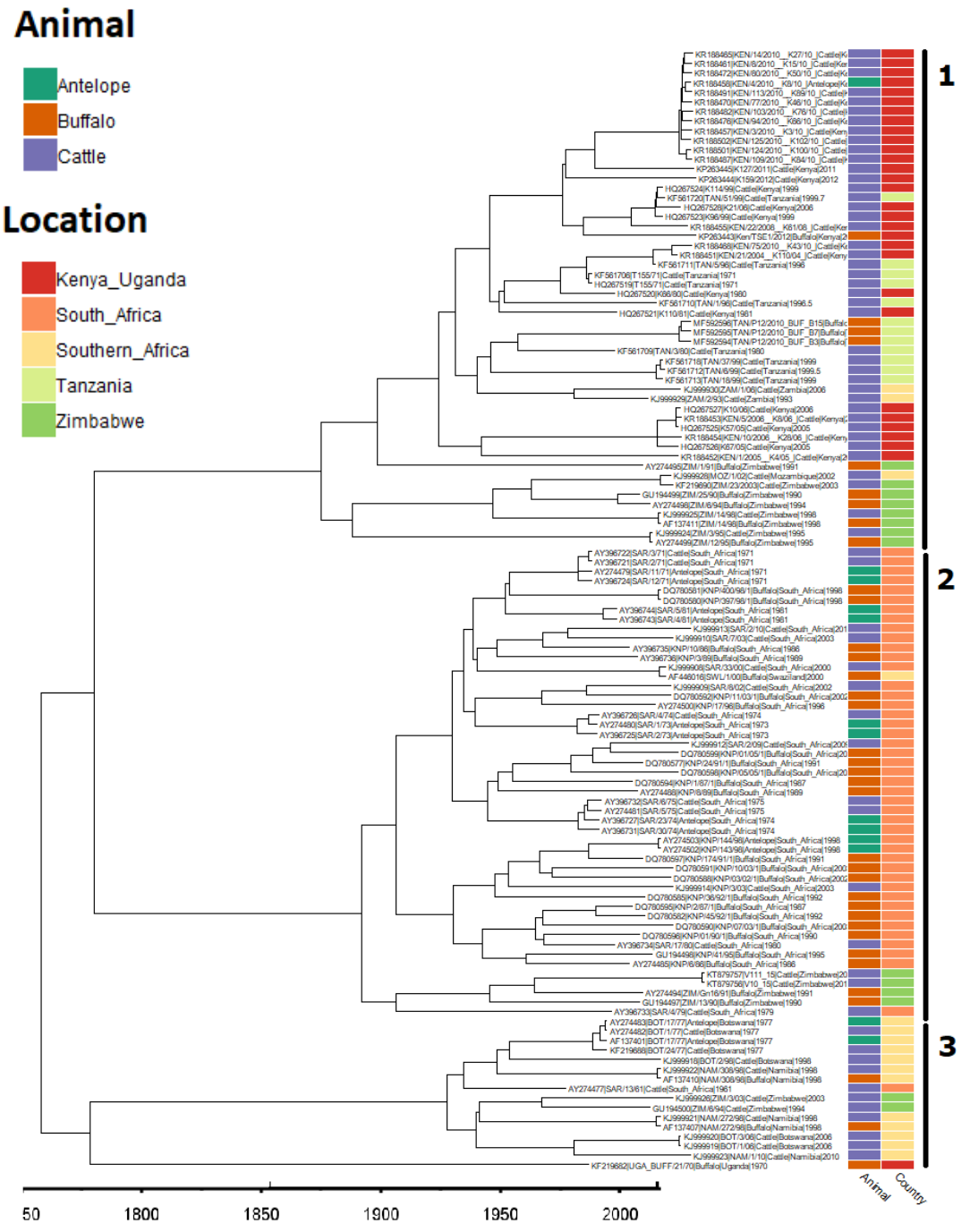
Supplementary table 8-42:Pearson correlation for the different predictors used in the discrete and continuous analysis

	Accessibility	**Cattle density**	**Presence of cropland**	**Elevation**	**Presence of forest**	**Presence of fragmented cropland**	**Human density**	**Logarithm cattle density**	**Logarithm human density**	**Average daily precipitation**	**Presence of herbaceous vegetation**	**Average daily temperature**
Accessibility	1.00	-0.09	-0.12	-0.17	0.10	-0.30	-0.19	-0.27	-0.50	-0.28	0.18	0.20
Cattle density	-0.09	1.00	0.05	0.05	-0.07	0.06	0.09	0.34	0.14	0.05	-0.05	-0.05
Presence of cropland	-0.12	0.05	1.00	-0.12	-0.01	0.04	0.08	0.19	0.18	0.08	-0.24	0.02
Elevation	-0.17	0.05	-0.12	1.00	0.10	0.16	0.12	0.09	0.38	0.35	-0.11	-0.90
Presence of forest	0.10	-0.07	-0.01	0.10	1.00	-0.37	-0.07	-0.28	-0.16	0.50	-0.54	-0.21
Presence of fragmented cropland	-0.30	0.06	0.04	0.16	-0.37	1.00	0.18	0.35	0.45	0.14	-0.36	-0.15
Human density	-0.19	0.09	0.08	0.12	-0.07	0.18	1.00	0.19	0.45	0.17	-0.13	-0.13
Logarithm cattle density	-0.27	0.34	0.19	0.09	-0.28	0.35	0.19	1.00	0.43	0.00	0.00	-0.11
Logarithm human density	-0.50	0.14	0.18	0.38	-0.16	0.45	0.45	0.43	1.00	0.40	-0.23	-0.39
Average daily precipitation	-0.28	0.05	0.08	0.35	0.50	0.14	0.17	0.00	0.40	1.00	-0.58	-0.45
Presence of herbaceous vegetation	0.18	-0.05	-0.24	-0.11	-0.54	-0.36	-0.13	0.00	-0.23	-0.58	1.00	0.21
Average daily temperature	0.20	-0.05	0.02	-0.90	-0.21	-0.15	-0.13	-0.11	-0.39	-0.45	0.21	1.00

Supplementary table 8-43: Results for effect of the cattle density on the virus circulation in a continuous setting

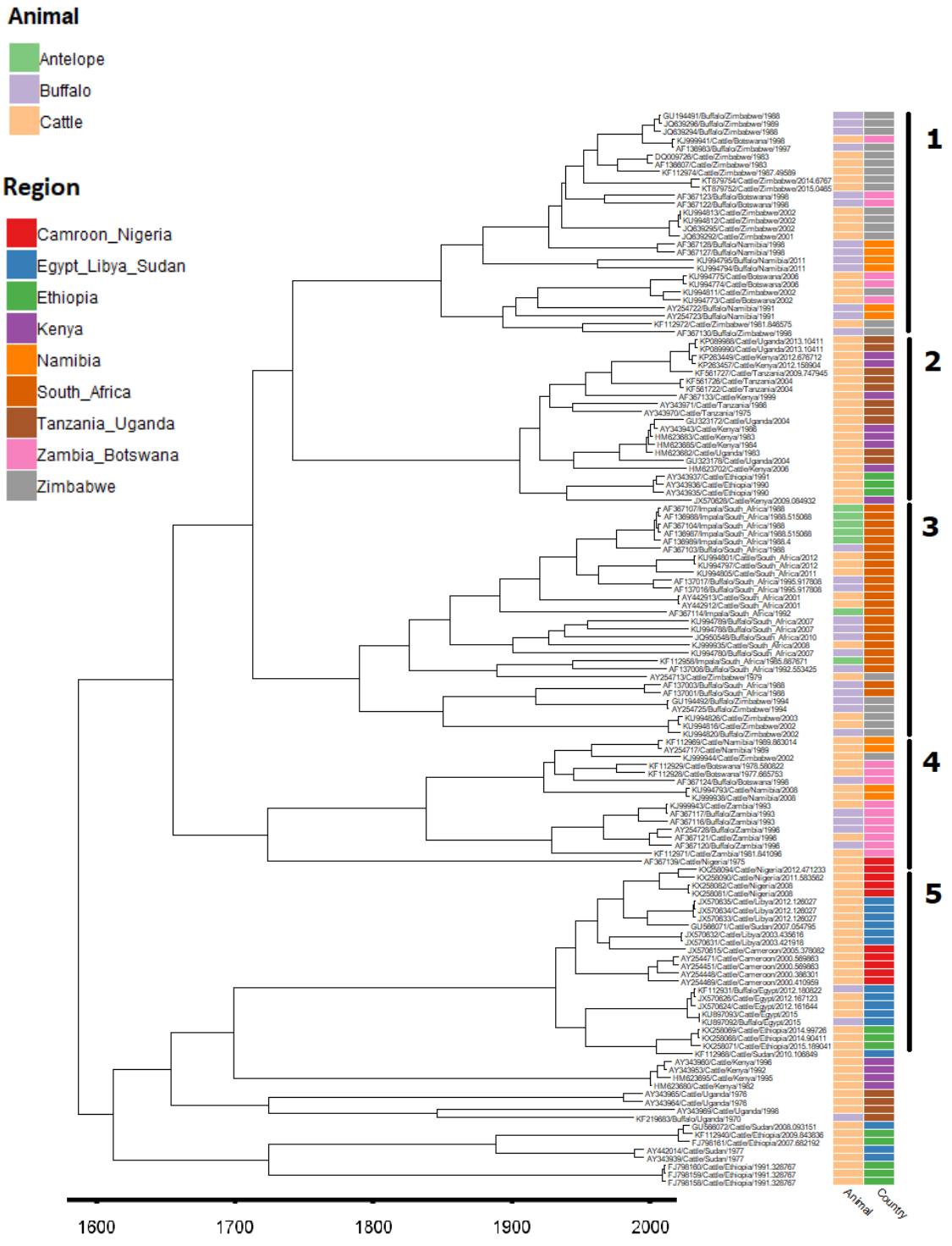
	Conductance	**Resistance**
Predictor	Bayes factor	Bayes factor
cattle above 25	0.36	0.68
cattle above 50	0.24	0.96
cattle above 100	0.08	1.76
cattle above 125	0.31	3.07

8.3 IMPORTANCE OF WILDLIFE IN THE CIRCULATION AND MAINTENANCE OF FOOT-AND-MOUTH DISEASE VIRUS SAT1 AND SAT2 IN AFRICA



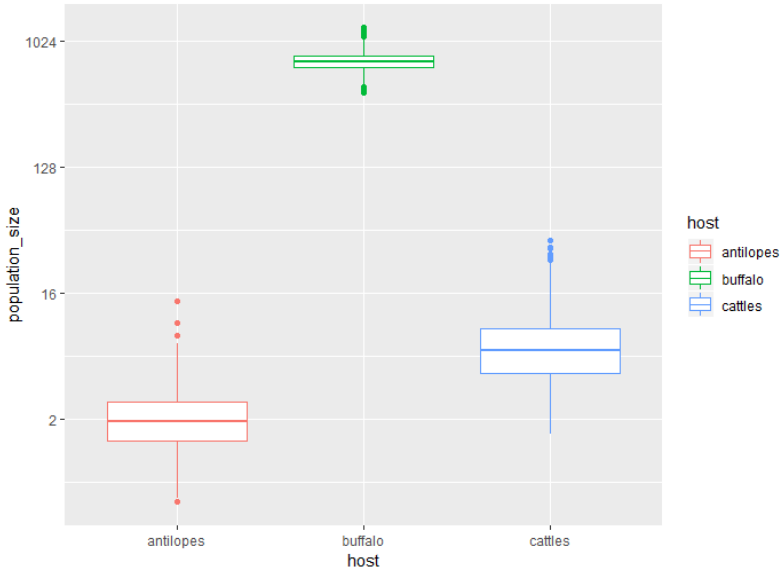
Supplementary figure 8-24: Annotated “mugration” tree for the SAT1 sequences with the Host and Location. The isolated main clades can be seen on the right of the figure

8.3. Importance of wildlife in the circulation and maintenance of Foot-and-Mouth disease virus SAT1 and SAT2 in Africa

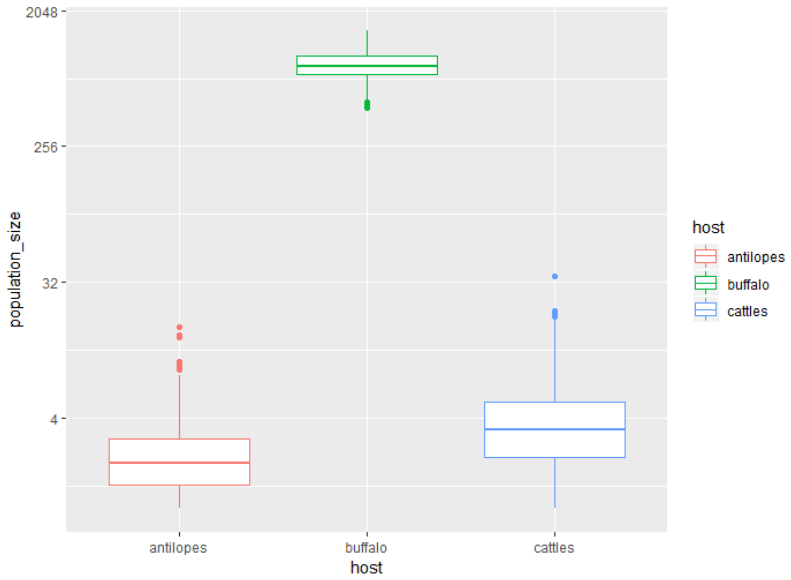


Supplementary figure 8-25: Annotated “mugration” tree for the SAT2 sequences with the Host and Location. The isolated main clades can be seen on the right of the figure

a.



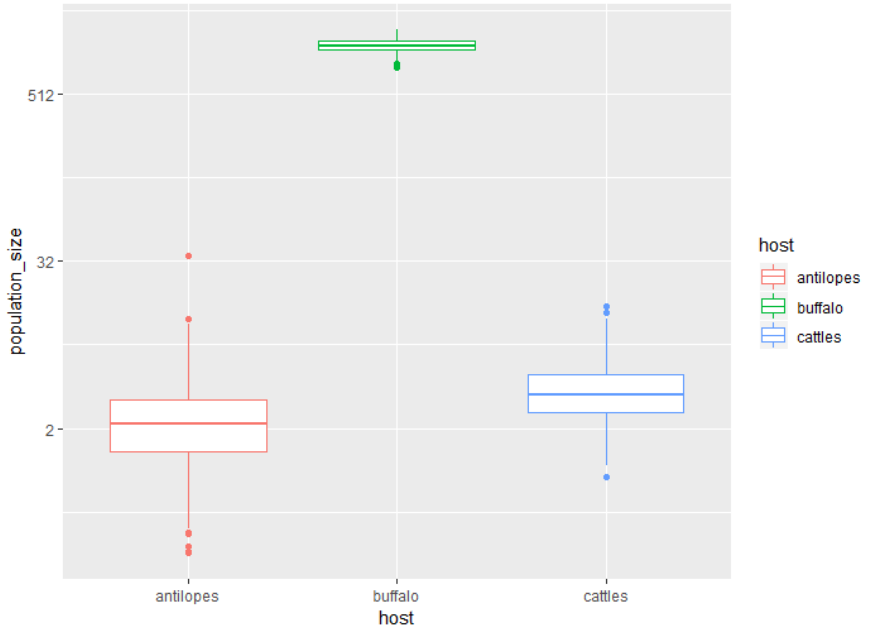
b.



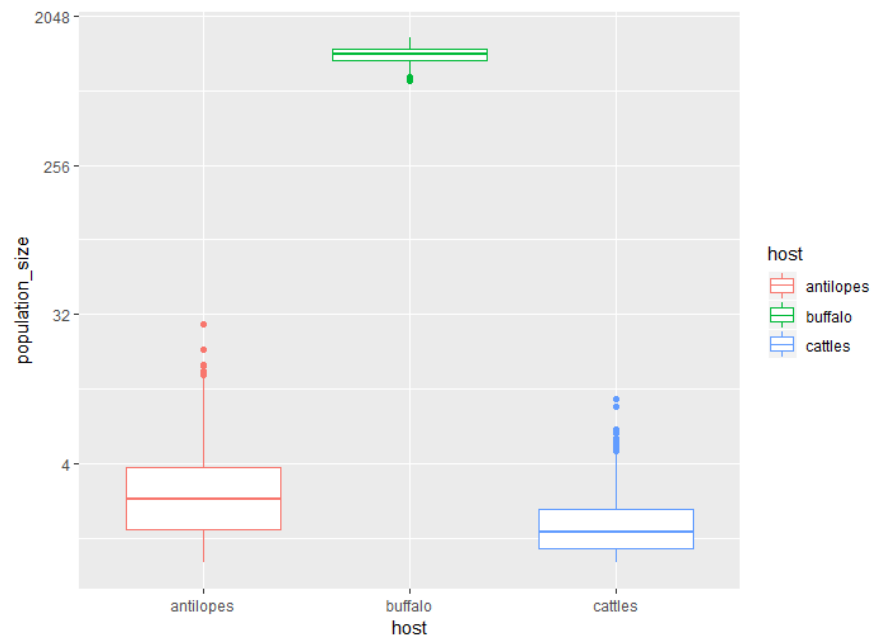
Supplementary figure 8-26: Boxplot showing the estimated viral population size present in all three analysed hosts for the serotype SAT1 using a. BASTA b. MASCOT

8.3. Importance of wildlife in the circulation and maintenance of Foot-and-Mouth disease virus SAT1 and SAT2 in Africa

a.



b.



Supplementary figure 8-27:Boxplot showing the estimated viral population size present in all three analysed hosts for the serotype SAT2 using a. BASTA b. MASCOT

Supplementary table 8-44: Parameters estimation for the reconstructed SAT1 tree

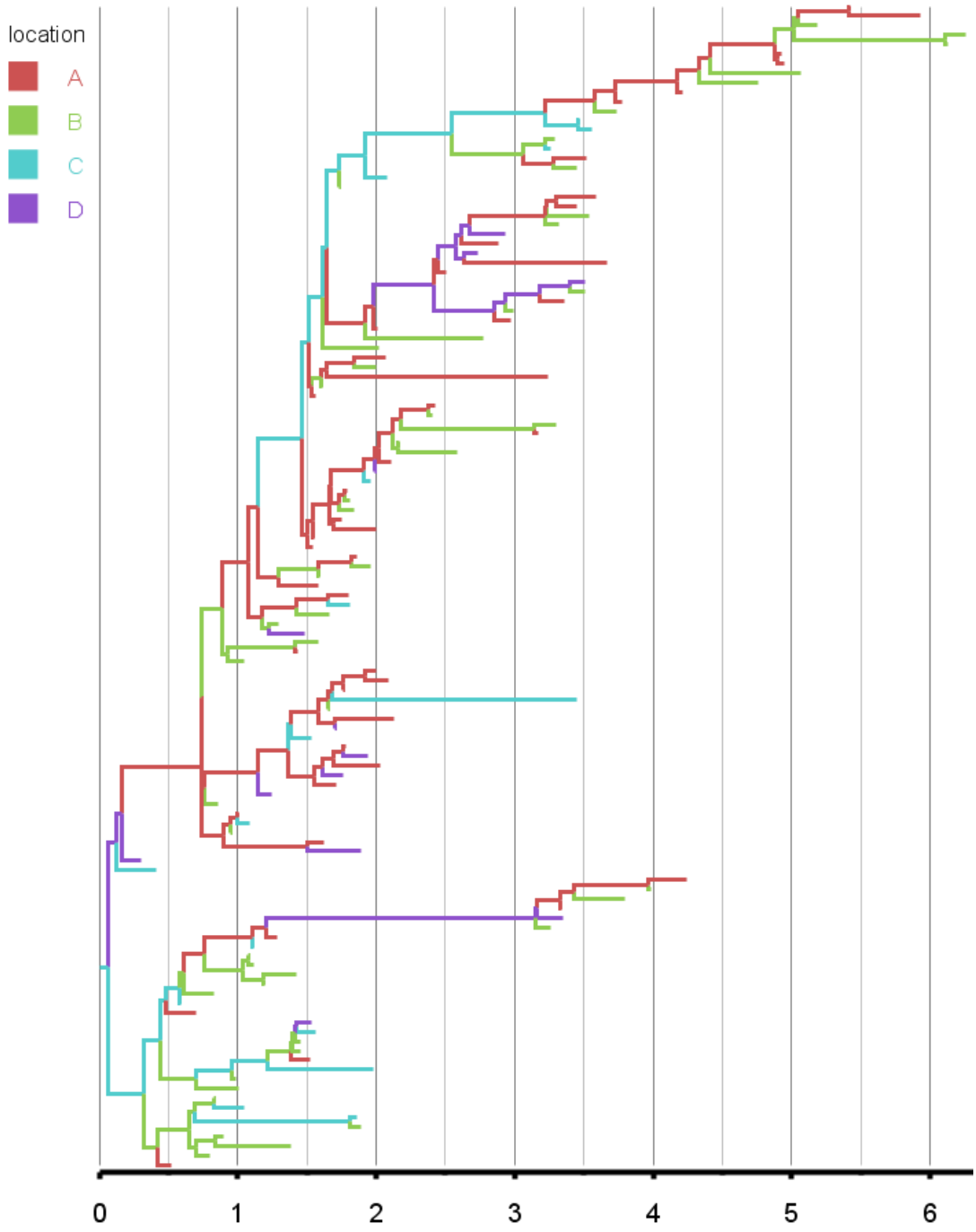
SAT1 parameters estimation	
MUGRATION	
Clock	1.71-3
Tree eight	253.262
Rate impala: buffalo	0.349
Rate impala: cattle	1.312
Rate buffalo: impala	0.7
Rate buffalo: cattle	1.917
Rate cattle: impala	0.617
Rate cattle: buffalo	1.183
BASTA	
Clock	2.079-3
Tree eight	206.771
Size antelope	2.43
Size buffalo	903.538
Size cattle	4.19
Rate Buffalo: Impala	0.26
Rate Cattle: Impala	0.33
Rate Impala: Buffalo	0.01
Rate Cattle: Buffalo	0.02
Rate Impala: Cattle	1.27
Rate Buffalo: Cattle	0.93
MASCOT	
Clock	1.95 -3
Tree eight	219.266
Size impala	2.16
Size buffalo	736.707
Size cattle	7.35
Rate Buffalo: Impala	0.6
Rate Cattle: Impala	0.46
Rate Impala: Buffalo	0
Rate Cattle: Buffalo	0
Rate Impala: Cattle	1.59
Rate Buffalo: Cattle	1.56

0.

Supplementary table 8-45: Parameters estimation for the reconstructed SAT2 tree

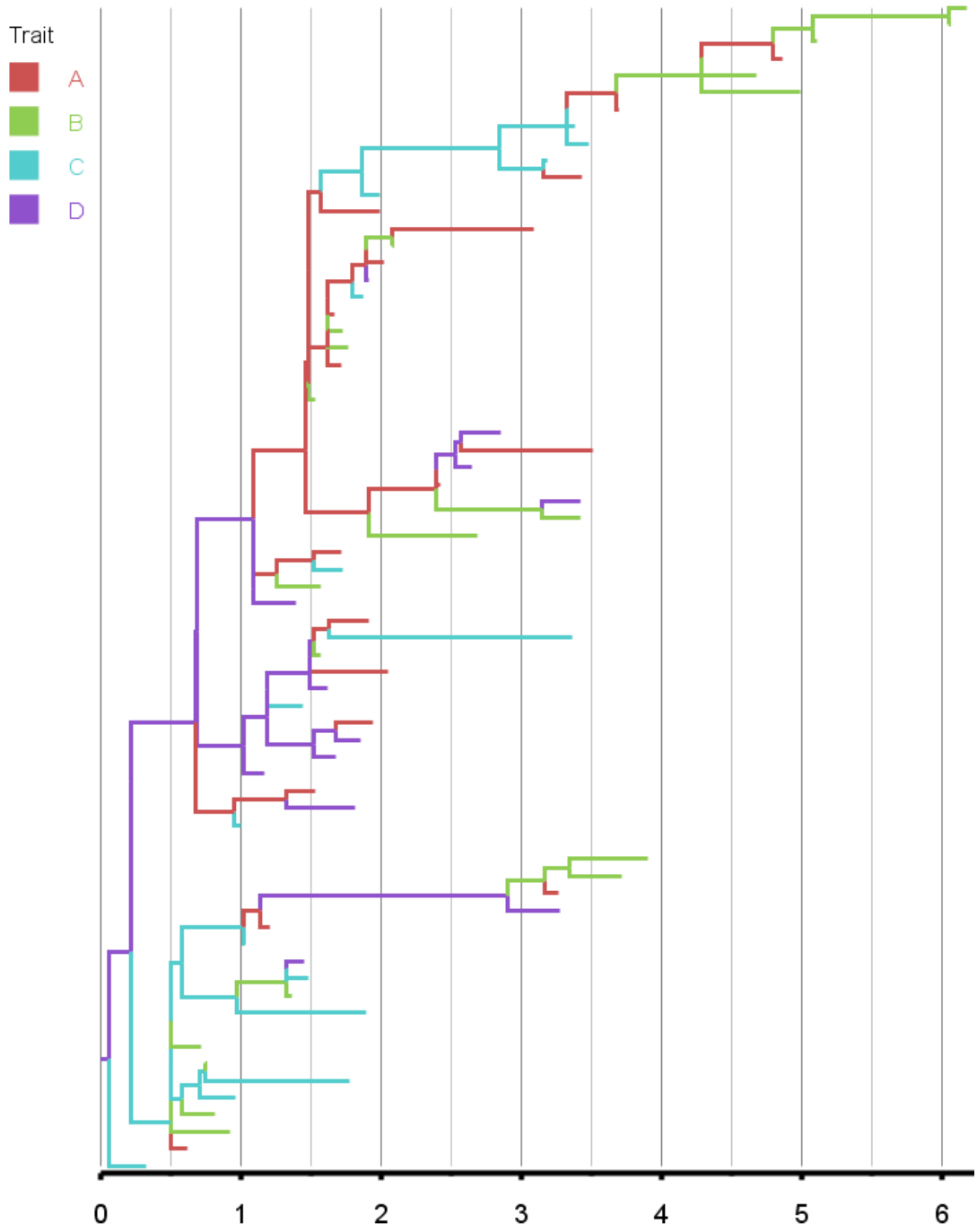
SAT2 parameters estimation	
MUGRATION	
Clock rate	1.1-3
Tree eight	431.38
Rate impala: buffalo	0.513
Rate impala: cattle	0.516
Rate buffalo: impala	0.582
Rate buffalo: cattle	2.781
Rate cattle: impala	0.643
Rate cattle: buffalo	0.955
MASCOT	
Clock rate	1.19 -3
Tree eight	355.41
Size impala	3.18
Size buffalo	1203.26
Size cattle	1.85
Rate Buffalo: Impala	0.47
Rate Cattle: Impala	0.25
Rate Impala: Buffalo	0
Rate Cattle: Buffalo	0.01
Rate Impala: Cattle	0.95
Rate Buffalo: Cattle	0.95
BASTA	
Clock rate	1.16 -3
Tree eight	353.11
Size impala	2.59
Size buffalo	1159.26
Size cattle	3.91
Rate Buffalo: Impala	1.07
Rate Cattle: Impala	0.25
Rate Impala: Buffalo	0
Rate Cattle: Buffalo	0
Rate Impala: Cattle	1.39
Rate Buffalo: Cattle	0.72

8.4 EPITREE-SIM: AN APPLICATION FOR EPIDEMIC SIMULATION AND PHYLOGENETIC TREE RECONSTRUCTION

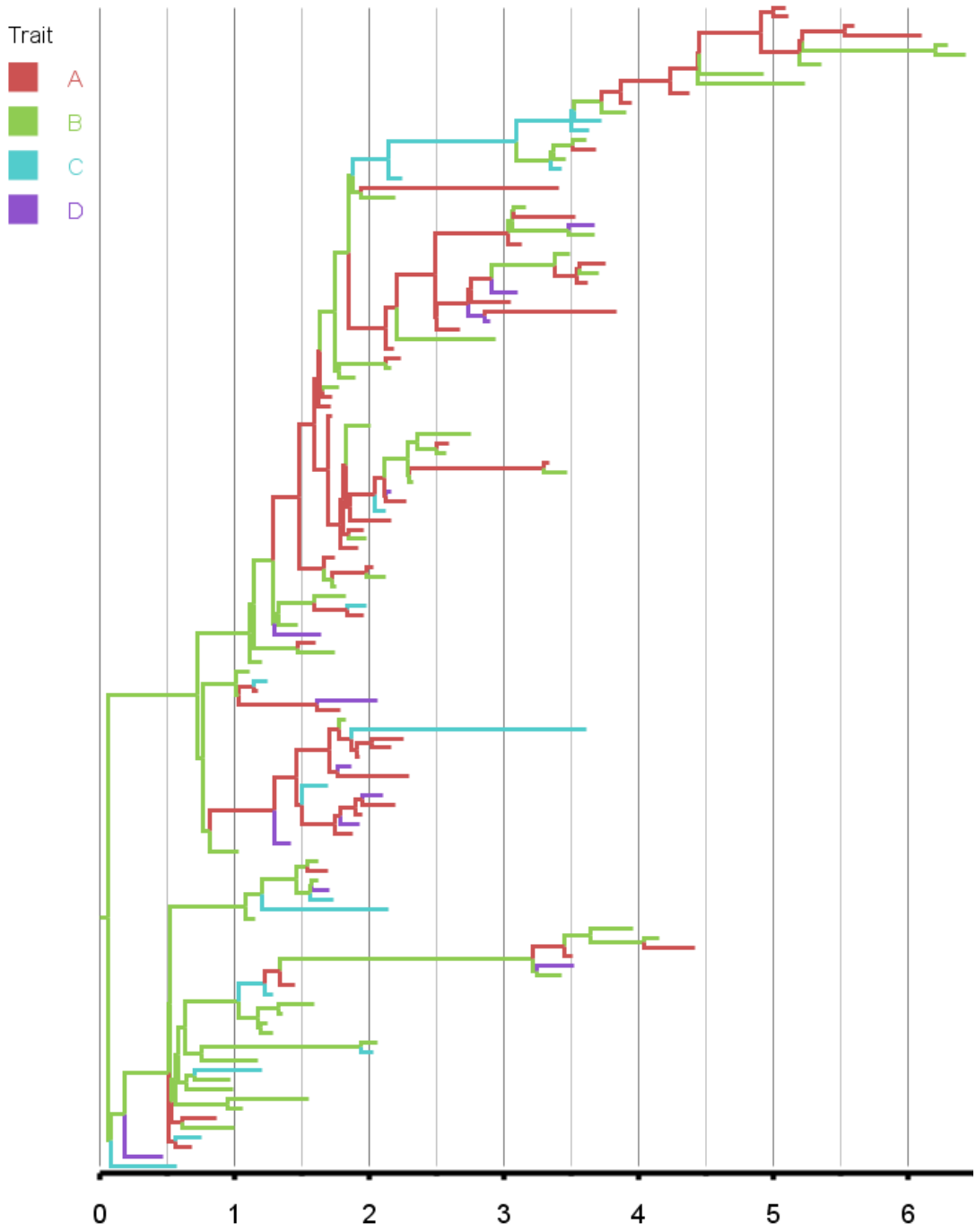


Supplementary figure 8-28: True phylogenetic tree for the first simulated epidemic. The phylogeny branches are coloured according to their respective host. The key for the graph is shown on the left.

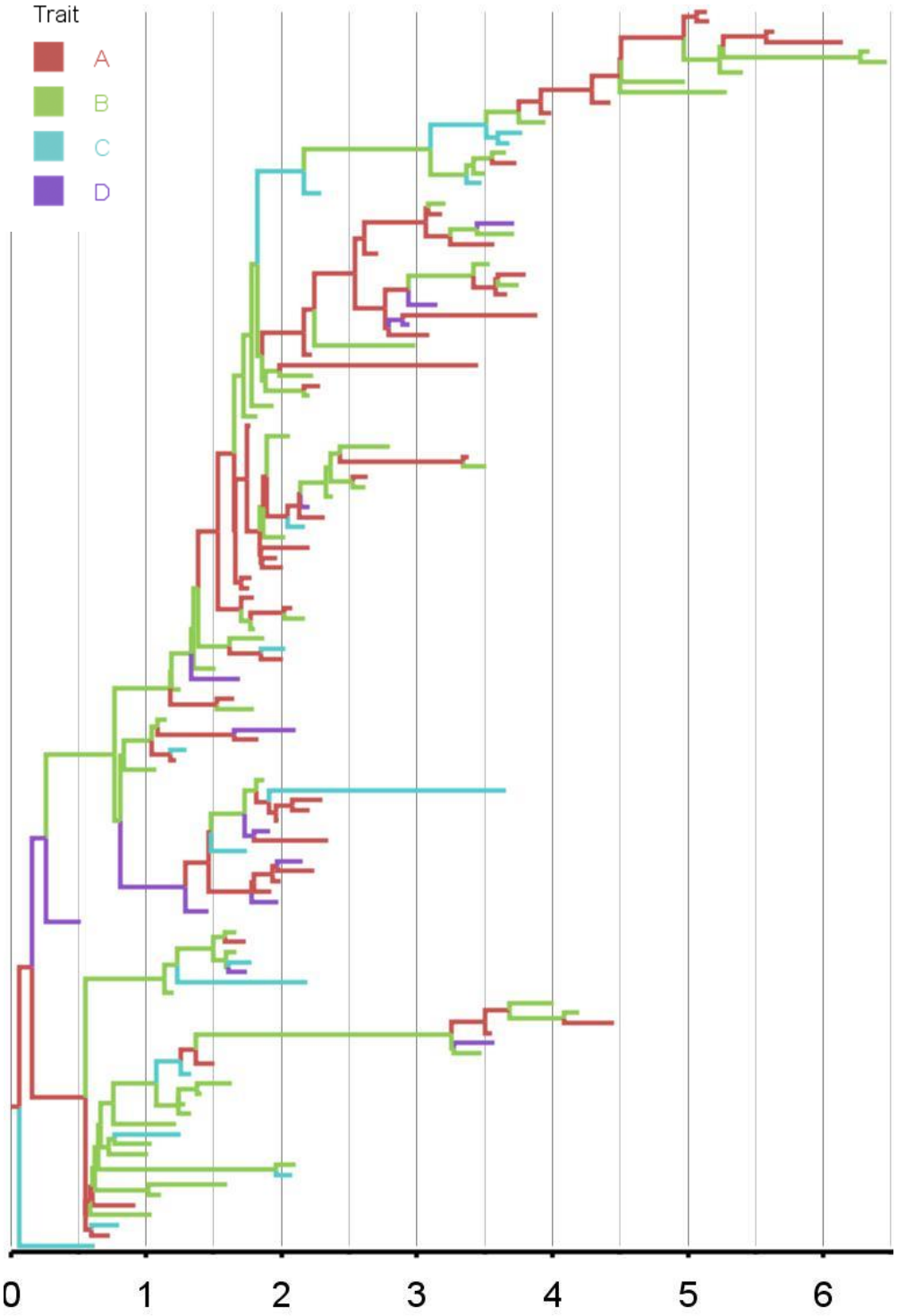
8.4. Epitree-sim: an application for epidemic simulation and phylogenetic tree reconstruction



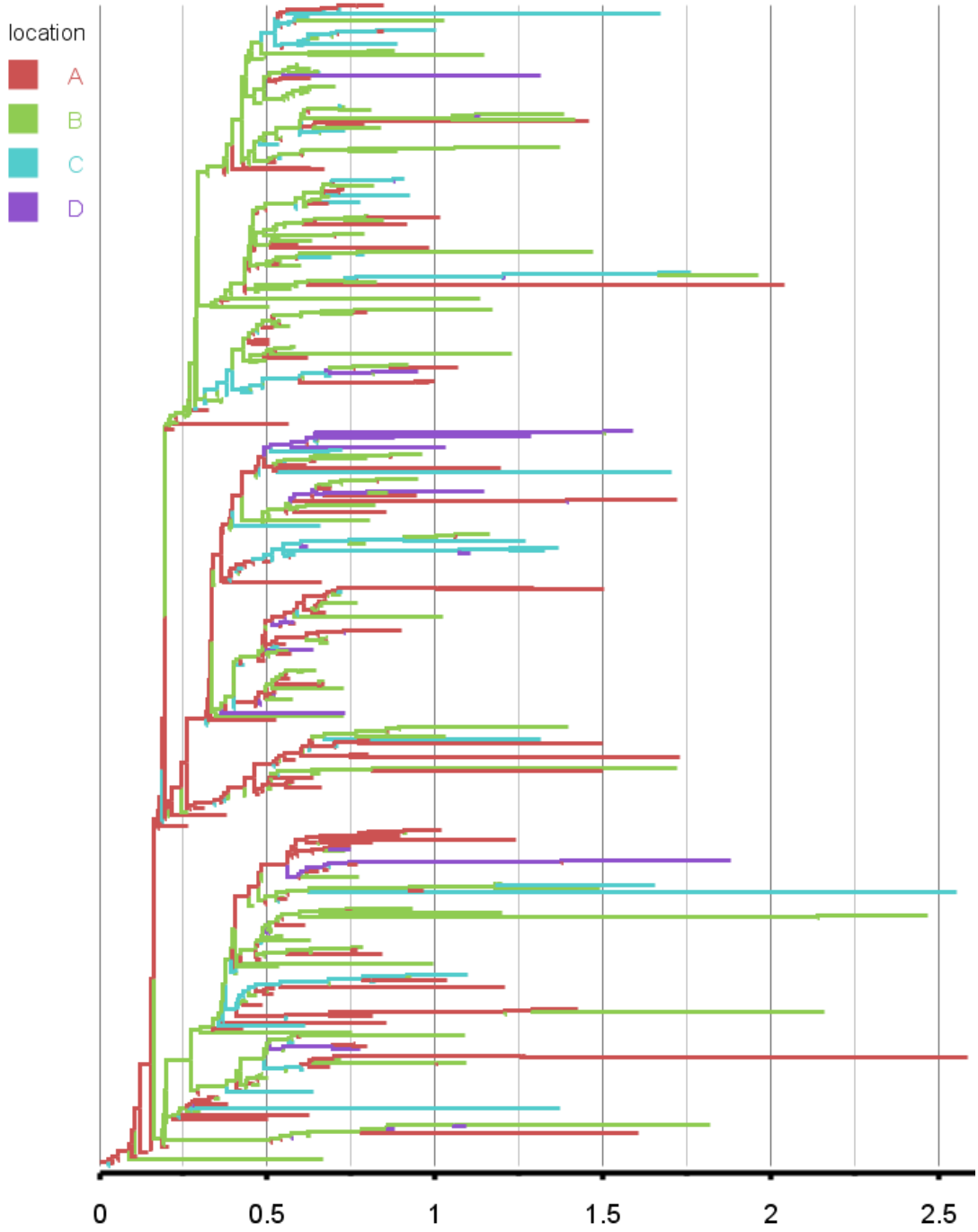
Supplementary figure-8-29: Reconstructed tree estimated through Epitree-sim for the first simulated epidemic. The phylogeny branches are coloured according to their respective host. The key for the graph is shown on the left.



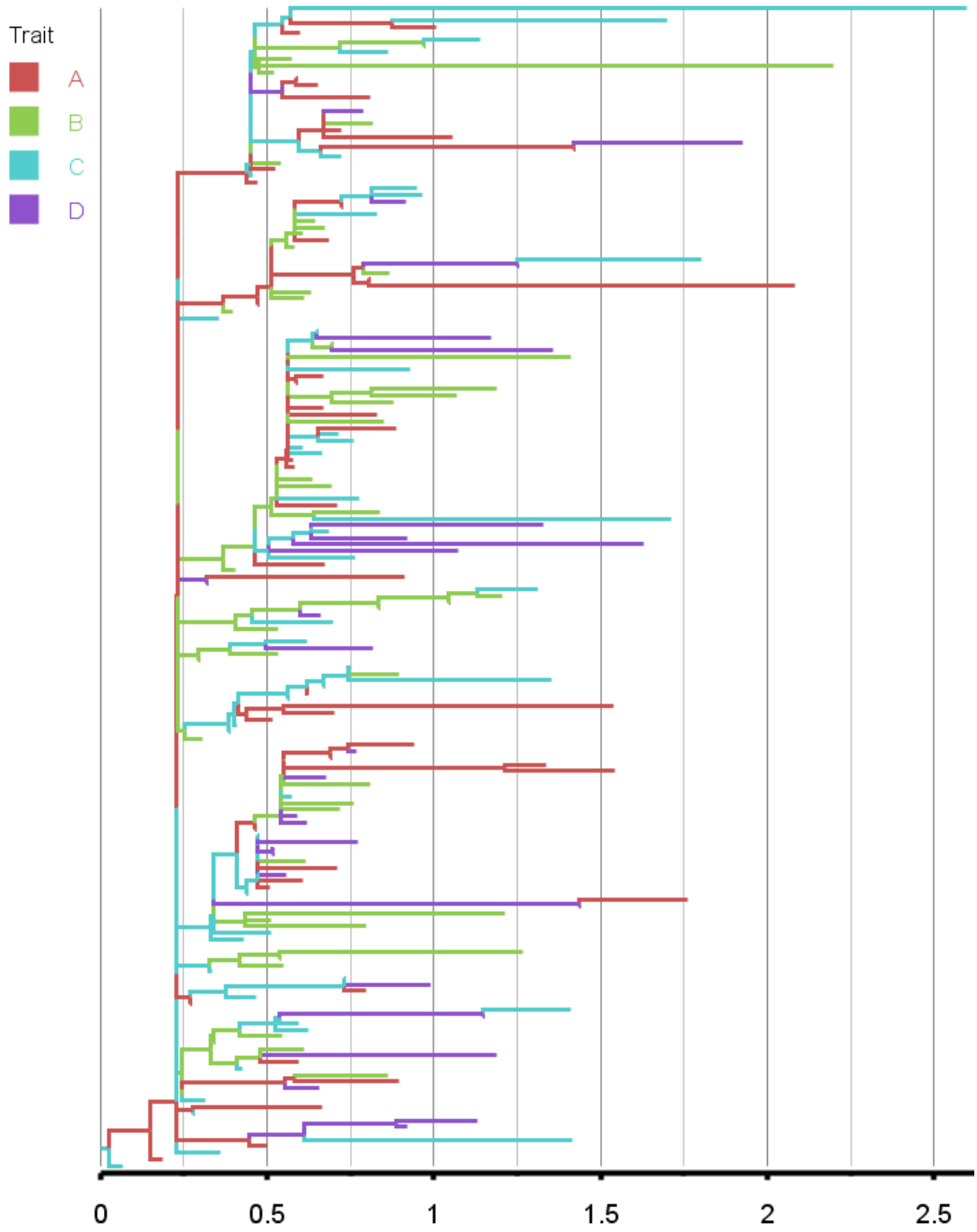
Supplementary figure 8-30: Estimated tree reconstructed through “mugration” approach for the first simulated epidemic. The phylogeny branches are coloured according to their respective host. The key for the graph is shown on the left.



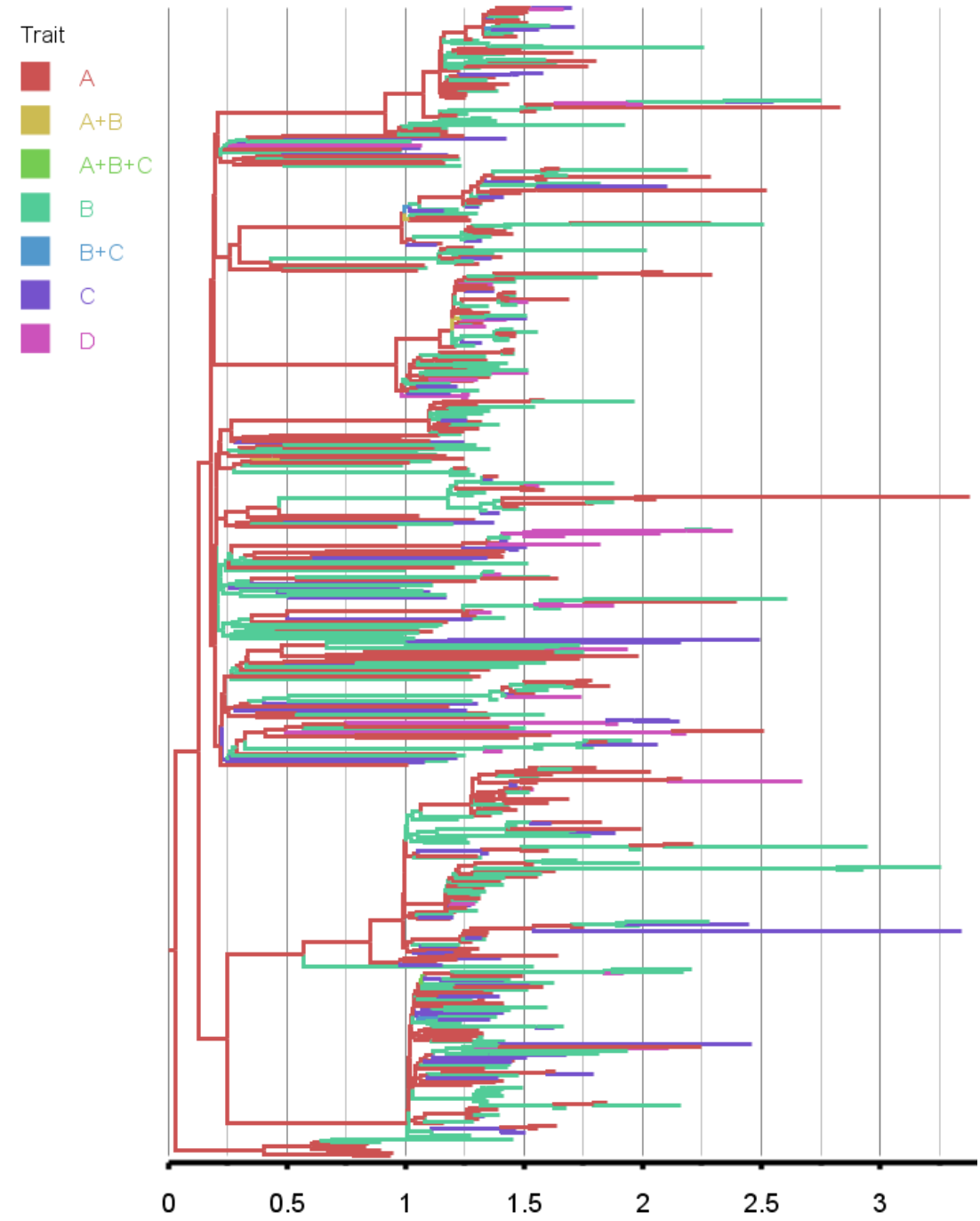
Supplementary figure 8-31: Estimated tree reconstructed through BASTA for the first simulated epidemic. The phylogeny branches are coloured according to their respective host. The key for the graph is shown on the left.



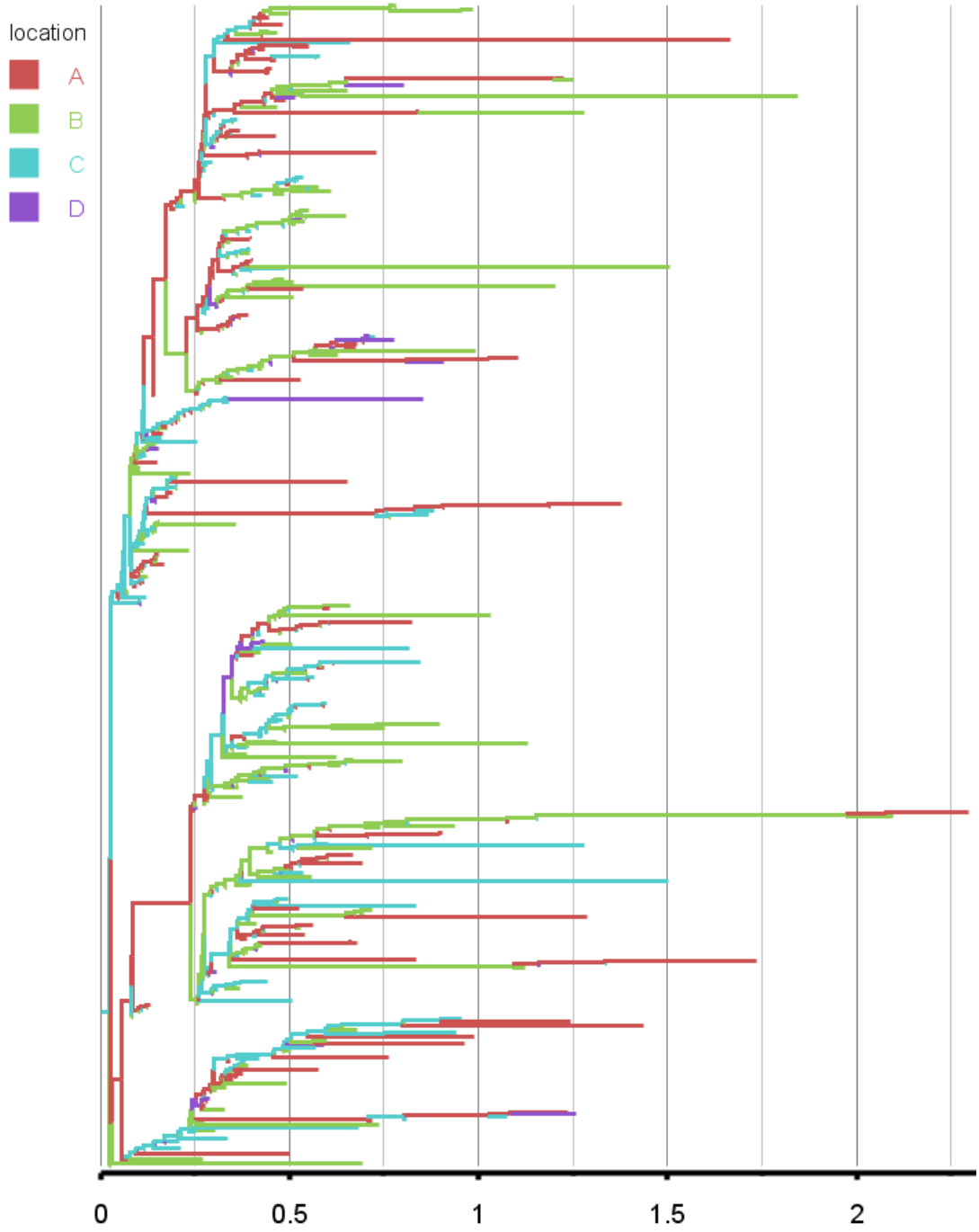
Supplementary figure 8-32: True phylogenetic tree for the second simulated epidemic. The phylogeny branches are coloured according to their respective host. The key for the graph is shown on the left.



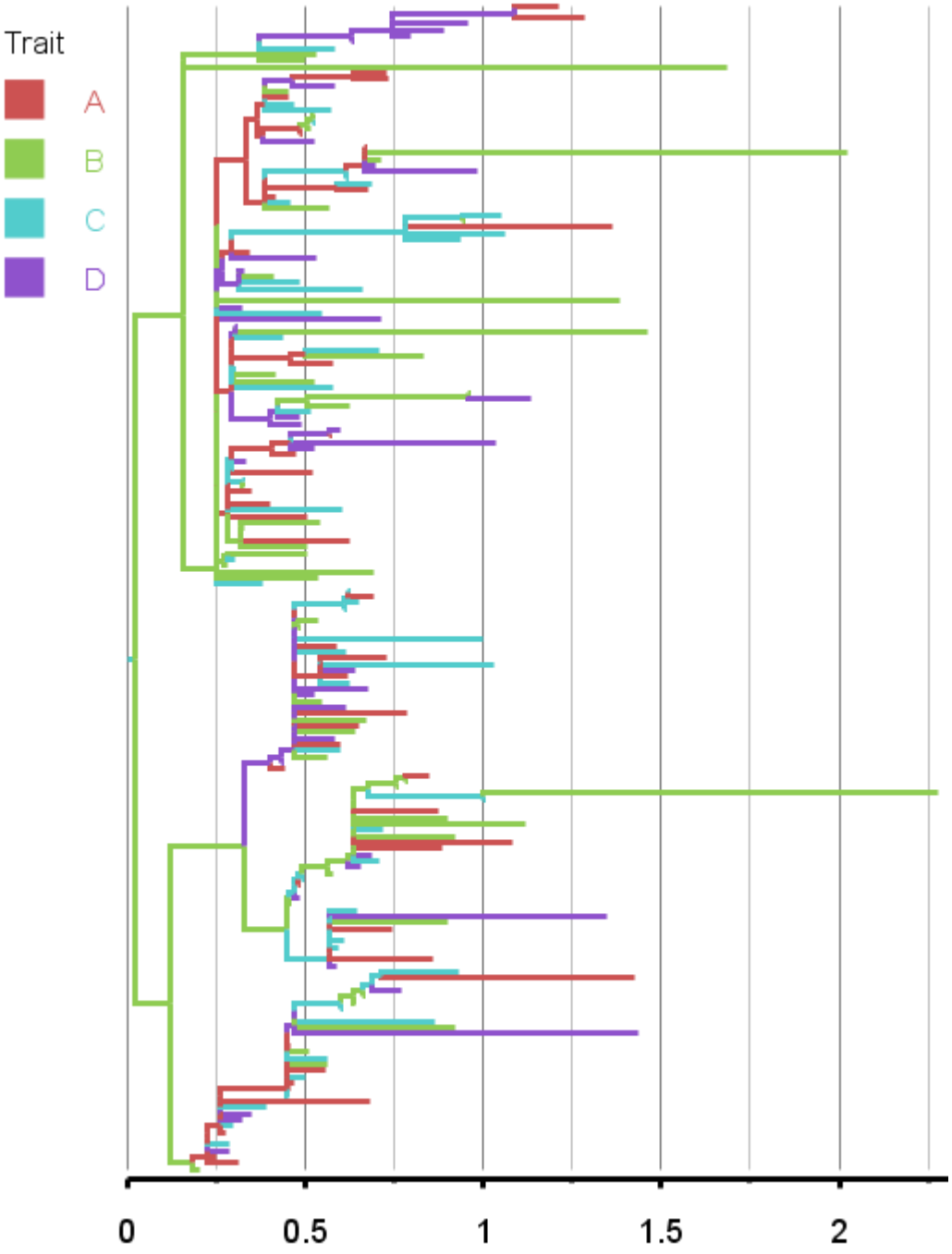
Supplementary figure 8-33: Estimated tree reconstructed by Epitree-sim for the second simulated epidemic. The phylogeny branches are coloured according to their respective host. The key for the graph is shown on the left.



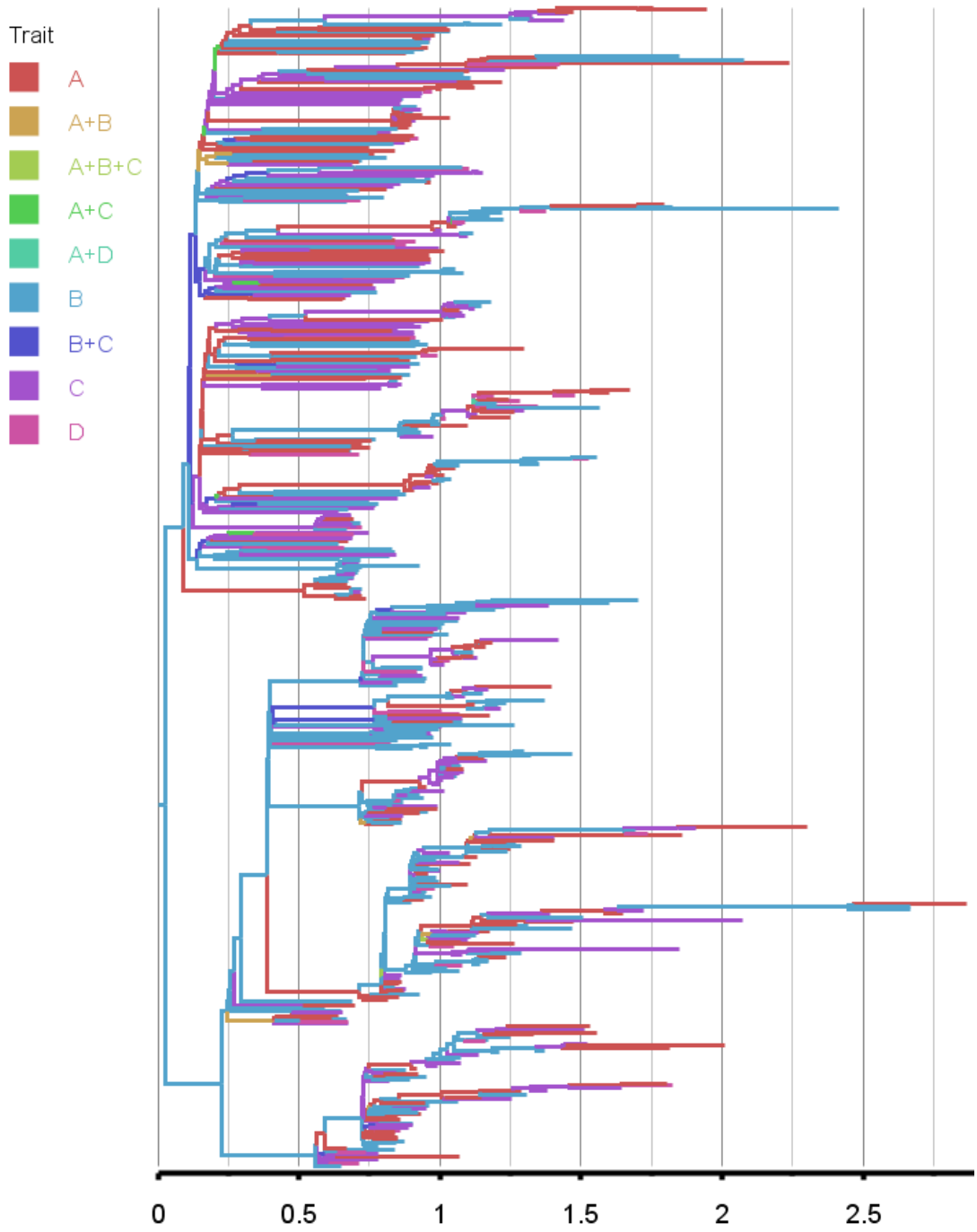
Supplementary figure 8-34: Estimated tree reconstructed through “mugration” approach for the second simulated epidemic. The phylogeny branches are coloured according to their respective host. The key for the graph is shown on the left.



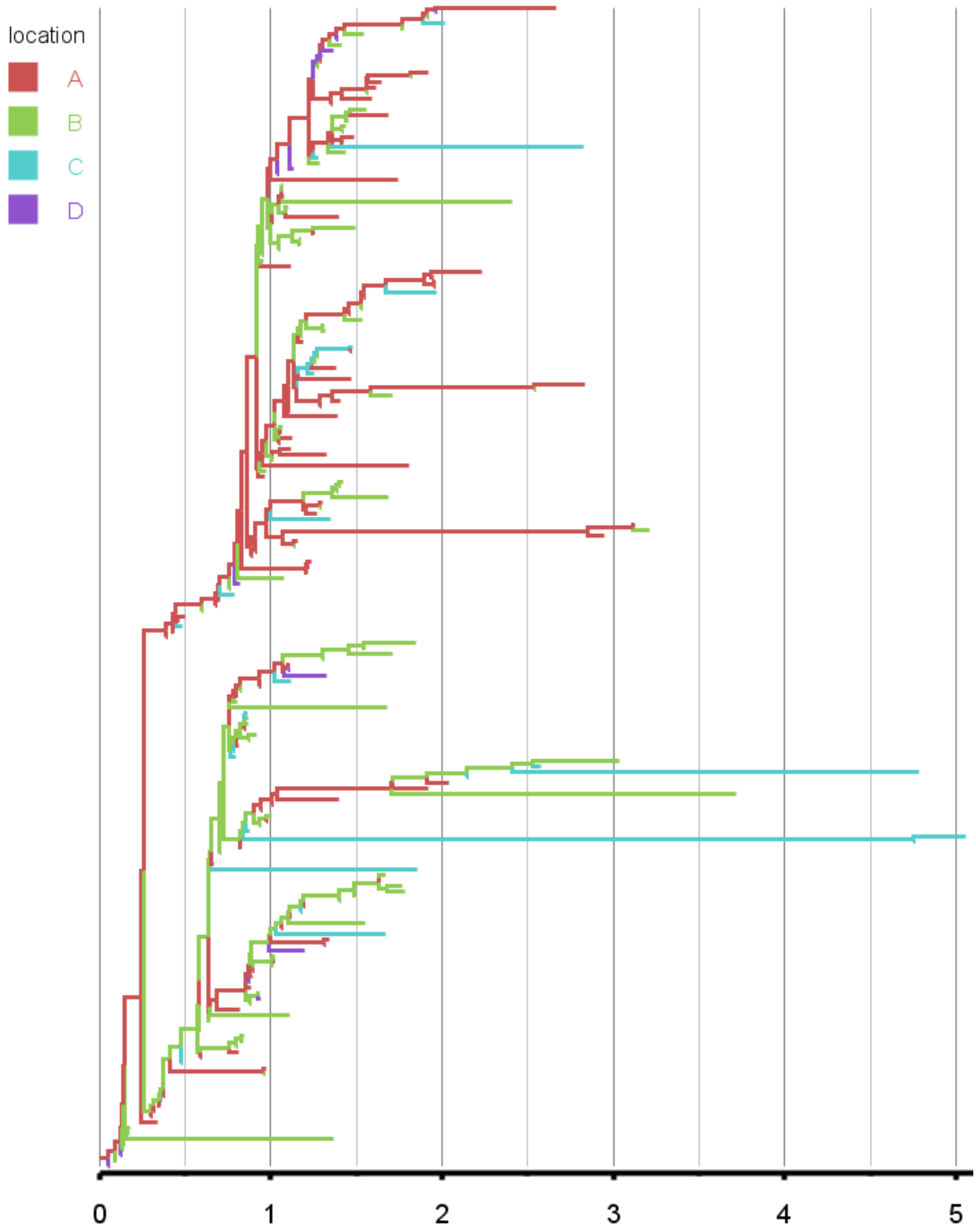
Supplementary figure 8-35: True phylogenetic tree for the third simulated epidemic. The phylogeny branches are coloured according to their respective host. The key for the graph is shown on the left.



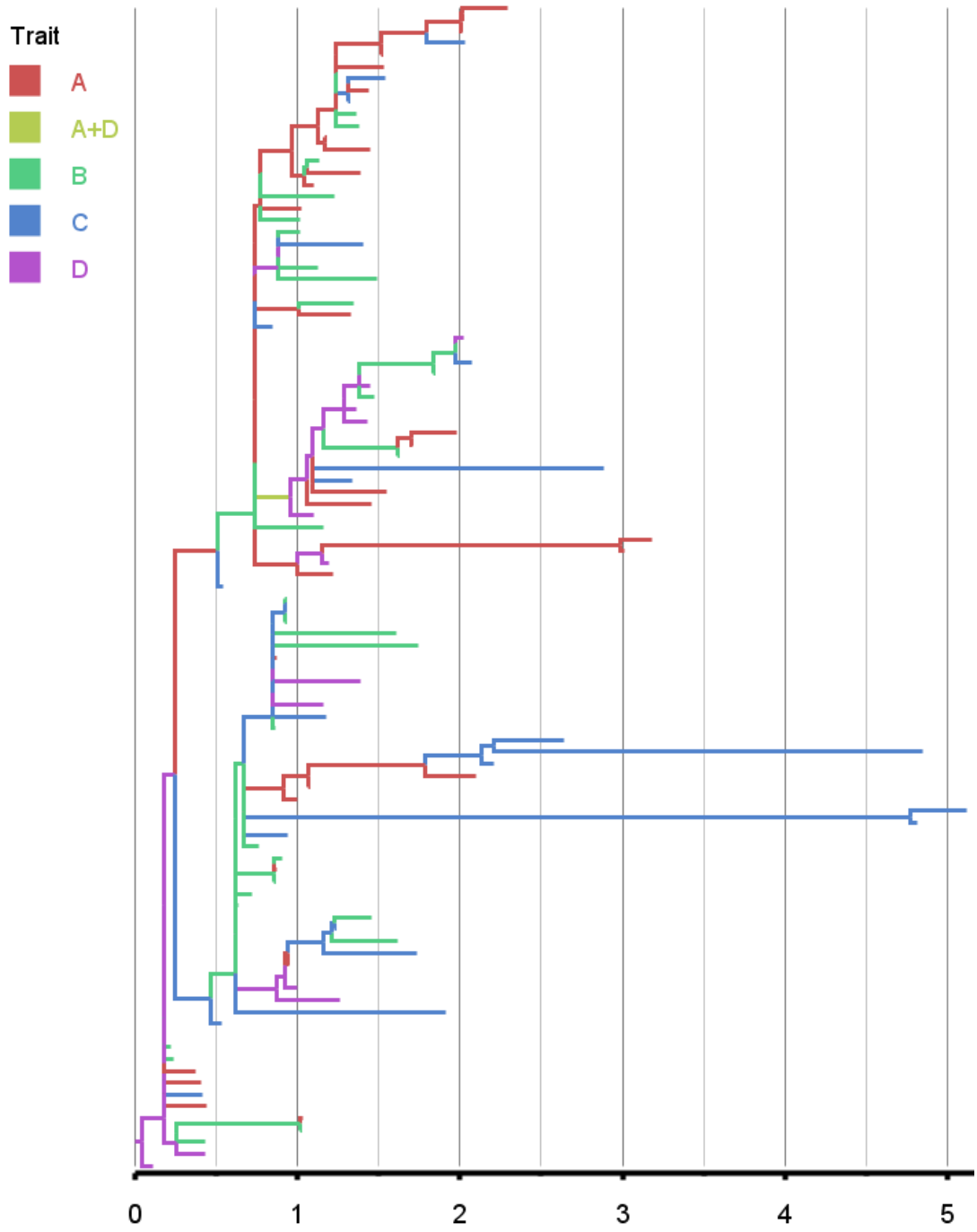
Supplementary figure 8-36: Estimated tree reconstructed by Epitree-sim for the third simulated epidemic. The phylogeny branches are coloured according to their respective host. The key for the graph is shown on the left.



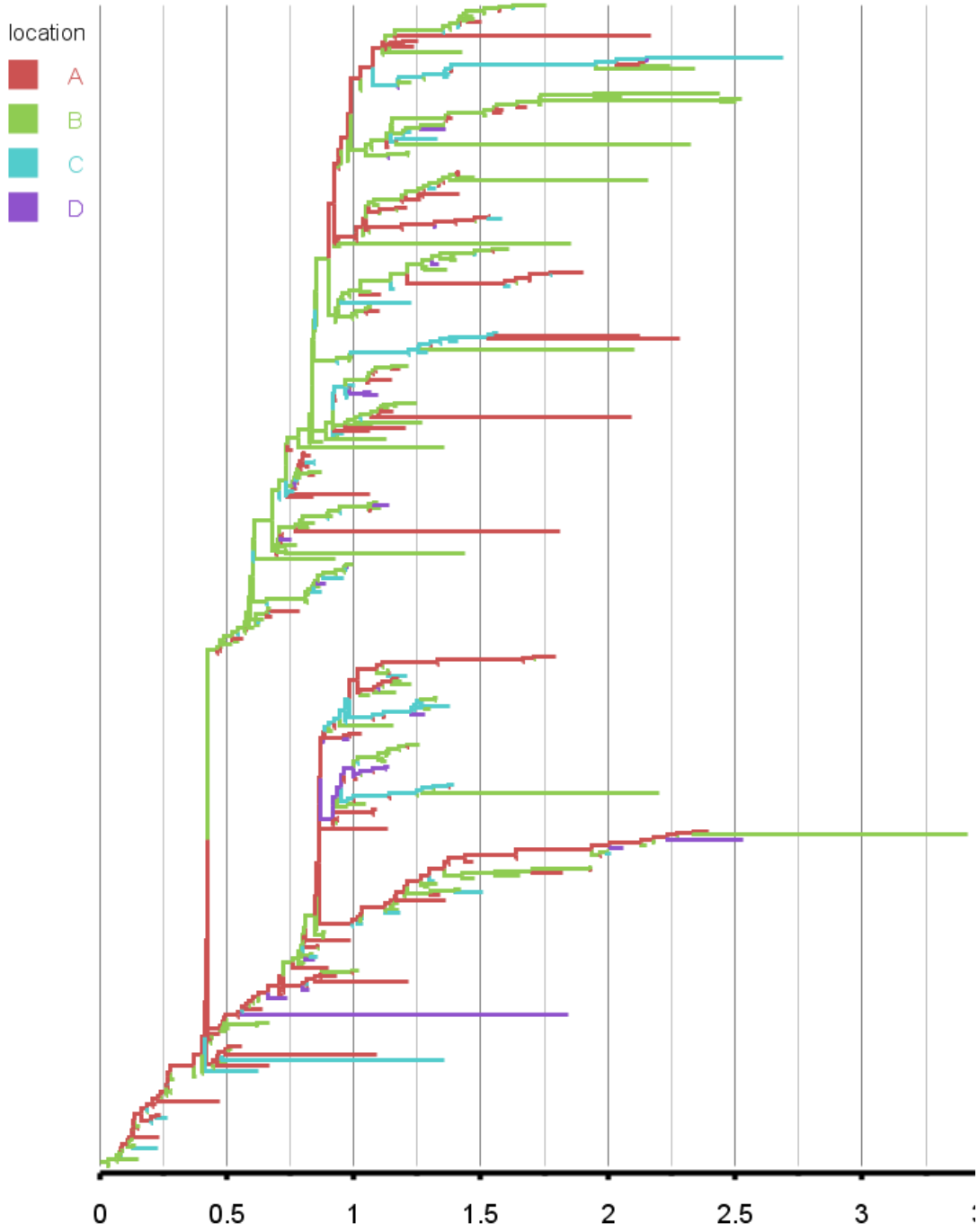
Supplementary figure 8-37: Estimated tree reconstructed through “mugration” approach for the third simulated epidemic. The phylogeny branches are coloured according to their respective host. The key for the graph is shown on the left.



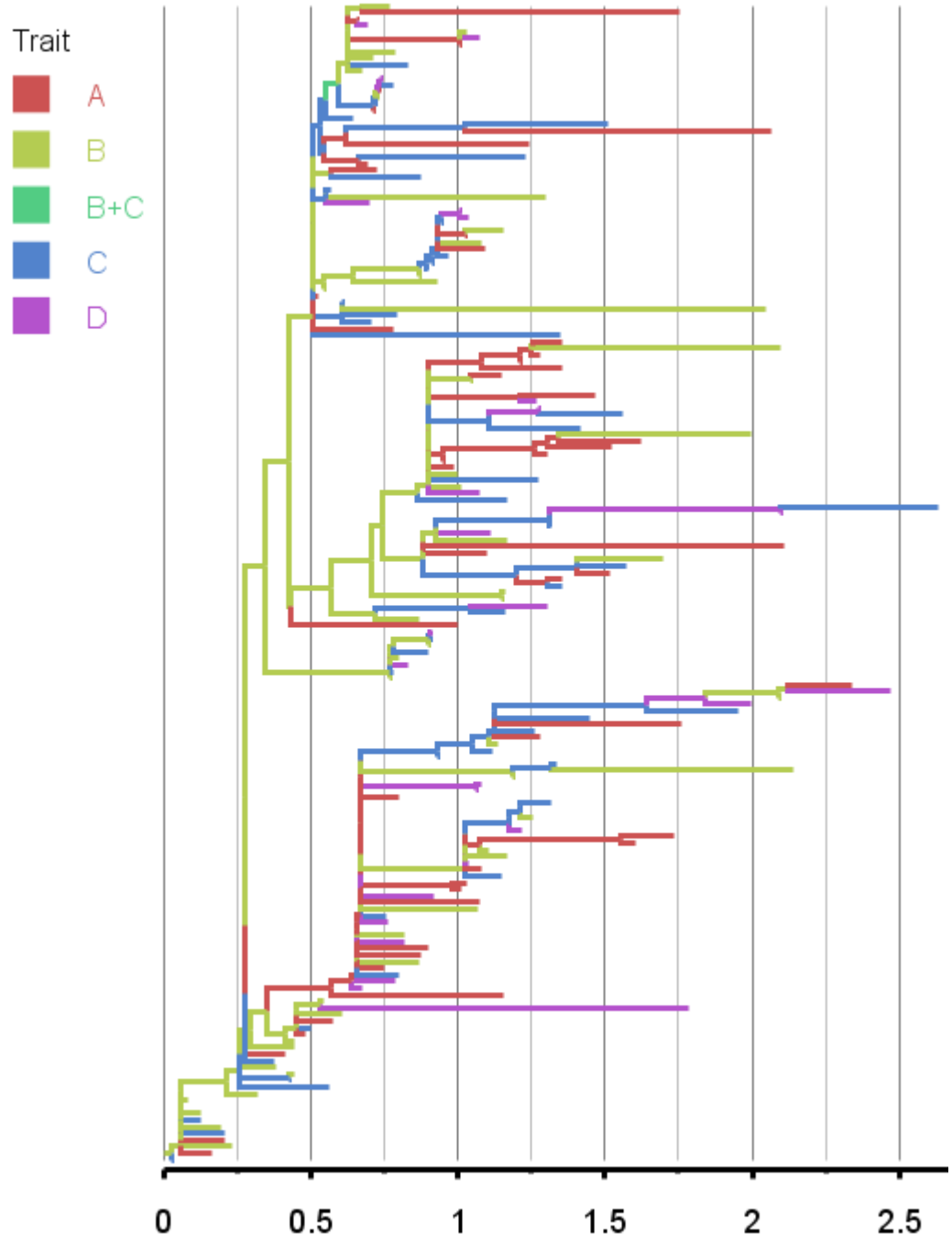
Supplementary figure 8-38: True phylogenetic tree for the fourth simulated epidemic. The phylogeny branches are coloured according to their respective host. The key for the graph is shown on the left.



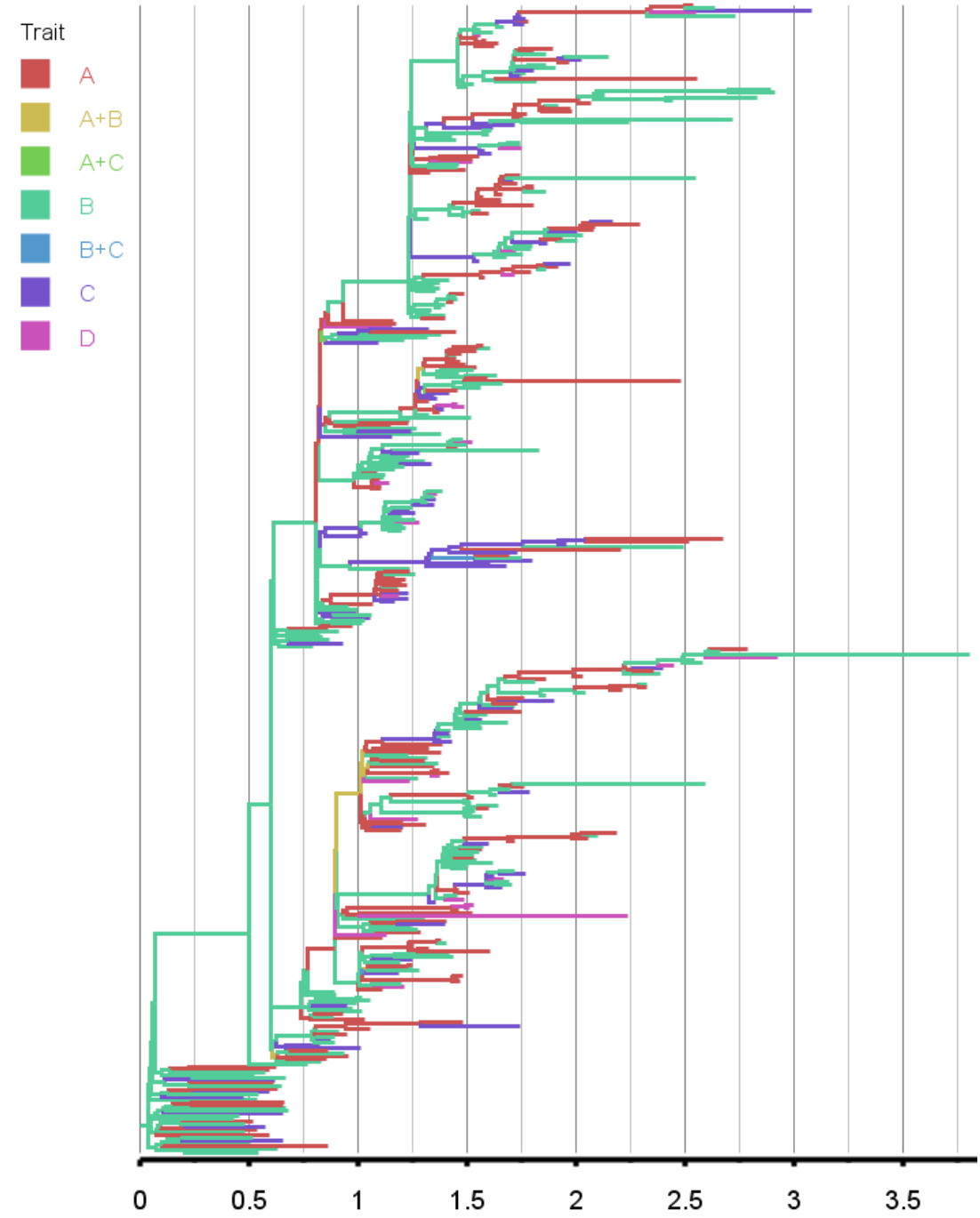
Supplementary figure 8-39: Estimated tree reconstructed through Epiteree-sim for the fourth simulated epidemic. The phylogeny branches are coloured according to their respective host. The key for the graph is shown on the left.



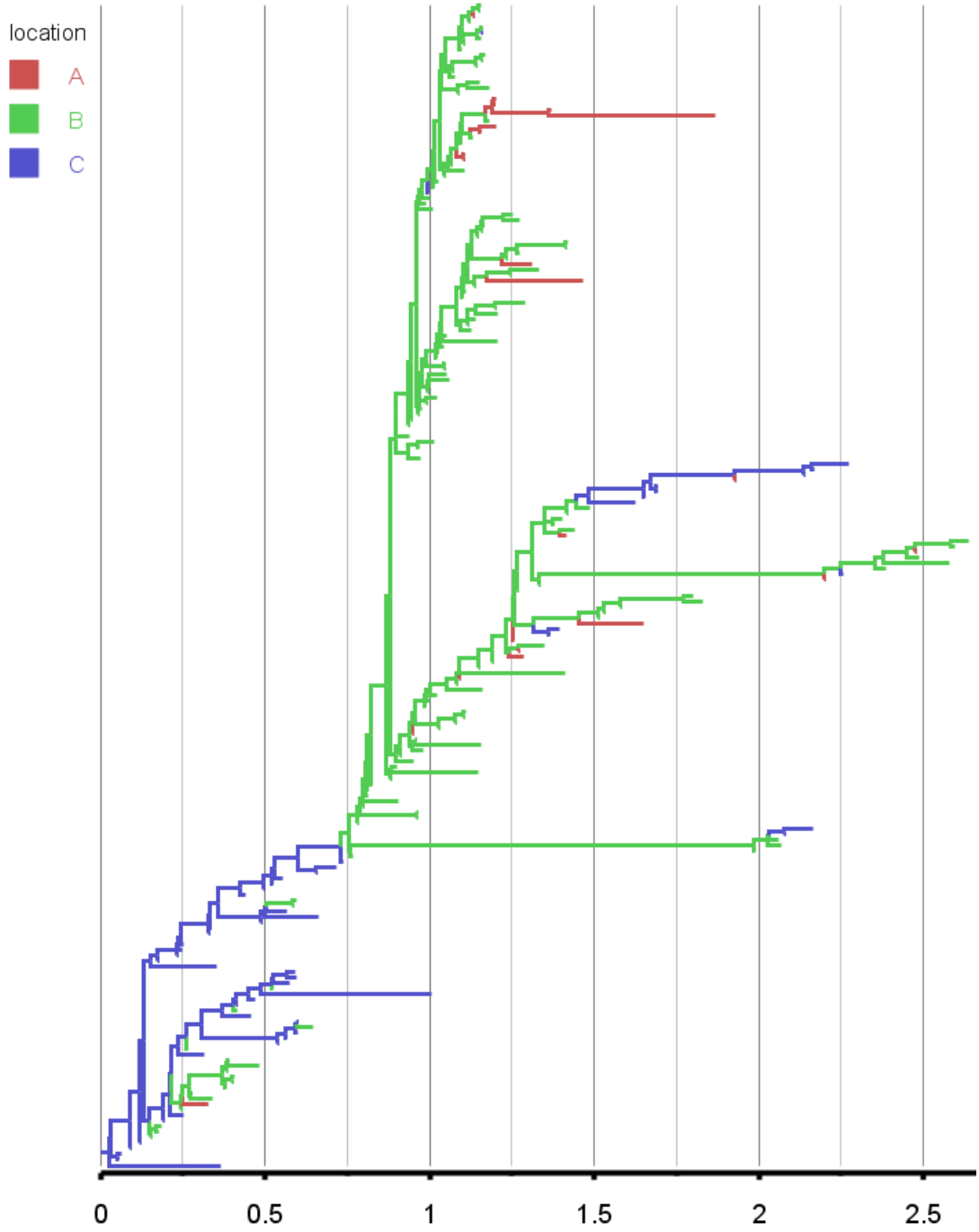
Supplementary figure 8-40: True phylogenetic tree for the fifth simulated epidemic. The phylogeny branches are coloured according to their respective host. The key for the graph is shown on the left.



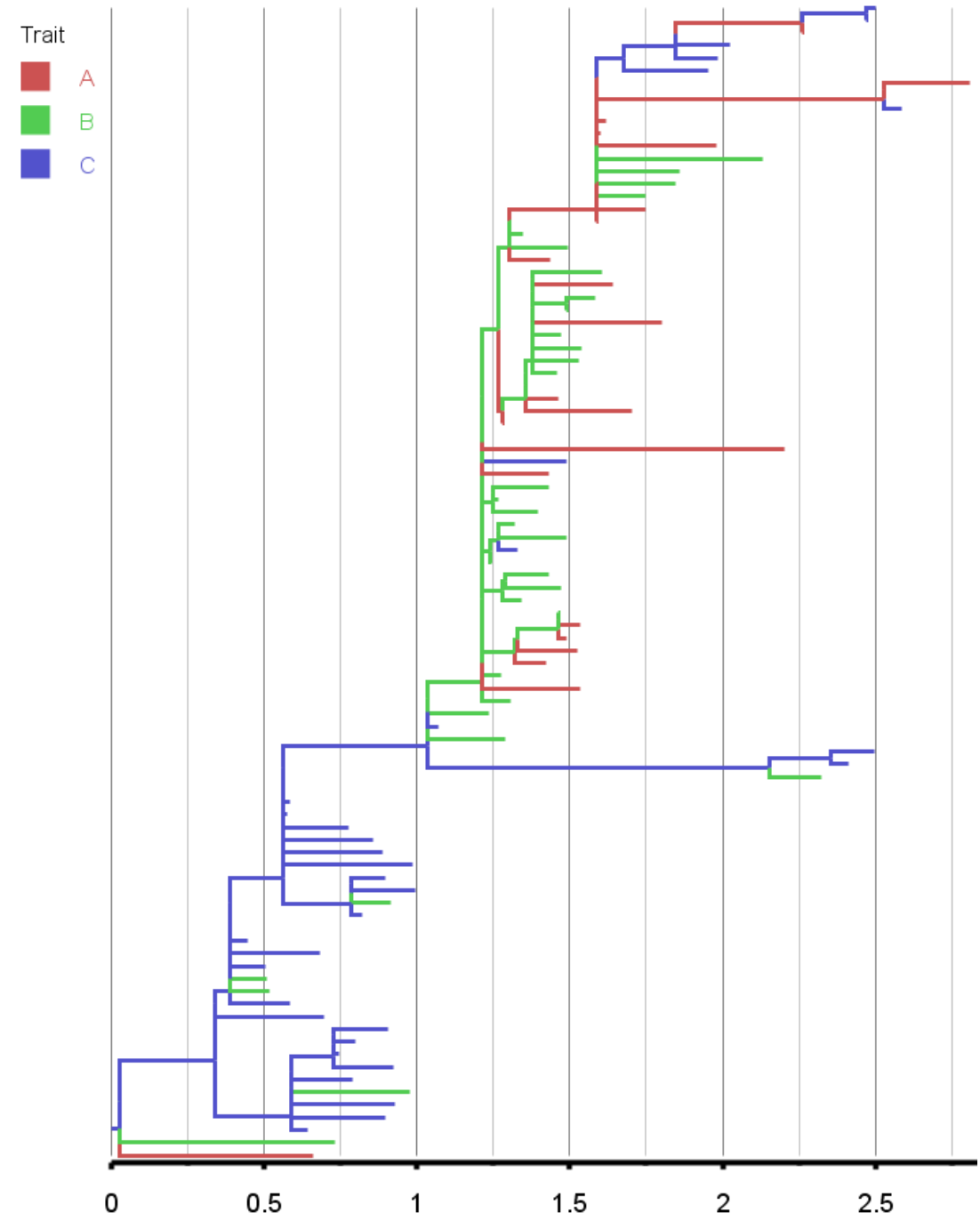
Supplementary figure 8-41: Estimated tree reconstructed through Epitree-sim for the fifth simulated epidemic. The phylogeny branches are coloured according to their respective host. The key for the graph is shown on the left.



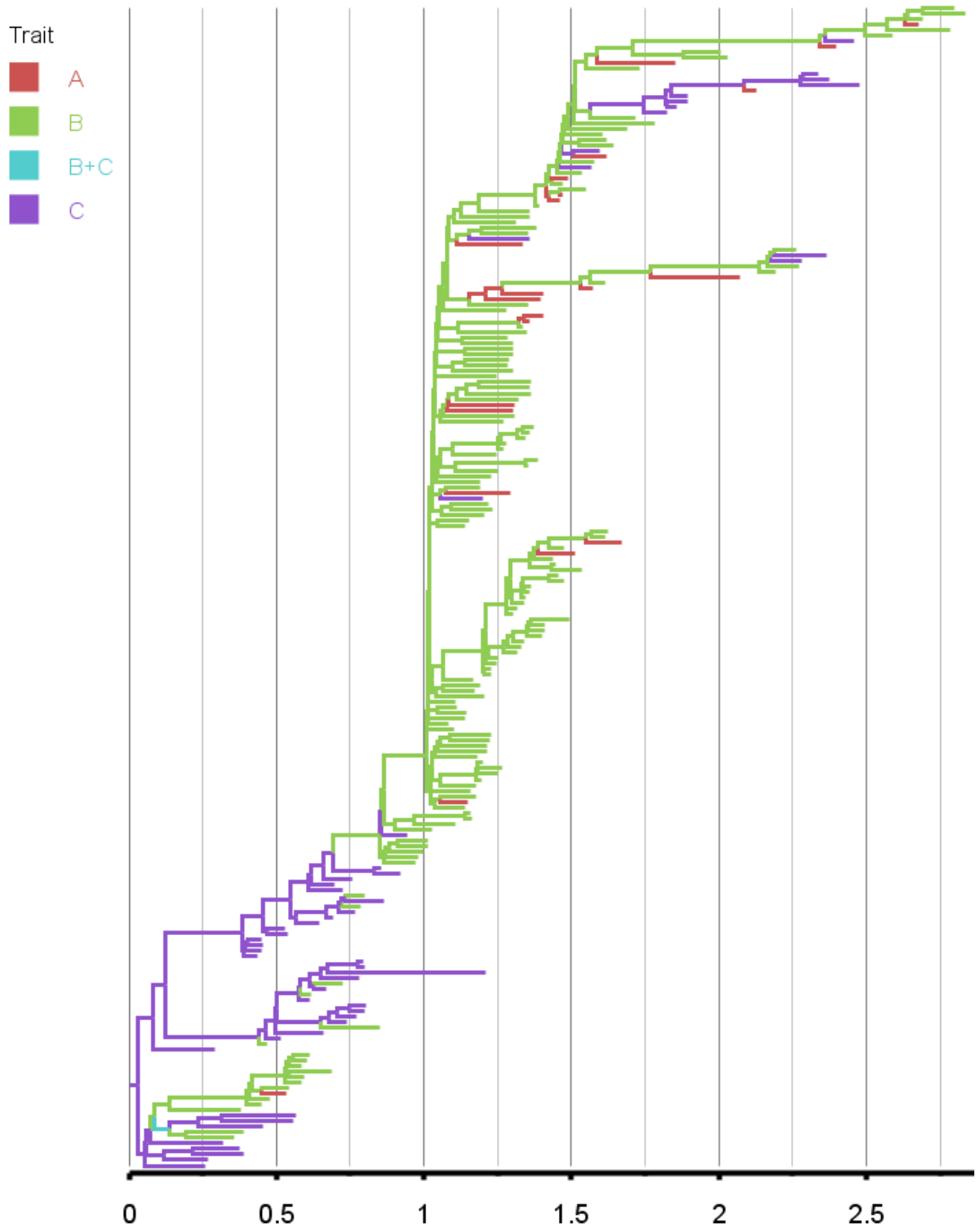
Supplementary figure 8-42: Estimated tree reconstructed through “mugration” approach for the fifth simulated epidemic. The phylogeny branches are coloured according to their respective host. The key for the graph is shown on the left.



Supplementary figure 8-43: True phylogenetic tree for the sixth simulated epidemic. The phylogeny branches are coloured according to their respective host. The key for the graph is shown on the left.



Supplementary figure 8-44: Estimated tree reconstructed through Epitree-sim for the sixth simulated epidemic. The phylogeny branches are coloured according to their respective host. The key for the graph is shown on the left.



Supplementary figure 8-45: Estimated tree reconstructed through “mugration” approach for the sixth simulated epidemic. The phylogeny branches are coloured according to their respective host. The key for the graph is shown on the left.

Supplementary table 8-46: Output of the BSSVS analysis for the simulated epidemic 1 showing the best supported rates of transition between the different populations for Epitree-sim.

Origin	Destination	Bayes factor	Posterior probability
A	B	167	0.99
A	C	42	0.95
A	D	57	0.96
B	C	19	0.89
B	D	25	0.92
C	D	15	0.87
B	A	65	0.97
C	A	17	0.88
D	A	110	0.98
C	B	125	0.98
D	B	13	0.85
D	C	30	0.93

Supplementary table 8-47: Output of the BSSVS analysis for the simulated epidemic 1 showing the best supported rates of transition between the different populations for the “mugration” approach.

Origin	Destination	Bayes factor	Posterior probability
A	B	3837	1.00
A	C	25	0.92
A	D	224	0.99
B	C	272	0.99
B	D	53	0.96
C	D	4	0.62
B	A	3837	1.00
C	A	9	0.81
D	A	12	0.84
C	B	20	0.90
D	B	8	0.77
D	C	5	0.70

Supplementary table 8-48: Output of the BSSVS analysis for the simulated epidemic 2 showing the best supported rates of transition between the different populations for Epitree-sim.

Origin	Destination	Bayes factor	Posterior probability
A	B	504	1.00
A	C	156	0.99
A	D	73	0.97
B	C	63	0.97
B	D	17	0.88

C	D	31	0.93
B	A	62	0.97
C	A	403	0.99
D	A	6	0.72
C	B	113	0.98
D	B	11	0.84
D	C	10	0.82

Supplementary table 8-49: Output of the BSSVS analysis for the simulated epidemic 2 showing the best supported rates of transition between the different populations for the “mugration” approach.

Origin	**Destination**	**Bayes factor**	**Posterior probability**
A	B	1012	1.00
A	C	1012	1.00
A	D	1012	1.00
B	C	1012	1.00
B	D	65	0.97
C	D	3	0.56
B	A	1012	1.00
C	A	21	0.90
D	A	3	0.58
C	B	9	0.80
D	B	8	0.77
D	C	2	0.50

Supplementary table 8-50: Output of the BSSVS analysis for the simulated epidemic 3 showing the best supported rates of transition between the different populations for Epitree-sim

Origin	**Destination**	**Bayes factor**	**Posterior probability**
A	B	48	0.96
A	C	49	0.96
A	D	47	0.95
B	C	60	0.96
B	D	47	0.95
C	D	89	0.98
B	A	42	0.95
C	A	209	0.99
D	A	43	0.95
C	B	223	0.99
D	B	25	0.92
D	C	23	0.91

Supplementary table 8-51: Output of the BSSVS analysis for the simulated epidemic 3 showing the best supported rates of transition between the different populations for the “mugration” approach

Origin	Destination	Bayes factor	Posterior probability
A	B	1012	1.00
A	C	1012	1.00
A	D	54	0.96
B	C	1012	1.00
B	D	1012	1.00
C	D	25	0.92
B	A	1012	1.00
C	A	1012	1.00
D	A	6	0.73
C	B	1012	1.00
D	B	5	0.68
D	C	4	0.63

Supplementary table 8-52: Output of the BSSVS analysis for the simulated epidemic 4 showing the best supported rates of transition between the different populations for Epitree-sim

Origin	Destination	Bayes factor	Posterior probability
A	B	30	0.93
A	C	373	0.99
A	D	9	0.79
B	C	48	0.96
B	D	35	0.94
C	D	20	0.90
B	A	154	0.99
C	A	31	0.93
D	A	103	0.98
C	B	209	0.99
D	B	271	0.99
D	C	31	0.93

Supplementary table 8-53: Output of the BSSVS analysis for the simulated epidemic 4 showing the best supported rates of transition between the different populations for the “mugration” approach

Origin	Destination	Bayes factor	Posterior probability
A	B	1012	1.00
A	C	1012	1.00
A	D	20	0.90
B	C	125	0.98

B	D	143	0.98
C	D	1	0.38
B	A	1012	1.00
C	A	3	0.57
D	A	4	0.64
C	B	10	0.82
D	B	5	0.71
D	C	3	0.56

Supplementary table 8-54; Output of the BSSVS analysis for the simulated epidemic 5 showing the best supported rates of transition between the different populations for Epitree-sim

Origin	**Destination**	**Bayes factor**	**Posterior probability**
A	B	58	0.96
A	C	154	0.99
A	D	175	0.99
B	C	264	0.99
B	D	23	0.91
C	D	264	0.99
B	A	458	1.00
C	A	46	0.95
D	A	21	0.90
C	B	287	0.99
D	B	13	0.85
D	C	17	0.88

Supplementary table 8-55: Output of the BSSVS analysis for the simulated epidemic 5 showing the best supported rates of transition between the different populations for the “mugration” approach

Origin	**Destination**	**Bayes factor**	**Posterior probability**
A	B	1012	1.00
A	C	1012	1.00
A	D	1012	1.00
B	C	1012	1.00
B	D	1012	1.00
C	D	4	0.65
B	A	1012	1.00
C	A	57	0.96
D	A	2	0.52
C	B	1012	1.00
D	B	3	0.58
D	C	3	0.55

Origin	Destination	Rate
A	B	1.869
A	C	1.252
A	D	1.309
B	C	1.061
B	D	1.064
C	D	0.784
B	A	1.606
C	A	0.985
D	A	1.703
C	B	1.552
D	B	0.982
D	C	1.24

Supplementary table 8-56: Output of the BSSVS analysis for the simulated epidemic showing the best supported rates of transition between the different populations for Epitree-sim

Origin	**Destination**	**Bayes factor**	**Posterior probability**
A	B	50	0.98
A	C	367	1.00
B	C	9	0.88
B	A	5525	1.00
C	A	3	0.68
C	B	501	1.00

Supplementary table 8-57: Output of the BSSVS analysis for the simulated epidemic showing the best supported rates of transition between the different populations for the “mugration” approach

Origin	**Destination**	**Bayes factor**	**Posterior probability**
A	B	11	0.91
A	C	7	0.86
B	C	552	1
B	A	552	1
C	A	4	0.77
C	B	552	1

Supplementary table 8-58: Comparison of the estimated rates values for the first simulated epidemic between Epitree-sim, “mugration” and BASTA analyses between the different existing populations

Origin	**Destination**	**Rate subsampling**	**Rate mugration**	**Rate BASTA**
A	B	1.869	2.721	2.188
A	C	1.252	0.682	1.082
A	D	1.309	0.965	1.213

B	C	1.061	1.034	2.302
B	D	1.064	0.662	1.195
C	D	0.784	0.573	0.946
B	A	1.606	2.613	1.034
C	A	0.985	0.672	1.134
D	A	1.703	0.777	0.96
C	B	1.552	0.795	1.428
D	B	0.982	0.64	1.299
D	C	1.24	0.602	0.81

Supplementary table 8-59: Comparison of the estimated rates values for the second simulated epidemic between Epitree-sim and "mugration" analysis between the different existing populations

Origin	**Destination**	**Rate Epitree-sim**	**Rate Mugration**
A	B	3.08	4.393
A	C	2.844	1.459
A	D	1.62	0.566
B	C	2.392	1.343
B	D	1.496	0.567
C	D	1.627	0.756
B	A	2.422	3.675
C	A	2.856	0.895
D	A	1.184	0.74
C	B	2.448	0.842
D	B	1.244	0.616
D	C	1.162	0.892

Supplementary table 8-60: Comparison of the estimated rates values for the third simulated epidemic between Epitree-sim and "mugration" analysis between the different existing populations

Origin	**Destination**	**Rate Epitree-sim**	**Rate Mugration**
A	B	2.27	2.351
A	C	2.386	1.679
A	D	1.898	0.564
B	C	2.725	2.93
B	D	2.097	0.859
C	D	2.306	0.565
B	A	2.318	2.293
C	A	3.037	2.218
D	A	2.094	0.719
C	B	3.106	2.619
D	B	1.994	0.678
D	C	2.018	0.749

Supplementary table 8-61: Comparison of the estimated rates values for the fourth simulated epidemic between Epitree-sim and “mugration” analysis between the different existing populations.

Origin	Destination	Rate Epitree-sim	Rate Mugration
A	B	1.599	3.447
A	C	1.844	1.13
A	D	0.896	0.652
B	C	1.448	0.818
B	D	1.13	0.892
C	D	0.895	0.796
B	A	2.259	3.913
C	A	1.178	0.731
D	A	1.799	0.672
C	B	1.842	0.551
D	B	1.793	0.688
D	C	1.129	0.677

Supplementary table 8-62: Comparison of the estimated rates values for the fifth simulated epidemic between Epitree-sim and “mugration” analysis between the different existing populations.

Origin	Destination	Rate Epitree-sim	Rate Mugration
A	B	2.058	2.603
A	C	2.277	1.265
A	D	1.637	0.616
B	C	2.981	1.729
B	D	1.112	0.546
C	D	1.453	0.615
B	A	2.84	3.085
C	A	1.985	1.087
D	A	1.376	0.737
C	B	2.456	1.784
D	B	1.353	0.619
D	C	1.308	0.843

Supplementary table 8-63: Comparison of the estimated rates values for the sixth simulated epidemic between Epitree-sim and “mugration” analysis between the different existing populations.

Origin	Destination	Rate Epitree-sim	Rate Mugration
A	B	1.334	0.845
A	C	1.235	0.613
B	C	0.732	1.032
B	A	2.547	2.05
C	A	0.657	0.478
C	B	1.141	1.808

Supplementary table 8-64: Output of the BSSVS analysis for the avian influenza dataset showing the best supported rates of transition between the sampled hosts for the subsampling approach

Origin	Destination	Bayes factor	Posterior probability
Domestic Anseriformes	Domestic Galliformes	138	0.98
Wild Anseriformes	Wild Charadriiforms	133	0.98
Domestic Anseriformes	Wild Anseriformes	74	0.96
Domestic Galliformes	Domestic Anseriformes	71	0.96
Wild Anseriformes	Domestic Anseriformes	68	0.95
Wild Anseriformes	Domestic Galliformes	58	0.95
Domestic Galliformes	Wild others	24	0.88
Domestic Galliformes	Wild Anseriformes	20	0.86
Domestic Anseriformes	Wild others	11	0.77

Supplementary table 8-65: Output of the BSSVS analysis for the avian influenza dataset showing the best supported rates of transition between the sampled hosts for the “mugration” approach

Origin	Destination	Bayes factor	Posterior probability
Domestic Anseriformes	Wild Anseriformes	35225	1.00
Wild Anseriformes	Wild Charadriiforms	35225	1.00
Wild Anseriformes	Domestic Anseriformes	35225	1.00
Wild Anseriformes	Domestic Galliformes	35225	1.00
Wild Anseriformes	Wild others	2707	1.00
Domestic Anseriformes	Domestic Galliformes	1528	1.00
Domestic Galliformes	Domestic Anseriformes	54	0.94
Domestic Anseriformes	Wild others	16	0.83
Domestic Galliformes	Wild others	13	0.80

Supplementary table 8-66: Comparison of the estimated rates values for the avian influenza dataset between the subsampling and “mugration” analysis for the host analysis

Origin	Destination	Value rate subsampling	Value rate “mugration”
Domestic Anseriformes	Domestic Galliformes	1.633	1.28
Domestic Anseriformes	Wild Anseriformes	2.954	3.26
Domestic Anseriformes	Wild Charadriiforms	0.919	0.86
Domestic Anseriformes	Wild others	0.68	0.53
Domestic Galliformes	Wild Anseriformes	1.168	0.81
Domestic Galliformes	Wild Charadriiforms	0.87	0.84
Domestic Galliformes	Wild others	0.976	0.69
Wild Anseriformes	Wild Charadriiforms	1.633	1.32
Wild Anseriformes	Wild others	0.902	0.63

Wild Charadriiforms	Wild others	0.861	0.78
Domestic Galliformes	Domestic Anseriformes	1.415	1.00
Wild Anseriformes	Domestic Anseriformes	2.288	2.55
Wild Charadriiforms	Domestic Anseriformes	0.804	0.78
Wild others	Domestic Anseriformes	0.89	0.66
Wild Anseriformes	Domestic Galliformes	2.215	1.93
Wild Charadriiforms	Domestic Galliformes	0.887	0.73
Wild others	Domestic Galliformes	0.887	0.71
Wild Charadriiforms	Wild Anseriformes	0.969	0.71
Wild others	Wild Anseriformes	0.82	0.77
Wild others	Wild Charadriiforms	0.966	0.80

Supplementary table 8-67: Comparison of the estimated indicator values for the avian influenza dataset between the subsampling and “mugration” analysis for the host analysis

Origin	**Destination**	**Value indicator subsampling**	**Value indicator “mugration”**
Domestic Anseriformes	Domestic Galliformes	0.956	1.00
Domestic Anseriformes	Wild Anseriformes	0.989	1.00
Domestic Anseriformes	Wild Charadriiforms	0.404	0.25
Domestic Anseriformes	Wild others	0.843	0.83
Domestic Galliformes	Wild Anseriformes	0.794	0.71
Domestic Galliformes	Wild Charadriiforms	0.499	0.27
Domestic Galliformes	Wild others	0.829	0.80
Wild Anseriformes	Wild Charadriiforms	1	1.00
Wild Anseriformes	Wild others	0.971	1.00
Wild Charadriiforms	Wild others	0.512	0.37
Domestic Galliformes	Domestic Anseriformes	0.958	0.94
Wild Anseriformes	Domestic Anseriformes	0.971	1.00
Wild Charadriiforms	Domestic Anseriformes	0.479	0.39
Wild others	Domestic Anseriformes	0.612	0.60
Wild Anseriformes	Domestic Galliformes	0.996	1.00
Wild Charadriiforms	Domestic Galliformes	0.486	0.45
Wild others	Domestic Galliformes	0.57	0.48

Wild Charadriiforms	Wild Anseriformes	0.55	0.56
Wild others	Wild Anseriformes	0.519	0.39
Wild others	Wild Charadriiforms	0.417	0.35

Supplementary table 8-68: Output of the BSSVS analysis for the avian influenza dataset showing the best supported rates of transition between the sampled locations for the subsampling approach

Origin	**Destination**	**Bayes factor**	**Posterior probability**
East China	Eastern Asia	210	0.97
Western Europe	Northern Europe	72	0.91
Western Europe	Eastern Europe	68	0.90
Eastern Europe	Southern Europe	56	0.88
East China	North-Central Asia	44	0.86
South Asia	Eastern Asia	32	0.81
Southern Europe	Western Europe	29	0.80
South Asia	Central Asia	23	0.76
Western Europe	Southern Europe	23	0.76
South Asia	East China	18	0.71
Western Europe	Central Asia	16	0.69
Eastern Europe	Northern Europe	16	0.69
South Asia	North-Central Asia	16	0.69
Central Asia	Eastern Europe	14	0.66

Supplementary table 8-69: Output of the BSSVS analysis for the avian influenza dataset showing the best supported rates of transition between the sampled locations for the “mugration” approach

Origin	**Destination**	**Bayes factor**	**Posterior probability**
East China	Eastern Asia	7464	1.00
Southern Europe	Western Europe	7464	1.00
Western Europe	Eastern Europe	7464	1.00
Western Europe	Northern Europe	7464	1.00
East China	North Central Asia	491	0.99
Eastern Europe	Southern Europe	318	0.98
South Asia	Eastern Asia	163	0.96
South Asia	Central Asia	96	0.93
Eastern Europe	Central Asia	89	0.92
Southern Europe	East China	89	0.92
East China	South Asia	82	0.92
East China	Western Europe	20	0.73
Eastern Europe	Northern Europe	13	0.64

Supplementary table 8-70: Comparison of the estimated rates values for the avian influenza dataset between the subsampling and “mugration” analysis for the location analysis

Origin	Destination	Value rate subsampling	Value rate “mugration”
Central Asia	East China	1.00	0.89
Central Asia	Eastern Asia	1.05	0.92
Central Asia	Eastern Europe	0.93	0.92
Central Asia	North Central Asia	0.94	0.93
Central Asia	Northern Europe	1.05	0.90
Central Asia	South Asia	1.01	0.93
Central Asia	Southern Europe	1.01	0.92
Central Asia	Western Europe	0.96	0.74
East China	Eastern Asia	1.77	1.54
East China	Eastern Europe	1.05	0.90
East China	North Central Asia	0.90	0.98
East China	Northern Europe	0.92	1.01
East China	South Asia	1.27	1.65
East China	Southern Europe	1.00	0.94
East China	Western Europe	1.01	0.79
Eastern Asia	Eastern Europe	1.00	0.97
Eastern Asia	North Central Asia	1.02	0.99
Eastern Asia	Northern Europe	1.05	1.00
Eastern Asia	South Asia	0.98	0.90
Eastern Asia	Southern Europe	0.99	0.95
Eastern Asia	Western Europe	1.04	0.96
Eastern Europe	North Central Asia	0.93	0.80
Eastern Europe	Northern Europe	1.06	0.85
Eastern Europe	South Asia	1.02	1.00
Eastern Europe	Southern Europe	1.93	1.81
Eastern Europe	Western Europe	1.28	0.96
North Central Asia	Northern Europe	1.01	0.93
North Central Asia	South Asia	1.03	0.81
North Central Asia	Southern Europe	1.02	0.96
North Central Asia	Western Europe	0.99	0.93
Northern Europe	South Asia	1.01	1.01
Northern Europe	Southern Europe	1.01	0.91
Northern Europe	Western Europe	0.95	0.94
South Asia	Southern Europe	0.98	0.93
South Asia	Western Europe	1.16	1.00
Southern Europe	Western Europe	1.40	2.08
EastChina	Central Asia	1.07	0.74
Eastern Asia	Central Asia	1.02	0.95

Eastern Europe	Central Asia	1.04	1.27
North Central Asia	Central Asia	1.01	0.90
Northern Europe	Central Asia	1.01	0.95
South Asia	Central Asia	1.39	1.16
Southern Europe	Central Asia	1.05	0.97
Western Europe	Central Asia	0.99	0.85
Eastern Asia	East China	0.92	0.80
Eastern Europe	East China	0.95	0.92
North Central Asia	East China	1.01	0.86
Northern Europe	East China	0.92	0.92
South Asia	East China	1.95	1.17
Southern Europe	East China	1.03	1.36
Western Europe	East China	0.94	0.81
Eastern Europe	Eastern Asia	1.05	1.01
North Central Asia	Eastern Asia	0.91	0.94
Northern Europe	Eastern Asia	1.05	0.94
South Asia	Eastern Asia	1.52	1.10
Southern Europe	Eastern Asia	1.03	0.95
Western Europe	Eastern Asia	1.07	0.99
North Central Asia	Eastern Europe	1.01	0.89
Northern Europe	Eastern Europe	0.95	0.88
South Asia	Eastern Europe	1.02	0.94
Southern Europe	Eastern Europe	1.02	0.81
Western Europe	Eastern Europe	2.63	3.30
Northern Europe	North Central Asia	1.01	0.95
South Asia	North Central Asia	1.12	0.95
Southern Europe	North Central Asia	1.03	0.92
Western Europe	North Central Asia	0.96	0.85
South Asia	Northern Europe	1.05	1.04
Southern Europe	Northern Europe	0.98	0.92
Western Europe	Northern Europe	2.18	1.84
Southern Europe	South Asia	1.01	0.98
Western Europe	South Asia	1.00	1.04
Western Europe	Southern Europe	1.44	0.87

Supplementary table 8-71: Comparison of the estimated indicator values for the avian influenza dataset between the subsampling and “mugration” analysis for the location analysis

Origin	**Destination**	**Value indicator subsampling**	**Value indicator “mugration”**
Central Asia	East China	0.15	0.21
Central Asia	Eastern Asia	0.09	0.09
Central Asia	Eastern Europe	0.74	0.18

Central Asia	North Central Asia	0.48	0.19
Central Asia	Northern Europe	0.10	0.07
Central Asia	South Asia	0.13	0.15
Central Asia	Southern Europe	0.13	0.13
Central Asia	Western Europe	0.33	0.52
East China	Eastern Asia	0.99	1.00
East China	Eastern Europe	0.14	0.17
East China	North Central Asia	0.92	0.99
East China	Northern Europe	0.21	0.08
East China	South Asia	0.30	0.94
East China	Southern Europe	0.17	0.10
East China	Western Europe	0.33	0.75
Eastern Asia	Eastern Europe	0.12	0.10
Eastern Asia	North Central Asia	0.18	0.08
Eastern Asia	Northern Europe	0.09	0.07
Eastern Asia	South Asia	0.24	0.23
Eastern Asia	Southern Europe	0.13	0.11
Eastern Asia	Western Europe	0.12	0.08
Eastern Europe	North Central Asia	0.33	0.50
Eastern Europe	Northern Europe	0.67	0.63
Eastern Europe	South Asia	0.13	0.05
Eastern Europe	Southern Europe	0.90	0.97
Eastern Europe	Western Europe	0.44	0.38
North Central Asia	Northern Europe	0.11	0.09
North Central Asia	South Asia	0.16	0.55
North Central Asia	Southern Europe	0.12	0.13
North Central Asia	Western Europe	0.21	0.27
Northern Europe	South Asia	0.13	0.09
Northern Europe	Southern Europe	0.16	0.15
Northern Europe	Western Europe	0.28	0.19
South Asia	Southern Europe	0.19	0.13
South Asia	Western Europe	0.63	0.33
Southern Europe	Western Europe	0.84	1.00
East China	Central Asia	0.27	0.58
Eastern Asia	Central Asia	0.11	0.10
Eastern Europe	Central Asia	0.60	0.93
North Central Asia	Central Asia	0.26	0.35
Northern Europe	Central Asia	0.14	0.11
South Asia	Central Asia	0.86	0.92
Southern Europe	Central Asia	0.10	0.07

Western Europe	Central Asia	0.76	0.34
Eastern Asia	East China	0.37	0.56
Eastern Europe	East China	0.28	0.11
North Central Asia	East China	0.22	0.33
Northern Europe	East China	0.36	0.20
South Asia	East China	0.82	0.34
Southern Europe	East China	0.33	0.94
Western Europe	East China	0.37	0.37
Eastern Europe	Eastern Asia	0.07	0.04
North Central Asia	Eastern Asia	0.42	0.20
Northern Europe	Eastern Asia	0.10	0.10
South Asia	Eastern Asia	0.91	0.95
Southern Europe	Eastern Asia	0.09	0.09
Western Europe	Eastern Asia	0.06	0.04
North Central Asia	Eastern Europe	0.17	0.16
Northern Europe	Eastern Europe	0.31	0.31
South Asia	Eastern Europe	0.25	0.16
Southern Europe	Eastern Europe	0.43	0.48
Western Europe	Eastern Europe	0.94	1.00
Northern Europe	North Central Asia	0.14	0.12
South Asia	North Central Asia	0.82	0.36
Southern Europe	North Central Asia	0.10	0.08
Western Europe	North Central Asia	0.25	0.35
South Asia	Northern Europe	0.07	0.06
Southern Europe	Northern Europe	0.19	0.16
Western Europe	Northern Europe	0.95	1.00
Southern Europe	South Asia	0.22	0.12
Western Europe	South Asia	0.18	0.04
Western Europe	Southern Europe	0.76	0.45

Supplementary table 8-72: Output of the BSSVS analysis for the foot and mouth disease virus serotype SAT1 showing the best supported rates of transition between the sampled hosts for the subsampling approach

Origin	**Destination**	**Bayes factor**	**Posterior probability**
Cattle	Buffalo	459	1.00
Antelope	Cattle	96	0.99
Buffalo	Cattle	71	0.98
Cattle	Antelope	51	0.98
Buffalo	Antelope	14	0.92

Supplementary table 8-73: Output of the BSSVS analysis for the foot and mouth disease virus serotype SAT1 showing the best supported rates of transition between the sampled hosts for the “mugration” approach

Origin	Destination	Bayes factor	Posterior probability
Buffalo	Cattle	5970	1
Cattle	Antelope	5970	0.99
Cattle	Buffalo	5970	1
Antelope	Cattle	313	0.99
Buffalo	Antelope	116	0.98

Supplementary table 8-74: Comparison of the estimated rate values for the foot and mouth disease virus dataset between the subsampling and “mugration” analysis for the host analysis

Origin	Destination	Value rate subsampling	Value rate “mugration”
Antelope	Buffalo	0.96	0.59
Antelope	Cattle	2.41	1.33
Buffalo	Cattle	2.00	1.84
Buffalo	Antelope	0.91	0.69
Cattle	Antelope	0.93	0.60
Cattle	Buffalo	1.61	1.16

Supplementary table 8-75: Comparison of the estimated indicator values for the foot and mouth disease virus dataset between the subsampling and “mugration” analysis for the host analysis

Origin	Destination	Value indicator subsampling	Value indicator “mugration”
Antelope	Buffalo	0.76	0.69
Antelope	Cattle	0.99	1.00
Buffalo	Cattle	0.98	1.00
Buffalo	Antelope	0.92	0.99
Cattle	Antelope	0.98	1.00
Cattle	Buffalo	1.00	1.00