Mathematical models in molecular epidemiology

Combining genetic and epidemiological data to unravel infectious disease dynamics

Rolf Joseph Ferdinand Ypma

ISBN 978-90-393-6035-4

Mathematical models in molecular epidemiology

Combining genetic and epidemiological data to unravel infectious disease dynamics

Wiskundige modellen in moleculaire epidemiologie

Het samenvoegen van genetische en epidemiologische gegevens om infectieziektedynamiek uit te pluizen

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op dinsdag 19 november 2013 des middags te 2.30 uur

door

Rolf Joseph Ferdinand Ypma

geboren op 22 maart 1986 te Barendrecht

Promotor:
Prof. dr. M.J.M. Bonten

Co-promotoren:
Dr. J. Wallinga
Dr. W.M. van Ballegooijen

Preface

Infectious diseases have plagued us since time immemorial, altering social structures and language itself. Long held to be an unassailable divine judgement, since the 19th century infectious diseases have increasingly been controlled, due to improvements in sanitation, vaccines and medicine. Yet despite this progress, total elimination is not feasible, as evidenced by the pandemics and emerging infections we've seen over the last few decades. How do we best combat these threats, where should we invest resources, what public health interventions are optimally (cost-)effective? Answers to such questions can only be found if we understand the dynamics of infectious diseases: how they spread. This thesis aims to provide methods to increase that understanding.

The thesis that lies before you is the result of work performed at the National Institute of Public Health and the Environment. I originally applied for the PhD position because I had always been interested in biology and wanted to apply my technical skills to solving actual and complex problems, while learning as much as possible in the process. Over the last few years, I feel these goals have been met. Studying infectious diseases is fascinating; I have learned much, both about the biology of viruses and bacteria and about science, collaborations and communication. I am grateful for the level of independence granted to me by both the Institute and my supervisors. Not only do I feel much of my work is 'my own', I have also had the opportunity to deploy a range of activities beyond publishing papers. These include extensive trips abroad, the organization of symposia and my board work for Proneri, the Institute's PhD network.

I have never intended to restrict myself to one disease or one method. Rather, my focus has been on developing a range of methods that are able to answer different questions using different amounts and quality of data. The two things these methods have in common is that they combine genetic and epidemiological data, and that they are intended to be applied. I have endeavoured to illustrate this latter point by giving applications to tuberculosis, methicillin-resistant *Staphylococcus aureus* and avian influenza. I value clear communication, and have written the introduction with both scientists and a general audience in mind. It is my hope that few readers have either insufficient or sufficient background to not learn anything new from the diverse range of subjects covered. I will briefly give my views on future research and developments in the general discussion.

Contents

Chapter 1 **Introduction** **1**
General background 4
Existing methodology 13
Thesis overview 21

Chapter 2 **A sign of superspreading in tuberculosis: highly skewed distribution of genotypic cluster sizes.** **23**

Chapter 3 **Finding evidence for local transmission of contagious disease in molecular epidemiological datasets.** **35**

Chapter 4 **Monitoring the spread of MRSA in the Netherlands: A reference laboratory perspective.** **49**

Chapter 5 **Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data.** **63**

Chapter 6 **Genetic data provide evidence for wind-mediated transmission of highly pathogenic avian influenza.** **79**

Chapter 7 **Relating phylogenetic trees to transmission trees of infectious disease outbreaks.** **89**

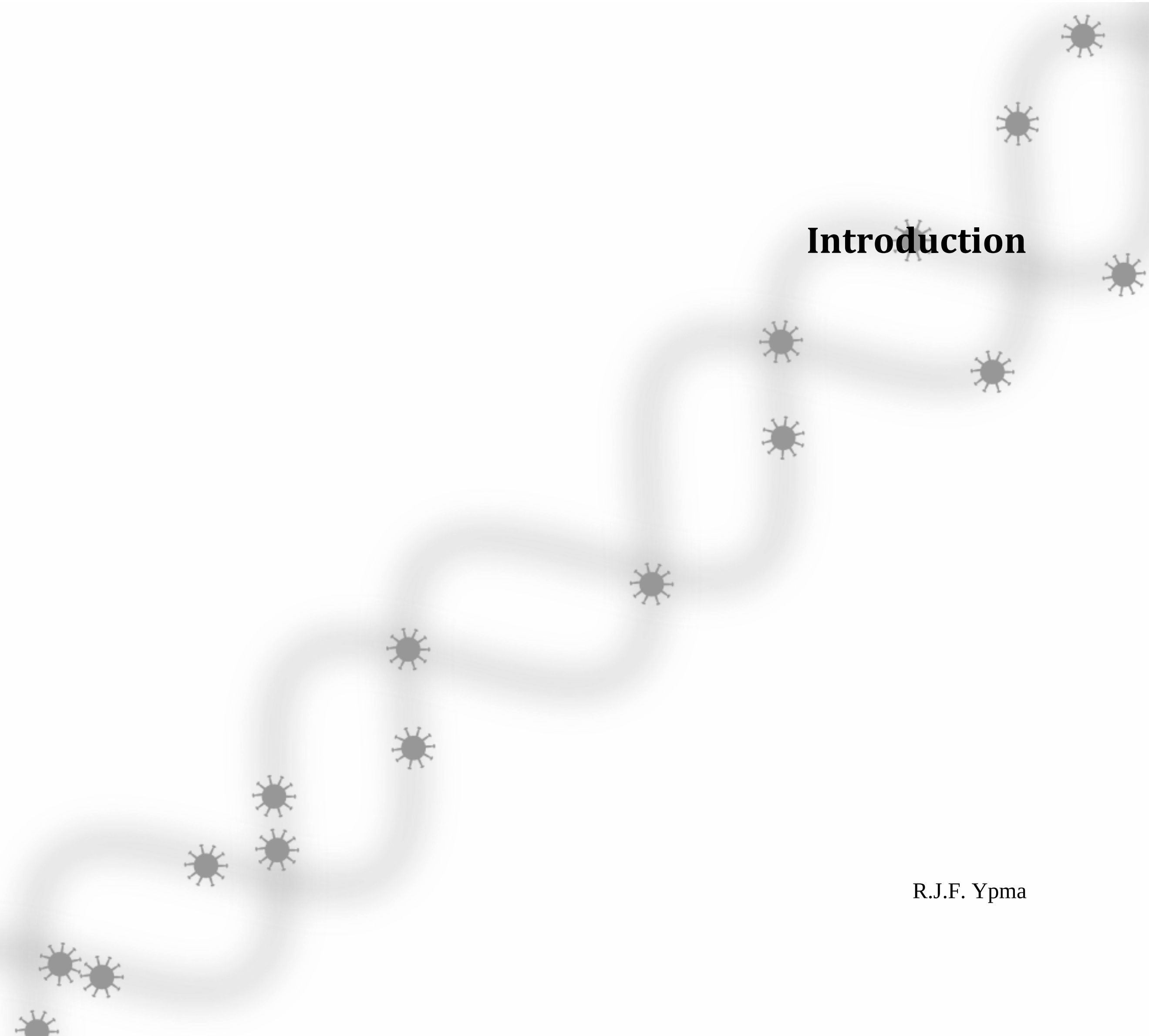
Chapter 8 **Discussion** **105**

Chapter 9 **To Conclude** **111**
Summary 112
Samenvatting 114
Dankwoord 116
Curriculum Vitae 118
References 119

Appendices **131**

Chapter 1

Introduction



R.J.F. Ypma

Living in the developed world, it is easy to feel that infectious diseases are a curiosity of history, rather than a serious threat today. As such, why would we study them? It is certainly true that mortality due to infectious diseases has dropped spectacularly over the last century.[1] However, one in five deaths worldwide are still caused by infectious diseases.[2] This might not be very visible in this part of the world, as the burden of infectious diseases is not evenly distributed; the highest burden of diseases such as HIV/AIDS,[3] tuberculosis[4] and malaria[5] is mainly found in Africa and Asia. Closer to home, a cause for worry is the increase in drug resistance in hospital-associated bacteria, which hampers our ability to treat patients.[6,7] Another reason to stay vigilant are emerging epidemics; the last decade alone has seen large outbreaks of 'new' diseases such as SARS,[8,9] the H1N1 pandemic[10,11] and the zoonotic avian influenza strains H5N1[12] and H7N9.[13] Besides morbidity and mortality caused, infectious diseases have a serious economic impact on society,[14,15] for example due to health care costs and productivity loss.

Although the need for control of infectious diseases seems clear, resources are always limited. We therefore want to know how effective interventions are, how effective they should be to stop spread, and how we can optimize them. When a vaccine is available, how expensive and effective is it? Should we vaccinate the whole population or certain groups, for example those most infectious or those most susceptible? In tuberculosis control, should we trace all contacts of all cases or focus on those we deem more infectious? For farm animal diseases, should we preventively cull animals, and if so how many and how fast do we have to be? To find answers to these questions, economic, ethical and practical considerations have to be taken into account. If we want this process to be at least partly objective, we need a quantitative understanding of the spread of infectious diseases.

Traditionally, quantifying parameters of disease dynamics has been done using epidemiological data such as time of symptom onset and recovery, and age, sex and location of cases. These can be used to estimate key parameters such as R_0, the basic reproduction number.[16-18] Monitoring such parameters allows for the evaluation of interventions.[19] As infectious disease dynamics will always be partly unobserved, e.g. due to asymptomatic or unreported cases, accurate and precise quantification of all parameters of interest can be difficult.[20]

A complementary source of information is provided by molecular biologists, who are increasingly apt at generating genetic sequences of pathogens, which form an alternative data source to quantify disease dynamics.[21] The separate analyses however might not be in agreement,[22,23] and only focusing on one of the two data types ignores the information present in the other. We would intuitively expect the best estimates to be given by an analysis that combines the two types of data. However, this combination of epidemiological and genetic data in one statistical analysis is still relatively rare; two reasons can be named. First, the quantity and quality of genetic data has been increasing rapidly in the last few years due to technical innovations and inference methods have not kept up. Second, the two types of data are generated in historically separated fields, complicating a combined approach.

How can we combine genetic and epidemiological data to infer infectious disease dynamics? There is unlikely to be one definitive answer; rather this is dependent on the specific question we are interested in and the type and amount of data available. For example, it is relevant whether we have observations on all cases or a subset, what percentage of cases was sampled, whether we established genotypes or full sequences, and how detailed the epidemiological data are. In general, of course, more detailed data allow for tackling of more detailed questions.

This thesis provides methods applicable to different datasets found in practice, and illustrates their use through application to data. The datasets span a wide range; in surveillance settings, there is often little more information than genotypes of a subset of cases and their sampling dates. From these datasets, high-level characteristics of infectious disease dynamics such as heterogeneity in infectiousness can be estimated (chapter 2) and local outbreaks can be detected by assessing correlations found (chapter 3 and 4). In detailed retrospective studies, it is possible that most cases are observed and both epidemiological data and sequence data are available. It then becomes feasible to reconstruct the full transmission tree of an outbreak (chapters 5-7).

It is my personal belief that a scientist should be able to translate or explain his or her work to a broad audience; one of the main goals of science is finding simple explanations for complex phenomena. Even more so for mathematicians, whose abstract of formal work may seem daunting at first glance, while their main ideas can often be explained in simple terms. Even more so for scientists working in an interdisciplinary field; as those working in different disciplines often have a different jargon, a different way of thinking and certainly a different set of basic knowledge. The research presented in this thesis can certainly be categorized as interdisciplinary; it uses ideas and concepts from mathematics, computer science, molecular biology, infectious disease epidemiology and population genetics. Finally, communication is, even more so, vital for scientists working on problems concerning our society as a whole, such as those faced in public health. It is useless to find all the answers if we cannot convince policy makers or the public of their worth.

The rest of the introduction consists of two parts; the first is a broad introduction for a broad audience into the biology and epidemiology of infectious diseases and mathematical models and their application. To be useful these sections are highly simplified and as such could be considered 'wrong'. I find this a fitting state of affairs for a thesis presenting models.[24] The second part is a brief scientific review of relevant literature and existing methodology.

General Background

Infectious diseases

The mathematical models that form the core of this thesis are tools to infer characteristics of infectious disease dynamics from information both on cases and pathogens. We will briefly discuss the biology of infectious diseases, and the fields that have traditionally studied these two distinct types of data: epidemiology and molecular biology.

Biology

Biology is the study of living things, and as such has always been fascinating for humans around the globe. Relatively few biologists are theorists; mathematics is not the essential tool it is in physics or engineering.[25] The reason is that biology isn't exactly rocket science; it's infinitely more complex. What follows is a simplification.

The basic building blocks of life are cells, small living units less than a tenth of a millimetre in diameter. Although simple life forms such as bacteria consist of one cell, we humans consist of trillions[26] (most of those however, again being bacteria[27]). Although the shape, size and content of cells varies enormously, they all use the same alphabet: DNA.

The `language of life' is written in a large molecule called deoxyribonucleic acid, or DNA. DNA usually consists of a double helix of two strands,[28] where a strand consists of a long chain of smaller molecules called nucleotides. These usually come in four varieties: adenine (A), cytosine (C), guanine (G) and thymine (T). Short sequences of these nucleotides, called genes, are read by the cell's machinery to form specific proteins, the molecules that do all the real work within the cell. Even the simplest mutation (alteration of the nucleotide sequence, or genetic code) can completely destroy the function of the protein that is coded for. Mutations, however, happen all the time when DNA is copied (for instance when the cell duplicates), as the wrong nucleotides can be accidentally inserted in the newly formed sequence. Some organisms, such as humans, employ ways to identify and repair such faulty sequences, leading to a lower 'mutation rate' (i.e. the per nucleotide probability of a mutation per replication or per unit of time). There are however two sides to the coin; although most mutations are deleterious, without them evolution could not have taken place.

Infectious diseases are caused by micro-organisms, such as parasites, fungi, bacteria and viruses, known as 'pathogens' (deriving from the Greek for 'the causing of suffering'). To be more precise, viruses are not usually considered an organism, and there are some infectious diseases caused by misfolded proteins ('prions'). Although microbes were discovered in the late seventeenth century by Anthonie van Leeuwenhoek,[29] a 'curious gentleman' from Delft whose letters had to be translated to English, their role in disease generation was not understood until the late 19th century. Most of the important human infectious diseases are caused by viruses and bacteria. The two exhibit marked differences in where they can be found and how they reproduce that are very relevant to molecular epidemiological inferences.

Viruses are often not considered alive, as without a host they are completely inert. They are not a cell, being much smaller, and usually consist of little more than a small container with some DNA (or RNA, a similar molecule with the interesting property that it usually exhibits higher mutation rates). When coming across a host cell, viruses have evolved a diverse array of techniques to enter the cell and use its machinery to produce a large amount of new viruses. This is usually bad news for the host, as resources are consumed and cells are destroyed. Furthermore, the viruses can induce symptoms, such as cough or diarrhoea, that are beneficial to virus spread, but not to the host. Most organisms therefore employ an arsenal of countermeasures, called the immune system, to find and destroy viruses and infected cells. Unfortunately, as the mutation rate of viruses is very high and their generation time is very low, new mutants can arise that are sufficiently different that the immune system cannot identify them.

Bacteria are single-celled organisms that come in a wide variety of sizes and shapes. Many species are spherical, and are called cocci, or rod-shaped, and called bacilli. They are ubiquitous on our planet, with total biomass roughly comparable to that of all other living organisms.[30] Only a few species cause disease in humans, and even many of those are harmless in usual circumstances and live on the skin or in the nose of healthy people. Most bacteria have a single piece of DNA called a chromosome, which they inherit from their single parent. Although their mutation rates are not as high as those seen in viruses, they can pick up DNA from the environment or other bacteria and incorporate it into their own. Furthermore, some bacteria have additional small DNA molecules called plasmids that can be shared with other bacteria. Such ‘horizontal gene transfer’ (the bacterial equivalent for sex) is relevant for public health, as it is an important way in which bacteria acquire genes that confer resistance to antibiotics.

Sequencing technology

The last few decades have seen an explosion in our ability to read DNA from biological samples. The oldest genetic techniques were able to 'type' organisms, for example by looking at variation within a specific region of the genome (the full DNA), or by counting the number of repetitions of a certain short sequence of DNA.[31] The first genome to be fully read was that of a virus in 1977,[32] the first human genome was fully sequenced in 2001.[33,34] The latest technologies have become so fast and inexpensive that soon medical doctors might sequence your genome for diagnostic purposes.[35,36] We might expect the abundance of genetic data being generated will tell us everything we want to know about any living being. Unfortunately, reading the code isn't the same as understanding it; although a given sequence of nucleotides always codes for a single protein, the actual structure or function of that protein is rarely clear.

The speed with which our ability to read sequences has increased has not been matched by our ability to derive meaningful answers from the generated data. One reason is that the sheer amount of data is staggering; a single run of a sequencing machine yields terabytes of data. Handling these large amounts of genetic data have become an important part of the field of ‘bio-informatics’, where the main focus is on trying to map differences in genomes to differences in organism function (such as the relation between particular mutations and hereditary diseases). Another reason could be that collaborations with different fields have to be established; genetic data might be most valuable when combined with different sources of information, as is the case in molecular epidemiology.

Epidemiology

The study of infectious diseases at the population level, or epidemiology (from epidemic: 'on the people'), is relatively young. John Snow is often credited as the father of (modern) epidemiology, due to his correct identification in 1854 of a contaminated water pump as the source of a cholera outbreak in London.[37] This credit might show the bias of the field towards physicians, as the Dutch-born mathematician Daniel Bernoulli had used statistical techniques and life expectancy tables to calculate the benefits of variolation[38] (deliberate infection with smallpox) almost a century before. Today much of infectious disease epidemiology is devoted to keeping a watchful eye on the prevalence of known diseases ('surveillance'), finding risk factors for disease and the implementation and evaluation of interventions, such as vaccine programs.[39]

Although the majority of mathematical tools used in infectious disease epidemiology consist of well-known statistical techniques such as regression analysis, more complex questions can be tackled using more specific mathematical modelling. Pioneers in this field were Ross and McKermack, who formulated the first Susceptible-Infected-Recovered (SIR) model,[40] which remains one of the most used mathematical models in infectious disease dynamics. Their 'SIR' model consisted of three coupled differential equations, describing the change in the number of susceptible (S), infectious (I) and recovered (R) (and assumed immune) individuals over time:

$$S' = -\beta SI$$
$$I' = \beta SI - \gamma I$$
$$R' = \gamma I$$

Key to the equations is the non-linear term βSI, which reflects the fact that for an infection to happen, you need both a susceptible and an infectious individual (reminiscent of the more widely known Lotka-Volterra equation describing predator-prey interactions[41,42]). A key quantity of this system is the so-called epidemiological threshold, or reproductive number at time 0:

$$R_0 = \frac{\beta}{\gamma} S(0)$$

which is the average number of infections caused by one infectious individual in a fully susceptible population; large outbreaks of an infectious disease can only occur if R_0>1. Using such equations, epidemiological data such as case counts over time can be coupled to relevant parameters such as R_0. This means we can inform our predictions about the future course of a disease or the impact of interventions using measurable quantities.

Molecular epidemiology

We can measure the similarity between genetic sequences, even if we do not understand the functional implications of their differences. For this reason, genetic data on pathogens can inform us on the epidemiological relationship between the infected individuals from whom the pathogens were sampled, a field of study called 'molecular epidemiology'. For example, genotyping has long been used to assess whether a number of observed tuberculosis cases could belong to the same transmission chain.[43]

How much information genetic data adds depends largely on how fast the pathogen mutates, and how densely a population has been sampled. Only when a mutant pathogen succeeds in begetting a large number of offspring, either through selection or chance, can we detect it. When a mutation is so successful that after some time it is present in the genome of all pathogens in a population, the mutation has been ‘fixed’ and is called a substitution. More information is present in genetic data when the substitution rate is high. For example, genetic differences between individuals can be used to reconstruct their ancestries or family tree (called phylogenetic tree in evolutionary studies); when the substitution rate is higher this tree can be reconstructed more precisely. Some viruses have a sufficiently high substitution rate that if we sample infected individuals in an outbreak, the genetic differences between the viruses inform us of the transmission dynamics of the outbreak. When the substitution rate is lower, we find many identical sequences, which yield less information. Likewise, when the sampling fraction decreases, our picture of the outbreak becomes less clear. The quality and quantity of the available data largely determine which questions can be answered, and what methods of analysis are appropriate.[44] For all but the most basic of these analyses, mathematical models and computer implementations are essential.

Mathematics

Mathematical models

A model is a simplified description of a system under study. A mathematical model is such a model formulated in mathematical language, which has the benefit of being very precise. As mathematics itself is rigid (it's either a tautology or it's wrong), it might come as a surprise that mathematical modelling can be considered closer to art than to science. Many opinions exist on what constitutes a 'good' model. Although formal model evaluation methods exist, confidence in a model is more often than not based on feeling. Perhaps the most important criterion for any model is its complexity; it should not be so *simple* that we cannot learn anything new from it or that it cannot capture some critical characteristic of the system studied, but should not be so *complex* that we cannot learn anything new from it.

Mathematical models can be used to investigate the behaviour of a system, or to quantify certain aspects of it. For example, using a model of infectious disease spread and vaccine effectiveness, it is possible to show that only part of a population has to be vaccinated to prevent spread of the disease.[18] Alternatively, a similar model could be used to estimate this vaccine efficiency from data on outbreaks.[45] Such quantifications crucially depend on the availability of sufficient and accurate data, and thus generally call for interdisciplinary approaches (mathematicians don't measure). The methods described in this thesis are data-driven, meant for quantifying infectious disease dynamics. It is the unique position of the Institute, where epidemiologists, microbiologists and mathematical modellers reside in adjacent buildings, that has made such analyses possible.

Likelihood

Central to many statistical methods used today, in particular the methods described in this thesis, are so-called likelihood equations. The general concept is not hard to understand, and we can explain it using an ornithological example.

When I walk through the meadow situated next to the Institute in which I work, I usually observe a number of birds. Let's call this number n. Of course, the number that I observe depends on the probability, for any given bird, that I spot it. Let's call this number p (it is a number between 0 and 1). As there are 100 birds present in the meadow, if the actual value would be $p=0.5$ (i.e. 50%), I would see $100\times0.5=50$ birds on average. The actual number of birds I see can vary quite a bit from day to day; anything between 40 and 60 seems likely, although even higher or lower numbers are still possible on rare occasions. We can write down a formula, the so-called probability distribution $P(n|p)$, that gives the probability that I observe any particular number of birds n, given my observation skills p. For example, with a p of 0.5, the probabilities of spotting 50, 60 or 90 birds are 0.08, 0.01 and 0.0000000000000001 (also see figure 1).

The probability distribution $P(n|p)$ is an example of a simple mathematical model. To build it, I assumed the value of n depends on p, which we call a parameter. I further assumed it does not depend on any other factor (like, say, the weather), which is a simplification. From these assumptions, I could derive the precise shape of the probability distribution. Even if I had made different assumptions, the derived probability distribution would always return zero for any n smaller than zero (I can't spot a negative number of birds), and the sum of all the probabilities are always one (I must spot some number).

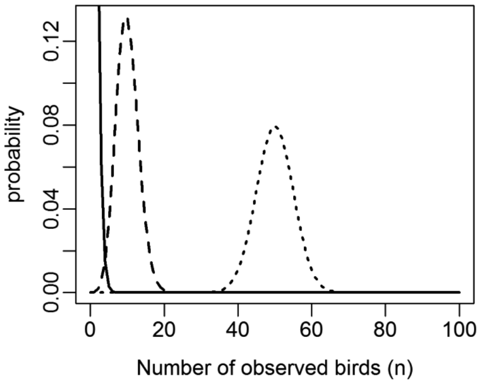


Figure 1. Graphical representation of $P(n|p)$, the probability of observing n birds given my observational skills p, the fraction of birds I normally spot. The distribution is given for (solid) p=0.01, (dashed) p=0.1 and (dotted) p=0.5. The expected number of birds spotted is equal to 100×p, so 1, 10 and 50 for the three p values.

When analysing data, we are not interested in the probability of the data given the parameters (we already know the data!), we are interested in what the data tell us about the actual values of the parameters. Rather than 'what is the probability of seeing n birds given my observation skills p?', the more sensible question in practice is 'what are my observation skills, given that I've seen n birds today?'. We can answer this question using the same formula $P(n|p)$ we had before, using a different interpretation. Whereas before we fixed p and put in an n to obtain its probability, we now fix n and put in a p to obtain its likelihood (this is sometimes denoted by $P(n|p)=L(p|n)$). The value obtained is no longer a probability, as probabilities necessarily sum to one. The likelihood values do not sum to one; in fact, adding the likelihoods obtained when putting in all possible values of p can yield any number. Any single likelihood obtained by putting in one p is therefore useless. However, the likelihood values for different p can be compared; these comparisons are often used to find the most likely p. For example, on my walk today I saw 10 birds. We can see from figure 1 that the most likely p value would be 0.1, as the corresponding distribution gives the highest likelihood value for n=10. That the most likely p value is 0.1 should not be a surprise, as the expected number of observed birds is exactly 100×0.1=10 when p=0.1.

Bayesian statistics

In the real world, there is often more information available than just the particular observations we are calculating with. For example, I am a mathematician, and it is common knowledge that all mathematicians generally look at their shoes whilst walking. This means we can safely assume my bird-spotting skills are rather poor, even before I’ve set foot in the meadow. Incorporating this type of knowledge that we have beforehand (or 'a priori') in our mathematical analysis is called Bayesian statistics (named after Thomas Bayes, an eighteenth century mathematician and Presbyterian minister who didn't formulate what is now known as Bayes' law, a simple theorem relating the probability and likelihood equations). Although no-one disputes Bayes' law, opinions differ as to if and how such a priori knowledge should be incorporated. Nevertheless, Bayesian statistics is nowadays used throughout science.

From a Bayesian point of view, the combination of data and a model are just something to change a prior distribution (what we believed a parameter to be) into a posterior distribution (our best guess as to the parameter's value given all information). For example, I am fairly confident I spot 50% of all birds; from this belief I can construct a prior distribution for p, specifying precisely how likely each value of p is a priori, where $p=0.5$ is the most likely value (see figure 2A). During my walk, I observed 10 birds (the data). This value is much lower than the value of $100\times0.5=50$ predicted by my prior; maybe I'm less observant than I thought. Mathematically, this means the value of 10 observed birds changes my prior assumption into a posterior distribution that deems lower p values more likely than the prior did (see figure 2A). If I repeat this measurement over several days, the actual value of p will become much clearer, and shape of the prior becomes less important (figure 2B). If I constructed a prior distribution that was more peaked, the posterior estimate would differ less. For example, if I were very confident in my personal abilities (hypothetically speaking, of course), only a few observations would not shake my faith, and I would cling to the value of $p=0.5$ (figure 2C).

What is important in practice is that the more informative the data are, the larger we can expect the difference between the prior and posterior to be. With abundant, precise data, it hardly matters what we put in as prior. With a small amount of data, the best guess we can make is effectively the prior; what we already thought before we started measurements. It is thus possible to get very strong results even with terrible data; just put in very strong prior beliefs! When we are more interested in what the data tell us (like in this thesis), we can use uninformative priors; distributions that effectively say we have no idea what the actual value is. This is a concept of considerable debate; it turns out to be surprisingly hard to assume nothing.[46]

Computation

In many applications in science we can write down how likely a certain set of parameters is, given our observed data. This has become a useful exercise in the last few decades, as computers allow us to find the most likely parameter values and associated uncertainty by computing the likelihood of all possible sets of parameters. Unfortunately, even with fast computers this is often infeasible due to the large number of possible parameter values that would have to be evaluated.

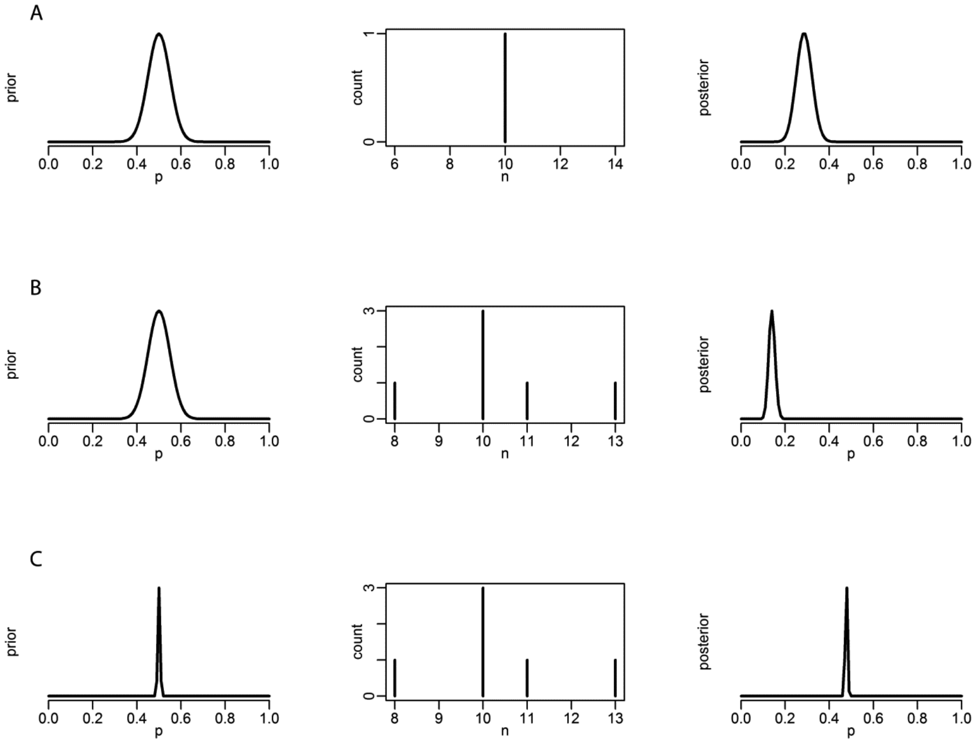


Figure 2. In Bayesian statistics, prior beliefs (left column) are transformed into posterior distributions (right column) by data (middle column). (A) I believe I can spot 50% of all birds, this translates into a prior distribution that is highest for *p*=0.5 (left). Today, I spotted 10 out of 100 birds (middle), which is indicative of an observational proficiency of *p*=0.1. The mathematics tell us the truth lies somewhere in between; the prior distribution and data together yield a posterior with a most likely value of *p*=0.3. (B) When I have the same prior distribution as before (left), but measurements over a whole week (middle), the posterior distribution differs more from the prior. This most likely value now is *p*=0.14. (C) Being more certain about a prior belief can be incorporated by using a more peaked prior (left). This peaked prior distribution shows only the value of *p*=0.5 is likely; other values are not. The same data as in (B) then have less impact: the posterior distribution (right) is very similar to the prior, having shifted only the tiniest amount to the left. This reflects the fact that more data are needed to convince someone very certain of his beliefs.

In many applications, varying a parameter by a tiny amount will also change the likelihood of that parameter by a tiny amount. As a simple example, consider three parameters a, b and c, all between 0 and 1. Let's use their sum as a simple example of a likelihood, so the likelihood of the set (a, b, c) is $L(a, b, c) = a + b + c$. Then all parameters close to (1,1,1) are pretty likely, whereas all parameter close to (0,0,0) are very unlikely. In general, the space of all possible parameters will have regions consisting of unlikely and (usually smaller) regions consisting of likely sets of parameters. A straightforward way to speed up computation is therefore to focus on those regions where parameters are likely. This is not always as simple as it sounds, as the space of all parameters can be high dimensional and we do not know beforehand where the likely parameters are.

One of the most popular techniques that makes use of this continuous nature of likelihoods to reduce computational burden is called Markov chain Monte Carlo (MCMC). An MCMC is a chain whose elements are sets of parameters. The chain is started at a random location in the parameter space, i.e. the first element is a randomly chosen set of parameters. In our example from above, let's start at $(0.37, 0.14, 0.42)$. The candidate for the next element is a different set of parameters obtained by changing all of the parameters in the first set a tiny amount. In the example, the candidate could be $(0.31, 0.11, 0.47)$. If this new set is more likely according to the likelihood equation and prior, the newly chosen parameters will be the next element in the chain. If the set is slightly worse than our current set, we might still pick it, but with a low probability. On rejection, we just stick with our current set. In our example, we would switch to the new proposal with probability $\frac{L(0.31,0.11,0.47)}{L(0.37,0.14,0.42)} = \frac{0.31+0.11+0.47}{0.37+0.14+0.42} = \frac{0.89}{0.93} \approx 0.96$. With probability $1 - 0.96 = 0.04$ we would stick with our old set. This simple procedure achieves two important things. First, by preferring likely sets of parameters it quickly finds those regions of the parameter space consisting of likely parameter sets. Second, due to the particular construction of the chain, the fraction of times a certain parameter value is found in the chain is proportional to how likely this value is. Thus the chain gives us the information we want: the posterior distribution of our parameters.

The advantage of MCMC's is that they guarantee to give the exact solution, no matter how complicated the problem. The disadvantage is that this exact solution is only reached after an infinite calculation time. For finite computation time, in general no guarantees can be given. The reason the technique is used is that it can often give good approximations to the actual solution in a reasonable short time. How reasonable this is, depends on the complexity of the problem and on a number of technical details in implementing the algorithm. MCMC's are used several times throughout this thesis.

Existing methodology

The methods presented in this thesis were explicitly designed to combine genetic and epidemiological data. It should come as no surprise that such methods build on, or are inspired by, existing methodology, much of which is focused on either of the two data types. Chapter 2 uses the theory of branching processes to analyse cluster sizes defined by genetic data. Chapters 3 and 4 are inspired by clustering techniques built to analyse spatiotemporal data, and extend them to identify outbreaks in molecular epidemiological datasets. Chapters 5-7 provide and illustrate a method to reconstruct transmission trees, which is feasible when all or most individuals in an outbreak have been sampled and sequenced. As chapter 7 explains, this method can be seen as a fusion of transmission tree reconstruction methods based on epidemiological data with phylodynamical methods based on genetic data. Here we will discuss each of these existing approaches.

Branching processes

A branching process is a stochastic process that models a reproducing population. Each individual in generation n begets a number of offspring in generation $n+1$, drawn from a discrete probability distribution $P(X)$. A branching process can be used to model transmission of infectious disease, when it is reasonable to assume that each infectious individual encounters the same number of susceptible individuals. Typical examples are a low endemic disease or the initial phase of an epidemic. The average number of new infections caused by an infectious individual in a fully susceptible population (R_0), is then equal to the expectation $E(X)$ of $P(X)$. The difference in infectiousness among infected individuals is related to the variance of the distribution.

Branching processes can be used to make inferences about infectious disease dynamics by considering the final sizes of epidemics (i.e. the total number of individuals infected in an outbreak). When $E(X)<1$ the process will die out: the size of the epidemic will be finite with probability 1. When $E(X)>1$ the size of the epidemic can become infinite with positive probability (which shows that a branching process should not be used to describe large epidemics). For a broad class of probability distributions, we can calculate the probability that the process will reach a certain size.[47] This feature can be very valuable when small outbreaks are observed, as the value of R_0 can be estimated from the outbreak sizes.[48] A branching process analysis was used, for example, to show that measles in Britain was close to causing large-scale epidemics following a drop in vaccination coverage.[49]

An observed distribution of final sizes can reveal the heterogeneity in infectiousness of infected individuals.[50] This concept is important, as a high heterogeneity ('superspreading') causes faster and more explosive outbreaks and calls for different intervention measures.[51] Genetic information here can play a valuable role; we can infer the offspring distribution required to quantify superspreading by analysing the size distribution of clusters of cases infected by pathogens with identical genotypes (chapter 2).

Outbreak detection

In surveillance of infectious disease, an important task is the detection of local outbreaks. A general statistical procedure to do this would be to compare the number of detected cases to the number expected when no outbreaks are present.[52-54] Such approaches can be extended by also taking the location of cases into account.[55-57] One well-known example of such a technique is the so-called space-time scan statistic,[58] in which the number of cases in a scan window is compared to the expected number when there are no local outbreaks. A scan window consists of all points within a maximum distance in space and time away from a certain source; the windows take the shape of cylinders in space-time. Comparison of scan windows with a range of sizes and locations will reveal the windows most likely to contain outbreaks.

Pathogens derived from the same outbreak will show identical or very similar genetic profiles. Thus, genetic data on pathogens can improve the detection of local outbreaks.[59,60] However, it is hard to say how large a genetic difference should be before we conclude that two cases are not infected in the same outbreak. In general, to answer this question requires detailed information on pathogen substitution rates and genetic diversity worldwide, information that is often not available. We can work around this problem by not focussing on absolute differences but on the correlation between distances in the different types of data we gather: temporal, spatial and genetic. Distances for cases infected in the same local outbreak will be small for all three data types. This principle can be exploited to identify local transmission using minimum a priori knowledge on the pathogen or population at risk (chapters 3-4).

Transmission tree reconstruction

One intuitively clear and powerful method to infer infectious disease dynamics at a very detailed scale is the reconstruction of transmission trees. A transmission tree is defined as the set of transmissions that gave rise to an outbreak; reconstruction of this tree means estimating who infected whom. When the tree is known, quantities such as the effective reproduction number,[19] heterogeneity in offspring distribution[51] and covariates of infectiousness can be easily estimated.

When there are no missing data, the mathematical estimation of a transmission tree T is relatively straightforward. Conditional independence of transmission events is assumed, meaning that the likelihood of a tree is given by the product of the likelihood of its links:

$$p(D|T) = \prod p(D_a, D_b | e_{ab} = 1)$$

where $D = \{D_i\}$ are the data for all cases i, $e_{ab} = 1$ denotes an edge from a to b (i.e. a infected b) and $p(D_a, D_b | e_{ab} = 1)$ denotes the likelihood of a having infected b. The exact formulation differs per application. For example, researchers can focus on the distribution of time between infections,[19,61] on geographical distance between cases[62] or on the intensity of contacts made between different groups of people.[63]

When information is available on susceptible individuals who escaped infection, the absolute infectiousness of infectious individuals can be estimated. This means we can compute, for any pair of infectious and susceptible individuals, the probability that an infection will occur between them within a defined time period. When information on susceptible individuals is unavailable, as is often the case, the infectiousness f is only known up to a constant. This means we can only look at the relative infectiousness of infectious individuals:

$$p(e_{ab} = 1|D) = \frac{f(a \to b|D_a, D_b)}{\sum_i f(i \to b|D_i, D_b)}$$

which can still inform us on covariates of infectiousness. For example, are people more infectious if they are older or have more severe symptoms?

When there are missing data, such as unobserved times of infection, the transmission events are no longer independent. For example, information on who infected an individual gives information on when the latter was infected, which gives information on whom the individual could have infected in turn. This is not a large problem for one-dimensional data like time, as infection times can be estimated simultaneously from dates of symptom onset or from sampling dates. Infection times are then considered known during computation steps and we can still use the equations above.

Genetic data on pathogens can inform us on the transmission tree if the pathogen measurably mutates during the outbreak. This concept was pioneered by Cottam et al.,[22] who restricted their epidemiological analysis of transmission trees for an outbreak of foot-and-mouth disease to those that corresponded to the phylogenetic tree (i.e. the tree of ancestral relationships among individuals) derived from viral sequence data. This sequential approach loses information present in the phylogenetic trees, as all trees with a likelihood above a certain cut-off are deemed equally likely, and all other trees are discarded. The problem can be overcome by considering the two data types simultaneously[64] (chapter 5). In this approach, the evolutionary process is simplified by assuming a mutation is fixed only at time of transmission. This means the transmission events can still be considered independent, which allows for faster calculation. Note that when sequences are missing, they cannot be efficiently estimated simultaneously like infection times, due to their high dimensionality (i.e. the number of base pairs sequenced).

The approach of considering all data simultaneously[64] was further built on by Morelli et al.,[65] who proposed a model that incorporated a mutation rate to model the dependence between time and genetic distance. In this model, the evolutionary time between to samples is taken to be the sum of the times between sampling and transmission, which is equivalent to the assumption that coalescent events of viral lineages coincide with transmission events between hosts. The actual relationship between host epidemiological and pathogen genetic data is slightly more involved, and depends on within-host pathogen dynamics[66] (chapter 7).

Phylogenetic tree reconstruction

When individuals reproduce, the genome of their offspring is rarely identical to their own. When the differences arise through mutations, it is possible to reconstruct the ancestry, or phylogenetic tree, of individuals through their genetic sequences. Reconstructing phylogenetic trees from sequences is a field of study in itself. The aim is to find the ancestry that best matches measured genetic distances between individuals.[67] This can be done by finding the likelihood of a phylogenetic tree P and mutational model μ given sequence data S. To find an expression for the likelihood, we make the assumption that mutations accumulate at a certain rate.[68] This rate may vary over parts of the tree,[69,70] and some mutations can be more likely to occur than others.[71] In particular, synonymous mutations, which do not alter the protein coded for, are assigned another rate than non-synonymous mutations. Furthermore, not all nucleotides are equally likely to mutate to all other nucleotides. Given their molecular size, the purines A and G are more likely to mutate from one to the other than into one of the pyrimidines C and T, and vice versa. These assumptions together form the substitution model μ, yielding an expression for the likelihood

$$L(P,\mu|S) = \prod_{l=1}^{M} \sum\nolimits_{\{A,C,G,T\}^N} \prod_{e\in edges} P(s_1 \rightarrow s_2|\Delta t,\mu,l)$$

where M is the number of nucleotides sequenced, N is the number of internal nodes of the tree, and $P(s_1 \rightarrow s_2|\Delta t,\mu,l)$ is the probability of mutating from one state (s_1) to another (s_2) over an edge e of the tree. This probability depends on the substitution model μ, the length of the edge Δt and the position of the nucleotide l.

The likelihood equation above can be used to sample the space of all trees and parameters using an MCMC. Calculation of the likelihood as written above is computationally costly; the number of elements over which the equation sums is 4^N. The so-called Felsenstein's pruning algorithm can substantially reduce this number using the tree structure of the phylogeny, by calculating the likelihood from the leaves up.[72] Even with this algorithm, however, searching the entire tree space is a costly procedure[73,74] (in fact, finding the maximum likelihood tree is NP-hard[75]); most researchers in practice use freely available software packages to do the calculations for them.[76]

Phylodynamics

Coalescent theory

The relationship between genetics and population biology is studied in the field of population genetics.[77] Founded at the start of the twentieth century, this mathematically oriented field attempts to answer questions on evolution, with much focus on change of frequencies of mutants in populations. An important breakthrough for our purposes was the development of coalescent theory,[78-80] which makes the connection to phylogenetic trees. The crucial idea of coalescent theory is that individuals will be more related (i.e. their most recent common ancestor will be closer in time) when the population to which they belong is smaller. For example, two people randomly selected from Bilthoven will be more likely to be related than two people randomly selected from Asia.

The relationship between population size and phylogenetic trees can be made more specific under some explicit assumptions. Let's consider a clonal population of size I with discrete generations, where each individual has one parent in the previous generation, chosen randomly. This is a simple form of the so called Wright-Fisher model,[81,82] named after two of the founders of the field. For any two individuals randomly chosen from the current generation, the probability that they have the same parent in the last generation is $1/I$. This means the time to the most recent common ancestor (MRCA) is geometrically distributed with parameter $1/I$; the expected number of generations we have to go back to find the MRCA is exactly the population size. For n individuals (where n is small compared to I) , the expected number of generations we have to wait for the first MRCA of any pair is $\binom{n}{2}\big/_{I}$, as $\binom{n}{2}$ pairs can be formed. This means that the branch lengths of the phylogenetic tree reconstructed from genetic data of a number of individuals can be directly related to the, possibly time-varying, size of the population to which they belong.[83-86]

Only relatively recently have coalescent approaches been applied to study infectious disease dynamics,[87-89] by analysing phylogenetic trees constructed from pathogen sequences (hence the field's name: phylodynamics[90]). These techniques have been used to address questions on the incidence and prevalence of a disease,[91-94] geographical spread of disease,[95,96] mode and timing of unobserved transmissions,[97,98] sexual risk behaviour[99] and population structure.[100-103] These quantities are important to public health but can be hard to measure by standard epidemiological approaches.

General framework

In phylodynamical methods, the objective is generally to derive the likelihood of a particular model of disease transmission, given genetic sequence data. The likelihood is usually separated into two parts, letting the phylogenetic tree P form an intermediate between the sequence data and the transmission parameters of interest.[20,65,93] First, the likelihood of a phylogenetic tree P given sequence data S and a mutational model μ is derived as above. Second, the likelihood of the transmission model parameters α given the phylogeny is derived. More precisely,[104,105]

$$L(\alpha,\mu,P|S) = L(P,\mu|S)L(\alpha|P,\mu,S) = L(P,\mu|S)L(\alpha|P)$$

where the second equality follows from conditional independence; when the phylogenetic tree is known, the sequence data and mutation model give no further information on the transmission model. The phylogeny is a nuisance parameter; it is usually integrated out to obtain the likelihood of the model parameters given the data.[106]

The part of phylodynamical methods most researchers focus on is the relationship between a mathematical model of infectious disease transmission and the phylogenetic tree: $L(\alpha|P)$. Early approaches were based on coalescent theory, where α consisted of the effective population size to be estimated from the sequence data. Subsequent models allowed for increasingly flexible population size models, going from simple constant and exponentially increasing sizes[88,97,107] to complex approaches using semi-parametric approaches and Gaussian Markov random fields.[108-111]

Infectious disease specific population models

Until recently, all population models $L(\alpha|P)$ used in phylodynamics were general models of effective population size, not specific for infectious diseases. In analysis of infectious disease dynamics, such models have two shortcomings. They do not take into account the depletion of susceptible individuals and they assume that the fraction of sampled infected hosts is very low, which need not be the case for infectious diseases. Two transmission processes have been proposed to take the place of the Wright-Fisher model; the SIR model[91] and the birth-death model.[112]

The central notion of the SIR model is that the number of new transmissions depends on both the number of infected and the number of susceptible individuals. Thus the rate $\lambda(t)$ at which coalescent events occur depends not only on the number of infected individuals $I(t)$, but also on the number of susceptible individuals $S(t)$. For the classic coalescent theory we have $\lambda(t) = \binom{n(t)}{2} \Big/ I(t)$, where by *n(t)* we denote the number of lineages at time *t* with extant progeny.

The more general form is[91,113,114]

$$\lambda(t) = \binom{n(t)}{2} \Big/ \binom{I(t)}{2} f(t)$$

where $f(t)$ is the rate of new infections per unit time. The term $\binom{n(t)}{2} \Big/ \binom{I(t)}{2}$ is the probability that two randomly selected infected individuals have sampled offspring, i.e. are represented by one of the branches of the phylogenetic tree. For SIR models, $f(t) = \beta I(t) S(t)$, yielding

$$\lambda(t) = \binom{n(t)}{2} \Big/ \binom{I(t)}{2} f(t) \approx 2\binom{n(t)}{2} \Big/ I(t)^2 \, \beta I(t) S(t) = 2\binom{n(t)}{2} \beta S(t) \Big/ I(t)$$

which is proportional to the rate used in the classical coalescent when S is constant. The main advantage of using this adjusted expression for the coalescent rate is that it forms a bridge between standard epidemiological models and phylodynamics. This makes it possible to estimate epidemiological parameters directly from sequence data.

In a birth-death model, we follow a set of individuals over time. These give birth to new individuals at rate g and are removed (i.e. die or recover) at rate b. At death, they can be observed (i.e. sampled) with probability s. The advantage of this very simple model is that the full joint probability distribution of a birth-death process and sampled phylogenetic tree can be specified.[112,115-117] Simultaneous estimation of g, b and s turns out to be impossible, but they can be estimated when additional information such as the time of most recent common ancestor is available.[118-120] The main advantage of this method is that the sampling fraction is no longer an issue, and parameters can be estimated even for densely sampled epidemics, such as HIV epidemics, in which sequences are available for more than half of the cases.[119,120]

Character states and population structure

A useful way to incorporate additional data into phylogenetic analyses is by considering character states.[121,122] A character is defined as any property of the pathogen sampled (or the individual it was sampled from) except the sequence itself. Using the character states at the tips of the phylogenetic tree, the states at the internal nodes can be estimated by constructing an appropriate model of the evolution of the character. A clear example of this approach is seen in the field of phylogeography, where the character studied is geographical location. If locations are known for all sampled cases, it becomes possible to estimate where a new epidemic arose, or what the important geographical paths of transmission are.[96,123-125] This technique has been used to track the spread of the 2009 H1N1 pandemic around the globe,[126] and to show that the spread of rabies virus in African dog populations follows human movement patterns.[98]

Another character state now being studied by phylodynamical approaches is the stage of the disease. This character is relevant for viruses such as HIV and the hepatitis B virus, where cases first go through an acute phase before developing a chronic infection. Relative to the chronic phase, the acute phase is characterised by higher viral load, distinct symptoms and possibly higher infectiousness.[127] As the fraction of new infections caused by cases in acute versus chronic phases affects the effectiveness of intervention strategies, it would be valuable to estimate the relative infectiousness of the two stages. Using a mathematical model that describes the process of acquiring a chronic infection after acute infection, it is theoretically possible to use phylogenies to estimate the infectiousness of cases in the different phases of a disease.[102]

Finally, a lot of attention has gone to inferring population structure from phylogenies. Structure here means either that a population can be divided into groups with different properties relevant to spread of a disease (e.g. certain age groups are more infectious) or that large heterogeneities in infectiousness exist among individuals. Analyses of population structure are performed by using character states, as mentioned above,[102,128] or by assessing the shape of the phylogenetic tree, as measured by various tree statistics.[101]

Thesis overview

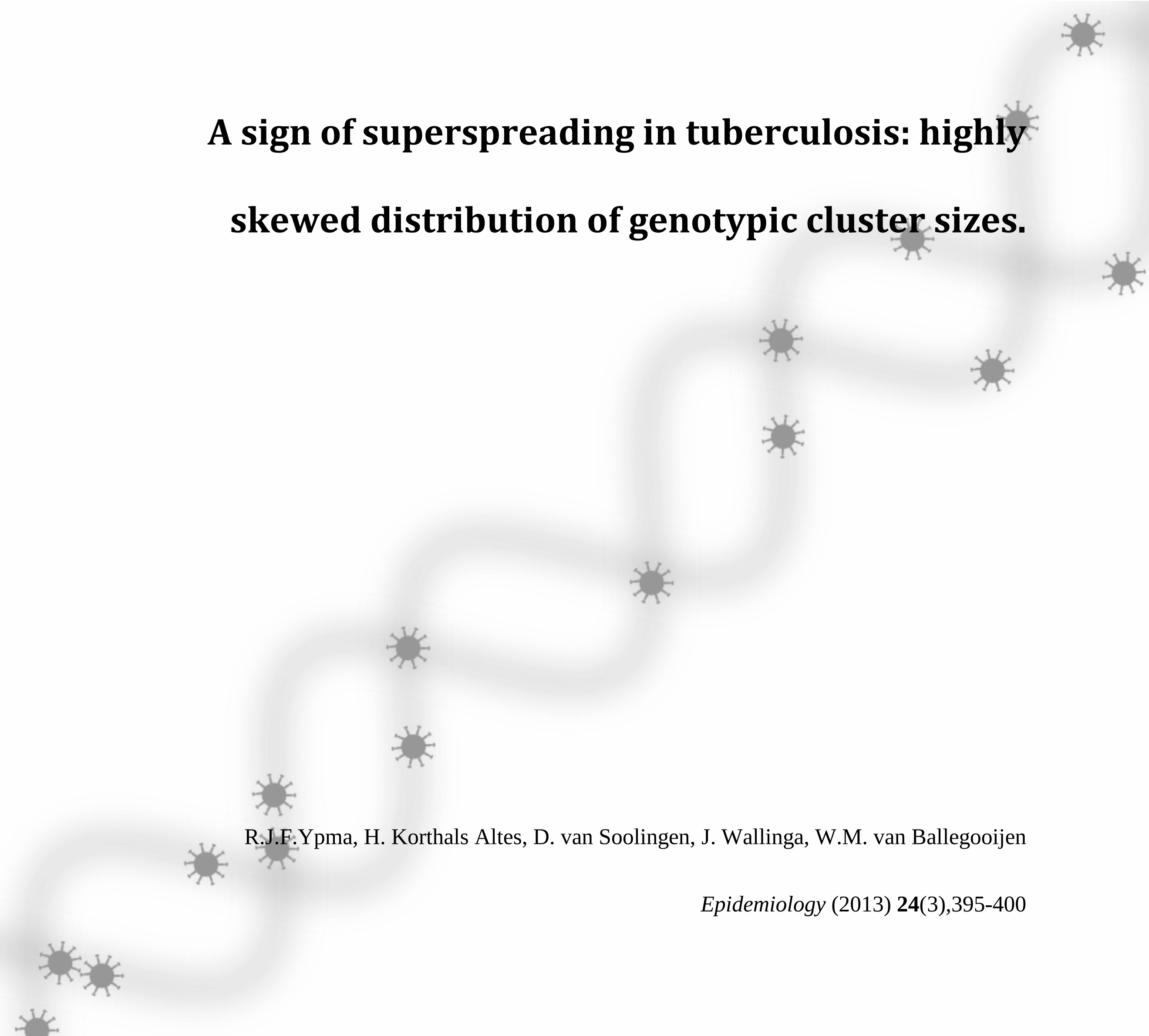
In chapter 2 we show how the notion of 'superspreading' (i.e. some infected individuals causing a disproportionately large number of secondary cases) can be quantified for tuberculosis using data on date of diagnosis, country of origin and genotype of the sampled bacterium. The genetic data here is of coarse resolution; although the population can be divided in groups of cases sharing the same genotype, no further distinction can be made within or among these groups. A mathematical model derived from branching processes allows for inference of heterogeneity in infectiousness based on the size distribution of these groups, indicating superspreading behaviour for tuberculosis.

In chapter 3 we develop and apply a method to identify cases of infectious diseases that are epidemiologically related, which allows for detection of local outbreaks and risk factors for local infection. The method builds on existing clustering methods that identify groups of related individuals, and adds to it by specifically including genetic data. Although the data used here are still genotypes rather than sequences, there is now a clear measure of similarity between the genotypes, which allows for inter-group comparisons. In chapter 4 we apply this method to a dataset on MRSA in Dutch hospitals, containing MLVA types, dates of sampling and geographical locations of the patients. The analysis shows marked differences between MLVA complexes, and suggests a high proportion of infections are imported into the hospital.

Chapter 5 describes a method to estimate the transmission tree of an outbreak of an infectious disease, based on epidemiological and genetic sequence data. The method is applied to an outbreak of avian influenza, and shows differences in infectiousness for different types of farms. Chapter 6 extends this analysis by including data on farms not infected, and shows a statistically significant correlation between wind direction and the direction of spread of the disease. In chapter 7, we treat genetic data in a more sophisticated way, allowing for within-host genetic diversity. This approach shows the relationship between transmission tree reconstruction and phylodynamics. The approach is tested on simulated data, and illustrated on data on a small outbreak of foot-and-mouth disease.

Chapter 2

A sign of superspreading in tuberculosis: highly skewed distribution of genotypic cluster sizes.



R.J.F.Ypma, H. Korthals Altes, D. van Soolingen, J. Wallinga, W.M. van Ballegooijen

Epidemiology (2013) **24**(3),395-400

Molecular typing is a valuable tool for gaining insight into spread of Mycobacterium tuberculosis. Typing allows for clustering of cases whose isolates share an identical genotype, revealing epidemiological relatedness. Observed distributions of genotypic cluster sizes of tuberculosis are highly skewed. A possible explanation for this skewness is the concept of 'superspreading': a high heterogeneity in secondary cases caused per infectious individual. Superspreading has been previously found for diseases such as SARS and smallpox, where the entire transmission tree is known. So far, no method exists to relate superspreading to the distribution of genotypic cluster sizes. We quantified heterogeneity in secondary infections per infectious individual by describing this number as a negative binomial distribution. The dispersion parameter k is a measure of superspreading; standard (homogeneous) models use values of $k \geq 1$, while small values of k imply superspreading. We estimated this negative binomial dispersion parameter for tuberculosis in the Netherlands, using the genotypic cluster size distribution for all 8330 cases of culture-confirmed, pulmonary tuberculosis diagnosed between 1993 and 2007 in the Netherlands. The dispersion parameter k was estimated at 0.10 (0.088, 0.12), well in the range of values consistent with superspreading. Simulation studies showed the method reliably estimates the dispersion parameter under a range of scenarios and parameter values. Heterogeneity in the number of secondary cases caused per infectious individual is a plausible explanation for the observed skewness in genotypic cluster size distribution of tuberculosis.

Introduction

Worldwide, about 9 million tuberculosis (TB) cases are reported yearly, leading to an estimated 1.4 million deaths per year.[129] Moreover, about a quarter of a million cases of multi-drug-resistant TB are notified annually, posing a new threat in TB control.

IS6110 restriction fragment length polymorphism (RFLP) typing distinguishes between different *M. tuberculosis* strains by visualizing repetition and genomic location of the IS6110 repetitive sequence.[130] Since the IS6110 transposition rate is low,[131] cases that are in the same transmission chain will usually have identical DNA fingerprints; therefore genotyping of *M. tuberculosis* is often used to trace presumed epidemiological links between TB cases. Furthermore, it is thought that the low genetic diversity among *M. tuberculosis* isolates in high-prevalence areas is related to high levels of recent transmission.[43]

There is thus a relationship between TB epidemiology and DNA fingerprints. We define a genotypic cluster as all isolates sharing an identical fingerprint. A large genotypic cluster is usually the result of local transmission, while rare genotypes are due to reactivation or import of cases or mutation. Luciani et al.[132] noted that the distribution of sizes of genotypic clusters for many datasets found in the literature is highly skewed, featuring some very large clusters and many of size one (i.e. unique DNA fingerprints). The authors suggested this skewness could be indicative of an increasing prevalence, with older clusters being exceptionally large, and small clusters the product of recent mutation.

An alternative explanation for the high skewness of the genotypic cluster size distribution is that a small proportion of all cases cause a large fraction of all infections, a phenomenon frequently called "superspreading".[51] There is anecdotal evidence that large TB outbreaks can be generated by a single case, diagnosed at a very late stage, causing many secondary cases.[133] Furthermore, Vynnycky et al.[134] estimated the variance to mean ratio of the number of individuals effectively contacted by a tuberculosis case in the Netherlands to be 20, indicating that some individuals could add disproportionately to the total number of infections caused. This heterogeneity in number of cases caused by an infectious individual has been observed for a range of infectious diseases.[51]

To examine whether heterogeneity in number of cases caused per infectious individual could explain the observed distribution of cluster sizes, we study data on all pulmonary cases from the Netherlands isolated between 1993 and 2007. For this population the explanation of a growing number of tuberculosis cases does not hold, as there is no sustained transmission of the disease in the Netherlands. Instead it is regularly introduced from abroad, mostly by immigrants from high-endemic countries, or caused by endogenous reactivation in (elderly) persons, often infected in their youth when TB prevalence was still high in the country.[135] In the Netherlands all cases of tuberculosis are tested since 1993 and typed using IS6110 RFLP typing when culture-confirmed.

Standard tools for population genetic inference, such as the coalescent,[132,136] assume random sampling from one large transmission tree, with a low sampling frequency. This does not apply to the Dutch tuberculosis data, which is sampled with a high sampling frequency from many small outbreaks. We therefore model spread of tuberculosis in the Netherlands as a subcritical branching process[48] where each reactivation or introduction from abroad can trigger a minor outbreak. We model the number secondary cases caused by one infectious individual by a negative binomial distribution; the dispersion parameter of this distribution gives a measure of superspreading. Adapting the branching process to allow for mutation during our study period, we derive likelihood equations for the distribution of genotypic cluster sizes and use these to estimate the negative binomial dispersion parameter from the observed genotypic cluster sizes.

Materials and methods

Data

In the Netherlands, between 1993 and 2007, 21,155 patients were diagnosed with tuberculosis, of which 14,818 (70%) were confirmed by culture. Samples from all of these cases were subjected to IS6110 RFLP typing; strains with fewer than five IS6110 copies were subtyped with the PGRS probe[137]. The Netherlands Tuberculosis Register, holding epidemiological data such as immigrant status, was matched with the database holding genetic data, on the basis of sex, date of birth, year of diagnosis, and postal code. This yielded a total of 12,222 (82%) culture-confirmed patients with typing information, of which 8,330 were pulmonary cases. We exclude non-pulmonary cases since they are not infectious, focusing on infections which lead to secondary pulmonary cases.

Transmission model

We model the spread of tuberculosis as a branching process, where the number of secondary cases caused by an infectious individual is given by a probability distribution, called the offspring distribution. Following Lloyd-Smith et al.[51] we assume a negative binomial offspring distribution with mean R and dispersion parameter $k \in (0,\infty)$. This distribution is widely used to model overdispersed count data, such as the number of social contacts made by people over a fixed timeperiod.[138,139] The parameter R could be considered an ‘effective reproduction number’. The dispersion parameter k governs the degree of superspreading; a lower value of k means a higher heterogeneity (figure 1). The variance-to-mean ratio of the negative binomial distribution is given by $1+R/k$. As $k \to \infty$, the offspring distribution becomes a Poisson distribution, which would result from all cases being equally infectious, having the same fixed infectious period, and making contact at random. k=1 yields a geometric distribution, which is used in many infectious disease models such as simple SIR models.[140] While this seems a plausible assumption for common diseases such as influenza,[141] for some diseases such as SARS and smallpox k has been estimated to be as extreme as 0.1 or less.[51,142,143]

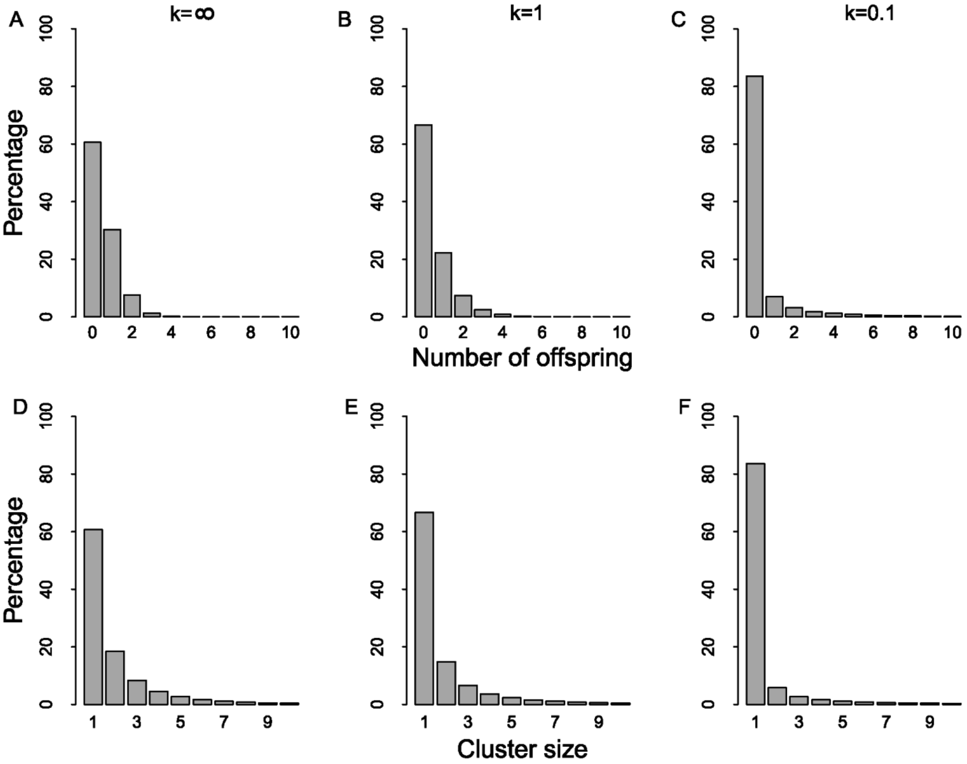


Figure 1. Probability distributions used in the inference procedure, for different values of the dispersion parameter *k*. (A-C) Distribution of the number of secondary infections caused per infectious individual, given by a negative binomial distribution with mean *R*=0.5 and (D-F) final size distributions for clusters. The columns show results for different values of the dispersion parameter *k*; (A,D) k=∞, as in a Poisson distribution (B,E) *k*=1, as in a geometric distribution and (C,F) k=0.1, a negative binomial distribution. The distributions become highly skewed for small *k*, e.g. while for the Poisson distribution (*k*=∞) 61 percent of all cases does not infect anyone else, for *k*=0.1 this number is 84 percent, even though the average number of infections remains the same.

We take a transmission cluster to be all infectious cases resulting from one index case; this could include infections caused by secondary, tertiary, etc. cases. When the mean number of infections caused by one infectious individual is smaller than one, such a transmission cluster will not go on indefinitely, but reach a certain final size. The probability distribution for the final size of a transmission cluster Y is given by[47,48] (appendix A.1):

$$P(Y = y) = \frac{1}{y}\binom{y + yk - 2}{yk - 1}\frac{R^{y-1}k^{ky}}{(k+R)^{(k+1)y-1}} \tag{1}$$

Deriving the transmission clusters

The observed genotypic clusters, i.e. all cases sharing the same DNA fingerprint, are not the same as the transmission clusters. This is because:

- multiple introductions into the country can share a fingerprint (without being epidemiologically linked);
- complete epidemiological and molecular data is not available for all cases in the study period, and no data is available for cases outside of the study period of 15 years, which could still belong to the same transmission cluster;
- due to mutation, epidemiologically linked patients could exhibit different DNA fingerprints.

To derive the transmission clusters from the genotypic clusters we use the rules described below.

Any case with a DNA fingerprint not seen in another patient in the preceding two years is taken to be an index case, since most infected persons that develop disease do so within two years.[144] This could underestimate the number of index cases: when two cases share a DNA fingerprint, there is a small chance they have introduced this fingerprint independently, even if they are found within two years. This is especially relevant for recently arrived immigrants, who are unlikely to have been infected in the Netherlands. We therefore assume immigrants that have been in the country for less than six months at date of diagnosis are index cases themselves, having brought the bacterium from abroad rather than having been infected in the Netherlands. This means that if a case found in the two years preceding the diagnosis of a recently arrived immigrant has the same DNA fingerprint as the immigrant, the genotypic cluster formed by all cases with this fingerprint actually consists of two transmission clusters, the immigrant being the index case of one of them (figure 2).

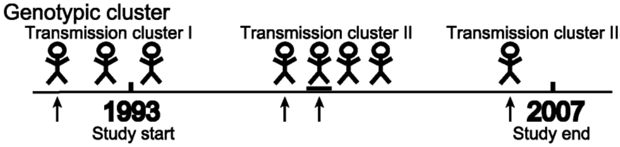


Figure 2. Cartoon illustrating the identification of transmission clusters from a genotypic cluster. Puppets represent patients infected with the same DNA fingerprint before, during or after the study period 1993-2007, the underlined puppet denotes a recently arrived immigrant. By defining a new transmission cluster when there are more than two years between consecutive cases, the method separates this genotypic cluster into three transmission clusters. Each transmission cluster has at least one index case, i.e. a case not infected by another case from the same transmission cluster. Index cases are denoted by arrows. Clusters near the edge of the study period may be partially observed, e.g. the left cluster in reality has size 3, rather than the observed 1. This is accounted for by assuming the sizes of the left and right clusters are at least 1, rather than exactly 1. Since the middle cluster has two index cases (the first case and a recently arrived immigrant), it actually consists of two overlapping transmission clusters.

Cases that belong to a transmission cluster but were diagnosed before or after the study period are censored[48] (figure 2). As linking of the two databases used lead to a match for 82% of cases, we assume that for each case in the study period full data is available with probability 0.82.

Transpositions of IS6110, resulting in a different DNA fingerprint, can occur during transmission in the Netherlands.[145] We denote the probability that an infected individual will show a different fingerprint than that of its infector when progressing to disease by p_m.[146] These mutations lead to an offspring distribution which is still negative binomially distributed with unchanged dispersion parameter, but with reduced mean $R(1\text{-}p_m)$ (appendix A.1.2) This means we cannot estimate R and p_m separately; we therefore introduce the fingerprint reproduction number R_m, defined as $R(1\text{-}p_m)$.

Inference of transmission parameters

The likelihood of parameters R_m and k when $a_{y,n}$ clusters of size y with n index cases have been fully observed, and $b_{y,n}$ censored clusters have been observed to have at least size y and n index cases is then

$$L(R_m, k \mid \vec{a}, \vec{b}) = \prod_y \prod_n \left(f(y \mid n) \right)^{a_{yn}} \prod_y \prod_n \left(1 - F(y-1 \mid n) \right)^{b_{yn}} \tag{3}$$

where $f(y|n)$ is the probability a cluster with n index cases has size y given the parameters, and $1\text{-}F(y\text{-}1|n)$ is the probability a cluster with n index cases has at least size y (see appendix A.1.3 for details). Using the likelihood function (equation 3), we can find maximum likelihood estimators for R_m and k and obtain confidence intervals by using profile likelihood.[147] To assess the impact of our assumptions, we perform sensitivity analyses for the number of years between time-separated clusters and for the percentage of cases observed.

Testing method performance on simulated data

To test the robustness of the inference procedure we use simulated data. We want to test whether the algorithm is able to differentiate between different levels of heterogeneity in the offspring distribution. We do this by generating datasets under a range of parameter values, and estimating the parameters R_m and k using the proposed approach.

We generate index cases, spread out uniformly over a time period starting 10 years before our study period and ending at the same time as our study period. Each case causes a random number of other cases, which is given by a negative binomial distribution with R=0.3, p_m=0.1 and k varying per simulation. We take the time between onset of disease of the source case and the infected case to be exponentially distributed with rate 0.8 per year.[148] We take a transposition rate of 0.1 per year.[131,149,150] In the Dutch dataset, 110 instances of an additional index case in a cluster were observed. These were recently arrived immigrants with a previously seen DNA fingerprint, resulting in genotypic clusters actually consisting of two transmission clusters. We therefore randomly take 110 simulated genotypic clusters and set their fingerprint to be identical to that of another 110 randomly selected genotypic clusters. From these simulated datasets we take all cases that fall in our study period spanning 15 years, and randomly remove 18% of these to mimic the fact that in reality we have information on 82% of cases. We stop generating index cases when we have observed 8000 cases. We then apply our inference method and compare estimated parameter values to the actual values we simulated with.

In the simulations, R_m is consistently overestimated and sensitive to the different values of additional variables (figure 3). However, the value of k is only slightly overestimated, and the estimation is not sensitive to the additional variables. There is a clear distinction between estimates of k from simulations where k=1 and simulations where k=0.1.

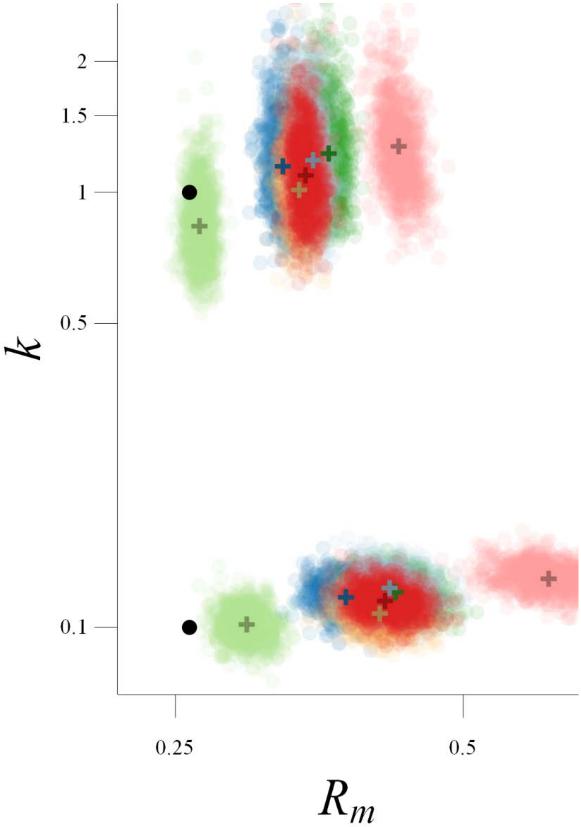


Figure 3. Performance of the method to distinguish between different levels of heterogeneity using simulated data. Point estimates are given for R_m and k for 1,000 simulations as described in the main text, using values of R_m=0.27 and k=0.1 and 1 (large dots). Colors denote different simulation scenarios, the median of the estimates for each of the estimates is given by a cross. Colors are as follows: the standard scenario (red), time between transmission clusters of one (light green) or three (pink) years, a transposition rate of 2 per year (dark blue), no (orange) or 400 (light blue) clusters with more than one index case, and an infection rate of 1.5 per year (dark green). There seems to be a small overestimation of k consistent over all simulations, and a large overestimation of R_m dependent on the simulation assumptions. Importantly, the method can clearly distinguish between simulations under k=0.1 and k=1.

Results

The 8,330 culture-confirmed pulmonary cases in the Netherlands formed 4,945 different genotypic clusters; figure 4 shows the distribution of their sizes. The observed genotypic cluster distribution is highly skewed. 81% of all clusters consist of only one case; these clusters contain 48% of all isolates. The largest cluster observed consisted of 119 cases.

We find estimates of the fingerprint reproduction number R_m of 0.48 (95% confidence interval: 0.44, 0.59) and the dispersion parameter k of 0.10 (95% confidence interval: 0.088, 0.12). This value for the dispersion parameter k is much lower than the values corresponding to the Poisson distribution (k=∞) and the geometric distribution (k=1) which are typically used in transmission models, and is suggestive of superspreading.

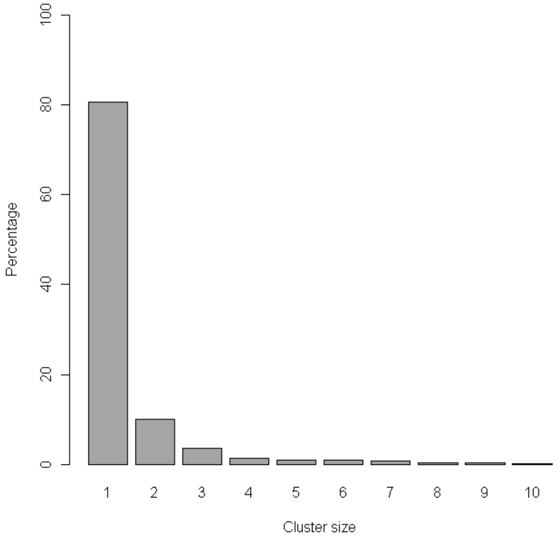


Figure 4. The percentage of genotypic clusters having size 1 to 10, for all cases of culture-confirmed, pulmonary tuberculosis in the Netherlands from 1993 to 2007. A cluster of size one is an isolate with a unique DNA fingerprint. There are 62 (1.3%) genotypic clusters with size greater than 10 (not shown), the largest of which has size 119.

To test for robustness, we showed that the estimation procedure is able to differentiate between values of k=1 and k=0.1 under a range of parameter values (figure 3), making it unlikely that the estimated low value could result from a transmission process without superspreading. In an additional sensitivity analysis, where we vary the minimum time between transmission clusters and the percentage of cases observed, we find the estimate of the fingerprint reproductive number to be sensitive to precise assumptions. However, the estimate of the dispersion parameter is insensitive (figure 5). Finally, we found this estimate to be insensitive to several of our assumptions regarding extra pulmonary cases, immigrants and recurrent mutations (appendix A.2).

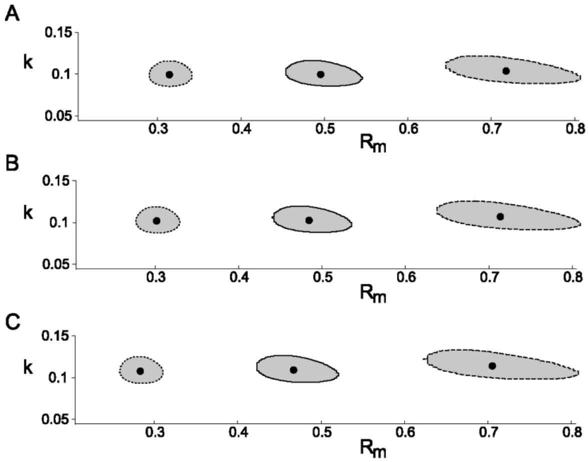


Figure 5. Estimates and confidence regions for the fingerprint reproduction number R_m and dispersion parameter *k* under different assumption. Estimates are given for a minimum time between clusters of 1 (dotted), 2 (solid) and 3 (dashed) years, and a percentage of cases with complete data of (A) 70%, (B) 82% and (C) 100%. The estimated value of R_m depends strongly on the minimum time between clusters, but *k* is consistently estimated at 0.1.

Discussion

The cluster size distribution given by RFLP typing of pulmonary cases in the Netherlands between 1993 and 2007 is highly skewed, and consistent with a value of the dispersion parameter *k* for the number of secondary cases caused per infectious individual of 0.10. This value indicates severe heterogeneity in number of secondary cases caused per infectious individual.

To the best of our knowledge, the dispersion parameter *k* has not been estimated before for tuberculosis. The low value of *k*=0.10 found is consistent with the range of values that have previously been inferred for other infectious diseases, such as smallpox, measles and SARS, and has been taken as an indication of superspreading.[51] The value found is also consistent with the range of values observed for the heterogeneity in the number of different social contacts that participants reported to have made during a day[139] or during a week.[138]

Alternative explanations for the high skewness in the distribution of tuberculosis genotypic cluster sizes exist. First, it has been argued that the skewness could be due to an increasing prevalence of the disease.[132] This explanation is implausible for our dataset, as the number of TB cases in the Netherlands is declining.[151] Second, the skewness could be due to bacteriological factors, such as differences in transmissibility or transposition rates between different strains. Although there is some evidence that such bacteriological differences exist,[152,153] the differences tend to be small or specific to certain types such as the Beijing genotype. It is therefore unlikely that bacteriological factors alone could result in the high skewness observed.

Simulations show our estimate of k to be robust under a wide range of assumptions. The actual value of k is slightly overestimated for most simulations, which makes the estimator conservative for our purpose. The value of the fingerprint reproduction number R_m, which is a lower bound for the effective reproduction number, is harder to estimate reliably; different assumptions lead to very different estimates. This is a consequence of the fact that the time between the onset of symptoms for a case and its infector can be much longer than the total study period.

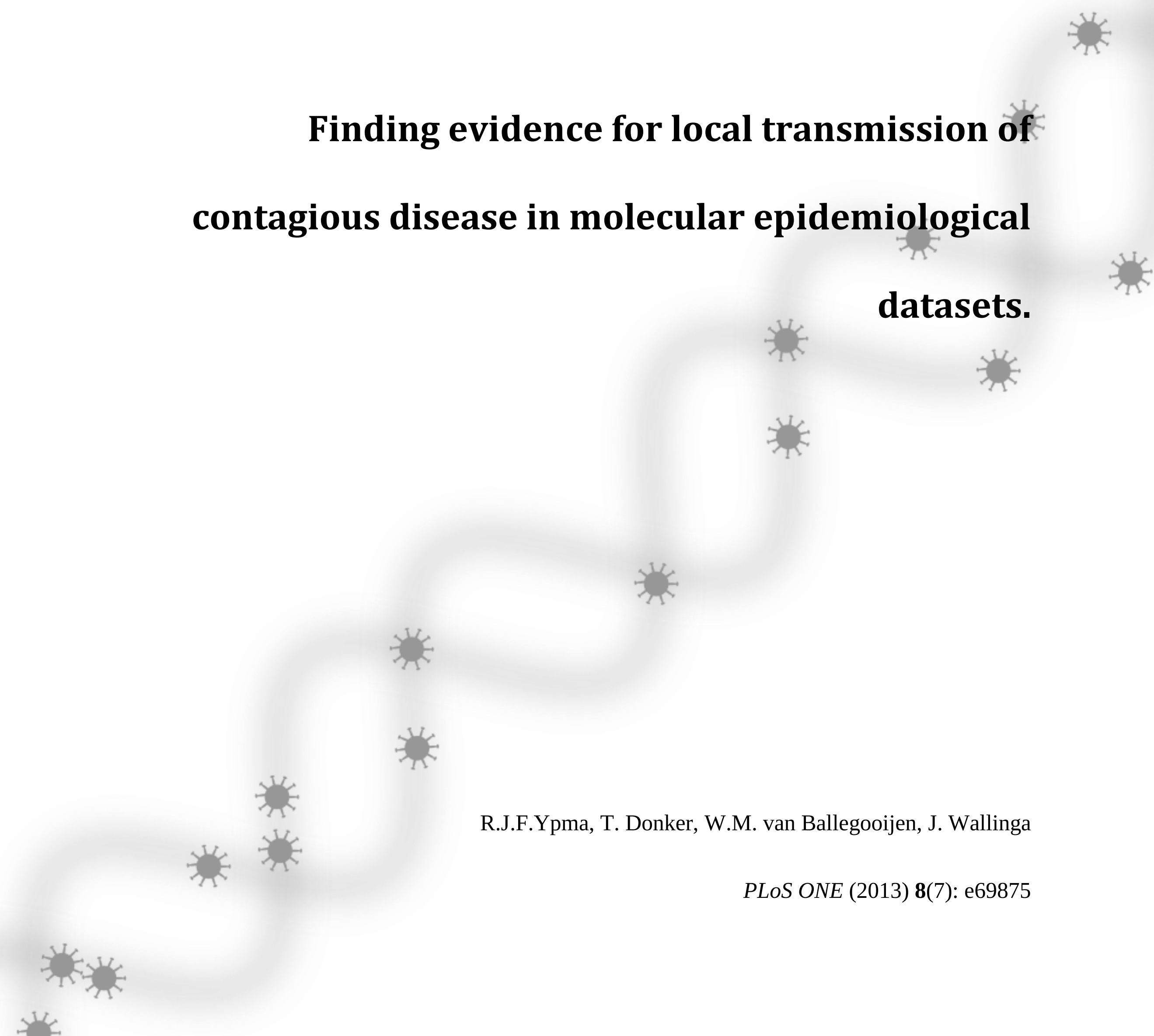
There are several factors which could possibly contribute to a high heterogeneity in secondary infections caused per infectious individual in tuberculosis, the separate contributions of which are hard to quantify.[154] For example, heterogeneity might arise because some hosts are more connected in a social network, because of shorter latent periods in different age groups, because of increased susceptibility in certain risk groups or because some cases are diagnosed at a very late stage. Estimating the effect of each of these factors is a challenge, since it is likely many of them are contributing simultaneously. One possible way to estimate these contributions is to incorporate all factors in one detailed model.[134] Another is to compare clustering patterns between high- and low-endemic countries, which share some factors (e.g. bacteriological) but differ in others (e.g. control measures and HIV prevalence).

A straightforward way of estimating the extent of superspreading for a given infectious disease is only available in those rare instances where the whole transmission tree for an outbreak is known. Recent studies on viral pathogens have focused on deriving related epidemiological measures from detailed phylogenetic trees.[101,155] We have shown this estimation can also be done using final size equations. Furthermore, these final sizes can be found from genotypic clusters, even when transmission clusters cannot be separated on purely epidemiological grounds.

Superspreading can be a major concern for disease control, especially for emerging diseases. When superspreading is present outbreaks tend to be explosive, and can be large even when $R<1$.[51] For tuberculosis, the main implication for public health is that the most cost-effective control strategy consists of highly targeted interventions. This provides a theoretical basis for the recent debate to adjust the guidelines for tuberculosis control in low incidence regions to focus resources on settings where transmission is more likely, such as highly infectious cases or highly susceptible persons surrounding a case.[156]

We conclude that the observed size distribution of clusters of tuberculosis cases with identical genotype in the Netherlands is consistent with a dispersion parameter that indicates superspreading. Although further research is needed to elucidate the causes of the observed heterogeneity, studies aiming to accurately describe the spread of tuberculosis will have to take this heterogeneity into account.

Chapter 3



Finding evidence for local transmission of contagious disease in molecular epidemiological datasets.

R.J.F.Ypma, T. Donker, W.M. van Ballegooijen, J. Wallinga

PLoS ONE (2013) **8**(7): e69875

Surveillance systems of contagious diseases record information on cases to monitor incidence of disease and to evaluate effectiveness of interventions. These systems focus on a well-defined population; a key question is whether observed cases are infected through local transmission within the population or whether cases are the result of importation of infection into the population. Local spread of infection calls for different intervention measures than importation of infection. Besides standardized information on time of symptom onset and location of cases, pathogen genotyping or sequencing offers essential information to address this question. Here we introduce a method that takes full advantage of both the genetic and epidemiological data to distinguish local transmission from importation of infection, by comparing inter-case distances in temporal, spatial and genetic data. Cases that are part of a local transmission chain will have shorter distances between their geographical locations, shorter durations between their times of symptom onset and shorter genetic distances between their pathogen sequences as compared to cases that are due to importation. In contrast to generic clustering algorithms, the proposed method explicitly accounts for the fact that during local transmission of a contagious disease the cases are caused by other cases. No pathogen-specific assumptions are needed due to the use of ordinal distances, which allow for direct comparison between the disparate data types. Using simulations, we test the performance of the method in identifying local transmission of disease in large datasets, and assess how sensitivity and specificity change with varying size of local transmission chains and varying overall disease incidence.

Introduction

An essential question in contagious disease surveillance settings is whether cases result from local transmission within a population or from importation of infection from outside the population. This distinction is of importance, as interrupting local transmission calls for different interventions than stopping importation of infection. Unfortunately, distinguishing cases related through local transmission from cases that result from importation of infection is difficult.

When occurrences of a contagious disease are monitored and stored in a standardized way, statistical algorithms can aid in identification of local spread of the disease. For example, drug-resistant pathogens found in hospitals either are due to nosocomial transmission or are brought into the hospital by the patient. The former can be identified using surveillance data by assessing the number of cases in a fixed time period,[59] this identification is essential in optimizing hospital control measures.

Genetic sequence data of pathogens provide an informative data source for distinguishing between local transmission and importation. Sampled pathogens are now routinely genotyped or sequenced in many settings, offering the potential to distinguish cases that were infected in a local transmission chain from those that were infected elsewhere by evaluating small genetic differences between sampled pathogens. However, existing algorithms to find clusters of related cases in large datasets focus only on temporal data[52-54,157-160] or on spatiotemporal data.[55,56,58] Although genetic data are already being used to distinguish between different strains of the same species,[59,161] the full potential offered by these data has so far not been utilized.[60]

In outbreaks of contagious diseases, cases are caused by other cases. This property results in clusters of cases due to local transmission of contagious diseases having a different mathematical structure in space and time than clusters of cases due to non-contagious diseases. Clusters due to contagious disease tend to have a more chain-like shape (figure 1). However, existing clustering algorithms that focus on spatiotemporal data do not account for this property, as they often have not been developed specifically for contagious diseases.

Here, we present a method that identifies locally infected cases from a dataset containing genetic, temporal and geographical data. For each pair of cases, we assess the distance between these cases with respect to their locations, their times of symptom onset, and the genetic sequences of the pathogens isolated from the cases. For a pair of cases not related through local transmission, we expect the distances in the separate data types to be independent. For cases that are part of a local transmission chain, we expect the distances between these cases to be small for each of the separate data types. We employ a form of hierarchical clustering that uses an ordinal distance between cases based on their genetic, temporal and geographical distances, and that reflects the fact that for contagious diseases, cases are caused by other cases. Clusters of cases resulting from local transmission are identified by testing whether they have smaller pairwise distances than would be expected under a null hypothesis of independence between the location, time of symptoms onset, and pathogen sequence of cases. As the purpose of the present paper is to introduce and explain the methodology, and to illustrate its use and limitations, we use simulated datasets. To test the ability of the method to detect local transmission of a contagious disease, by assessing the sensitivity and false positive rate of assigning cases to local transmission clusters.

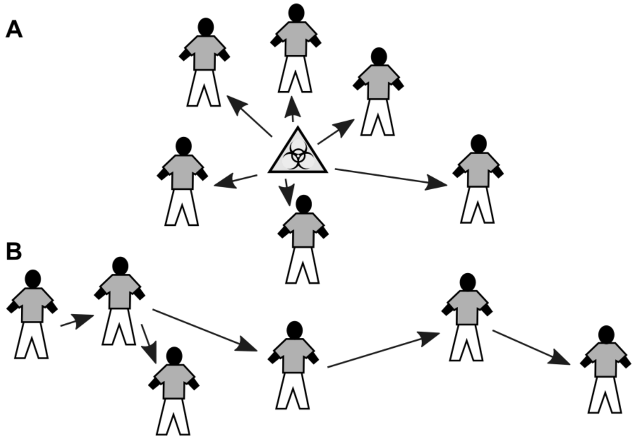


Figure 1. Different patterns of disease clusters. Clusters of disease cases caused by a point source (A) show a different pattern than clusters caused by human-to-human transmission of a contagious disease (B). (A) When there is a point source cases tend to be found in the region around it. Modern scan statistics exploiting this pattern have been developed to find evidence of point sources causing disease. (B) When contagious diseases are spread by human-to-human transmission, clusters tend to be more chain-like; the relevant distances are those between pairs of cases rather than between case and point source. Although it is still possible to find clusters in situation (B) with algorithms developed for (A), the problem can be handled more naturally by taking into account the different cluster pattern.

Methods

We consider a contagious disease surveillance dataset that consists of a large number of cases. We assume that for each case we know the date of symptom onset or sampling date, the geographical location, and the genetic type or sequence of the pathogen. Some of the cases might be infected within the time and region of the study, while others are infected elsewhere. Our objective will be to identify transmission clusters; sets of cases related through a local transmission chain.

It is infeasible to consider every possible subset of cases in the dataset as a possible transmission cluster, because the number of subsets grows exponentially with the number of cases. We adopt a hierarchical clustering approach; here the dataset is sequentially divided into subsets of increasing size, yielding a tree-like structure, or dendogram (figure 2). The subsets encountered in this way are the most plausible local transmission clusters.

To perform hierarchical clustering one needs a measure of dissimilarity between sets of cases. We construct such a measure using both a measure of dissimilarity between individual cases and a linkage criterion that gives the similarity of two subsets as a function of the pairwise dissimilarities of their elements.

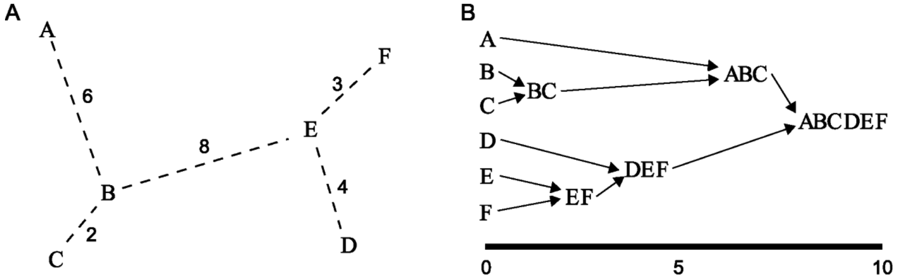


Figure 2. Graphical representation of hierarchical clustering. Hierarchical clustering sequentially clusters together elements of a set, based on inter-element distances. (A) Representation of a set of six elements. Shown is a minimal spanning tree: the tree that connects all elements minimizing total distance. (B) The clustering provided by hierarchical clustering when using single linkage clustering. Sequentially, the two current subsets with smallest distance are joined together, where the initial subsets are the six elements. This means the distances of clustering on the x-axis in (B) are the distances of the minimal spanning tree in (A). In total five distinct clusters are passed before all elements cluster together.

Linkage criterion

We use single linkage clustering,[162-164] the oldest and arguably simplest linkage criterion, which states that the distance between two sets of cases is the minimum of the pairwise inter-element distances. A commonly cited drawback of this criterion is that it tends to create chain-like clusters, with a high average distance within the cluster. Contagious disease epidemiology seems one of the few settings where this property is actually an advantage. Since contagious disease cases are caused by other cases the distance between two sets is well described by be the smallest pairwise distance and chain-like clusters are very plausible; for instance outbreaks spreading to another location, or viruses mutating in a certain direction over time.

Pairwise dissimilarity

It is not immediately obvious how a dissimilarity should be defined between two individual cases. We have to combine a temporal, a geographical and a genetic distance, comparing days with kilometers and mutations. Furthermore, the absolute values of these distances are not directly informative. First, because we assume no knowledge on pathogen characteristics we cannot interpret any absolute value. Second, because many cases in one geographical, temporal or genetic region might be the result of a high population density, seasonality or higher pathogen fitness, respectively, rather than of local transmission. The relevant notion of dissimilarity between two cases is therefore not an absolute distance, but the number of other cases found in between the two cases,[165] i.e. closer to both the two cases than the two cases are to each other. For example, two cases living a kilometer apart are more likely to be related when this is in a rural area than in a large city, two cases infected at the same day are more likely to be related when they are infected during an off-season than during an epidemic, and two identical strains are more likely to be related when this is a rare sequence than when this sequence is ubiquitous. This relevant notion of number of cases in between two cases is not pathogen specific and allows for combination of the three disparate data types. We define the dissimilarity d_i for a given data type i between two cases a and b as (figure 3):

$$d_i(a,b)=|\{p: D_i\,(a,p) \leq D_i\,(a,b) \wedge D_i\,(b,p) \leq D_i\,(a,b)\}| - 1 \qquad (1)$$

where |.| denotes the number of elements of a set, ∧ the logical AND operator, D_i the absolute distance for data type i (time, location or genetic) and the '-1' ensures $d_i(a,a)=0$.

Under our null hypothesis of all cases being unrelated, the dissimilarities in the different data types are independent. In contrast, dissimilarities between two cases infected in the same local transmission chain will be small for each of the data types. We obtain the full dissimilarity d between two points a and b as the expected number of cases in between them under the null hypothesis; the product of the data type specific dissimilarities (figure 3)

$$d(a,b) \ = d_{gen}(a,b) \times d_{geo}(a,b) \times d_{time}(a,b) \qquad (2)$$

When the data are continuous (i.e. all observed values are unique) it is possible to analytically obtain the full distribution for d under the null hypothesis. For instance, d_{time} is distributed as the absolute value of the difference of two independent random variables following a discrete uniform distribution on $[1,N]$, with N the number of cases. When cases are infected locally there will be more small dissimilarities than under the null hypothesis. When a data type is discrete (as genetic data always are) several cases can have identical values for this data type. In this case we propose an extension to $d_i(a,b)$ in which the dissimilarity between two cases is the expected number of cases in between them if this identical value was due to small measurement error (appendix B.1).

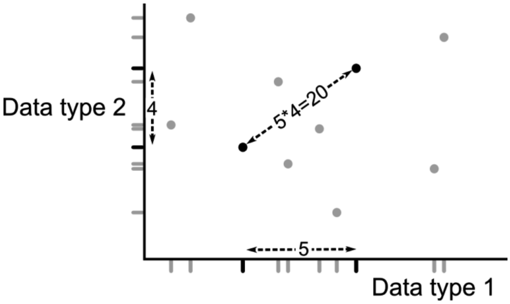


Figure 3. Graphical representation of the dissimilarity measure between cases. Shown is a dataset of nine cases and two (one-dimensional) data types. For each of the two data types, the dissimilarity between the two black cases is given as the number of cases in between them (for that data type), including one of the two black cases. This definition ensures the dissimilarity between a case and itself is zero. The total dissimilarity between the black cases is then given as the product of these, here 5×4=20.

Finding putative transmission clusters

We want to assess for a given subset *S* of the dataset *D* whether it is a local transmission cluster, i.e. whether the cases in *S* are closer together than would be expected under the null hypothesis. Define the unique 'weakest link' *l(S)* of a cluster as the largest dissimilarity in the minimal spanning tree of *S*; the larger the dissimilarity the less likely it is that all cases in the cluster are part of a local transmission chain. *l(S)* increases with *S*; we therefore compare *l(S)* to the value we would expect under the null hypothesis for a cluster of at least this size. We call *S* a putative transmission cluster (PTC) if the probability of observing a cluster of at least this size with weakest link at most *l(S)* under the null hypothesis is less than 0.001. This probability can be obtained by permuting the dataset (see appendix B.2 for details). The upper bound for the probability (here 0.001) should be small but other than that is arbitrary, and could be changed depending on the application.

It is important to note that the tests applied to each of the clusters encountered in the hierarchical clustering scheme are not independent. For example, if a set of ten cases is found to be a PTC then the set of eleven cases constructed by adding one random nearby case will probably also be a PTC. This dependence is inherent to clustering algorithms and not necessarily a problem, as finding the cluster is usually more important than uniquely identifying all the cases that belong to it. It could, however, lead to a high false positive rate when assessing whether cases are correctly assigned to a PTC.

Testing on simulated datasets

To gauge the strengths and limitations of the algorithm, we tested it on simulated datasets where we know precisely which cases were part of a local transmission chain and which were import. The performance of any clustering algorithm depends on how strongly the clusters are separated in the data. For instance, if a dataset consists of several outbreaks clearly distinguishable in time and place, we expect an algorithm to do well. On the other hand, when cases belonging to one outbreak can be found throughout the spatiotemporal and genetic space, any algorithm will struggle to identify the outbreak. The simulations are thus focused on the intermediate region, where clustered cases are not easily distinguishable based on separate data types, but are still close enough that the combined information from the data types yields enough information for clustering.

We use two measures of the performance of the proposed method. First, we take the percentage of locally infected cases correctly assigned to a PTC. Second, we take the percentage of imported cases incorrectly assigned to a PTC. The former is a measure of the sensitivity, the latter of the false positive rate. If sensitivity is high whilst the false positive rate remains low, the method could be suitable for use in outbreak detection. As we assess performance of the method at the case level, while statistical tests are performed at the cluster level, the false positive rate is not guaranteed to be beneath the p-value of 0.001 used. If the false positive rate becomes too large the method would be unsuitable for outbreak detection, as too many false alarms would be given, but might still be useful in assessing properties of locally infected cases. We performed simulations under different incidence rates and with different sizes of the transmission clusters.

Each of the simulated datasets consisted of many unrelated cases and a small number of local transmission clusters, containing in expectation ten percent of the total number of cases for each of the simulations. All of the cases have a time, position and genotype associated with them. The geographical position of an imported case A is given by x_{A} ~ (Uniform(0,100), Uniform(0,100)), its time of sampling by t_A ~ Uniform(0,100), and its genotype gen_A is represented by a string of 8 random bits (each can be 0 or 1 with equal probability). Locally infected cases were simulated with infectors chosen as specified below. An infected case B and its infecting case A are related as follows; x_B~x_A+(Normal(0,4), Normal(0,4)), t_B~t_A+Exponential(1), and gen_B is generated from gen_A by flipping each bit with probability 1/16.

The absolute spatial distance between two cases was taken to be the usual Euclidean distance, with distances in both dimensions the minimum of $|x_1-x_2|$ and $100-|x_1-x_2|$. This makes the geographical region a torus, ensuring all clusters are fully observed. The absolute genetic distance was calculated as the number of different bits, leading to an expected genetic distance of 0.5 between infector and infected. The absolute temporal distance was the absolute value of the time difference. From these absolute distances dissimilarities were calculated for each data type, and combined into pairwise dissimilarities using equation (2).

In a first scenario all locally infected cases belong to the same outbreak, with an index case randomly chosen from the unrelated cases. In a second scenario we generated smaller transmission clusters, by letting 1/9 of all cases generate secondary cases according to a geometric distribution with mean R=0.5. These were themselves equally infectious, yielding transmission clusters of expected size 2. In the epidemiological literature this scenario is known as ‘stuttering transmission chains’:[166] small outbreaks occur but large outbreaks do not, since the mean number of secondary cases per infectious case, or effective reproduction number, R, is smaller than one. In a third scenario we generated even more and smaller outbreaks, with each case generating new cases according to a geometric distribution with mean R=0.1, yielding a dataset with many very small transmission clusters of expected size 1.11. For all of these scenarios, we performed simulations with an initial number of unrelated cases of 90, 450 or 900, representing different incidence rates, yielding nine simulation scenarios in total. The expected total number of cases for these simulations are thus 100, 500 and 1000. For each of these nine scenarios, we simulated 100 datasets, and applied the methodology to identify significant clusters.

The results of the clustering algorithm depend on how related infector-infected pairs are in each of the data types. When this relation is stronger, we expect clustering results to improve. We therefore performed additional simulations where the distances between infector-infected pairs are smaller (appendix B.3).

Many actual surveillance datasets face the problem of missing or unobserved cases. This is similar to a scenario where all cases are observed, but the relation between infector-infected pairs is weakened. To illustrate this, we performed analyses on simulated datasets from which we randomly discarded 20% of all cases (appendix B.4).

Results

Figure 4 gives a graphical representation of a typical simulated dataset with small outbreaks for each of the separate data types, figure S1 (appendix B) gives the same information for a simulated large outbreak. Under the chosen parameter settings, identifying the different clusters by only looking at the separate data types is challenging.

The results for finding local transmission for all simulation scenarios are given in figure 5 and table 1. For each simulation scenario, we report the distribution and median of the percentage of cases assigned to a putative transmission cluster (PTC), both for locally infected and for imported cases. For all scenarios, the percentage of locally infected assigned to a PTC is higher than the percentage of import cases assigned to a PTC. This means that for all scenarios the three data types, when combined, provide sufficient statistical signal to identify local transmission.

In general, the method assigns outbreak cases to PTC's more often when transmission clusters are larger. This is no surprise as the strength of the statistical signal increases with outbreak size. This higher sensitivity comes at the cost of a lower specificity; when assessing the large outbreaks of expected size 100 the median false positive rate is 0.16, while it is near 0 for most other scenarios. The false positives here are cases that are, coincidentally, close to the actual cluster; the number of such cases increases with the size of the outbreak.

The method has a lower sensitivity and specificity when the incidence rate is higher. This is because there are more unrelated cases per unit of space, time and genetics, while the absolute inter-outbreak case distances remain the same. Therefore, the ordinal distance between outbreak cases becomes larger when incidence rates are higher, which makes it harder to identify transmission clusters (appendix B figure S3, table S2). When inter-outbreak case distances become smaller, outbreaks are easier to detect (appendix B figure S2, table S1).

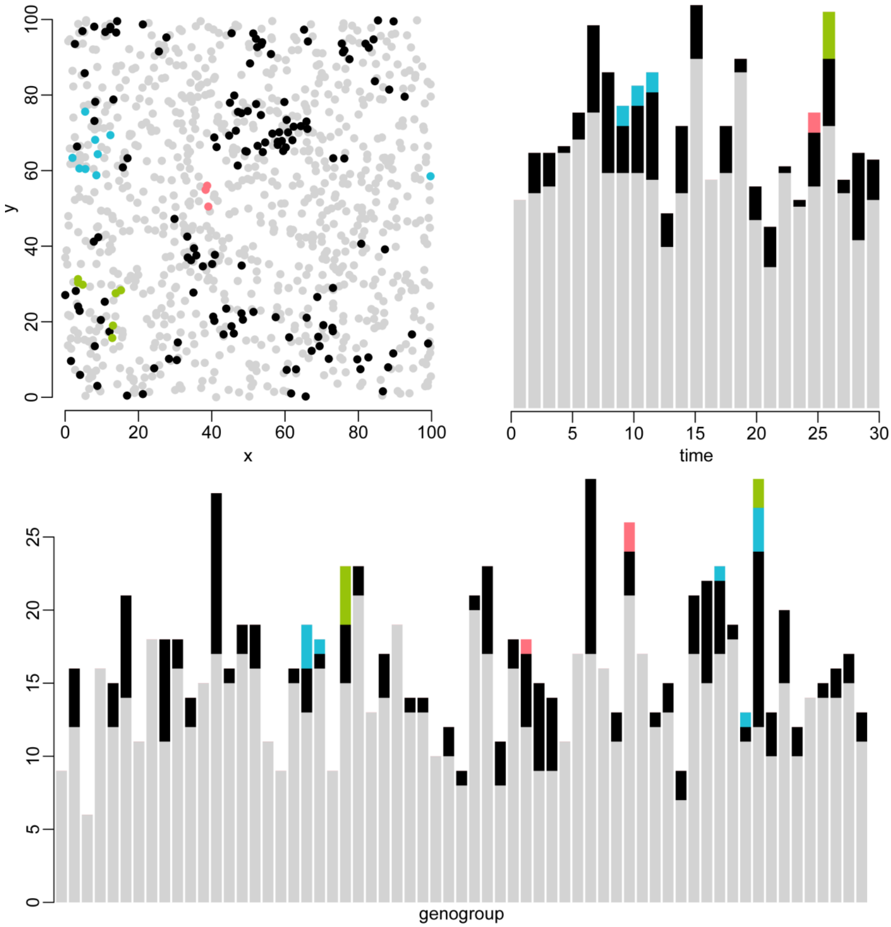


Figure 4. Graphical representation of the three data types for a typical simulation. This simulation consisted of 1019 cases of which 119 (12%) were infected by other cases. In total, there were 158 related cases belonging to 39 transmission chains, and 861 unrelated cases (grey). To visualize individual transmission chains, three chains were chosen at random and drawn in blue, green and pink. (A) Geographical location of all simulated cases. The geography is a torus, so the right side is equated with the left side, and the top side is equated with the bottom side. (B) Simulated cases over time. (C) Simulated cases have one of 28=256 possible genotypes. For clarity, the distribution of cases over 64 genogroups is plotted; a genogroup is defined as a set of four genotypes that are identical up to the last two digits. The order of the genogroups on the x-axis does not reflect genetic distance. Note that outbreaks cannot be accurately identified using only one of these data types.

Table 1. Median of sensitivity/false positive rate of assigning locally infected cases to a putative transmission cluster for simulated datasets.

	High incidence	**Low incidence**	**Very low incidence**
Large outbreak	1.00/0.16	1.00/0.06	1.00/0.01
Small outbreaks	0.49/0.01	0.58/0.01	0.60/0.00
Very small outbreaks	0.06/0.00	0.08/0.00	0.11/0.00

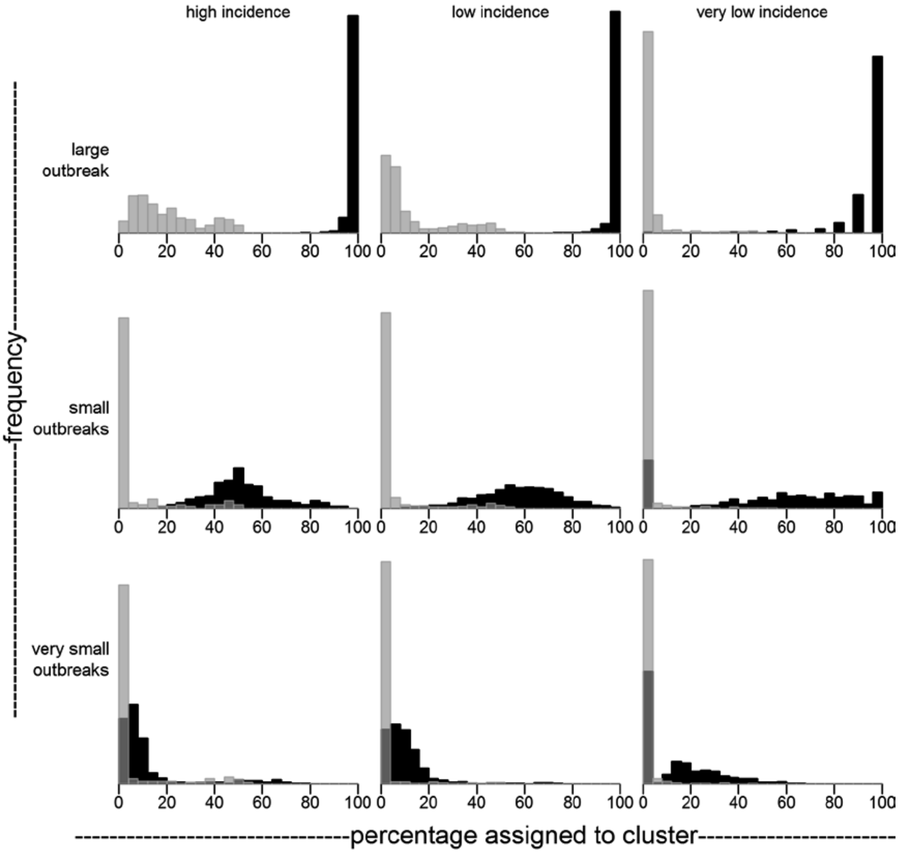


Figure 5. Sensitivity (black) and false positive rate (grey) for analyses on simulated datasets. For each of nine simulation scenarios, percentage of outbreak (black) and non-outbreak (grey) cases assigned to a putative transmission cluster are shown. In each scenario, ten percent of all cases are an outbreak case. Total expected number of cases is (left column) 1000, (middle column) 500 or (right column) 100. Outbreak cases belong to (top row) one large outbreak, (middle row) small outbreaks caused by 1/10 of cases being contagious with basic reproduction number R=0.5, (bottom row) very small outbreaks caused by all cases being contagious with R=0.1. For all scenarios, outbreak cases are distinguishable from unrelated cases. Sensitivity increases with outbreak size, at the cost of an increased false positive rate. Sensitivity and false positive rate improve when the incidence, or equivalently the number of cases in the same region of spacetime, decreases. Figure 4 corresponds to a simulation from the middle left panel.

Discussion

We have presented a method to identify transmission chains of locally infected contagious disease cases in large databases containing temporal, geographical and genetic data. The method does not require assumptions on population at risk or pathogen-specific properties. The method is novel in explicitly incorporating the genetic distances measured between sampled pathogens, and in accounting for the chain-like structure of transmission chains of contagious diseases.

Several methods to find locally infected cases in large datasets have been published and some are commonly used in epidemiological investigations.[58-60,161] However, many of these methods were not explicitly developed for analysis of contagious diseases, and ignore the fundamental characteristic of contagious diseases: that each infected case can itself be a source for new cases. The method presented here does take this into account by focusing on the distances between cases, rather than on the number of cases in a particular region of space-time. This latter approach is suitable when cases are caused by one common source, rather than by the cases themselves.

The ability of the method to correctly cluster together cases belonging to the same transmission chain depends on how ‘close’ these cases are to each other in time, space and genetics, relative to non-related cases. This depends both on the properties of the pathogen studied, the incidence of disease, and the size of the study region and duration. For example, a dataset resulting from a study period of one year on a pathogen with an average serial interval of half a week would for our method, due to the ordinal distances used, be equivalent to a dataset resulting from a study period of ten years on a pathogen with a serial interval of five weeks and ten times lower incidence. Shorter serial intervals, lower incidence rates and longer study periods allow for more accurate identification of outbreak cases.

In our simulations, we have taken the Euclidean distance between geographical locations. However, this is not always the most relevant distance metric. For example, people are more likely to travel between densely populated areas than between sparsely populated areas.[167] Thus, when a study region encompasses both urban and rural areas, more relevant measures of distance could be given by mobility patterns[168] or road distances.[98] Note that no extra correction is needed to adjust for the higher urban population densities leading to more cases: this is taken into account by the ordinal distance used.

The non-parametric method introduced is able to identify locally infected cases when little is known about the pathogen studied. When precise pathogen-specific information or information on population at risk is available, a more precise description of the system can be given. More specific methods can be used and information can be obtained from the data types separately, which should lead to better identification of transmission clusters. The non-parametric method could still be of value in such as a scenario, as it provides a simple first-try approach: results can be compared to those obtained from an analysis that uses more information, and assumptions made about pathogen characteristics and population at risk can be tested.

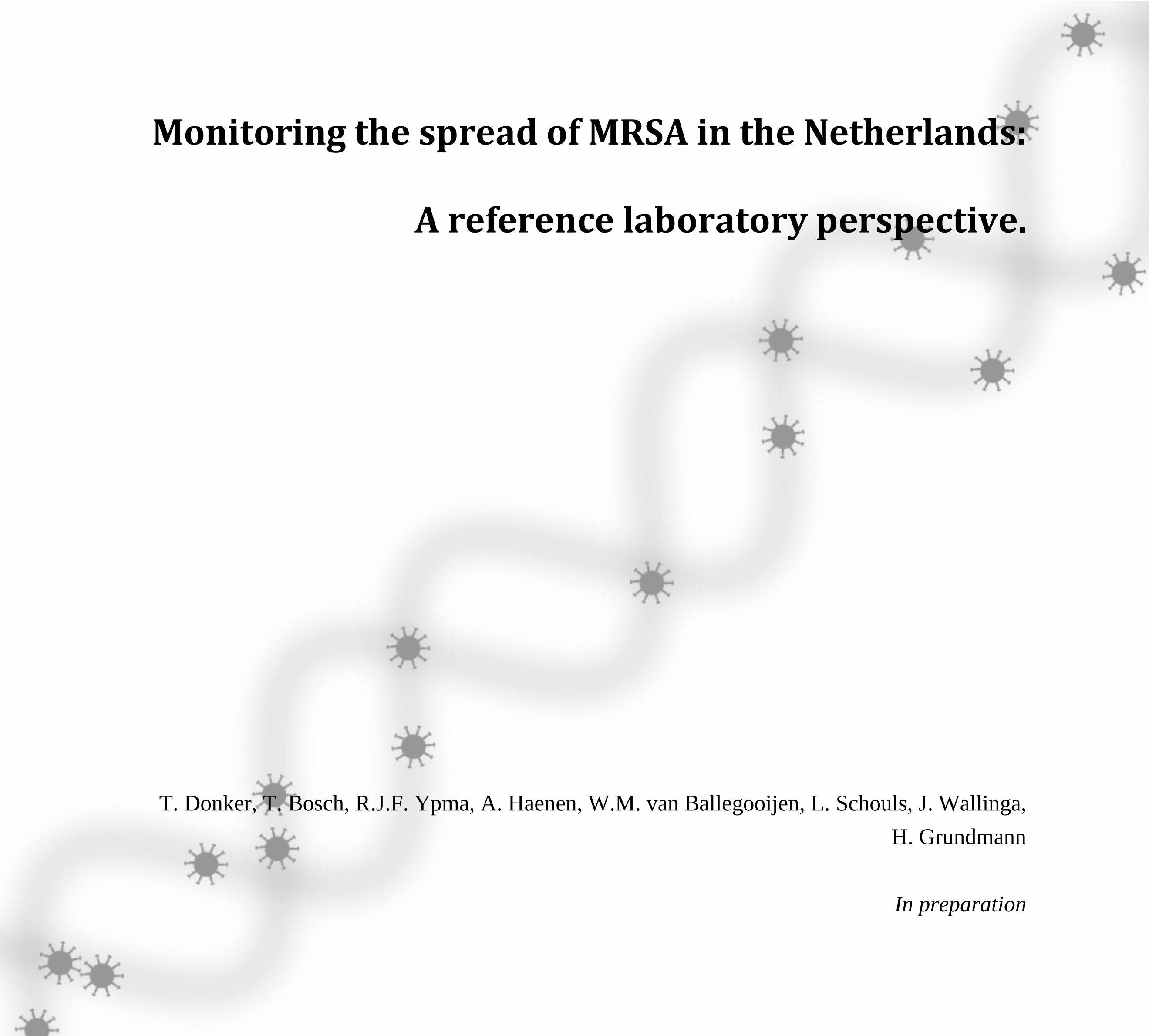
Here we performed validation of the method using simulated datasets, in which the origin of cases is known. Results obtained give confidence the method can be sensibly applied to actual surveillance datasets. Examples of existing large molecular epidemiological databases include VNTR typing datasets of tuberculosis,[169] spa typing datasets of MRSA[170] and short read sequencing datasets of hepatitis B, hepatitis A and norovirus.[99,171,172] Note that the relevant spatial information differs for these datasets; we might focus on place of residence for tuberculosis, but on hospital, ward or even bed for MRSA. Future work will have to focus on applying the method presented here to such datasets.

The method presented has some drawbacks. First, the null hypothesis states that all cases in the dataset are independent. This leads to a bias when many transmission clusters are present; as the locally infected cases cluster together, the remaining independent cases will themselves lie closer together (in ordinal distance) than under the null hypothesis. Thus, especially when a high percentage of cases are locally infected, the statistical test could overestimate the percentage of clustered cases. This overestimation might be alleviated if prior information on the percentage of cases locally infected were available. Second, we assumed independence between the different data types for unrelated cases. These data are never truly independent, as all cases belong to the same phylogenetic tree acting over a long time scale. The local outbreaks we are interested in can be seen as local tips of these large trees. Whether the data can be approximated as being independent depends on the spatial dynamics of the pathogen, the evolutionary time separating sampled pathogens and the size of the region studied. For example, the approximation might be valid when studying MRSA at the hospital level, but not at the level of a continent as geographical structure can be seen in genotypes sampled.[170] Third, we have not taken into account the boundaries of our datasets, i.e. the edges of the geographical area and time window studied. This might decrease the sensitivity of finding clusters near the start or end of the study period, and should be addressed when the method is applied prospectively.

Results of clustering methods such as the one presented are important in epidemiological investigations for a number of reasons. First, they provide a measure of how much local transmission takes place. For example, if many putative transmission clusters are found in a hospital setting, infection control measures have to be intensified. Second, the algorithm can be used as a tool to find transmission clusters in large databases that can then be further investigated, removing the need for a detailed analysis by hand of the complete database. Third, properties of clustered cases can be compared to non-clustered cases. This will, for example, allow researchers to test whether patients of a particular age are more prone to transmit disease, or whether certain genotypes are more likely to spread in a hospital setting. These applications differ in their requirements on the sensitivity and specificity of the algorithm, where generally the second application will have the most stringent, and the third application the most relaxed requirements.

With the decreasing cost of sequencing and genotyping techniques, the availability of genetic data continues to grow. In particular, in many surveillance settings large molecular epidemiological databases have been set up. As the size and complexity of these databases grows, we can only expect the usefulness of automated methods such as these to assist in answering public health questions will grow concordantly.

Chapter 4



Monitoring the spread of MRSA in the Netherlands:

A reference laboratory perspective.

T. Donker, T. Bosch, R.J.F. Ypma, A. Haenen, W.M. van Ballegooijen, L. Schouls, J. Wallinga,
H. Grundmann

In preparation

In the Netherlands efforts to control methicillin-resistant *Staphylococcus aureus* (MRSA) in hospitals have been largely successful due to stringent admission screening and isolation of patients that fall into defined risk groups. However, Dutch hospitals are not free of MRSA, whereby an increasing number of cases are coincidental findings that do not belong to any of the defined risk groups. Some may still be the result from undetected nosocomial transmission, while the origin of others remains unknown. Combining available information such as date of isolation, geographical origin and MLVA typing data for all MRSA isolates submitted between 2008-2011 to the National Institute for Public Health and the Environment, we applied a novel clustering algorithm to map the geo-temporal distribution of MRSA by MLVA clonal complex over the entire Dutch health care network. Of the 2966 isolates reported as coincidental findings, 579 were part of geo-temporal clusters, while 2387 were classified as MRSA of unknown origin (MUO). We also observed marked differences in the proportion of isolates from different MLVA types associated with clusters among hospital patients (MC45 46%, MC22 35%, MC8 27%, MC398 4%) indicating differential transmissibility. The majority of clustered isolates (74%) were reported from more than one laboratory. The frequency of MRSA of unknown origin among patients with coincidental MRSA is an indication of a largely undefined extra-institutional but genetically highly diverse reservoir. Efforts to understand the emergence and spread of high risk clones require the pooling of standard epidemiological information and typing data into central databases.

Introduction

Occurrence of MRSA in hospitals differs markedly between countries.[173] The low levels in the Netherlands have conventionally been attributed to the so-called Dutch search and destroy policy,[174] which stipulates that all patients who have had MRSA in the past or who are regarded at high risk of being colonised or infected are screened on admission and treated in strict isolation until screening results become available.[175] Despite of these efforts, hospitals in the Netherlands are not free of MRSA. Time and again, MRSA is isolated from hospitalised patients, considerable time after admission without obvious risk factors. This leads to extensive screening of contact patients and possibly even hospital staff which can be rather disruptive.

Patients with coincidental MRSA findings pose another epidemiological and public health challenge. They either represent cases of primary introduction which are regarded as MRSA of unknown origin (MUO) or are the result of unobserved secondary spread. The first would be suggestive of a tip of the iceberg phenomenon whereby the frequency with which MUOs are isolated in hospitals would be a reflection of an undetermined extra-institutional reservoir, the second an indication of hidden intra- or inter-institutional transmission chains.

Conventional hospital-based investigational epidemiology has its limitations as it can only link cases with an apparent epidemiological association, for instance if patients had shared a room, whereas data available at national public health institutes or reference laboratories contain too little information to make these obvious epidemiological links. However, molecular typing data, and information about the location and time of isolation may provide sufficient detail to address some of the above mentioned challenges.

By combining data available at the National Institute for Public Health and the Environment of the Netherlands we map the distribution of MLVA types over time and space and provide clues to the size and clonal composition of institutional and multi-institutional MRSA clusters. We further determine the number and genetic diversity of MUOs as an estimate of the frequency of primary hospital introductions by patients without the obvious risk factors of MRSA carriage. Using the results of these analyses, we draw conclusions about the ability of the search and destroy policy to control MRSA in hospitals using the current guidelines, and the existence of unknown reservoirs outside of hospitals.

Methods

Data and algorithm

MRSA Surveillance

Primary MRSA isolates from patients admitted to Dutch hospitals are routinely sent to the National Institute for Public Health and the Environment by all microbiological laboratories in the

Netherlands for reference typing. Each isolate is investigated by Multi Locus Variable Number of Tandem Repeat Analysis (MLVA) and *spa* typing, *mecA* and *lukS lukF* PCR as part of the standard reference service. The health-care institutions are asked to complete a questionnaire about epidemiological metadata for each isolate. This questionnaire includes questions about the origin of the isolate, demographics of the patient, the reason for sampling, and whether or not they belonged to one of the risk-groups defined by the Workgroup Infection Prevention (WIP).[175] We used data collected between 2008 and 2011 (4 years), and included only the isolates for which information for the following variables was available: date of isolation, MLVA type, residence postal code of patients, and institution. We excluded duplicate isolates from the same patient, same institution and same year. We used typing data partitioned at the level of MLVA clonal complex (MC). Because of the large number of isolates belonging to MC398, and because the algorithm is computationally intensive, we only used data from 2009 for this clonal complex.

Cluster Algorithm

To identify clusters of isolates with higher than expected geo-temporal occurrence in the MRSA surveillance database, a non-parametric clustering algorithm proposed by Ypma et al.[176] was used. In short, this algorithm uses the ranked distance between any two samples for each of the variables, to calculate the combined distance between them. Each isolate is assigned to its nearest neighbour, being the closest isolate in the combined ranked distance. The resulting tree is subsequently recursively split up by its weakest link, resulting in N-1 possible clusters, where N is the number of isolates. The combination of cluster size and the strength of the weakest link in the cluster is then tested against 1000 random generated transmission trees to assess whether its weakest link was stronger than could be expected for a cluster of the same size under random assumption.

Ranked distance

For each variable, the pairwise ranked distance is calculated. The distance between isolate a and b is defined as the number of isolates between them:

$$d_i(a,b)=|\{p\text{: } D_i\ (a,p) \leq D_i\ (a,b) \wedge D_i\ (b,p) \leq D_i\ (a,b)\}| - 1$$

Where |.| denotes the number of elements of a set, D_i the absolute distance in this data type and the -1 ensures $d_i(a,a)=0$. We obtain the full distance d between two points a and b as the product of the data type specific distances:

$$d(a,b)\ = d_{gen}(a,b) \times d_{geo}(a,b) \times d_{time}(a,b)$$

Permutation test

To assess the distribution of the weakest link-strength for each cluster size under random assumption, we performed a permutation test. For each of the variables, a random order of samples was chosen, creating random combinations of the variables per isolate. Subsequently, all weakest link-cluster size combinations were determined as described above. For each cluster size, we selected the strongest link out of the set of weakest links found for this permutation dataset. This process was repeated on 1000 permutation datasets.

Distance measures

For each of the variables, a pairwise distance metric (D_i) needs to be defined. For date and residence postal code the difference between the dates of isolation in days, and Euclidean distance between the centroids of the postal code areas were used. For the MLVA data, the number of VNTR loci difference between the isolates was used (range 0-8).

As distance between health-care institutions, we used the referral distance. This is the shortest path between two hospitals through the national patient referral network,[177] formed by the exchange of patients between hospitals. The shortest paths between all hospitals were determined using a simulated patient referral network (see appendix C).

Epidemiological validation

We first performed the cluster algorithm using isolation date, residence postal code and MLVA, ignoring the health-care institution. This was done because patients tend to be admitted to their local hospitals,[178] using both residence postal code and health care institution would therefore result in geographical clusters even if temporal and genetic structure was absent. As a consequence, we performed a second algorithm using isolation date, institution and MLVA in parallel. For each MLVA clonal complex, we tracked the number of isolates that fall inside and outside the clusters, the total number of clusters, and the number of clusters that included isolates from more than one hospital.

False positive findings of clusters, and in particular large clusters, can skew the results dramatically. We therefore assessed the influence of single isolates on the clustering results by repeating the analysis on randomly chosen subsets of the data. We selected 90% of the isolates and repeated the cluster algorithm and permutation test. This jack-knife process was repeated 1000 times. The distribution of the percentage clustered isolates was compared with the point-estimate, using the entire dataset.

Certain risk groups defined by the WIP are more likely associated with nosocomial transmission; their isolates can be expected to cluster more often than those from risk groups associated with introductions from reservoirs outside the national health care system. The results of the clustering algorithm should support this intuition. For instance, if patients were screened because they had previously been hospitalised outside the Netherlands, the isolates can be expected to be introductions, and should therefore not be clustered. Conversely, having had contact with a known MRSA carrier can be regarded as risk categories that would typically be part of a transmission chain. Isolates with multiple –opposing– risk factors, such as patients who have been in contact with farm animals and have had contact with a known MRSA carrier were gathered in the “ambiguous” group, because no definite expectation can be determined. We compared the odds of an isolate belonging to a cluster, stratified by risk groups, to test whether the assignment of the isolates to a cluster coincided with the expectations based on epidemiological profile.

Results

MRSA Surveillance

Between 2008 and 2011, 14042 MRSA isolates were submitted to the National Institute for Public Health and the Environment (RIVM) for reference confirmation and typing. Of these, 2864 were duplicate isolates and 2226 lacked information about the patient’s place of residence and were excluded from the analysis. Molecular typing grouped 7720 (86.2%) isolates into five dominant MLVA clonal complexes (MC, table 1). The remaining 1232 isolates belonged to an additional 11 MCs. The MC that contained most isolates (3852, 43.0%) was MC398 coinciding with sequence type (ST)398 which represents livestock-associated MRSA in the Netherlands.[179] Since cluster analysis scales exponentially with the number of isolates, inclusion of all MC398 isolates became computationally too intensive and only MC398 isolates for 2009 were included into the current study resulting in a total of 6295 isolates.

For 1561 of the 6295 isolates (24.8%), no epidemiological metadata were available. A total of 1405 isolates (22.3%) represented coincidental finding without details of their WIP risk category, 818 (13.0%) were identified as the result of screening efforts related to contact tracing, and 754 (12.0%) were reported as expected cases. Of the remaining 1757 (27.9%) isolates the patient’s WIP defined risk group was known. The most frequent risk group was “contact with farm animals” which coincided with MC398. Among the other MCs the most common risk groups were “admission from a foreign hospital” and “known MRSA carrier” (see table 1).

Table 1. The number of isolates assigned to a cluster (total number of isolates), by MLVA clonal complex (MC) and epidemiological information, such as the risk groups defined by the Workgroup Infection Prevention (WIP).

	All*	MC398*	MC5	MC8	MC45	MC22
All isolates	1724 (6295)	52 (1195)	571 (1416)	356 (1292)	285 (611)	193 (549)
WIP Introductions	94 (1064)	33 (507)	23 (158)	12 (122)	8 (49)	9 (86)
Unexpected Cases	276 (1405)	2 (72)	77 (321)	69 (384)	43 (139)	22 (118)
No Information	303 (1561)	8 (362)	117 (338)	40 (305)	51 (114)	33 (120)
WIP Ambiguous	34 (153)	1 (19)	17 (48)	5 (30)	5 (120	2 (13)
Expected Cases	184 (754)	6 (201)	53 (146)	35 (127)	28 (59)	14 (62)
WIP Transmission	354 (540)	1 (16)	151 (213)	77 (115)	39 (67)	57 (69)
Contact Tracing	479 (818)	1 (18)	133 (192)	118 (209)	111 (171)	56 (81)
Contact with Farm Animals	*46 (543)*	*33 (497)*	*5 (12)*	*3 (8)*	*3 (7)*	*0 (4)*
Admitted from foreign hospital	*31 (389)*	*0 (7)*	*12 (109)*	*4 (87)*	*4 (37)*	*9 (80)*
Adopted child	*11 (112)*	*0 (2)*	*4 (29)*	*5 (25)*	*0 (2)*	*0 (2)*
>2 months ago in foreign hospital	*7 (22)*	*0 (1)*	*4 (11)*	*0 (2)*	*1 (3)*	*0 (0)*
Foreign dialysis patient	*2 (3)*	*0 (0)*	*1 (2)*	*0 (0)*	*0 (0)*	*0 (0)*
Known MRSA carrier	*31 (149)*	*1 (20)*	*15 (46)*	*4 (29)*	*5 (11)*	*2 (12)*
Known HCW MRSA carrier	*2 (5)*	*0 (0)*	*0 (0)*	*2 (2)*	*0 (1)*	*0 (1)*
Protected contact with MRSA carrier	*10 (21)*	*1 (2)*	*1 (4)*	*4 (6)*	*0 (1)*	*4 (7)*
Unprotected contact with MRSA carrier	*115 (194)*	*0 (5)*	*49 (77)*	*24 (34)*	*14 (30)*	*18 (21)*
Admitted to hospital with known MRSA problem	*121(178)*	*0 (5)*	*61 (81)*	*31 (48)*	*10 (14)*	*13 (18)*
Shared room with MRSA patient	*110 (152)*	*0 (4)*	*42 (55)*	*18 (27)*	*15 (22)*	*22 (23)*

*MC398 only includes isolates from 2009, all others 2006-2009. “All” denotes the sum of all included isolates.

Clusters

Combining the distances between isolation date, residence postal code and MLVA variables, we found 152 clusters consisting of 1724 isolates in total that were closer together than could be expected by chance (figure 1). Between different MLVA clonal complexes, the proportion of isolates that formed clusters differed considerably: only 4.4% of the MC398 isolates were part of a cluster, whereas 46.6% of the MC45 isolates were part of a cluster (figure 2). The dominant community and animal-associated MLVA complexes show lower proportions of clustered isolates than the hospital-associated MLVA complexes. The identified clusters were robust since point estimates were within the interquartile range of the jack-knife estimates for the majority of dominant MLVA clonal complexes.

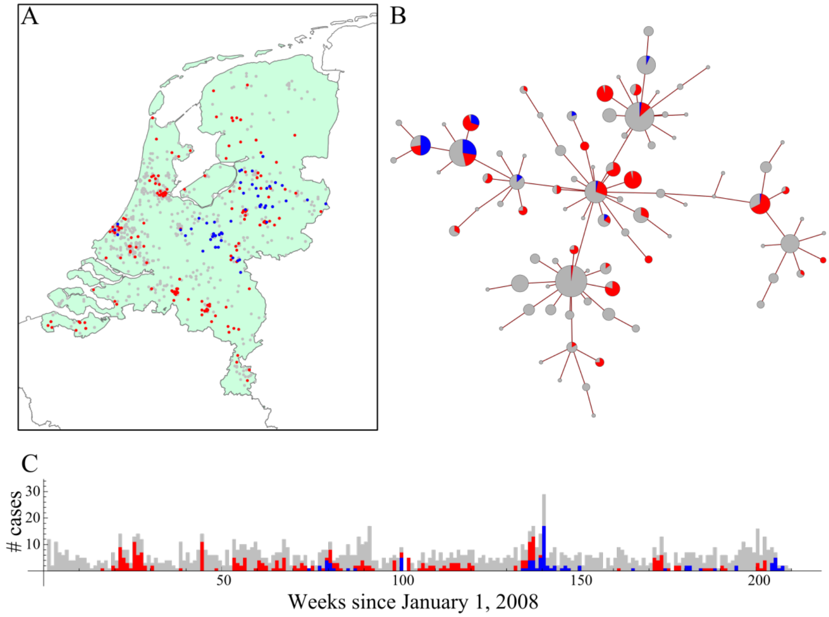


Figure 1. The result of the clustering algorithm for MC8, showing the three data sources used: A) The postal code of the patients' home addresses, B) the MLVA cluster, here shown as a minimum spanning tree, and C) time at which the sample was received, here shown in weekly aggregate numbers. Grey denotes isolates outside clusters, blue and red those within clusters. Blue shows the isolates present in the single largest cluster. The three sources individually show no clear clustered pattern, and clusters only emerge during combined analysis.

Sixty-eight clusters (45%) included isolates that originated from more than one health care institution. Most of the larger clusters include multiple institutions, and 74% of the clustered isolates are present in a multi-institutional cluster.

When using the health care institution instead of residence postal code, we found 142 clusters, consisting of 1420 isolates, across all MLVA clonal complexes. In 47 clusters, isolates from more than one institution were included. The clusters in the residence postal code analysis and the health care institution analysis largely overlap; 1120 isolates clustered in both analyses, and the proportions of clustered isolates per MLVA clonal complex show a similar pattern in both analyses (appendix C figure S1).

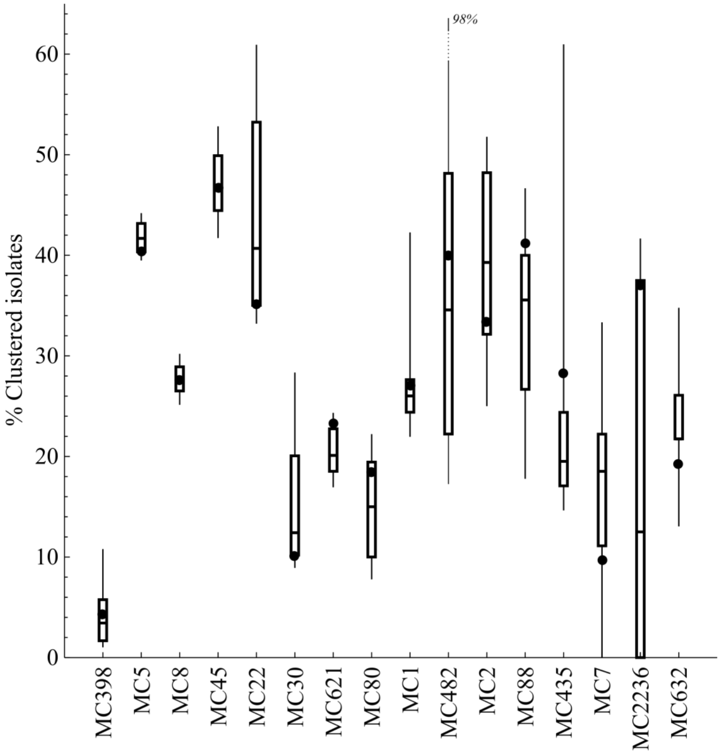


Figure 2. The proportion of isolates part of a cluster according to the algorithm, dots show the point estimate, box plots (median, IQR, 95% interval) are the result of the 90% jack-knife analysis. Common hospital-associated MRSA strains (MC5, MC45, MC22) show higher proportions of clustered cases than community- (MC8) or livestock-associated (MC398) MRSA.

Epidemiological validation

The results from the cluster analysis largely correspond with the available epidemiological information. The majority (59%) of isolates identified as part of contact screening were part of a cluster, while of the isolates from patients who have been admitted after hospitalisation abroad, only 8% clustered. Taking these numbers, we observe that the odds of isolates being clustered overlaps with the expectation of cases being part of a transmission chain (figure 3). Of the coincidental finding, both unexpected findings and those without epidemiological information, respectively 20% and 19% of the isolates clustered.

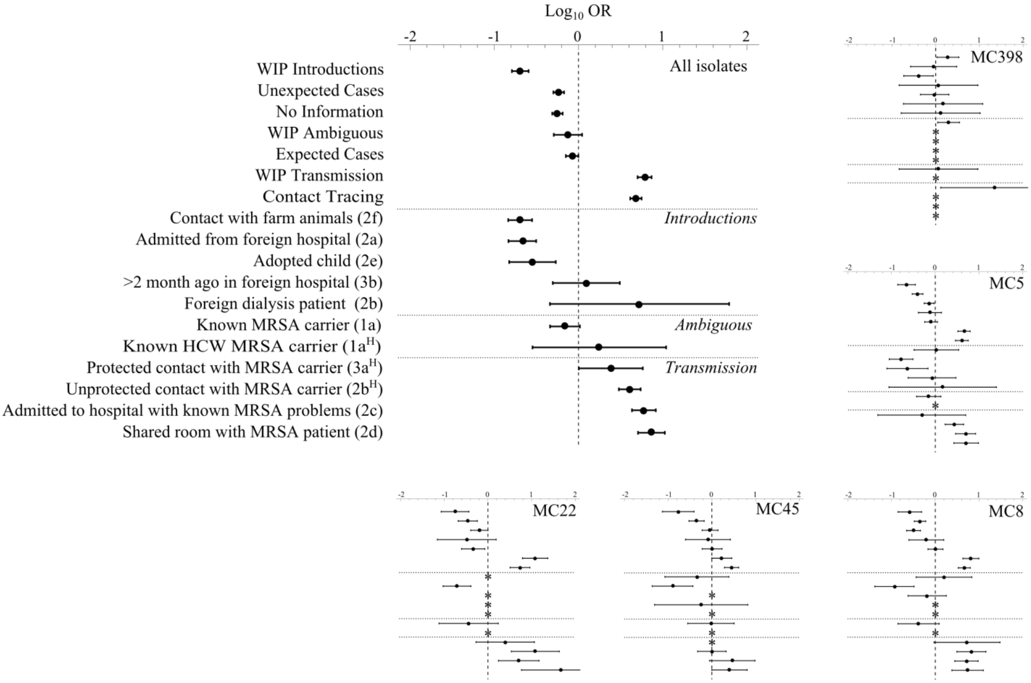


Figure 3. The odds of belonging to a cluster of isolates differ considerably between epidemiological risk groups. WIP risk groups were gathered on the basis of the expectation of belonging to a transmission chain (introductions, ambiguous and transmission). The algorithm results overlap with the predictions on the basis of WIP risk group. The unexpected cases and isolates without epidemiological information more often fall outside the clusters.

The pattern in isolates positive for the Panton-Valentine Leukocidin (PVL$^+$) gene also follows expectation. This marker, commonly associated with community origin,[180-182] was found significantly more often in isolates that did not cluster than in those that did cluster (figure 4). In MC8, which harbours the largest part of the PVL$^+$ isolates (40.5%), 12% of the clustered isolates were PVL$^+$ against 50% among the other isolates. Taking into account that the proportion of clustered isolates is already low in MC8, this means that just 8% (43 of 515) of the PVL$^+$ MC8 isolates is present in a cluster. The proportion of risk group categories and PVL$^+$ isolates is not stable over the study period (table 2). The proportion PVL$^+$ isolates, unexpected cases, and isolates collected through contact tracing clearly increased over time, while the proportion of isolates in a transmission-associated risk group decreased. Despite the significant trends in a number of risk groups, the proportion of clustered cases remained virtually unchanged over time.

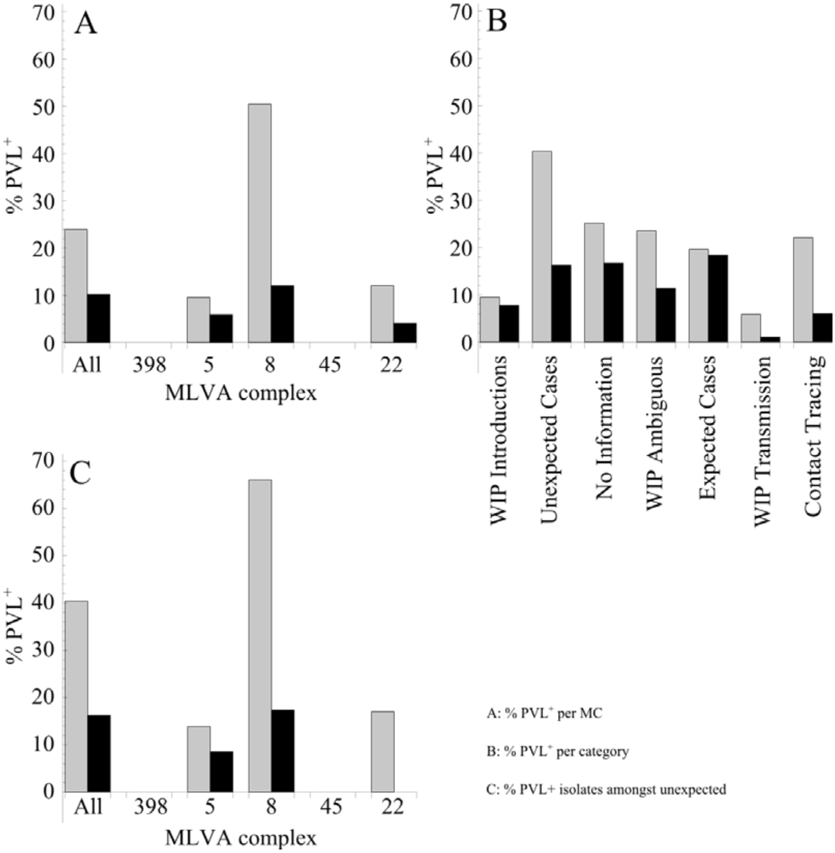


Figure 4. The proportion of isolates positive for Panton-Valentine Leukocidin (PVL) is higher amongst cases outside the clusters. Black bars show clustered isolates, grey bars show non-clustered isolates. A) The proportion of PVL positive cases for the largest MLVA clonal complexes (MCs), B) for each of the risk groups and C) in the unexpected cases of the largest MCs. Unexpected cases outside the clusters, especially in MC8 isolates, contain the highest proportion of PVL$^+$ isolates. It is likely that these cases are introductions into the hospitals of community-acquired MRSA.

Table 2. The trend of categories of isolates over time, showing the Spearman's rank correlation coefficient (ρ) for the monthly proportions of each of the categories from January 2008 to December 2011. (*p<0.05 **p<0.01)

	All	**MC5**	**MC8**	**MC45**	**MC22**
Clustered Cases	-0,036	0,185	-0,156	-0,071	0,158
PVL+	0,640 **	0,433 **	0,612 **	-	0,251
WIP Introductions	-0,081	-0,195	-0,090	-0,058	0,209
Unexpected Cases	0,355 *	0,124	0,205	0,071	0,114
No Information	-0,314	-0,158	-0,217	-0,186	-0,343 *
WIP Ambiguous	0,278	0,101	0,097	0,107	0,155
Expected Isolates	-0,208	0,021	-0,163	-0,139	-0,347 *
WIP Transmission	-0,371 **	-0,325 *	-0,180	-0,217	0,007
Contact Tracing	0,495 **	0,619 **	0,283	0,315*	0,281

Discussion

Thanks to the stringent screening of risk groups, many MRSA colonised patients are identified and isolated on admission to Dutch hospitals. However, a large fraction of the patients found to be colonised with MRSA did not belong to any of the defined risk groups and were coincidental findings. The corresponding isolates are often just marked as unexpected findings, or have no epidemiological information available at all. Some of these coincidental isolates were the result of unobserved transmission while others were introduced from hitherto undiscovered reservoirs. Isolates in the last group can be gathered under the term “MRSA of Unknown Origin” (MUO).[183]

Utilising a novel clustering algorithm,[176] we were able to map the distribution of MLVA types over time and space across the entire Dutch health care system, and thus determine the size and clonal composition of clusters of related isolates. This allowed us to infer what proportion of isolates belonging to any specific clone is typically associated with intra- and inter-hospital transmission clusters and determine the number and genetic diversity of MUOs as an estimate of the frequency of primary introductions.

The algorithm assigned 579 of the 2966 (19.5%) coincidental findings to clusters, indicating that these are closely related to other isolates, probably through transmission. The other coincidental findings (80.5%) did not significantly cluster with any other isolates, and could be considered to be MUOs. The exact reservoir remains elusive, although some indications are visible in the results. The high proportion of PVL$^+$ isolates among the MUOs in MLVA clonal complex 8 (MC8) for instance hints towards a community origin.[180-182] The likely fact that a number of isolates resulting from transmission were initially classified as coincidental findings illustrates the inability of the surveillance system to identify all nosocomial transmission events.

The shortcomings in the epidemiological data are a reflection of how conventional epidemiological methods rely heavily on data collected on-site by doctors and nursing staff. Although the gathering of data for a single case may seem a small task, the combined task for all cases may prove difficult to incorporate in the tight schedule of a hospital. This reduces the sensitivity of the standard surveillance and investigational epidemiology because cases will be missed and direct links disappear. The autonomous nature of the clustering algorithm is therefore one of its key advantages, it only needs the most basic information: time, place and genetic profile. It can find related cases, even -and especially- when their relation is not obvious from the separate data sources.

The results indicate that cluster algorithm correctly identifies the related isolates; those generally considered to be the result of independent introduction are more often found outside the clusters, while isolates taken as part of contact tracing efforts, typically related, are often found within clusters. Despite the obvious ascertainment bias in the contact tracing isolates, these results do show that the algorithm correctly identifies epidemiologically linked isolates. However, the sensitivity of our algorithm depends on the sampling effort in hospitals. If fewer isolates from a single outbreak are collected, it will become harder for the algorithm to distinguish it from the independent introductions. This would mainly apply to small clusters; large transmission chains will still be detected, because the chance of finding two related cases increases with the number cases in the chain. It is therefore possible that some of the isolates assigned as independent introductions are part of a small, but largely unobserved, transmission chain which has been missed.

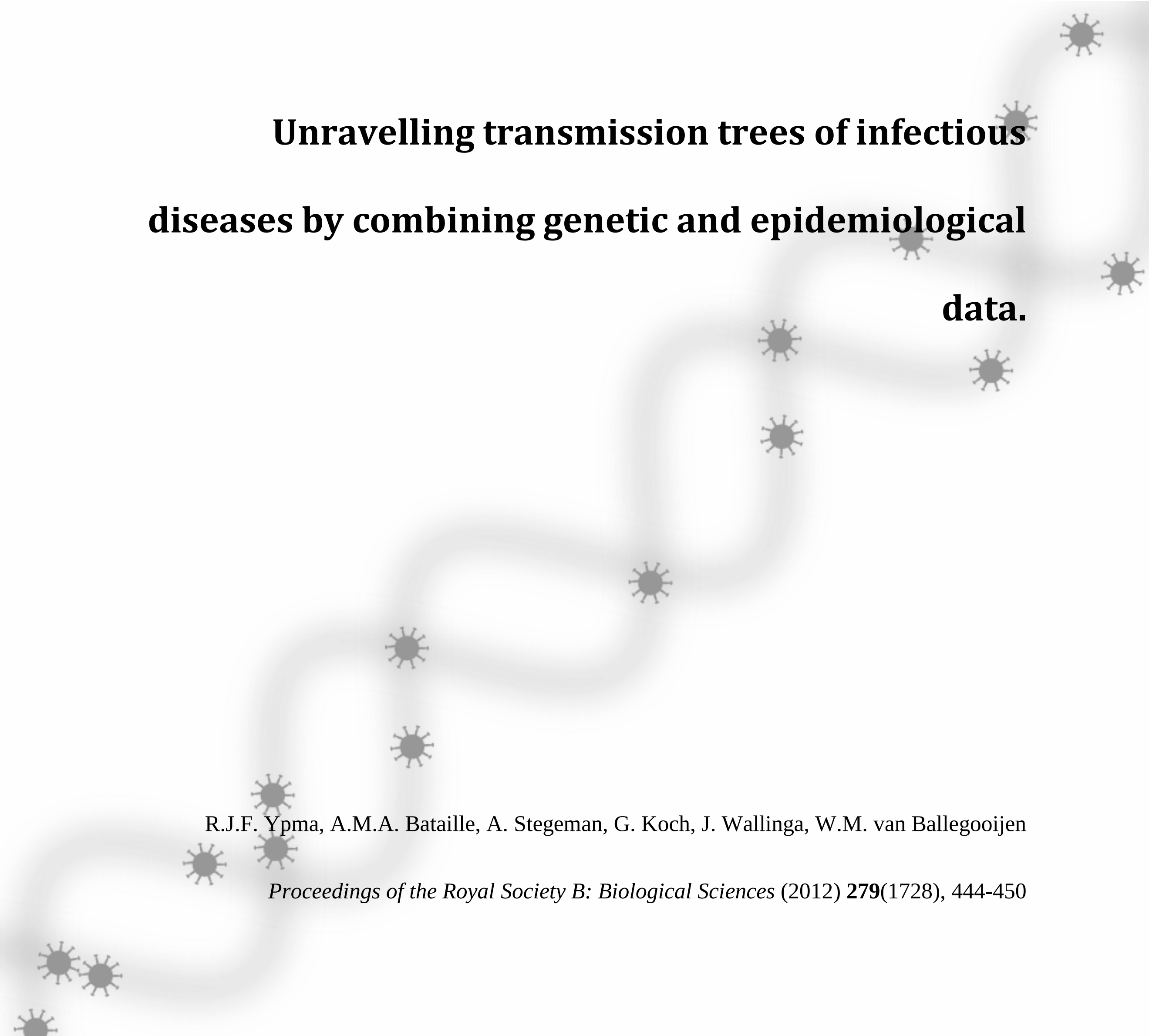
Using a cluster algorithm to identify clusters of related isolates also delivers novel insights in the dynamics of MRSA in the Netherlands that can aid the design of novel intervention strategies and guidelines. Firstly, many of the clusters comprise of isolates from multiple health-care institutions, which would have remained hidden if only the isolates of a single institution were analysed. These multi-institutional clusters are likely caused by transmission through exchanged patients transferring MRSA between hospitals. It shows how the occurrence of nosocomial pathogens in hospitals cannot be viewed as happening in single, independent units. Rather, hospitals are interconnected in a larger, nation-wide, network. Any interventions strategy or surveillance system can only deliver sensible results if it is organised at national or regional level.

Secondly, the proportion of clustered isolates reflects propensity of a clone to spread through the hospital population, as nosocomial transmission will result in clustered isolates. The isolates belonging to MC398, which include ST398, show very few clusters. These Livestock-Associated MRSA isolates are considered to spread easily within animal herds,[184] but apparently only spread through the human population on a very limited scale. Most of the MC398 isolates would therefore be the result of independent introductions from the animal reservoir. This same pattern is seen in MRSA strains described to spread in the community, MC8, albeit to a lesser extent. More often than in the hospital-associated strains, the isolates seem to be the result of introductions, in this case probably from the general population. Novel intervention strategies could take this differential transmissibility into account and differentiate patients according to the clonal type of their MRSA.

Finally, the large fraction of MUOs among the coincidental findings paints a bleak future for the Dutch “search and destroy” policy. If MRSA is increasingly spreading through the community, as seems to be the case with PVL$^+$ MC8, hospitals will have to cope with an ever-increasing amount of MRSA introductions. These unidentified introductions will lead to more transmission, because the patients without obvious risk factors admitted from the community are not isolated and screened on admission, and as a consequence lead to more MRSA of unknown origin. This may ultimately lead to the Dutch search and destroy success story to collapse on itself. The fact that the risk factors for community-acquisition of MRSA are diffuse complicates screening efforts, as no clear risk group can be identified. To break this circle of increasing transmission, more efforts have to be made to identify the source of the MRSA infections of unknown origin. We have shown that in order to be effective, these efforts require pooled epidemiological information and typing data, gathered in central databases.

Chapter 5

Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data.



R.J.F. Ypma, A.M.A. Bataille, A. Stegeman, G. Koch, J. Wallinga, W.M. van Ballegooijen

Proceedings of the Royal Society B: Biological Sciences (2012) **279**(1728), 444-450

Knowledge on the transmission tree of an epidemic can provide valuable insights into disease dynamics. The transmission tree can be reconstructed by analysing either detailed epidemiological data (e.g. contact tracing) or, if sufficient genetic diversity accumulates over the course of the epidemic, genetic data of the pathogen. We present a likelihood-based framework to integrate these two data types, estimating probabilities of infection by taking weighted averages over the set of possible transmission trees. We test the approach by applying it to temporal, geographic and genetic data on the 241 poultry farms infected in an epidemic of avian influenza A (H7N7) in the Netherlands in 2003. We show that the combined approach estimates the transmission tree with higher correctness and resolution than analyses based on genetic or epidemiological data alone. Furthermore, the estimated tree reveals the relative infectiousness of farms of different types and sizes.

Introduction

Estimating the transmission tree for an epidemic of an infectious disease can provide valuable insights; it has been used to evaluate effectiveness of intervention measures,[19,61,185,186] to quantify superspreading,[51] to identify mechanisms of transmission[187] and to study viral evolutionary patterns.[188,189]

Unfortunately, estimating the transmission tree is rarely trivial. Generally one needs detailed epidemiological data, such as contact structures, while for many epidemics only general statistics, such as case lists with time of symptom onset, are known. Furthermore, data is often missing for some cases. Luckily, another valuable source of information, genetic data on the pathogen, is becoming increasingly available. The amount of genetic diversity observed between samples taken from different cases informs us of the distance in the transmission tree between these cases.

Estimates of the transmission tree will be best when all available data is combined in one analysis; however methods to achieve this are still largely lacking.[66] One approach proposed by Cottam et al.[22] is to use genetic data to exclude certain potential transmission trees, and then evaluate the remaining possible trees with epidemiological data. Jombart et al.[190] followed a similar approach, contrasting the construction of transmission trees with the construction of phylogenetic trees.

We argue that a more consistent approach can be obtained by combining both genetic and epidemiological data in one likelihood function when reconstructing the transmission tree. This likelihood function can then be used in a Bayesian setting to sample from the space of all transmission trees, allowing the simultaneous estimation of both the tree and the parameters of the function itself. These parameters themselves can be used to describe the dynamics of the epidemic. Such an approach would be able to handle missing data, e.g. cases for which no genetic data is known.

We develop the approach to obtain a more detailed understanding of an epidemic of avian influenza A (H7N7) in different types of poultry farms in the Netherlands in 2003. This epidemic spread over a large area and infected 241 farms and 89 humans in total, with one human fatality,[191-193] even though a movement ban was promptly imposed upon confirmation and both infected and suspected farms were culled. The epidemic also spread abroad, infecting 8 farms in Belgium and 1 in Germany. This specific H7N7 strain has not been detected since, indicating the epidemic formed a dead end for virus spread.

Using only epidemiological data, previous studies showed there is a strong spatial component to spread[194] and small hobby farms are less susceptible to infection than large commercial farms.[195] However it remains unclear how the virus spread from farm to farm or how farm size and type relate to infectiousness. Previous studies for other farm animal disease such as foot-and-mouth-disease found a clear relation between the number and type of animals on a farm and its infectiousness.[185,196,197] We might expect similar differences in infectiousness for avian influenza H7N7 as well, for example because the transport structures are different for different farm types (e.g. food transports, egg rearing) and the different host species react differently to the virus. Here we quantify the infectiousness of farms of different types and sizes, using transmission trees we reconstructed from genetic, geographic and temporal data.

Methods

Data

Influenza A (H7N7) virus was detected on in total 241 farms in the Netherlands. These consisted of 205 commercial chicken farms, 14 hobby farms (defined as flocks of less than 300 animals), 19 turkey farms and 3 duck farms. The 9 farms infected abroad have not been included in this analysis because virus sequences of these outbreaks, exact location and or date of infection were not known to us and the farms are considered dead ends for this epidemic.

For all 241 farms the geographical location and the date of culling have been recorded, the date of infection has been estimated using animal mortality data[198] for all farms but the hobby farms, for which we take the estimates from Boender et al.[194] Results in this paper were found to be robust to small variations in infection dates. Data on the number of animals kept was available for 220 farms. All available data are provided online (although we are only allowed to specify the geographical location in terms of the closest town, due to privacy restrictions).

For 185 farms the RNA consensus sequence of the haemagglutinin, neuriminidase and polymerase PB2 genes was determined from a pooled sample of five infected animals[199] (GISAID accession numbers EPI_ISL_68268-68352 and EPI_ISL_82373-82472). We will refer to these farms as sequenced farms, to the remaining 56 farms as unsequenced farms.

Model

The first farm to be infected in the epidemic is taken as the index case, infected by some unknown source; all other farms are assumed to be infected by a previously infected farm through an unknown route. We make the simplifying assumption that all three types of data are independent from each other, e.g. knowing the geographical distance between a farm and the farm that infected it tells you nothing about the genetic distance between the sequences sampled at those farms or the time between infections of the two farms. Figure 1a illustrates how the approach works. For every farm B we evaluate which of the previously infected farms could have been the source. The likelihood L that a certain farm A infected B increases if A is not yet culled when B is infected, if A is located close to B, if the sequence taken from farm A is similar to that taken at farm B, and if no other viable candidates exist that could have infected farm B. When genetic information is unavailable for one of the farms, we look at the sequence of the farm that infected this farm (figure 1b).

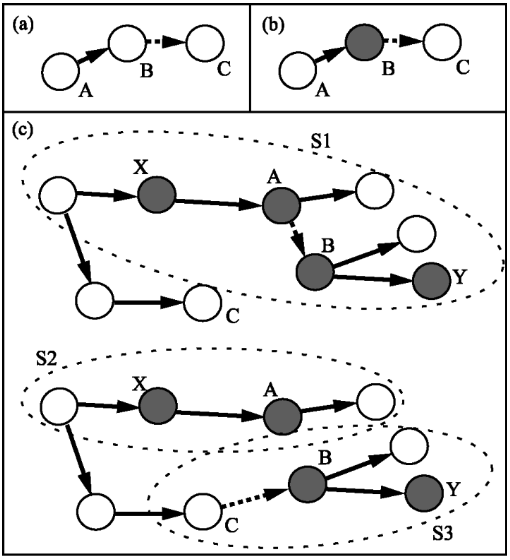


Figure 1. Examples of transmission trees. Circles denote farms, full arrows denote estimated infections, dotted arrows denote possible infections. Filled circles denote farms from which no sequence data was available. (a) When all data is available, the probability that B infected C is proportional to the likelihood $L(\delta_{BC})$ given in (1). (b) When genetic data is missing for B, we can infer this by looking at the sequence of the farm A that infected B. We then assess the likelihood for the whole subtree: A infecting B, infecting C. (c) When genetic data is missing for multiple farms, we look at the subtrees containing those farms. Here two transmission trees are possible; B is infected by either A or C. The likelihood that A infected B is proportional to the likelihood of S1, while the likelihood that C infected B is proportional to the product of the likelihoods of S2 and S3.

Likelihood without missing data

When there is no missing data, the likelihood of a transmission tree is simply the product of the likelihoods of the links it consists of. Therefore we construct a likelihood function L which gives the likelihood for our set of parameters $\boldsymbol{w} = (b, r_0, \alpha, \boldsymbol{p})$ and the event δ_{AB} that a certain farm A infected another farm B, given our data D consisting of a temporal, geographical and genetic component, $\boldsymbol{t}$, $\boldsymbol{x}$ and RNA respectively. This function consists of a product of contributions given by the temporal, geographical and genetic data.

$$L(\delta_{AB}, \mathrm{w}|D) = L_t(\delta_{AB}, b|\mathbf{t}_A, \mathbf{t}_B) L_{geo}(\delta_{AB}, r_0, \alpha|\mathrm{x}_A, \mathrm{x}_B) L_{gen}(\delta_{AB}, \mathrm{p}|\mathrm{RNA}_A, \mathrm{RNA}_B) \tag{1}$$

Time

We assume farms are infectious starting one day after infection due to a latent period,[198,200,201] and remain equally infectious until they are culled (a sensitivity analysis for the length of the latent period can be found in appendix D.3). Date of infection and culling are denoted by t^{inf} and t^{cull} respectively. Due to decay of contaminated particles, and the possibility of an infection mechanism which introduces time lag, infectiousness drops exponentially after culling with rate of decline b. This gives us the likelihood of the parameter b and farm A infecting another farm B given their temporal data:

$$L_t(\delta_{AB}, b|\mathbf{t}_A, \mathbf{t}_B) = \begin{cases} 0 & t_B^{inf} <= t_A^{inf} \\ 1 & t_A^{inf} < t_B^{inf} <= t_A^{cull} \\ e^{-b(t_B^{inf} - t_A^{cull})} & t_B^{inf} > t_A^{cull} \end{cases} \qquad (2)$$

Geography

We assume farms are more prone to infect farms close by than far away, with the likelihood contribution of infecting a farm a certain distance $|x_A\text{-}x_B|$ kilometers away given by the best-fitting distance kernel taken from Boender et al.:[194]

$$L_{geo}(\delta_{AB}, r_0, \alpha|\mathbf{x}_A, \mathbf{x}_B) = \frac{1}{1 + \left(\frac{||\mathbf{x}_A - \mathbf{x}_B||}{r_0}\right)^{\alpha}} \qquad (3)$$

We performed the analysis with a wide range of different shapes for this kernel, and found the results to be robust to the specific choice of kernel (see appendix D.2)

Genetics

We assume the sequence sampled to be the most prevalent sequence on the whole farm. Additional cloning experiments performed by Bataille et al.[199] showed this to be very probable for four out of five farms tested, where the fifth was the first farm to be infected when the epidemic spread to the south of the country, thus having an exceptionally long infectious period. We assume there is a fixed number of nucleotides N that can mutate independently of each other, and any mutations get fixed during or shortly after infection. Since many mutations will drastically lower the fitness of the virus, we set N at one third of the total number of nucleotides sequenced. We take p_{ts} and p_{tv} to be the expected number of transitions and transversions per infection respectively. We assume there is a fixed probability p_{del} of a deletion occurring during any infection. Although such a deletion could decrease N, the length of the deletions is small enough to ignore this effect. Let the number of transitions and transversions needed to go from the sequence of A to that of B be denoted as d_{ts} and

d_{tv} respectively, and let $\mathbf{1}_{\text{del}}$ be an indicator function: 1 if a deletion occurred, 0 if it did not. We then have for the likelihood contribution

$$L_{gen}(\delta_{AB}, \mathbf{p}|\text{RNA}_A, \text{RNA}_B) = \frac{\left(\frac{p_{ts}}{N}\right)^{d_{ts}}}{(1-\left(\frac{p_{ts}}{N}\right))^{d_{ts}-N}} \frac{\left(\frac{p_{tv}}{N}\right)^{d_{tv}}}{(1-\left(\frac{p_{tv}}{N}\right))^{d_{tv}-N}} p_{del}^{\mathbf{1}_{del}}(1-p_{del})^{1-\mathbf{1}_{del}} \quad (4)$$

Likelihood with missing data

If all data were available for each farm, the likelihood of any transmission tree would be the product of the likelihoods of the links it consists of. However, when data is missing for a certain farm, we have to incorporate information on the neighbouring farms in the tree to assess the likelihood. Figure 1c illustrates this concept. To make it precise, let *T* be the transmission tree to be evaluated. *T* consists of a set of links, one for each of the farms except the index case. Let *i* be a certain data type in the analysis, and let S_i be the largest partition of *T* into subsets, called subtrees, such that for each farm that misses data type *i* all links connected to that farm are in the same subtree. Then the likelihood of *T* is the product over all data types *i* of the product of the likelihoods of each of the subtrees in S_i:

$$\begin{aligned} L(T, \mathbf{w}|D) &= L_t(T, b|\mathbf{t}_T) L_{geo}(T, r_0, \alpha|\mathbf{x}_T) L_{gen}(T, \mathbf{p}|\text{RNA}_T) \\ &= \prod_{S \in S_t} L_t(S, b|\mathbf{t}_S) \prod_{S \in S_{geo}} L_{geo}(S, r_0, \alpha|\mathbf{x}_S) \prod_{S \in S_{gen}} L_{gen}(S, \mathbf{p}|\text{RNA}_S) \quad (5) \end{aligned}$$

The likelihood of a subtree is again the product of the likelihoods of its links, where for each of the farms with missing data we sum over all possible data values. Note that if all data of a certain type is known for all farms, which for the epidemic of avian influenza is the case for both temporal and geographical data, the subtrees for that data type are just the individual links.

Calculation

We construct a Monte Carlo Markov chain (MCMC) to sample from the space of all possible transmission trees and parameters $\boldsymbol{w} = (b, r_0, \alpha, p_{ts}, p_{tv}, p_{del})$, using flat priors for all parameters on the positive real numbers and for all transmission links. Subtrees containing unsequenced farms, i.e. farms without genetic data, are handled by summing over all data values consistent with the least amount of mutations needed to explain the subtree. Note that we don’t have to sum over all unsequenced farms: firstly genetic data is irrelevant for farms that infect no other farms (Y in figure 1). Secondly farms that infect only one other farm (X in figure 1) can be handled by extending (4) to allow for a farm indirectly infecting another farm where there are x transmissions in the chain:

$$L_{gen}(\delta_{AB}, \mathbf{p}|\text{RNA}_A, \text{RNA}_B) = \frac{x^{d_{ts}}\left(\frac{p_{ts}}{N}\right)^{d_{ts}}}{(1-\left(\frac{p_{ts}}{N}\right))^{d_{ts}-xN}} \frac{x^{d_{tv}}\left(\frac{p_{tv}}{N}\right)^{d_{tv}}}{(1-\left(\frac{p_{tv}}{N}\right))^{d_{tv}-xN}} p_{del}^{\mathbf{1}_{del}}(1-p_{del})^{1-\mathbf{1}_{del}} \quad (6)$$

To correctly assess the likelihood of the parameters we have to normalize the likelihood functions, i.e. divide the likelihood by the integral over the entire parameter range. Since the geographical locations of the farms are fixed we normalize (3) for each farm by dividing by the total infecting potential for that farm, i.e. the sum over all other farms of the product of (2) and (3).

We take a burn-in of 500.000 iterations, checking for convergence, and then sample ten thousand times, at every 500th iteration. Averaging over the posterior density over the space of trees gives the probability for each possible infection event. We obtain a point estimate for parameters by taking the median of their posterior densities and construct a 95% credibility interval by taking 95% of the posterior probability mass.

For each of the sampled trees we can estimate the infectiousness of each farm by dividing the number of infections caused by this farm by the length of its infectious period. Taking the average for each farm type gives us an estimate of relative infectiousness of different types of farms.

Evaluation of the estimated tree

To evaluate how much information is contained in the geographical and genetic data respectively we exclude the geographical (3) or the genetic (4) likelihood from the full likelihood function (1), rerun the analyses, and see how much our results differ from the full picture obtained by using all data. Two important properties of the estimated tree to look at are correctness and resolution; how close the tree is to the actual transmission tree and how many links can be established at or above a certain probability level respectively. Although correctness is the most important one, it is also the hardest to measure. Resolution tells us how well we can distinguish between potential source farms; if our assumptions on the disease dynamics are reasonable an increased resolution should also be indicative of an increased correctness.

To test whether the method correctly handles cases with missing data we analyse the position of the unsequenced farms in the estimated transmission tree.

Results

From our analysis we obtain an estimate of the transmission tree; for each pair of farms A and B we get the probability that A infected B. A graphical representation of the estimated tree is given in figure 2a, which shows the high-probability transmissions on a map of the region. Figures 2b and 2c do the same for the trees obtained when excluding geographical and genetic data respectively from the analysis. The decrease in estimated transmissions relative to figure 2a shows that combining all data leads to an increase in resolution. An increase in correctness is also shown: the combined approach suggests only one introduction of the disease into the southern part of the country (figure 2a), while the analysis based on genetic data predicts multiple introductions (figure 2b). However, additional cloning experiments[199] have made clear that the disease was transmitted to the south of the country only once and then spread locally.

The resolution of the estimated trees is depicted in more detail in figure 3: it shows more resolution can be contained when using the genetic than when using the geographic data, but combining these gives the highest resolution. Furthermore the figure makes clear that the 56 unsequenced farms cannot be placed in the tree with high confidence.

There is a subtle point that can be made about the placement of unsequenced farms in the estimated tree. That is, if an unsequenced farm in reality infected another farm that we did sequence, we would observe a relatively large genetic distance between the farm that infected the unsequenced farm and the farm that was infected by it. Therefore we would expect a sequenced farm that has a large genetic distance to all other farms to be actually infected by an unsequenced farm with higher probability, although in general we cannot identify which of the unsequenced farms was responsible. Indeed, in the estimated tree, we see that sequenced farms that have a large genetic distance to other farms are more likely to be infected by an unsequenced farm (figure 4).

Table 1 gives our estimates of the parameters used to describe the dynamics of the disease. The estimated value of b=0.28 shows infectiousness drops by (1-exp(-0.28))*100% ≈ 26% every day after culling. This low decay rate could be partly due to errors in the estimation of the infection dates. The values of r_0 and α give the infection pressure exerted by a farm to farms a certain distance away: the pressure exerted at 1 km is 24 times higher than that exerted at 10 km. The expected number of point mutations per transmission is $p_{ts}+p_{tv} \approx 1.4$, reflecting the high genetic diversity found in this epidemic.

Table 1. Estimates of the parameters used, with 95% credibility intervals (CI).

Parameter (units)	**Interpretation**	**Estimated value (95% CI)**
b (day^{-1})	Rate of decline of infectiousness	0.28 (0.23, 0.34)
r_0 (km)	Scale parameter of spatial kernel	2.4 (1.2, 3.7)
α	Shape parameter of spatial kernel	2.3 (1.7, 2.8)
p_{ts}	Average number of transitions	1.1 (0.88, 1.3)
p_{tv}	Average number of transversions	0.32 (0.22,0.43)
p_{del}	Probability of deletion	0.069 (0.027, 0.13)

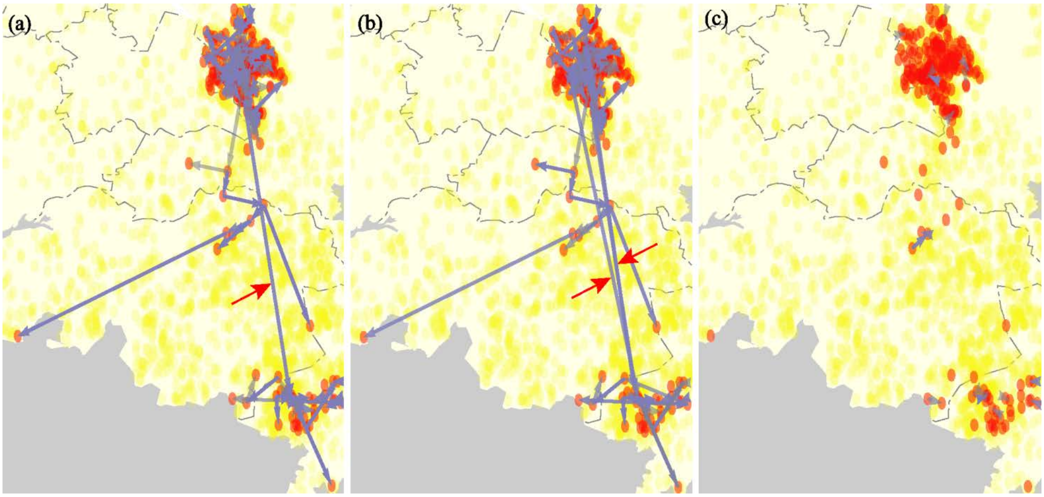


Figure 2. Infection events with posterior probability >0.5. The analysis was run separately three times using (a) temporal, genetic and geographic, (b) temporal and genetic or (c) temporal and geographic data. Red dots denote infected farms, yellow dots denote farms not infected in the epidemic. A higher opacity of arrows corresponds to a higher estimated probability. The lack of arrows in (c) tells us geographical information alone is not enough to establish transmission links. Although genetic information yields quite accurate results, we obtain more certainty for many links when also incorporating the geographical data. Furthermore in the combined analysis we can correctly show there was only one introduction into the southern province of the Netherlands rather than multiple as suggested by the genetic data (indicated by the red arrows). A zoomed in version of the northern part of the outbreak is given in appendix D figure S1.

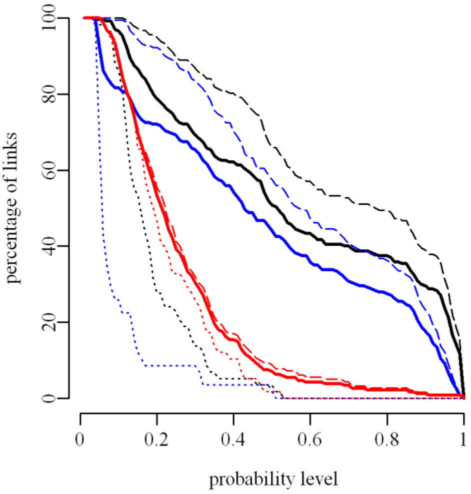


Figure 3. Resolution of the estimated trees. For each level of probability, the percentage of farms for which the farm that infected it can be estimated at or above this level is plotted. Estimates are based on temporal and (blue) genetic, (red) geographical or (black) both data. Striped and dotted lines give the same information for sequenced and unsequenced farms respectively. The graph for genetic data (blue) is much higher than that for the geographic data (red), which shows the former yields more resolution than the latter. This is confirmed by the fact that unsequenced farms are hard to place accurately in the transmission tree, resulting in a small surface under the dotted lines. However, combination of all data types results in the highest resolution, shown by the largest surface being under the black line.

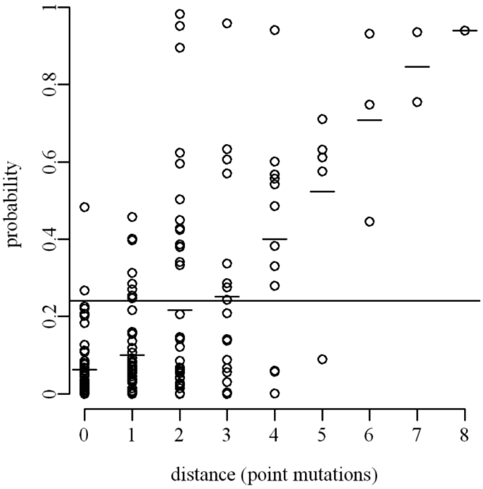


Figure 4. Scatterplot illustrating the position of unsequenced farms in the estimated tree. For each sequenced farm, we plot the minimum of the genetic distances to the sequenced farms infected earlier than this farm (horizontal axis) against the probability it was infected by an unsequenced farm (vertical axis). The small horizontal lines give the average probability per distance, the large horizontal line gives the fraction of unsequenced farms (=0.23). The increasing trend shows the placement of unsequenced farms in the estimated tree is more likely to be between farms whose genetic distance is relatively large. This is probably due to large genetic distances being indicative of farms in the actual transmission tree not being sequenced.

The transmission tree can be used to estimate the infectiousness of the different types of farms involved in the epidemic, and to look at the relation between farm size and infectiousness (figure 5). We see that hobby farms on average caused less, and turkey farms caused more infections per day than chicken farms. This last observation could be partly explained by the fact that turkey farms were among the first to be infected when the epidemic spread to the south of the country, where the infections remained undetected for some time. Although there is a relation between farm size and infectiousness, it takes a peculiar form. It seems farms with less than 1000 animals (this includes all hobby farms) are hardly infectious when compared to the larger farms, but above this boundary infectiousness does not increase with size.

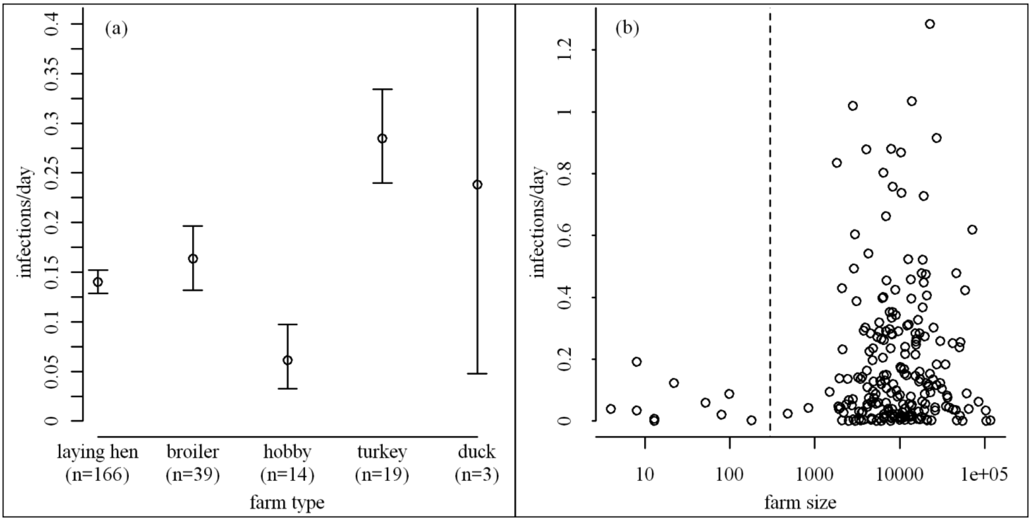


Figure 5. Estimated average infectiousness for (a) different types of farms and (b) number of animals on the farm, as measured by number of infections caused divided by the time period in days between infection and culling of the farm. All farms to the left of the dashed line in (b) are hobby farms by definition (these are farms with less than 300 animals). (a) We see hobby farms are less infectious than other types of farms, while turkey farms are more infectious than chicken farms. This can possibly be explained by the fact that many of the turkey farms in the epidemic were among the first to be infected when the disease spread to the southern part of the country, where control measures were not yet in place. (b) There is a correlation between total animals present on a farm and its infectiousness, however the exact relationship remains unclear.

Discussion

We have estimated the transmission tree for an epidemic of avian influenza by combining genetic, geographic and temporal data using one likelihood function. We have shown the increased correctness and resolution obtained by using all data types in one analysis, illustrated that the approach correctly handles missing data, and estimated the infectiousness of farms of different types and sizes.

The results we obtained on the epidemic of avian influenza fit well into previous results. Bavinck et al.[195] found hobby farms to be less susceptible, but lacked the data to assess infectiousness, which we additionally showed to be lower than that of other farms. The dichotomous relation we found between farm size and infectiousness might be explained by large farms being detected and culled before a substantial number of animals could get infected, or by a mechanism of spread which is not very sensitive to the amount of infectious particles present in a farm. Boender et al.[194] used epidemiological data on all poultry farms in the Netherlands to estimate the parameters of the distance kernel (3). Even though they did not have any genetic information available, their estimates of r_0 and α, 1.9 (1.1, 2.9) and 2.1 (1.8, 2.4), are very close to our estimates, 2.4 (1.2, 3.7) and 2.3 (1.7, 2.8). The similarity of these estimates, based on different datasets, increases our confidence in both analyses. Further research on this H7N7 epidemic should focus on how the virus spread; knowledge on which farm infected which coupled with information on for example human movement and/or historical wind direction could answer long-standing questions on what mechanisms are responsible for spread of avian influenza. Furthermore, it will be interesting to look at different predictors for infectiousness; although we lacked the data to do so, it might be very worthwhile to look at the impact of for example different contact structures or trading practices.

In modelling evolutionary drift, the usual assumption is that of a (relaxed) molecular clock, i.e. a constant rate at which new mutations fix in the population. This assumption appears valid on long timescales; however it is unclear whether this assumption holds on short timescales such as the duration of one avian influenza epidemic. We make the simplifying assumption that mutations only fix in the viral population during or shortly after infection, independent of time. The rationale behind our assumption is that the infection of a farm constitutes a population bottleneck for the virus. If infectious periods are short, these bottlenecks are exactly where the emergence of new dominant strains is to be expected, even though the particular mutations could have been present before the bottleneck. A more sophisticated evolutionary model that depends both on time and on infection events would be highly desirable, constructing and fitting one will be a challenging improvement to the analysis in this paper.

When using different types of data, a natural question is how to ‘weigh’ these different types. We argue that the most natural approach to weighing is by constructing intuitively plausible likelihood functions for each of the data types and then combining these in a straightforward matter such as in our analysis. There is then a weighing of data types implicitly done by the choice of function. For example the number of nucleotides that could mutate, N in (4), is not known exactly; increasing or decreasing this value puts more or less weight on the genetic data respectively. Plausible numbers for N range from the total amount of nucleotides seen to mutate (214) to the total amount of nucleotides sampled (5354). We found the results described in this paper to be robust to changes in N for this range.

An alternative approach to estimating transmission trees using genetic and epidemiological data that has been suggested by Cottam et al.[22] is to assess the data types sequentially; first using genetic information to exclude possible transmission trees, then assessing the likelihood of these using epidemiological data. In our terminology this is equivalent to assigning each possible transmission tree a likelihood based solely on genetic data, and discarding the trees with likelihood values below a certain level. Not only is this choice of level arbitrary, but since the remaining trees are only evaluated using epidemiological data any further information in the likelihoods that was derived from the genetic data is lost. A third approach was presented by Jombart et al.,[190] who nicely contrasted the construction of transmission trees with the construction of phylogenetic trees, making the link to the field of phylogeography. They however only look at epidemiological data when genetic sequences for multiple cases are identical, disregarding this data otherwise. Furthermore the method lacks a way to handle missing data.

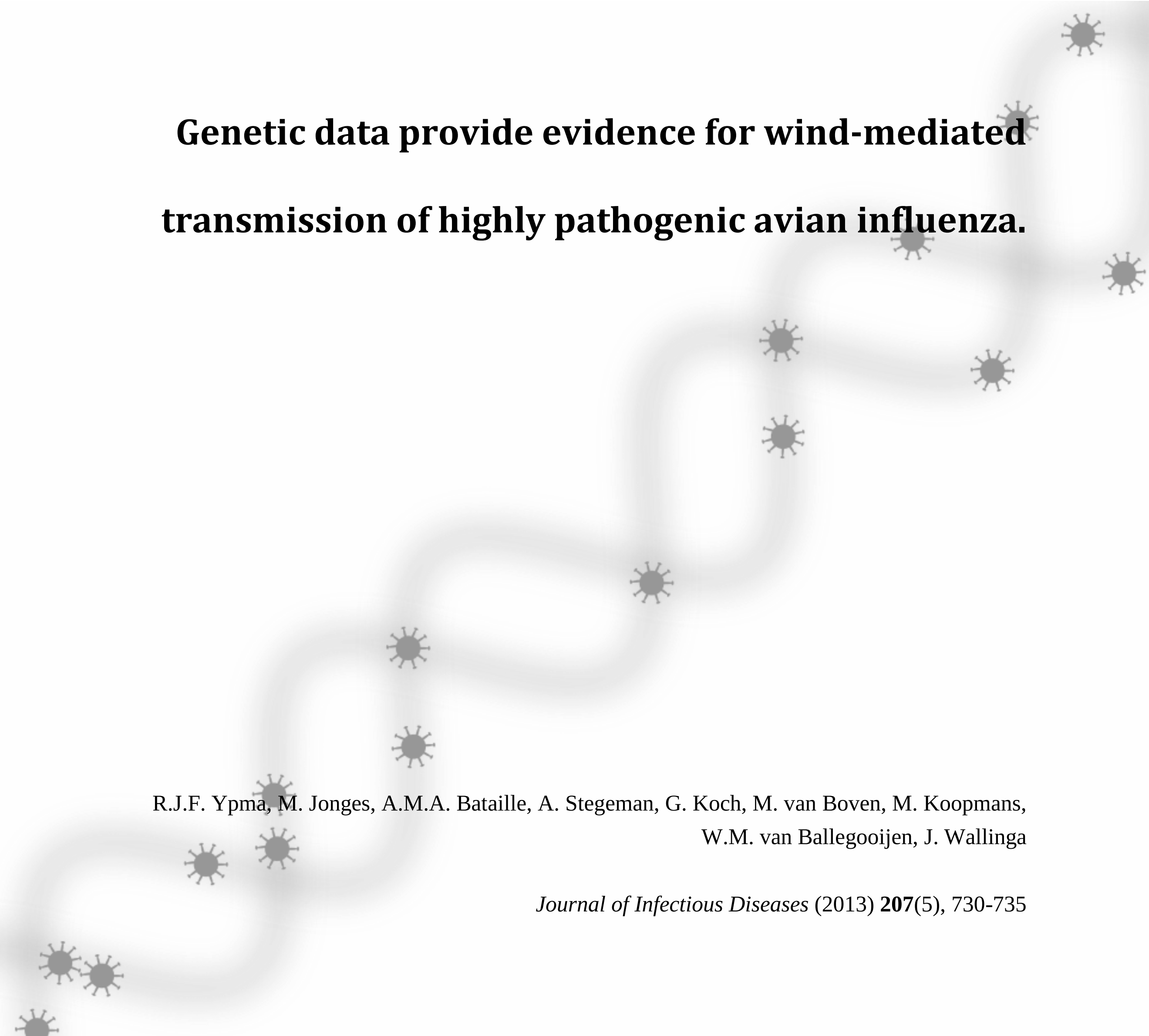
The general framework presented in this article is applicable in a wide range of settings. Although we have looked at infections between farms, the same techniques are also applicable when the infections are between individual host organisms, such as humans, animals or plants. Different types of epidemiological data such as day of symptom onset, age or area code could be handled by constructing an appropriate likelihood function.[19,139] Furthermore, the ability to handle missing data makes the method suitable for use in many practical settings. Since the information contained in genetic data increases with genetic diversity, the method is probably more suitable to analyse epidemics of rapidly mutating pathogens such as viruses; however there are no theoretical objections to apply it to epidemics of other types of pathogens, which could be worthwhile if much epidemiological data is available.

The framework presented in this paper reflects our belief that all data gathered on an epidemic can provide clues to what actually happened. How much information each data type holds is usually hard to determine, but the better we can extract this information, the clearer our picture of the epidemic becomes. Ultimately, techniques such as presented here coupled with our increasing capacity to build up large (genetic) datasets will allow for the accurate and correct estimation of transmission trees in a variety of settings, increasing our understanding of disease dynamics. The real challenge then lies in not only understanding these dynamics, but in also putting this knowledge to use in our efforts to combat diseases.

Acknowledgments

We would like to thank Annemarie Bouma, Gert-Jan Boender, Thomas Hagenaars and Michiel van Boven for critically reading an earlier version of this manuscript. We would like to thank the Associate Editor and two anonymous referees for useful comments.

Chapter 6



Genetic data provide evidence for wind-mediated transmission of highly pathogenic avian influenza.

R.J.F. Ypma, M. Jonges, A.M.A. Bataille, A. Stegeman, G. Koch, M. van Boven, M. Koopmans,
W.M. van Ballegooijen, J. Wallinga

Journal of Infectious Diseases (2013) **207**(5), 730-735

Outbreaks of highly pathogenic avian influenza in poultry can cause severe economic damage, and represent a public health threat. Development of efficient containment measures requires an understanding of how these influenza viruses are transmitted from one farm to the next. However, the actual mechanisms of inter-farm transmission are largely unknown. Dispersal of infectious material by wind has been suggested, but never demonstrated, as a possible cause of transmission between farms. Here we provide statistical evidence that the direction of spread of avian influenza A(H7N7) is correlated with the direction of wind at date of infection. We find the direction of spread by reconstructing the transmission tree for a large outbreak in the Netherlands in 2003, using detailed genetic and epidemiological data. We conservatively estimate the contribution of a possible wind-mediated mechanism to the total amount of spread during this outbreak to be around 18%.

Introduction

Avian influenza is endemic in many wild bird species, which harbour all known subtypes of influenza A viruses. The virus can be transmitted from wild birds to poultry, thereby crossing the species boundary. Although most virus strains cause no or few clinical symptoms in poultry, highly pathogenic (HPAI) variants can arise through mutation.[202-204] These highly virulent strains, the most notorious of which is HPAI H5N1, can cause large outbreaks with high mortality, posing a major (economic) threat to poultry farming around the globe. The virus can cross over to human hosts, potentially resulting in severe disease or even death.[205] Therefore, HPAI is considered to be a serious public health threat.[206,207]

There have been several large outbreaks of avian influenza in Western countries, involving clusters of large commercial poultry farms. Due to the ease with which the disease seems to spread between farms, the high mortality rates among poultry, and the public health threat posed by an outbreak,[208] rigorous control measures have to be implemented. These typically consist of a complete transport ban of poultry and depopulation of all farms that either have infected animals or are at risk of infection, resulting in substantial economic losses.

Poultry farms emit large quantities of particulate matter,[209,210] which could be driven by wind to transport viable virus from an infected to an uninfected farm.[211] However, opinions differ widely on whether this actually causes new infections during an outbreak; the mechanism has never been demonstrated conclusively.[211-214] Humans, trucks or wild birds could also act as a vector, carrying the virus from one farm to the next.[215] Knowledge of the actual transmission mechanisms and their relative importance could lead to more efficient and more effective control strategies. For example, spread by humans could be reduced by stricter enforcements of biosecurity procedures, whereas wind-mediated spread could be combated by adjusted ventilation systems. Insight as to the mechanism of spread would also lead to more precise estimates of which farms are at risk of infection. This knowledge is highly valuable during an outbreak, for example when planning the order in which to cull farms.[216]

Here, we use detailed genetic and epidemiological data from an outbreak of HPAI A(H7N7) in the Netherlands in 2003 to test the hypothesis that wind aided in transmission of the pathogen. In this outbreak 241 poultry farms were infected (confirmed by virus isolation), 30 million birds were culled, and there was one human fatality.[191,192] For 231 of the infected farms isolated virus RNA has been sequenced.[199,208] These unique genetic data, in combination with time of infection and time of culling, allow us to reconstruct which farm infected which. To test for the role of wind, we compare the direction of these individual farm-to-farm transmission events to the wind direction at the date of infection, accounting for any bias induced by the geography of the farm locations. We conclude by giving an estimate for the percentage of infections that can be attributed to wind-mediated transmission.

Materials and methods

There were 5360 poultry farms in the Netherlands in 2003, for all of which geographical information is available. For 1531 farms the flocks were culled, for all of these the date of culling is known. For 227 of the 241 infected farms the date of infection has been estimated, based on mortality data.[198] The remaining 14 farms are hobby farms, defined as farms with less than 300 animals, for which no mortality data are available. For these we use the infection date as estimated by Boender et al.[194] The HA, NA and PB2 genes of viral samples from 231 farms have previously been sequenced.[199,208] Sequence data can be found in the GISAID database under accession numbers EPI_ISL_68268-68352, EPI_ISL_82373-82472 and EPI_ISL_83984-84031. Available meteorological data include wind speed and direction (with a ten degree precision) for every hour of every day of the outbreak, measured at five weather stations close to the infected farms (figure 1). These data are available from the Royal Dutch Meteorological Institute at www.knmi.nl.

Estimation of transmission events

To estimate which farm infected which, we used the genetic and temporal data on the infected farms, following the method described by Ypma et al.[64] We describe the likelihood of a possible transmission tree given the data, by arguing that the tree is more likely if the source farms are more infectious at the putative dates of infection, and if the total number of mutations needed to explain the genetic data is lower, using a simple substitution model which differentiates between transitions and transversions.[64,199] We then sampled from the space of all transmission trees using a Markov chain Monte Carlo approach, and obtained the probability of a certain transmission event by the proportion of sampled trees that includes this event. We denote transmissions with a posterior probability of at least 0.9 as observed transmissions; results for different cut-off values are similar (appendix E.5).

Correlation of wind and transmission directions

To measure whether observed transmissions are in the same direction as the wind, we compared the direction of transmissions with the wind direction. We took the vector average wind direction as measured at the station closest to the infecting farm at the date of infection. To check for a correlation between the wind direction and direction of transmission, we calculated the circular correlation coefficient,[217] and compared the value of this coefficient with values obtained under the null hypothesis that wind direction and direction of transmission are independent and uniformly distributed over all directions.

A correlation found between wind and transmission could arise as an artefact of the geographical location of poultry farms. For instance, if the index farm of the outbreak lay in the west of an area dense with poultry farms, and the prevailing wind in this region was a western wind, we could get a correlation even in the absence of any causal relationship. To construct the correct null model for the relation between direction of transmissions and direction of wind when wind plays no role, we used simulations. We used the coordinates of poultry farms corresponding to the Netherlands in 2003, and infected the farm corresponding to the index farm in the real outbreak. Every farm infected in the 10 ten days of the simulation was culled (and thus removed from the simulation) 10 days after transmission, as were all farms in a 1-km region around it 2 days later. Farms infected later than 10 days into the simulation were culled after 7 days, ring culling again following 2 days later. The probability of infecting another farm decreased with distance, as estimated previously.[64,194] Only simulations that led to a total number of infections between 200 and 280 were used for the subsequent analysis. From each of these simulated outbreaks we randomly sampled as many transmission links as there were observed transmissions in the actual data. We used these sampled transmissions from 1000 simulations to calculate correlation with wind direction. Furthermore, we compared the simulation results to the actual data using two additional statistics that are relevant for a possible wind-mediated mechanism of spread. First, we looked at the angle between wind and transmission by taking the cumulative distribution function of $P(x)$, the probability that the angle between transmission and wind direction is x, and performed a one-sided Kolmogorov-Smirnov test to test if this distribution was significantly larger for the actual dataset. Second, we compared the average number of hours that wind coincided with the direction of a transmission, which could be interpreted as the time window for transmission due to wind.

Quantification of wind contribution to spread

We quantify what proportion of the transmission events could be attributed to a wind-related mechanism of transmission, assuming that such a mechanism exists. Here we use the term "wind-mediated transmission" to denote transmission events that can be attributed to a wind-related mechanism of transmission if such a mechanism would exist.

An observed transmission was defined to be in the direction of the wind when the average wind direction was in the direction of spread, up to five degrees, for at least one hour on the date of infection. The proportion of transmissions mediated by wind can then be estimated by comparing the fraction of observed transmission events that are in the direction of wind, to the fraction of transmissions expected to be in the direction of the wind if wind played no role. This expected fraction can be found from our simulations. The observed fraction will be a combination of transmissions actually due to wind, and some that were in the same direction as wind due to chance. From this we estimate the percentage of transmissions mediated by wind (see appendix E.4.1 for details).

To test the robustness of this simple estimation procedure, we also performed a more detailed analysis that uses all available data, but needs additional assumptions. In this approach, uncertainty in estimated infection dates was accounted for by using a prior of several days centred on the estimated date. The probability per day of infecting a farm at a certain distance can be found by looking at the ratio of the number of farms at that distance that were and could have been infected.[194] Here, we take this distance-related probability to consist of a ‘wind-related part’ W, assuming a Gaussian plume model for wind-related spread,[218] and an 'unknown mechanism part' *U*. The percentage of transmissions related to wind can then be estimated as the *W/(U+W)* (see appendix E.4.2 for details).

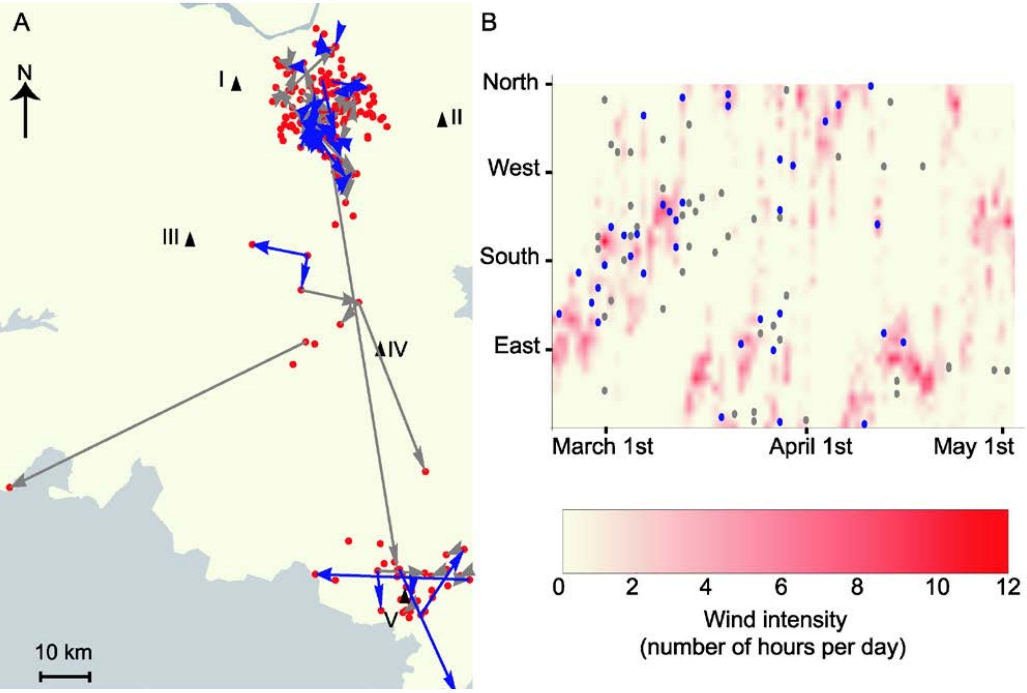


Figure 1. Observed transmission events of avian influenza A(H7N7) between farms in the Netherlands, 2003. Transmissions that coincide with wind direction (measured at the closest meteorological station) are in blue. Observed transmissions are defined as transmissions estimated at a posterior probability of at least 0.9. (A) Transmission events on a map of the region. Dots denote infected poultry farms, triangles denote the five closest meteorological measuring stations. (B) Density plot of wind direction (measured at station IV) against time, with red regions indicating wind is predominantly in one direction. Dots denote estimated time and direction of transmissions. Wind direction and direction of transmissions appear to be correlated. For example at the start of the outbreak, around March 1st, wind direction is predominantly south east (red region); in this time period most transmission events are also in this direction (dots).

Results

For all farms infected in the avian influenza A(H7N7) outbreak, we identified the most probable infecting farm using infection date, culling date, and viral RNA sequence data of the HA, NA, and PB2 genes. For 83 farms we could identify a single infector farm with a probability of at least 0.9; we call these 83 pairs of infected and infector farms the 'observed transmissions'. Figure 1A shows the location of the farms, the distance over which the observed transmissions occurred, and the direction of observed transmissions on a map of the region. Most transmissions occurred over short distances in a central high-density farm area. Figure 1B shows the wind direction over the course of the outbreak, together with the direction of the observed transmissions.

The circular correlation coefficient between direction of observed transmissions and wind direction was 0.051, significantly higher than expected when directions were uniform and independent (p=0.01) (appendix E figure S1). The circular correlation coefficient was also significantly higher than expected based on geography of farm locations, as we found the 0.975 quantile for the circular correlation coefficient under the simulations to be 0.043. We further found that the angles between the direction of observed transmissions and the vector average wind direction at the date of infection are significantly smaller than the angles between the direction of simulated transmissions and wind direction (one-sided Kolmogorov-Smirnov test p<0.01) (figure 2). Likewise, the average number of hours for which wind is in the same direction as the observed transmission at the date of infection is significantly higher for the actual dataset than for the simulations. We therefore conclude that the correlation between direction of transmission and direction of wind is higher than can be explained by chance and location of farms.

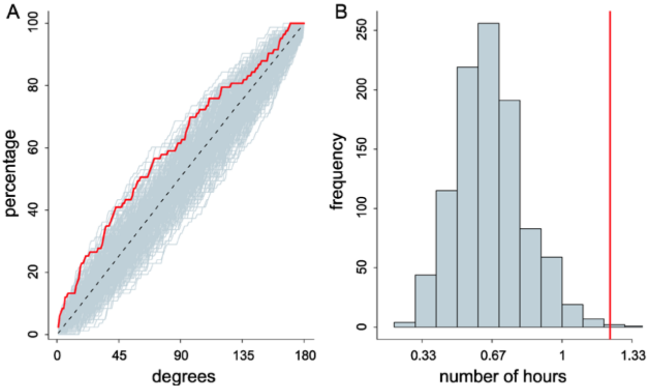


Figure 2. Comparison of observed transmissions with simulations in which transmission was unrelated to wind. (A) The cumulative percentage of transmissions having an angle with the vector average wind direction of a certain degree, for observed transmission events (red) and 1000 simulations (grey). The distribution for the dataset lies significantly outside the range generated by the null hypothesis given by the simulations (one-sided Kolmogorov-Smirnov test, p<0.01). (B) The average number of hours for which wind direction was equal to direction of spread at the date of infection. Since there are 24 hours in a day, and hourly average wind directions are available with 10-degree precision, we expect the mean number of hours when wind plays no role to be 24/(360/10)=2/3. The histogram gives the values for the simulations; the red line gives the value for the observed transmission events. Both graphs show that the correlation between wind direction and direction of spread cannot be explained by chance or farm geography alone.

The strong positive correlation between wind direction and direction of influenza transmission suggests that a substantial proportion of the transmission events are mediated by wind. To estimate this proportion of transmissions mediated by wind we compared the percentage of observed transmissions in the direction of the wind with the percentage expected by chance. In our analysis of the actual data, 34% of the posterior probability was on transmissions in the direction of the wind. We assumed this 34% was made up of transmissions that were and transmission that were not mediated by wind. The first would all be observed to be in the direction of the wind, while of the latter only a percentage would be observed to be in the direction of the wind by chance. From our simulations we know that on average 24% of transmissions not related to wind will be in the direction of the wind by chance. Using these numbers and maximizing a likelihood equation (see appendix E.2), we estimate the percentage of transmissions caused by wind at 18% (95% confidence interval (CI): 6.3, 30). To test for robustness, we also performed a more detailed analysis, which assumes a Gaussian plume model for wind spread, puts a prior on the infection dates, and takes uninfected farms into account. From this model we obtain an estimate of the percentage of transmissions of 20% (95% CI: 9.8, 29), consistent with the first analysis (figure 3).

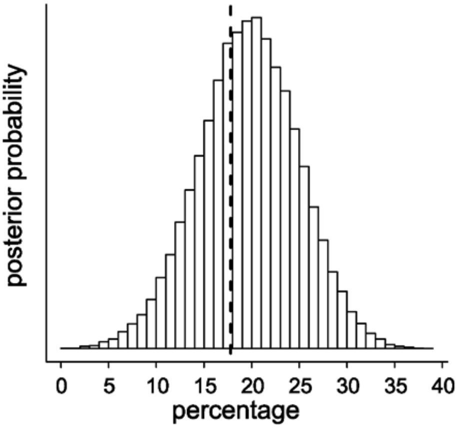


Figure 3. Estimation of the percentage of transmissions mediated by wind. Comparing the fraction of observed transmissions in the direction of wind to the fraction expected when wind plays no role yields an estimate of 18% (dashed line). Using a mechanistic model, which assumes a Gaussian plume model for wind-related spread, yields an estimate of 20%. The full posterior distribution is given by the bars.

Discussion

We have shown that for the outbreak of HPAI A(H7N7) in the Netherlands in 2003, inter-farm transmissions are more often in the direction of the wind than can be explained by chance and coordinates of poultry farms. Based on the proportion of estimated transmissions for which wind direction was observed to be in the same direction of spread, we estimated the percentage of transmission related to wind at 18%.

Wind-related spread of avian influenza has direct consequences for containment efforts. Farms emit vast quantities of particulate matter,[209] which could well carry viable virus.[214] Several systems, such as air scrubbers, water or oil sprinkling, changes in ventilation rate, and ionization systems have been shown to reduce dust concentrations[209,219] and could be an efficient way to stop infectious particles from getting in or out. Alternative wind-related mechanisms cannot be excluded on the basis of our analysis. Wild birds or insects acting as vectors for the disease,[220] flying preferentially in the direction of the wind,[221,222] would explain our observations as well, but call for different control strategies. Furthermore, culling strategies may take into account the role of wind. First, care should be taken to ensure contaminated material does not get into the environment during culling activity. Second, wind direction should be taken into account when estimating the risk of infection for farms; the most efficient culling order will first target those farms that are at higher risk of infection and, when infected, will pose higher risks to other farms themselves.[216] Thus increased knowledge of which farms are at risk, provided by current and forecasted wind direction, allows for a more efficient culling strategy.

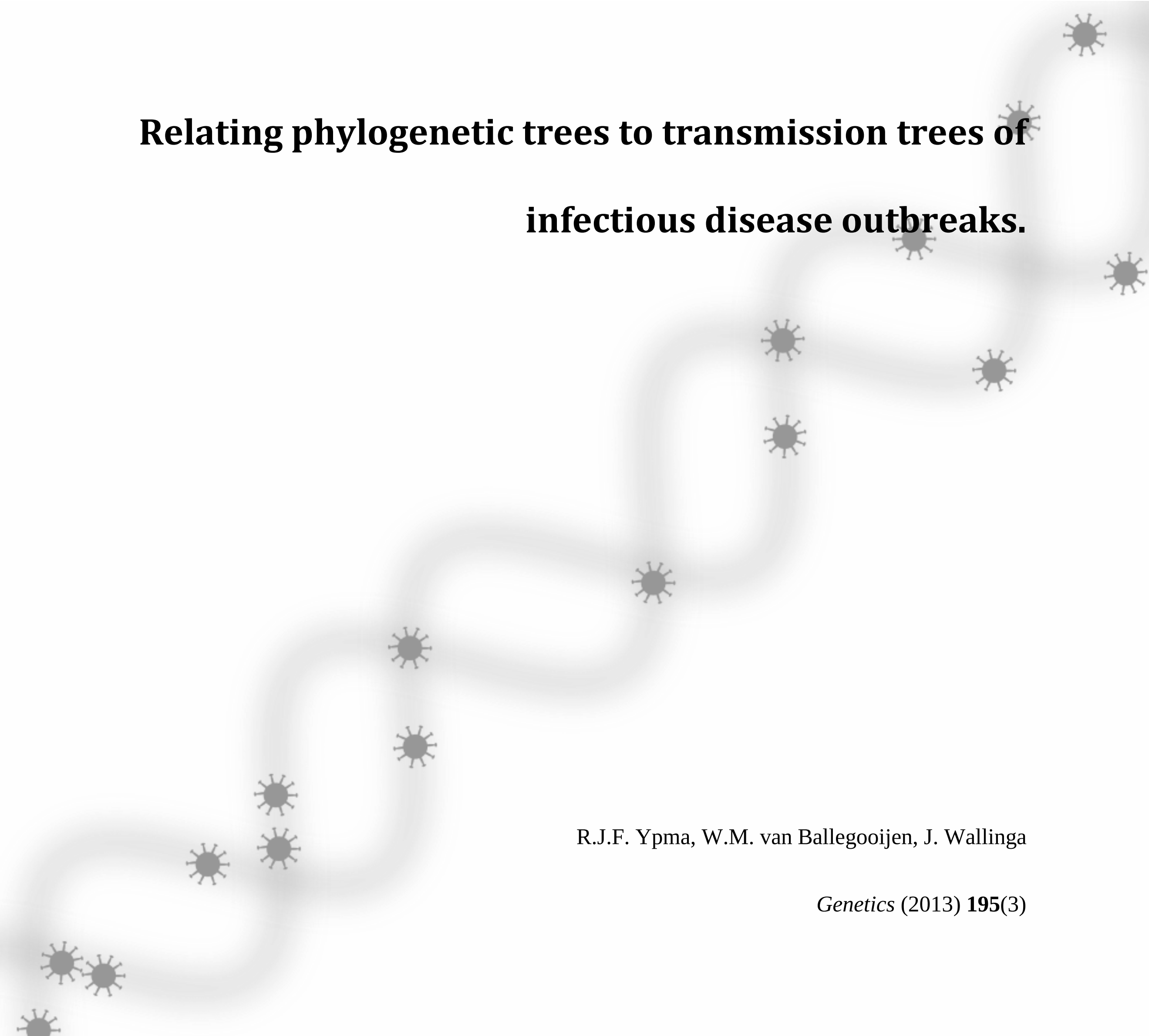
There are several sources of error in our estimation of the transmission tree. The assumed constant infectiousness of farms over the course of infection, the substitution model and the assumption of independence between mutations and time are simplifications. Furthermore, there is uncertainty in the estimation of the infection dates, and there may be errors in the geographical and genetic data. These limitations will lead to errors in the inference of transmission events. Counter intuitively, these limitations only strengthen our conclusion. If wind-mediated transmission played only a minor role, there is a negligible probability that the cumulated small errors in the data and in the inferential procedures could have produced the observed strong positive correlation between wind direction and direction of transmission. The main reason for this is that the transmission tree was reconstructed without using the meteorological data; only afterwards were transmissions compared to wind directions. It is much more likely that the cumulated errors would reduce any existing strong positive correlation. Therefore our conclusion of wind-mediated spread holds in the presence of small errors and the value of 18% should serve as a lower bound for the actual percentage of transmissions related to wind.

The type of analysis presented here may have potential to identify and quantify transmission mechanisms for other farm animal diseases. However, the resolution we obtained here, tracking individual transmissions, could only be achieved through the high percentage of farms sampled and high genetic diversity found. This resolution was necessary; an analysis looking solely at prevailing wind direction and farm coordinates would not have shown any significant effects. We therefore argue efforts should be made to gather genetic data for outbreaks of other infectious diseases as well. We do however note that usefulness of such data will depend on the genetic diversity that accumulates over the course of the outbreak, which might be lower for other pathogens, and depends on the methods used.

Identifying the mechanisms responsible for the transmission of livestock disease between farms is challenging: first, because data have to be collected during the outbreak, when the first priority will be control rather than research; second, because there are probably several different mechanisms at play, making their identification troublesome. The key to identifying transmission mechanisms is the reconstruction of detailed transmission networks, made possible by the joint analysis of detailed genetic and epidemiological data.

Chapter 7

Relating phylogenetic trees to transmission trees of infectious disease outbreaks.



R.J.F. Ypma, W.M. van Ballegooijen, J. Wallinga

Genetics (2013) **195**(3)

Transmission events are the fundamental building blocks of the dynamics of any infectious disease. Much about the epidemiology of a disease can be learned when these individual transmission events are known or can be estimated. Such estimations are difficult and generally only feasible when detailed epidemiological data are available. The genealogy estimated from genetic sequences of sampled pathogens is another rich source of information on transmission history. Optimal inference of transmission events calls for the combination of genetic data and epidemiological data into one joint analysis. A key difficulty is that the transmission tree, which describes the transmission events between infected hosts, differs from the phylogenetic tree, which describes the ancestral relationships between pathogens sampled from these hosts. The trees differ both in timing of the internal nodes and in topology. These differences become more pronounced when a higher fraction of infected hosts is sampled. We show how the phylogenetic tree of sampled pathogens is related to the transmission tree of an outbreak of an infectious disease, by the within-host dynamics of pathogens. We provide a statistical framework to infer key epidemiological and mutational parameters by simultaneously estimating the phylogenetic tree and the transmission tree. We test the approach using simulations, and illustrate its use on an outbreak of foot-and-mouth disease. The approach unifies existing methods in the emerging field of phylodynamics with transmission tree reconstruction methods that are used in infectious disease epidemiology.

Introduction

Estimating who infected whom for an outbreak of an infectious disease can provide valuable insights. Estimated transmission trees have been used to evaluate effectiveness of intervention measures,[19,61,185,196] to quantify superspreading,[51] to estimate key parameters[62,223,224] and to identify mechanisms of transmission.[187,225] Transmission trees can be statistically reconstructed using epidemiological data from outbreak investigations, such as time of symptom onset, geographical location and social ties; these data generally have to be very detailed to allow for accurate reconstructions.

For many pathogens, in particular RNA viruses, evolutionary processes occur on the same timescale as epidemiological processes.[66,226] This makes it possible to draw conclusions about epidemiology from genetic analysis. The field that infers epidemiological characteristics from genetic sequences by simultaneously considering host dynamics and pathogen genetics has been dubbed ‘phylodynamics’.[227] In practical applications researchers have considered a specific epidemiological model dependent on the phylogenetic tree inferred from sequence data, simultaneously estimating mutational and epidemiological parameters. This allowed them to answer questions on relative population sizes and dates of introduction of pathogens. Initially the epidemiological models used were classical models from population genetics such as the Wright-Fisher model.[228] Recently more realistic epidemiological models such as the SIR model[91,229] and birth-death model[112] have been suggested. In these methods, the sampled hosts are thought of as the leaves of the phylogenetic tree, while the internal nodes coincide with transmission times to unobserved hosts. The phylogenetic tree is thus equated with the partially observed transmission tree.

Although the transmission tree and the phylogenetic tree of an outbreak may appear as two incarnations of the same tree, they are in fact different in interpretation and in local characteristics. The phylogenetic tree represents the clonal ancestry of sampled pathogens; its leaves are sampled pathogens, its internal nodes are most recent common ancestors of the sampled and transmitted pathogens (figure 1).[66] As a pair of lineages corresponding to two transmitted pathogens can coalesce together before coalescing to the lineage sampled from the infecting host, the topology of the two trees need not be the same (figure 1A). The difference between phylogenetic trees and transmission trees is closely related to the difference between phylogenetic trees and species trees; in the latter context this phenomenon is known as ‘incomplete lineage sorting’.[106,230] While the timing of nodes in the transmission tree corresponds to transmission times, the timing of internal nodes of the phylogenetic tree corresponds to coalescent events which take place prior to transmission. The absolute difference in branch length between the two trees depends on the epidemiological generation interval and within-host dynamics. As the branch length of the phylogenetic tree decreases with sampling rate of hosts, the relative difference between the partially observed transmission tree and the phylogenetic tree is largest when the sampling rate is high (figure 1).

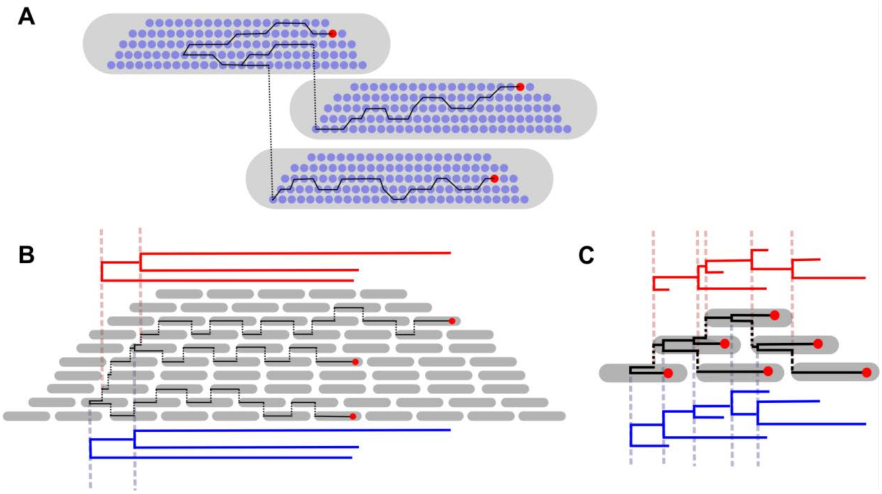


Figure 1. Schematic for viral dynamics. In all panels, time progresses from left to right. Hosts are depicted as grey pods, virus particles as blue dots and sampled virus particles as red dots. (A) The timing of coalescence of viral lineages depends on within-host viral dynamics. Virus (blue) numbers within hosts (grey) rapidly increase at onset of infection and decrease near the end of the infection, influencing coalescent rates. A possible ancestry between sampled viruses (red) is given in black. Although the initial host infects the latter two, the sampled viruses from these latter two are more closely related, as they coalesce with each other before coalescing with the virus sampled from the initial host. (B) When viruses are sampled from only a few hosts in a large outbreak, the timing of the coalescence of the sampled viruses is nearly identical to the timing of transmission immediately following the coalescence. The timing of coalescent events is mainly governed by inter-host infection dynamics, and the phylogenetic tree derived from the sequences (blue) is very similar to the one derived when internal node times are equated with transmission times (red). (C) When viruses are sampled from all hosts in an outbreak, coalescent times and transmission times are very different. The phylogenetic tree derived when approximating coalescent times by transmission times (red) is very different from the actual phylogenetic tree (blue).

Recently, methods have been proposed that focus on including genetic information in transmission tree reconstructions.[64,65] However, these approaches either ignore the phylogenetic tree, or they assume internal nodes of the phylogenetic trees coincide with transmission events. As these approaches require a high sampling fraction, the within-host genetic diversity can contribute to the genetic diversity observed between the sampled sequences. Because this contribution is ignored in the analyses, it can lead to incorrect inference of the transmission tree and biased estimates of parameters.

Here, we present a consistent way to use pathogen genetic sequence data in transmission tree reconstruction, by simultaneously estimating the phylogenetic tree of the pathogens. This requires inclusion of a model of within-host pathogen dynamics. In this paper we will describe the likelihood framework for such a joint estimation of the transmission tree and the phylogenetic tree. We investigate the performance and robustness of the methodology using simulations of an influenza outbreak in a confined setting. In these simulations, we also investigate the sensitivity of the outcome to unsampled or unobserved hosts. Finally, we illustrate the use of the method by applying it to a previously published dataset on an outbreak of foot-and-mouth disease (FMD).

Methods

Joint likelihood of the transmission tree, phylogenetic tree and within-host dynamics

We focus on outbreaks for which nearly all cases are observed, host characteristics such as time of symptom onset are known and sequences are obtained from pathogens sampled from a proportion of the hosts. The transmission events are not observed. From these data we will simultaneously estimate both the transmission tree and the phylogenetic tree, by writing down the likelihood for any pair of these trees. Throughout, we assume the first infected host, or index case, introduced infection, and all other infected hosts are infected by another host in this outbreak. We assume every host is infected at most once.

Define a transmission tree *T* as the set of all transmissions between infected hosts, including transmission times.[19] The transmission times could be observed or unknown, in which case we will estimate them simultaneously. The phylogenetic tree *P* is the usual dichotomous tree, with timed internal nodes. The function *W(t,h)* gives a measure of within-host genetic diversity, as the product of the pathogen generation time and the within-host effective pathogen population size in host *h* at time *t*. *W(t,h)* can be thought of as being proportional to the total number of virus particles in host *h* at time *t*. Let $\boldsymbol{\theta}$ be the epidemiological parameters and $\boldsymbol{\mu}$ the mutational parameters, and let the data *D* consist of genetic sequences D_G and epidemiological data D_E.

The probability of the transmission tree *T*, phylogenetic tree *P*, within-host dynamics *W* and parameters $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ can be found by first applying first Bayes' theorem, and then the chain rule of probability:

$$\begin{aligned} p(T,\theta,W,P,\mu|D_E,D_G) &\propto p(D_E,D_G|T,\theta,W,P,\mu)\pi(T,\theta,W,P,\mu) \\ &= p(D_E|T,\theta,W,P,\mu)p(D_G|D_E,T,\theta,W,P,\mu)p(P|T,\theta,W,\mu)\pi(T,\theta,W,\mu) \end{aligned}$$

where π denotes the prior probability. We can further simplify this equation by using conditional dependencies. In particular, if we know the full transmission tree, epidemiological parameters and within-host dynamics, the phylogenetic tree and mutational parameters give no further information on the probability of the epidemiological data:

$$p(D_E|T,\theta,W,P,\mu) = p(D_E|T,\theta,W)$$

Likewise, when the ancestry of sampled pathogens and the mutational parameters are known, the epidemiological data and parameters give no further information on the sequence data:

$$p(D_G|D_E,T,\theta,W,P,\mu) = p(D_G|P,\mu)$$

Also, the mutational and epidemiological parameters give no further information on the ancestry when the transmission tree and within-host dynamics are known (we assume there is no information on selection pressures):

$$p(P|T,\theta,W,\mu) = p(P|T,W)$$

We thus get

$$p(T,\theta,W,P,\mu|D_E,D_G) \propto p(D_E|T,\theta,W)p(D_G|P,\mu)p(P|T,W)\pi(T,\theta,W,\mu) \qquad (1)$$

The amount of prior information will differ per application; prior information on parameters and within-host dynamics might be available due to previous studies, prior information on the transmission tree might be available through contact tracing.

We point out the similarity of equation (1) to previous methodologies. The first term on the right hand side is, up to the inclusion of the within-host model *W*, identical to the likelihood equations found in transmission tree reconstruction methods.[19] The second and third terms together resemble the likelihood equations found in many phylodynamic approaches,[21,66] with the difference that the epidemiological model in such approaches is replaced here by the combination of the transmission tree and the within-host model.

Numerical implementation: sampling from the probability distribution

Equation (1) defines (up to a constant) a probability distribution on the space of transmission trees, phylogenetic trees and parameters. We can sample from this distribution using Markov Chain Monte Carlo (MCMC) methods. The initial state consists of a transmission tree, phylogenetic and parameters with probability larger than 0. We construct the initial state as follows. We first ensure every host has a time of infection, by assigning one if the time was unobserved. We then construct a transmission tree by assigning an infector to all infected hosts, apart from the host infected first. The infecting hosts are randomly taken from the set of hosts that are infectious at the time of infection of the infected host. We construct a phylogenetic tree that is consistent with this tree, and take initial values for the parameters from their prior distributions.

In each iteration of the MCMC, the trees and parameters are updated. An iteration consists of five different steps:

- for a host infected in the outbreak, pick a new infector. In this step we also alter the phylogenetic tree, to make sure it’s compatible with the proposed transmission tree;
- consider a new root for the transmission tree, switching infection times between the current and proposed root;
- update infection times;

- per host, update the phylogenetic tree contained in that host;
- update parameter values.

See appendix F.3 for technical details. Below, we illustrate this general concept using simulated and real data.

Testing the approach on simulated datasets

To illustrate the method and assess robustness to missing data, we apply it to simulated data on an influenza outbreak in a confined school-based population of 200 individuals. Such confined outbreaks are amenable to analysis as, due to the low number of cases, individual infections can be tracked using only epidemiological data.[63,224]

Each simulation starts with one infected and 199 susceptible individuals, half of them children; only outbreaks with at least 50 cases were used for further analysis. We use a latent period of two days, and a gamma distributed infectious period with a mean of three days and a variance of two. During their infectious period, individuals exert an equal and constant force of infection on all susceptible individuals, i.e. we assume homogeneous mixing, such that the expected number of infections caused by an infectious individual is 0 for adults and 2 for children at the start of the outbreak. We assume that symptom onset coincides with the start of the infectious period. See appendix F.1 for details.

We take a simple analytically tractable function, a quartic kernel, for the product of pathogen generation time and effective population size:

$$W(t,h) = 1 + 1000(1 - (\frac{2(t-t_h)}{T_h} - 1)^2)^2$$

where T_h is the time between infection and recovery of host h, and (t-t_h) is the time since infection t_h of h. Note that the model assumes an effective size of 1 at transmission, ensuring only one strain can infect a host. We take a substitution rate of 0.003 substitutions/site/year. [231]

For each case, we record the time of symptom onset, the recovery time, a sequence sampled one day after symptom onset, and whether the case is a child or adult. These data are then used for inference, as described below.

Inference

We estimate the transmission tree and parameters $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ using the proposed likelihood equation (1). To this end, we need to specify the three components of the likelihood.

Likelihood component for the transmission tree

We assume the incubation period, defined as the time between infection and developing symptoms, is gamma distributed with mean 2 and variance 1. The likelihood of the transmission tree then becomes the product over all infected hosts of the infectiousness of the infecting host relative to the total infectiousness of all hosts times the probability distribution *s* for the length of the incubation period:

$$p(D_E|T,\theta,W) = \prod_{x\in H-} s(o_x - t_x)\frac{I_{t_x}(v(x)|\theta, D_E)}{\sum_{y\in H} I_{t_x}(y|\theta, D_E)}$$

where *H* is the set of all infected hosts, H^- is the set of all infected hosts minus the index case, t_x and o_x are the start of the infection and the start of the infectious period of host *x*, *v(x)* is the infector of host *x*, and $I_t(a|\theta,D_E)$ is the infectiousness of host *a* at time *t*. Here we take infectiousness to be the rate at which a host infects any other host. The infectiousness $I_t(a|\theta,D_E) = \theta$ if $t>s_a$ and *a* has not recovered at time *t*, 0 otherwise. Note that the infectiousness does not depend on the infected host, as we assume homogeneous mixing and we do not want to a priori assume a difference in infectiousness between adults and children.

Likelihood component for the phylogenetic tree

The likelihood of the phylogenetic tree can be obtained from coalescent theory for haploid organisms. Going backwards in time, two events can take place that alter the number of lineages of the phylogenetic tree contained within one host. First, the number can decrease by one due to a coalescent event. Second, the number can increase by one due to an incoming lineage: a transmission event from this host to another, akin to a new sample in standard coalescent models. The first case is described by the likelihood that the lineages did not coalesce for a time, and finally coalesced, the second case is described by the likelihood that the lineages did not coalesce. Using forward time we get

$$p(P|T,W) = \prod_{x\in H} \prod_{[\tau_1,\tau_2]\in C_x} W(\tau_1, x)^{-\mathbb{1}_{coal}} e^{-\binom{n_\tau}{2}\int\limits_{\tau_1}^{\tau_2} \frac{1}{W(t,x)}dt}$$

where *H* is the set of all infected hosts and C_x is the set of all intervals $\tau = [\tau_1, \tau_2]$ where the number n_τ of viral lineages within host *x* with sampled offspring is constant and larger than 1. The indicator function $\mathbb{1}_{coal}$ is 1 if the interval starts with a coalescent, 0 otherwise. Here, we assume *W* to be fully known.

Likelihood component for the mutational parameters

We take the simplest feasible substitution model; all mutations are equally likely and happen with rate μ. We therefore get

$$p(D_G|P,\mu) = \prod_{bases} \sum_{\{A,C,G,T\}^N} \prod_{edges} (1-e^{-\mu t})^{\mathbb{1}_{mut}} + (e^{-\mu t})^{(1-\mathbb{1}_{mut})}$$

where the first product is over all base pairs, the sum is over all possible assignments of each of the nucleotides to each of the *N* internal nodes, the second product is over all branches of the phylogenetic tree, *t* denotes branch length, and the indicator denotes whether a mutation occurred on that branch. We omit the Jukes-Cantor correction because time-scales are very short and selection pressures absent, and hence the probability of multiple mutations at the same locus is negligible (see appendix F.1.1); including other, perhaps more realistic substitution models would be straightforward.

Evaluating performance

We examine how well the transmission tree can be reconstructed by evaluating the probability assigned to the actual transmission events. We estimate the probability that host *j* infected host *i* by the proportion of sampled trees in which *j* infected *i*. We assess false positives by evaluating, for each infected host, whether the infector assigned the highest probability is the actual infector. We further evaluate how well the method estimates the substitution rate, and whether the method is able to find the reduced infectiousness of adults. The latter is done by counting the fraction of infections caused by adults among transmission events estimated at probability at least 0.9. As we are interested in the parameters and transmission tree, the phylogenetic tree can be considered a nuisance parameter, and we do not further investigate its estimation.

To assess robustness of the estimation procedure, we simulate data, estimate parameters and evaluate performance for seven different simulation scenarios. In the baseline scenario, all data are available. In the second and third scenario, we examine the impact of incomplete sampling by randomly discarding 0 per cent or 50 per cent of sequences. In a fourth scenario, we examine the impact of unobserved hosts by randomly discarding 20 per cent of infected hosts. To keep the number of observed hosts comparable, we start these simulations with 20 per cent more susceptibles. In a fifth scenario, we examine sensitivity to an increased substitution rate of 0.01 substitutions/site/year. In a sixth scenario, we examine sensitivity to a decreased substitution rate of 0.001 substitutions/site/year. In a seventh scenario, we examine the impact of a misspecified within-host model, by setting *W(t,h)*=1 in the analysis. Specifying this within-host model is equivalent to making the incorrect assumption that coalescent events coincide with transmission events.

Application to an outbreak of foot-and-mouth disease

To illustrate the use of the method in practice, we re-analyze data on an outbreak of foot-and-mouth disease (FMD) in 2001 in Durham county, England.[22,65,232] Here, for a cluster of 12 infected farms, spatial information, date of symptom onset, culling date and one full genome sequence is known for all 12 farms in the cluster. Both the phylogenetic tree and transmission tree have been estimated before for this outbreak separately, giving inconsistent results.[22] We apply our method to illustrate how to estimate both trees simultaneously, using the same epidemiological model and substitution model as were used in a previous study.[65] We assume an exponentially increasing function for W (see appendix F.2 for details), as at time of culling the disease is still spreading within the farms.

Results

We first test the proposed method by estimating the transmission tree from the simulated datasets. Figure 2 shows how well individual transmissions can be estimated, table 1 shows the average probability assigned to the actual transmissions. When all data are available on average half of the transmissions in the reconstructed transmission tree are correct (figure 2), this number is almost one if we restrict ourselves to transmission events assigned a high probability. Few transmissions can be correctly estimated when the percentage of infected hosts sampled decreases. When instead the percentage of hosts observed decreases transmissions can be estimated correctly quite often, but the percentage of incorrectly estimated transmissions increases strongly. More transmission events can be estimated when the substitution rate is higher, fewer transmission events can be estimated when the substitution rate is lower. Nearly as many transmission events can be correctly estimated when we assume coalescent events coincide with transmission events as when we know the correct within-host model. However, the percentage of transmissions that is incorrectly estimated increases.

We now turn to estimation of mutational parameters from the simulated datasets. The substitution rate can be estimated well even when data are missing (figure 3A); the mean estimated rate was 0.0033, the actual value of 0.003 was contained in the 95% credibility interval for 90% of the simulations. Assuming coalescent events coincide with transmission leads to a large overestimation (mean 0.0067, coverage 15%); this is because the total branch length of the phylogenetic tree is underestimated, leading to an overestimate of the substitution rate (figure 1).

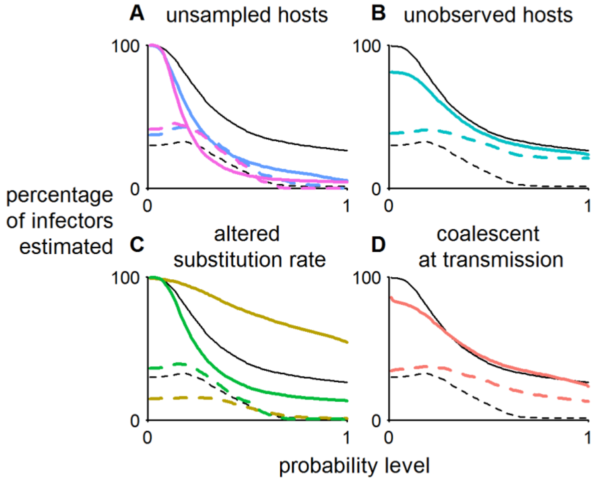


Figure 2. Accuracy of estimating transmission trees using genetic sequences of pathogens, for different simulation scenarios. Solid lines give the average percentage of infected hosts for which the actual infector has been assigned a probability of at least the level indicated on the x-axis. Dashed lines give the average percentage of infected hosts for which the infector has been incorrectly identified, with a probability at least the level indicated on the x-axis. (A) Results when 100% (black), 50% (blue) or 0% (purple) of infected hosts have been sampled. When fewer hosts are sampled, only a few infectors can be identified at a high probability level. (B) Results when all (black) or 80% (turquoise) of hosts are observed. When fewer hosts are observed fewer infectors are identified correctly, and incorrect inferences are made even at high probability levels. (C) Results when substitution rate is 3×10^{-3} (black), an increased 1×10^{-2} (yellow) or a decreased 1×10^{-3} substitutions/site/year (green). At higher substitution rates the inference is more accurate. (D) Results when coalescent events are allowed to differ from transmission events (black) or when coalescent events are incorrectly assumed to coincide with transmission events (pink). The incorrect assumption leads to incorrect estimations even at a high probability level.

Next, we focus on performance in estimating the relative infectiousness of adults from the simulated datasets. Figure 3B gives, for each of 100 simulations under each of the seven scenarios, the point estimate and confidence interval for p, the fraction of infections caused by adults. Most estimates of p are close to the correct value of 0, except for the scenarios where only 80% of hosts are observed (mean p=0.05), or when transmission times are equated with coalescent times (mean p=0.11). More importantly, the confidence interval for p becomes larger when there is more uncertainty about the transmission tree, most notable under the two scenarios with decreased sampling rates.

Having tested the method, we apply it to infer transmission trees from epidemiological and genetic sequence data collected during an FMD outbreak in Durham county, England, in 2001. Figure 4A illustrates a typical sample from the MCMC, consisting of a transmission chain connecting the infected farms and a phylogenetic tree connecting the sampled sequences. The results are broadly similar to those obtained from previous analyses on the same dataset (appendix F.2). The mean latency period was estimated at 7.8 (95% credibility interval (CI): 4.9, 12) days (figure 4), which is in agreement with a typical latency period of 5 days (95% CI: 1-12 days) of FMD virus.[233-235] Ignoring the within-host genetic diversity would have led to an unrealistically large estimate for the latency period of 24 (95% CI: 17, 35) days.[65] The substitution rate was estimated at 1.1×10^{-2} (95%

CI: 8.7×10^{-3}, 1.5×10^{-2}) substitutions per site per year, which is higher than a typical value of 7.7×10^{-3} based on genetic data only.[22] We found the value to be sensitive to the precise specification of the within-host dynamics. See appendix F.2.3 for further results and comparison to previous estimates. Together, these results confirm the method is able to estimate plausible parameter values, whilst reconstructing the transmission tree and a consistent phylogenetic tree of the outbreak.

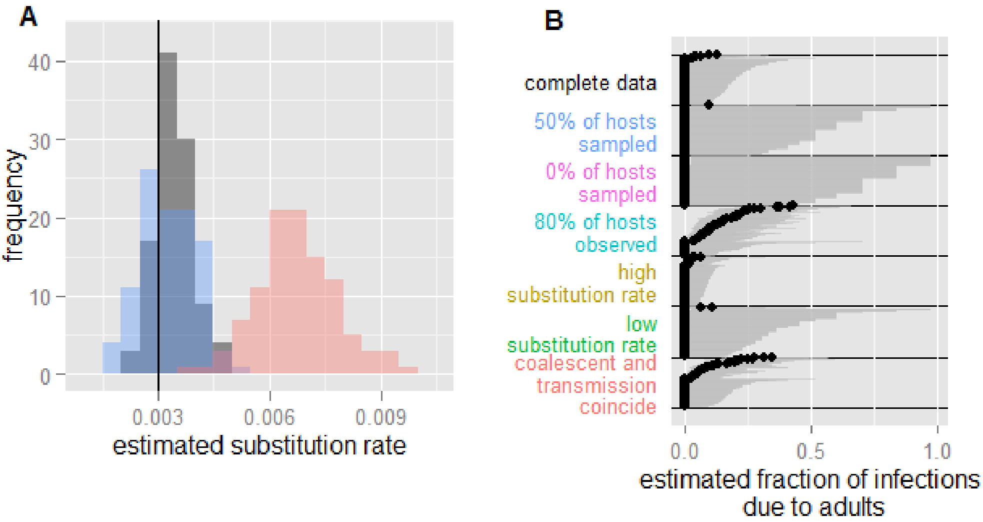


Figure 3. Robustness of estimates of genetic and epidemiological parameters under various scenarios. (A) Distribution of point estimates of the substitution rates for 100 simulations, for three simulation scenarios. Actual value is 0.003 (black line). Estimates are accurate when all information is available (black) and when 50% of hosts are sampled (blue), although the latter leads to a broader distribution. Assuming coalescent events coincide with transmission events (pink) however leads to a large overestimation, since the total branch length of the phylogenetic tree is underestimated. Mean estimates are 3.3×10^{-3}, 3.3×10^{-3} and 6.7×10^{-3} respectively. (B) Point estimate (black) and 95% confidence interval (grey) of the fraction of infections due to adults, actual value is 0. Shown are 100 sorted estimates, for each of seven scenarios (complete data, missing sequences, unobserved hosts, altered substitution rate and incorrect within-host model). Estimates are away from the actual value of 0 when only 80% of hosts are observed, or when coalescent times are equated with transmission times. The width of the confidence interval largely depends on the amount of information available, e.g. when less genetic information is available due to incomplete sampling the point estimates are accurate, but the confidence interval can become very broad.

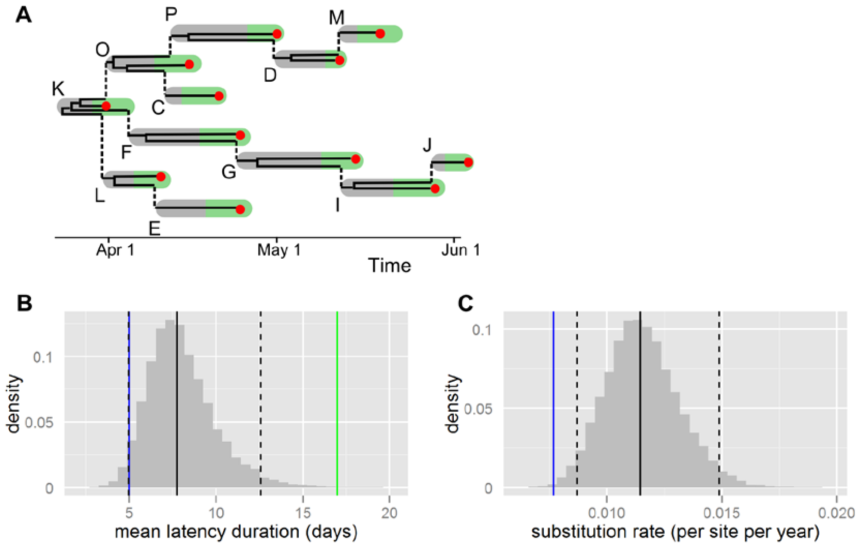


Figure 4. Results from the analysis on the foot-and-mouth disease datasets. (A) A typical transmission tree sampled from the MCMC. Shown are infected farms (labelled pods), their latent periods (grey) and infectious periods (green), samples viruses (red) and the phylogenetic tree connecting these viruses (black). The phylogenetic tree is contained within the transmission tree; due to the exponentially increasing within-host effective pathogen population size assumed, most coalescents occur early during an infection. (B) Posterior distribution for the mean latency period β_1. Solid black line gives the median, striped black lines give the 2.5th and 97.5th percentile. Blue line gives a previous estimate from the literature, green line gives the estimate derived from the same dataset in a previous study that ignored within-host genetic diversity. The estimate (solid black) is higher than we would expect from the literature (blue). The overestimation could be due to unobserved infected farms. Not allowing for within-host genetic diversity gives an overestimation (green). (C) Posterior distribution for the substitution rate μ. Solid black line gives the median, striped black lines give the 2.5th and 97.5th percentile, blue line gives a previous estimate from the literature. The higher estimate we obtained could be due to an overly simplified within-host model.

Table 1. Accuracy of estimating transmission trees from genetic sequences of pathogens average probability assigned to actual transmission events.

Scenario	Percentage correct
all data	50
50% sampled	31
0% sampled	24
80% observed	44
high substitution rate	77
low substitution rate	37
coalescent at transmission	48

Discussion

We have shown how the within-host dynamics of pathogens relate the phylogenetic tree of sampled pathogens to the transmission tree of an outbreak of an infectious disease. We use this relationship to estimate key parameters by combining genetic data on the pathogen with epidemiological data. The advantage of this estimation procedure is that estimates are more accurate, and estimation is feasible even when hosts are unsampled or unobserved.

The method is able to correctly estimate epidemiological and mutational parameters, and can infer individual transmission events. Although the methods performs best when all infected hosts have been observed and sampled, simulations show that even when 20 per cent of infected hosts are unobserved the transmission tree can still be estimated reasonably well. Note however, that estimation of some epidemiological parameters might become biased. For example, in the FMD example, the slight overestimation of both duration of the latency period and substitution rate could well be due to unobserved farms.[65] As expected, the accuracy of the estimates increases with substitution rate. The method therefore seems most suited to analyse datasets on RNA viruses. Analysis for DNA viruses or bacteria might still be feasible if the times between infection are large enough, as sufficient genetic diversity might still accumulate over the course of the outbreak.

To correctly estimate both the transmission tree and the phylogenetic tree, knowledge on the within-host effective pathogen population size and pathogen generation time is needed. This knowledge will however not be available in general. On the short timescales we are considering, we expect it will be challenging to estimate the shape of the time-varying population size from the data. Conversely, the impact of a misspecification of the within-host effective pathogen population size is minor. Even the extreme case of taking a constant population size of 1 (i.e. equating transmission times with coalescent times) leads to reasonable estimation of the transmission tree. We see this in our application to FMD; although the substitution rate is overestimated, the estimated epidemiological parameters agree very well with previous estimates. It is possible that within-host dynamics are easier to estimate for chronic infections. However, for chronic infections the epidemiological data are usually less informative of the transmission tree.

In proposing this methodology, we extend previously proposed methods that aim to reconstruct transmission trees using both epidemiological and genetic data. The field was pioneered by Cottam et al.,[22] who considered a sequential estimation procedure for the phylogenetic tree and the transmission tree; an approach that leads to loss of information contained in the phylogenetic tree. More recent approaches considered both data types simultaneously,[64,65,190] but implicitly or explicitly associated sequences with hosts, rather than with individual pathogens within the host. This means transmission times coincide with coalescent times, an approximation that is inaccurate when the sampling fraction is high. Our simulations show this leads to suboptimal inference of the transmission tree, and to a large overestimation of the substitution rate.

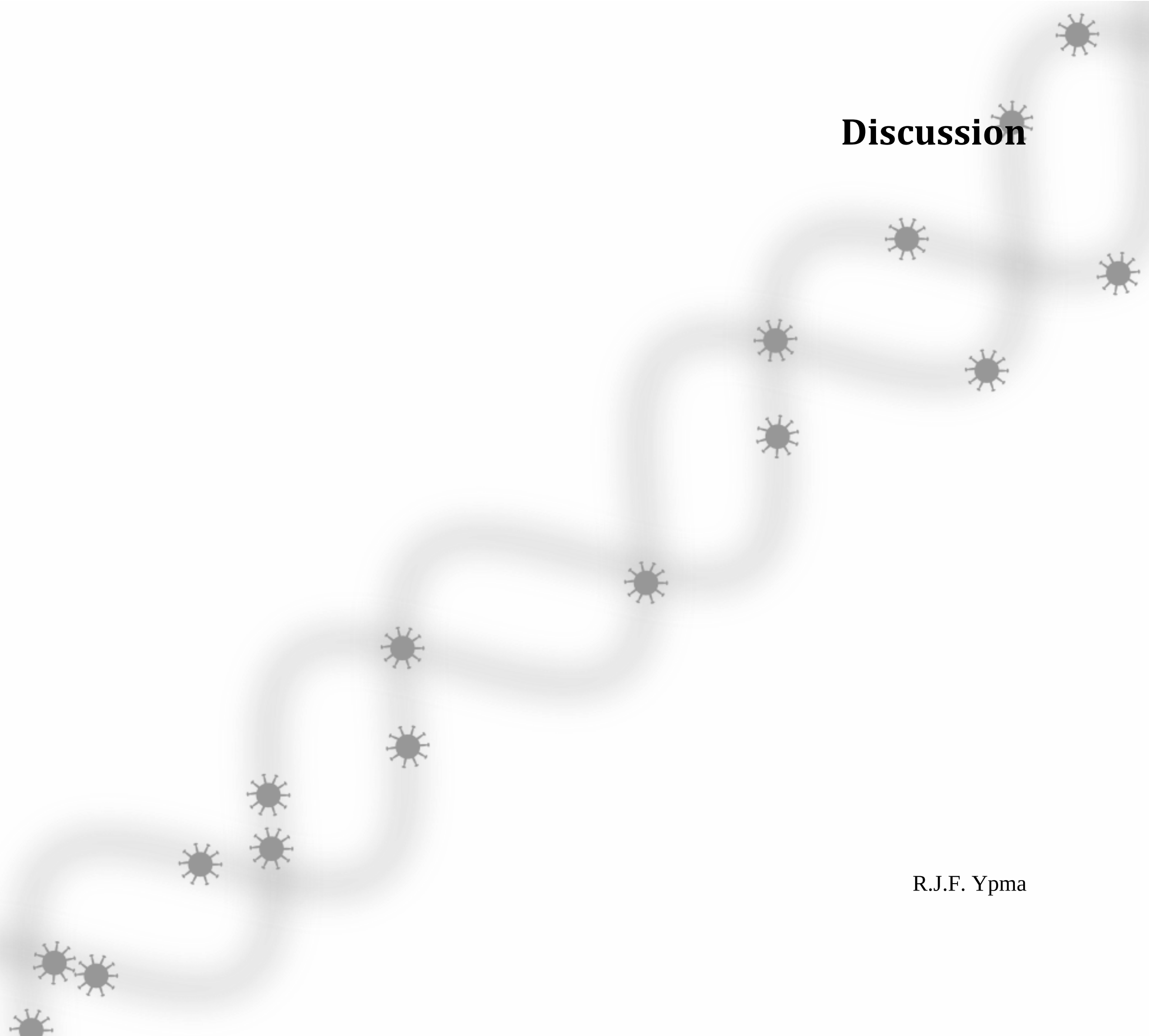
Transmission tree reconstruction methods focus on individual transmission events, whereas many published phylodynamical approaches[66] typically focus on finding general characteristics of an outbreak, e.g. epidemiological parameters, from sampled sequences. To estimate individual transmissions detailed data are needed; we expect that the number of outbreaks that are studied in such detail will increase. More importantly, transmission tree reconstruction allows for an understanding of the outbreak at a high level of detail; for example hypotheses regarding the transmission mechanism can be tested using the reconstructed transmission tree.[225]

Reconstructing large outbreaks at the detailed level of individual transmissions is only feasible when highly informative data are available. These could either take the form of detailed epidemiological data on who infected whom, informative genetic data, i.e. a large number of sampled sequences exhibiting high genetic diversity, or a combination of both. The method we have presented uses both data types to estimate simultaneously the transmission tree and the phylogenetic tree, acknowledging the fact that these two are, in fact, different. With the decreasing cost of sequencing technologies it is likely that more and more of such detailed molecular epidemiological datasets will become available for a range of pathogens, a trend we have already seen in the last few years.[199,236,237] We therefore expect that the usefulness of this combination of two historically separated fields will become even more apparent in years to come.

Chapter 8

Discussion

R.J.F. Ypma



The aim of this thesis has been to develop methodology to quantify infectious disease dynamics by combining genetic and epidemiological data. In chapter 2 we showed how data on date of diagnosis and genotype can be used to quantify the notion of 'superspreading' (i.e. some infected individuals causing a disproportionately large number of secondary cases). In chapters 3 and 4 we developed and applied methodology to detect local outbreaks of infectious diseases from large databases containing temporal, spatial and genetic data. In chapters 5-7 we showed how the transmission tree of an outbreak can be estimated, using information on times of infection, case characteristics and genetic sequences of the pathogen.

The studies presented in this thesis have offered several new insights into the dynamics of the diseases studied. First, the analysis in chapter 2 revealed a high heterogeneity in the number of secondary cases of tuberculosis caused by one infectious individual in the Netherlands. This has implications for allocation of resources in tuberculosis control, since an optimal strategy might be to focus resources on highly infectious cases. Second, the analysis in chapter 4 suggested that MRSA carriers for whom health care staff could not provide a risk category were likely to have acquired their MRSA in the community. Third, the analysis in chapter 5 showed that infectiousness of farms harbouring avian influenza increases with the number of animals. Lastly, the analysis presented in chapter 6 showed a likely role of wind in spreading avian influenza between farms. In more general terms, the methods developed in this thesis provide ways to detect outbreaks (chapter 3), follow infections (chapters 5, 7), elucidate correlates of infectiousness (chapters 4, 5) and find improvements for control (chapters 2, 6). As the methods are generically applicable, more results might be added to this list in the near future.

An analysis combining genetic and epidemiological data can be said to yield more than the sum of its parts;[238] it can give results beyond those obtained from the separate analyses. This is particularly clear for the branching process method and clustering method presented in chapters 2 and 3 of this thesis. In chapter 2, we used genetic and epidemiological data to separate transmission chains for tuberculosis in the Netherlands. Such an analysis would have been impossible using epidemiological data alone, as many transmission chains occur at the same time, and we would have no way to separate them. The genetic data alone is enough to separate the dataset into genotypic clusters, but these are not equal to the transmission clusters needed as final sizes of the branching process. Thus, the separate data sources would not have yielded an estimate of the heterogeneity in infectiousness, whereas the combination of the two does.

The analysis presented in chapter 3 makes use of the correlations found between geographical, temporal and genetic data. Although it can be carried out by using only two of the three,[58] performance would no doubt strongly decrease; small distances between two cases are much more likely to occur simultaneously by chance in two data types than in three data types. This leads to a strongly reduced signal for finding related cases when only two of the three data types are used. As an illustration, not distinguishing the various clonal complexes of MRSA in chapter 4 would have led to a hugely decreased signal. This decrease is due to both the percentage of clustered cases varying strongly per clonal complex, and to the discarding of the small genetic differences within clonal complexes, which hold much of the information.

Likewise, the transmission tree reconstruction methods from chapters 5-7 can be performed using just epidemiological[19] or just genetic[190] data. However, the former yields results only when very few cases are involved, the latter only when the substitution rate is extremely high. Using both types of data, the analysis of transmission events in an outbreak of avian influenza (chapter 5) yielded reliable estimates for almost 50% of the edges in the transmission tree; using epidemiological data alone, this fraction decreased to less than 10%. Similarly, in chapter 6, the combined data were sufficient to identify a wind-mediated mechanism of disease spread, but either type alone was not. Extensive simulations in chapter 7 also show that combining the two data sources in one analysis markedly improves results.

The pattern that arises is the following: epidemiological data are informative on small outbreaks, whereas genetic data allow us to differentiate between different outbreaks. The concept of outbreak I use here is context-dependent. An epidemic can be considered one outbreak; indeed, epidemiological data such as case counts over time can inform us of the general characteristics of an epidemic. However, an epidemic itself consists of several outbreaks which occur simultaneously. Epidemiological data cannot distinguish between them, but genetic data can. Once an epidemic has been subdivided into separate outbreaks, these can be further analysed using epidemiological data. Mathematically, the reason for this ‘separating ability’ of genetic data is its high dimensionality (i.e. the number of base pairs sequenced). Such separation of outbreaks hinges on two factors. First, the evolutionary time separating outbreaks has to be long enough for mutations to have fixed (this need not be the case during fast outbreaks, see chapter 7). Second, the dimensionality of the genetic data has to be high enough that back mutations do not play a large role (this need not be the case for genotyping data, see chapter 2, or when there is strong positive selection).

To predict future developments in the field of molecular epidemiology, we should note that historically, new developments in the field of molecular biology and, by extension, molecular epidemiology, have been technology driven. In recent years, we have seen relatively coarse genotyping techniques being replaced by more advanced sequencing methodology. It is natural to assume that in a few years, studies will focus on whole-genome sequencing or even deep sequencing.[35] Whole-genome sequencing reads every base pair of a genome. Deep sequencing reads all base pairs several times and can thereby establish the frequency of any mutation, giving information on the complete population of pathogens present in a sample. Whole-genome sequencing is increasingly applied to bacteria, which have a large genome and low mutation rate relative to viruses; this technique holds the promise to allow some of the more detailed phylodynamic methods developed for viruses to be applied to bacteria.[236,239-241] Deep sequencing provides information on intra-host genetic diversity. Whether this technique will also increase our understanding of inter-host transmission, compared to conventional sequencing, will depend on the speed with which individual mutations fix.[242,243] For example, the data will not add anything if the fixation times are shorter than the times between sampling. In general, as sequencing techniques become faster and cheaper, we expect an increasing role for analyses that focus on densely sampled outbreaks, such as those in chapters 5-7.

One interesting development in the field of reconstructing transmission trees from epidemiological and genetic data (cf. chapters 5-7) is that new methods seem to fall into one of two classes. The first class of methods focuses on epidemiological data and adds genetic data in a very simplified way (as in chapter 5). The second class focusses on phylogenetic trees constructed from advanced mutational models and adds epidemiological data (similar to chapter 7). It remains to be seen if the extra realism added in these latter analyses by allowing for dependence between time and genetic distance is worth the corresponding increase in computational effort. I would conjecture this to be the case for chronic infections, where times between subsequent infections can be variable, but maybe not for acute infections, where generation intervals are more predictable. Although the observed mutations are a clear marker of transmission tree topology, the short timescales seen in outbreaks can lead to a high variance in the number of mutations among individuals. The mutations then are not very informative of actual timing, particularly as we seldom know the within-host dynamics needed for a proper calculation of evolutionary time.

As the methods described in this thesis are data-driven, it is important to consider the variety of data sources. We can expect the highest quality data to be generated from research projects set up specifically to study a pathogen or outbreak in detail. However, the largest amount of data is generated in surveillance and clinical settings. For example, the large number of HIV sequences available to researchers have been generated because of their clinical relevance in assessing drug-resistance markers.[244-246] Furthermore, initiatives are being raised that do not target a specific outbreak or epidemic, but aim to sequence an enormous amount of 'background' microbes.[248,249] Finding ways to convert these data into useful information will present many interesting challenges.

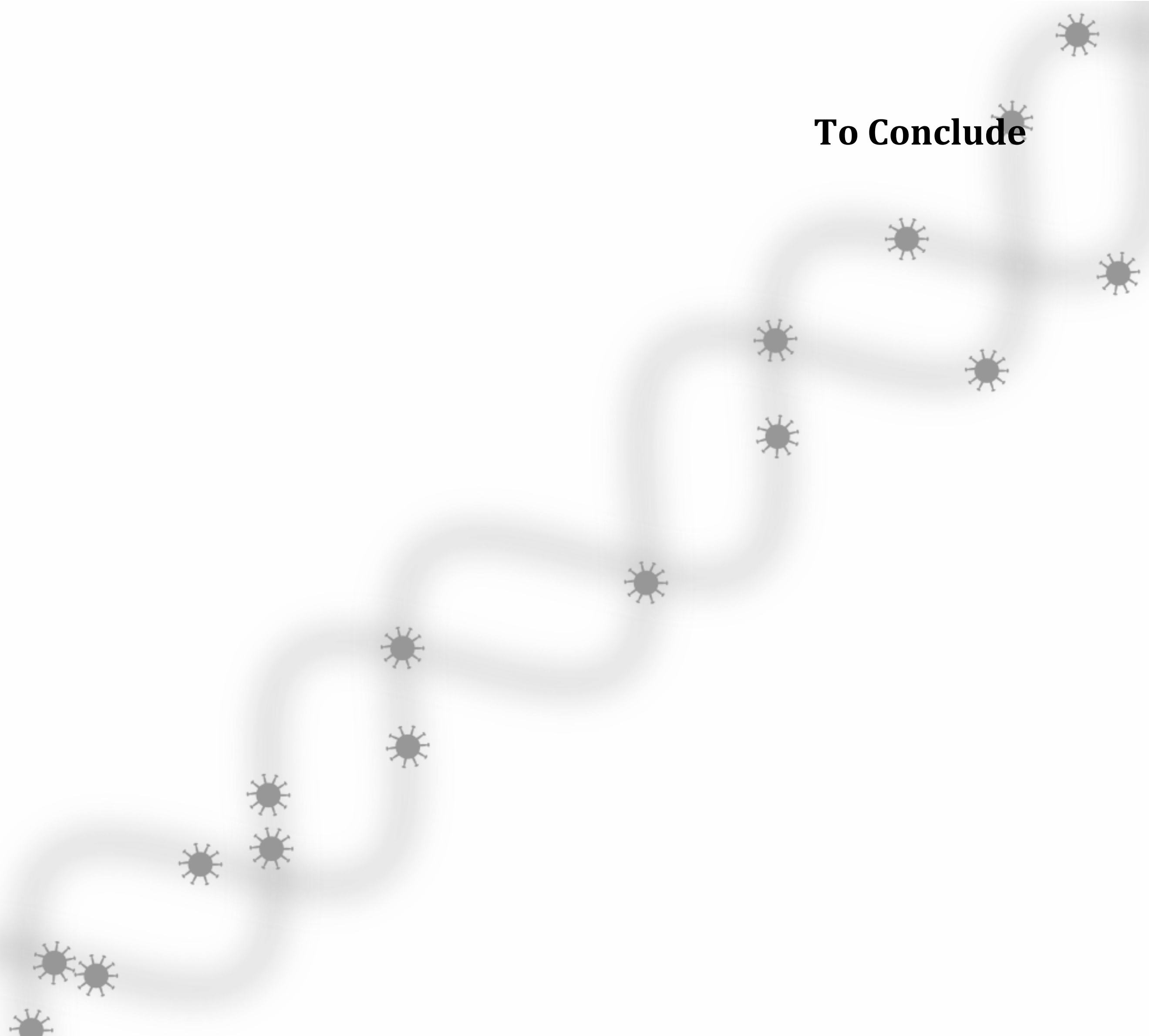
The scientific standard in the field today is that sequences described in the literature should be freely available through public repositories, which is a great way to facilitate the sharing of data. However, as noted before[250] and emphasized by this thesis, research could benefit even more if available 'metadata' (e.g. epidemiological data) were likewise shared on a routine basis. Information on sampling time is invaluable and not always reported. Additionally, sampling location and demographic and clinical host characteristics could be very useful, but are only sporadically available. Focus should be given to encourage researchers to collect and share these additional data.

One aspect that has not been touched upon in this thesis is the need for ethical reflection.[251,252] When genetic data become so informative that it becomes feasible to estimate who-infected-whom, research and data intended to aid public health can raise legal and societal issues. For example, HIV sequence data have been used in a court case where the defendants were accused of deliberately transmitting the virus. Scientists working with pathogen genetic data should consider not only the scientific merit of their research, but also the societal effects it could have.

Combining genetic and epidemiological data in one analysis can be a powerful way to learn more about spread of infectious diseases. As the technology that drives this field is moving fast, future research should focus not only on refining existing methodology, but on anticipating the high quantity and quality of data that will be generated in the next few years. Extracting useful information from these data truly is an interdisciplinary endeavour, with the potential to unravel infectious disease dynamics at unprecedented levels.

Chapter 9

To Conclude



Summary

Infectious diseases are studied both to cure the ill and to prevent new infections from occurring. To achieve the latter, it is often very valuable to know how, when, where and between whom infectious diseases spread. For example, if we knew hospital bacteria were transmitted among patients by health care workers we could increase hygiene measures. If we knew young people were responsible for most new tuberculosis infections, we could increase contact tracing efforts around them. And if we knew avian influenza was predominantly spread by wind, new filters could be installed in farm ventilation systems. Unfortunately, such knowledge can be hard to come by, as transmission of pathogens is often invisible and we generally only observe them indirectly, through the illness they cause. Fortunately, even indirect observations hold valuable information. The field of infectious disease dynamics focusses on obtaining this information from such observations.

A relatively new source of information is formed by genetic data. It has become possible to isolate pathogens from samples taken from cases, such as blood or saliva, and establish the DNA sequence of the pathogens. These sequences can tell us what type of pathogen we are dealing with, and whether we have found a new mutant or a well-known strain. The sequences also hold information on the relationships between the different samples taken, and can thus also inform us on the dynamics of the disease. However, combining these genetic data with more traditional epidemiological data (such as time of symptom onset or location of cases) to unravel infectious disease dynamics is not straightforward.

The goal of this thesis has been to devise mathematical methods to combine epidemiological and genetic data to unravel infectious disease dynamics. The thesis presents three methods, differing in how much information is assumed to be present in the genetic data. An important criterion has been whether methods are applicable in practice; all methods have been applied to actual datasets to illustrate their use.

The first method assumes that genetic data is very coarse: able only to separate distinct clusters of pathogens of the same species that are closely related. The sizes of these clusters are then informative of the dynamics of the disease. In chapter 2 we used such genetic data on tuberculosis in the Netherlands, together with temporal data on cases, to infer groups of cases probably linked through transmission. The presence of both many small and some very large groups indicated that the disease spreads in a very heterogeneous way. This means that most infected people cause no new infections, but some people will cause many. An effective way of combatting the disease might thus be to allocate most resources to identifying these disproportionally infectious individuals.

In the method described in chapter 3, we again assumed the genetic data to consist of genotypes which separate the dataset into clusters, but we additionally assumed a measure of genetic similarity between these clusters to be available. We further assumed there was temporal and geographical information on clusters. We then devised a statistical method that could identify cases that were

more closely related, based on the three data types, than would be expected by chance. In practice, this means we might be able to identify chains of disease transmissions, in a setting where the infectious agent is usually imported from outside. One example of this is drug-resistant pathogens in hospitals that either can be carried into the hospital by the patient or result from spread within the hospital. We applied the method to data on MRSA in Dutch hospitals to find patient characteristics that are associated with within-hospital spread of this pathogen (chapter 4).

The final method we presented has the most stringent demands on data quality and quantity. Both sequence and epidemiological data on a large fraction of individuals infected in an outbreak are assumed available. We then use these data to estimate who-infected-whom (chapter 5). We applied this method to the outbreak of avian influenza A(H7N7) in the Netherlands to show that the direction of transmission coincided with wind direction, which indicates that wind plays a role in spreading the disease (chapter 6). Finally, we show how full phylogenetic trees can be incorporated in the analysis by duly taking into account the within-host genetic diversity that can accumulate during an infection (chapter 7). The final model builds a bridge between transmission tree reconstruction and phylodynamics, between the epidemiology of infectious diseases and the population genetics of pathogens.

Samenvatting

We bestuderen infectieziekten niet alleen om zieken te genezen, maar ook om nieuwe infecties te voorkomen. Om effectieve maatregelen die infecties voorkomen te verzinnen is het nuttig te weten hoe, wanneer, waar en tussen wie infecties plaatsvinden. Bijvoorbeeld: als we wisten dat ziekenhuisinfecties tussen patiënten verspreid werden door verplegers en artsen konden we de hygiëne-maatregelen aanscherpen, als we wisten dat tuberculose-infecties vooral door jongere mensen veroorzaakt werden konden we het contactopsporingsonderzoek rond jonge mensen verzwaren en als we wisten dat aviaire influenza tussen boerderijen vooral verspreid werd door wind konden we filters installeren om de uitstoot van virusdeeltjes tegen te gaan. Dit soort kennis is echter lastig te vergaren, aangezien transmissie van pathogenen in het algemeen onzichtbaar is en we het pathogeen zelf vaak pas zien als mensen er ziek van worden. Dit soort indirecte observaties kunnen echter nog steeds erg waardevol zijn. In het veld van de infectieziektedynamiek proberen onderzoekers deze observaties te gebruiken om nuttige informatie over ziekteprocessen te verkrijgen.

Genetische data vormt een relatief nieuwe bron van informatie. In de afgelopen jaren zijn technieken ontwikkeld en verbeterd die in staat zijn de DNA sequentie van een pathogeen te bepalen van monsters genomen van ziektegevallen, zoals bloed- of speekselmonsters. Deze sequenties kunnen niet alleen gebruikt worden om het pathogeen te identificeren, maar ook om te bepalen of het een bekende variant of een mutant betreft. Verder bevatten de sequenties informatie over de relatie tussen de verschillende monsters, en dus over de verspreiding van de ziekte. Idealiter zouden we deze genetische data willen combineren met meer traditionele epidemiologische data (zoals eerste ziektedag of locatie van gevallen), hoe deze combinatie eruit zou moeten zien is echter niet voor de hand liggend.

Het doel van dit proefschrift is het ontwikkelen van wiskundige methodes om genetische en epidemiologische data te combineren, om hiermee de verspreiding van infectieziekten beter te begrijpen. In het proefschrift worden drie methodes gepresenteerd, die verschillen in het soort genetische data dat ze gebruiken. Een belangrijke eigenschap van elk van deze methodes is dat ze verder gaan dan alleen theorie; alle drie de methodes worden in het proefschrift toegepast op echte data om het gebruik te illustreren.

De eerste methode gaat uit van genetische data met een lage resolutie, slechts in staat pathogenen in te delen in clusters van nauw verwante pathogenen. De verdeling van groottes van de clusters blijkt informatief over de verspreiding van de ziekte. Hoofdstuk 2 richt zich op tuberculose in Nederland. Met behulp van genetische en temporele data worden gevallen van tuberculose ingedeeld in groepen waarbinnen de ziekte zich vermoedelijk verspreid heeft. Het feit dat er heel veel hele kleine, maar ook een aantal zeer grote groepen zijn geeft aan dat de ziekte zich op heterogene wijze verspreid. Hiermee wordt bedoeld dat de meeste gevallen bijna niemand infecteren, maar sommigen juist vele nieuwe gevallen veroorzaken. Een effectieve wijze van bestrijding van de ziekte zou dan

ook kunnen zijn om de meeste middelen te richten op deze buitensporig besmettelijke ziektegevallen.

De tweede methode (hoofdstuk 3) richt zich wederom op genetische data die gevallen opdeelt in clusters, maar nu is er ook een maat van genetische verwantschap tussen de clusters bekend. Verder zijn de bemonsterdag en locatie van gevallen bekend. De methode is dan in staat om gevallen te onderscheiden die nauwer verwant zijn, op basis van de drie datatypes, dan op grond van toeval verwacht kan worden. Hierdoor is het mogelijk om lokale uitbraken van een ziekte te detecteren, in omgevingen waar de meeste gevallen hun infectie elders opgelopen hebben. Een goed voorbeeld van zo’n omgeving is het ziekenhuis, waar veel antibioticaresistente bacteriën door patiënten binnengebracht worden, maar patiënten elkaar ook kunnen besmetten. In hoofdstuk 4 passen we de methode toe op MRSA in Nederlandse ziekenhuizen, om patiëntkarakteristieken te identificeren die gerelateerd zijn aan binnen-ziekenhuis verspreiding van dit pathogeen.

De derde methode stelt de hoogste eisen aan de kwaliteit en kwantiteit van de data; de methode gaat uit van beschikbaarheid van epidemiologische en sequentie data van een groot gedeelte van de gevallen besmet in een uitbraak. Hiermee kan berekend worden wie wie infecteerde (hoofdstuk 5). Door deze methode toe te passen op een grote uitbraak van aviaire influenza A(H7N7) konden we laten zien dat windrichting gecorreleerd was met richting van verspreiding van de ziekte, wat een rol voor wind in de verspreiding impliceert (hoofdstuk 6). De methode kan verder verfijnd worden door ook binnen-gastheer genetische diversiteit mee te nemen (hoofdstuk 7). Deze verfijning slaat een brug tussen transmissieboomreconstructie en phylodynamica, twee velden binnen de wetenschappelijke disciplines van infectieziekte epidemiologie en populatiegenetica.

Dankwoord

Uiteraard zijn een hoop mensen van belang geweest in de totstandkoming van dit proefschrift; zij verdienen dank. Allereerst zijn dit Jacco en Marijn, mijn directe begeleiders en geestelijk vaders van dit project. Marijn, ik heb altijd het gevoel gehad dat we een sterk team zijn, door onze complementaire vaardigheden. Waar ik uren kan besteden aan het (bijna) oplossen van een complex probleem, zie jij de weg om het probleem heen. Waar ik direct en onverstaanbaar ben, heb jij politiek gevoel en weet je mensen te overtuigen. Je bent kritisch waar nodig, maar vooral ook empathisch, rationeel, realistisch en idealistisch. Je keuze voor beleid verbaasde niemand, en ik denk dat we blij mogen zijn dat je het ver gaat schoppen.

Jacco, er zijn weinig wetenschappers van wie ik zo’n hoge dunk heb als van jou. Je bent inhoudelijk erg sterk, hebt een zeer brede interesse, een uitzonderlijk gevoel voor humor en een extreem subtiele wijze van kritiek leveren, en bent toegankelijk en perfectionistisch. Deze eigenschappen hebben geleid tot een verbetering van elk van de artikelen in dit proefschrift, soms meer nog dan mij lief was. Je gewoonte manuscripten met vulpen te becommentariëren is buitengewoon charmant, en compleet onpraktisch. Je bent een ijzersterke begeleider, het was een genoegen met je te werken.

Marc, je rol als promotor vervulde je met verve. Immer positief liet je me veel vrijheid, maar tegelijk wist je kritische vragen te stellen en doorgrondde je snel de essentiële punten van mijn modellen. En dat voor een arts. Bedankt voor onze gesprekken, je suggesties en je vertrouwen.

Twee coauteurs wil ik met name bedanken. Remko, we deelden squash, Proneri, een kamer, voetbal, fitness, Stata en statistische problemen, en nog een hoop zaken die ik hier niet noem. Tjibbe, je lach en ogen zijn onvergetelijk, samen white-boarden was genieten en roadtripping USA!

Ik wil de hele modelleurgroep en EPI, met name Annelie, Madelief, Irene, Patricia, Michiel, Dennis, Anna en Hester, bedanken voor de stimulerende en gezellige omgeving om in te werken. Ik dank al mijn Proneri bestuursgenootjes voor vele leuke vergaderingen en geslaagde evenementen. I thank all of my students; I have learned a lot from you guys, and enjoyed the process. In het bijzonder Rob, met wie ik ook op het persoonlijke vlak nog veel heb. Ik dank Els en Els voor de feilloze ondersteuning. I’d like to thank all those abroad that have sparred with me, and often have been my gracious hosts; Katia, Marc, Pleuni, Thibaut, Marc and many others.

Na deze lange opsomming, die ongetwijfeld nog velen mist, bedank ik in het bijzonder nog jou, mijn lezer. Schrijven is slechts zinnig als er ook gelezen wordt en het feit dat je tot het dankwoord gekomen bent betekent ongetwijfeld dat je dit gehele proefschrift vol aandacht gelezen hebt. Hiervoor mijn dank.

Traditioneel wordt dan als laatste het vriendinnetje bedankt; mijn grote steun en toeverlaat, met wie ik alles delen kan. Wegens promotie doe ik dat hier niet.

Omdat jij me begrijpt,

dank,

mijn Lot.

Curriculum Vitae

Rolf Joseph Ferdinand Ypma was born on March 22, 1986, in Barendrecht. He finished his gymnasium at the Grotius College, Delft, in both the technical and biological tracks. After a brief spell of Aerospace Engineering, he studied Applied Mathematics at the Technical University Delft. He was active in several organizational and cultural committees during his studies, and graduated in 2009. After travelling for some months in South-America, he started his PhD position at the National Institute of Public Health and the Environment (RIVM) and University Medical Center Utrecht in the same year. During his PhD, he headed the Institute’s PhD network, travelled to meet fellow scientists at institutions such as Duke University, Harvard University and Imperial College, and organized several symposia and workshops. He is currently at Cambridge University, studying brain networks.

References

1 Armstrong, G. L., Conn, L. A. & Pinner, R. W. Trends in infectious disease mortality in the United States during the 20th century. *JAMA : the journal of the American Medical Association* **281**, 61-66 (1999).

2 World Health Organization. The 10 leading causes of death by broad income group (2008). (2011).

3 Mathers, C. D. & Loncar, D. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS medicine* **3**, e442-e442, doi:10.1371/journal.pmed.0030442 (2006).

4 World Health Organization. Global tuberculosis report 2012. (2012).

5 World Health Organization. World Malaria Report 2012. (2012).

6 Control, E. C. f. D. p. a. The bacterial challenge: time to react. (2009).

7 Levy, S. B. & Marshall, B. Antibacterial resistance worldwide: causes, challenges and responses. (2004).

8 Fouchier, R. A. M. *et al.* Aetiology: Koch's postulates fulfilled for SARS virus. *Nature* **423**, 240-240, doi:10.1038/423240a (2003).

9 Peiris, J. S. M. *et al.* Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet* **361**, 1319-1325 (2003).

10 Leung, G. M. & Nicoll, A. Reflections on pandemic (H1N1) 2009 and the international response. *PLoS medicine* **7**, 6-6, doi:10.1371/journal.pmed.1000346 (2010).

11 Centers for Disease, C. & Prevention. Swine Influenza A(H1N1) Infection in Two Children --- Southern California, March--April 2009. (2009).

12 Claas, E. C. *et al.* Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus. *Lancet* **351**, 472-477, doi:10.1016/s0140-6736(97)11212-0 (1998).

13 Gao, R. *et al.* Human infection with a novel avian-origin influenza A (H7N9) virus. *N Engl J Med* **368**, 1888-1897, doi:10.1056/NEJMoa1304459 (2013).

14 International Federation for Animal, H. The Costs of Animal Disease. (2012).

15 Fonkwo, P. N. Pricing infectious disease. The economic and health implications of infectious diseases. *EMBO reports* **9 Suppl 1**, S13-17, doi:10.1038/embor.2008.110 (2008).

16 Wallinga, J. & Lipsitch, M. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences* **274**, 599-604, doi:10.1098/rspb.2006.3754 (2007).

17 Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Dynamics and Control*. (Oxford University Press, 1992).

18 Diekmann, O. & Heesterbeek, J. A. P. *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*. (John Wiley & Sons, New York, NY, 2000).

19 Wallinga, J. & Teunis, P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am J Epidemiol* **160**, 509-516 (2004).

20 Russell, C. A. *et al.* The potential for respiratory droplet-transmissible A/H5N1 influenza virus to evolve in a mammalian host. *Science* **336**, 1541-1547, doi:10.1126/science.1222526 (2012).

21 Volz, E. M., Koelle, K. & Bedford, T. Viral Phylodynamics. *PLoS Computational Biology* **9**, e1002947-e1002947, doi:10.1371/journal.pcbi.1002947 (2013).

22 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).

23 Liu, J. *et al.* SARS transmission pattern in Singapore reassessed by viral sequence variation analysis. *PLoS medicine* **2**, e43-e43, doi:10.1371/journal.pmed.0020043 (2005).

24 Box, G. E. P. & Draper, N. R. *Empirical Model-Building and Response Surfaces*. (John Wiley & Sons, New York, NY, 1987).

25 May, R. M. Uses and abuses of mathematics in biology. *Science* **303**, 790-793, doi:10.1126/science.1094442 (2004).

26 Lodish. *Molecular Cell Biology*. (W.H.Freeman and Company, 2007).

27 Berg, R. The indigenous gastrointestinal microflora. *Trends in Microbiology* **4**, 430-435, doi:10.1016/0966-842x(96)10057-3 (1996).

28 Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737-738, doi:10.1038/171737a0 (1953).

29 van Leeuwenhoek, A. Observations, communicated to the publisher by Mr. Antony van Leewenhoeck, in a Dutch letter of the 9 th of Oct. 1676, here English 'd: concerning little animals by him observed in rain-well-sea-and snow-water; as also in water wherein pepper had lain inf. *Philosophical transactions of the Royal Society of London.* **12**, 821-833 (1677).

30 Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 6578-6583 (1998).

31 van Soolingen, D., de Haas, P. E., Hermans, P. W., Groenen, P. M. & van Embden, J. D. Comparison of various repetitive DNA elements as genetic markers for strain differentiation and epidemiology of Mycobacterium tuberculosis. *J Clin Microbiol* **31**, 1987-1995 (1993).

32 Sanger, F. *et al.* Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature* **265**, 687-695, doi:10.1038/265687a0 (1977).

33 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).

34 Venter, J. C. *et al.* The sequence of the human genome. *Science (New York, N.Y.)* **291**, 1304-1351, doi:10.1126/science.1058040 (2001).

35 Wetterstrand, K. A. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at: www.genome.gov/sequencingcosts* (2013).

36 Mardis, E. R. A decade's perspective on DNA sequencing technology. *Nature* **470**, 198-203, doi:10.1038/nature09796 (2011).

37 Snow, J. in *Medical Times and Gazette* 321-322 (1854).

38 Bernoulli, D. Essai d'une nouvelle analyse de la mortalite causee par la petite verole et des avantages de l'inoculation pour la prevenir. *Mem. Math. Phys. Acad. Roy. Sci., Paris*, 1-45 (1766).

39 Bonita, R., Beaglehole, R. & Kjellström, T. *Basic Epidemiology*. (World Health Organization, 2006).

40 Kermack, W. O. & McKendrick, A. G. A Contribution to the Mathematical Theory of Epidemics. *Proc. Roy. Soc. Lond. A* **115**, 700-721 (1927).

41 Lotka, A. J. *Elements of Physical Biology*. (Williams and Wilkins Company, 1925).

42 Volterra, V. Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. *Mem. Acad. Lincei Roma* **2**, 31-113 (1926).

43 Small, P. M. *et al.* The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. *The New England journal of medicine* **330**, 1703-1709, doi:10.1056/nejm199406163302402 (1994).

44 Archie, E. a., Luikart, G. & Ezenwa, V. O. Infecting epidemiology with genetics: a new frontier in disease ecology. *Trends in ecology & evolution (Personal edition)* **24**, 21-30, doi:10.1016/j.tree.2008.08.008 (2009).

45 van Boven, M. *et al.* Estimation of measles vaccine efficacy and critical vaccination coverage in a highly vaccinated population. *Journal of the Royal Society, Interface / the Royal Society* **7**, 1537-1544, doi:10.1098/rsif.2010.0086 (2010).

46 Robert, C. P. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation* (Springer Verlag, New York, 2007).

47 Becker, N. On parametric estimation for mortal branching processes. *Biometrika* **61**, 393-399 (1974).

48 Farrington, C. P., Kanaan, M. N. & Gay, N. J. Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics (Oxford, England)* **4**, 279-295, doi:10.1093/biostatistics/4.2.279 (2003).

49 Jansen, V. A. A. *et al.* Measles outbreaks in a population with declining vaccine uptake. *Science* **301**, 804-804 (2003).

50 Ferguson, N. M., Fraser, C., Donnelly, C. A., Ghani, A. C. & Anderson, R. M. Public Health Risk from the Avian H5N1 Influenza Epidemic. *Science* **304**, 968-969 (2004).

51 Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355-359 (2005).

52 Farrington, C. P., Andrews, N. J., Beal, A. D. & Catchpole, M. A. A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease. *J. R. Statist. Soc. A* **159**, 547-563 (1996).

53 Stroup, D. F., Williamson, G. D., Herndon, J. L. & Karon, J. M. Detection of aberrations in the occurrence of notifiable diseases surveillance data. *Statistics in medicine* **8**, 323-329; discussion 331-322 (1989).

54 Le Strat, Y. & Carrat, F. Monitoring epidemiologic surveillance data using hidden Markov models. *Stat. Med.* **18**, 3463-3478 (1999).

55 Hossain, M. M. & Lawson, A. B. Space-time Bayesian small area disease risk models: development and evaluation with a focus on cluster detection. *Environmental and ecological statistics* **17**, 73-95, doi:10.1007/s10651-008-0102-z (2010).

56 Watkins, R. E., Eagleson, S., Veenendaal, B., Wright, G. & Plant, A. J. Disease surveillance using a hidden Markov model. *BMC Med Inform Decis Mak* **9**, 39 (2009).

57 Kulldorff, M. A spatial scan statistic. *Communications in Statistics-Theory and Methods* **26**, 1481-1496 (1997).

58 Kulldorff, M., Heffernan, R., Hartman, J., Assuncao, R. & Mostashari, F. A space-time permutation scan statistic for disease outbreak detection. *PLoS Med* **2**, e59 (2005).

59 Huang, S. S. *et al.* Automated detection of infectious disease outbreaks in hospitals: a retrospective cohort study. *PLoS medicine* **7**, e1000238-e1000238, doi:10.1371/journal.pmed.1000238 (2010).

60 O'Brien, T. F. & Stelling, J. Integrated Multilevel Surveillance of the World's Infecting Microbes and Their Resistance to Antimicrobial Agents. *Clinical microbiology reviews* **24**, 281-295, doi:10.1128/cmr.00021-10 (2011).

61 Heijne, J. C. *et al.* Enhanced hygiene measures and norovirus transmission during an outbreak. *Emerg Infect Dis* **15**, 24-30 (2009).

62 Haydon, D. T. *et al.* The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak. *Proceedings. Biological sciences / The Royal Society* **270**, 121-127, doi:10.1098/rspb.2002.2191 (2003).

63 Cauchemez, S. *et al.* Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 2825-2830, doi:10.1073/pnas.1008895108 (2011).

64 Ypma, R. J. F. *et al.* Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings. Biological sciences / The Royal Society* **279**, 444-450, doi:10.1098/rspb.2011.0913 (2012).

65 Morelli, M. J. *et al.* A Bayesian Inference Framework to Reconstruct Transmission Trees Using Epidemiological and Genetic Data. *PLoS Computational Biology* **8**, e1002768-e1002768, doi:10.1371/journal.pcbi.1002768 (2012).

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

67 Baum, D. Reading a phylogenetic tree: The meaning of monophyletic groups. *Nature Education* **1** (2008).

68 Zuckerkandl, E. & Pauling, L. B. in *Horizons in Biochemistry* (eds M. Kasha & B. Pullman) 189-225 (New York: Academic Press, 1962).

69 Thorne, J. L., Kishino, H. & Painter, I. S. Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol* **15**, 1647-1657 (1998).

70 Sanderson, M. J. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular biology and evolution* **14**, 1218-1231 (1997).

71 Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**, 111-120 (1980).

72 Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **17**, 368-376 (1981).

73 Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution* **29**, 1969-1973, doi:10.1093/molbev/mss075 (2012).

74 Suchard, M. A. & Rambaut, A. Many-core algorithms for statistical phylogenetics. *Bioinformatics (Oxford, England)* **25**, 1370-1376, doi:10.1093/bioinformatics/btp244 (2009).

75 Roch, S. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* **3**, 92-94, doi:10.1109/TCBB.2006.4 (2006).

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

77 Ewens, W. J. *Mathematical Population Genetics (2nd Edition)*. (Springer-Verlag, New York, 2004).

78 Kingman, J. F. C. The coalescent. *Stochastic processes and their applications* (1982).

79 Kingman, J. F. C. On the genealogy of large populations. *Journal of Applied Probability* (1982).

80 Kingman, J. F. Origins of the coalescent. 1974-1982. *Genetics* **156**, 1461-1463 (2000).

81 Fisher, R. A. *The genetical theory of natural selection*. (Clarendon Press, Oxford, 1930).

82 Wright, S. Evolution in Mendelian populations. *Genetics* (1931).

83 Griffiths, R. C. & Tavaré, S. Sampling theory for neutral alleles in a varying environment. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **344**, 403-410, doi:10.1098/rstb.1994.0079 (1994).

84 Donnelly, P. & Tavaré, S. Coalescents and genealogical structure under neutrality. *Annual review of genetics* **29**, 401-421, doi:10.1146/annurev.ge.29.120195.002153 (1995).

85 Marjoram, P. & Tavaré, S. Modern computational approaches for analysing molecular genetic variation data. *Nature reviews. Genetics* **7**, 759-770, doi:10.1038/nrg1961 (2006).

86 Nee, S., Holmes, E. C., Rambaut, A. & Harvey, P. H. *Inferring population history from molecular phylogenies*. 66-80 (Oxford University Press, Oxford, 1996).

87 Rodrigo, A. G. & Felsenstein, J. Coalescent approaches to HIV-1 population genetics. in press in Molecular Evolution of HIV, ed. KA Crandall. (1998).

88 Drummond, A. J., Nicholls, G. K., Rodrigo, A. G. & Solomon, W. Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data. *Genetics* **161**, 1307-1320 (2002).

89 Pybus, O. G., Charleston, M. A. & Gupta, S. The epidemic behavior of the hepatitis C virus. *Science* (2001).

90 Grenfell, B. T. *et al.* Unifying the epidemiological and evolutionary dynamics of pathogens. *Science (New York, N.Y.)* **303**, 327-332, doi:10.1126/science.1090727 (2004).

91 Volz, E. M., Kosakovsky Pond, S. L., Ward, M. J., Leigh Brown, A. J. & Frost, S. D. Phylodynamics of infectious disease epidemics. *Genetics* **183**, 1421-1430 (2009).

92 Siebenga, J. J. *et al.* Phylodynamic reconstruction reveals norovirus GII.4 epidemic expansions and their molecular determinants. *PLoS pathogens* **6**, e1000884-e1000884, doi:10.1371/journal.ppat.1000884 (2010).

93 Sprong, H. *et al.* Circumstantial evidence for an increase in the total number and activity of borrelia-infected ixodes ricinus in the Netherlands. *Parasites & vectors* **5**, 294-294, doi:10.1186/1756-3305-5-294 (2012).

94 Biek, R., Drummond, A. J. & Poss, M. A virus reveals population structure and recent demographic history of its carnivore host. *Science (New York, N.Y.)* **311**, 538-541, doi:10.1126/science.1121360 (2006).

95 Biek, R., Henderson, J. C., Waller, L. A., Rupprecht, C. E. & Real, L. A. A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 7993-7998, doi:10.1073/pnas.0700741104 (2007).

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

97 Pybus, O. G., Drummond, a. J., Nakano, T., Robertson, B. H. & Rambaut, a. The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. *Molecular biology and evolution* **20**, 381-387 (2003).

98 Talbi, C. *et al.* Phylodynamics and human-mediated dispersal of a zoonotic virus. *PLoS pathogens* **6**, e1001166-e1001166, doi:10.1371/journal.ppat.1001166 (2010).

99 van Ballegooijen, W. M. *et al.* Molecular sequence data of hepatitis B virus and genetic diversity after vaccination. *American journal of epidemiology* **170**, 1455-1463, doi:10.1093/aje/kwp375 (2009).

100 Magiorkinis, G. *et al.* Integrating phylodynamics and epidemiology to estimate transmission diversity in viral epidemics. *PLoS computational biology* **9**, e1002876-e1002876, doi:10.1371/journal.pcbi.1002876 (2013).

101 Leventhal, G. E. *et al.* Inferring epidemic contact structure from phylogenetic trees. *PLoS Comput Biol* **8**, e1002413 (2012).

102 Frost, S. D. W. & Volz, E. M. Modelling tree shape and structure in viral phylodynamics. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **368**, 20120208-20120208, doi:10.1098/rstb.2012.0208 (2013).

103 Goodreau, S. M. Assessing the effects of human mixing patterns on human immunodeficiency virus-1 interhost phylogenetics through social network simulation. *Genetics* **172**, 2033-2045, doi:10.1534/genetics.103.024612 (2006).

104 Felsenstein, J. Phylogenies from molecular sequences: inference and reliability. *Annu Rev Genet* **22**, 521-565, doi:10.1146/annurev.ge.22.120188.002513 (1988).

105 Kuhner, M. K., Yamato, J. & Felsenstein, J. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* **140**, 1421-1430 (1995).

106 Rosenberg, N. A. & Nordborg, M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat Rev Genet* **3**, 380-390, doi:10.1038/nrg795 (2002).

107 Kuhner, M. K., Yamato, J. & Felsenstein, J. Maximum Likelihood Estimation of Population Growth Rates Based on the Coalescent. *Genetics* **149**, 429-434 (1998).

108 Strimmer, K. & Pybus, O. G. Exploring the Demographic History of DNA Sequences Using the Generalized Skyline Plot. *Molecular biology and evolution* **18**, 2298-2305, doi:10.1093/oxfordjournals.molbev.a003776 (2001).

109 Opgen-Rhein, R., Fahrmeir, L. & Strimmer, K. Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. *BMC evolutionary biology* **5**, 6-6, doi:10.1186/1471-2148-5-6 (2005).

110 Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular biology and evolution* **22**, 1185-1192, doi:10.1093/molbev/msi103 (2005).

111 Minin, V. N., Bloomquist, E. W. & Suchard, M. A. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular biology and evolution* **25**, 1459-1471, doi:10.1093/molbev/msn090 (2008).

112 Stadler, T. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *Journal of theoretical biology* **261**, 58-66, doi:10.1016/j.jtbi.2009.07.018 (2009).

113 Frost, S. D. W. & Volz, E. M. Viral phylodynamics and the search for an 'effective number of infections'. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **365**, 1879-1890, doi:10.1098/rstb.2010.0060 (2010).

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

115 Gernhard, T. New analytic results for speciation times in neutral models. *Bulletin of mathematical biology* **70**, 1082-1097, doi:10.1007/s11538-007-9291-0 (2008).

116 Gernhard, T. The conditioned reconstructed process. *Journal of theoretical biology* **253**, 769-778, doi:10.1016/j.jtbi.2008.04.005 (2008).

117 Stadler, T. Lineages-through-time plots of neutral models for speciation. *Mathematical biosciences* **216**, 163-171, doi:10.1016/j.mbs.2008.09.006 (2008).

118 Stadler, T. *et al.* Estimating the basic reproductive number from viral sequence data. *Mol Biol Evol* **29**, 347-357 (2012).

119 Stadler, T. & Steel, M. Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models. *Journal of theoretical biology* **297**, 33-40, doi:10.1016/j.jtbi.2011.11.019 (2012).

120 Stadler, T., Kühnert, D., Bonhoeffer, S. & Drummond, A. J. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences of the United States of America* **110**, 228-233, doi:10.1073/pnas.1207965110 (2013).

121 Jermann, T. M., Opitz, J. G., Stackhouse, J. & Benner, S. A. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature* **374**, 57-59, doi:10.1038/374057a0 (1995).

122 Schluter, D. Uncertainty in ancient phylogenies. *Nature* **377**, 108-110, doi:10.1038/377108a0 (1995).

123 Lemey, P., Rambaut, A., Welch, J. J. & Suchard, M. A. Phylogeography takes a relaxed random walk in continuous space and time. *Molecular biology and evolution* **27**, 1877-1885, doi:10.1093/molbev/msq067 (2010).

124 Pybus, O. G. *et al.* Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 15066-15071, doi:10.1073/pnas.1206598109 (2012).

125 Cybis, G. B., Sinsheimer, J. S., Lemey, P. & Suchard, M. A. Graph hierarchies for phylogeography. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **368**, 20120206-20120206, doi:10.1098/rstb.2012.0206 (2013).

126 Lemey, P., Suchard, M. & Rambaut, A. Reconstructing the initial global spread of a human influenza pandemic: A Bayesian spatial-temporal model for the global spread of H1N1pdm. *PLoS currents* **1**, RRN1031, doi:10.1371/currents.RRN1031 (2009).

127 Pilcher, C. D. *et al.* Brief but efficient: acute HIV infection and the sexual transmission of HIV. *The Journal of infectious diseases* **189**, 1785-1792, doi:10.1086/386333 (2004).

128 Stadler, T. & Bonhoeffer, S. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **368**, 20120198-20120198, doi:10.1098/rstb.2012.0198 (2013).

129 WHO. Global tuberculosis report. Report No. ISBN 978 92 4 156438 0, (WHO, 2011).

130 van Embden, J. D. *et al.* Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol* **31**, 406-409 (1993).

131 de Boer, A. S. *et al.* Analysis of rate of change of IS6110 RFLP patterns of Mycobacterium tuberculosis based on serial patient isolates. *J Infect Dis* **180**, 1238-1244 (1999).

132 Luciani, F., Francis, A. R. & Tanaka, M. M. Interpreting genotype cluster sizes of Mycobacterium tuberculosis isolates typed with IS6110 and spoligotyping. *Infect Genet Evol* **8**, 182-190 (2008).

133 Kiers, A., Drost, A. P., van Soolingen, D. & Veen, J. Use of DNA fingerprinting in international source case finding during a large outbreak of tuberculosis in The Netherlands. *Int J Tuberc Lung Dis* **1**, 239-245 (1997).

134 Vynnycky, E. *et al.* The effect of age and study duration on the relationship between 'clustering' of DNA fingerprint patterns and the proportion of tuberculosis disease attributable to recent transmission. *Epidemiology and infection* **126**, 43-62 (2001).

135 Vynnycky, E., Borgdorff, M. W., van Soolingen, D. & Fine, P. E. Annual Mycobacterium tuberculosis infection risk and interpretation of clustering statistics. *Emerg Infect Dis* **9**, 176-183 (2003).

136 Stadler, T. Inferring epidemiological parameters on the basis of allele frequencies. *Genetics* **188**, 663-672, doi:10.1534/genetics.111.126466 (2011).

137 Kremer, K. *et al.* Comparison of methods based on different molecular epidemiological markers for typing of Mycobacterium tuberculosis complex strains: interlaboratory study of discriminatory power and reproducibility. *J Clin Microbiol* **37**, 2607-2618 (1999).

138 Wallinga, J., Teunis, P. & Kretzschmar, M. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *Am J Epidemiol* **164**, 936-944 (2006).

139 Mossong, J. *et al.* Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med* **5**, e74 (2008).

140 Dietz, K. The estimation of the basic reproduction number for infectious diseases. *Stat Methods Med Res* **2**, 23-41 (1993).

141 Fraser, C., Cummings, D. A., Klinkenberg, D., Burke, D. S. & Ferguson, N. M. Influenza transmission in households during the 1918 pandemic. *Am J Epidemiol* **174**, 505-514 (2011).

142 Li, Y. *et al.* Predicting super spreading events during the 2003 severe acute respiratory syndrome epidemics in Hong Kong and Singapore. *Am J Epidemiol* **160**, 719-728 (2004).

143 Lipsitch, M. *et al.* Transmission dynamics and control of severe acute respiratory syndrome. *Science* **300**, 1966-1970 (2003).

144 Borgdorff, M. W., van der Werf, M. J., de Haas, P. E., Kremer, K. & van Soolingen, D. Tuberculosis elimination in the Netherlands. *Emerg Infect Dis* **11**, 597-602 (2005).

145 Schurch, A. C. *et al.* High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. *J Clin Microbiol* **48**, 3403-3406 (2010).

146 Eilers, P. H., Van Soolingen, D., Thi Ngoc Lan, N., Warren, R. M. & Borgdorff, M. W. Transposition rates of Mycobacterium tuberculosis IS6110 restriction fragment length polymorphism patterns. *J Clin Microbiol* **42**, 2461-2464 (2004).

147 Venzon, D. J. & Moolgavkar, S. H. A method for computing profile-likelihood-based confidence intervals. *Appl. Statist.* **37**, 87-94 (1988).

148 Borgdorff, M. W. *et al.* The incubation period distribution of tuberculosis estimated with a molecular epidemiological approach. *Int J Epidemiol* **40**, 964-970 (2011).

149 Tanaka, M. M. & Rosenberg, N. A. Optimal estimation of transposition rates of insertion sequences for molecular epidemiology. *Statistics in medicine* **20**, 2409-2420 (2001).

150 Warren, R. M. *et al.* Calculation of the stability of the IS6110 banding pattern in patients with persistent Mycobacterium tuberculosis disease. *J Clin Microbiol* **40**, 1705-1708 (2002).

151 Borgdorff, M. W. *et al.* Progress towards tuberculosis elimination: secular trend, immigration and transmission. *Eur Respir J* **36**, 339-347 (2010).

152 Verhagen, L. M. *et al.* Mycobacterial factors relevant for transmission of tuberculosis. *J Infect Dis* **203**, 1249-1255 (2011).

153 Caminero, J. A. *et al.* Epidemiological evidence of the spread of a Mycobacterium tuberculosis strain of the Beijing genotype on Gran Canaria Island. *Am J Respir Crit Care Med* **164**, 1165-1170 (2001).

154 Murray, M. Determinants of cluster distribution in the molecular epidemiology of tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 1538-1543, doi:10.1073/pnas.022618299 (2002).

155 Leigh Brown, A. J. *et al.* Transmission network parameters estimated from HIV sequences for a nationwide epidemic. *The Journal of infectious diseases* **204**, 1463-1469, doi:10.1093/infdis/jir550 (2011).

156 Erkens, C. G. *et al.* Tuberculosis contact investigation in low prevalence countries: a European consensus. *Eur Respir J* **36**, 925-949 (2010).

157 Hutwagner, L. C., Maloney, E. K., Bean, N. H., Slutsker, L. & Martin, S. M. Using laboratory-based surveillance data for prevention: an algorithm for detecting Salmonella outbreaks. *Emerg Infect Dis* **3**, 395-400 (1997).

158 Murphy, S. P. & Burkom, H. Recombinant temporal aberration detection algorithms for enhanced biosurveillance. *J Am Med Inform Assoc* **15**, 77-86 (2008).

159 Nobre, F. F. & Stroup, D. F. A monitoring system to detect changes in public health surveillance data. *Int J Epidemiol* **23**, 408-418 (1994).

160 Stern, L. & Lightfoot, D. Automated outbreak detection: a quantitative retrospective analysis. *Epidemiol Infect* **122**, 103-110 (1999).

161 Stelling, J. *et al.* Automated use of WHONET and SaTScan to detect outbreaks of Shigella spp. using antimicrobial resistance phenotypes. *Epidemiol Infect* **138**, 873-883 (2010).

162 Florek, K., Lukaszewicz, J., Perkal, J., Steinhaus, H. & Zubrzycki, S. K. Florek, J. Lukaszewicz, J. Perkal, H. Steinhaus, S. Zubrzycki, Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicum 2*, 282-285 (1951).

163 McQuitty, L. L. Elementary linkage analysis for isolating orthogonal and oblique types and typal relevancies. *Educ. Psychol. Measmt.* **17**, 207-222 (1957).

164 Sneath, P. H. A. The application of computers to taxonomy. *J. gen. Microbiol.* **17**, 201-226 (1957).

165 Jacquez, G. M. A k nearest neighbour test for space-time interaction. *Statistics in medicine* **15**, 1935-1949 (1996).

166 Lloyd-Smith, J. O. *et al.* Epidemic dynamics at the human-animal interface. *Science* **326**, 1362-1367, doi:10.1126/science.1177345 (2009).

167 Xia, Y., Bjornstad, O. N. & Grenfell, B. T. Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. *The American naturalist* **164**, 267-281, doi:10.1086/422341 (2004).

168 Balcan, D. *et al.* Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc Natl Acad Sci U S A* **106**, 21484-21489, doi:10.1073/pnas.0906910106 (2009).

169 de Beer, J. L. *et al.* Comparative study of IS6110 RFLP and VNTR typing of Mycobacterium tuberculosis in the Netherlands, based on a five year nationwide survey. *J Clin Microbiol*, doi:10.1128/JCM.03061-12 (2013).

170 Grundmann, H. *et al.* Geographic distribution of Staphylococcus aureus causing invasive infections in Europe: a molecular-epidemiological analysis. *PLoS Med* **7**, e1000215, doi:10.1371/journal.pmed.1000215 (2010).

171 Verhoef, L. *et al.* An integrated approach to identifying international foodborne norovirus outbreaks. *Emerg Infect Dis* **17**, 412-418, doi:10.3201/eid1703.100979 (2011).

172 Mukomolov, S. *et al.* Increased circulation of hepatitis A virus genotype IIIA over the last decade in St Petersburg, Russia. *Journal of medical virology* **84**, 1528-1534, doi:10.1002/jmv.23378 (2012).

173 European Antimicrobial Resistance Surveillance Network (EARS-Net), Annual Report. ECDC. Acquired from http://www.ecdc.europa.eu/en/publications/Publications/antimicrobial-resistance-surveillance-europe-2011.pdf (2011).

174 Vandenbroucke-Grauls, C. M. Methicillin-resistant Staphylococcus aureus control in hospitals: the Dutch experience. *Infection control and hospital epidemiology : the official journal of the Society of Hospital Epidemiologists of America* **17**, 512-513 (1996).

175 Infection Prevention Working Party (WIP), Hospital MRSA guideline. Acquired from http://www.wip.nl/UK/free_content/Richtlijnen/MRSA hospital.pdf (2012).

176 Ypma, R. J., Donker, T., van Ballegooijen, W. M. & Wallinga, J. Finding evidence for local transmission of contagious disease in molecular epidemiological datasets. *PloS one* **8**, e69875, doi:10.1371/journal.pone.0069875 (2013).

177 Donker, T., Wallinga, J. & Grundmann, H. Patient referral patterns and the spread of hospital-acquired infections through national health care networks. *PLoS computational biology* **6**, e1000715-e1000715, doi:10.1371/journal.pcbi.1000715 (2010).

178 Donker, T., Wallinga, J. & Grundmann, H. Dispersal of antibiotic-resistant high-risk clones by hospital networks: changing the patient direction can make all the difference. *Journal of Hospital Infection* **in press.** (2013).

179 de Neeling, A. J. *et al.* High prevalence of methicillin resistant Staphylococcus aureus in pigs. *Vet Microbiol* **122**, 366-372, doi:10.1016/j.vetmic.2007.01.027 (2007).

180 Otter, J. A. & French, G. L. Molecular epidemiology of community-associated meticillin-resistant Staphylococcus aureus in Europe. *Lancet Infect Dis* **10**, 227-239, doi:10.1016/S1473-3099(10)70053-0 (2010).

181 Francis, J. S. *et al.* Severe community-onset pneumonia in healthy adults caused by methicillin-resistant Staphylococcus aureus carrying the Panton-Valentine leukocidin genes. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **40**, 100-107, doi:10.1086/427148 (2005).

182 Zetola, N., Francis, J. S., Nuermberger, E. L. & Bishai, W. R. Community-acquired meticillin-resistant Staphylococcus aureus: an emerging threat. *Lancet Infect Dis* **5**, 275-286, doi:10.1016/S1473-3099(05)70112-2 (2005).

183 Lekkerkerk, W. S. *et al.* Emergence of MRSA of unknown origin in the Netherlands. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* **18**, 656-661, doi:10.1111/j.1469-0691.2011.03662.x (2012).

184 van Loo, I. *et al.* Emergence of methicillin-resistant Staphylococcus aureus of animal origin in humans. *Emerg Infect Dis* **13**, 1834-1839, doi:10.3201/eid1312.070384 (2007).

185 Keeling, M. J., Woolhouse, M. E., May, R. M., Davies, G. & Grenfell, B. T. Modelling vaccination strategies against foot-and-mouth disease. *Nature* **421**, 136-142 (2003).

186 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 (New York, N.Y.)* **292**, 1155-1160, doi:10.1126/science.1061020 (2001).

187 Spada, E. *et al.* Use of the minimum spanning tree model for molecular epidemiological investigation of a nosocomial outbreak of hepatitis C virus infection. *J Clin Microbiol* **42**, 4230-4236 (2004).

188 Zhang, L. *et al.* Host-specific driving force in human immunodeficiency virus type 1 evolution in vivo. *J Virol* **71**, 2555-2561 (1997).

189 Leitner, T. & Albert, J. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc Natl Acad Sci U S A* **96**, 10752-10757 (1999).

190 Jombart, T., Eggo, R. M., Dodd, P. J. & Balloux, F. Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity* **106**, 383-390, doi:10.1038/hdy.2010.78 (2011).

191 Stegeman, A. *et al.* Avian influenza A virus (H7N7) epidemic in The Netherlands in 2003: course of the epidemic and effectiveness of control measures. *J Infect Dis* **190**, 2088-2095 (2004).

192 Fouchier, R. A. *et al.* Avian influenza A virus (H7N7) associated with human conjunctivitis and a fatal case of acute respiratory distress syndrome. *Proc Natl Acad Sci U S A* **101**, 1356-1361 (2004).

193 Koopmans, M. *et al.* Transmission of H7N7 avian influenza A virus to human beings during a large outbreak in commercial poultry farms in the Netherlands. *Lancet* **363**, 587-593 (2004).

194 Boender, G. J. *et al.* Risk maps for the spread of highly pathogenic avian influenza in poultry. *PLoS Comput Biol* **3**, e71-e71 (2007).

195 Bavinck, V. *et al.* The role of backyard poultry flocks in the epidemic of highly pathogenic avian influenza virus (H7N7) in the Netherlands in 2003. *Prev Vet Med* **88**, 247-254 (2009).

196 Ferguson, N. M., Donnelly, C. A. & Anderson, R. M. Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. *Nature* **413**, 542-548, doi:10.1038/35097116 (2001).

197 Bessell, P. R., Shaw, D. J., Savill, N. J. & Woolhouse, M. E. Estimating risk factors for farm-level transmission of disease: foot and mouth disease during the 2001 epidemic in Great Britain. *Epidemics* **2**, 109-115, doi:10.1016/j.epidem.2010.06.002 (2010).

198 Bos, M. E. H. *et al.* Estimating the day of highly pathogenic avian influenza (H7N7) virus introduction into a poultry flock based on mortality data. *Veterinary research* **38**, 493-504, doi:10.1051/vetres (2007).

199 Bataille, A., van der Meer, F., Stegeman, A. & Koch, G. Evolutionary analysis of inter-farm transmission dynamics in a highly pathogenic avian influenza epidemic. *PLoS pathogens* **7**, e1002094-e1002094, doi:10.1371/journal.ppat.1002094 (2011).

200 de Jong, M. C., Stegeman, A., van der Goot, J. & Koch, G. Intra- and interspecies transmission of H7N7 highly pathogenic avian influenza virus during the avian influenza epidemic in The Netherlands in 2003. *Rev Sci Tech* **28**, 333-340 (2009).

201 Spekreijse, D., Bouma, A., Stegeman, J. A., Koch, G. & de Jong, M. C. The effect of inoculation dose of a highly pathogenic avian influenza virus strain H5N1 on the infectiousness of chickens. *Vet Microbiol* **147**, 59-66, doi:10.1016/j.vetmic.2010.06.012 (2011).

202 Capua, I. *et al.* The 1999-2000 avian influenza (H7N1) epidemic in Italy: veterinary and human health implications. *Acta Trop* **83**, 7-11 (2002).

203 Hirst, M. *et al.* Novel avian influenza H7N3 strain outbreak, British Columbia. *Emerg Infect Dis* **10**, 2192-2195 (2004).

204 Suarez, D. L. *et al.* Recombination resulting in virulence shift in avian influenza outbreak, Chile. *Emerg Infect Dis* **10**, 693-699 (2004).

205 World Health Organization. Cumulative number of confirmed human cases of avian influenza A/(H5N1) reported to WHO, confirmed human cases of avian influenza A(H5N1). (2012).

206 Palese, P. Influenza: old and new threats. *Nat Med* **10**, S82-87 (2004).

207 Watanabe, Y., Ibrahim, M. S., Suzuki, Y. & Ikuta, K. The changing nature of avian influenza A virus (H5N1). *Trends Microbiol* **20**, 11-20 (2012).

208 Jonges, M. *et al.* Comparative Analysis of Avian Influenza Virus Diversity in Poultry and Humans during a Highly Pathogenic Avian Influenza A (H7N7) Virus Outbreak. *J Virol* **85**, 10598-10604 (2011).

209 Cambra-Lopez, M., Aarnink, A. J., Zhao, Y., Calvet, S. & Torres, A. G. Airborne particulate matter from livestock production systems: a review of an air pollution problem. *Environ Pollut* **158**, 1-17 (2010).

210 Takai, H. *et al.* Concentrations and emissions of airborne dust in livestock buildings in Northern Europe. *Journal of Agricultural Engineering Research* **10**, 59-77 (1998).

211 Sedlmaier, N. *et al.* Generation of avian influenza virus (AIV) contaminated fecal fine particulate matter (PM(2.5)): genome and infectivity detection and calculation of immission. *Vet Microbiol* **139**, 156-164 (2009).

212 Power, C. A. An Investigation into the Potential Role of Aerosol Dispersion of Dust from Poultry Barns as a Mode of Disease Transmission during an Outbreak of Avian Influenza (H7:N3) in Abbotsford, BC in 2004 *Bulletin of the Aquaculture Association of Canada* **105**, 7-14 (2005).

213 Davis, J., Garner, M. G. & East, I. J. Analysis of local spread of equine influenza in the Park Ridge region of Queensland. *Transbound Emerg Dis* **56**, 31-38 (2009).

214 Spekreijse, D., Bouma, A., Koch, G. & Stegeman, J. A. Airborne transmission of a highly pathogenic avian influenza virus strain H5N1 between groups of chickens quantified in an experimental setting. *Vet Microbiol* **152**, 88-95 (2011).

215 Te Beest, D. E., Stegeman, J. A., Mulder, Y. M., van Boven, M. & Koopmans, M. P. Exposure of Uninfected Poultry Farms to HPAI (H7N7) Virus by Professionals During Outbreak Control Activities. *Zoonoses Public Health* **58**, 493-499 (2011).

216 Te Beest, D. E., Hagenaars, T. J., Stegeman, J. A., Koopmans, M. P. & van Boven, M. Risk based culling for highly infectious diseases of livestock. *Vet Res* **42**, 81 (2011).

217 Fisher, N. I. & Lee, A. J. A Correlation Coefficient for Circular Data. *Biometrika* **70**, 327-332 (1983).

218 Ssematimba, A., Hagenaars, T. J. & De Jong, M. C. M. Modelling the wind-borne spread of highly pathogenic avian influenza between farms. *PLoS ONE* **7**, e31114 (2012).

219 Aarnink, A. J. A. *et al.* in *CIGR International Symposium on 'Agricultural Technologies in a Changing Climate'.*

220 Sawabe, K. *et al.* Detection and isolation of highly pathogenic H5N1 avian influenza A viruses from blow flies collected in the vicinity of an infected poultry farm in Kyoto, Japan, 2004. *Am J Trop Med Hyg* **75**, 327-332 (2006).

221 Lawson, J. R. & Gemmell, M. A. The potential role of blowflies in the transmission of taeniid tapeworm eggs. *Parasitology* **91 (Pt 1)**, 129-143 (1985).

222 Able, K. P. Environmental influences on the orientation of free flying nocturnal bird migrants. *Animal Behaviour* **22**, 224-238 (1974).

223 Heijne, J. C. M. *et al.* Quantifying transmission of norovirus during an outbreak. *Epidemiology (Cambridge, Mass.)* **23**, 277-284, doi:10.1097/EDE.0b013e3182456ee6 (2012).

224 Hens, N., Calatayud, L., Kurkela, S., Tamme, T. & Wallinga, J. Robust reconstruction and analysis of outbreak data: influenza A(H1N1)v transmission in a school-based population. *Am J Epidemiol* **176**, 196-203, doi:10.1093/aje/kws006 (2012).

225 Ypma, R. J. F. *et al.* Genetic data provide evidence for wind-mediated transmission of highly pathogenic avian influenza. *The Journal of infectious diseases*, doi:10.1093/infdis/jis757 (2013).

226 Holmes, E. C., Nee, S., Rambaut, A., Garnett, G. P. & Harvey, P. H. Revealing the history of infectious disease epidemics through phylogenetic trees. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **349**, 33-40, doi:10.1098/rstb.1995.0088 (1995).

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

228 Pybus, O. G. *et al.* The epidemic behavior of the hepatitis C virus. *Science* **292**, 2323-2325 (2001).

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

230 Maddison, W. P. & Knowles, L. L. Inferring phylogeny despite incomplete lineage sorting. *Systematic biology* **55**, 21-30, doi:10.1080/10635150500354928 (2006).

231 Jenkins, G. M., Rambaut, A., Pybus, O. G. & Holmes, E. C. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *Journal of molecular evolution* **54**, 156-165, doi:10.1007/s00239-001-0064-3 (2002).

232 Cottam, E. M. *et al.* Molecular epidemiology of the foot-and-mouth disease virus outbreak in the United Kingdom in 2001. *J Virol* **80**, 11274-11282, doi:10.1128/JVI.01236-06 (2006).

233 Keeling, M. J. *et al.* Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science* **294**, 813-817, doi:10.1126/science.1065973 (2001).

234 Gibbens, J. C. & Wilesmith, J. W. Temporal and geographical distribution of cases of foot-and-mouth disease during the early weeks of the 2001 epidemic in Great Britain. *The Veterinary record* **151**, 407-412 (2002).

235 Charleston, B. *et al.* Relationship between clinical signs and transmission of an infectious disease and the implications for control. *Science* **332**, 726-729, doi:10.1126/science.1199884 (2011).

236 Gardy, J. L. *et al.* Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *The New England journal of medicine* **364**, 730-739, doi:10.1056/NEJMoa1003176 (2011).

237 Harris, S. R. *et al.* Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. *The Lancet Infectious Diseases*, doi:10.1016/s1473-3099(12)70268-2 (2012).

238 Αριστοτέλης. *Metaphysics VIII*. 1045a 8-10 (350 B.C.).

239 Croucher, N. J., Harris, S. R., Grad, Y. H. & Hanage, W. P. Bacterial genomes in epidemiology--present and future. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **368**, 20120202-20120202, doi:10.1098/rstb.2012.0202 (2013).

240 Walker, T. M. *et al.* Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet infectious diseases* **13**, 137-146, doi:10.1016/s1473-3099(12)70277-3 (2013).

241 Schürch, A. C. *et al.* High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. *Journal of clinical microbiology* **48**, 3403-3406, doi:10.1128/jcm.00370-10 (2010).

242 Wright, C. F. *et al.* Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. *Journal of virology* **85**, 2266-2275, doi:10.1128/jvi.01396-10 (2011).

243 Hughes, J. *et al.* Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks. *PLoS pathogens* **8**, e1003081-e1003081, doi:10.1371/journal.ppat.1003081 (2012).

244 Ledergerber, B., Overbeck, J., Egger, M. & Lüthy, R. The Swiss HIV Cohort Study: Rationale, organization and selected baseline characteristics. *Sozial- und Präventivmedizin SPM* **39**, 387-394, doi:10.1007/bf01299670 (1994).

245 Yahi, N. *et al.* Use of drug resistance sequence data for the systematic detection of non-B human immunodeficiency virus type 1 (HIV-1) subtypes: how to create a sentinel site for monitoring the genetic diversity of HIV-1 at a country scale. *The Journal of infectious diseases* **183**, 1311-1317, doi:10.1086/319859 (2001).

246 Bezemer, D. *et al.* Transmission networks of HIV-1 among men having sex with men in the Netherlands. *AIDS (London, England)* **24**, 271-282, doi:10.1097/QAD.0b013e328333ddee (2010).

247 Weimer, B. C. *100k genome project*.

248 MalariaGen. Community project on population genomics of Plasmodium falciparum. (2012).

249 Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174-180, doi:10.1038/nature09944 (2011).

250 Holmes, E. C. & Grenfell, B. T. Discovering the phylodynamics of RNA viruses. *PLoS Computational Biology* **5**, e1000505-e1000505, doi:10.1371/journal.pcbi.1000505 (2009).

251 Rump, B., Cornelis, C., Woonink, F. & Verweij, M. The need for ethical reflection on the use of molecular microbial characterisation in outbreak management. *Euro surveillance : bulletin Européen sur les maladies transmissibles = European communicable disease bulletin* **18**, 20384-20384 (2013).

252 Pybus, O. G., Fraser, C. & Rambaut, A. Evolutionary epidemiology: preparing for an age of genomic plenty. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **368**, 20120193-20120193, doi:10.1098/rstb.2012.0193 (2013).

Appendices

Contents

A A Sign of Superspreading in Tuberculosis: Highly Skewed Distribution of Genotypic Cluster Sizes 135
A.1 Likelihood derivation . . . 136
A.1.1 Final size distribution . . . 136
A.1.2 Mutation . . . 136
A.1.3 Dealing with unobserved cases . . . 137
A.1.4 Full likelihood . . . 137
A.2 Sensitivity analysis . . . 138
A.2.1 Recently arrived immigrants . . . 138
A.2.2 Extra pulmonary tuberculosis cases . . . 138
A.2.3 Homoplasy/recurrent mutation . . . 139

B Finding evidence for local transmission of contagious disease in molecular epidemiological datasets 143
B.1 Pairwise dissimilarities . . . 144
B.2 Putative transmission clusters . . . 144
B.3 Details of simulations . . . 145
B.3.1 Generating simulated datasets . . . 145
B.3.2 Final size calculations . . . 146
B.4 Additional simulation results . . . 146
B.4.1 Small distances . . . 146
B.4.2 Unobserved cases . . . 146

C Monitoring the spread of MRSA in the Netherlands: A reference laboratory perspective 153
C.1 Simulating the hospital network . . . 154

D Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data 163
D.1 Representing the transmission tree . . . 164
D.2 Testing different spatial kernels . . . 164
D.3 Latent period . . . 165

E Genetic data provide evidence for wind-mediated transmission of highly pathogenic avian influenza 171
E.1 Overview of data . . . 172
E.1.1 Epidemiological data . . . 172
E.1.2 Genetic data . . . 172
E.1.3 Meteorological data . . . 172
E.2 Inference of transmission events . . . 173

E.2.1 Likelihood given temporal data 173
E.2.2 Likelihood given genetic data 174
E.2.3 Estimation of the transmission tree 174
E.2.4 Double infections 174
E.3 Correlation between wind direction and direction of transmission 175
E.4 Quantification of the effect of wind 175
E.4.1 A heuristic approach 176
E.4.2 A mechanistic approach 176
E.5 Sensitivity analyses 179
E.5.1 Cut-off value for observed transmissions 179
E.5.2 Metereological stations 179
E.5.3 Quantifications 179

F Relating phylogenetic trees to transmission trees of infectious disease outbreaks 185
F.1 Test of estimation procedure on simulated data. 186
F.1.1 Simulating outbreaks 186
F.1.2 Simulated data 186
F.2 Application of the estimation procedure to data on foot-and-mouth disease. . . . 187
F.2.1 Data 187
F.2.2 Model 187
F.2.3 Results 188
F.3 Sampling from the joint posterior distribution using MCMC 188
F.3.1 Initial state 188
F.3.2 Update phylogenetic tree 188
F.3.3 Update transmission tree 190
F.3.4 Update infection times 191
F.3.5 Update epidemiological parameters 191
F.3.6 Update mutational parameters 191
F.3.7 Convergence 191
F.3.8 Estimation 192

References 194

A

A Sign of Superspreading in Tuberculosis: Highly Skewed Distribution of Genotypic Cluster Sizes

Rolf J.F. Ypma, Hester Korthals Altes, Dick van Soolingen, Jacco Wallinga, W. Marijn van Ballegooijen

A.1 Likelihood derivation

A.1.1 Final size distribution

We consider a branching process with a certain distribution for the number of offspring X. It is well-known that this process dies out with probability 1 if $E(X) < 1$. Let $\Pi(s) = \sum_{x=0}^{\infty} f(x)s^x$ be the probability generating function of X, we can then write $\Pi^y(s) = (\sum_{x=0}^{\infty} f(x)s^x)^y = \sum_{x=0}^{\infty} g(x)s^x$ for some function $g(x)$. Let Y denote the total number of individuals generated by the branching process, including the starting individuals. Becker [2] showed that the probability that $Y = y$, when the process started with n individuals, is given by the coefficient of s^{y-n} in $\Pi^y(s)$ multiplied by $\frac{n}{y}$:

$$P(Y = y|n) = \frac{n}{y} g(y - n)$$

This result is useful when $\Pi(s)$ has a mathematically tangible form. This is the case for the so-called power series distribution, given by

$$P(X = x) = \frac{a(x)\theta^x}{A(\theta)}$$

with $\sum a(x)\theta^x = A(\theta)$; here $\Pi(s) = \frac{A(\theta s)}{A(\theta)}$. The negative binomial distribution has two parameters, p and k, and is a special case of this distribution; $\theta = p$, $a_k(x) = \binom{x+k-1}{k-1}$ and $A_k(\theta) = (1-p)^{-k}$. Let Π_{NB} denote its generating function, we then have

$$\Pi_{NB}^y(s) = \left(\frac{A_k(sp)}{A_k(p)}\right)^y = \frac{(1-p)^{ky}}{(1-sp)^{ky}} = (1-p)^{ky} A_{ky}(sp) = (1-p)^{ky} \sum_{x \geq 0} \binom{x + yk - 1}{yk - 1} p^x s^x$$

where for the last equality we used $A_{ky}(sp) = \sum a_{ky}(x)(sp)^x$. Taking the coefficient of s^{y-n} and multiplying by $\frac{n}{y}$ gives, for $y \geq n$,

$$\begin{aligned} P(Y = y|n) &= \frac{n}{y}\binom{y - n + yk - 1}{yk - 1} p^{y-n}(1-p)^{ky} \\ &= \frac{n}{y}\binom{y + yk - n - 1}{yk - 1} \frac{R^{y-n}k^{ky}}{(k+R)^{(k+1)y-n}} \end{aligned} \tag{A.1}$$

with $R = E(X) = \frac{kp}{1-p}$. Obviously, $P(Y = y|n) = 0$ for $y < n$. Note that this distribution is only proper if $R \leq 1$. However it still holds for larger R if we define $P(Y = \infty|n) = 1 - \sum_{y \in \mathbb{N}} P(Y = y|n)$ [11].

A.1.2 Mutation

It is possible that mutation (e.g. a transposition) occurs during infection in the Netherlands, resulting in offspring belonging to a different genotypic cluster. We model this by saying that for each infection event a mutation occurs with probability p_m. Let X_{cl} denote the number of offspring of a certain individual having the same DNA fingerprint as that individual. Since each of the offspring has a different type with probability p_m, having x offspring of the same type means either having x offspring and no mutations, or $x+1$ offspring and one mutation, or $x+2$ offspring and two mutations, etc. This is the convolution of a negative binomial distribution with a binomial distribution

$$P(X_{cl} = x) = \sum_{i \geq x} P(X = i) \binom{i}{x} p_m^{i-x}(1-p_m)^x$$

which is again a negative binomial distribution with parameters $R(1-p_m)$ and k (Lemma 1).

Thus the final size of each of the clusters is described by equation (A.1), with R replaced by $R(1-p_m)$. This means it will be impossible to estimate R and p_m separately from the dataset. Note that, although the value of the dispersion parameter k does not change, the value of the variance-to-mean ratio $1+\frac{R}{k}$ does.

Lemma 1. *The convolution of a negative binomial distribution with a binomial distribution is again a negative binomial distribution.*

If we write $A^{(x)}(\theta)$ for the x-th derivative of $A(\theta)$, we have that for the negative binomial distribution

$$\begin{aligned}
P(X_{cl}=x) &= \sum_{i\geq x} P(X=i)\binom{i}{x} p_m^{i-x}(1-p_m)^x \\
&= \sum_{i=x}^{\infty}\binom{i+k-1}{k-1}(\frac{k}{k+R})^k(\frac{R}{k+R})^i\binom{i}{x}p_m^{i-x}(1-p_m)^x \\
&= \sum_{i=x}^{\infty}\binom{i+k-1}{k-1}(1-p)^k p^i\binom{i}{x}p_m^{i-x}(1-p_m)^x \\
&= (1-p_m)^x p^x\frac{(1-p)^k}{x!}\sum_{i=x}^{\infty}\binom{i+k-1}{k-1}\frac{(p_m p)^{i-x}}{(i-x)!}i! \\
&= (1-p_m)^x p^x\frac{(1-p)^k}{x!}A^{(x)}(p_m p) \\
&= (1-p_m)^x p^x\frac{(1-p)^k}{x!}\frac{\Gamma(x+k-1)}{\Gamma(k-1)}(1-p_m p)^{-k-x} \\
&= \binom{x+k-1}{k-1}(\frac{1-p}{1-p_m p})^k(\frac{p-p_m p}{1-p_m p})^x \\
&= \binom{x+k-1}{k-1}(\frac{k}{k+R(1-p_m)})^k(\frac{R(1-p_m)}{k+R(1-p_m)})^x
\end{aligned}$$

which is just the negative binomial distribution again, but now with mean $E(X_{cl})=R(1-p_m)$. ☐

A.1.3 Dealing with unobserved cases

We assume each individual which is not an index case is seen with probability p_{seen}. The probability distribution for the observed cluster size Y_{obs} given n index cases becomes

$$P(Y_{obs}=y|n)=\frac{1}{\sum_{i=1}^{\infty}P(Y_{cl}=i|n)(1-p_{seen})^i}\sum_{i=y}^{\infty}P(Y_{cl}=i|n)\left(\begin{array}{c} i \\ y \end{array}\right)p_{seen}^y(1-p_{seen})^{i-y}$$

for $y>0$, since clusters of size 0 are unobserved. It would be more elegant to allow for unobserved index cases as well, however this would need additional assumptions on the distribution of index cases over clusters. This is not straightforward, particulary because these additional index cases in our study are immigrants, typically from a few countries with high prevalence of tuberculosis; their DNA fingerprints are highly correlated.

A.1.4 Full likelihood

The likelihood of parameters R, k, p_m when $a_{y,n}$ clusters of size y with n index cases have been fully observed, and $b_{y,n}$ censored clusters have been observed to have at least size y and n index cases is

$$
\begin{aligned}
L(R,k,p_m|\mathbf{a},\mathbf{b}) &= \textstyle\prod_y \prod_n P(Y_{obs}=y|n)^{a_{y,n}} P(Y_{obs} \geq y|n)^{b_{y,n}} \\
&= \textstyle\prod_y \prod_n \left(\dfrac{\sum_{i=y}^{\infty} P(Y_{cl}=i|n) \begin{pmatrix} i \\ y \end{pmatrix} p_{seen}^{y}(1-p_{seen})^{i-y}}{\sum_{i=1}^{\infty} P(Y_{cl}=i|n)(1-p_{seen})^{i}} \right)^{a_{y,n}} \\
& \textstyle\prod_y \prod_n \left(1 - \sum_{j=1}^{y-1} \left[\dfrac{\sum_{i=j}^{\infty} P(Y_{cl}=i|n) \begin{pmatrix} i \\ j \end{pmatrix} p_{seen}^{j}(1-p_{seen})^{i-j}}{\sum_{i=1}^{\infty} P(Y_{cl}=i|n)(1-p_{seen})^{i}} \right] \right)^{b_{y,n}}
\end{aligned}
$$

and

$$
P(Y_{cl}=i|n) = \begin{cases} \dfrac{n}{i}\begin{pmatrix} i+ik-n-1 \\ ik-1 \end{pmatrix} \dfrac{(R(1-p_m)^{i-n}k^{ki}}{(k+R(1-p_m))^{(k+1)i-n}} & \text{if } n \leq i \\ 0 & \text{if } n > i \end{cases}
$$

A.2 Sensitivity analysis

To construct our model we made several assumptions regarding TB epidemiology, here we will explore the sensitivity of our results to these assumptions.

A.2.1 Recently arrived immigrants

In our model we assumed that the 933 immigrants that had been in the Netherlands for less than six months at date of diagnosis were infected elsewhere. We consider three alternatives to this assumption, and estimate the parameters R_m and k under these alternative assumptions.

First, we assume all recently arrived immigrants were infected in the Netherlands, and treat them the same as all other cases in our dataset. We would expect that this assumption slightly increases our estimate of R_m, as more infections are needed to explain all cases. Indeed, the assumption leads to estimates of $R_m = 0.5, k = 0.11$, where our original estimates were $R_m = 0.48, k = 0.10$. The slightly larger value of k is probably due to an increase in the number of 'average sized' clusters (i.e. size 2,3,4,...), which point at moderate values of k. The increase in k however, is minor.

Second, we exclude all transmission clusters that contain a recently arrived immigrant from our dataset. As larger clusters are more likely to contain such an immigrant, we expect to find a lower value of R_m. Indeed, this assumption leads to estimates of $R_m = 0.39, k = 0.10$. Note that the estimated value for k is identical to our original estimate.

Third, again assuming that the recently arrived immigrants were infected abroad, we include any additional cases abroad that never immigrated to the Netherlands by regarding all clusters containing a recently arrived immigrant as censored. Our interpretation of the estimated parameters would then change slightly, as infections abroad are now also included. With additional censoring, we find estimates of $R_m = 0.61, k = 0.09$.

Summarizing, we find that different assumptions regarding the role of recently arrived immigrants lead to different estimations of R_m, but nearly identical estimates of k.

A.2.2 Extra pulmonary tuberculosis cases

In our model we assumed extra pulmonary tuberculosis (EPTB) cases were non-infectious, and discarded them from the analysis. Thus the parameters estimated in the main text describe the number of pulmonary cases caused by one pulmonary case. It is possible that some infections have been caused by (possibly misdiagnosed) EPTB cases. We consider the extreme scenario in which EPTB cases are as infectious as pulmonary TB cases.

We added the 3892 EPTB cases to our dataset and analysed the full dataset of 12222 cases. We found 8221 clusters, 7070 (86%) of which consisted of only one case. We find estimates of the parameters of $R_m = 0.50, k = 0.094$. The results appears to be insensitive to the assumption regarding EPTB cases.

A.2.3 Homoplasy/recurrent mutation

In our model we assumed that all RFLP types generated through a mutation are unique, i.e. lead to new clusters. However, it is possible that several mutation events lead to the same RFLP type, either through homoplasy or recurrent mutations. The net result of such mutations would be one large genotypic cluster, rather than several small clusters. Thus any given case might not have been infected within its genotypic cluster but could be the result of transmission from a different cluster and a recurrent mutation; it is then an additional index case for its genotypic cluster.

Each cluster of size y with n_{obs} observed index cases contains $y - n_{obs}$ cases that could actually be an index case, due to the mechanism described above. Assume this happens for each case with probability p_r. Then the actual number of index cases n for a cluster of size y with n_{obs} observed index cases is distributed as $n_{obs} + n_r$, with $n_r \sim Bin(y - n_{obs}, p_r)$.

The likelihood equation in section (A.1.4) that we use to obtain estimates is based on $\{a_{y,n}\}$ and $\{b_{y,n}\}$, the number of clusters completely/partially observed having precisely/at least y cases and n index cases. Taking recurrent mutations into account, we note that some of the clusters will in reality have more index cases than observed. We define $A_{y,n}$ as the expected number of clusters of size y and n index cases, given the observed clusters. We get

$$A_{y,n} = \begin{cases} 0 & \text{if } y < n \\ a_{y,n} + \sum_{n_{obs}=1}^{n-1} a_{y,n_{obs}} \begin{pmatrix} y - n_{obs} \\ n - n_{obs} \end{pmatrix} p_r^{n-n_{obs}} (1-p_r)^{y-n} & \text{if } y = n \\ \sum_{n_{obs}=1}^{n} a_{y,n_{obs}} \begin{pmatrix} y - n_{obs} \\ n - n_{obs} \end{pmatrix} p_r^{n-n_{obs}} (1-p_r)^{y-n} & \text{if } y > n \end{cases}$$

and equivalently we define $B_{y,n}$:

$$B_{y,n} = \begin{cases} 0 & \text{if } y < n \\ b_{y,n} + \sum_{i=1}^{n-1} b_{y,n_{obs}} \begin{pmatrix} y - n_{obs} \\ n - n_{obs} \end{pmatrix} p_r^{n-n_{obs}} (1-p_r)^{y-n} & \text{if } y = n \\ \sum_{i=1}^{n} b_{y,n_{obs}} \begin{pmatrix} y - n_{obs} \\ n - n_{obs} \end{pmatrix} p_r^{n-n_{obs}} (1-p_r)^{y-n} & \text{if } y > n \end{cases}$$

We then replace $a_{y,n}$ and $b_{y,n}$ in the likelihood equations of section (A.1.4) by $A_{y,n}$ and $B_{y,n}$ to obtain estimates for R_m and k. For $p_r \in [0, 0.5]$ the adjusted likelihood equation yields estimates of R_m in $[0.40, 0.48]$ and k in $[0.10, 0.052]$ (figure S1). This means the method is insensitive to recurrent mutations.

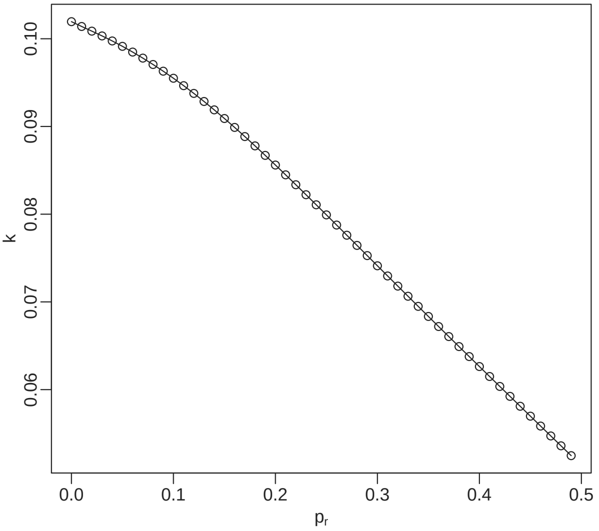


Figure S1. Estimates of the dispersion parameter k under different values for p_r, the probability that a non-index case in an genotypic cluster was infected by a case with a different RFLP type. In the main text $p_r = 0$, yielding an estimate of $k = 0.10$. For higher values of p_r, i.e. higher rates of recurrent mutations, the estimate of k decreases. This shows the initial estimate of k was conservative.

B

Finding evidence for local transmission of contagious disease in molecular epidemiological datasets

Rolf J.F. Ypma, Tjibbe Donker, W. Marijn van Ballegooijen, Jacco Wallinga

B.1 Pairwise dissimilarities

Let $D_i^m(a,b)$ be the measured distance between two cases a and b for data type i. There could be identical values in our dataset (i.e. $D_i^m(a,b) = 0$). As we use ordinal distances to detect cases lying close together, detection of local transmission clusters will be more challenging when many values are identical; no ordering exists on these. To be able to make comparison between cases with identical values and those with distinct values, we will assume that for all cases for which the same value was measured, the actual value lies a random infinitesimal distance away from this measured value. This is actually true for the temporal data, which is always interval censored, as dates but not exact times are given. It is not true for genetic data, which are discrete. However, these can be seen as a proxy for evolutionary time separating two samples, which is again continuous.

We define the dissimilarity $d_i(a,b)$ between two cases a and b as the expected number of cases between them, plus one:

$$\begin{aligned} d_i(a,b) &= |\{p : D_i^m(a,p) < D_i^m(a,b) \wedge D_i^m(b,p) < D_i^m(b,a)\}| \\ &\qquad + \tfrac{|\{p:D_i^m(a,p)=0\}-1| + |\{p:D_i^m(b,p)=0\}-1|}{2} + 1 \\ &= |\{p : D_i^m(a,p) < D_i^m(a,b) \wedge D_i^m(b,p) < D_i^m(b,a)\}| \\ &\qquad + \tfrac{|\{p:D_i^m(a,p)=0\}| + |\{p:D_i^m(b,p)=0\}|}{2} \end{aligned}$$

when $D_i^m(a,b) \neq 0$, and

$$d_i(a,b) = \frac{|\{p : D_i^m(a,p) = 0\}| - 2}{3} + 1 = \frac{|\{p : D_i^m(a,p) = 0\}| + 1}{3}$$

when $D_i^m(a,b) = 0$. Here '$\wedge$' denotes the logical AND operator; $A \wedge B$ is true if and only if both A and B are true. To see that the definition above coincides with the expected value plus one, consider three points a, b and c, with a and b having the same observed value $()D_i^m(a,b) = 0)$, each lying a random infinitesimal distance away from their measured values. Then the probability that any two of the actual pairwise distances are equal is zero. Therefore, if $D_i^m(a,c) \neq 0$, the probability that b is in between a and c is $\frac{1}{2}$. If $D_i^m(a,c) = 0$, the probability that b is in between a and c is $\frac{1}{3}$. Further note that, for distinct cases, when identical values do not occur in our dataset the definition above is equivalent to equation (1) in the main text.

As in the main text, the full dissimilarity between two cases a and b is given by

$$d(a,b) = \Pi_i d_i(a,b)$$

B.2 Putative transmission clusters

For any subset $S \subseteq D$, define $l(S)$ as the largest dissimilarity in the minimum spanning tree of S. Note that several minimum spanning trees can exist, but $l(S)$ is unique (see lemma 2). To test the null hypothesis of independence between data types, we construct the set D' from D by randomly permuting the values of the data types. D' is identical to D for each of the data types, but satisfies the null hypothesis. We then define the p-value for S as the probability that a subset with at least that size and at most that largest dissimilarity exists under the null hypothesis:

$$P(\exists S' \subseteq D' : |S'| \geq |S|, l(S') \leq l(S))$$

and we call S a putative transmission cluster (PTC) if this p-value is beneath a threshold of 0.001.

We can limit the number of clusters we have to test using hierarchical clustering, a technique that yields a dendogram of the dataset. A dendogram can be defined as a function $h : [0, \infty) \rightarrow$

P_D, where P_D is the set of all partitions of the dataset D, with the properties that $m \leq m'$ implies $h(m) \leq h(m')$ (i.e. every element of $h(m)$ is a subset of an element of $h(m')$), and h is eventually the whole dataset ($h(m) = D$ for sufficiently large m). $h(m)$ here is the set of subsets S of D such that $l(S) \leq m$ and the only set S_2 that contains S and has $l(S_2) \leq m$ is S itself. Let $\mathcal{S}$ be the set of subsets of D that are in $h(m)$ for some m. By lemma 3, subsets of D that are a PTC are always contained in an element of $\mathcal{S}$ which is also a PTC. Since we are interested in whether cases belong to a cluster or not, we only have to test the elements of $\mathcal{S}$ for being a PTC.

Lemma 2. *For any weighted graph G, all minimal spanning trees have the same maximum edge weight.*

To prove this, let's assume T_1 and T_2 are minimal spanning trees of G, such that their maximum edge weights are different. Without loss of generality, let the maximum edge weight of T_1 be larger, and let $e \in T_1$ be an edge with this weight. Now select an edge e' from T_2 such that e' is in the cut induced by e in T_1. As the maximum edge weight of T_2 is smaller than that of T_1 by assumption, the weight of e' is smaller than that of e. The tree $(T_1 - \{e\}) \cup e'$ is a spanning tree of G, with total weight less than T_1. This is a contradiction, as T_1 was a minimal spanning tree. □

Lemma 3. *If $S \subseteq D$ is a putative transmission cluster (PTC), $\exists T \in \mathcal{S}$ with $S \subseteq T$ and T a PTC.*

Either $S \in h(l(S)) \subseteq \mathcal{S}$ and we are done, or $\exists T \in h(l(S)) \subseteq \mathcal{S}$, with $S \subset T$. Because S is a PTC we have

$$P(\exists S' \subset D' : |S'| \geq |S|, l(S') \leq l(S)) < 0.001$$

since furthermore $|T| > |S|$, $l(T) = l(S)$ and l is monotonically increasing in cluster size, we have that

$$\begin{aligned} P(\exists S' \subset D' : |S'| \geq |T| > |S|, l(S') \leq l(T) = l(S)) &<= \\ P(\exists S' \subset D' : |S'| \geq |S|, l(S') \leq l(S)) &< 0.001 \end{aligned}$$

which shows that T is also a PTC. □

B.3 Details of simulations

B.3.1 Generating simulated datasets

In our first simulation scenario, all locally infected cases belong to one large outbreak. One index case was generated near the start of the study period so the outbreak would be completed within the time window. As the variance in the final size of a large outbreak generated by branching processes is quite large, we restricted this outbreak to be of size exactly one tenth of the size of the total simulated dataset. We therefore generated cases until the number was reached, and picked an infector for each from the set of previously generated cases. As assigning cases randomly from the set of already generated cases would amount to strong superspreading behavior (the index case would get a large number of infectees assigned), we preferentially picked more recently generated cases. In particular, we set the probability for any generated case to be assigned as an infector as twice the probability of picking the case generated before. For example, when three cases had been generated, they would be picked as an infector with probabilities 1/7, 2/7 and 4/7. This procedure is arbitrary, but simple and keeps the expected number of infections per infected individual bounded. For example, the expected number of infections caused by the index case would be $\sum_{i=1}^{N} \frac{1}{2^i - 1} \approx 1.61$.

The small and very small outbreaks were generated using branching processes, as explained in the main text. Below we calculate the expected size of the outbreaks generated in this way.

B.3.2 Final size calculations

To find the expected value of the final size S of the outbreaks in the second and third scenario, let $f(x)$ be the probability that one infectious case infects x others. For the geometric distribution we use, $f(x) = p^x(1-p)$, where $p = \frac{R}{1+R}$ and R is the expected number of infections per infectious case. As each case infected again infects new cases, we have

$$E(S) = 1 + \sum_{x=0}^{\infty} f(x)xE(S) = 1 + RE(S)$$

which simplifies to $E(S) = \frac{1}{1-R}$, yielding expected sizes 2 and $\frac{10}{9}$ for the R values of 0.5 and 0.1 used.

As only outbreaks of value of at least 2 are characterized as clusters in our analysis, we might also want to find the expected size of these clusters: $E(S|S > 1)$. This is equivalent to conditioning on the index case causing at least one infection. We get

$$E(S|S > 1) = 1 + \sum_{x=1}^{\infty} \frac{f(x)}{1-f(0)}xE(S) = 1 + \frac{R}{1-f(0)}E(S) = 1 + \frac{1+R}{1-R}$$

yielding 4 and 20/9≈2.22 for the two scenarios.

B.4 Additional simulation results

In this section we give the results obtained by applying the proposed method to simulations where the absolute distances between infector-infected pairs are smaller than in the main text, and to simulations where 20% of cases are unobserved.

B.4.1 Small distances

The statistical signal left by clusters of cases depends for a large part on the relation for each of the data types between infector-infected pairs. When distances in these data types are smaller, the statistical signal is stronger. To illustrate this, we performed additional simulations in which these distances are smaller. We simulate as described in the main text, but the time distance is now exponentially distributed with expectation 0.5 (1 in the main text), the geographical distance is $N(0, 2)$ ($N(0, 4)$ in the main text), and the expected number of mutations is now 0.1 (0.5 in the main text). Clustering performance is given in figure S2 and table S1. As the statistical signal is much stronger, the distinction between outbreak and unrelated cases is much clearer than in the main text.

B.4.2 Unobserved cases

Many datasets face the problem of missing or unobserved cases. Here, we tested the performance of our method when facing unobserved cases. We do this by performing simulations as described in the main text, and then separately discarding each of the cases with probability 0.2, thus discarding 20% of cases at random. We applied our method to this reduced datasets, results are given in figure S3 and table S2. Datasets with missing cases are similar to complete datasets with larger distances between the cases; thus this scenario constitutes the opposite of the one in the previous section. As expected, clustering performance decreases for most scenarios. A notable exception are the very small clusters, where sensitivity actually increases. As these transmission clusters are mainly of size two, discarding a case does not lead to larger distances, but to elimination of the cluster. Thus the number of cases and transmission clusters is affected, but not the intra-cluster distances. Outbreak cases and unrelated cases can be distinguished for all scenarios, showing that the method can provide useful results even when cases are unobserved.

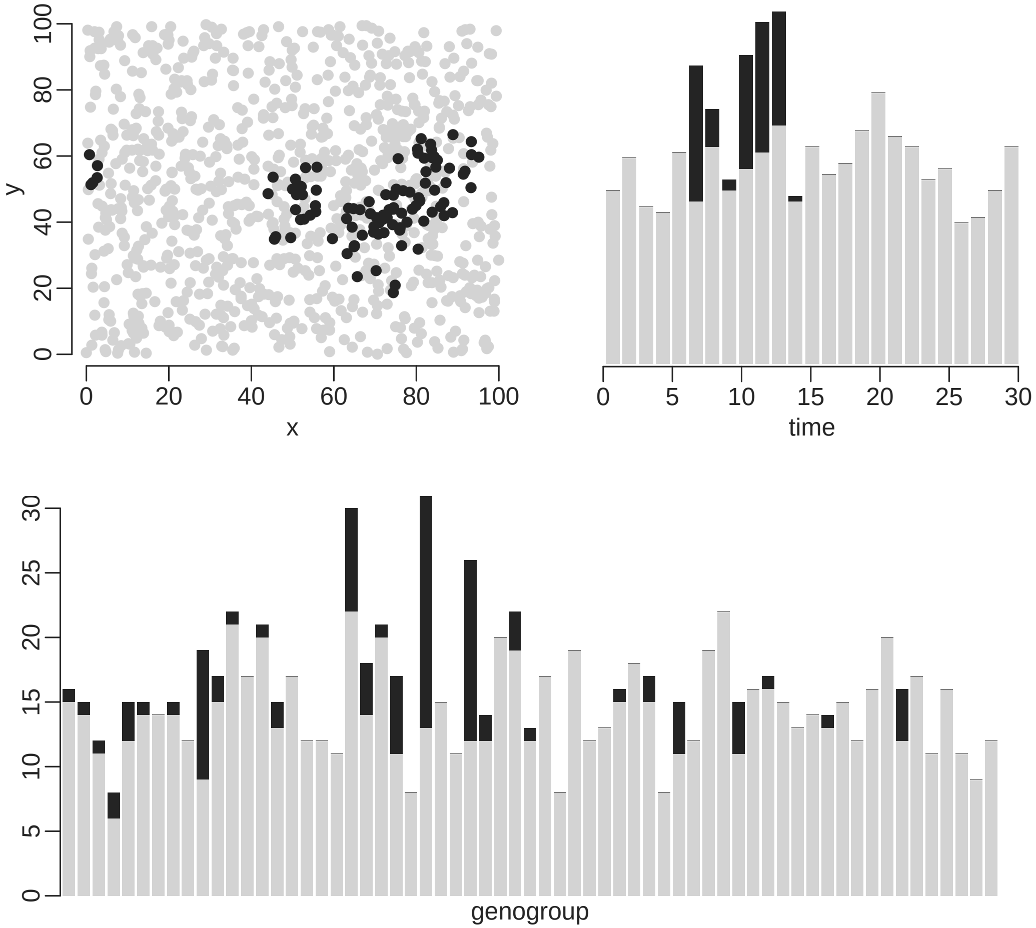


Figure S1. Graphical representation of the three data types for a typical simulation containing one large outbreak. This simulation consisted of 1000 cases of which 10% pertained to one large outbreak (black). (top left) Geographical location of all simulated cases. The geography is a torus, so the right side is equated with the left side, and the top side is equated with the bottom side. (top right) Simulated cases over time. (bottom) Simulated cases have one of 28=256 possible genotypes. For clarity, the distribution of cases over 64 genogroups is plotted; a genogroup is defined as a set of four genotypes that are identical up to the last two digits. The order of the genogroups on the x-axis does not reflect genetic distance.

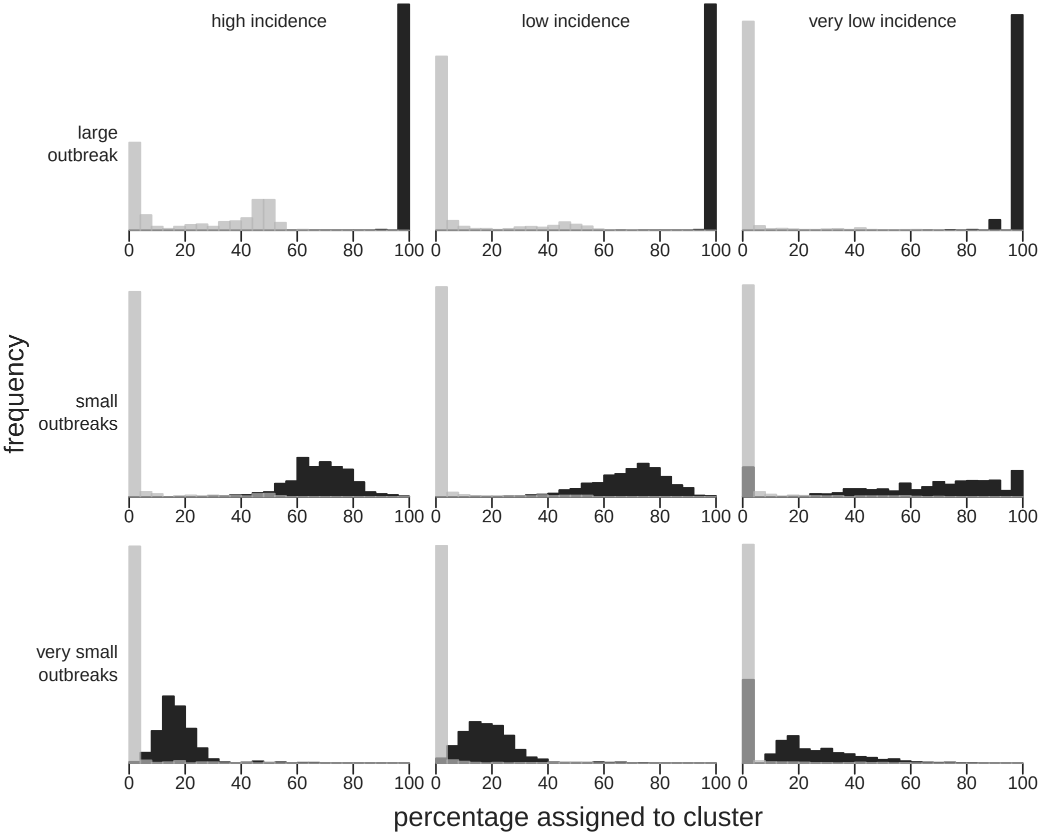


Figure S2. Sensitivity (black) and false positive rate (gray) when distances between pairs of infector and infected are small. Percentage of (black) outbreak and (gray) non-outbreak cases assigned to putative transmission clusters for simulations under nine different scenarios, when the distance between a locally infected case and its infector is smaller than in the simulations given in the main text. In each scenario, ten percent of all cases is an outbreak case. Total expected number of cases is (left column) 1000, (middle column) 500 or (right column) 100. Outbreak cases belong to (top row) one large outbreak, (middle row) small outbreaks caused by 1/10 of cases being infectious with R_0=0.5, (bottom row) minor outbreaks caused by all cases being infectious with R_0=0.1. Sensitivity and specificity increase with respect to simulations in the main text, as smaller distances lead to a stronger statistical signal.

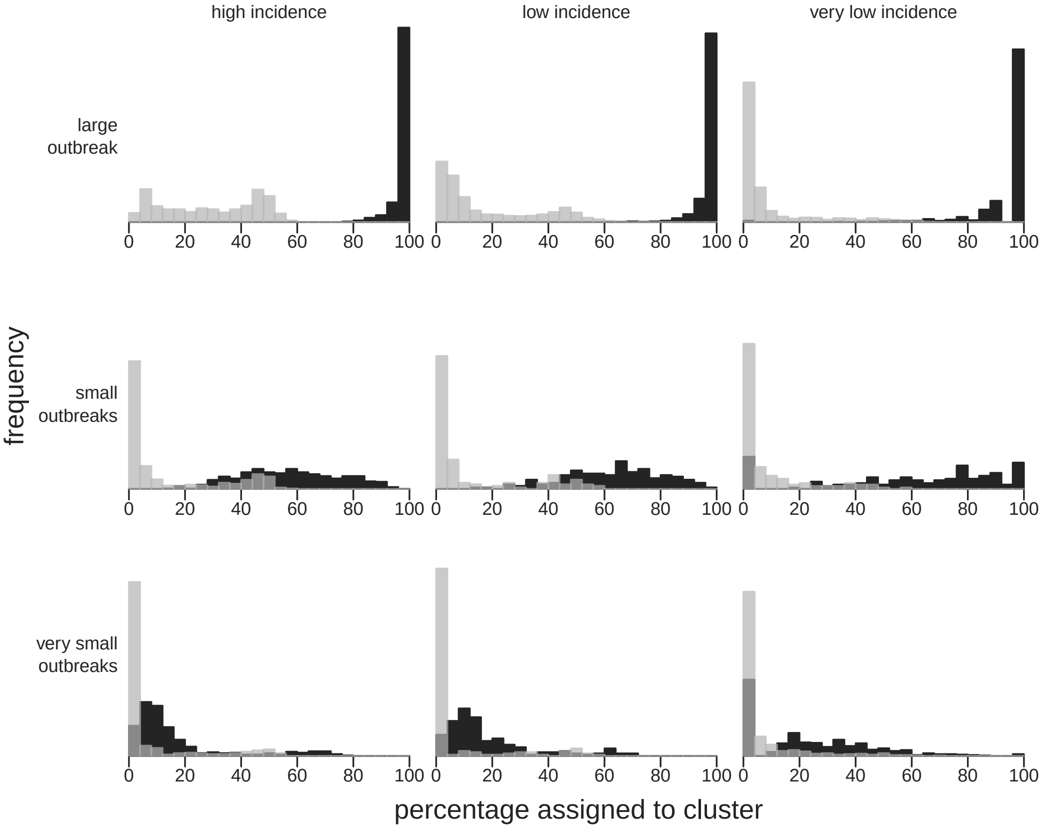


Figure S3. Sensitivity (black) and false positive rate (gray) when 20% of cases are unobserved. Percentage of (black) outbreak and (gray) non-outbreak cases assigned to putative transmission clusters for simulations under nine different scenarios, when 20 % of cases is unobserved. In each scenario, ten percent of all cases is an outbreak case. Total expected number of cases is (left column) 1000, (middle column) 500 or (right column) 100. Outbreak cases belong to (top row) one large outbreak, (middle row) small outbreaks caused by 1/10 of cases being infectious with R_0=0.5, (bottom row) minor outbreaks caused by all cases being infectious with R_0=0.1. As expected, performance decreases when the distance between cases increases. A notable exception are the very small clusters, where sensitivity actually increases. As these transmission clusters are mainly of size two, discarding a case does not lead to larger distances, but to elimination of the cluster. Thus the number of cases and clusters is affected, but the intra-cluster distances are not.

Table S1. Median of sensitivity/false positive rate of assigning locally infected cases to a putative transmission cluster for simulated datasets where distances for infector-infected pairs are small.

	high incidence	**low incidence**	**very low incidence**
large cluster	1/0.22	1/0.01	1/0
small clusters	0.68/0.01	0.71/0	0.71/0
very small cluster	0.16/0	0.18/0	0.16/0

Table S2. Median of sensitivity/false positive rate of assigning locally infected cases to a putative transmission cluster for simulated datasets where 20% of cases are unobserved.

	high incidence	**low incidence**	**very low incidence**
large cluster	1/0.3	1/0.09	1/0.03
small clusters	0.57/0.03	0.64/0.03	0.69/0.01
very small cluster	0.10/0	0.13/0	0.2/0

C

Monitoring the spread of MRSA in the Netherlands: A reference laboratory perspective

T. Donker, T. Bosch, R.J.F. Ypma, A. Haenen, W.M. van Ballegooijen, L. Schouls,
J. Wallinga, H. Grundmann

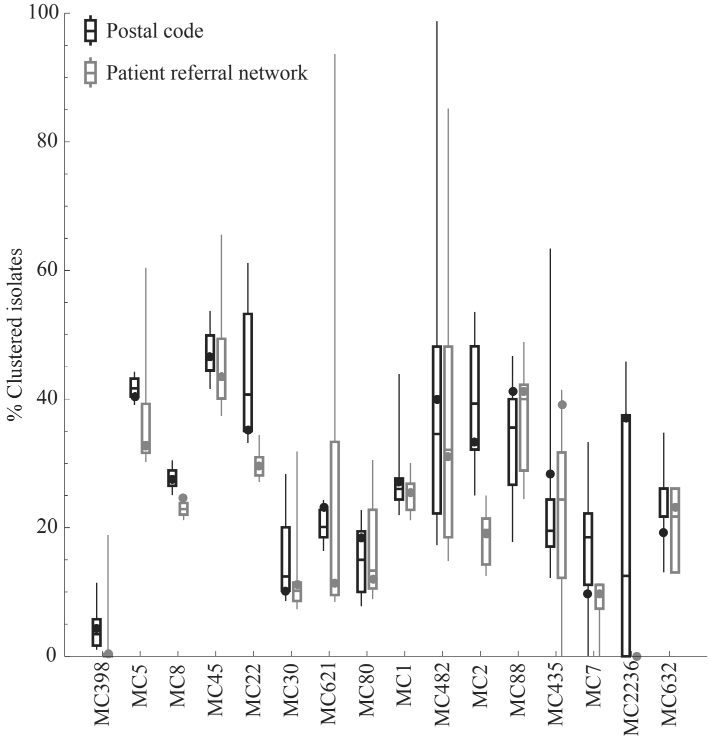


Figure S1. The proportion of isolates part of a cluster according to the algorithm. The results of the analysis using the postal code are shown in black, results from the analysis using the patient referral network in grey. Using the network as distance measure generally delivers less and smaller clusters.

C.1 Simulating the hospital network

In order to measure the distance between hospitals, we simulated the national patient referral network of the Netherlands. The general structure of this network is known, however, the exact position of each of the hospitals is not known, because they are only known by their semi-anonymous identifier. We therefore calculated the pairwise distance between the coordinates of all hospitals, measured on the Dutch national grid. These distances (D_{ij}) were then used to calculate the link strength between hospitals:

$$L_{ij} = \frac{w_{ij}}{D_{ij}}$$

Where wij is a weight factor (See table S1) to compensate for the disassortative mixing between hospital categories. The resulting metric is similar to the patient flow between hospitals, a high value indicates hospitals close together, a low value hospitals further apart. It is, however, not a direct reflection of the number of exchanged patients, but the resulting network has a similar structure to the original measured one (Figure S2).

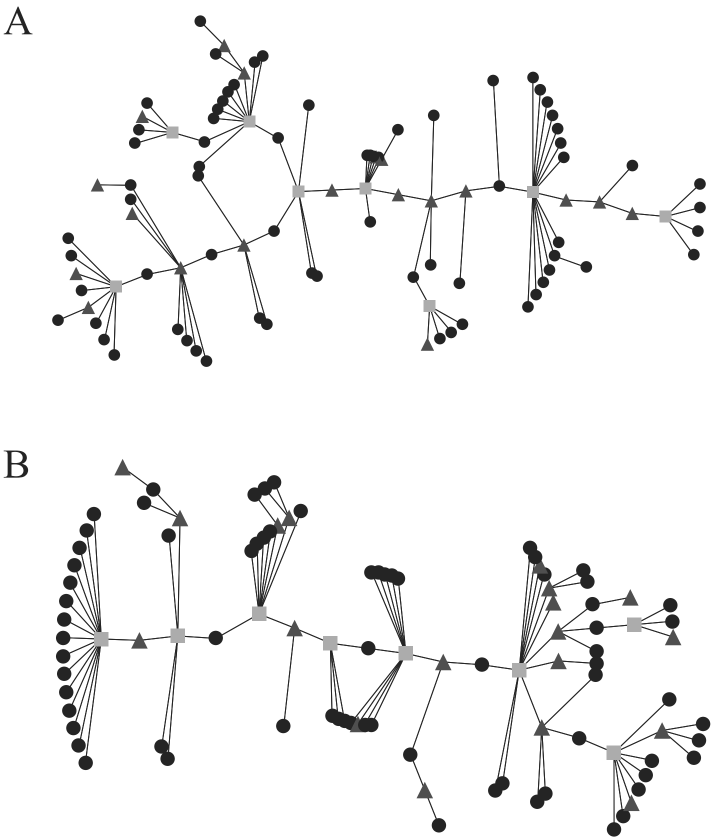


Figure S2. The structure of the patient referral network shown as a minimum spanning tree. A) The patient referral network constructed from the Dutch Medical Registry 2004. B) The network reconstructed using geographical distances between hospitals, weighted by a hospital category-specific factor. Squares denote university hospitals, triangles teaching hospitals, and black dots general hospitals. The structure of both networks is more or less the same, with groups of general hospitals around university hospitals, and a back-bone structure consisting of mostly University and teaching hospitals.

Table S1. To simulate the patient referral network, the geographical distance between hospitals was compensated by a weight depending on the types of hospitals, to account for the disassortative mixing between hospitals in the network (i.e. general hospitals tend to refer to teaching or university hospitals, not to other general hospitals).

	General Hospital	Teaching Hospital	University Hospital
General Hospital	1	3	5
Teaching Hospital	3	1	7
University Hospital	5	7	1

Table S2a. The number of isolates, split up by MLVA clonal complex (MC) and epidemiological information, such as the risk groups defined by the Workgroup Infection Prevention (WIP).

	All*	MC398*	MC5	MC8	MC45	MC22	MC30	MC621	MC80
All	6295	1195	1416	1292	611	549	349	211	201
WIP Introductions	1064	507	158	122	49	86	26	41	16
Unexpected Cases	1405	72	321	384	139	118	128	44	69
No Information	1561	362	338	305	114	120	92	56	66
WIP Ambiguous	153	19	48	30	12	13	12	5	6
Expected Cases	754	201	146	127	59	62	40	37	25
WIP Transmission	540	16	213	115	67	69	13	7	1
Contact Tracing	818	18	192	209	171	81	38	21	18
Contact with Farm Animals	*543*	*497*	*12*	*8*	*7*	*4*	*4*	*4*	*3*
Admitted from foreign hospital	*389*	*7*	*109*	*87*	*37*	*80*	*19*	*6*	*13*
Adopted child	*112*	*2*	*29*	*25*	*2*	*2*	*2*	*31*	*0*
¿2 months ago in foreign hospital	*22*	*1*	*11*	*2*	*3*	*0*	*1*	*0*	*0*
Foreign dialysis patient	*3*	*0*	*2*	*0*	*0*	*0*	*0*	*0*	*0*
Known MRSA carrier	*149*	*20*	*46*	*29*	*11*	*12*	*13*	*5*	*6*
Known HCW MRSA carrier	*5*	*0*	*0*	*2*	*1*	*1*	*0*	*0*	*0*
Protected contact with MRSA carrier	*21*	*2*	*4*	*6*	*1*	*7*	*0*	*0*	*0*
Unprotected contact with MRSA carrier	*194*	*5*	*77*	*34*	*30*	*21*	*10*	*2*	*1*
Admitted to hospital with known MRSA problem	*178*	*5*	*81*	*48*	*14*	*18*	*2*	*1*	*0*
Shared room with MRSA patient	*152*	*4*	*55*	*27*	*22*	*23*	*1*	*4*	*0*

Table S2b. The number of isolates, split up by MLVA clonal complex (MC) and epidemiological information, such as the risk groups defined by the Workgroup Infection Prevention (WIP).

	All*	MC1	MC482	MC2	MC88	MC435	MC7	MC2236	MC632
All	6295	137	90	63	51	46	31	27	26
WIP Introductions	1064	20	7	12	4	3	9	2	2
Unexpected Cases	1405	38	26	6	18	17	7	12	6
No Information	1561	36	28	4	10	13	8	2	7
WIP Ambiguous	153	2	1	0	1	1	0	2	1
Expected Cases	754	17	15	3	9	8	2	2	1
WIP Transmission	540	8	2	20	2	1	2	0	4
Contact Tracing	818	16	11	18	7	3	3	7	5
Contact with Farm Animals	*543*	*1*	*2*	*0*	*0*	*0*	*0*	*1*	*0*
Admitted from foreign hospital	*389*	*9*	*3*	*12*	*3*	*3*	*0*	*0*	*1*
Adopted child	*112*	*9*	*0*	*0*	*1*	*0*	*9*	*0*	*0*
¿2 months ago in foreign hospital	*22*	*1*	*1*	*0*	*0*	*0*	*0*	*1*	*1*
Foreign dialysis patient	*3*	*0*	*1*	*0*	*0*	*0*	*0*	*0*	*0*
Known MRSA carrier	*149*	*2*	*1*	*0*	*1*	*0*	*0*	*2*	*1*
Known HCW MRSA carrier	*5*	*0*	*0*	*0*	*0*	*1*	*0*	*0*	*0*
Protected contact with MRSA carrier	*21*	*0*	*0*	*1*	*0*	*0*	*0*	*0*	*0*
Unprotected contact with MRSA carrier	*194*	*2*	*0*	*5*	*2*	*1*	*0*	*0*	*4*
Admitted to hospital with known MRSA problem	*178*	*5*	*1*	*3*	*0*	*0*	*0*	*0*	*0*
Shared room with MRSA patient	*152*	*1*	*1*	*12*	*0*	*0*	*2*	*0*	*0*

Table S3a. The number of isolates assigned to a cluster using the postal code of the patient's residential address as location data, split up by MLVA clonal complex (MC) and epidemiological information, such as the risk groups defined by the Workgroup Infection Prevention (WIP). *MC398 only includes isolates from 2009, all others 2006-2009. "All" denotes the sum of all included isolates.

	All*	MC398*	MC5	MC8	MC45	MC22	MC30	MC621	MC80
All	1724	52	571	356	285	193	35	49	37
WIP Introductions	94	33	23	12	8	9	1	3	0
Unexpected Cases	276	2	77	69	43	22	6	9	13
No Information	303	8	117	40	51	33	3	7	15
WIP Ambiguous	34	1	17	5	5	2	0	1	1
Expected Cases	184	6	53	35	28	14	9	12	3
WIP Transmission	354	1	151	77	39	57	4	7	0
Contact Tracing	479	1	133	118	111	56	12	10	5
Contact with Farm Animals	*46*	*33*	*5*	*3*	*3*	*0*	*0*	*2*	*0*
Admitted from foreign hospital	*31*	*0*	*12*	*4*	*4*	*9*	*1*	*0*	*0*
Adopted child	*11*	*0*	*4*	*5*	*0*	*0*	*0*	*1*	*0*
¿2 months ago in foreign hospital	*7*	*0*	*4*	*0*	*1*	*0*	*0*	*0*	*0*
Foreign dialysis patient	*2*	*0*	*1*	*0*	*0*	*0*	*0*	*0*	*0*
Known MRSA carrier	*31*	*1*	*15*	*4*	*5*	*2*	*0*	*1*	*1*
Known HCW MRSA carrier	*2*	*0*	*0*	*2*	*0*	*0*	*0*	*0*	*0*
Protected contact with MRSA carrier	*10*	*1*	*1*	*4*	*0*	*4*	*0*	*0*	*0*
Unprotected contact with MRSA carrier	*115*	*0*	*49*	*24*	*14*	*18*	*3*	*2*	*0*
Admitted to hospital with known MRSA problem	*121*	*0*	*61*	*31*	*10*	*13*	*1*	*1*	*0*
Shared room with MRSA patient	*110*	*0*	*42*	*18*	*15*	*22*	*0*	*4*	*0*

Table S3b. The number of isolates assigned to a cluster using the postal code of the patient's residential address as location data, split up by MLVA clonal complex (MC) and epidemiological information, such as the risk groups defined by the Workgroup Infection Prevention (WIP). *MC398 only includes isolates from 2009, all others 2006-2009. "All" denotes the sum of all included isolates.

	All*	MC398*	MC482	MC2	MC88	MC435	MC7	MC2236	MC632
All	1724	37	36	21	21	13	3	10	5
WIP Introductions	94	1	1	1	0	0	0	1	1
Unexpected Cases	276	9	11	0	8	0	1	6	0
No Information	303	12	6	1	2	6	0	1	1
WIP Ambiguous	34	0	1	0	1	0	0	0	0
Expected Cases	184	4	12	0	3	5	0	0	0
WIP Transmission	354	6	1	8	0	1	2	0	0
Contact Tracing	479	5	4	11	7	1	0	2	3
Contact with Farm Animals	*46*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
Admitted from foreign hospital	*31*	*0*	*0*	*1*	*0*	*0*	*0*	*0*	*0*
Adopted child	*11*	*1*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
¿2 months ago in foreign hospital	*7*	*0*	*0*	*0*	*0*	*0*	*0*	*1*	*1*
Foreign dialysis patient	*2*	*0*	*1*	*0*	*0*	*0*	*0*	*0*	*0*
Known MRSA carrier	*31*	*0*	*1*	*0*	*1*	*0*	*0*	*0*	*0*
Known HCW MRSA carrier	*2*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
Protected contact with MRSA carrier	*10*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
Unprotected contact with MRSA carrier	*115*	*1*	*0*	*3*	*0*	*1*	*0*	*0*	*0*
Admitted to hospital with known MRSA problem	*121*	*4*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
Shared room with MRSA patient	*110*	*1*	*1*	*5*	*0*	*0*	*2*	*0*	*0*

Table S4a. The number of isolates assigned to a cluster using the position of the health care institution in the patient referral network as location data, split up by MLVA clonal complex (MC) and epidemiological information, such as the risk groups defined by the Workgroup Infection Prevention (WIP). *MC398 only includes isolates from 2009, all others 2006-2009. “All” denotes the sum of all included isolates.

	All*	MC398*	MC5	MC8	MC45	MC22	MC30	MC621	MC80
All	2110	158	531	330	328	179	119	195	49
WIP Introductions	230	92	19	9	14	16	8	39	0
Unexpected Cases	351	6	77	53	56	20	33	38	17
No Information	420	35	112	48	55	17	30	51	22
WIP Ambiguous	35	1	14	4	5	2	1	5	0
Expected Cases	229	18	44	24	30	15	18	36	5
WIP Transmission	330	3	139	66	36	56	8	7	0
Contact Tracing	515	3	126	126	132	53	21	19	5
Contact with Farm Animals	*112*	*92*	*5*	*2*	*5*	*1*	*1*	*3*	*0*
Admitted from foreign hospital	*69*	*0*	*11*	*4*	*9*	*14*	*7*	*5*	*0*
Adopted child	*46*	*0*	*4*	*2*	*0*	*1*	*0*	*31*	*0*
¿2 months ago in foreign hospital	*3*	*0*	*0*	*1*	*0*	*0*	*0*	*0*	*0*
Foreign dialysis patient	*2*	*0*	*1*	*0*	*0*	*0*	*0*	*0*	*0*
Known MRSA carrier	*34*	*2*	*13*	*3*	*5*	*2*	*1*	*5*	*0*
Known HCW MRSA carrier	*2*	*0*	*0*	*2*	*0*	*0*	*0*	*0*	*0*
Protected contact with MRSA carrier	*11*	*0*	*1*	*4*	*0*	*6*	*0*	*0*	*0*
Unprotected contact with MRSA carrier	*105*	*1*	*41*	*18*	*14*	*17*	*5*	*2*	*0*
Admitted to hospital with known MRSA problem	*113*	*1*	*57*	*27*	*8*	*12*	*2*	*1*	*0*
Shared room with MRSA patient	*103*	*1*	*42*	*17*	*14*	*21*	*1*	*4*	*0*

Table S4b. The number of isolates assigned to a cluster using the position of the health care institution in the patient referral network as location data, split up by MLVA clonal complex (MC) and epidemiological information, such as the risk groups defined by the Workgroup Infection Prevention (WIP). *MC398 only includes isolates from 2009, all others 2006-2009. “All” denotes the sum of all included isolates.

	All*	MC398*	MC482	MC2	MC88	MC435	MC7	MC2236	MC632
All	2110	102	33	12	25	22	18	0	9
WIP Introductions	230	17	4	6	2	3	0	0	1
Unexpected Cases	351	24	6	2	8	6	4	0	1
No Information	420	27	7	2	2	4	7	0	1
WIP Ambiguous	35	2	0	0	1	0	0	0	0
Expected Cases	229	14	11	1	5	6	2	0	0
WIP Transmission	330	7	1	1	0	1	2	0	3
Contact Tracing	515	11	4	0	7	2	3	0	3
Contact with Farm Animals	*112*	*1*	*1*	*1*	*0*	*0*	*0*	*0*	*0*
Admitted from foreign hospital	*69*	*8*	*2*	*5*	*1*	*3*	*0*	*0*	*0*
Adopted child	*46*	*7*	*0*	*0*	*1*	*0*	*0*	*0*	*0*
¿2 months ago in foreign hospital	*3*	*1*	*0*	*0*	*0*	*0*	*0*	*0*	*1*
Foreign dialysis patient	*2*	*0*	*1*	*0*	*0*	*0*	*0*	*0*	*0*
Known MRSA carrier	*34*	*2*	*0*	*0*	*1*	*0*	*0*	*0*	*0*
Known HCW MRSA carrier	*2*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
Protected contact with MRSA carrier	*11*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
Unprotected contact with MRSA carrier	*105*	*2*	*0*	*1*	*0*	*1*	*0*	*0*	*3*
Admitted to hospital with known MRSA problem	*113*	*5*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
Shared room with MRSA patient	*103*	*0*	*1*	*0*	*0*	*0*	*2*	*0*	*0*

D

Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data

R.J.F. Ypma, A.M.A. Bataille, A. Stegeman, G. Koch, J. Wallinga, W.M. van Ballegooijen

D.1 Representing the transmission tree

A graphical representation of the estimated transmission tree was given in the main text (figure 2). To increase clarity for the area with highest farm-density, figure S1 gives a zoomed-in version of figure 2 for this area.

D.2 Testing different spatial kernels

The form of the spatial kernel used in the main text was taken from Boender et al. [3], who tested a number of parameterisations and found this kernel to fir the data best. To see whether this choice of parametric form influences our results, we test a range of kernels found in the literature (table S1, second column). The first five kernels tested are taken from Boender et al. [3], where kernel 5 is the kernel from the main article, and kernel 1 is identical to the analysis given in the main text using only genetic and temporal data. Kernels 6-8 are from a family of kernels featuring an exponential decay term. These have for instance been used for modelling the spread of foot-and-mouth disease [14, 16] and the dispersal of plant seeds [23, 21]. To calculate the effect of using different kernels, we substitute them in equation (3) in the main text, and rerun the analysis.

Table S1 gives estimates of parameter values for all kernels tested, figure S2 gives the resolution of the estimated trees (comparable to figure 3 in the main text) and the estimated average infectiousness of different types of farms (comparable to figure 5 in the main text).

We see that, although analyses using different kernels result in different values for the probabilities of transmissions, the results on resolution of the tree and infectiousness of farm types remain the same. Resolution is slightly lower for kernels 1 and 6. The first is the analysis without geographical data, the second comes close to this analysis since its two parameters are both estimated to be very close to zero (table S1). These parameters c and d in the exponential decay factor are consistently estimated very low, which is an indication that the exponential decaying function does not fit this epidemic well. The high estimated value for d in kernel 7 is probably due to very low estimates for c. Estimates of parameter values for genetic and temporal parameters are robust for different spatial kernels (table S2), which is probably due to the relatively low amount of information contained in the geographical data (main text).

Table S1. Spatial parameter estimates for models with different spatial kernels

	Spatial kernel	α (95% CI)	r_0 (95% CI)	c (95% CI)	d (95% CI)
1	1	-	-	-	-
2	$(1+x)^{-1}$	-	-	-	-
3	$(1+x^2)^{-1}$	-	-	-	-
4	$(1+x^{\alpha})^{-1}$	1.7 (1.5,2.0)	-	-	-
5	$(1+\frac{x}{r_0}^{\alpha})^{-1}$	2.3 (1.7, 2.8)	2.4 (1.2, 3.7)	-	-
6	$e^{(} - cx^d)$	-	-	0.039 (0.00075, 0.65)	0.12 (0.00087, 1.2)
7	$(1+x^{\alpha})^{-1}e^{-cx^d}$	2.3 (1.7, 2.9)	-	0.046 (0.0037, 0.13)	10 (4.0, 13)
8	$(1+\frac{x}{r_0}^{\alpha})^{-1}e^{-cx^d}$	2.4 (1.8, 3.1)	2.3 (1.3, 4.7)	0.14 (0.0051, 0.71)	0.46 (0.0064, 2.3)

Table S2 Temporal and genetic parameter estimates for models with different spatial kernels

	Spatial kernel	b (95% CI)	p_{ts} (95% CI)	p_{tv} (95% CI)	p_{del} (95% CI)
1	1	0.25 (0.20, 0.31)	1.1 (0.86, 1.3)	0.30 (0.20, 0.41)	0.060 (0.028, 0.11)
2	$(1+x)^{-1}$	0.28 (0.22, 0.35)	1.1 (0.88, 1.3)	0.31 (0.22, 0.42)	0.063 (0.031, 0.11)
3	$(1+x^2)^{-1}$	0.29 (0.24, 0.35)	1.1 (0.89, 1.3)	0.32 (0.22, 0.42)	0.068 (0.030, 0.12)
4	$(1+x^{\alpha})^{-1}$	0.29 (0.23, 0.35)	1.1 (0.88, 1.3)	0.31 (0.21, 0.41)	0.066 (0.029, 0.12)
5	$(1+\frac{x}{r_0}^{\alpha})^{-1}$	0.28 (0.23, 0.34)	1.1 (0.88, 1.3)	0.32 (0.22, 0.43)	0.069 (0.027, 0.13)
6	$e^{(} - cx^d)$	0.24 (0.20, 0.30)	1.1 (0.87, 1.3)	0.30 (0.20, 0.41)	0.063 (0.029, 0.12)
7	$(1+x^{\alpha})^{-1}e^{-cx^d}$	0.29 (0.23, 0.35)	1.1 (0.89, 1.3)	0.31 (0.22, 0.42)	0.068 (0.031, 0.11)
8	$(1+\frac{x}{r_0}^{\alpha})^{-1}e^{-cx^d}$	0.28 (0.23, 0.34)	1.1 (0.88, 1.3)	0.31 (0.22, 0.42)	0.066 (0.030, 0.12)

Another possibility is to use distance over road rather than Euclidean distance as a measure for geographical distance, such as done for foot and mouth disease by Savill et al. [22]. Unfortunately we lack the data to do this analysis thoroughly; to our knowledge there is no extensive database containing all (minor) roads in the particular area. However, we believe the analysis would yield results similar to the one we performed using Euclidean distance, as this is a very road-dense area with no large geographical obstacles. Furthermore it is quite likely that several different mechanisms were responsible for spread of the disease, some following the road- and others the Euclidean distance. For FMD this mixture of transmission mechanisms was given as a possible explanation for the fact that Euclidean distance gave the better fit, unless a geographical obstacle was present [22].

D.3 Latent period

Although the assumption of a latent period seems valid for the avian influenza epidemic [4, 10, 9], the exact length for this latent period is not known. In the main article we assumed a length of one day; here we test this assumption for robustness. We do this by increasing the length of the latent period in equation (1) in the main text, and rerunning the full analysis.

Table S3 gives estimates of parameter values for the different analyses, figure S3 gives the resolution of the estimated trees (comparable to figure 3 in the main text) and the estimated average infectiousness of different types of farms (comparable to figure 5 in the main text).

Although analyses using different latent periods result in different values for the probabilities of transmissions, the estimates of parameter values as well as the resolution of the tree remain the same. Estimates of infectiousness of farm types vary slightly, especially for the types that included fewer farms. This is probably due to the fact that some transmissions become impossible in the model when the latent period is increased, which can have a large impact on

Table S3 Comparison between models with different length of latent period

Length of latent period (days)	b (95% CI)	α (95% CI)	r_0 (95% CI)
1	0.28 (0.23, 0.34)	2.3 (1.7, 2.8)	2.4 (1.2, 3.7)
2	0.29 (0.24, 0.35)	2.3 (1.8, 2.9)	2.4 (1.3, 3.9)
3	0.22 (0.18, 0.26)	2.4 (2.0, 3.0)	2.4 (1.5, 3.5)

Length of latent period (days)	p_{ts} (95% CI)	p_{tv} (95% CI)	p_{del} (95% CI)
1	1.1 (0.88, 1.3)	0.32 (0.22,0.43)	0.069 (0.027, 0.13)
2	1.1 (0.91, 1.4)	0.30 (0.21, 0.42)	0.074 (0.034, 0.12)
3	1.2 (1.0, 1.5)	0.33 (0.24, 0.45)	0.078 (0.040, 0.11)

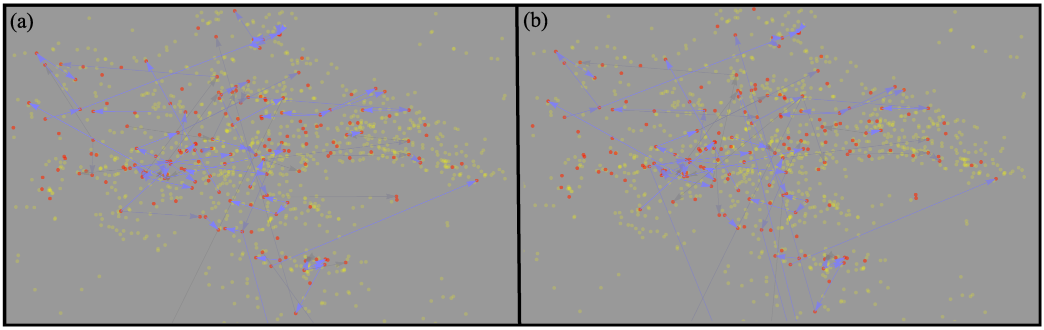


Figure S1. Infection events with posterior probability >0.5, in the northern part of the outbreak. This is a zoomed-in version of figure 2 in the main text. Results are for a model using (a) temporal, genetic and geographic, and (b) temporal and genetic data. Red dots denote infected farms, yellow dots denote farms not infected in the epidemic. A higher opacity of arrows corresponds to a higher estimated probability. We see the two analyses give roughly the same transmission links, although the addition of geographic data allows for more precise estimations.

the average infectiousness of a small group of farms. However all conclusions remain unchanged.

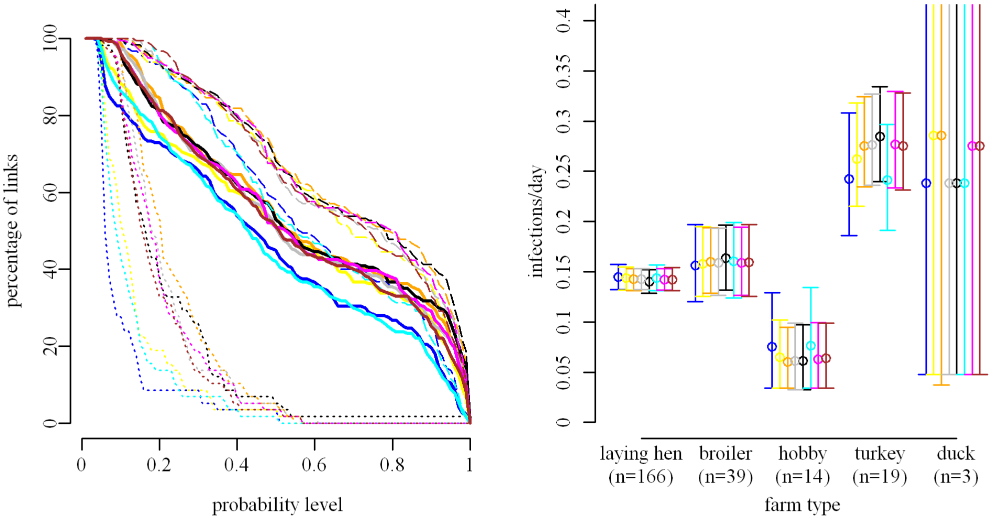


Figure S2. (Left) Resolution of the estimated trees, using different spatial kernels. For each level of probability, the percentage of farms for which the farm that infected it can be estimated at or above this level is plotted. Kernels 1-8 (table S1) are coloured blue, yellow, orange, grey, black, cyan, magenta and brown respectively. Kernels 5 and 1 (black and blue) are the analyses from the main text using all data and only using genetic and temporal data. Striped and dotted lines give the same information for sequenced and unsequenced farms respectively. All trees have roughly the same resolution, apart from those resulting from kernels 1 and 6 (blue and cyan lines). The former uses no spatial information, while the latter comes close to this analysis since the kernel solely consists of an exponential decrease, whose parameters have been estimated close to zero. This is an indication exponential decay does not fit this data well.
(Right) Estimated average infectiousness for different types of farms, as measured by number of infections caused divided by the time period in days between infection and culling of the farm. Results are for models using different spatial kernels, kernels 1-8 (table S1) are coloured blue, yellow, orange, grey, black, cyan, magenta and brown respectively. Kernels 5 and 1 (black and blue) are the analyses from the main text using all data and only using genetic and temporal data. Estimates are similar for all kernels, kernels 1 and 6 (blue and cyan) have less discriminatory power and therefore less variation in their estimates.

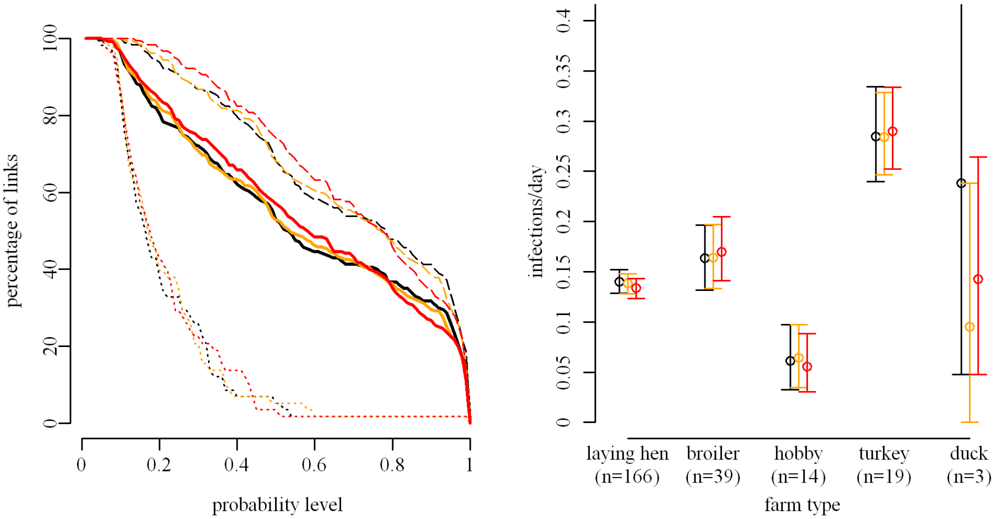


Figure S3. (Left) Resolution of the estimated trees, assuming the latent period is one (black), two (orange) or three (red) days. For each level of probability, the percentage of farms for which the farm that infected it can be estimated at or above this level is plotted. Striped and dotted lines give the same information for sequenced and unsequenced farms respectively. All trees have roughly the same resolution.
(Right) Estimated average infectiousness for different types of farms, as measured by number of infections caused divided by the time period in days between infection and culling of the farm. Results are for models using a latent period of one (black), two (orange) or three (red) days. We can see some small differences, especially for duck farms. This is probably because some transmissions become impossible in the model when a longer latent period is assumed. However the conclusions of low infectiousness of hobby farms and high infectiousness of turkey farms remain unaltered.

E

Genetic data provide evidence for wind-mediated transmission of highly pathogenic avian influenza

R.J.F. Ypma, M. Jonges, A. Bataille, A. Stegeman, G. Koch, M. van Boven, M. Koopmans, W.M. van Ballegooijen, J. Wallinga

Table S1. Overview of data used.

variable	**description**	**available for**
$\mathbf{x}$	farm location	all 5360 poultry farms
t^{cull}	date of culling	all 1531 culled farms
t^{inf}	estimated date of infection	all 241 infected farms
RNA	sequences of the HA, NA and PB2 genes	231 of the 241 infected farms
w_{dir}	average wind direction (precision of 10 degrees)	every hour of every day of 2003
w_v	average wind speed (precision of 1 m/s)	every hour of every day of 2003
r	fraction of time without precipitation	every hour of every day of 2003

E.1 Overview of data

For the outbreak of avian influenza A(H7N7) in the Netherlands in 2003, much data are available. Here we give a description of the data used in our analyses, and give an overview in table S1.

E.1.1 Epidemiological data

There were 5360 poultry farms in the Netherlands in 2003, for all of which geographical information $\mathbf{x}$ is available. For 1531 farms the flocks were culled, for all of these the date of culling T^{cull} is known. For 227 of the 241 infected farms the date of infection t^{inf} has been estimated, based on mortality data [4]. The remaining 14 farms are hobby farms, defined as farms with less than 300 animals, for which no mortality data are available. For these we use the infection date as estimated by Boender et al. [3].

The geographic and temporal data together have previously been used to estimate the critical farm density, i.e. above what density of farms outbreaks are can occur [3].

E.1.2 Genetic data

The HA, NA and PB2 genes of viral samples from 231 farms have previously been sequenced [1, 17]. Sequence data RNA can be found in the GISAID database under accession numbers EPI_ISL_68268-68352, EPI_ISL_82373-82472 and EPI_ISL_83984-84031.

These data have previously been used to give general characteristics of the outbreak [1], to reconstruct the transmission tree [26] and to assess the public health threat due to mutations of the virus in the animal host [17].

E.1.3 Meteorological data

Available meteorological data include wind speed w_v and direction w_{dir} (with a ten degree precision) and the fraction of time r without precipitation for every hour of every day of the outbreak, measured at five weather stations close to the infected farms (figure 1). These data are available from the Royal Dutch Meteorological Institute at www.knmi.nl.

E.2 Inference of transmission events

Our objective is to estimate which farm infected which in the outbreak of avian influenza A(H7N7) in the Netherlands in 2003, using temporal data $\mathbf{t} = (t^{inf}, t^{cull})$ on the infected farms and genetic data RNA on virus sampled from 231 of the 241 infected farms. Following the approach by Ypma et al.[26], we will obtain joint estimates for the transmission tree and the parameters **par** describing the transmission process. To do this we need the likelihood of the transmission tree and parameters, given the available data **D**. If we had complete data, the likelihood for the whole transmission tree T would be the product of the likelihoods for the individual transmission events δ:

$$L(T, \mathbf{par}|\mathbf{D}) = \prod_{\delta \in T} L(\delta, \mathbf{par}|\mathbf{D})$$

Denote the event that a certain farm A infected another farm B by δ_{AB}. We then take the likelihood L of this event to be the product of the likelihood given the temporal data, L_t, and the likelihood given the genetic data, L_{gen}:

$$L(\delta_{AB}, \mathbf{par}|\mathbf{D}_{AB}) =$$

$$L_t(\delta_{AB}, b|\mathbf{t}_A, \mathbf{t}_B) L_{gen}(\delta_{AB}, \mathbf{p}|\mathrm{RNA}_A, \mathrm{RNA}_B)$$

where b and $\mathbf{p}$ are parameters explained below.

Note that we could include geographical information in this likelihood to better estimate the transmission tree [26]. However, we want to use the transmission events estimated here by comparing them with wind directions. Since spread due to wind will be distance-dependent, already including geographical information in estimating the transmissions could lead to circular results. To check whether this would influence results, we also estimated the transmission tree including geographical information. We still found a significant correlation between wind direction and direction of transmissions as in the main text. The estimate of the percentage of transmissions (described in section E.4.1) due to wind changed from 18% to 17% (figure S3).

E.2.1 Likelihood given temporal data

After the date of infection t^{inf}, we assume a latent period of one day, and then a constant infectiousness until the date of culling t^{cull}. After culling infectiousness decreases exponentially with rate b. The likelihood of A infecting B based on temporal data is given by the infectiousness I_A of A on the infection date t_B^{inf} of B:

$$L_t(\delta_{AB}, b|\mathbf{t}_A, \mathbf{t}_B) = I_A(t_B^{inf}) = \begin{cases} 0 & t_B^{inf} \leq t_A^{inf} \\ 1 & t_A^{inf} < t_B^{inf} \leq t_A^{cull} \\ e^{-b(t_B^{inf} - t_A^{cull})} & t_B^{inf} > t_A^{cull} \end{cases}$$

Note that we are only looking at relative infectiousness here, the absolute infectiousness could be determined by adding information on farms that could have been infected but were not [3].

It was previously shown that changing the latent period to 2 or 3 days did not much influence the estimation of the parameters and the transmission tree [26]. However, uncertainty in the transmission dates could influence the estimate of the effect of wind (due to changing wind directions). Here we use the estimated dates to test for a significant correlation with wind, later we explore the influence of changing the transmission dates (section E.4.2).

E.2.2 Likelihood given genetic data

For each infection, we assume a deletion can occur with probability p_{del}, and for each of N nucleotides a transition or transversion can occur with probabilities p_{ts} and p_{tv} [18]:

$$L_{gen}(\delta_{AB}, \mathbf{p}|\mathrm{RNA}_A, \mathrm{RNA}_B) =$$

$$\frac{\left(\frac{p_{ts}}{N}\right)^{d_{ts}}}{(1-\left(\frac{p_{ts}}{N}\right))^{d_{ts}-N}} \frac{\left(\frac{p_{tv}}{N}\right)^{d_{tv}}}{(1-\left(\frac{p_{tv}}{N}\right))^{d_{tv}-N}} p_{del}^{\mathbf{1}_{del}} (1-p_{del})^{1-\mathbf{1}_{del}}$$

Here $\mathbf{p} = (p_{ts}, p_{tv}, p_{del})$, d_{ts} and d_{tv} are the number of transitions and transversions between the sequences found at farm A and B, and $\mathbf{1}_{del}$ is 1 when a deletion has occured, 0 otherwise. More elaborate substitution models can also be incorporated, but here could easily lead to over-parametrization [1].

E.2.3 Estimation of the transmission tree

For some farms no genetic data is available. We can still evaluate the transmission tree when it contains such a farm F, by looking at the genetic data of the farm that infects F, and the farms infected by F. We then sum over all possible sequences F could have had which are consistent with the least amount of mutations needed to explain this subtree. Thus we split the transmission tree T into the set of smallest subtrees $\mathbf{S}_{gen}$ such that the root and leaves of all $S \in \mathbf{S}_{gen}$ are farms for which genetic data is available. The likelihood for a transmission tree T and parameters $\mathbf{par}$ is then given by

$$\begin{aligned} L(T, \mathbf{par}|\mathbf{D}) &= L_t(T, b|\mathbf{t}) L_{gen}(T, \mathbf{p}|\mathrm{RNA}) \\ &= \prod_{\delta_{AB} \in T} L_t(\delta_{AB}, b|\mathbf{t}_A, \mathbf{t}_B) \prod_{S \in \mathbf{S}_{gen}} L_{gen}(S, \mathbf{p}|\mathrm{RNA}_S) \end{aligned}$$

We sample from the space of all possible transmission trees and parameters using a Monte Carlo Markov Chain (MCMC), taking flat priors for all parameters on the positive real numbers and uniform priors for all transmission links. We take a burn-in of 500 000 iterations, checking for convergence, and then sample 10 000 times, at every 500th iteration.

For any two farms A and B, the probability that A infected B can then be obtained as the fraction of sampled transmission trees in which A infects B.

E.2.4 Double infections

A critical assumption in the approach described is that the 241 farms (apart from the index farm) were infected exactly once, by a previously infected farm. Farms could have been infected multiple times. However, due to the high virulence of the virus, any second introduction would have to follow the first very swiftly to still be able to spread within the already infected farm. Cloning performed on samples from 6 farms showed no signs of multiple introductions [1]. However, from one farm in the beginning of the outbreak, sequences have been determined from two available samples. These two sequences differ by 9 SNPs, while for both sequences related viruses (1 SNP difference) have been sequenced from other farms. These genetic distances are indicative of two separate introductions; we have taken this farm to be infected twice. Thus we have 241 transmission events. We redid our analysis excluding the second sample of the doubly infected farm; this does not influence our results (figure S3).

E.3 Correlation between wind direction and direction of transmission

The correlation between directions can be computed using circular correlation coefficients. These are extensions of the concept of correlation to the circle, e.g. a value of 1 indicates complete correlation. Several versions exist, we take the often used coefficient proposed by Fisher and Lee [15]

$$c(a,b) = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sin(a_i - a_j)\sin(b_i - b_j)}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sin^2(a_i - a_j)\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sin^2(b_i - b_j)}}$$

where a and b are vectors of directions, expressed in radians, and $c(a,b)$ is the correlation between them.

We take a to be the vector of directions of the 83 observed transmissions found, and b the vector of wind directions at the date of infections, with a time lag $\Delta \in \{-5,-4,...,4,5\}$. As wind speeds are usually sufficient to cover a typical transmission distance within an hour, we expect the wind direction at the date of infection to be the most relevant, and thus the correlation to be highest for time lag 0. We compare the values found against the null hypothesis of no correlation, which we find by calculating the correlation between two vectors of 83 randomly generated directions, and repeating this 1000 times. We also compare the correlation to the simulation results described in the main text.

Results are shown in figure S1. The correlations found are significantly higher than expected by chance for all non-positive time lags. Interestingly, the correlation is highest for a lag of -3, meaning that the correlation is highest for wind directions three days prior to the estimated date of infection. This could indicate that there either is a bias in the estimation of the dates of infection, or there is a lag inherent in the infection mechanism (e.g. transported particles are only taken up by animals a few days after wind-mediated transportation). Note that, if wind-mediated transmission indeed took place several days before the estimated date of infection, our estimation of the contribution of wind to disease transmission is actually an underestimation (also see Discussion of main text).

E.4 Quantification of the effect of wind

The main result of our study is that direction of transmissions and wind coincide more often than can be explained by chance. This could be due to a causal relationship; a wind-mediated mechanism of spread. Assuming such a wind-mediated mechanism of spread indeed exists, we estimate the percentage of transmissions attributable to this mechanism. We propose two methods.

The first method makes minimal assumptions on how the mechanism works. Rather, the method compares the fraction of transmissions that are in the direction of the wind with the fraction expected when wind plays no role (obtained from simulations). The difference between these fractions is informative of the percentage of transmissions caused by wind.

The second method assumes a specific mechanistic model for wind-mediated spread. This method repeats the whole inference procedure, jointly estimating the transmission tree and parameters (which now include parameters describing wind-mediated spread). Although this is a much more refined model, a mismatch between the assumed mechanistic model and the actual mechanism of wind-mediated spread could lead to an underestimation of the percentage of transmissions due to wind.

E.4.1 A heuristic approach

We estimate the percentage of transmission events due to wind. To do this, we compare the fraction of estimated transmissions (weighted by their probability) that are in the direction of the wind to the fraction expected if none or all of the transmissions were due to wind. The finest resolution metereological data available gives average wind directions for every hour, up to ten degrees. We take a transmission event from farm A to B to be 'in the direction of the wind' if at the date of infection of B at least one of the hourly average wind directions is the same as the direction from A to B, allowing for a maximum deviation of five degrees.

We say each transmission has occurred due to wind with probability p, and due to any other mechanism with probability $(1-p)$. Let p_w be the probability that a transmission caused by wind was also observed to be in the direction of the wind. We expect this probability p_w to be close to one, although it can be smaller due to noise in the data and variability in wind direction.

Let p_{nw} be the probability that a transmission not caused by wind was observed to be in the direction of the wind. If the average wind direction in a certain hour were unrelated to the wind direction the hour before, p_{nw} would be $1-(1-\frac{1}{36})^{24} \approx 0.49$, since there are 24 hours in a day and 36 possible wind directions for each hour. However, since wind direction is heavily autocorrelated the actual value of p_{nw} will be lower. To obtain its actual value we use simulations in which wind plays no role. In silico, we infect the index case of the outbreak, and use the transmission parameters previously estimated [3] to simulate an outbreak. We do this 1000 times, and for each of the simulations count the fraction of transmissions in the direction of the wind. This value was found to be 0.244, which we take as an estimate for p_{nw}.

If farm A actually infected farm B, the probability of wind to be observed in the direction of A to B on the estimated infection date of B is $pp_w + (1-p)p_{nw}$. If A did not infect B wind plays no causal role and we approximate the probability of observing wind to be in the direction of A to B as p_{nw}. Let $\mathbf{F}$ be the set of all farms, and $P(\delta_{AB})$ be the posterior probability for the event that farm A infected B. Let $\mathbf{1}_{w(AB)}$ be 1 if wind was observed to be in the direction of A to B at the date of infection of B, 0 otherwise. Then the probability of observing and estimated transmission to be in the direction of the wind is

$$f(A,B) = P(\delta_{AB})(pp_w + (1-p)p_{nw}) + (1-P(\delta_{AB}))p_{nw}$$

and the likelihood of p given our posterior probabilities $\mathbf{P}$ and the wind data $\mathbf{w}$ is

$$L(p|\mathbf{P},\mathbf{w}) \quad = \quad \prod_{A\in\mathbf{F}}\prod_{B\in\mathbf{F}}\left(\mathbf{1}_{w(AB)}f(A,B) + (1-\mathbf{1}_{w(AB)})(1-f(A,B))\right)$$

Now p can be estimated by maximum likelihood, and confidence intervals can be obtained by using profile likelihood. When we set p_w conservatively to 1 and p_{nw} to 0.244, p is estimated at 0.18 (95%CI 0.063,0.30). For lower values of p_w this value increases, making our estimate conservative.

E.4.2 A mechanistic approach

We now extend the transmission tree estimation from chapter (E.2) by assuming a specific model to describe the spread of disease due to wind. This allows us to differentiate between spread due to wind and spread due to an unknown mechanism, and to estimate the proportion of transmission events due to wind.

Likelihood given epidemiological data

To describe the spread of infectious particles by wind we take a Gaussian plume model. This type of model is often used to describe spread of particles by wind. We use the model by

Ssematimba et al. [25], which was specifically fitted to this outbreak. This model assumes particles leave a farm through a chimney, and gradually descend due to gravity. The particles randomly drift in all directions, but their movement will be dominated by the current wind; movement will predominantly be in the direction of the wind w_{dir} at the wind speed w_v. This results in a plume of particles in the direction of the wind (hence 'plume' model).

We take wind to be constant in speed and direction during each hour of every day. The total force of infection exerted by a farm A to another farm B is then the sum of two parts. The first part is due to an unknown mechanism independent of direction, well described by a hyperbolic relation [3]

$$\lambda_{AB}^{U} = \frac{h_0}{1 + \left(\frac{||\mathbf{x}_A - \mathbf{x}_B||}{r_0}\right)^{\alpha}}$$

where $||\mathbf{x}_A - \mathbf{x}_B||$ is the distance in km between the two farms, and h_0, α and r_0 are parameters.

The second part is due to wind, which we take to be linearly related to the number of particles settled according to the Gaussian plume model G

$$\lambda_{AB}^{W} = h_w \sum_{h=1}^{24} \mathbf{1}_{dir(A,B)=dir_h} r_h G(||\mathbf{x}_A - \mathbf{x}_B||, v_h)$$

where $dir(A, B)$ is the direction from A to B, and dir_h and v_h are the wind direction and speed at a certain hour h. r_h is the fraction of time without precipitation during a certain hour, which we put in since dust will settle quickly when it is raining. Excluding this factor does not influence results. Note that in the previous analysis wind speed was not included; as we could not discern a clear pattern from the data and it is not clear how wind speed influences transmission [25, 24], we opted for the most parsimonious model and left this variable out.

We take the formulation above since dispersion in the lateral direction will be small compared to the 10 degree precision of the data [25]. The complete likelihood of parameters $\mathbf{h} = (h_0, r_0, \alpha, h_w, b)$ and farm A infecting farm B given the temporal and geographical data is then

$$L_{t+geo}(\delta_{AB}, \mathbf{h}|\mathbf{x}_A, \mathbf{x}_B, \mathbf{w}) \quad = \quad 1 - exp\left[-I_A(t_B^{inf})(\lambda_{AB}^{U} + \lambda_{AB}^{W})\right]$$

where

$$I_A(t_B^{inf}) = \begin{cases} 0 & t_B^{inf} \leq t_A^{inf} \\ 1 & t_A^{inf} < t_B^{inf} \leq t_A^{cull} \\ e^{-b(t_B^{inf} - t_A^{cull})} & t_B^{inf} > t_A^{cull} \end{cases}$$

as before (section E.2.1).

Estimating infection dates

The dates of infections were estimated based on mortality data [4]. These estimates can be incorrect, which might have a large effect on the estimated contribution of wind, although this effect will be negated by the fact that wind directions on two consecutive days are correlated. We jointly estimate the dates of infection, taking a prior ranging from four days before to four days after the estimated date. We took a triangle-shaped distribution, where the originally estimated date is most likely and the other dates are 2^n less likely, where n is the number of days difference with the originally estimated date. For the 14 farms for which no mortality data was available, we took a uniform prior on the 9 days. To test the dependence of results on this prior, we performed the same analysis with the different priors. Firstly, by taking the prior to be 5 days wide, rather than 9. Thirdly, by taking a uniform prior of 9 days for all farms.

Taking the prior too wide could lead to an overestimation of the contribution of wind, since for any putative transmission the algorithm can put the infection date of the receiving farm on a day when the wind is blowing in the direction of the transmission. Figure S3 shows that taking wider and flatter priors increases our estimate of the contribution of wind.

Estimation of the transmission tree

There is additional information in farms that were not infected, i.e. when wind plays a role in transmission we expect these to be underrepresented in the downwind direction. We therefore take into account farms that could have been infected but were not [3]. Let $\mathbf{F}$ be the set of all poultry farms in the Netherlands during the outbreak and $\mathbf{F}_I$ the set of all infected poultry farms, such that $\mathbf{F}_I \subset \mathbf{F}$. Let t_F be the date for which farm F was no longer at risk: this is its culling date, or ∞ if F was never culled. The total likelihood for a transmission tree T, the estimated infection dates $\mathbf{t}_{F_I}$ and parameters $\mathbf{par}$ given the data D is then given by the likelihood of the tree times the likelihood that the uninfected farms did not get infected:

$$\begin{aligned} L(T, \mathbf{t}_{F_I}, \mathbf{par}|D) \;\;=\;\; & \textstyle\prod_{S \in S_{gen}} L_{gen}(S, \mathbf{p}|\mathrm{RNA}_S) \displaystyle\prod_{\delta_{AB} \in T} 1 - e^{-I_A(t_B^{inf})(\lambda_{AB}^U + \lambda_{AB}^W)} \times \\ & \prod_{F \in \mathbf{F}} \prod_{F_I \in \mathbf{F}_I} \prod_{t=0}^{t_F - 1} e^{-I_{F_I}(t)(\lambda_{F_I F}^U + \lambda_{F_I F}^W)} \end{aligned}$$

For computational purposes, we work with the log likelihood, turning the above products into sums. Furthermore, we first precompile a matrix of transmission events that could have happened but did not, specified for distance and wind speed, so the triple product above for uninfected farms does not have to be calculated fully in every iteration [6].

Quantification of the effect of wind

To quantify the effect of wind, we note that for any transmission event the probability this event was due to wind is given by the proportion of the force of infection from the infecting to the infected farm which is due to wind. For any iteration of the MCMC, the proportion of transmissions mediated by wind can then be estimated as the average of these probabilities

$$\frac{1}{\#T} \sum_{\delta_{AB} \in T} \frac{\lambda_{AB}^W}{\lambda_{AB}^U + \lambda_{AB}^W}$$

where by $\#T$ we denote the number of transmission events. Figure S3 gives the posterior densities of the estimates of the proportion of transmission events due to wind.

E.5 Sensitivity analyses

E.5.1 Cut-off value for observed transmissions

In the main text a significant correlation between transmission events and wind direction is shown using ‘observed transmission events’, defined as transmission events estimated at a posterior probability of at least 0.5. Since this is an arbitrary cut-off value, we performed the same analysis with different cut-off values to test for robustness. Figure S2 gives the same information as figure 2 in the main text, with cut-off values for observed transmission events of 0.1, 0.2,..., 0.9. Since we do not use any weighing of these observed transmission, we would expect the correlation between observed transmissions and wind direction to increase with the cut-off value. This is because at low cut-off values many of the observed transmissions will have a low posterior probability and are likely to be estimated incorrectly. As described in the main text, we performed a one-sided Kolmogorov-Smirnov test to test if the angles between transmission events and the vector average wind direction were significantly smaller for the actual dataset than for the simulations (figure S2, left graphs). We also compared the average number of hours that wind coincided with the direction of a transmission, which could be interpreted as the time window for spread due to wind (figure S2, right graphs). For both, we found a significant ($p < 0.01$) difference between the actual data and the simulations for cut-off values of 0.4 and greater.

Note that the variance of the simulation results increases with cut-off value. This is because we sample, from each simulation, as many transmissions as we have observed transmissions. This sampling reflects the fact that we need a more stringent test when we have less measurements.

E.5.2 Metereological stations

In the main text, we took the wind direction at the date of transmission, as measured at the metereological station closest to the transmitting farm. Wind directions measured where very similar for all five stations; results changed little when taking the meteorological station closest to the receiving farm, or when taking all wind directions from one of the five metereological stations.

E.5.3 Quantifications

In our heuristic approach to quantify the percentage of transmissions mediated by wind, we assumed p_w, the probability of observing a transmission due to wind to be in the direction of the wind, to be 1. It could be lower, due to uncertainty in the infection date and wind variability. The value of 1 is actually conservative; the estimate of the magnitude of the role of wind increases when p_w decreases.

Figure S3 gives the same quantification under various assumption; estimates from the heuristic approach when excluding the double infection (section E.2.4), and when including the geographic information in reconstructing the transmission tree, and posterior distributions under the mechanistic model using different priors for the infection dates. Although the spread of these estimates shows it is hard to pinpoint the exact number, the range of plausible values (6%-30%) seems well supported.

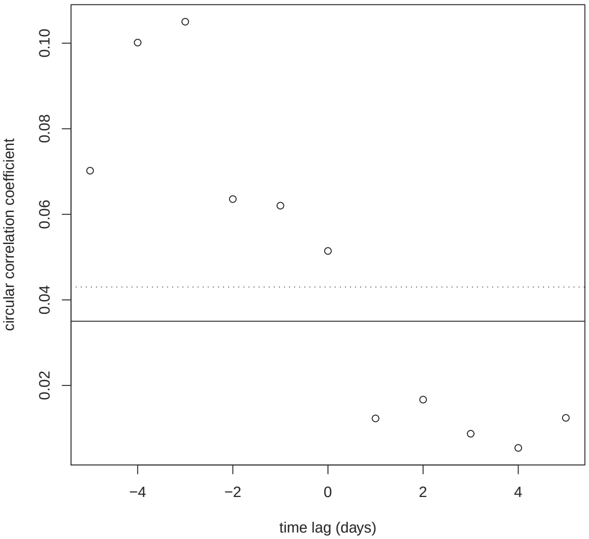
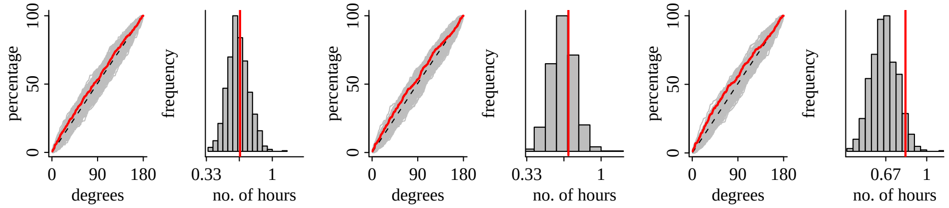
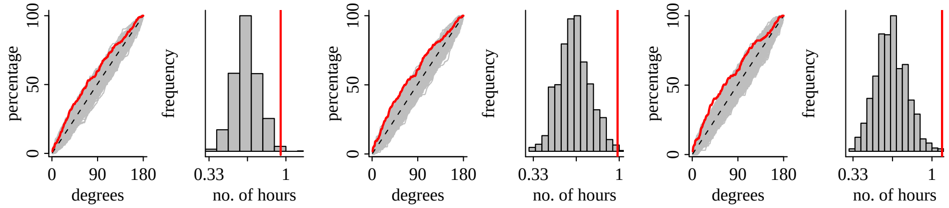


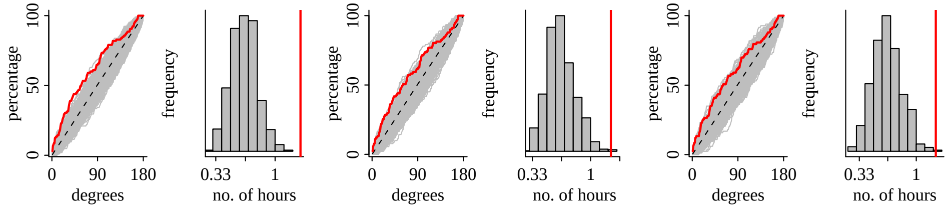
Figure S1. Circular correlation coefficient of directions of observed transmission and wind directions and, at estimated date of infection (time lag = 0) and five days before or after. The horizontal solid line indicates the 0.975 percentile for the distribution of the correlation coefficient under the null hypothesis of two independent uniform distributions, the dotted line indicates the same quantile under a second null hypothesis explored using simulations that takes into account farm geography (main text). As the value for time lag 0 is above both lines, wind direction and direction of transmission are statistically significantly correlated. The correlation is highest for wind directions three days prior to the estimated date of infection. This could indicate a wind-mediated transmission mechanism that is not instantaneous but takes some time to lead to infection, or a systematic bias in the estimation of the infection dates.



cut-off value 0.1 cut-off value 0.2 cut-off value 0.3



cut-off value 0.4 cut-off value 0.5 cut-off value 0.6



cut-off value 0.7 cut-off value 0.8 cut-off value 0.9

Figure S2. Correlation between observed transmissions and wind direction for (red) actual data and (grey) simulations. Left graphs give the cumulative distribution for the angle between transmission events and the vector average wind direction at the estimated date of transmission. Right graphs give the average number of hours at the date of transmission for which wind was in the direction of the transmission. Observed transmissions are those transmission events estimated at a posterior probability larger than a certain cut-off value. Panels give results for cut-off values of 0.1, 0.2,..., 0.9. The value of 0.9 is used in the main text. Observed transmissions are significantly correlated with wind direction ($p < 0.01$) for cut-off values 0.4 and greater.

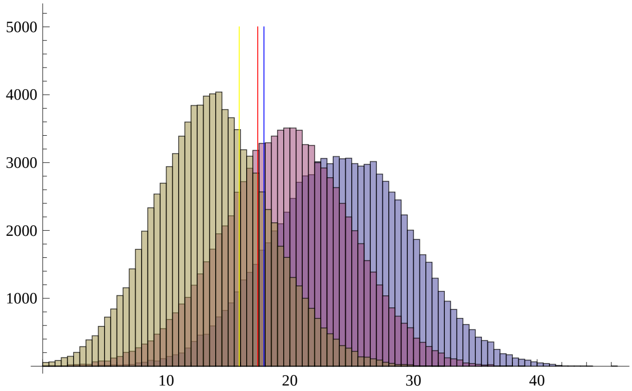


Figure S3. Estimates of the percentage of transmissions due to a wind-mediated mechanism, under various assumptions. The blue line gives the estimate under the heuristic approach (section E.4.1). The yellow line gives the same estimate when geographic data is included in estimating the transmission events, the red line gives the estimate when the putative double infection is discarded (section E.2.4). Histograms give posterior density when assuming a specific wind model (section E.4.2), with a (yellow) narrow prior on the transmission date, (pink) wider prior on the transmission date, (blue) flat prior of 9 days on the transmission date. See section E.4.2 for details. The blue line and pink distribution are given in the main text (figure 3).

F

Relating phylogenetic trees to transmission trees of infectious disease outbreaks

R.J.F. Ypma, W.M. van Ballegooijen, J. Wallinga

F.1 Test of estimation procedure on simulated data.

F.1.1 Simulating outbreaks

To illustrate the method and assess robustness to missing data, we apply it to simulated data based on an influenza outbreak in a confined school-based population. A common question in such a setting is whether one group of individuals (e.g. children) is more infectious than another (e.g. adults). To emphasize the difference, we performed simulations where only children were infectious.

We start with a population of 200 hosts, of which one is infected. Each individual belongs to one of two groups (children and adults). The outbreak is generated using a simple stochastic SIR model with homogeneous mixing in continuous time. For each infected host x an infectious period of length ip_x is drawn from a gamma distribution with a mean of three days and a variance of two days [5]. The infectiousness of host x is then

$$B_t(x) = \begin{cases} \beta_x & \text{if} \quad t_x + 2 < t < t_x + 2 + ip_x \\ 0 & \text{else} \end{cases}$$

where $\beta_x = \frac{2}{3*200}$ if x is a child (i.e. $R_0 = 2$), 0 otherwise.

Using the (known) transmission tree, we generate sequences as follows; we assume each infected individual is sampled one day after symptom onset ($t_x + 2 + 1$). We construct a phylogenetic tree using these timepoints as tips, by backwards simulation of the tree using $p(P|T, W)$ as specified in the main text. We then generate point mutations on this tree using a simple molecular clock; the number of mutations per edge of the phylogenetic tree is Poisson distributed with expected value $L \times \mu \times t = 10000 \times 0.003 \times t$ for the baseline scenario, with $L = 10000$ the number of base pairs and $\mu = 0.003$ the substitution rate. Note that the approximation of a Poisson distribution is valid here due to the short timescales. In fact, the probability of a recurrent mutation ocurring is approximately equal to the probability any nucleotide mutates twice in 5 days, times the number of nucleotides, which is $(0.003 \times \frac{5}{365})^2 \times 10000 \approx 1.7 \times 10^{-5}$.

A quantity of interest is the genetic distance between the sequences sampled from an infected host and its infector, as this is ultimately the data we are working with. The evolutionary time separating the samples taken from a host and its infector is roughly between 3 and 11 days, so the expected number of mutations separating the sequences is between $10000 * 0.003 * \frac{3}{365} = 0.25$ and $10000 * 0.003 * \frac{11}{365} = 0.9$, which seems plausible for an RNA-virus [19, 1, 7].

We investigated a total of seven scenarios. To improve comparability between results from the different scenarios, we re-used the same simulated outbreaks for the different scenarios. This re-usage eliminates variation resulting from the stochastic nature of the simulations. For example, the same phylogenetic tree was used in different scenarios to generate sequences based on different substitution rates. The exception is formed by the scenario where 20% of cases were discarded. As we wanted to keep the number of infected individuals comparable, separate simulations were performed for this scenario, which started with an increased susceptible population of $200/(1 - 0.2) = 250$.

F.1.2 Simulated data

For each case, we assume we know the time of symptom onset, the recovery time, a sequence sampled one day after symptom onset, and whether the case is a child or adult. The likelihood equations used in the MCMC are given in the main text. For efficient calculation of the likelihood $p(D_G|P, \mu)$ of the phylogenetic tree and mutation rate given the sequences we use Felsenstein's pruning algorithm [12]. We use a Uniform(0,1000) prior for all parameters.

F.2 Application of the estimation procedure to data on foot-and-mouth disease.

F.2.1 Data

In 2001, a large epidemic of foot-and-mouth disease (FMD) occured in the United Kingdom. A subset of 15 farms of this large epidemic, the so-called ‘Darlington cluster’ has been extensively studied [7, 8, 20]. Three of the farms are not epidemiologically linked to the other 12, and subsequently dropped from the analysis [20]. The remaining 12 farms are labelled $\mathcal{F} = \{C, D, E, F, G, H, I, J, K, L, M, O\}$. For each of the farms we have

- T_i^{obs}, the date of detection of the virus,
- D_i^{obs}, the estimated age of infection at date of infection, as assessed by a visual exam of the clinical state of lesions found,
- T_i^{end}, the date of culling,
- $\mathbf{x}_i$, the spatial location (as latitude/longitude),
- S_i^{obs}, an 8000 bp DNA sequence sampled at T^{obs}.

All these data are freely available from the open-access publication by Morelli et al. [20].

F.2.2 Model

The likelihood component for the transmission tree is based on an epidemiological model used by Morelli et al. [20]. In particular, after infection farms enter a latent period, whose duration L_i is gamma distributed with expectation β_1 and variance β_2^2. After the latent period, the farms enter an infectious period, as infected animals develop lesions. After a certain time D_i, the farm is detected as being infected. The data contain an estimate of D_i, which is assumed to be gamma distributed with mean D_i^{obs} and variance $D_i^{obs}/4$. Infectiousness of farms is assumed to stay constant during the infectious period, but to decrease exponentially with distance, with a mean transmission distance of $2\alpha_2$. A vague prior is assumed for all epidemiological parameters; an exponential distribution with mean 100. Letting $D_E = (T^{obs}, D^{obs}, T^{end}, \mathbf{x})$, we get

$$p(D_E|T, \theta, W) = \prod_{i \in \mathcal{F}^-} \frac{e^{-\frac{|\mathbf{x}_i - \mathbf{x}_{v(i)}|}{\alpha_2}}}{2\pi\alpha_2^2} \mathbb{1}_{t_{v(i)}+L_{v(i)}<t_i<T^{end}_{v(i)}} \prod_{i \in \mathcal{F}} g(L_i, \beta_1, \beta_2^2) g(D_i, D_i^{obs}, \frac{D_i^{obs}}{4})$$

where $\mathcal{F}^-$ is the set of all farms except the index case, t_i is the time of infection of farm i, $v(i)$ is the infector of farm i, and $g(x, m, w)$ is the probability density function of the gamma distribution with mean m and variation w.

In this application, the hosts we consider are actually farms. These farms themselves contain a large number of animals. Per farm, several animals were infected [8]. We can thus view the host (i.e. farm) to be strongly compartmentalized, resulting in most coalescents occuring just after infection. We therefore take the within-host pathogen effective population size times pathogen generation time W at time t in host h to be exponentially increasing

$$W(t, h) = e^{r(t-t_h)}$$

if h is infected at time t, 0 otherwise. We set the growth rate r at 1.02.

With this model, we get

$$\begin{aligned} L(P|T,W) &= \prod_{x\in\mathcal{F}}\prod_{[\tau_1,\tau_2]\in C_x} W(\tau_1,x)^{-\mathbb{1}_{coal}} e^{-\binom{n_\tau}{2}\int\limits_{\tau_1}^{\tau_2}\frac{1}{W(t,x)}dt} \\ &= \prod_{x\in\mathcal{F}}\prod_{[\tau_1,\tau_2]\in C_x} e^{-(r(\tau_1-t_x)\mathbb{1}_{coal})} e^{-\binom{n_\tau}{2}(\frac{1}{r}e^{-r(\tau_1-t_x)}-\frac{1}{r}e^{-r(\tau_2-t_x)})} \end{aligned}$$

where τ_1 and τ_2 are the start and end of coalescent intervals.

Following [20] we use a one parameter substitution model, yielding

$$p(D_G|P,\mu) = \prod_{bases}\sum_{\{A,C,G,T\}^N}\prod_{edges}(1-e^{-\mu t})^{\mathbb{1}_{mut}} + (e^{-\mu t})^{(1-\mathbb{1}_{mut})}$$

which can be efficiently evaluated using Felsenstein's pruning algorithm [12]. The prior for the substitution rate μ is uniformly distributed on (0,1000).

F.2.3 Results

Posterior distributions for the transmission tree and epidemiological parameters (α_2, β_2) are given in figure S1. To allow for comparison to previous studies, we also show the two transmission trees estimated before [8, 20], and the point estimates previously obtained for the parameters [20].

F.3 Sampling from the joint posterior distribution using MCMC

F.3.1 Initial state

As an initial state, a random permissable state was chosen, as follows. For each host x but the first, we chose a random infector $v(x)$ from the set of hosts infectious at the time of infection of x. We then constructed a phylogenetic tree consistent with this transmission tree (thus ignoring sequence data). Inital values for parameters were randomly sampled from their prior distributions.

F.3.2 Update phylogenetic tree

We update the phylogenetic tree P per host. We choose a host x at random, and update both the topology of the part of P contained within x, and the timing of the internal nodes contained in x. The number of internal nodes n_{i_x} of P contained in x is equal to the number of pathogen lineages that coalesce within x minus one, which is equal to the number of sequences n_{s_x} sampled from x plus the number of hosts infected by x that have a sequence minus one. More formally, let $T(x)$ be the smallest subtree of T such that

- $x \in T(x)$,
- for any node i other than x, if its infector $v(i) \in T(x)$, then $i \in T(x)$.

So $T(x)$ is the subtree that consists of x and the hosts directly or indirectly infected by x. Let $l(T)$ be 1 if T contains at least one host which has at least one sampled sequence, 0 otherwise. Then the number of coalescent events c_x that happen within host x is

$$c_x = \max\{0, n_{s_x} - 1 + \sum_{y\in T}\mathbb{1}_{v(y)=x}l(T(y))\}$$

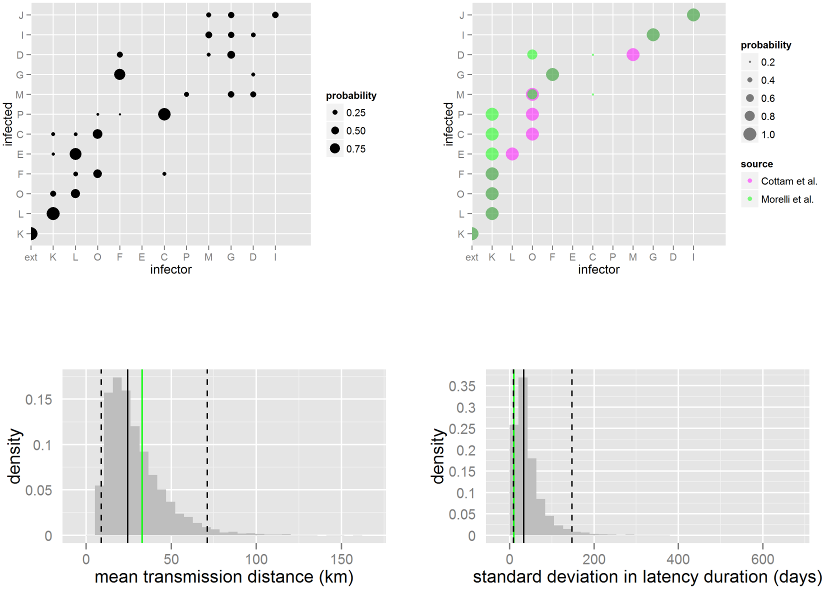


Figure S1. Posterior distribution for the transmission tree, mean transmission distance $2\alpha_2$ and standard deviation of latency duration β_2. (Top) Posterior distributions for the transmission tree, with infecing farm on the x-axis and infected farm on the y-axis. The size of the blobs corresponds to the posterior probability assigned to this pair. Estimates are given for the proposed method (left) and taken from previous studies (right) by Cottam et al. (pink) and Morelli et al. (green)[8, 20] which ignore within-host dynamics. The transmission tree cannot be established with high certainty for this outbreak, possibly due to unobserved infected farms. Allowing for within-host dynamics captures this uncertainty, in contrast to previous analyses (top right) which yield trees mainly containing pairs with posterior probability close to one. (Bottom) Posterior distributions for the mean transmission distance $2\alpha_2$ (left) and standard deviation of latency duration β_2 (right). Lines are point estimate (black solid), 95% credibility interval (black striped) and previous estimates obtained by Morelli et al. (green)

If $c_x = 0$, we do nothing. Otherwise, we propose a phylogenetic tree P^* which differs from P only in the nodes contained in x, where we sample these nodes and their topology from the probability density function

$$f(x) = \prod_{[\tau_1,\tau_2]\in C_x} W(\tau_1, x)^{-\mathbb{1}_{coal}} e^{-\binom{n_\tau}{2}\int\limits_{\tau_1}^{\tau_2} \frac{1}{W(t,x)}dt}$$

where C_x is the set of all intervals $[\tau_1, \tau_2]$ where the number n_τ of viral lineages within host x with sampled offspring is constant and larger than 1. We then accept P^* with probability

$$\begin{aligned} p(accept) &= \min\{1, \tfrac{p(D_E|T,\theta,W)p(D_G|P^*,\mu)p(P^*|T,W)\pi(T,\theta,W,\mu)}{p(D_E|T,\theta,W)p(D_G|P,\mu)p(P|T,W)\pi(T,\theta,W,\mu)} \tfrac{q(P|P^*)}{q(P^*|P)}\} \\ &= \min\{1, \tfrac{p(D_G|P^*,\mu)f^*(x)}{p(D_G|P,\mu)f(x)} \tfrac{f(x)}{f^*(x)}\} \\ &= \min\{1, \tfrac{p(D_G|P^*,\mu)}{p(D_G|P,\mu)}\} \end{aligned}$$

where $\frac{q(P|P^*)}{q(P^*|P)}$ is the Metropolis-Hastings ratio.

F.3.3 Update transmission tree

We randomly choose a host x. If x is not the index case, let $v(x)$ be its infector. We propose a new infector $v^*(x)$ randomly from the set of hosts that are

- infectious at time of infection t_x,
- not $v(x)$,
- not x (although this is usually already ensured by the first condition).

We could then try to construct the new transmission tree T^* generated by accepting $v^*(x)$ as the new infector of x. Note that T^* contains no cycles, as hosts are never infectious before being infected.

However, the phylogenetic tree will in general no longer correspond to $T*$. More precisely, if both $l(T(x)) = 1$ and $l(T - T(x)) = 1$, then there must be an edge of the phylogenetic tree with one node contained in a host in $T - T(x)$, and one in $T(x)$. If all hosts have sequences, the nodes are contained in $v(x)$ and x. In general, this edge causes P to be inconsistent with the new tree T^*. We therefore simultaneously propose a new tree P^* that is consistent with T^*, by removing the older of the two nodes of this edge from its current location, and relocating it to the first host y in the infection chain leading up from $v^*(x)$ (i.e. $\{v^*(x), v(v^*(x)), v(v(v^*(x)))\}$) that contains at least one pathogen lineage. This is a particular form of a pruning and regrafting operator [13], also see figure S2. If all hosts are sequenced, $y = v^*(x)$. We then update the phylogenetic tree contained in y as described above. The proposed pair of trees (T^*, P^*) is accepted with probability

$$\begin{aligned} p(accept) &= \min\{1, \tfrac{p(D_E|T^*,\theta,W)p(D_G|P^*,\mu)p(P^*|T^*,W)\pi(T^*,\theta,W,\mu)}{p(D_E|T,\theta,W)p(D_G|P,\mu)p(P|T,W)\pi(T,\theta,W,\mu)} \tfrac{q(T,P|T^*,P^*)}{q(T^*,P^*|T,P)}\} \\ &= \min\{1, \frac{s(o_x-t_x)\frac{I_x(v^*(x)|\theta,D_E)}{\sum_{y\in H} I_x(y|\theta,D_E)}p(D_G|P^*,\mu)f^*(y)f^*(y^*)\pi(T^*)}{s(o_x-t_x)\frac{I_x(v(x)|\theta,D_E)}{\sum_{y\in H} I_x(y|\theta,D_E)}p(D_G|P,\mu)f(y)f(y^*)\pi(T)} \frac{f(y)f(y^*)}{f^*(y)f^*(y^*)}\} \\ &= \min\{1, \tfrac{I_x(v^*(x)|\theta,D_E)p(D_G|P^*,\mu)}{I_x(v(x)|\theta,D_E)p(D_G|P,\mu)}\} \end{aligned}$$

as we take $\pi(T)$ equal for all trees.

If x is the index case, we let $v^*(x)$ be the first host it infects, and we switch infection times for the two hosts [20]. We then follow the procedure above.

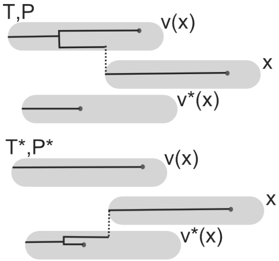


Figure S2. Example of a proposal transmission tree/phylogenetic tree (T^*, P^*) pair. A new infecting host $v^*(x)$ is proposed for host x. A new phylogenetic tree is then also proposed, by removing the branch from $v(x)$ and drawing a new phylogenetic subtree for $v^*(x)$ which includes the branch.

F.3.4 Update infection times

If not all infection times are observed, pick a host x at random and propose a new infection time t_x^*, sampled from the latency period distribution s. This creates the proposal transmission tree T^*. If $v(x)$ is not infectious at t_x^*, or T^* is inconsistent with P, reject. If not, accept with probability

$$\begin{aligned}
p(accept) &= \min\{1, \frac{p(D_E|T^*,\theta,W)p(D_G|P,\mu)p(P|T^*,W)\pi(T^*,\theta,W,\mu)}{p(D_E|T,\theta,W)p(D_G|P,\mu)p(P|T,W)\pi(T,\theta,W,\mu)}\frac{q(T|T^*)}{q(T^*|T)}\} \\
&= \min\{1, \frac{s(o_x-t_x^*)\frac{I_x^*(v(x)|\theta,D_E)}{\sum_{y\in H} I_x^*(y|\theta,D_E)}f^*(x)f^*(v(x))\pi(T^*)}{s(o_x-t_x)\frac{I_{t_x}(v(x)|\theta,D_E)}{\sum_{y\in H} I_{t_x}(y|\theta,D_E)}f(x)f(v(x))\pi(T)}\frac{s(o_x-t_x)}{s(o_x-t_x^*)}\} \\
&= \min\{1, \frac{\frac{I_x^*(v(x)|\theta,D_E)}{\sum_{y\in H} I_x^*(y|\theta,D_E)}f^*(x)f^*(v(x))}{\frac{I_{t_x}(v(x)|\theta,D_E)}{\sum_{y\in H} I_{t_x}(y|\theta,D_E)}f(x)f(v(x))}\}
\end{aligned}$$

F.3.5 Update epidemiological parameters

We create a proposal θ^* by choosing i from 1,...,$|\theta|$ and adding $Y \sim Normal(0, \sigma_i)$ to θ_i. The value of σ_i is calibrated in the burn-in period. We accept with probability

$$\begin{aligned}
p(accept) &= \min\{1, \frac{p(D_E|T,\theta^*,W)p(D_G|P,\mu)p(P|T,W)\pi(T,\theta^*,W,\mu)}{p(D_E|T,\theta,W)p(D_G|P,\mu)p(P|T,W)\pi(T,\theta,W,\mu)}\frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}\} \\
&= \min\{1, \frac{s(o_x-t_x)\frac{I_{t_x}(v(x)|\theta^*,D_E)}{\sum_{y\in H} I_{t_x}(y|\theta^*,D_E)}\pi(\theta^*)}{s(o_x-t_x)\frac{I_{t_x}(v(x)|\theta,D_E)}{\sum_{y\in H} I_{t_x}(y|\theta,D_E)}\pi(\theta)}\}
\end{aligned}$$

F.3.6 Update mutational parameters

We create a proposal μ^* by choosing i from 1,...,$|\mu|$ and adding $Y \sim Normal(0, \delta_i)$ to μ_i. The value of δ_i is calibrated in the burn-in period. We accept with probability

$$\begin{aligned}
p(accept) &= \min\{1, \frac{p(D_E|T,\theta,W)p(D_G|P,\mu^*)p(P|T,W)\pi(T,\theta,W,\mu^*)}{p(D_E|T,\theta,W)p(D_G|P,\mu)p(P|T,W)\pi(T,\theta,W,\mu)}\frac{q(\theta|\theta^*)}{q(\mu^*|\mu)}\} \\
&= \min\{1, \frac{p(D_G|P,\mu^*)\pi(\mu^*)}{p(D_G|P,\mu)\pi(\mu)}\}
\end{aligned}$$

F.3.7 Convergence

For all analyses, we ran an inital burn-in period of 10^5 iterations. Convergence of parameters was checked by eye. The variance of the proposal distributions of the parameters was adjusted during this time, such that the acceptance probability was around 0.5. The chain was then run for 3×10^5 iterations, and sampled every 200th iteration. This yields a chain of 1500 samples from the posteriod distribution. Increasing these values gave no observable difference in results.

F.3.8 Estimation

Point estimates of parameters were obtained as the median of the posterior distribution sampled from the Markov chain. 95% credibility interval boundaries were computed as the 2.5th and 97.5th percentile.

References

[1] A. Bataille, F. van der Meer, A. Stegeman, and G. Koch. Evolutionary analysis of inter-farm transmission dynamics in a highly pathogenic avian influenza epidemic. *PLoS Pathogens*, 7(6), 2011.

[2] N. Becker. On parametric estimation for mortal branching processes. *Biometrika*, 61(3):393–399, 1974.

[3] G.J. Boender, T.J. Hagenaars, A. Bouma, G. Nodelijk, A.R.W. Elbers, M.C.M. De Jong, and M. Van Boven. Risk maps for the spread of highly pathogenic avian influenza in poultry. *PLoS Computational Biology*, 3(4):704–712, 2007.

[4] M.E.H. Bos, M. Van Boven, M. Nielen, A. Bouma, A.R.W. Elders, G. Nodelijk, G. Koch, A. Stegeman, and M.C.M. De Jong. Estimating the day of highly pathogenic avian influenza (h7n7) virus introduction into a poultry flock based on mortality data. *Veterinary Research*, 38(3):493–504, 2007.

[5] S. Cauchemez, F. Carrat, C. Viboud, A.J. Valleron, and P.Y. Boëlle. A bayesian mcmc approach to study transmission of influenza: Application to household longitudinal data. *Statistics in Medicine*, 23(22):3469–3487, 2004.

[6] S. Cauchemez and N.M. Ferguson. Methods to infer transmission risk factors in complex outbreak data. *J R Soc Interface*, 9(68):456–469, 2012.

[7] E.M. Cottam, D.T. Haydon, D.J. Paton, J. Gloster, J.W. Wilesmith, N.P. Ferris, G.H. Hutchings, and D.P. King. Molecular epidemiology of the foot-and-mouth disease virus outbreak in the united kingdom in 2001. *Journal of Virology*, 80(22):11274–11282, 2006.

[8] E.M. Cottam, G. Thébaud, J. Wadsworth, J. Gloster, L. Mansley, D.J. Paton, D.P. King, and D.T. Haydon. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings of the Royal Society B: Biological Sciences*, 275(1637):887–895, 2008.

[9] Spekreijse D., Bouma A., Stegeman J.A., Koch G., and de Jong M.C.M. The effect of inoculation dose of a highly pathogenic avian influenza virus strain h5n1 on the infectiousness of chickens. *Veterinary Microbiology*, 147:59–66, 2010.

[10] de Jong M.C., Stegeman A., van der Goot J., and Koch G. Intra- and interspecies transmission of h7n7 highly pathogenic avian influenza virus during the avian influenza epidemic in the netherlands in 2003. *Rev. Sci. Tech.*, 28:333–340, 2009.

[11] C.P. Farrington, M.N. Kanaan, and N.J. Gay. Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics*, 4(2):279–295, 2003.

[12] J. Felsenstein. Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[13] J. Felsenstein. *Inferring Phylogenies.* Sinauer, Sunderland, MA, 2004.

[14] N.M. Ferguson, C.A. Donnely, and Andersion R.M. The foot-and-mouth epidemic in great britain: Pattern of spread and impact of interventions. *Science*, 292:1155–1160, 2001.

[15] N.I. Fisher and A.J. Lee. A correlation coefficient for circular data. *Biometrika*, 70(2):327–332, 1983.

[16] Thornley J.H.M. and France J. Modelling foot and mouth disease. *Prev. Vet. Med.*, 89:139–154, 2009.

[17] M. Jonges, A. Bataille, R. Enserink, A. Meijer, R.A.M. Fouchier, A. Stegeman, G. Koch, and M. Koopmans. Comparative analysis of avian influenza virus diversity in poultry and humans during a highly pathogenic avian influenza a (h7n7) virus outbreak. *Journal of Virology*, 85(20):10598–10604, 2011.

[18] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120, 1980.

[19] J. Liu, S.L. Lim, Y. Ruan, A.E. Ling, L.F.P. Ng, C. Drosten, E.T. Liu, L.W. Stanton, and M.L. Hibberd. Sars transmission pattern in singapore reassessed by viral sequence variation analysis. *PLoS Medicine*, 2:0162–0168, 2005.

[20] M.J. Morelli, G. Thébaud, J. Chadœuf, D.P. King, D.T. Haydon, and S. Soubeyrand. A bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Computational Biology*, 8(11), 2012.

[21] Shaw M.W., Harwood T.D., Wilkinson M.J., and Elliot L. Assembling spatially explicit landscape models of pollen and spore dispersal by wind for risk assessment. *Proc. R. Soc. B*, 273:1705–1713, 2006.

[22] Savill N.J., Shaw D.J., Deardon R., Tildesley M.J., Keeling M.J., Woolhouse M.E.J., Brooks S.P., and Grenfell B.T. Topographic determinants of foot and mouth disease transmission in the uk 2001 epidemic. *BMC Veterinary Research*, 2, 2006.

[23] Portnoy S. and Willson M.F. Seed dispersal curves: behaviour of the tail of the distribution. *Evolutionary Ecology*, 7:25–44, 1993.

[24] J.H. Srensen, D.K.J. Mackay, C. Jensen, and A.I. Donaldson. An integrated model to predict the atmospheric spread of foot-and-mouth disease virus. *Epidemiology and Infection*, 124(3):577–590, 2000.

[25] A. Ssematimba, T.J. Hagenaars, and M.C.M. De Jong. Modelling the wind-borne spread of highly pathogenic avian influenza between farms. *PLoS One*, 2012.

[26] R.J.F. Ypma, A. Bataille, A. Stegeman, G. Koch, J. Wallinga, and W.M. van Ballegooijen. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proc Biol Sci*, 279(1728):444–450, 2012.