UPTEC X 22008

Examensarbete 30 hp

Juni 2022

Exploring methods for detecting super-spreaders using molecular data

A literature study and case study of VTEC O157:H7 in dairy calves

Amanda Wallskog

Civilingenjörsprogrammet i molekylär bioteknik

Exploring methods for detecting super-spreaders using molecular data – A literature study and case study of VTEC O157:H7 in dairy calves

Amanda Wallskog

Abstract

Verotoxin-producing *Escherichia coli* (VTEC) of serotype O157:H7 is a pathogen causing illness in humans worldwide. The path and nature of transmission from and among cattle is important knowledge when it comes to preventing cases of disease in humans. Two concepts potentially playing an important role in transmission of VTEC O157:H7 are super-shedding and super-spreading. Super-shedders are individuals (here calves) shedding a high amount of bacteria. Super-spreaders are individuals (here calves) spreading the disease in a higher extent compared to the rest of the population investigated. Little is known about these phenomenons' effect on transmission as well as the relation between them. Therefore, it is important to investigate this further.

The purpose of this master thesis was to get a better understanding of how super-spreaders can be identified. One way to identify super-spreaders and explore the transmission of a pathogen is to investigate molecular data using computational methods. Here, a literature study with a systematic approach was conducted in order to scan the literature for such methods. In this first phase of the master thesis three methods, all constructing transmission trees, were identified as relevant methods for the second phase. These methods are called outbreaker2, phybreak and TransPhylo.

In the second phase of the master thesis, 32 whole genome sequences of VTEC O157:H7 collected from four different cattle farms were investigated using the methods outbreaker2 and phybreak. Both methods were able to identify samples infecting more secondary cases compared to the rest of the investigated population. Some of these samples came from the environment, possibly shedding light on the importance of the pathogen's ability to survive outside of the host, and therefore playing an important role in transmission of the disease. The rest of the samples infecting more secondary cases were from calves, and a minority of these were super-shedders. From this the importance of the relation between super-shedders and super-spreaders can neither be confirmed nor denied.

Outbreaker2 suggested that the spread of the pathogen is frequently occurring between the four neighbouring farms, while phybreak instead suggested that the spread mostly occurs within the farms. From this, a scenario explaining that the transmission possibly occurs within farms is presented.

Teknisk-naturvetenskapliga fakulteten

Uppsala universitet, Utgivningsort Uppsala

Handledare: Lena-Mari Tamminen Ämnesgranskare: Ulf Emanuelson

Examinator: Johan Åqvist

Populärvetenskaplig sammanfattning

Superspridare är ett begrepp som i spåren av coronapandemin knappast undgått någon. En superspridare definieras som en individ som sprider smittan vidare till fler jämfört med andra infekterade individer. Då en superspridande individ kan ligga bakom en betydande del av smittspridningen av ett sjukdomsutbrott är det viktigt att kunna identifiera dessa, för att i sin tur kunna begränsa smittspridningen. Ett annat begrepp som ibland nämns i samband med superspridare är superutsöndrare. Dessa individer utsöndrar en stor mängd smittoämne, och det har föreslagits att individer som utsöndrar i hög grad löper större risk att bli en superspridare.

Syftet med detta examensarbete var att få en bättre förståelse för hur superspridare kan bli identifierade. Den första fasen bestod av en litteraturstudie med en systematisk ansats. Där undersöktes det vilka slags metoder som forskare använder sig av för att undersöka förekomsten av superspridare vid utbrott av olika infektionssjukdomar. Den andra fasen bestod av att applicera relevanta metoder på ett dataset med helgenomsekvenser av Verotoxinproducerande *Escherichia coli* (VTEC) O157:H7, från fyra svenska gårdar med nötkreatur. Helgenomsekvenser är den data som beskriver en organisms arvsmassa. Arvsmassan består av deoxyribonukleinsyra (DNA) som är en unik kod uppbyggd av fyra olika baser; A, T, C och G. Genom denna unika kod kan man särskilja på olika arter, men även på individer inom en art.

I litteraturstudien identifierades tre lovande metoder för att undersöka förekomsten av superspridare; outbreaker2, phybreak och TransPhylo. Dessa metoder använder helgenomsekvenser och datum för provtagning för att generera så kallade transmission trees. Transmission trees beskriver smittspridningstillfällen mellan infekterade värdar och kan på så sätt avslöja vem som har smittat vem. I den andra fasen applicerades helgenomsekvenserna från VTEC-proverna, samt datum för dess provtagning, i outbreaker2 och phybreak. Även om outbreaker2 och phybreak har samma input och output finns det vissa skillnader metoderna emellan. Den största skillnaden är att phybreak tar hänsyn till att genomuppsättningen av patogenerna inom en infekterad värd är mångfaldig, samtidigt som outbreaker2 antar att alla patogener har samma genomuppsättning.

Resultatet från outbreaker2 föreslog att smittspridningen mestadels sker mellan de olika provtagna gårdarna, medan phybreak föreslog att smittspridningen mestadels sker inom gårdarna. Båda metoderna identifierade potentiella superspridare och av dessa totalt elva superspridare var två superutsöndrare. Det faktum att det finns stora skillnader mellan metoderna gör att resultaten om individuell smittspridning skall tolkas med försiktighet. Därför kan inte heller relationen om superspridning och superutsöndring fastställas eller dementeras i detta examensarbete.

Contents

Abbreviations

DNA	Deoxyribonucleic acid
E. coli	*Escherichia coli*
EHEC	Enterohemorrhagic *Escherichia coli*
HUS	Hemolytic-uremic syndrome
MCMC	Markov chain Monte Carlo
M. tuberculosis	*Mycobacterium tuberculosis*
PEO	Population, Exposure, Outcome
PICO	Population, Intervention, Comparison, Outcome
SARS-CoV-2	Severe Acute Respiratory Syndrome Corona Virus 2
SNP	Single nucleotide polymorphism
STEC	Shiga toxin-producing *Escherichia coli*
TMRCA	The most recent common ancestor
VTEC	Verotoxin-producing *Escherichia coli*
WGS	Whole Genome Sequencing
WoS	Web of Science

1 Introduction

Verotoxin-producing *Escherichia coli* (VTEC) of serotype O157:H7 is a zoonotic pathogen causing illness worldwide. This type of *E. coli* causes symptoms like diarrhea in humans, which sometimes can develop into the severe condition hemolytic-uremic syndrome (HUS). (Kaper & O'Brien, 2014) Humans most often get infected via contaminated food or via direct environmental exposure. In order to establish health-care measurements to prevent illness, more knowledge about the route and nature of the transmission of the pathogen is needed. The usual reservoir for VTEC O157:H7 is cattle and the calves and cows carrying the pathogen are usually asymptomatic, making it difficult to detect the presence of the pathogen and therefore hinder transmission. (L. Gally & P. Stevens, 2016)

One way to investigate the transmission of a pathogen is to explore the sequences of the whole genome of the pathogen. All organisms has a genome consisting of a unique code, called deoxyribonucleic acid (DNA). DNA consists of four bases; A, T, C and G. Whole genome sequencing (WGS) is a method that can determine the order of these bases, revealing the unique genome sequence of an organism. (CDC, 2019b) WGS methods have become much faster and cost efficient in the last decades, making them more accessible (NHGRI, 2021). In response to a lot of data being generated, methods used to analyze this kind of data are being continuously developed and more routinely used for outbreak investigations.

2 Background

2.1 Introduction to VTEC

VTEC O157:H7 is as mentioned before a zoonotic pathogen causing illness in humans worldwide. It is a bacteria that causes diarrhea, sometimes with severe complications like HUS. This syndrome is characterized by three components; renal failure, hemolytic anemia and thrombocytopenia (Nataro & Kaper, 1998). This pathogen was first discovered in the late 70s and early 80s and it was discovered independently by two different research groups. In 1977 Konowalchuk *et al.* discovered an *E. coli* that produced toxins against vero cells, therefore calling it verotoxin-producing *E. coli* (VTEC) Konowalchuk *et al.* (1977). In 1983 O'Brien & LaVeck discovered the same toxin, but they discovered that it could be neutralized with antitoxin prepared against Shiga toxin produced from *Shigella dysenteriae*, making them call the pathogen Shiga toxin-producing *E. coli* (STEC) O'Brien & LaVeck (1983). O'Brien *et al.* (1983) later showed that these toxins in fact were the same, and since then the names has been used interchangeably (O'Brien *et al.*, 1983). For the rest of this report the term VTEC will be used since that is traditionally the term that has been used within the veterinary field.

Another term frequently used when talking about pathogenic *E. coli* is enterohemorrhagic *E. coli* (EHEC). EHEC is a group of *E. coli* causing bloody diarrhea and severe complications, such as HUS, in humans. EHEC is therefore a subset to VTEC, since all EHEC are considered to be pathogens, but not all VTEC are considered pathogens (Nataro & Kaper, 1998).

VTEC consists of different subtypes, and in 2008 Manning *et al.* defined nine evolutionary different clades using a single nucleotide polymorphism (SNP) typing system (Manning *et al.*, 2008). One of the subtypes that causes severe disease in humans is the subtype clade 8. Clade 8 is the main cause of food borne outbreaks in Sweden, making it an important subtype to keep track of. (Söderlund *et al.*, 2014).

2.2 Reservoir and Transmission of VTEC

The usual reservoir for VTEC O157:H7 is cattle. The infected animals do not show any symptoms of disease, which makes it difficult to detect if the pathogen is present on a farm. Humans can be infected via direct or indirect contact with an infected animal, from the environment or from contaminated food or water. Transmission from person to person could also occur. (Ameer *et al.*, 2021) Therefore, the transmission among cattle plays an important role for the transmission to humans as well. The prevalence of VTEC O157:H7 on farms in Sweden has increased which contributes to a higher risk of infection in humans. (Eriksson *et al.*, 2005; Söderlund *et al.*, 2014). By reducing the transmission among cattle on farms, the transmission to humans could also be reduced.

The transmission among cattle has been associated with several different risk factors. Examples of such risk factors are; the introduction of new animals to an already existing herd (Wilson *et al.*, 1998), redeployment of individuals within a herd of animals (Chase-Topping *et al.*, 2008), densely grouped animals, large groups of animals, (Vidovic & Korber, 2006) the spread of slurry on grazing land and having pigs in close contact with a herd of cattle (Gunn *et al.*, 2007). Even though several risk factor have been identified, more knowledge about the transmission among cattle is needed in order to reduce the transmission, both between cattle but also to humans.

2.3 Super-shedders vs Super-spreaders

Super-shedders, i.e. cattle shedding more than 10^3 colony forming units/g feces, have been proposed to play an important role in the prevalence and transmission of VTEC O157:H7 (Spencer *et al.*, 2015). It is not yet known exactly what factors are responsible for super-shedding, although several studies have been performed to understand the phenomenon (D. Munns *et al.*, 2015). While the concept of super-shedding has received attention, another concept also playing a potentially important role in the transmission of VTEC, has received less attention; super-spreaders. Super-spreaders are defined as individuals that have more opportunities to infect others with some kind of pathogen, both through direct and indirect contact (D. Munns *et al.*, 2015). The concepts are closely related, for example super-shedding may increase the risk of super-spreading, but it may be possible to be a super-shedder but not a super-spreader. Individual differences, such as behavior or social contacts, may be more important than the amount of bacteria shed, for becoming a super-spreader.

A model related to this subject is the so called 20/80 rule, which suggests that 20% of a host population contributes to 80% of the spread of some infectious disease (Woolhouse *et al.*, 1997). This is a rule that seems to be applicable to a variety of different infectious diseases, and in the case with the transmission of VTEC O157:H7 it would mean that 20% of the calves shedd 80% of the bacteria. This brings additional light on the importance of being able to identify super-shedders as well as super-spreaders in order to control and hinder the transmission of infectious diseases.

2.4 Transmission Trees

Sequences from the genome of pathogens together with epidemiological data can be used to create so called transmission trees. These kind of trees can provide information on how strains of the pathogen, sampled from different individuals, is related. Therefore information of who infected whom can be interpreted as well as information on the nature of the outbreak. Transmission trees reveals the history of the host, carrying the pathogen, which means that they illustrate the event of transmission that occurs between a primary and secondary case of infection. This tree differs from a phylogenetic tree, which instead describes the ancestral relationship of pathogens that are sampled and do not reveal who infected whom. (Didelot *et al.*, 2021) These different trees should therefor not be confused. The major differences between these trees is how the nodes and leaves should be interpreted. In transmission trees the nodes correspond to the event of transmission (where a leaf is only transmitted but does not transmit), while in a phylogenetic tree the leaves corresponds to the sampled pathogens and the internal nodes corresponds to the most recent common ancestor (TMRCA). Another difference is that the timing of the nodes in a transmission tree reveals the time point of transmission, whereas in phylogenetic trees it corresponds to the coalescent events the takes place before the transmission. (Ypma *et al.*, 2013)

2.5 Methods for Generation of Transmission Trees

Two methods used to create transmission trees are outbreaker2 and phybreak. These are flexible and relatively recently developed methods designed to manage a variety of different pathogens. The packages are available in R (R Core Team, 2021) and use sequencing data and sampling dates as their input while the output is a transmission tree.

2.5.1 outbreaker2

Outbreaker2 (Campbell *et al.*, 2018) is an extended version of outbreaker (Jombart *et al.*, 2014) and it is a discrete-time stochastic model that construct a transmission tree (the output) based on the genetic data of some pathogen and date of collection (the input). This framework is flexible and it provides a tool that researchers can use to reconstruct outbreaks of some infectious pathogen. It is a method implemented in a Bayesian framework, taking in different parameters, likelihood and movement functions and describe prior distributions for these. It uses a Markov chain Monte Carlo (MCMC) method to update the parameters implemented in the Bayesian framework and does so from one iteration to the other. Advantages of this method is that it can infer different R numbers (the basic reproduction number describing how many individuals that can be infected by an infectious individual (Holme & Masuda, 2015)) at an individual level, allowing the discovery of heterogeneous transmission and therefor the detection of super-spreaders. Another advantage is that this method allows for multiple introductions of the pathogen to some population. Introductions can be identified as genetic outliers, the problem though is that they could also arise due to some other reason, such as sequencing errors or recombination. This means that the interpretation of this should be made carefully. Also assuming that all different introductions of the pathogen is genetically different is not correct, since they could come from closely related lineages. A way of dealing with this in outbreaker2 is the feature that known introduction cases can be fixed. Additional limitation of this method is that it requires the outbreak to be densely sampled (cannot detect unobserved cases), and is not suitable to apply to diseases where asymptomatic carriers are a common phenomenon. Other limitations is that the method assumes a single pathogen lineage within the host as well as that the date of the event of transmission is not inferred. (Campbell *et al.*, 2018)

2.5.2 phybreak

Phybreak is a method that, like previous described method, uses genetic data and sampling dates as input, while the output is a transmission tree. It also uses a Bayesian framework and MCMC as previously described. Out of the two methods described, phybreak is the most novel and refined. It takes into account the transmission, case observation, within-host pathogen dynamics and mutation, which outbreaker2 does not. The transmission model assumes that all cases have been sampled as well as that the outbreak is over, and therefore the mean number of secondary cases is one. To interpret this in phybreak, a gamma distribution of the generation period, the time interval between a primary and secondary case of infection, is constructed. The case observation parameter also consists of a gamma distribution for the sampling interval, which is the time period between the onset of the infection and the sampling. The within-host pathogen dynamics is a model that lets us model the dynamic of the pathogen within the host, and it is modeled to simulate coalescence events. Therefore assuming clonal lineages and eventually this model ends up in a bottleneck at transmission of one lineage. The mutation model is a parameter that takes in the mutations rate per site per time unit and is based on a Jukes-Cantor model. (Klinkenberg *et al.*, 2017)

Advantages with phybreak is that it is a fast method and that it can be used for different kinds of pathogens as well as several different settings. Disadvantages with this method is that, as mentioned with the transmission parameter, all cases needs to be sampled and the outbreak has to have come to an end. If not all cases have been sampled and included, this could limit the identification of transmission clusters, and paths of transmission could be obscured due to missing cases in the data. The method does not either take heterogeneous infectiousness between individuals into account. (Klinkenberg *et al.*, 2017)

2.6 Project Aims

This master thesis was performed at the Swedish University of Agricultural Sciences (SLU) in Uppsala. The purpose of the project was to get a better understanding of how super-spreaders can be identified. This was done in two steps. Phase one included reviewing previous efforts to investigate the occurrence of super-spreaders. The aim of this part was to provide an overview of previous studies investigating super-spreading using molecular methods (pathogen examined and methods used). This was done through a literature study based on a systematic approach. The second phase included using whole genome sequence data from an outbreak of VTEC O157:H7. The aim of this second part was to explore the presence of super-spreaders of VTEC O157:H7 in this dataset to provide insight on whether super-spreading is a phenomenon that one specific individual obtains or if it is a state that different individuals can obtain at different time points during an outbreak. The results from the literature study were also used to provide input for choice of methods in the second phase of the study.

3 Methods

3.1 Phase 1 - Literature Study with a Systematic Approach

The first phase of this master thesis was to conduct a literature study with a systematic approach. The course of action is visualized as a flowchart in Figure 1.

3.1.1 Formulate Research Question

The very first step of the literature search was to define the research question. To be able to define this, some simple searches were performed and a variety of articles on the subject were read. This preparatory searching and reading did not only contribute to formulating a research question but also gave a good perception on what search terms to use. In order to formulate a demarcated and well defined research question, different frameworks can be used. Two popular frameworks are PICO (Population, Intervention, Comparison, Outcome) and PEO (Population, Exposure, Outcome) where the first one is usually used for quantitative questions and the later for qualitative questions (Karolinska Institutet, 2022). None of these frameworks could be followed meticulously, due to that not all parts of the frameworks could be applied to the research question, but were used as an inspiration for the formulation of the research question, presented below.
Research question: What characterizes studies investigating super-spreading in humans or animals of any infectious pathogen using molecular sequence data?

3.1.2 Find Search Terms, Create Search Blocks and Conduct the Search

Search terms for the search were decided on from doing multiple different searches, so called test searches. Reading particularly important articles for the subject, called key articles, also contributed to finding the suitable search terms. For this subject, the articles describing ”outbreaker” (Jombart *et al.*, 2014) and ”phybreak” (Klinkenberg *et al.*, 2017) were considered key articles, since they describe the methods to be used unless more specific and relevant methods appeared from the literature study. Also investigating words that are important and frequently occurring in relevant abstracts and titles laid the foundation in building the search. Lastly all terms that were decided to be used for the search were placed in so called search blocks, in order to achieve a structured final search and these blocks were combined with boolean operators, such as AND and OR. (Karolinska Institutet, 2022)

The final search, with the corresponding search blocks, that were carried out in PubMed and Web of Science (WoS) (all databases), can be found in Table 1 and 2 in Appendix A. These two databases were chosen due to that they are relevant for this subject, as well as both can be accessed through Uppsala University. Both searches were performed in the advanced search field in the respective databases. This made the search more specific by searching for MeSH terms and for Title/Abstract in PubMed as well as for Topic (TS) in WoS. In order to find more variants of a word, truncation using asterisks were used. The search was performed on February 22nd 2022. No filters were applied. The full list of articles resulting from the search can be found in Appendix B.

3.1.3 Define Inclusion and Exclusion Criteria

To define, delimit and decide what articles to include to the more thorough examination, called characterization (3.1.5), inclusion and exclusion criteria were decided. These criteria were used as guidelines when scanning through the results from the search. (Redaktionen, 2022)

The inclusion and exclusion criteria used for this study were the following;
Inclusion: Studies that aim to identify super-spreaders of infectious pathogens using genomic data.
Exclusion: Studies that only investigate models based on simulated outbreaks and that do not use real life genomic data.

3.1.4 Screening for Relevance

After the search the results were reviewed. In the first step of reviewing, a more simple screening for relevance was performed. This screening included going through the titles and abstracts of all hits from all databases. From here, based on relevance and the pre-defined inclusion and exclusion criteria, the articles were either included or excluded to the next step of validation, the characterization. A rule of thumb usually used for systematic literature searches is that about 10 % of the articles from the search should be of relevance and included (Redaktionen, 2022). See Table 3 in Appendix C for a list of what articles that were included or excluded.

3.1.5 Characterization

The second part of the reviewing was the characterization. This step was more comprehensive compared to the first step and for this all included articles were read as a whole. There were 14 articles included from the screening and therefore reviewed in the characterization. For this, a number of questions were asked and answered for every article, see Table 4 and 5 in Appendix D, to find different trends and patterns among the articles.

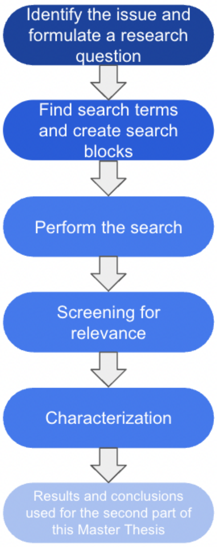

Identify the issue and formulate a research question
Find search terms and create search blocks
Perform the search
Screening for relevance
Characterization
Results and conclusions used for the second part of this Master Thesis

Figure 1: A flowchart representing the workflow for the literature study.

3.2 Phase 2 - Exploring Super-spreaders of VTEC O157:H7

In the second phase of the project a statistical analysis was performed in R using outbreaker2 and phybreak. Data from 32 cases, taken from four different cattle farms with an aggressive type of VTEC O157:H7, were used in both methods. The sequences used were previously sampled by Dr. Lena-Mari Tamminen and were a part of a larger dataset, with samples collected from a total of 12 cattle farms. This larger dataset has been used in a study investigating the behaviour of dairy calves and the animal welfare, in order to explore risk factors and what factors that possibly drives the colonisation of VTEC O157:H7 in calves (Tamminen *et al.*, 2020). However, these sequences has not previously been used to investigate the presence of super-spreaders, which is what will be investigated in this master thesis. Both samples from the environment and from calves were used. A table compiling this information, as well as information about if the calf was a shedder or not and the date of the sampling, can be found in Appendix E, Table 6.

Hybrid assemblies were generated from short and long reads. The hybrid assemblies were created using Unicycler v. 0.4.8 (Wick *et al.*, 2017) by applying default parameters. The short reads were trimmed using Trimmomatic v. 0.36 (Bolger *et al.*, 2014) and the long reads were trimmed using the short reads as a reference.

3.2.1 outbreaker2

In outbreaker2 both genetic information and temporal information about the outbreak is needed. Here, the 32 whole genome sequences of VTEC O157:H7 was stored as DNAbin objects and the sampling dates corresponding to the DNA sequences was stored as a vector of dates. In addition to this data, information about the distributions of the incubation period as well as the generation time of VTEC O157:H7 was needed. The distribution used to model these time periods are typically gamma distributions and were used here accordingly. (Thibaut, 2018) If nothing else is stated, the default values has been used to run outbreaker2.

The incubation period is defined as the time between exposure to the pathogen and symptom onset (Awofisayo-Okuyelu *et al.*, 2019). This kind of data do not exist for calves infected with VTEC O157:H7 since cattle do not show symptoms, but it does exist for when VTEC O157:H7 infects humans. Since we assume that the time it takes for the *E. coli* bacteria to establish itself in the gastrointestinal system is the same in humans as in calves, this data is used. For humans the incubation period is usually between 3-4 days, but can span from 1-10 days (CDC, 2019a), and this information was used to create a gamma distribution for the incubation period. A gamma distribution can be created using two parameters; the shape and scale parameters. These were generated from the mean values and standard deviation from the incubation period interval. For the incubation period the parameters were; shape = 0.30 and scale = 18.15. This distribution was then used in the function outbreaker_data and applied to the f_dens parameter. An additional distribution, using a shape value of 3 was also performed. This was done in order to be able to compare the trees generated from outbreaker2 with the trees generated from phybreak.

The generation time is defined as the time between a primary and secondary infection (Thibaut, 2018). The generation time of VTEC O157:H7 is estimated to be between 2-8 days (Spencer *et al.*, 2015). The mean and standard deviation from this interval was used to create the gamma parameters shape and scale, which were used to created a gamma distribution. For the generation time the parameters were; shape = 0.19 and scale = 26.79. This distribution was then used in the function outbreaker_data and applied to the w_dens parameter. An additional distribution, using a shape value of 3 was also performed. This was done in order to be able to compare the trees generate from outbreaker2 with the trees generated from phybreak.

Additional parameters that can be specified in the method can be found in the create_config function in outbreaker2 (Thibaut *et al.*, 2021). Here three parameters were specified for the run and the rest were set to default values. The first parameter to be specified was the number of iterations that should be used for the MCMC, and this was set to 10^5. The second parameter to specify was what tree should be used to initialize the MCMC in outbreaker. The tree ”random” was selected, meaning that the ancestors were randomly selected from the preceding cases. Lastly the initial value for the mutation rate of VTEC O157:H7 was specified. Different sources presents different values for this and thereby two different rates were used and compared. From Gibson *et al.* (2018) 1.44^{-7} mutations per site per year was presented as the mutation rate of *E. coli*. This number was divided by 365 days to get 3.95^{-10} mutations per site per day which were the number used in outbreaker2. Reeves *et al.* (2011) presented 2.26^{-7} mutations per site per year as the mutation rate for *E. coli*. This number was also divided by 365 days to get the correct number and unit; 6.19^{-10} mutations per site per day.

3.2.2 phybreak

In phybreak, just as in outbreaker2, genetic information and temporal information about the outbreak is needed. The same sequences and sampling times were used in phybreak as in outbreaker2. They were also stored in the same way, as a DNAbin and vector of dates respectively. In phybreak four additional parameters to the function also called phybreak has to be defined. These are the transmission model, the sampling model, the within-host model and the mutation model, all briefly described in the background. If nothing else is stated, the default values have been used to run phybreak.

The transmission model describes the time interval between a primary and secondary case of infection. This is the same definition as the generation time used in outbreaker2. Therefore the same interval, 2-8 days (Spencer *et al.*, 2015) was planned to be used for phybreak. This however did not work, due to that the shape value generated from the gamma distribution, which is the input needed in phybreak, could not be handled in the following generation of transmission trees. Due to this the default value for this parameter was used; gen_shape = 3. (Klinkenberg *et al.*, 2017)

The sampling model describes the time between onset of infection and the time for sampling, slightly different from the incubation period in outbreaker2, which was defined as the time between onset of infection and symptoms. For this project though they were assumed to be the same, and the interval 1-10 days was therefore planned to be used for phybreak. Unfortunately this value did not work either, due to that the shape value generated from the gamma distribution, which is the input needed in phybreak, could not be handled in the following MCMC iterations. Due to this the default value for this parameter was used as well; sample_shape = 3. (Klinkenberg *et al.*, 2017)

The within-host model is the model that describes the phylogenetic mini-trees inside each host, which is an interpretation of how the *E. coli* population grows over time within the host. For this project model number three was chosen. This model assumes that the within host population of *E. coli* grows linearly, and for this an initial value for that slope has to be provided. The values used were wh.model = 3 and wh.slope = 1.

The values used for the mutation model in phybreak were the same values used for the mutation rate in outbreaker2; 3.95^{-10} mutations per site per day (Gibson *et al.*, 2018) and 6.19^{-10} mutations per site per day (Reeves *et al.*, 2011).

After storing all the necessary data and setting all parameters (priors) needed, MCMC iterations were run in order to sample from the posterior. First a burnin (burnin.phybreak) iteration was run for 5000 cycles, and thereafter a sample (sample.phybreak) iteration were run for 25'000 cycles. The burnin only return a phybreak object (updated priors), while the sample method also samples from the chain and return the samples stored in the phybreak object (posterior). After the MCMC iterations, the results were analyzed and a transmission tree was created. First a summary of the phybreak object containing the posteriors were created using the "edmonds" method in the transtree function. Thereafter the "edmonds" method was once again used, but in the plotTrans function, in order to plot the created transmission tree. The method "edmonds" selects the infector (transmitting a pathogen) of a infectee (receiving a pathogen) that is most frequently sampled in the MCMC, therefore having the highest support. (Klinkenberg *et al.*, 2017)

4 Results

4.1 Phase 1 - Literature Study with a Systematic Approach

4.1.1 The Result in Numbers

From the literature study, a total of 74 articles were yielded from the search; 16 from Web of Science and 58 from PubMed. After removal of duplicates, a total of 62 articles were left for the screening for relevance. From the screening for relevance, 48 articles were excluded, based on the previously decided inclusion and exclusion criteria. Reason for exclusion could for example be that they did not investigate super-spreaders or that they only did some kind of modeling not using any real molecular data. See section 3.1.3 for inclusion and exclusion criteria and Appendix C for a list of what articles that were excluded/included. This left 14 articles for the characterization, about 23% of all articles. See Figure 2 for an illustration of the results from the literature study.

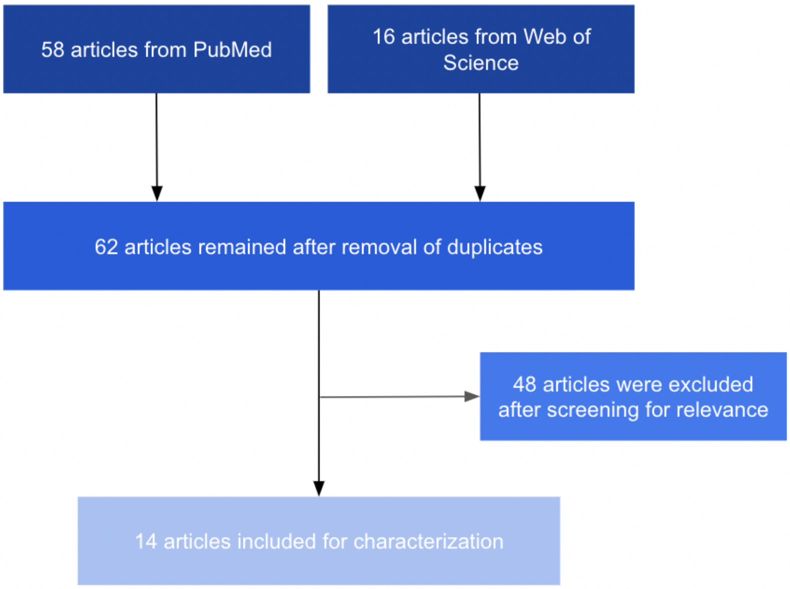

58 articles from PubMed
16 articles from Web of Science
62 articles remained after removal of duplicates
48 articles were excluded after screening for relevance
14 articles included for characterization

Figure 2: A flow diagram of the results from the literature study.

The two articles considered to be key articles were found in the search and were both included from the screening for relevance to the characterization.

The majority of the articles, 9 out of 14, were published 2020 or later, see Figure 3. This suggests that this is an area of research that is expanding and receiving more attention. This could be due to that this kind of data is more routinely being produced and investigated when an outbreak of some infectious disease occurs, but also the Covid-19 pandemic has most likely contributed to a lot of articles being published on the subject.

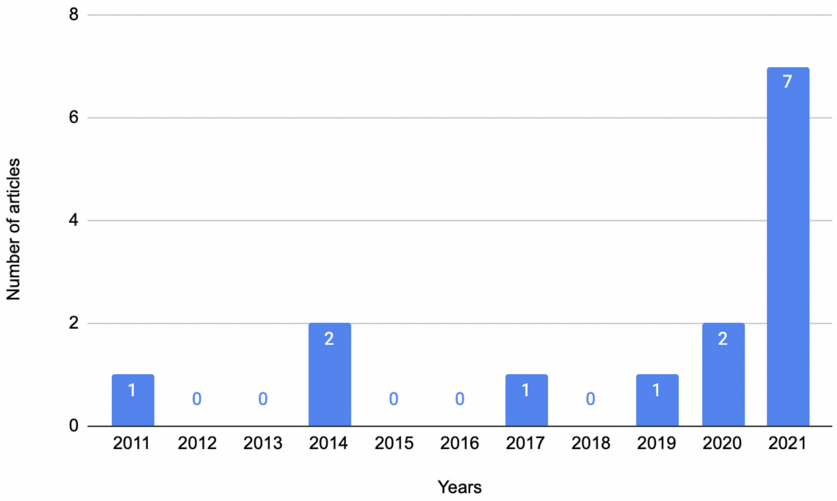


Figure 3: A bar chart representing the distribution of included articles published over time.

Another pattern found among the articles included for the characterization were that Severe Acute Respiratory Syndrome Corona Virus 2 (SARS-CoV-2) and *Mycobacterium tuberculosis* were by far the most common pathogens investigated among the results, see Figure 4. The many articles about SARS-CoV-2 is most likely a direct cause of the Covid-19 pandemic. The many *M. tuberculosis* could be due to that this disease is often carefully monitored when detected. This often provides a lot of epidemiological data useful in these kind of studies, making it available for research. Interestingly, no articles on *E. coli* were found among the included articles.

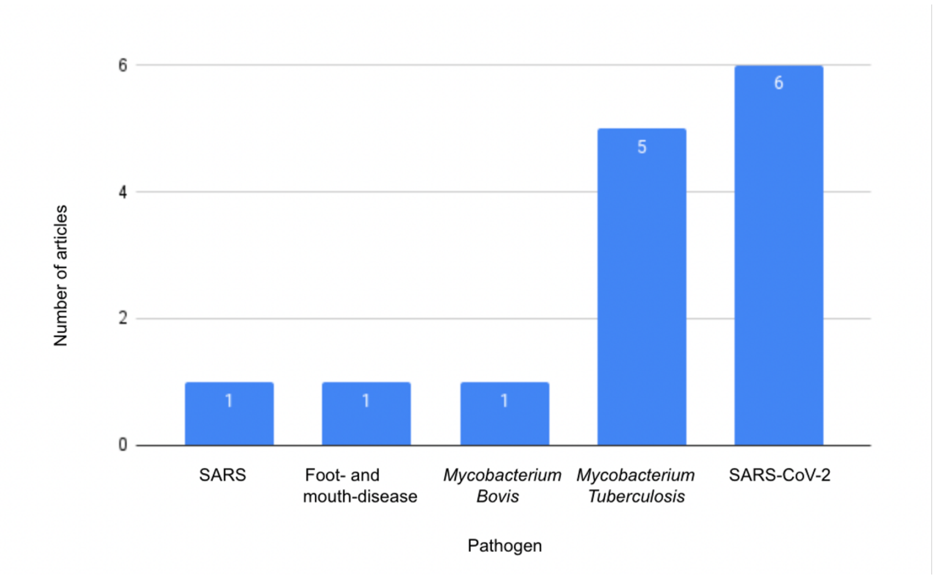


Figure 4: A bar chart representing the number of articles investigating each pathogen.

4.1.2 The Data and Software

All articles included in the characterization used WGS data for their investigations. In addition to the molecular data all articles included also used some kind of epidemiological data in combination with the genetic data. Types of epidemiological data that could be used were temporal and spacial data about the outbreak, as well as information about contacts to create contact networks.

Most articles (11/14) constructed a phylogenetic tree using some software. A list of all articles with their output and what software they used can be found in Appendix D, Table 4. The software listed were not all used for the construction of a phylogenetic tree or the corresponding output, but were also used for alignments, to study geographical data, to visualize and annotate trees etc. The other outputs generated from the resulting articles were transmission trees and networks of transmission events.

Some of the studies (#2 for example) used code and packages written in house, but most of the studies used a variety of different widely available software in order to analyze their data and construct some kind of network/tree. Most of the workflows appeared to be rather specific and tailored for the data available as well as the settings of the article. Two of the studies (#6 and #7) each presented a newly developed tool, both implemented as a package in R. The presented tools were outbreaker (Jombart *et al.*, 2014) and phybreak (Klinkenberg *et al.*, 2017), both widely applicable for a variety of different infectious pathogen outbreaks. Among the software in Table 4 an additional tool, also being widely applicable and a package implemented in R like outbreaker and phybreak was found; TransPhylo (Didelot *et al.*, 2021) found in article #12.

There were mainly two different kinds of studies identified. The first kind was studies that investigated an outbreak of some infectious disease and were interested in the patterns of transmission and if super-spreaders could be identified. The second kind of study were about developing/testing a method (that possibly could be used for studies like in the first category) and therefore were more interested in how the method performed. This was then applied to simulated and/or data from existing studies to validate the efficiency of the methods. The studies belonging to the first category were; #1, #2, #4, #5, #9, #10, #11, #12, #13, and #14 and the studies belonging to the second category were; #3, #6, #7 and #8, also see Table 5 in Appendix D.

4.1.3 The Super-spreaders and Super-shedders

All but one study (#8) were able to identify super-spreaders, see Table 5. The identification could be either that they identified one or several super-spreaders in the investigated outbreak, or that they found that their method of interest were able to identify super-spreaders.

In all but one articles investigating an outbreak (#2), the effect super-spreaders had on the outbreak were mentioned. For example in study #1 super-spreading, in combination with some socioeconomic factors, could explain the high prevalence of disease in the investigated population. Another study, #5, identified that the super-spreaders had an effect on the spread on the investigated hospital. They found that 80% of the cases of transmission were caused by 21% of the individuals, the 20/80 rule. In study #10, they suggest that super-spreaders can have an effect on what lineages of the pathogen (SARS-CoV-2) that becomes most successful in the spread of the disease.

Two of the studies (#1 and #5) addressed the relationship between being a super-spreader and shedding/having a high bacillary load. In study #1 they found that individuals who were considered to be super-spreaders also had a significant bacillary load of *M. tuberculosis*. In study #5 they found indications that increased shedding of the pathogen, SARS-CoV-2, would also increase the likelihood of that individual becoming a super-spreader.

4.1.4 The Summary of Phase One

In this literature study, patterns indicating that articles answering the stated research question is being published more frequently in the last couple of years were found.

A pattern over what pathogens that were more frequently investigated (SARS-CoV-2 and *M. tuberculosis*) were also found. Another pattern found was that the most common output of the articles were phylogenetic trees, but that transmission trees also occurred. These were produced using a variety of different methods and software. Most of the articles identified super-spreaders, some articles also looked at the effect of super-spreaders but only two articles looked at the relation between super-spreaders and super-shedders. Since no study investigating super-spreaders among individuals infected with VTEC O157:H7 were found, the more general methods; outbreaker, phybreak, were considered good options to apply to the molecular data for the second phase of this project. In addition to these two methods, another general method, TransPhylo, was identified and could have been a possible option for the second phase.

4.2 Phase 2 - Exploring Super-spreaders of VTEC O157:H7

In the second phase of this master thesis, 32 sequences of VTEC O157:H7 were explored for super-spreaders using different software, identified in the first phase.

4.2.1 outbreaker2

Outbreaker is a package first presented in 2014 and has since then been updated, nowadays called outbreaker2. Therefore, outbreaker2 is the name used for this method throughout the project.

Outbreaker2, as mentioned before, generates transmission trees as its output. In the transmission tree the circles represent the samples, the ID written next to it shows which sample it is and the colors represents different samples. The arrows represent the path of transmission, where a bigger arrow indicates a higher probability of that transmission event compared to a smaller arrow (the support is also illustrated in Figures 12 - 15 in Appendix F). In Appendix E, Table 6, a list of all samples and their ID, what farm they belong to, if the sample was an environment or individual sample, if they were super-shedders as well as the date of sampling can be found. The four trees generated can be seen in Figure 5 - 8.

Figure 5 and 6 shows the transmission trees where the generation time and incubation period were used to create the gamma distribution. Figure 5 shows the transmission tree generated when using the mutation rate 3.95^{-10} mutations per site per day and Figure 6 shows the tree generated when using the mutation rate 6.19^{-10} mutations per site per day. As seen by following the arrows, the trees in Figure 5 and 6 show the exact same events of transmission. This indicates that the two different mutation rates used do not generate any differences in the path of transmission. In Figure 5 and 6 three different samples which infect three secondary cases were identified; F5-env-2, F9-env-2 and F9-2466-2. Two samples taken from the environment and one sample taken from a calf which was not a super-shedder.

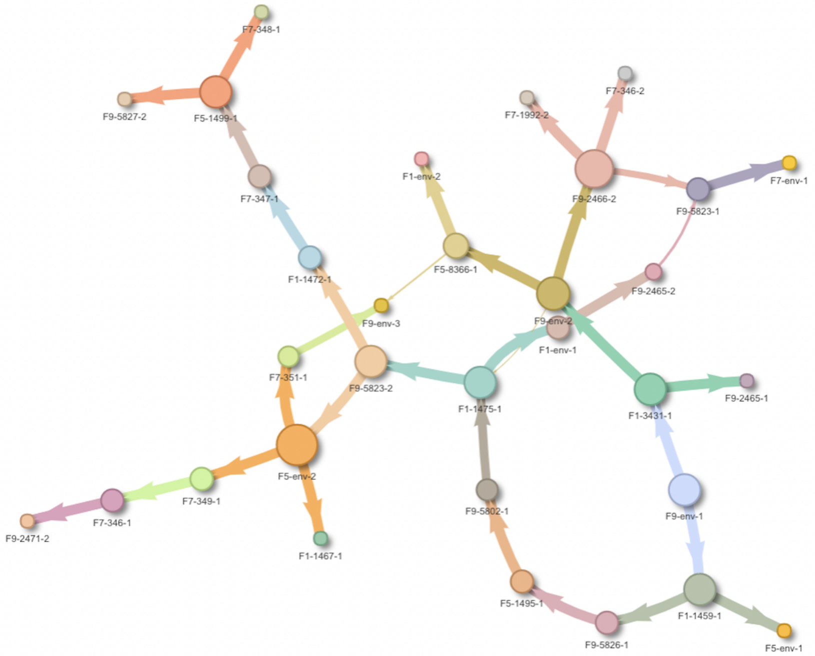


Figure 5: Transmission tree generated from outbreaker2 using the mutation rate 3.95^{-10} mutations per site per day and using the generation time and incubation period to create the gamma distribution.

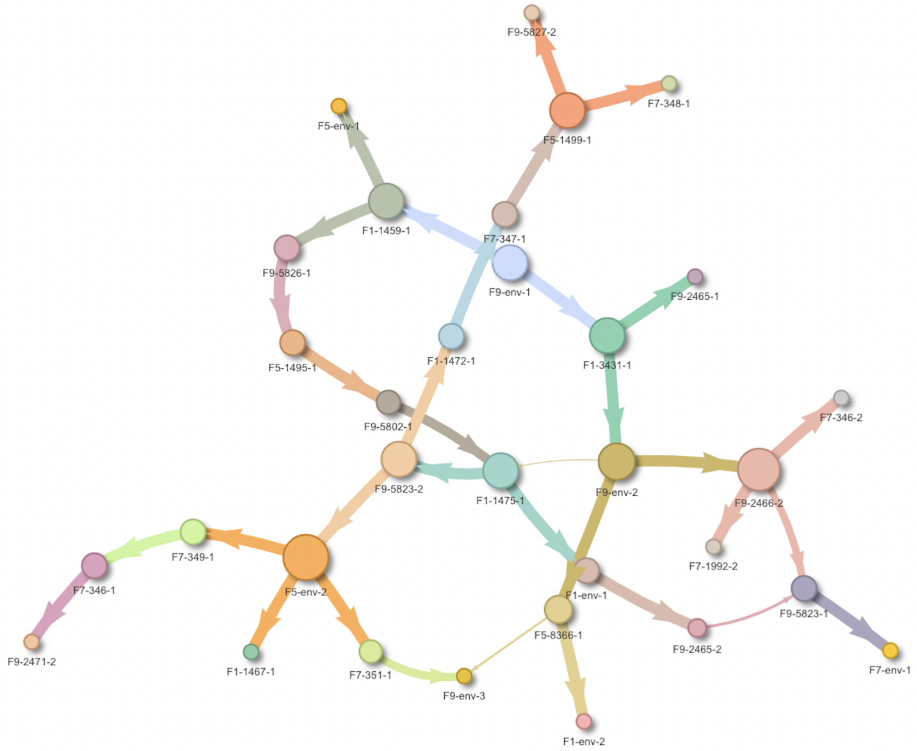


Figure 6: Transmission tree generated from outbreaker2 using the mutation rate 6.19^{-10} mutations per site per day and using the generation time and incubation period to create the gamma distribution.

In Figure 7 and Figure 8 the trees generated using the default parameters for the gamma distribution is shown. Figure 7 shows the transmission tree generated when using the mutation rate 3.95^{-10} mutations per site per day and Figure 8 shows the tree generated when using the mutation rate 6.19^{-10} mutations per site per day. No differences in the path of transmission could be identified between these trees generated using different mutation rates, also an indication that these mutations rates do not generate any differences in the path of transmission. In Figure 7 and 8 three different samples which infect three secondary cases were identified; F9-2466-2, F5-8366-1 and F1-1475-1, where F9-2466-2 and F5-8366-1 were not super-shedders, but F1-1475-1 was a super-shedder.

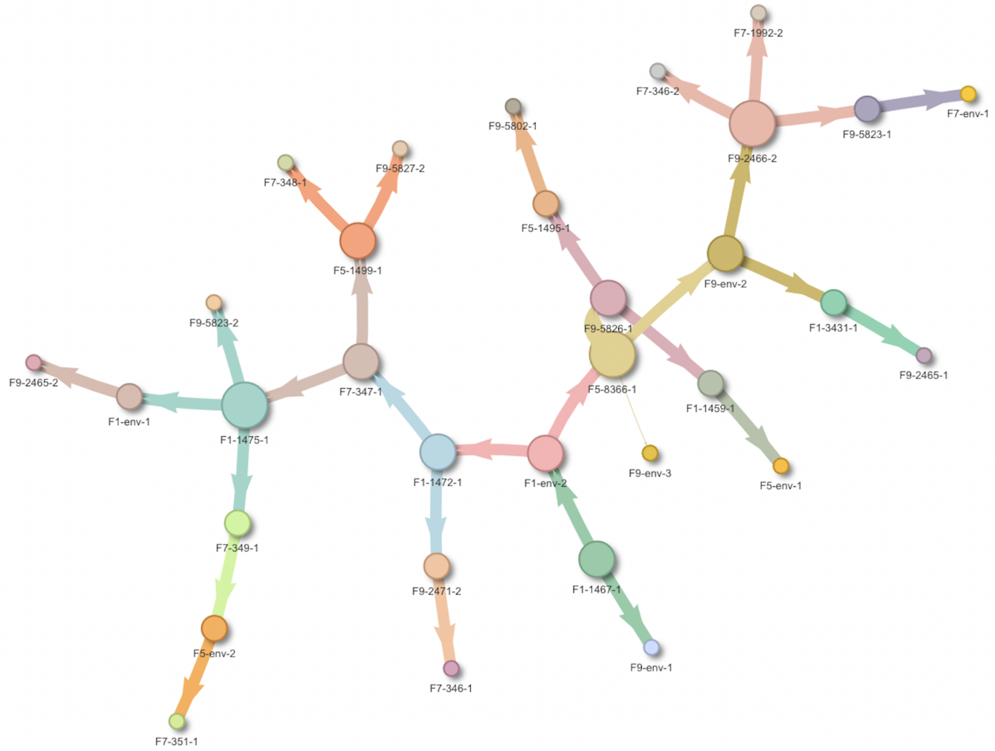


Figure 7: Transmission tree generated from outbreaker2 using the mutation rate 3.95^{-10} mutations per site per day and using the default gamma shape to create the gamma distribution.

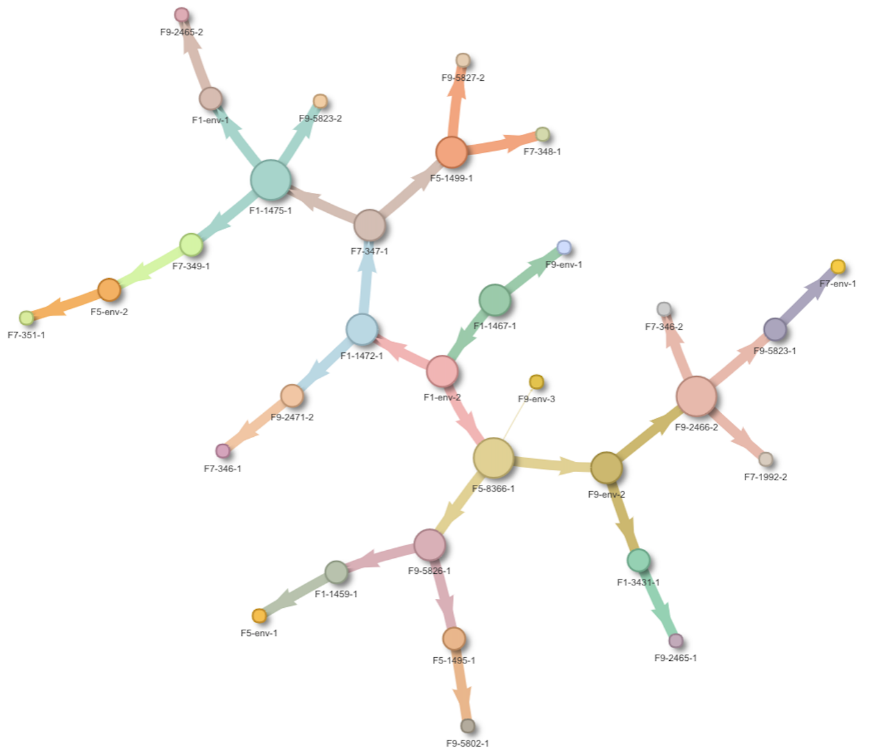


Figure 8: Transmission tree generated from outbreaker2 using the mutation rate 6.19^{-10} mutations per site per dayand using the default gamma shape to create the gamma distribution.

From the transmission trees generated by outbreaker2 it can be identified that the transmission of VTEC O157:H7 follows no specific pattern, but transmission both within farms and between the different farms were observed. In Figure 11, found in Appendix E, the distances between the four different farms are illustrated. These rather small distances between the farms makes it possible to identify all the farms as one epidemiological unit.

As stated, the transmission trees generated from different mutations rates were identical in their path of transmission. When comparing the path of transmission between the trees generated from different gamma distributions, it was found that these trees were not identical.

Some events stayed consistent between the different trees. These events were;
F5-1499-1 transmitting to both F9-5827-2 and F7-348-1,
F9-2466-2 transmitting to F7-1992-2, F7-346-2 and F9-5823-1 and

F9-5823-1 transmitting to F7-env-1.

Individuals infecting more secondary cases compared to others were found in the two different resulting trees, but only one of these, F9-2466-2, were found in both.

In some cases, the direction of the transmission were the opposite between two cases between the tree generated from the two different gamma distributions. Example of events like this can be found in Figure 5 where F1-1459-1 transmit F9-5826-1, but in Figure 7 F9-5826-1 transmit F1-1459-1.

In Appendix F, Figure 16 - 19 plots of the probability of the generation time and incubation period for the different gamma distributions. From Figure 16 and Figure 17 one can see that both the generation time and incubation period have the highest probabilities between 0 and 10 days.
In Figure 18 and Figure 19 the highest probability of the generation time is between 40 and 60 days, whereas for the incubation period the highest probability is seen between day 25 and day 55.

In Appendix F, Figure 12 - 15, plots representing the support for the consensus ancestry for the four transmission trees can be found. The x-axis shows the support and the y-axis shows the number of transmission events. It is the support for each transmission event that is plotted. These plots shows that almost all events has rather high support.

4.2.2 phybreak

The output from phybreak is also transmission trees. The transmission trees generated from the 32 sequences can be found in Figure 9 and 10. In the transmission trees, the IDs to the right corresponds to the sample, the grey blobs shows the median posterior generation interval distribution, the black crosses shows when the samples were taken and the colored arrows indicate a transmission event where the different colors indicate different posterior infector probability; purple arrow <100 %, red arrow <80 %, yellow arrow <60 % and green arrow <20 %. (Klinkenberg *et al.*, 2017)

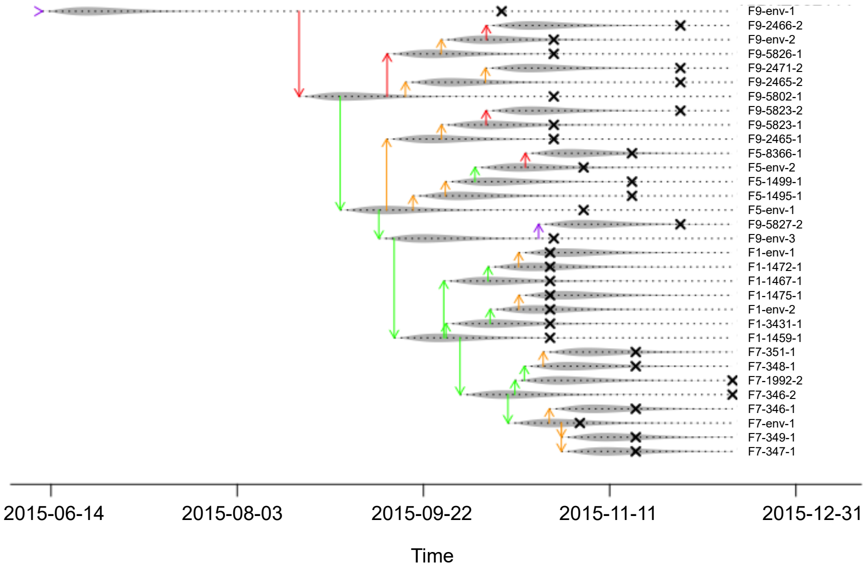


Figure 9: Transmission tree generated from phybreak with mutations rate 3.95^{-10} mutations per site per day.

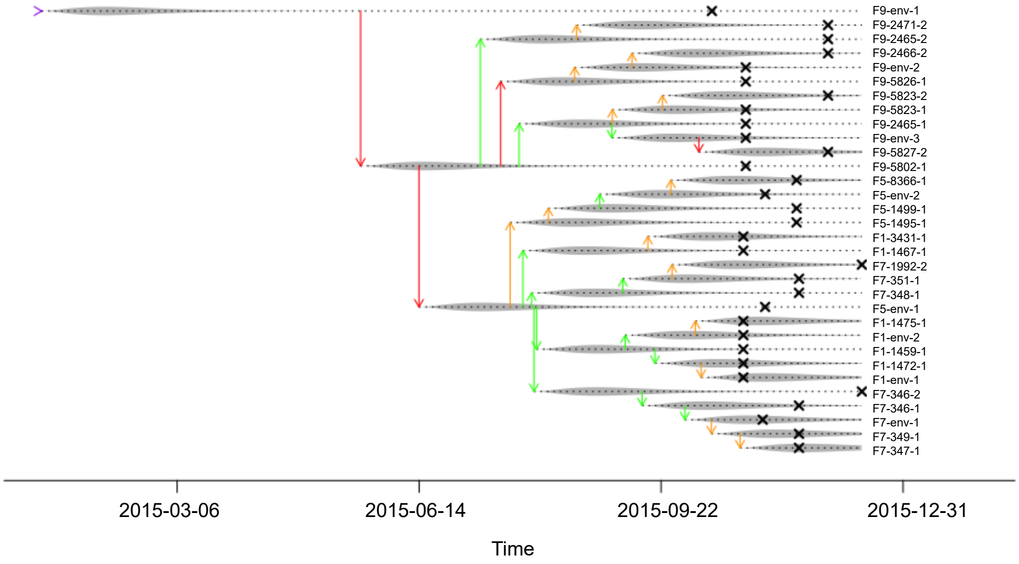


Figure 10: Transmission tree generated from phybreak with mutations rate 6.19^{-10} mutations per site per day.

For phybreaker, trees were only generated with a gamma distribution based on default gamma parameters. Two different mutation rates were used. In Figure 9 the generated transmission tree using the mutation rate 3.95^{-10} mutations per site per day can be found and in Figure 10 the tree generated using the mutation rate 6.19^{-10} mutations per site per day can be found. Unlike in the case using outbreaker2, differences between the trees generated from different mutations rate were identified for phybreak. These differences were mostly a change in the order of the samples in the trees (eg. F9-env-2 being the third case from the top in Figure 9, but the fifth case from the top in Figure 10). Another difference was that some cases transmitted to different individuals in the different trees, the infector and infectee changed. Examples of such differences are;
F1-1467-1 transmitting to F1-1472-1 in Figure 9 but
F1-1467-1 transmitting to F1-3431-1 in Figure 10.
Another such case is;
F1-1459-1 transmitting to F1-3431-1, F1-1467-1 and F7-346-2 in Figure 9 but
F1-1459-1 transmitting to F1-env-2 and F1-1472-1 in Figure 10.

In the phybreak trees, some samples infecting more secondary cases compared to others were found. In Figure 9 such identified cases were;
F9-5802-1, F5-env-1, F1-1459-1, F7-346-2 and F7-env-1.
In Figure 10 such cases were;
F9-2465-1, F9-5802-1, F5-env-1 and F1-1459-1.
Three of these cases were found in both trees; F9-5802-1, F5-env-1 and F1-1459-1. Among these F9-5802-1 shedd bacteria.

In the transmission trees generated from phybreak it can be found that the samples from the different farms cluster with samples from the same farm, observing only a few events of transmission between farms.

From the MCMC iterations, information on the parsimony score is presented, see Figure 20 - 23 in Appendix G. The parsimony score describes the minimum number of mutations that can describe the generated transmission tree (or the tree at the different state of the iteration). The dataset of 32 sequences used here contains 719 SNPs, therefore the most parsimonious tree should have a parsimony score of 719. None of the generated trees reach this score, but reached a parsimony score of 1312 and 1314 respectively. This is an indication that the generated trees do not fully converge.

4.2.3 Comparison between outbreaker2 and phybreak

The biggest differences between the transmission trees generated form the different methods were that phybreak generated trees where the time of sampling were visualized whereas outbreaker2 generated transmission trees in a network form, not visualizing the sampling time of the sequences.

The resulting transmission trees generated from outbreaker2 and phybreak differed. The results from outbreaker2 suggested that the spread of the disease were local and transmission occurred between the farms. The results from phybreak on the other hand suggested that the spread of the disease mostly happened within the farms, with only a few cases of transmission identified between the farms.

Even though the transmission trees generated from the two methods were different, some events still seem to be somewhat linked. Showing that even though the same events of transmission were not identified in the different methods, there were some close connections between the infected individuals from the different methods. For example in outbreaker2, see Figure 7, F5-8366-1 transmit F9-5826-1, F9-env-2 and F9-env-3 while in phybreak, see Figure 9, F9-5826-1 transmitt F9-env-2.

Both outbreaker2 and phybreak could identify samples transmitting more secondary cases compared to the rest of the samples. When investigating these samples between the trees generated from outbreaker2 and phybreak, completely different samples were suggested to spread the disease more than the rest of the samples. A such example is F9-5802-1, which in phybreak causes three secondary cases (see Figure 9) while in outbreaker2 cause no secondary case (see Figure 7).

5 Discussion

5.1 Phase 1 - Literature Study with a Systematic Approach

The aim of the first phase of this project was to provide an overview of previous studies investigating super-spreading using molecular methods. This was accomplished through a literature study with a systematic approach. The articles resulting from the search were few, but many of them were relevant to the stated research question. The reason that relatively few articles were found could either be because that relatively few studies have been published on the subject, or that the search did not manage to capture all of the relevant studies.

The two considered key articles were found in the search, which is a good indication on that relevant studies were captured in the search. However, two relevant articles presenting outbreaker2 and TransPhylo respectively, were not captured by the search. This is an indication that the search was too narrow, not capturing all relevant articles on the subject. Another indication that the search was too narrow was that the yield of articles included for the characterization were higher than recommended, 23 % instead of 10 %.

An additional result from the literature search was that no articles investigating the spread of *E. coli* were found. This indicates that more research within this field is needed in order to gain more knowledge about the transmission of VTEC O157:H7 and to fight the spread of it.

Other limitations with the literature study were that the search was performed in English, only getting articles written in English as the result. Though this could be a problem, other languages would not be of interest in this master thesis, due to language barrier. The search was limited to the databases PubMed and WoS. This contributes to the risk of relevant studies, present only in other databases, not being identified.

Limitations like this would be expected from a literature study with a systematic approach. In order to deal with these limitations a more thorough literature search would have to be performed, a systematic literature review. A complete search like this could however not be performed within the frames of this master thesis. One reason for this was that there was not enough time to perform a more thorough search (e.g. including more databases would generate more hits, would mean more articles to scan through, which there were no time for). Another parameter in a systematic literature review is that several people have to go through the resulting articles, which were not possible in this master thesis.

The two key articles describing outbreaker and phybreak were known before the literature search, and were planned to be used if no more relevant methods were found from the search. Considering that the most promising methods found in the search were outbreaker, phybreak and TransPhylo, it is appropriate that two of these were used in the second phase of the project.

5.2 Phase 2 - Exploring Super-spreaders of VTEC O157:H7

The aim of the second phase of this project was to explore the presence of super-spreaders of VTEC O157:H7 in a dataset. This was done using outbreaker2 and phybreak, also identified from previous studies in the literature review.

In outbreaker2, four different transmission trees were generated. The trees with different mutation rates did not show any differences in events of transmission, but the trees generated with different gamma distributions parameters did show differences in the transmission events. It could possibly be that the difference in mutation rate were to small to create an effect in outbreaker2, while the gamma distribution parameters had a quite large difference, causing an effect on the transmission tree. In a transmission point of view, this could mean that the incubation period and generation time of a pathogen effect the transmission more compared to the mutation rate, at least with the different values used in outbreaker2 in this project.

In phybreak, two transmission trees were created. Here, the only difference between the two generated trees were the mutation rate. Interestingly, in phybreak this change did have an effect on the events of transmission. This could indicate that phybreak is a more sensitive method, at least when it comes to the sensitivity to change in mutation rate.

In outbreaker2, a total of five cases were found to transmit more secondary cases compared to the other samples. In Figure 5 and 6 two environmental samples and one individual sample were identified. That two environmental samples were found to contribute a lot to the transmission could possibly be an indication on the importance of the pathogens presence in the environment. In Figure 7 and 8 three individual samples were identified, one of them being a super-shedder. One of the individual samples were found in both Figure 5 and 6 as well as in Figure 7 and 8. The fact that one out of five potential super-spreaders identified in outbreaker2 were a super-shedder is interesting results. However, it can not be considered an indication that super-spreaders also are super-shedders, but the relation between the two phenomenons can not be dismissed either.

In phybreak a total of six samples infecting more secondary cases compared to the others were identified. Among these, two samples came from the environment and four were samples from individuals. Among the individual samples, only one shedd bacteria. From these results the hypothesis that super-spreaders also are super-shedders can not be established nor dismissed, just like in outbreaker2.

These cases infecting more secondary cases compared to the rest are potential super-spreaders. But this is only an indication and it cannot be determine whether they are actual super-spreaders or not from the results generated in this study.

The environmental samples suggested to play an important role of the transmission could possibly be bacteria that are shedded from super-shedders. If that would be the case, the importance of the environmental samples could possibly indicate an important role of super-shedders as well. In contrast to this, a study by Spencer *et al.* (2015) suggests that shedding individuals have a limited influence on the transmission and that it instead is the environment samples themselves that work as a reservoir for the infection and therefore effect the transmission. A limitation to discuss in this context is that the dataset used in this master thesis consists of rather few samples, taken only at a few different time points. This does not give a full representation of the outbreak. A snapshot of the transmission on a farm could possibly look very different at different time points, especially if shedding is a dynamic phenomenon. Therefore it cannot be stated from these results whether it is the super-shedders or the environmental samples that plays an important role of transmission or not.

The results from outbreaker2 suggest that the transmission of VTEC O157:H7 occurs between farms as well as within farms. This indication of local transmission of the pathogen could be explained by that the four farms investigated are located close to each other, and can be considered an epidemiological unit. A previous study by Lena-Mari Tamminen *et al.* (2019) identified this local transmission of the pathogen when investigating 80 farms on a Swedish island. Tamminen *et al.* (2019) The result from outbreaker2 potentially strengthen the results found in this study.

On the other hand, the results from phybreak suggests that the spread of the disease mostly occurs within the different farms, with only a few cases of between farm transmission identified in the transmission trees. This result suggest the opposite compared to the result generated from outbreaker2. One explanation for this could be that phybreak takes the within-host diversity into consideration. Phybreak could potentially therefore catch smaller differences in the investigated genomes and interpret this in the generated transmission tree. This implementation of within-host diversity could possibly be the reason why the change in mutation rate affect the trees in phybreak.

A possible scenario, explaining the results in both outbreaker2 and phybreak, is that most of the transmission occurs within farms and that only phybreak with its within-host diversity feature can distinguish it. An explanation for this could be that if two cases from the same clone of bacteria start an outbreak on two different farms they will spread and diversify on those two different farms. These bacteria will diversify in different directions, but by chance they could diversify in the same direction. As a result to this, their genome would be very similar and from this they could possibly be identified as event of transmission when in reality they have just evolved in the same direction. Potentially, outbreaker2 is a method that can not distinguish these small differences, but phybreak can due to taking the within-host diversity of a pathogen into consideration. Interestingly, the results from the MCMC iterations in phybreak indicates that some homoplasy is occurring in the trees. Homoplasy is when similar traits of two different species or lineages has evolved independently (Campbell *et al.*, 2015). Exactly what the above scenario describes.

There are several limitations of these methods and results that has to be discussed. Firstly, both methods assumes that all cases of infection are sampled. Since this is not the case, both methods will find infectors and infectees in this dataset that are not true infectors and infectees in real life.

In the MCMC iterations performed in phybreak, the parsimony score did not decrease to the SNP value. This is an indication that the method did not fully converge, another indication of homoplasy (as discussed above). (Klinkenberg *et al.*, 2017)

In phybreak only one gamma distribution were used, the one using default parameters. This was due to that the gamma parameter called shape, based on the generation time and incubation period of VTEC O157:H7, could not be run in phybreak without crashing. An explanation to why this happened could be that the parameter values were to small, since testing bigger values worked but smaller values did not work. Using default values creates pretty wide distributions, therefore being less informative. This will of course affect the resulting trees in phybreak, and as seen in Figure 9 and Figure 10 the majority of arrows showing the transmission events have a rather low posterior infection probability. Therefore the results should be interpreted carefully.

The information used to the generate gamma distributions of the incubation period and generation time were approximated from existing literature. Since calves do not show any symptoms when infected with this pathogen, it is difficult to generate this kind of data. Information about the incubation period were therefore taken from data describing the infection in humans. This is not a representation of the dynamics and time frames in a calf or on a farm, but as stated an approximation. The generation time was also approximated from the literature based on data how many days passed between infections of VTEC O157:H7 in calves on a farm. It is important to take these approximations, as well as the fact that the transmission trees are just results of statistical modelling, in mind when interpreting the results of this project. The results do not tell the truth about the transmission on and between the farms, but use statistics to model likely scenarios.

5.3 Future Work

In future work, it would be interesting to apply the dataset to TransPhylo as well, and compare its result with the results from outbreaker2 and phybreak. Since outbreaker2 and phybreak came up with very different results, it would be interesting to see what results a third method would generate.

It would also be interesting to generate phylogenetic trees from the dataset and compare with the transmission trees. In the results from the literature search a lot of studies generated phylogenetic trees to investigate the spread of some infectious disease. Phylogenetic trees and transmission trees do not say the same thing, but it would still be interesting to investigate if transmission trees are better at revealing super-spreaders and tell us stories about who transmitted whom, or if phylogenetic trees also could be a good approach for this.

In future investigations it would be interesting to perform some more extended error search of the phybreak method, in order to be able to run it with the gamma parameters generated from the incubation period and generation time. The default values used generate a wide distribution and it would be interesting if the distribution became less wide using the distribution parameters from the incubation period and generation time.

In the future, it would also be interesting to compare the findings of super-spreaders with the findings from the study from Lena-Mari Tamminen *et al.* (2020), to see what traits and behaviours calves identified as super-spreaders have. It would also be interesting to further investigate the relation between super-shedders and super-spreaders, in connection with calves behaviour and animal welfare.

6 Conclusions

From the literature study performed in the first phase of this master thesis, relevant articles answering the research question were found. From scanning through and validating these articles it was recognized that outbreaker2, phybreak and TransPhylo were the most suitable methods to apply to the second phase of the project.

In the second phase, from exploring the sequences of VTEC O157:H7 using outbreaker2 and phybreak, interesting findings can be presented. Firstly, it is clear that the different methods provide different transmission scenarios and the differences implicate different transmission dynamics between nearby farms. This is likely a result of the within host variation modelled in phybreak and due to the large impact on the most likely transmission tree this effect should be further explored in future studies. Secondly, there were indications of super-spreading events among the calves. However, these were relatively rare and often associated with environmental samples. Thus, the role of individual super-spreaders and the association between super-shedding requires further investigation.

7 Ethical Approval

The samples used in this master thesis were previously collected by Dr. Lena-Mari Tamminen. The sampling and handling was carried out in accordance with the ethical approval granted by the regional ethical committee (Uppsala Djurförsöksetiska Nämnd, Dnr: C 85/15). All methods were carried out in accordance with relevant guidelines and regulations. (Tamminen *et al.*, 2020)

8 Acknowledgement

This project would not have been possible without my fantastic supervisor, Lena-Mari Tamminen. Thank you for guiding me throughout this project and helping me when I needed it. I would also like to give thanks to my subject reviewer Ulf Emanuelson, my deputy supervisor Robert Söderlund, my examinator Johan Åqvist and my opponent Jenny Eriksson. Also thank you Lena Henriksson, coordinator of this course, for everything you do and have done for us students on X.

I would like to thank my family, for always supporting and believing in me. Additional thanks to my little brother, Carl Wallskog, for helping me with LaTeX.

TWDG, thank you for all late study nights. Thank you for all the unforgettable moments and all memories we have created during my six years at the university. Here's to the future.

Till alla jag har fått priviligeriet att lära känna i Uppsala under mina sex år; stort tack för alla tankar, känslor och minnen!

References

Ameer MA, Wasey A, Salen P. 2021. *Escherichia coli* (*E Coli* O157 H7). In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan–. PMID: 29939622.

Awofisayo-Okuyelu A, Brainard J, Hall I, McCarthy N. 2019. Incubation Period of Shiga Toxin–Producing *Escherichia coli*. Epidemiologic Reviews doi 10.1093/epirev/mxz001.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

Campbell F, Didelot X, Fitzjohn R, Ferguson N, Cori A, Jombart T. 2018. outbreaker2: a modular platform for outbreak reconstruction. BMC Bioinformatics 19: 363.

Campbell NA, Reece JB, Urry LA, Cain ML, Wasserman SA, Minorsky PV, Jackson RB. 2015. Biologi A Global Approach, volume 10th. PEARSON.

CDC. 2019a. Questions and Answers | *E. coli* | CDC. URL: https://www.cdc.gov/ecoli/general/index.html. Retrieved 2022-04-26.

CDC. 2019b. Whole Genome Sequencing (WGS) | PulseNet Methods| PulseNet | CDC. URL: https://www.cdc.gov/pulsenet/pathogens/wgs.html. Retrieved 2022-05-19.

Chase-Topping M, Gally D, Low C, Matthews L, Woolhouse M. 2008. Super-Shedding and the Link Between Human Infection and Livestock Carriage of *Escherichia coli* O157. Nature reviews. Microbiology 6: 904–912.

D Munns K, Selinger LB, Stanford K, Guan L, R Callaway T, A McAllister T. 2015. Perspectives on Super-Shedding of *Escherichia coli* O157:H7 by Cattle. FOODBORNE PATHOGENS AND DISEASE 12: 89–103.

Didelot X, Kendall M, Xu Y, White PJ, McCarthy N. 2021. Genomic Epidemiology Analysis of Infectious Disease Outbreaks Using TransPhylo. Current Protocols 1: e60.

Eriksson E, Aspan A, Gunnarsson A, Vågsholm I. 2005. Prevalence of Verotoxin-Producing *Escherichia coli* (VTEC) O157 in Swedish Dairy Herds. Epidemiology and Infection 133: 349–358. Publisher: Cambridge University Press.

Gibson B, Wilson DJ, Feil E, Eyre-Walker A. 2018. The distribution of bacterial doubling times in the wild. Proceedings of the Royal Society B: Biological Sciences 285: 20180789.

Gunn GJ, McKendrick IJ, Ternent HE, Thomson-Carter F, Foster G, Synge BA. 2007. An investigation of factors associated with the prevalence of verocytotoxin producing *Escherichia coli* O157 shedding in Scottish beef cattle. The Veterinary Journal 174: 554–564.

Holme P, Masuda N. 2015. The Basic Reproduction Number as a Predictor for Epidemic Outbreaks in Temporal Networks. PLoS ONE 10: e0120567.

Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. 2014. Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. PLOS Computational Biology 10: e1003457. Publisher: Public Library of Science.

Kaper JB, O’Brien AD. 2014. Overview and Historical Perspectives. Microbiology Spectrum 2: 2.6.16. Publisher: American Society for Microbiology.

Karolinska Institutet. 2022. Systematisk litteraturöversikt som examensarbete | Karolinska Institutet Universitetsbiblioteket. https://kib.ki.se/soka-vardera/systematiska-oversikter/systematisk-litteraturoversikt-som-examensarbete. Retrieved 2022-02-01.

Klinkenberg D, Backer JA, Didelot X, Colijn C, Wallinga J. 2017. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. PLOS Computational Biology 13: e1005495. Publisher: Public Library of Science.

Konowalchuk J, Speirs JI, Stavric S. 1977. Vero response to a cytotoxin of *Escherichia coli*. Infection and Immunity 18: 775–779. Publisher: American Society for Microbiology.

L Gally D, P Stevens M. 2016. Microbe Profile: *Escherichia coli* O157 : H7 – notorious relative of the microbiologist's workhorse. Microbiology 163: 1–3.

Manning SD, Motiwala AS, Springman AC, Qi W, Lacher DW, Ouellette LM, Mladonicky JM, Somsel P, Rudrik JT, Dietrich SE, Zhang W, Swaminathan B, Alland D, Whittam TS. 2008. Variation in virulence among clades of *Escherichia coli* O157:H7 associated with disease outbreaks. Proceedings of the National Academy of Sciences 105: 4868–4873. Publisher: Proceedings of the National Academy of Sciences.

Nataro JP, Kaper JB. 1998. Diarrheagenic *Escherichia coli*. Clinical Microbiology Reviews 11: 142–201. Publisher: American Society for Microbiology.

NHGRI. 2021. DNA Sequencing Costs: Data. URL: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data. Retrieved 2022-04-18.

O'Brien AD, LaVeck GD. 1983. Purification and characterization of a *Shigella dysenteriae* 1-like toxin produced by *Escherichia coli*. Infection and Immunity 40: 675–683.

O'Brien A, Lively T, Chen M, Rothman S, Formal S. 1983. ESCHERICHIA COLI 0157:H7 STRAINS ASSOCIATED WITH HAEMORRHAGIC COLITIS IN THE UNITED STATES PRODUCE A SHIGELLA DYSENTERIAE 1 (SHIGA) LIKE CYTOTOXIN. The Lancet 321: 702.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Redaktionen UUB. 2022. Uppsala universitetsbibliotek - Uppsala universitet. URL: https://libguides.ub.uu.se/c.php?g=693032p=4969441 Retrieved 2022-05-07.

Reeves PR, Liu B, Zhou Z, Li D, Guo D, Ren Y, Clabots C, Lan R, Johnson JR, Wang L. 2011. Rates of Mutation and Host Transmission for an *Escherichia coli* Clone over 3 Years. PLoS ONE 6: e26907.

Spencer SEF, Besser TE, Cobbold RN, French NP. 2015. 'Super' or just 'above average'? Supershedders and the transmission of *Escherichia coli* O157:H7 among feedlot cattle. Journal of the Royal Society Interface 12: 20150446.

Söderlund R, Jernberg C, Ivarsson S, Hedenström I, Eriksson E, Bongcam-Rudloff E, Aspán A. 2014. Molecular Typing of *Escherichia coli* O157:H7 Isolates from Swedish Cattle and Human Cases: Population Dynamics and Virulence. Journal of Clinical Microbiology 52: 3906–3912.

Tamminen LM, Hranac CR, Dicksved J, Eriksson E, Emanuelson U, Keeling LJ. 2020. Socially engaged calves are more likely to be colonised by VTEC O157:H7 than individuals showing signs of poor welfare. Scientific Reports 10: 6320.

Tamminen LM, Söderlund R, Wilkinson D, Torsein M, Eriksson E, Churakov M, Dicksved J, Keeling L, Emanuelson U. 2019. Risk factors and dynamics of verotoxigenic *Escherichia coli* O157:H7 on cattle farms: An observational study combining information from questionnaires, spatial data and molecular analyses. Preventive Veterinary Medicine 170: 104726.

Thibaut J. 2018. Ebola simulation part 2: outbreak reconstruction · RECON learn. URL: https://reconlearn.org/post/practical-ebola-reconstruction.html Retrieved 2022-05-02.

Thibaut J, Finlay C, Rich F. 2021. Set and check parameter settings of outbreaker — create_config. URL: https://www.repidemicsconsortium.org/outbreaker2/reference/create_config.html. Retrieved 2022-05-02.

Vidovic S, Korber DR. 2006. Prevalence of *Escherichia coli* O157 in Saskatchewan Cattle: Characterization of Isolates by Using Random Amplified Polymorphic DNA PCR, Antibiotic Resistance Profiles, and Pathogenicity Determinants. Applied and Environmental Microbiology 72: 4347–4355.

Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLOS Computational Biology 13: e1005595. Publisher: Public Library of Science.

Wilson JB, Renwick SA, Clarke RC, Rahn K, Alves D, Johnson RP, Ellis AG, McEwen SA, Karmali MA, Lior H, Spika J. 1998. Risk factors for infection with verocytotoxigenic *Escherichia coli* in cattle on Ontario dairy farms. Preventive Veterinary Medicine 34: 227–236.

Woolhouse M, Dye C, Etard JF, Smith T, Charlwood J, Garnett G, Hagan P, Hii J, Ndhlovu P, Quinnell R, Watts C, Chandiwana S, Anderson R. 1997. Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. Proceedings of the National Academy of Sciences of the United States of America 94: 338–342.

Ypma RJF, van Ballegooijen WM, Wallinga J. 2013. Relating Phylogenetic Trees to Transmission Trees of Infectious Disease Outbreaks. Genetics 195: 1055–1062.

A Literature Study - The Search

Table 1: A table showing the search block applied in PubMed. The column ”Number” describes the order that the blocks were done in, block number 11 being the final search. The column ”Query” displays the words and boolean operators applied, the column ”Filters” shows if any filters were applied, and lastly the column ”Results” shows how many hits the query yielded.

Number	Query	Filters	Results
11	(((((”disease transmission, infectious”[MeSH Terms]) OR (”carrier state”[MeSH Terms])) OR (”super-spread*”[Title/Abstract] OR superspread*[Title/Abstract] OR ”super spread*”[Title/Abstract] OR ”high risk contact*”[Title/Abstract])) AND (”transmission dynamic*”[Title/Abstract] OR ”transmission pattern*”[Title/Abstract] OR ”transmission cluster*”[Title/Abstract] OR ”transmission event*”[Title/Abstract] OR ”transmission driver*”[Title/Abstract] OR ”spreader event*”[Title/Abstract])) AND ((Whole genome sequencing[MeSH Terms]) OR (”genome sequencing whole”[Title/Abstract] OR ”sequencing whole genome”[Title/Abstract] OR ”Complete Genome Sequencing”[Title/Abstract] OR ”genome sequencing complete”[Title/Abstract] OR ”sequencing complete genome”[Title/Abstract] OR ”molecular data”[Title/Abstract] OR ”WGS”[Title/Abstract] OR ”whole genome sequenc*”[Title/Abstract] OR ”genetic data”[Title/Abstract] OR ”genome data”[Title/Abstract] OR ”dna sequenc*”[Title/Abstract] OR ”genomic data”[Title/Abstract]))) AND (”identif*”[Title/Abstract] OR ”understand*”[Title/Abstract] OR ”pinpoint*”[Title/Abstract] OR ”recogni*”[Title/Abstract] OR ”explor*”[Title/Abstract])	None	58
10	”identif*”[Title/Abstract] OR ”understand*”[Title/Abstract] OR ”pinpoint*”[Title/Abstract] OR ”recogni*”[Title/Abstract] OR ”explor*”[Title/Abstract]	None	6,033,313

Continued on next page

Continued from previous page

Number	Query	Filters	Results
9	((((”disease transmission, infectious”[MeSH Terms]) OR (”carrier state”[MeSH Terms])) OR (”super-spread*”[Title/Abstract] OR superspread*[Title/Abstract] OR ”super spread*”[Title/Abstract] OR ”high risk contact*”[Title/Abstract])) AND (”transmission dynamic*”[Title/Abstract] OR ”transmission pattern*”[Title/Abstract] OR ”transmission cluster*”[Title/Abstract] OR ”transmission event*”[Title/Abstract] OR ”transmission driver*”[Title/Abstract] OR ”spreader event*”[Title/Abstract])) AND ((Whole genome sequencing[MeSH Terms]) OR (”genome sequencing whole”[Title/Abstract] OR ”sequencing whole genome”[Title/Abstract] OR ”Complete Genome Sequencing”[Title/Abstract] OR ”genome sequencing complete”[Title/Abstract] OR ”sequencing complete genome”[Title/Abstract] OR ”molecular data”[Title/Abstract] OR ”WGS”[Title/Abstract] OR ”whole genome sequenc*”[Title/Abstract] OR ”genetic data”[Title/Abstract] OR ”genome data”[Title/Abstract] OR ”dna sequenc*”[Title/Abstract] OR ”genomic data”[Title/Abstract]))	None	78
8	(Whole genome sequencing[MeSH Terms]) OR (”genome sequencing whole”[Title/Abstract] OR ”sequencing whole genome”[Title/Abstract] OR ”Complete Genome Sequencing”[Title/Abstract] OR ”genome sequencing complete”[Title/Abstract] OR ”sequencing complete genome”[Title/Abstract] OR ”molecular data”[Title/Abstract] OR ”WGS”[Title/Abstract] OR ”whole genome sequenc*”[Title/Abstract] OR ”genetic data”[Title/Abstract] OR ”genome data”[Title/Abstract] OR ”dna sequenc*”[Title/Abstract] OR ”genomic data”[Title/Abstract])	None	164,404

Continued on next page

Continued from previous page

Number	Query	Filters	Results
7	”genome sequencing whole”[Title/Abstract] OR ”sequencing whole genome”[Title/Abstract] OR ”Complete Genome Sequencing”[Title/Abstract] OR ”genome sequencing complete”[Title/Abstract] OR ”sequencing complete genome”[Title/Abstract] OR ”molecular data”[Title/Abstract] OR ”WGS”[Title/Abstract] OR ”whole genome sequenc*”[Title/Abstract] OR ”genetic data”[Title/Abstract] OR ”genome data”[Title/Abstract] OR ”dna sequenc*”[Title/Abstract] OR ”genomic data”[Title/Abstract]	None	155,094
6	Whole genome sequencing[MeSH Terms]	None	14,696
5	(((”disease transmission, infectious”[MeSH Terms]) OR (”carrier state”[MeSH Terms])) OR (”super-spread*”[Title/Abstract] OR superspread*[Title/Abstract] OR ”super spread*”[Title/Abstract] OR ”high risk contact*”[Title/Abstract])) AND (”transmission dynamic*”[Title/Abstract] OR ”transmission pattern*”[Title/Abstract] OR ”transmission cluster*”[Title/Abstract] OR ”transmission event*”[Title/Abstract] OR ”transmission driver*”[Title/Abstract] OR ”spreader event*”[Title/Abstract])	None	1,366
4	”transmission dynamic*”[Title/Abstract] OR ”transmission pattern*”[Title/Abstract] OR ”transmission cluster*”[Title/Abstract] OR ”transmission event*”[Title/Abstract] OR ”transmission driver*”[Title/Abstract] OR ”spreader event*”[Title/Abstract]	None	8,042
3	((”disease transmission, infectious”[MeSH Terms]) OR (”carrier state”[MeSH Terms])) OR (”super-spread*”[Title/Abstract] OR superspread*[Title/Abstract] OR ”super spread*”[Title/Abstract] OR ”high risk contact*”[Title/Abstract])	None	99,834

Continued on next page

Continued from previous page

Number	Query	Filters	Results
2	”super-spread*”[Title/Abstract] OR superspread*[Title/Abstract] OR ”super spread*”[Title/Abstract] OR ”high risk contact*”[Title/Abstract]	None	767
1	(”disease transmission, infectious”[MeSH Terms]) OR (”carrier state”[MeSH Terms])	None	99,198

Table 2: A table showing the search block applied in Web of Science. The column ”Number” describes the order that the blocks were done in, block number 7 being the final search. The column ”Query” displays the words and boolean operators applied, the column ”Filters” shows if any filters were applied, and lastly the column ”Results” shows how many hits the query yielded.

Number	**Query**	**Filters**	**Results**
7	#5 AND #6	None	15
6	TS=(Identif* OR understand* OR pinpoint* OR recogni* OR explor*)	None	12,951,756
5	#3 AND #4	None	17
4	TS=(”genome sequencing whole” OR ”sequencing whole genome” OR ”Complete Genome Sequencing” OR ”genome sequencing complete” OR ”sequencing complete genome” OR ”molecular data” OR ”WGS” OR ”whole genome sequenc*” OR ”genetic data” OR ”genome data” OR ”dna sequenc*” OR ”genomic data”)	None	414,652
3	#1 AND #2	None	147
2	TS=(”transmission dynamic*” OR ”transmission pattern*” OR ”transmission cluster*” OR ”transmission event*” OR ”transmission driver*” OR ”spreader event*”)	None	12,716
1	TS=(”super-spread*” OR ”superspread*” OR ”super spread*” OR ”high risk contact*”)	None	1,102

B Literature Study - The Results

List of references, both included and excluded articles.

1. Klinkenberg Don, Backer Jantien A, Didelot Xavier, Colijn Caroline, Wallinga Jacco. 2017. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. PLOS Computational Biology. 13: 1-32.

2. Böhmer Merle M, Buchholz Udo, Corman Victor M, Hoch Martin, Katz Katharina, Marosevic Durdica V, Böhm Stefanie, Woudenberg Tom, Ackermann Nikolaus, Konrad Regina, Eberle Ute, Treis Bianca, Dangel Alexandra, Bengs Katja, Fingerle Volker, Berger Anja, Hörmansdorfer Stefan, Ippisch Siegfried, Wicklein Bernd, Grahl Andreas, Pörtner Kirsten, Muller Nadine, Zeitlmann Nadine, Boender T. Sonia, Cai Wei, Reich Andreas, An der Heiden Maria, Rexroth Ute, Hamouda Osamah, Schneider Julia, Veith Talitha, Mühlemann Barbara, Wölfel Roman, Antwerpen Markus, Walter Mathias, Protzer Ulrike, Liebl Bernhard, Haas Walter, Sing Andreas, Drosten Christian, Zapf Andreas. 2020. Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: a case series. The Lancet, Infectious Diseases. 20: 920-928.

3. Humphreys H, Coleman D. C. 2019. Contribution of whole-genome sequencing to understanding of the epidemiology and control of meticillin-resistant *Staphylococcus aureus*. The Journal of Hospital Infection 102: 189-199

4. Santibanez S, Hübschen J. M, Ben Mamou M. C, Muscat M, Brown K. E, Myers R, Donoso Mantke O, Zeichhardt H, Brockmann D, Shulga S. V, Muller C. P, O'Connor P. M, Mulders M. N, Mankertz A. 2017. Molecular surveillance of measles and rubella in the WHO European Region: new challenges in the elimination phase. Clinical Microbiology. 23: 516-523

5. Kong Ling Yuan, Eyre David W, Corbeil Jacques, Raymond Frederic, Walker A. Sarah, Wilcox, Mark H, Crook Derrick W, Michaud Sophie, Toye Baldwin, Frost Eric, Dendukuri Nandini, Schiller Ian, Bourgault Anne-Marie, Dascal Andrew, Oughton Matthew, Longtin Yves, Poirier Louise, Brassard Paul, Turgeon Nathalie, Gilca Rodica, Loo Vivian G. 2019. Clostridium difficile: Investigating Transmission Patterns Between Infected and Colonized Patients Using Whole Genome Sequencing. Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America. 68: 204-209

6. Balaji Aakash, Ozer Egon A, Kociolek Larry K. 2019. *Clostridioides difficile* Whole-Genome Sequencing Reveals Limited Within-Host Genetic Diversity in a Pediatric Co-

hort. Journal of Clinical Microbiology. 57: e00559-19

7. Komissarov Andrey B, Safina Ksenia R, Garushyants Sofya K, Fadeev Artem V, Sergeeva Mariia V, Ivanova Anna A, Danilenko Daria M, Lioznov Dmitry, Shneider Olga V, Shvyrev Nikita, Spirin Vadim, Glyzin Dmitry, Shchur Vladimir, Bazykin Georgii A. 2021. Genomic epidemiology of the early stages of the SARS-CoV-2 outbreak in Russia. Nature Communications. 10.1038/s41467-020-20880-z

8. Du Jiteng, Xia Jing, Li Shuyun, Shen Yuxi, Chen Wen, Luo Yuwen, Zhao Qin, Wen Yiping, Wu Rui, Yan Qigui, Huang Xiaobo, Cao Sanjie, Han Xinfeng, Cui Min, Huang Yong. 2020. Evolutionary dynamics and transmission patterns of Newcastle disease virus in China through Bayesian phylogeographical analysis. PloS One. 10.1371/journal.pone.0239809

9. Letizia, Andrew G, Ramos Irene, Obla Ajay, Goforth Carl, Weir Dawn L, Ge Yongchao, Bamman Marcas M, Dutta Jayeeta, Ellis Ethan, Estrella Luis, George Mary-Catherine, Gonzalez-Reiche Ana S, Graham William D, van de Guchte Adriana, Gutierrez Ramiro, Jones Franca, Kalomoiri Aspasia, Lizewski Rhonda, Lizewski Stephen, Marayag Jan, Marjanovic Nada, Millar Eugene V, Nair Venugopalan D, Nudelman German, Nunez Edgar, Pike Brian L, Porter Chad, Regeimbal James, Rirak Stas, Santa Ana Ernesto, Sealfon Rachel S. G, Sebra Robert, Simons Mark P, Soares-Schanoski Alessandra, Sugiharto Victor, Termini Michael, Vangeti Sindhu, Williams Carlos, Troyanskaya Olga G, van Bakel Harm, Sealfon Stuart C. SARS-CoV-2 Transmission among Marine Recruits during Quarantine. 2020. The New England Journal of Medicine. 383: 2407-2416

10. Holt Deborah C, Harris Tegan M, Hughes Jaquelyne T, Lilliebridge Rachael, Croker David, Graham Sian, Hall Heather, Wilson Judith, Tong Steven Y. C, Giffard Phillip M. 2021. Longitudinal whole-genome based comparison of carriage and infection associated *Staphylococcus aureus* in northern Australian dialysis clinics. PloS One. 10.1371/journal.pone.0245790

11. Henderson Alasdair D, Kama Mik, Aubry Maite, Hue Stephane, Teissier Anita, Naivalu Taina, Bechu Vinaisi D, Kailawadoko Jimaima, Rabukawaqa Isireli, Sahukhan Aalisha, Hibberd Martin L, Nilles Eric J, Funk Sebastian, Whitworth Jimmy, Watson Conall H, Lau Colleen L, Edmunds W. John, Cao-Lormeau Van-Mai, Kucharski Adam J. 2021. Interactions between timing and transmissibility explain diverse flavivirus dynamics in Fiji. Nature Communications. 10.1038/s41467-021-21788-y

12. Leavitt Sarah V, Lee Robyn S, Sebastiani Paola, Horsburgh C. Robert, Jenkins Helen E, White Laura F. 2020. Estimating the relative probability of direct transmission between infectious disease patients. International Journal of Epidemiology. 49: 764-775

13. Yang Chongguang, Lu Liping, Warren Joshua L, Wu Jie, Jiang Qi, Zuo Tianyu, Gan Mingyu, Liu Mei, Liu Qingyun, DeRiemer Kathryn, Hong Jianjun, Shen Xin, Colijn Caroline, Guo Xiaoqin, Gao Qian, Cohen Ted. 2018. Internal migration and transmission dynamics of tuberculosis in Shanghai, China: an epidemiological, spatial, genomic analysis. The Lancet, Infectious Diseases. 18: 788-795

14. Donskey Curtis J, Sunkesula Venkata C. K, Stone Nimalie D, Gould Carolyn V, McDonald L. Clifford, Samore Matthew, Mayer JeanMarie, Pacheco Susan M, Jencson Annette L, Sambol Susan P, Petrella Laurica A, Gulvik Christopher A, Gerding Dale N. 2018. Transmission of *Clostridium difficile* from asymptomatically colonized or infected long-term care facility residents. Infection Control and Hospital Epidemiology. 39: 909-916

15. Jajou Rana, Kohl Thomas A, Walker Timothy, Norman Anders, Cirillo Daniela Maria, Tagliani Elisa, Niemann Stefan, de Neeling Albert, Lillebaek Troels, Anthony Richard M, van Soolingen Dick. 2019. Towards standardisation: comparison of five whole genome sequencing (WGS) analysis pipelines for detection of epidemiologically linked tuberculosis cases. Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin.
10.2807/1560-7917.ES.2019.24.50.1900130

16. Mekonnen Daniel, Derbie Awoke, Chanie Asmamaw, Shumet Abebe, Biadglegne Fantahun, Kassahun Yonas, Bobosha Kidist, Mihret Adane, Wassie Liya, Munshea Abaineh, Nibret Endalkachew, Yimer Solomon Abebe, Tønjum Tone, Aseffa Abraham. 2019. Molecular epidemiology of *M. tuberculosis* in Ethiopia: A systematic review and meta-analysis. Tuberculosis (Edinburgh, Scotland). 10.1016/j.tube.2019.101858

17. Illingworth Christopher Jr, Hamilton William L, Warne Ben, Routledge Matthew, Popay Ashley, Jackson Chris, Fieldman Tom, Meredith Luke W, Houldcroft Charlotte J, Hosmillo Myra, Jahun Aminu S, Caller Laura G, Caddy Sarah L, Yakovleva Anna, Hall Grant, Khokhar Fahad A, Feltwell Theresa, Pinckert Malte L, Georgana Iliana, Chaudhry Yasmin, Curran Martin D, Parmar Surendra, Sparkes Dominic, Rivett Lucy, Jones Nick K, Sridhar Sushmita, Forrest Sally, Dymond Tom, Grainger Kayleigh, Workman Chris, Ferris Mark, Gkrania-Klotsas Effrossyni, Brown Nicholas M, Weekes Michael P, Baker Stephen, Peacock Sharon J, Goodfellow Ian G, Gouliouris Theodore, de Angelis Daniela, Török M. Estée. 2021. Superspreaders drive the largest outbreaks of hospital onset COVID-19 infections. eLife. 10.7554/eLife.67308

18. Bousali Maria, Dimadi Aristea, Kostaki Evangelia-Georgia, Tsiodras Sotirios, Nikolopoulos Georgios K, Sgouras Dionyssios N, Magiorkinis Gkikas, Papatheodoridis George, Pogka Vasiliki, Lourida Giota, Argyraki Aikaterini, Angelakis Emmanouil, Sourvinos George, Beloukas Apostolos, Paraskevis Dimitrios, Karamitros Timokratis.

2021. SARS-CoV-2 Molecular Transmission Clusters and Containment Measures in Ten European Regions during the First Pandemic Wave. Life (Basel, Switzerland). 10.3390/life11030219

19. Nutman A, Marchaim D. 2019. How to: molecular investigation of a hospital outbreak. Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases. 25: 688-695

20. Stimson James, Gardy Jennifer, Mathema Barun, Crudu Valeriu, Cohen Ted, Colijn Caroline. 2019. Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions. Molecular Biology and Evolution. 36: 587-603

21. Hassan Brekhna, Ijaz Muhammad, Khan Asadullah, Sands Kirsty, Serfas Georgios-Ion, Clayfield Liam, El-Bouseary Maisra Mohammed, Lai Giulia, Portal Edward, Khan Afifah, Watkins William J, Parkhill Julian, Walsh Timothy R. 2021. A role for arthropods as vectors of multidrug-resistant Enterobacterales in surgical site infections from South Asia. Nature Microbiology. 6: 1259-1270

22.Chow Nancy A, Gade Lalitha, Tsay Sharon V, Forsberg Kaitlin, Greenko Jane A, Southwick Karen L, Barrett Patricia M, Kerins Janna L, Lockhart Shawn R, Chiller Tom M, Litvintseva Anastasia P, US Candida auris Investigation Team. 2018. Multiple introductions and subsequent transmission of multidrug-resistant *Candida auris* in the USA: a molecular epidemiological survey. The Lancet, Infectious Diseases. 18: 1377-1384

23. Bhowmick Biswajit, Zhao Jianguo, Øines Øivind, Bi Tianlin, Liao Chenghong, Zhang Lei, Han Qian. 2019. Molecular characterization and genetic diversity of *Ornithonyssus sylviarum* in chickens (*Gallus gallus*) from Hainan Island, China. Parasites & Vectors. 10.1186/s13071-019-3809-9

24. Abbasi Ibrahim, Nasereddin Abdelmajeed, Warburg Alon. 2019. Development of a next generation DNA sequencing-based multi detection assay for detecting and identifying Leishmania parasites, blood sources, plant meals and intestinal microbiome in phlebotomine sand flies. Acta Tropica. 10.1016/j.actatropica.2019.105101

25. Alvarez Gonzalo G, Zwerling Alice A, Duncan Carla, Pease Christopher, Van Dyk Deborah, Behr Marcel A, Lee Robyn S, Mulpuru Sunita, Pakhale Smita, Cameron D. William, Aaron Shawn D, Patterson Michael, Allen Jean, Sullivan Kathryn, Jolly Anne, Sharma Meenu K, Jamieson Frances B. 2021. Molecular Epidemiology of *Mycobacterium tuberculosis* To Describe the Transmission Dynamics Among Inuit Residing in Iqaluit Nunavut Using Whole-Genome Sequencing. Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America. 72: 2187-2195

26. Gorrie Claire L, Mirceta Mirjana, Wick Ryan R, Edwards David J, Thomson Nicholas R, Strugnell Richard A, Pratt Nigel F, Garlick Jill S, Watson Kerri M, Pilcher David V, McGloughlin Steve A, Spelman Denis W, Jenney Adam W. J, Holt Kathryn E. 2017. Gastrointestinal Carriage Is a Major Reservoir of *Klebsiella pneumoniae* Infection in Intensive Care Patients. Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America. 65: 208-215

27. Lau 2, Grenfell Bryan T, Worby Colin J, Gibson Gavin J. 2019. Model diagnostics and refinement for phylodynamic models. PLoS computational biology. 10.1371/journal.pcbi.1006955

28. Jombart Thibaut, Cori Anne, Didelot Xavier, Cauchemez Simon, Fraser Christophe, Ferguson Neil. 2014. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS computational biology.
10.1371/journal.pcbi.1003457

29. Buckley Cameron, Forde Brian M, Trembizki Ella, Lahra Monica M, Beatson Scott A, Whiley David M. 2018. Use of whole genome sequencing to investigate an increase in *Neisseria gonorrhoeae* infection among women in urban areas of Australia. Scientific Reports. 10.1038/s41598-018-20015-x

30. Giovanetti Marta, de Mendonça Marcos Cesar Lima, Fonseca Vagner, Mares-Guia Maria Angélica, Fabri Allison, Xavier Joilson, de Jesus Jaqueline Goes, Gräf Tiago, Dos Santos Rodrigues Cintia Damasceno, Dos Santos Carolina Cardoso, Sampaio Simone Alves, Chalhoub Flavia Lowen Levy, de Bruycker Nogueira Fernanda, Theze Julien, Romano Alessandro Pecego Martins, Ramos Daniel Garkauskas, de Abreu Andre Luiz, Oliveira Wanderson Kleber, do Carmo Said Rodrigo Fabiano, de Alburque Carlos F. Campelo, de Oliveira Tulio, Fernandes Carlos Augusto, Aguiar Shirlei Ferreira, Chieppe Alexandre, Sequeira Patrícia Carvalho, Faria Nuno Rodrigues, Cunha Rivaldo Venâncio, Alcantara Luiz Carlos Junior, de Filippis Ana Maria Bispo. 2019. Yellow Fever Virus Reemergence and Spread in Southeast Brazil, 2016-2019. Journal of Virology. 10.1128/JVI.01623-19

31. Giske C. G, Dyrkell F, Arnellos D, Vestberg N, Hermansson Panna S, Fröding I, Ullberg M, Fang H. 2019. Transmission events and antimicrobial susceptibilities of methicillin-resistant *Staphylococcus argenteus* in Stockholm. Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases. 25: 1289.e5-1289.e8

32. Croucher Nicholas J, Didelot Xavier. 2015. The application of genomics to tracing bacterial pathogen transmission. Current Opinion in Microbiology. 23: 62-67

33. Gardy Jennifer L, Johnston James C, Ho Sui Shannan J, Cook Victoria J, Shah Lena,

Brodkin Elizabeth, Rempel Shirley, Moore Richard, Zhao Yongjun, Holt Robert, Varhol Richard, Birol Inanc, Lem Marcus, Sharma Meenu K, Elwood Kevin, Jones Steven J. M, Brinkman Fiona S. L, Brunham Robert C, Tang Patrick. 2011. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. The New England Journal of Medicine. 364: 730-739

34. Nurjadi Dennis, Eichel Vanessa M, Tabatabai Patrik, Klein Sabrina, Last Katharina, Mutters Nico T, Pöschl Johannes, Zanger Philipp, Heeg Klaus, Boutin Sébastien. 2021. Surveillance for Colonization, Transmission, and Infection With Methicillin-Susceptible *Staphylococcus aureus* in a Neonatal Intensive Care Unit. JAMA network open. 10.1001/jamanetworkopen.2021.24938

35. van Tonder Andries J, Thornton Mark J, Conlan Andrew J. K, Jolley Keith A, Goolding Lee, Mitchell Andrew P, Dale James, Palkopoulou Eleftheria, Hogarth Philip J, Hewinson R. Glyn, Wood James L. N, Parkhill. 2021. Inferring *Mycobacterium bovis* transmission between cattle and badgers using isolates from the Randomised Badger Culling Trial. PLoS pathogens. 10.1371/journal.ppat.1010075

36. Wittwer Matthias, Altpeter Ekkehard, Pilo Paola, Gygli Sebastian M, Beuret Christian, Foucault Frederic, Ackermann-Gäumann Rahel, Karrer Urs, Jacob Daniela, Grunow Roland, Schürch Nadia. 2018. Population Genomics of *Francisella tularensis* subsp. *holarctica* and its Implication on the Eco-Epidemiology of Tularemia in Switzerland. Front Cell Infect Microbiol. 10.3389/fcimb.2018.00089

37. Alisjahbana Bachti, Koesoemadinata Raspati Cundarani, Hadisoemarto Panji Fortuna, Lestari Bony Wiem, Hartati Sri, Chaidir Lidya, Huang Chuan-Chin, Murray Megan, Hill Philip Campbell, McAllister Susan Margaret. 2021. Are neighbourhoods of tuberculosis cases a high-risk population for active intervention? A protocol for tuberculosis active case finding. PloS One. 10.1371/journal.pone.0256043

38. Borland Erin M, Ledermann Jeremy P, Powers Ann M. 2016. Culex Tarsalis Mosquitoes as Vectors of Highlands J Virus. Vector Borne and Zoonotic Diseases (Larchmont, N.Y.). 16: 558-565

39. Bataille Arnaud, Fournié Guillaume, Cruz Marilyn, Cedeño Virna, Parker Patricia G, Cunningham Andrew A, Goodman Simon J. 2012. Host selection and parasite infection in *Aedes taeniorhynchus*, endemic disease vector in the Galápagos Islands. Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases. 12: 1831-1841

40. Pinholt Mette, Bayliss Sion C, Gumpert Heidi, Worning Peder, Jensen Veronika V. S, Pedersen Michael, Feil Edward J, Westh Henrik. 2019. WGS of 1058 *Enterococcus faecium* from Copenhagen, Denmark, reveals rapid clonal expansion of vancomycin-

resistant clone ST80 combined with widespread dissemination of a vanA-containing plasmid and acquisition of a heterogeneous accessory genome. The Journal of Antimicrobial Chemotherapy. 74: 1776-1785

41. Lee Robyn S, Proulx Jean-François, McIntosh Fiona, Behr Marcel A, Hanage William P. 2020. Previously undetected super-spreading of *Mycobacterium tuberculosis* revealed by deep sequencing. eLife. 10.7554/eLife.53245

42. Harris Simon R, Cartwright Edward J. P, Török M. Estée, Holden Matthew T. G, Brown Nicholas M, Ogilvy-Stuart Amanda L, Ellington Matthew J, Quail Michael A, Bentley Stephen D, Parkhill Julian, Peacock Sharon J. 2013. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. The Lancet, Infectious Diseases. 13: 130-136

43. Stucki David, Ballif Marie, Egger Matthias, Furrer Hansjakob, Altpeter Ekkehardt, Battegay Manuel, Droz Sara, Bruderer Thomas, Coscolla Mireia, Borrell Sonia, Zürcher Kathrin, Janssens Jean-Paul, Calmy Alexandra, Mazza Stalder Jesica, Jaton Katia, Rieder Hans L, Pfyffer Gaby E, Siegrist Hans H, Hoffmann Matthias, Fehr Jan, Dolina Marisa, Frei Reno, Schrenzel Jacques, Böttger Erik C, Gagneux Sebastien, Fenner Lukas. 2016. Standard Genotyping Overestimates Transmission of *Mycobacterium tuberculosis* among Immigrants in a Low-Incidence Country. Journal of Clinical Microbiology. 54: 1862-1870

44. Vasconcelos Dos Santos Thiago, de Pita-Pereira Daniela, Araújo-Pereira Thais, Britto Constança, Silveira Fernando Tobias, Póvoa Marinete Marins, Rangel Elizabeth Ferreira. 2019. *Leishmania* DNA detection and species characterization within phlebotomines (Diptera: Psychodidae) from a peridomicile-forest gradient in an Amazonian/Guianan bordering area. PloS One. 10.1371/journal.pone.0219626

45. Morelli Marco J, Thébaud Gaël, Chadœuf Joël, King Donald P, Haydon Daniel T, Soubeyrand Samuel. 2012. A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. PLoS computational biology. 10.1371/journal.pcbi.1002768

46. Anderson Laura F, Tamne Surinder, Brown Timothy, Watson John P, Mullarkey Catherine, Zenner Dominik, Abubakar Ibrahim. 2014. Transmission of multidrug-resistant tuberculosis in the UK: a cross-sectional molecular and epidemiological study of clustering and contact tracing. The Lancet, Infectious Diseases. 14: 406-415

47. Crellen Thomas, Turner Paul, Pol Sreymom, Baker Stephen, Nguyen Thi Nguyen To, Stoesser Nicole, Day Nicholas Pj, Turner Claudia, Cooper Ben S. 2019. Transmission dynamics and control of multidrug-resistant *Klebsiella pneumoniae* in neonates in a developing country. eLife. 10.7554/eLife.50468

48. Gehre Florian, Antonio Martin, Faïhun Frank, Odoun Mathieu, Uwizeye Cecile, de Rijk Pim, de Jong Bouke C, Affolabi Dissou. 2013. The first phylogeographic population structure and analysis of transmission dynamics of *M. africanum* West African 1– combining molecular data from Benin, Nigeria and Sierra Leone. PloS One. 10.1371/journal.pone.0077000

49. Thiemann T. C, Brault A. C, Ernest H. B, Reisen W. K. 2012. Development of a high-throughput microsphere-based molecular assay to identify 15 common bloodmeal hosts of *Culex* mosquitoes. Molecular Ecology Resources. 12: 238-146

50. Auty Harriet K, Picozzi Kim, Malele Imna, Torr Steve J, Cleaveland Sarah, Welburn Sue. 2012. Using molecular data for epidemiological inference: assessing the prevalence of *Trypanosoma brucei rhodesiense* in tsetse in Serengeti, Tanzania. PLoS neglected tropical diseases. 10.1371/journal.pntd.0001501

51. Mollentze Nardus, Nel Louis H, Townsend Sunny, le Roux Kevin, Hampson Katie, Haydon Daniel T, Soubeyrand Samuel. 2014. A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. Proceedings Biological Sciences. 10.1098/rspb.2013.3251

52. Colijn Caroline, Gardy Jennifer. 2014. Phylogenetic tree shapes resolve disease transmission patterns. Evolution, Medicine, and Public Health. 2014: 96-108

53. Loftus R. W, Dexter F, Robinson A. D. M, Horswill A. R. 2018. Desiccation tolerance is associated with *Staphylococcus aureus* hypertransmissibility, resistance and infection development in the operating room. The Journal of Hospital Infection. 100: 299-308

54. Roe Chandler C, Horn Kimberly S, Driebe Elizabeth M, Bowers Jolene, Terriquez Joel A, Keim Paul, Engelthaler David M. 2016. Whole genome SNP typing to investigate methicillin-resistant *Staphylococcus aureus* carriage in a health-care provider as the source of multiple surgical site infections. Hereditas. 10.1186/s41065-016-0017-x

55. Rice Benjamin L, Golden Christopher D, Anjaranirina Evelin Jean Gasta, Botelho Carolina Mastella, Volkman Sarah K, Hartl Daniel L. 2016. Genetic evidence that the Makira region in northeastern Madagascar is a hotspot of malaria transmission. Malaria Journal. 10.1186/s12936-016-1644-4

56. Yan Zhongqiang, Zhou Yu, Du Mingmei, Bai Yanling, Liu Bowei, Gong Meiliang, Song Hongbin, Tong Yigang, Liu Yunxi. 2019. Prospective investigation of carbapenem-resistant *Klebsiella pneumonia* transmission among the staff, environment and patients in five major intensive care units, Beijing. The Journal of Hospital Infection. 101: 150-157

57. Delwart 1, Busch M. P, Kalish M. L, Mosley J. W, Mullins J. I. 1995. Rapid molecular epidemiology of human immunodeficiency virus transmission. AIDS research and human retroviruses. 11: 1081-1093

58. Clarke J. R, Anderson T. J. C, Bandi C. 2004. Sexual transmission of a nematode parasite of Wood Mice (Apodemus sylvaticus)?. Parasitology. 128: 561-568

59. Yang Xuemei, Dong Ning, Chan Edward Wai-Chi, Chen Sheng. 2020. Genetic cluster analysis of SARS-CoV-2 and the identification of those responsible for the major outbreaks in various countries. Emerging Microbes & Infections. 9: 1287-1299

60. Tasakis Rafail Nikolaos, Samaras Georgios, Jamison Anna, Lee Michelle, Paulus Alexandra, Whitehouse Gabrielle, Verkoczy Laurent, Papavasiliou F. Nina, Diaz Marilyn. 2021. SARS-CoV-2 variant evolution in the United States: High accumulation of viral mutations over time likely through serial Founder Events and mutational bursts. Plos One. 10.1371/journal.pone.0255169

61. Lopez Mariana G, Chiner-Oms Alvaro, Garcia de Viedma Dario, Ruiz-Rodriguez Paula, Alma Bracho Maria, Cancino-Munoz Irving, D'Auria Giuseppe, de Marco Griselda, Garcia-Gonzalez Neris, Goig Galo Adrian, Gomez-Navarro Inmaculada, Jimenez-Serrano Santiago, Martinez-Priego Llucia, Ruiz-Hueso Paula, Ruiz-Roldan Lidia, Torres-Puente Manuela, Alberola Juan, Albert Eliseo, Aranzamendi Zaldumbide Maitane, Pilar Bea-Escudero Maria, Antonio Boga Jose, Bordoy Antoni E,Canut-Blasco Andres, Carvajal Ana, Cilla Eguiluz Gustavo, Cordon Rodriguez Maria Luz, Costa-Alcalde Jose J, de Toro Maria, de Toro Peinado Inmaculada, Luis del Pozo Jose, Duchene Sebastian, Fernandez-Pinero Jovita, Fuster Escriva Begona, Gimeno Cardona Concepcion, Gonzalez Galan Veronica, Gonzalo Jimenez Nieves, Hernaez Crespo Silvia, Herranz Marta, Antonio Lepe Jose, Lopez-Causape Carla, Luis Lopez-Hontangas Jose, Martin Vicente, Martro Elisa, Milagro Beamonte Ana, Montes Ros Milagrosa, Moreno-Munoz Rosario, Navarro David, Maria Navarro-Mari Jose, Not Anna, Oliver Antonio, Palop-Borras Begona, Grande Monica Parra, Pedrosa-Corral Irene, Perez Gonzalez Maria Carmen, Perez-Lago Laura, Perez-Ruiz Mercedes, Pineiro Vazquez Luis, Rabella Nuria, Rezusta Antonio, Robles Fonseca Lorena,Sanbonmatsu-Gamez Sara, Sicilia Jon, Soriano Alex, Tirado Balaguer Maria Dolores, Torres Ignacio, Tristancho Alexander, Marimon Jose Maria, Coscolla Mireia, Gonzalez-Candelas Fernando, Comas Inaki. 2021. The first wave of the COVID-19 epidemic in Spain was associated with early introductions and fast spread of a dominating genetic variant. Nature Genetics. 10.1038/s41588-021-00936-6

62. Zeller Mark, Gangavarapu Karthik, Anderson Catelyn, Smither Allison R, Vanchiere John A, Rose Rebecca, Snyder Daniel J, Dudas Gytis, Watts Alexander, Matteson Nathaniel L, Robles-Sikisaka Refugio, Marshall Maximilian, Feehan Amy K, Sabino-

Santos Gilberto, Bell-Kareem Antoinette R, Hughes Laura D, Alkuzweny Manar, Snarski Patricia, Garcia-Diaz Julia, Scott Rona S, Melnik Lilia I, Klitting Raphaëlle, McGraw Michelle, Belda-Ferre Pedro, DeHoff Peter, Sathe Shashank, Marotz Clarisse, Grubaugh Nathan, Nolan David J, Drouin Arnaud C, Genemaras Kaylynn J, Chao Karissa, Topol Sarah, Spencer Emily, Nicholson Laura, Aigner Stefan, Yeo Gene W, Farnaes Lauge, Hobbs Charlotte A, Laurent Louise C, Knight Rob, Hodcroft Emma B, Khan Kamran, Fusco Dahlene N, Cooper Vaughn S, Lemey Phillipe, Gardner Lauren, Lamers Susanna L, Kamil Jeremy P, Garry Robert F, Suchard Marc A, Andersen Kristian G. 2021. Emergence of an early SARS-CoV-2 epidemic in the United States. medRxiv. 10.1101/2021.02.05.21251235

C Literature Study - The Screening for Relevance

Table 3: The resulting articles with information on authors, year of publication, database, and if the article was excluded or included to the characterization.

Author	**Year**	**Database**	**Included/Excluded**
Abbasi, Ibrahim *et al.*	2019	PubMed	Excluded
Alisjahbana, Bachti *et al.*	2021	PubMed	Excluded
Alvarez, Gonzalo G. *et al.*	2021	PubMed, WoS	Included
Anderson, Laura F *et al.*	2014	PubMed	Excluded
Auty, Harriet K. *et al.*	2012	PubMed	Excluded
Balaji, Aakash. *et al.*	2019	PubMed	Excluded
Bataille, Arnaud *et al.*	2012	PubMed	Excluded
Bhowmick, Biswajit *et al.*	2019	PubMed	Excluded
Boehmer, Merle M. *et al.*	2020	PubMed, WoS	Excluded
Borland, Erin M. *et al.*	2016	PubMed	Excluded
Bousali, Maria *et al.*	2021	PubMed, WoS	Included
Buckley, Cameron *et al.*	2018	PubMed	Excluded
Chow, Nancy A. *et al.*	2018	PubMed	Excluded
Clarke, J. R. *et al.*	2004	PubMed	Excluded
Colijn, Caroline *et al.*	2014	PubMed, WoS	Included
Crellen, Thomas *et al.*	2019	PubMed	Excluded
Croucher, Nicholas J. and Didelot, Xavier	2015	PubMed	Excluded
Delwart, E. L. *et al.*	1995	PubMed	Excluded
Donskey, Curtis J. *et al.*	2018	PubMed	Excluded
Du, Jiteng *et al.*	2020	PubMed	Excluded
Gardy, Jennifer L. *et al.*	2011	PubMed, WoS	Included
Gehre, Florian *et al.*	2013	PubMed	Excluded
Giovanetti, Marta *et al.*	2019	PubMed	Excluded
Giske, C. G. *et al.*	2019	PubMed	Excluded
Gorrie, *et al.*	2017	PubMed	Excluded
Harris, Simon R. *et al.*	2013	PubMed	Excluded
Hassan, Brekhna *et al.*	2021	PubMed	Excluded
Henderson, Alasdair D. *et al.*	2021	PubMed	Excluded
Holt, Deborah C. *et al.*	2021	PubMed	Excluded
Humphreys, H. and Coleman, D. C.	2019	PubMed	Excluded

Continued on next page

Continued from previous page

Author	**Year**	**Database**	**Included/Excluded**
Illingworth, Christopher J. R. *et al.*	2021	PubMed, WoS	Included
Jajou, Rana *et al.*	2019	PubMed	Excluded
Jombart, Thibaut *et al.*	2014	PubMed, WoS	Included
Klinkenberg, Don *et al.*	2017	PubMed	Included
Komissarov, Andrey B. *et al.*	2021	PubMed	Excluded
Kong, Ling Yuan *et al.*	2019	PubMed, WoS	Excluded
Lau, Max S. Y. *et al.*	2019	PubMed, WoS	Included
Leavitt, Sarah V. *et al.*	2020	PubMed	Excluded
Lee, Robyn S. *et al.*	2020	PubMed, WoS	Included
Letizia, Andrew G. *et al.*	2020	PubMed	Excluded
Loftus, R. W.	2018	PubMed	Excluded
Lopez, Mariana G. *et al.*	2021	WoS	Included
Mekonnen, Daniel *et al.*	2019	PubMed, WoS	Excluded
Mollentze, Nardus *et al.*	2014	PubMed	Excluded
Nurjadi, Dennis *et al.*	2021	PubMed	Excluded
Nutman, A. and Marchaim, D.	2019	PubMed	Excluded
Pinholt, Mette *et al.*	2019	PubMed	Excluded
Rice, Benjamin L. *et al.*	2016	PubMed	Excluded
Roe, Chandler C. *et al.*	2016	PubMed	Excluded
Santibanez, S. *et al.*	2017	PubMed	Excluded
Stimson, James *et al.*	2019	PubMed	Excluded
Stucki, *et al.*	2016	PubMed	Excluded
Tasakis, Rafail Nikolaos *et al.*	2021	WoS	Included
Thiemann, T. C. *et al.*	2012	PubMed	Excluded
van Tonder, Andries J. *et al.*	2021	PubMed, WoS	Included
Vasconcelos Dos Santos, Thiago *et al.*	2019	PubMed	Excluded
Wittwer, Matthias *et al.*	2018	PubMed	Excluded
Yan, Zhongqiang *et al.*	2019	PubMed	Excluded
Yang, Chongguang *et al.*	2018	PubMed	Excluded
Yang, Xuemei *et al.*	2020	WoS	Included
Zeller, Mark *et al.*	2021	WoS	Included
Morelli, Marco J. *et al.*	2012	PubMed	Excluded

D Literature Study - The Characterization

Table 4: Topics from the characterization concerning what pathogen was investigated, what kind of data was used as well as the output and software used.

#	Author (Year)	Pathogen (Host)	Output	Software
1	Alvarez, Gonzalo G. *et al.* (2021)	*Mycobacterium Tuberculosis* (Human)	Phylogenetic trees	Maximum likelihood method (using a Tamura-Nei model), Interactive Tree Of Life, SMALT version 0.7.6, SAMtools version 1.4, FreeBayes version 1.1.0, SAMtools mpileup
2	Bousali, Maria *et al.* (2021)	SARS-CoV-2 (Human)	Phylogenetic trees	IQ-TREE, TreeTime, Nextstrain's "augur" pipeline(involves sequence alignment with MAFFT), in house written programs in R to analyze the phylogenetic tree, utilized the libraries "ape", "phangorn", "ggtree" and tidyverse package.
3	Colijn, Caroline *et al.* (2014)	*Mycobacterium Tuberculosis* (Human)	Phylogenetic trees	Matlab's seqpdist and seqneighjoin functions, ClassificationKNN.fit and SVMtrain methods in Matlab, "phyloTop" and "e1071" package in R, Burrows-Wheeler Aligner, samtools mpileup.

Continued on next page

Continued from previous page

#	Author (Year)	Pathogen (Host)	Output	Software
4	Gardy, Jennifer L. *et al.* (2011)	*Mycobacterium Tuberculosis* (Human)	Phylogenetic trees	Maximum Likelihood method GARLI at the CIPRES portal, Bayesian Markov chain Monte Carlo method MrBayes 3.14 as implemented in Geneious 4.7.6, SSAHA v.21, ClustalX 2.0.
5	Illingworth, Christopher J. R. *et al.* (2021)	SARS-CoV-2 (Human)	Maximum likelihood transmission networks	A2B-COVID software package, SQL v18.5.1 and FoodChain-Lab.
6	Jombart, Thibaut *et al.* (2014)	SARS (Human)	Transmission trees and the method outbreaker in R	MUSCLE, DiscrSI from the R package EpiEstim.
7	Klinkenberg, Don *et al.* (2017)	*Mycobacterium Tuberculosis* (Human, animals)	Transmission trees, phylogenetic trees and the method phybreak in R	-
8	Lau, Max S. Y. *et al.* (2019)	Foot- and mouth-disease (Animals)	Model-diagnostc framework for phylodynamic models	-

Continued on next page

Continued from previous page

#	Author (Year)	Pathogen (Host)	Output	Software
9	Lee, Robyn S.*et al.* (2020)	*Mycobacterium Tuberculosis* (Human)	Maximum likelihood trees	IQ-Tree v.1.6.8, Interactive Tree of Life, FastQC v.0.11.5, Trimmomatic v.0.36, miniKraken, Seqtk v.1.2, Burrows Wheeler Aligner MEM algorithm v.0.7.15, Samtools v.1.5, Picard MarkDuplicates v.2.9.0, Genome Analysis ToolKit, snpEff v.4.3t, custom Python scripts v.3.6, Tablet v.1.17.08.17, snp-sites -c v.2.4.0, snp-dists v.0.6, Stata v.15.

Continued on next page

Continued from previous page

#	Author (Year)	Pathogen (Host)	Output	Software
10	Lopez, Mariana G. *et al.* (2021)	SARS-CoV-2 (Human)	Maximum likelihood trees	IQ-Tree with GTR model, iTOL tool, pipeline based on IVAR (Kraken, fastp v 0.20.1, bwa and IVAR v 1.2 , MultiQC, MAFFT, MEGA software, QGIS v.3.14.16-Pi, TempEst v 1.5.3, Beast 2.6, LogCombiner v 2.6.3, Treeannotator v 2.6.3, FigTree v 1.4.3, TreeSlicer, Tracer v 1.7.1, R packages “ape”, “treeio”, “doParallel”, “foreach”, “geosphere”, “lwgeom”, “sp”, “sf”, “rgeos” and “ggplot2”.
11	Tasakis, Rafail Nikolaos *et al.* (2021)	SARS-CoV-2 (Human)	Time-scaled phylogenetic trees	IQ-TREE, VIRULIGN, R (4.0.2) script, pangolin, R stats package and Tidyverse v. 1.3.0, pheatmap v. 1.0.12, dendextent v. 1.14.0,, msa, treeio and ggtree packages from R.

Continued on next page

Continued from previous page

#	Author (Year)	Pathogen (Host)	Output	Software
12	van Tonder, Andries J. *et al.* (2021)	*Mycobacterium Bovis* (Cattle, Badger)	Maximum likelihood phylogenetic trees	IQ-tree v1.6.5, SpoTyping v2.0, RD-analyzer v1.0, Trimmomatic v0.33, BWA mem v0.7.17, SAMtools v1.2 mpileup, BCFtools v1.2, pairsnp v1.0, R library iGRAPH, TransPhylo, R library "phytools", BEAST v1.8.4, LogCombiner v1.8.4, TreeAnnotator v1.8.4, R library TIPDATINGBEAST, R library coda , EAST2 package BASTA, pyjar, R libraries treeio and ggtree, PostgreSQL, R library geosphere, R libraries maps and mapdata.
13	Yang, Xuemei *et al.* (2020)	SARS-CoV-2 (Human)	Phylogenetic tree	RAxML version 8.2.4, iTOL, Nextstrain pipeline (MAFFT, IQ-TREE, Treetime, Augur, Auspice and Inkscape 0.91), MAFFT v7.310, Snippy.

Continued on next page

Continued from previous page

#	Author (Year)	Pathogen (Host)	Output	Software
14	Zeller, Mark *et al.* (2021)	SARS-CoV-2 (Human)	Maximum likelihood tree	HKY nucleotide substitution model, BEAST v1.10.5pre, Apache Spark v2.4.6, PySpark v2.4.6, R package "Epidemia", "outbreak.info.

Table 5: Topics from the characterization concerning super-spreaders and super-shedders as well as the objectives of the study.

#	Identification of super-spreaders?	Effect of super-spreaders?	Relation between super-spreaders and super-shedders?	Objective
1	Yes	Yes	Yes	Outbreak investigation
2	Yes	No	No	Outbreak investigation
3	Yes	Yes	No	Method development /testing
4	Yes	Yes	No	Outbreak investigation
5	Yes	Yes	Yes	Outbreak investigation
6	Yes	No	No	Method development /testing
7	Yes	No	No	Method development /testing
8	No	No	No	Method development /testing
9	Yes	Yes	No	Outbreak investigation
10	Yes	Yes	No	Outbreak investigation
11	Yes	Yes	No	Outbreak investigation
12	Yes	Yes	No	Outbreak investigation
13	Yes	Yes	No	Outbreak investigation
14	Yes	Yes	No	Outbreak investigation

E Information about VTEC O157:H7 Sequences

Table 6: Table showing information about the whole genome sequences of VTEC O157:H7 used in this master thesis. In total 32 sequences from 4 different cattle farms were used, both environmental (8) and samples from calves (24) were sequenced and used.

Sequence ID	Farm	Environment sequence	Calf sequence	Super-shedder (cfu/g feces)	Date of sampling
F9-env-1	Farm 9	X			2015-10-13
F1-1472-1	Farm 1		X		2015-10-26
F1-1475-1	Farm 1		X	X (16000)	2015-10-26
F1-3431-1	Farm 1		X		2015-10-26
F1-1467-1	Farm 1		X		2015-10-26
F1-1459-1	Farm 1		X		2015-10-26
F1-env-1	Farm 1	X			2015-10-26
F1-env-2	Farm 1	X			2015-10-26
F9-5826-1	Farm 9		X	X (900)	2015-10-27
F9-2465-1	Farm 9		X		2015-10-27
F9-5823-1	Farm 9		X	X (15100)	2015-10-27
F9-5802-1	Farm 9		X	X	2015-10-27
F9-env-2	Farm 9	X			2015-10-27
F9-env-3	Farm 9	X			2015-10-27
F7-env-1	Farm 7	X			2015-11-03
F5-env-1	Farm 5	X			2015-11-04
F5-env-2	Farm 5	X			2015-11-04
F5-1499-1	Farm 5		X		2015-11-17
F5-1495-1	Farm 5		X	X (143000)	2015-11-17
F5-8366-1	Farm 5		X		2015-11-17
F7-351-1	Farm 7		X		2015-11-18
F7-349-1	Farm 7		X		2015-11-18
F7-348-1	Farm 7		X		2015-11-18
F7-347-1	Farm 7		X		2015-11-18
F7-346-1	Farm 7		X	X (185000)	2015-11-18
F9-2465-2	Farm 9		X		2015-11-30
F9-2466-2	Farm 9		X		2015-11-30
F9-2471-2	Farm 9		X		2015-11-30

Continued on next page

Continued from previous page

Sequence ID	Farm	Environment sequence	Calf sequence	Super-shedder (cfu/g feces)	Date of sampling
F9-5823-2	Farm 9		X	X (28500)	2015-11-30
F9-5827-2	Farm 9		X		2015-11-30
F7-1992-2	Farm 7		X		2015-12-14
F7-346-2	Farm 7		X		2015-12-14

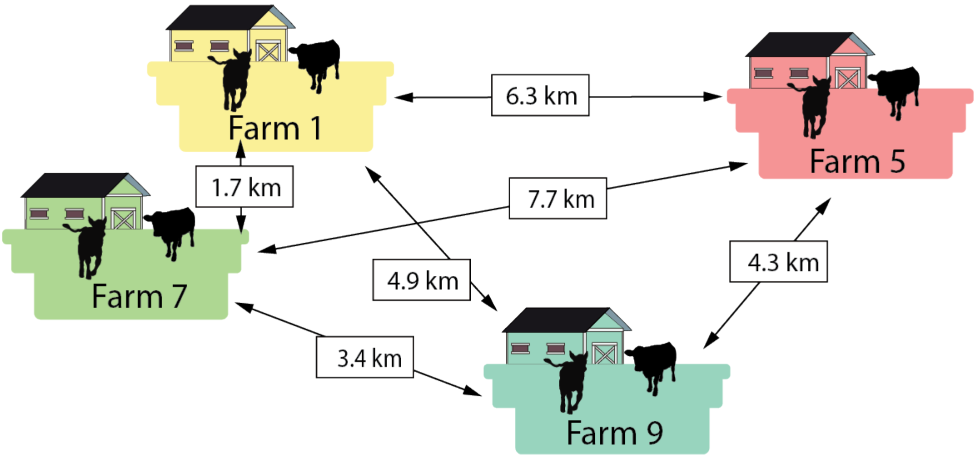


Figure 11: Distances between the four different farms. Illustration used with permission from Lena-Mari Tamminen.

F outbreaker2 - Additional Figures

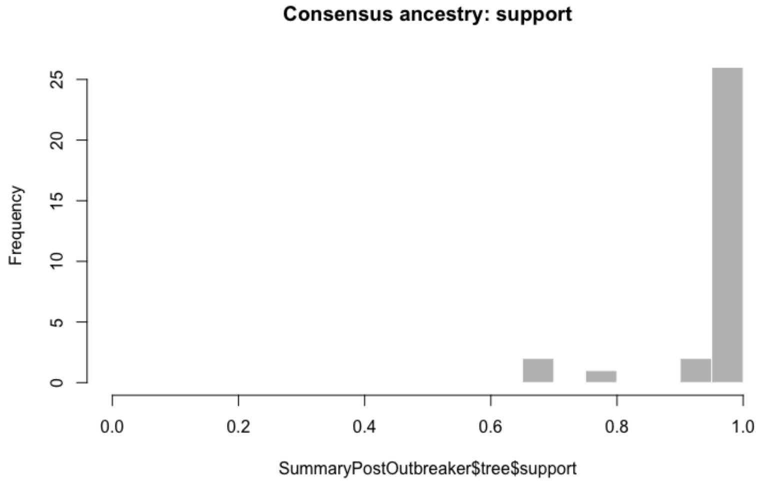


Figure 12: Consensus Ancestry Support for the transmission tree generated from outbreaker2 found in Figure 5.

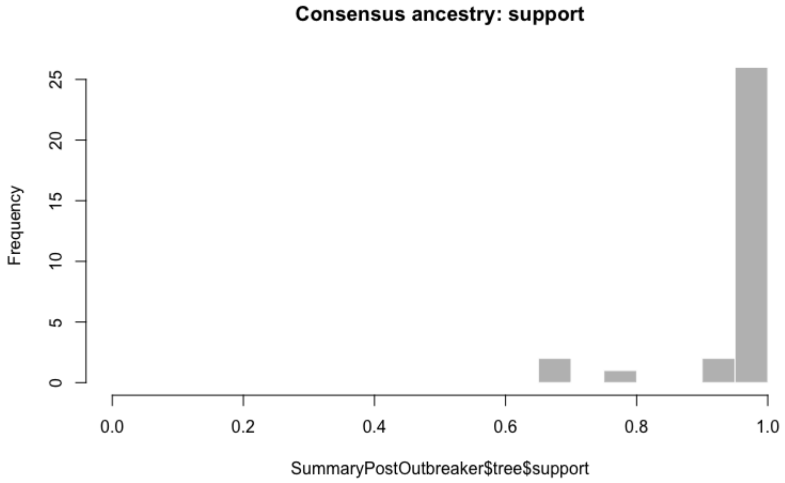


Figure 13: Consensus Ancestry Support for the transmission tree generated from outbreaker2 found in Figure 6.

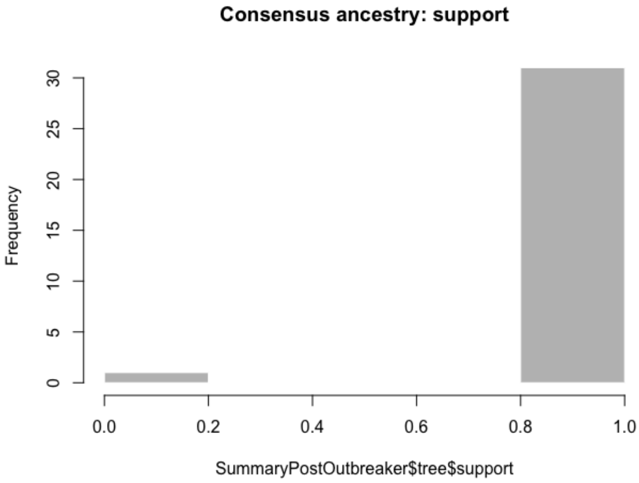


Figure 14: Consensus Ancestry Support for the transmission tree generated from outbreaker2 found in Figure 7.

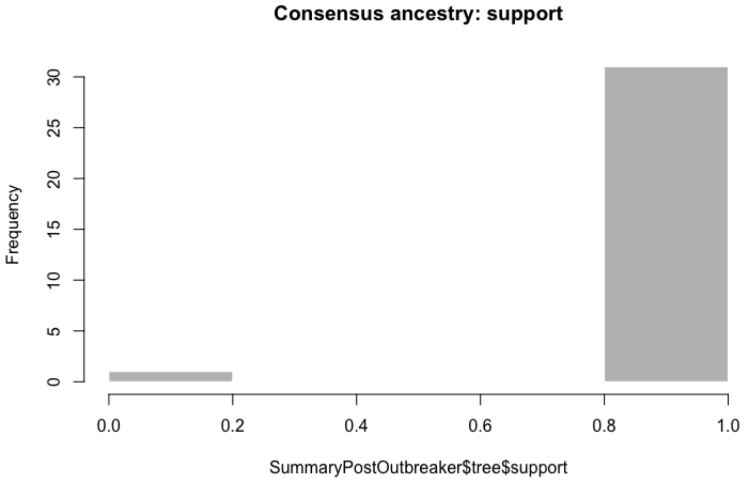


Figure 15: Consensus Ancestry Support for the transmission tree generated from outbreaker2 found in Figure 8.

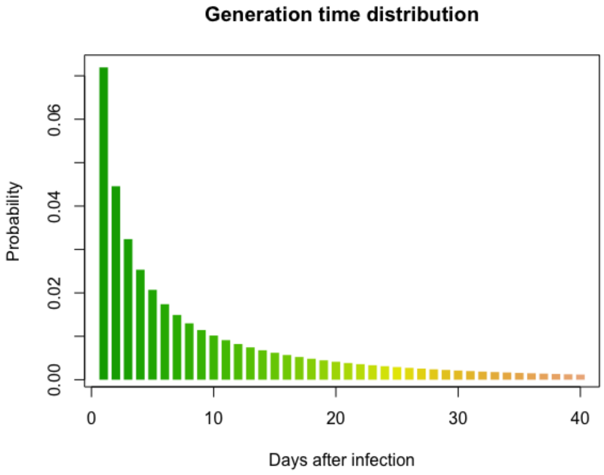


Figure 16: Generation time distribution, (a gamma distribution) based on the generation time interval.

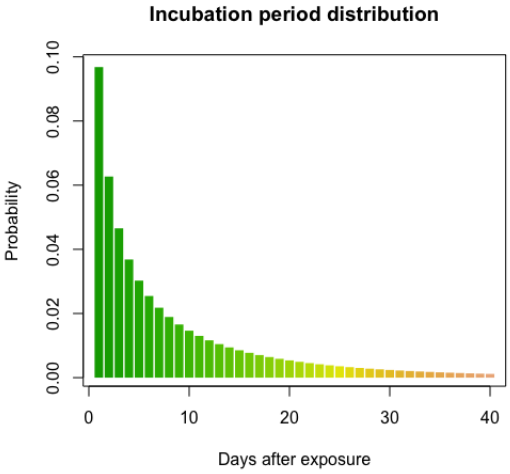


Figure 17: Incubation period distribution, (a gamma distribution) based on the incubation period interval.

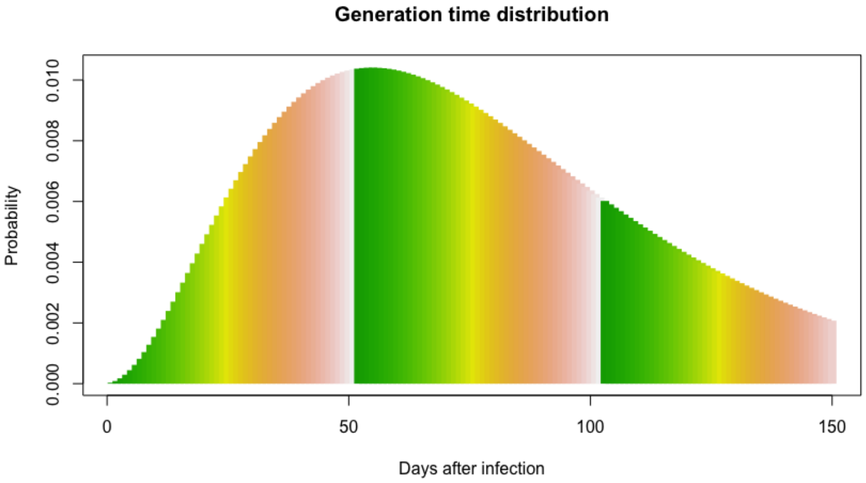


Figure 18: Generation time distribution, (a gamma distribution) based on the default gamma shapes values.

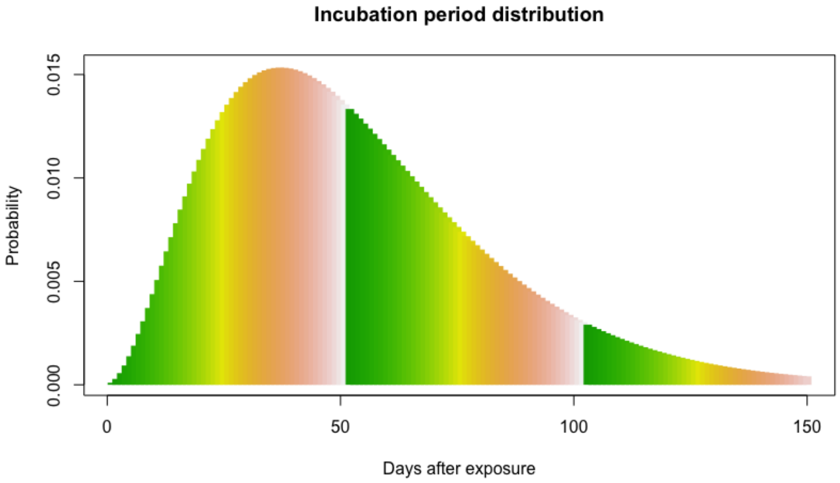
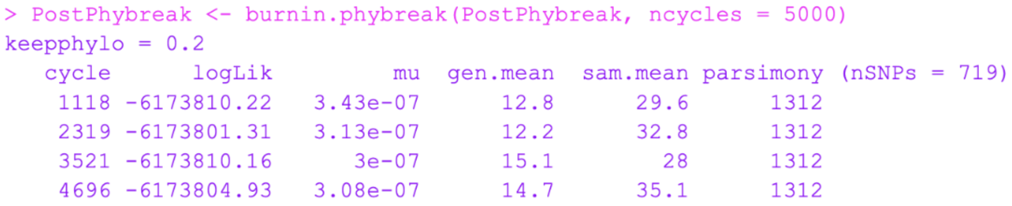


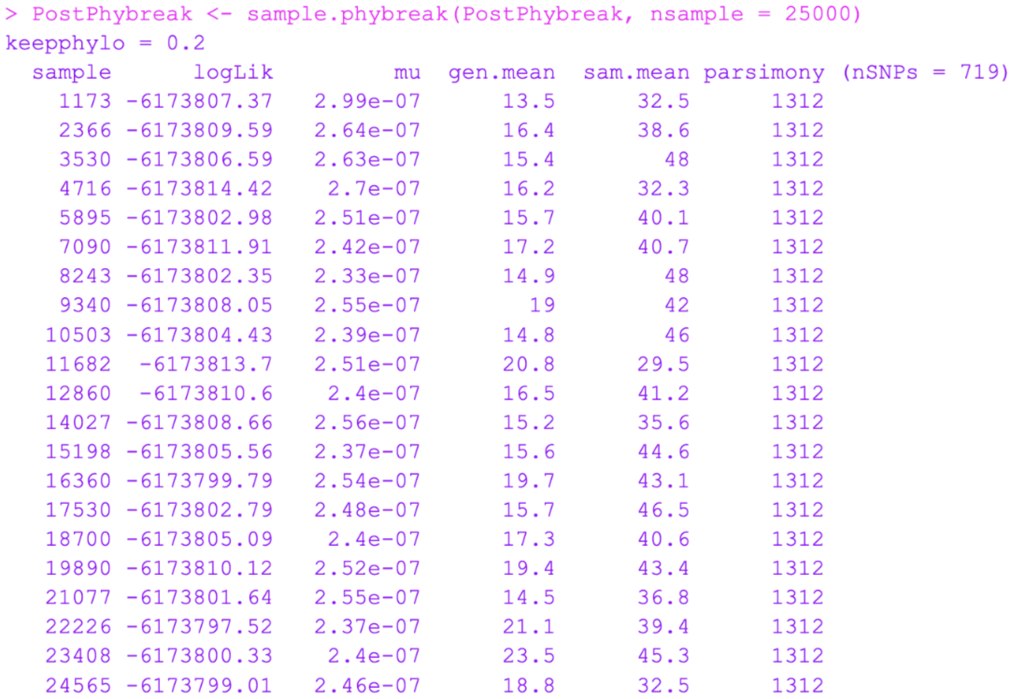
Figure 19: Incubation period distribution, (a gamma distribution) based on the default gamma shapes values.

G phybreak - Additional Figures



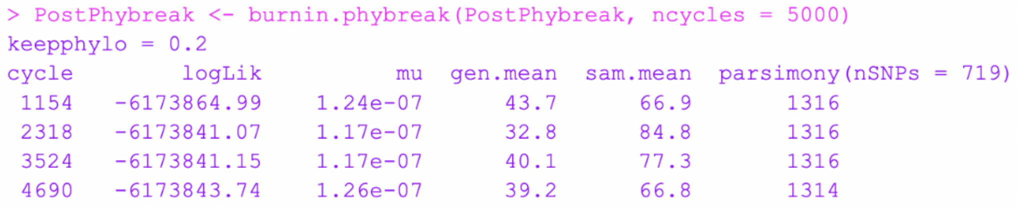
```
> PostPhybreak <- burnin.phybreak(PostPhybreak, ncycles = 5000)
keepphylo = 0.2
   cycle      logLik          mu  gen.mean  sam.mean parsimony (nSNPs = 719)
    1118 -6173810.22   3.43e-07      12.8      29.6      1312
    2319 -6173801.31   3.13e-07      12.2      32.8      1312
    3521 -6173810.16      3e-07      15.1        28      1312
    4696 -6173804.93   3.08e-07      14.7      35.1      1312
```

Figure 20: Result from the MCMC burnin iteration in phybreak, using the mutation rate 3.95^{-10} mutations per site per day.



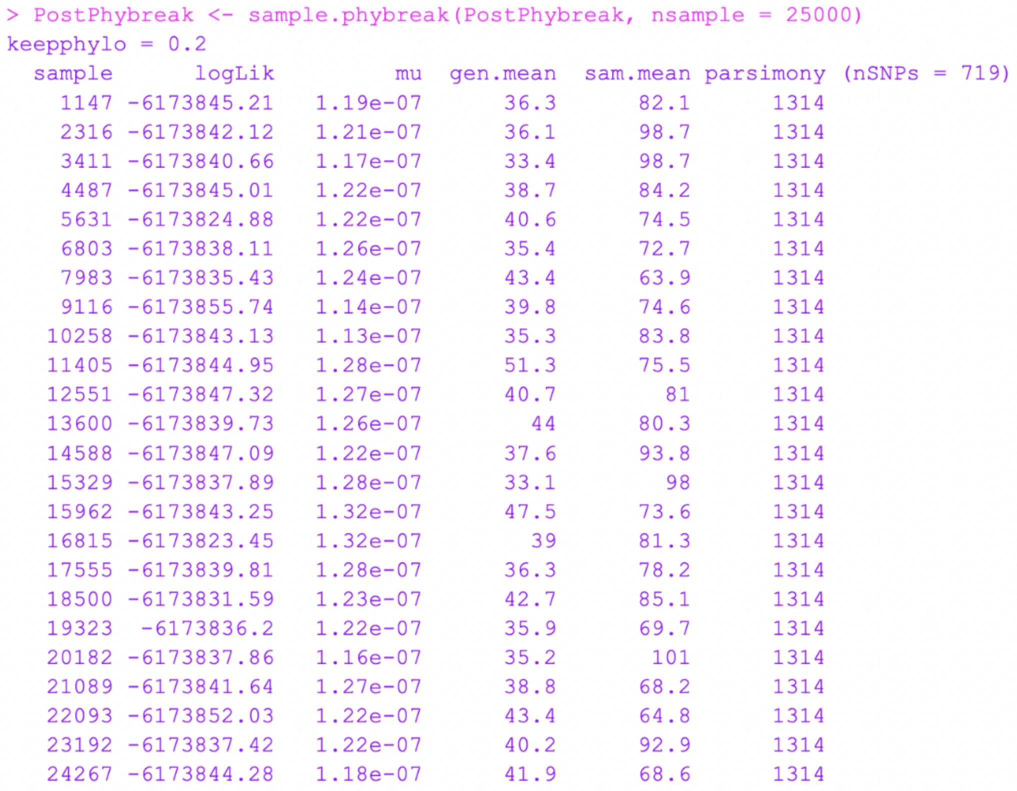
```
> PostPhybreak <- sample.phybreak(PostPhybreak, nsample = 25000)
keepphylo = 0.2
  sample      logLik          mu  gen.mean  sam.mean parsimony (nSNPs = 719)
    1173 -6173807.37   2.99e-07      13.5      32.5      1312
    2366 -6173809.59   2.64e-07      16.4      38.6      1312
    3530 -6173806.59   2.63e-07      15.4        48      1312
    4716 -6173814.42    2.7e-07      16.2      32.3      1312
    5895 -6173802.98   2.51e-07      15.7      40.1      1312
    7090 -6173811.91   2.42e-07      17.2      40.7      1312
    8243 -6173802.35   2.33e-07      14.9        48      1312
    9340 -6173808.05   2.55e-07        19        42      1312
   10503 -6173804.43   2.39e-07      14.8        46      1312
   11682  -6173813.7   2.51e-07      20.8      29.5      1312
   12860  -6173810.6    2.4e-07      16.5      41.2      1312
   14027 -6173808.66   2.56e-07      15.2      35.6      1312
   15198 -6173805.56   2.37e-07      15.6      44.6      1312
   16360 -6173799.79   2.54e-07      19.7      43.1      1312
   17530 -6173802.79   2.48e-07      15.7      46.5      1312
   18700 -6173805.09    2.4e-07      17.3      40.6      1312
   19890 -6173810.12   2.52e-07      19.4      43.4      1312
   21077 -6173801.64   2.55e-07      14.5      36.8      1312
   22226 -6173797.52   2.37e-07      21.1      39.4      1312
   23408 -6173800.33    2.4e-07      23.5      45.3      1312
   24565 -6173799.01   2.46e-07      18.8      32.5      1312
```

Figure 21: Result from the MCMC sample iteration in phybreak, using the mutation rate 3.95^{-10} mutations per site per day.



```
> PostPhybreak <- burnin.phybreak(PostPhybreak, ncycles = 5000)
keepphylo = 0.2
cycle         logLik          mu  gen.mean  sam.mean  parsimony(nSNPs = 719)
 1154   -6173864.99   1.24e-07      43.7      66.9       1316
 2318   -6173841.07   1.17e-07      32.8      84.8       1316
 3524   -6173841.15   1.17e-07      40.1      77.3       1316
 4690   -6173843.74   1.26e-07      39.2      66.8       1314
```

Figure 22: Result form the MCMC burnin iteration in phybreak, using the mutation rate 6.19^{-10} mutations per site per day.



```
> PostPhybreak <- sample.phybreak(PostPhybreak, nsample = 25000)
keepphylo = 0.2
  sample       logLik          mu  gen.mean  sam.mean parsimony (nSNPs = 719)
    1147 -6173845.21  1.19e-07      36.3      82.1      1314
    2316 -6173842.12  1.21e-07      36.1      98.7      1314
    3411 -6173840.66  1.17e-07      33.4      98.7      1314
    4487 -6173845.01  1.22e-07      38.7      84.2      1314
    5631 -6173824.88  1.22e-07      40.6      74.5      1314
    6803 -6173838.11  1.26e-07      35.4      72.7      1314
    7983 -6173835.43  1.24e-07      43.4      63.9      1314
    9116 -6173855.74  1.14e-07      39.8      74.6      1314
   10258 -6173843.13  1.13e-07      35.3      83.8      1314
   11405 -6173844.95  1.28e-07      51.3      75.5      1314
   12551 -6173847.32  1.27e-07      40.7        81      1314
   13600 -6173839.73  1.26e-07        44      80.3      1314
   14588 -6173847.09  1.22e-07      37.6      93.8      1314
   15329 -6173837.89  1.28e-07      33.1        98      1314
   15962 -6173843.25  1.32e-07      47.5      73.6      1314
   16815 -6173823.45  1.32e-07        39      81.3      1314
   17555 -6173839.81  1.28e-07      36.3      78.2      1314
   18500 -6173831.59  1.23e-07      42.7      85.1      1314
   19323  -6173836.2  1.22e-07      35.9      69.7      1314
   20182 -6173837.86  1.16e-07      35.2       101      1314
   21089 -6173841.64  1.27e-07      38.8      68.2      1314
   22093 -6173852.03  1.22e-07      43.4      64.8      1314
   23192 -6173837.42  1.22e-07      40.2      92.9      1314
   24267 -6173844.28  1.18e-07      41.9      68.6      1314
```

Figure 23: Result from the MCMC sample iteration in phybreak, using the mutation rate 6.19^{-10} mutations per site per day.