Examining the multitude of available methods for attributing sources to molecular infection and antimicrobial resistance

Name: Tristan Schadron (6669980)

First Reviewer: Dr A.L. Zomer

Second Reviewer: Dr. L. Mughini-Gras

Date: 2/4/2022

Abstract

In order to counteract disease outbreaks, monitor pathogen populations, and allow preventive measures to be put in place against pathogens, pathogens need to be attributed to putative sources. To this extent source attribution may look at phenotypical and genotypical characteristics of the pathogen to link it to a source. Proper designation to a source requires overcoming problems related to the pathogen and sources characteristic, which may erase recognizable patterns differentiating one serovar strain from another. However, no standard approach to source attribution exists, which overcomes the problems and limitations inherent therein. No standard approach to all source attribution tasks is likely to exist, however by combining different genotyping approaches using WGS data pathogens can be attributed with a higher resolution. Here biological problems and technical problems associated with source attribution, among which host range, host switching behavior, genome plasticity, source designation, metadata annotation, problems with data, and spatio-temporal dynamics are evaluated. These technical and biological problems are placed in context of different phenotyping, genotyping and genotype-based microbial source attribution approaches to give an intuitive overview of the strengths and weaknesses of the aforementioned approaches. Agreeing with previous papers, we find that a combination of genotyping approaches is the best way forward. However, WGS genotyping approaches require standardization before universal application. We hope to highlight possible research directions, such as to what extent genetic signals are associated with adaptation, and by proxy attributable to a source. Additionally, we stressed the relevance of spatio-temporal data to expand source attribution capabilities.

Layman's summary

Disease outbreaks can occur sporadically, during which one of the first steps taken is the identification of the possible source of the outbreak. For example, a *Salmonella* outbreak can originate from infected meat, which in turn could originate from some chickens on a random farm located somewhere far off in the hinterland. Source attribution refers to the identification of this source from which pathogen originates. Human cases of foodborne diseases may be attributed to animals, food, and environmental reservoirs. Through the transmission routes, direct exposure and risk factors are also often considered.

The relevance of source attribution lies not only in its' ability to determine how outbreaks can be counteracted, but also in the prevention of the spread of pathogens. Knowledge of the source of a pathogen allows relevant authorities to perform food safety interventions and measure the impact of such interventions. Such intervention may include the removal of products from shelves, temporary closure of affected animal farms, and improved hygiene in slaughterhouses.

In order to get data on these pathogens, samples are taken from the human cases as well as possible sources. Genetical information of the pathogen and observable characteristics of the pathogen are used to determine species and perhaps even the specific strain; e.g. proteins expressed on the surface of the pathogen can be used as a characteristic. Similarities and differences between pathogens from the sources and human cases can then be used to infer the relatedness and origin of the pathogen.

Source attribution may sound intuitive, but specific characteristics of pathogens and sources may complicate matters. Certain pathogens or strains often switch hosts, some have broad-host ranges, and some are more inclined to have their genetic information shuffled and changed. Additionally, for a link between source and pathogen to be determined the source should be included and clearly defined. All these characteristics and considerations muddle a clear inference of a source.

To overcome these problems many different techniques were either developed or applied. All these techniques have their specific downsides. However, genetic information, through the sheer quantity of available information contained within a single genome by means of nucleotide sequences, can reveal many more patterns of similarity and dissimilarity.

I propose that different approaches to retrieve these patterns/signals from genomic nucleotide sequences should thus be applied together to compare one approach to another. Moreover, metadata may be applied to further distinguish both source and pathogens. Metadata refers to specific information pertaining to data acquisition, such as the host the sample was taken from, the place, and time of acquisition. However, research is still required to reveal the temporal and geographical effects on source attribution, and to reveal relevant patterns within different pathogens and pathogen strains. These patterns may change through time, thus a continuously updated reference frame is needed to standardize source attribution and allow tracking of signatures associated with particular strains.

Introduction

In the year 2019 a sudden gulf of disease whipped across Germany, even potentially being the cause of death for two elderly persons. As to the root of this disease; sausages contaminated with *Listeria*, a bacteria that causes gastroenteritis, meningitis and sepsis (*Deutsche Presse-Agentur, 2019*).

The quick identification of origin of the outbreak, preventing further disease and potentially deaths, was thanks to source attribution. Source attribution allows identification of the sources of the outbreak, which in turn allows a quick response from the relevant authorities to counteract the outbreak. Source attribution is defined as the subdivision of human cases caused by foodborne pathogens to their sources (*Mughini-gras et al., 2019*).

The term source, in relation to source attribution, denotes not only the reservoir the disease originates from, such as for example a variety of farm animals, but oftentimes includes the route of transmission, the direct exposure and the risk factors (*Mughini-gras et al., 2019*). Examples of transmission routes are both food and the environment, whilst exposure may occur through meat or milk, and risk factors might be the consumption thereof or the proximity to livestock (*Wagenaar., 2015*).

There are a plethora of different methodologies used for source attribution. These techniques commonly fall either within the categories; microbiological methods, epidemiological methods, or a combination of both (*Mughini-gras et al., 2019*). Whereas, the epidemiological approach to source attribution looks at the distribution and determinants to infer sources, the microbiological methods use phenotype or genotype data to characterize the pathogen and trace the source of the infection.

Source attribution is, however, still far from perfect. There are no universal consensus approaches for attributing sources to most foodborne diseases, since information regarding their source is lacking, which complicates understanding. This is partly due to the majority (around 95%) of these cases being sporadic, causing no wider outbreak (*Zhang et al., 2019*). Additionally, improving source attribution would lessen the number of cases and economical burden associated with an outbreak through prevention and mitigation. For example, *Salmonella enterica*, a prevalent foodborne pathogen, infects over a million humans and causes considerable economic damages of around $ 3.7 billion a year (*Zhang et al., 2019 & Hoffmann et al., 2015*). Identification of the sources responsible for disease is therefore crucial.

Despite still being imperfect there are ample opportunities for the improvement of source attribution. The advent of next-generation sequencing, specifically whole genome sequencing (WGS), allows greater resolution as compared to traditional genetic approaches to source attribution (*Franz et al., 2016*). In addition, the continued improvement of WGS and the lessening costs support innovative new ways to tackle source attribution. For example, larger amounts of data have made the application of the increasingly popular field of machine learning in the context of source attribution possible (*Arning et al., 2021, Lupolova et al., 2019 & Munck et al., 2020*). In addition, the bipartite nature of the association between sample and source, allows for a variety of network

related analyses (*Merlotti et al., 2020*). These techniques and others open up a variety of avenues for the development of source attribution

However, dually there is a need to overcome a myriad of problems inherent in source attribution and the types of data used. For example, pathogen behavior, such as host switching (*Dearlove et al., 2016 & Woodcock et al. 2017*), genome recombination (*Woodcock et al., 2017*), and a generalist lifestyle (*Dearlove et al., 2016*), affect source attribution. Additionally, current gaps in knowledge, such as unknown sources (*Wilson et al., 2008*), suboptimal source clustering (Griekspoor et al., 2013), the inability to sample everything (*Lupolova et al., 2019*), and the lack of spatio-temporal data (*Smid et al., 2013*) hamper precise source attribution. Lastly, there appears to be no standardized protocol for any of these methods encompassing different pathogens.

The great variety of ways to tackle source attribution, with their tacit benefits and shortcomings, the various problems related to source attribution, and the availability and nature of the data begs the question; what methods or combinations of methods are best able to overcome the problems and limitations inherent in source attribution to perform accurate and precise attribution to the correct sources?

All by all, no standard approach to all source attribution tasks exists, and there probably will not be a general fit-all approach since pathogen behavior differs. Therefore, decisions on a case-by-case basis need to be made considering behavior and available data. Nevertheless, increasingly the existing gaps in knowledge are bridged allowing more sophisticated approaches to be taken. Despite this, much research is still required to evaluate and optimize the best decision making for different pathogens and choice of sources.

Main body of text

1. Biological problems related to source attribution

In order to figure out the best methods to apply to a source attribution problem, it is important to understand the biological problems associated therewith. Pathogen behavior influences the genetic makeup and plasticity of the genome. For example, the host range of a given pathogen influences the level of specialization. The wider the host range the less specialized a pathogen is and the more it conforms to a generalist lifestyle. Some species of pathogen like *Campylobacter coli* and *C. jejuni* have a wide host range, including cattle, sheep, poultry, pigs, wild birds and wild mammals (*Dearlove et al., 2015*). Whereas, live-stock associated lineages of *Staphylococcus aureus* are mostly host-restricted (*Woodcock et al., 2017*). However, even within species there may be differences up to serovar level in their host restriction. To illustrate, whilst *Salmonella enterica* serovar Typhimurium has a rather broad host range, the serovar Typhi is restricted to fewer hosts (*Lupolova et al., 2019*).

Possible links were made between a generalist lifestyle and genome plasticity. A study by *Woodcock et al., 2017* inferred that outcompeting of *Campylobacter* generalists by specialists, who have a fitness advantage in a specified niche, was prevented by the introduction of new genetic variance. Their model specifies that horizontal gene transfer, frequent host switching and recombination can promote host generalism by enhancing the standing variation in populations and enhancing the efficiency of selection of combinations of beneficial alleles.

Rapid host switching is problematic for source attribution. Namely, the transient presence of a pathogen within a host can cause errors in the acquired metadata associated with genomic techniques (*Lupolova et al., 2019*). Wrong labeling decreases the source attribution capacities and for models requiring training with data may introduce faulty patterns. Furthermore, studies on the capacity of whole-genome sequencing of *Campylobacter* strain's broad host range found that the rapid rates of host switching erased the phylogenetic association between host and pathogen complicating source attribution (*Dearlove et al., 2015*). After all, the genetic signatures of such strains were omnipresent and were not limited to a specific host. Overall it might be worth investigating the commonalities and differences between pathogens adapting to a single host and broad host ranges, such as patterns of gene degradation (*Wheeler et al., 2018*).

Affinity for a host does not only play into host switching, but also plays into the ability of a pathogen to infect a certain host. After all, the contributions of a host to human infection do not necessarily reflect the composition of the sampled source (*Arning et al., 2021*).

Concerning the gathering of data there are also a few things to be kept in mind; only a fraction of information is gathered, there may be unknown sources, and samples should be representative of the sources. The fraction of information gathered is but a mere fraction of the in total existing data available. Therefore, the different steps between the original source and the eventually diseased individual need to be inferred or circumvented. This means that the source identified may not fit one-on-one with the disease isolate, since the real source was never sampled. *Wilson et al., 2008* used re-sampling to show that no major unknown sources for *C. jejuni* remained to be discovered.

However, even smaller sources can underlie a large reassignment of putative sources (*David et al., 2013b*). In addition, as only a fraction is sampled it does not represent the entire population structure, which means the data is biased to some extent (*Lupolova et al., 2019*)

Wrongly designated source, which should be multiple genetically diverse sources may affect performance and cause ambiguity. For instance many studies erroneously group wild birds into a single group, however *Griekspoor et al., 2013* showed a strong differentiation between *C. jejuni* populations in wild birds on a nucleotide level. These differences, however, were obscured on an allele level, which comparatively lacked resolution.

Probabilistic source attribution at times is preferred over discrete source attribution, because it is able to reflect the uncertainty in designation of the source. For example, probabilistic source attribution helps in the attribution of *Campylobacter* isolates to either sheep or cattle is iffy, since there is a large overlap in the gene pool, due to frequent transmission between cattle and sheep, which may reflect the physiological makeup of the gastrointestinal tract of ruminants (*Wilson et al., 2008*). This is related to source clustering where in a perfect situation low diversity within sources and high diversity between sources is expected. However, the close proximity and regular contact between some host species and broad-host ranges muddle separation between sources. Techniques, such as AMOVA, exist to examine this variance (*Excoffier et al., 1992*).

In some cases the addition of metadata helps elucidate source separation. However, for the majority of available sequence data metadata is not readily available (*Lupolova et al., 2019*). Additional information such as the geographical location of the sample help further source prediction and data separation.

Palma et al., 2018 found that geographically segregated *S. enterica* Typhimurium isolates formed distinct monophyletic groups, suggesting that the spatial factor may influence source attribution. For example, a mix of datasets disregarding geographical or temporal information may increase the data available, but may increase within cluster variance. Geographically separated population may evolve independently from each other even within similar niches (*Lupolova et al., 2019*). The decision of local vs. global sampling thus reflects choices in general and more distinct data patterns. The same might be true for some hosts. However, factors like travel and trade may also influence this (*Smid et al., 2013*). *Griekspoor et al., 2013* noticed a less pronounced geographical pattern in wild bird populations, as compared to different livestock animals. This is especially relevant to pathogens undergoing fast genomic change where global sampling obscures source attribution. On a temporal scale fast genomic changes can cause distinct patterns between pathogen populations, which ideally should be accounted for. *Smid et al., 2013* looked into the effects of differing spatio-temporal data of *C. jejuni* and *C. coli*. They found that even on a small time-scale, dissimilarity was introduced in sources. Ignoring temporal differences between data may thus cause a temporal bias. Introduction of a temporal aspect allows evaluation of contributions of sources and the impacts of intervention as well (*Mughini-gras et al., 2019*).

2. The various methodologies for source attribution

2.1. Epidemiological methods

Among the various methodologies used for source attribution are the so-called epidemiological approaches. Better known epidemiological approaches include; case-control studies, cohort studies, and outbreak analysis.

Case-control studies compare the frequencies of exposure to a factor between a control group and the group of diseased. Whereas, a cohort study compares an unexposed group to an exposed group and based thereon compares the frequency of the cases (*Mughini-gras et al., 2019*). With this in mind, data is collected through means of questionnaires or interviews. There are a variety of problematic factors involved in case-control and other epidemiological studies; concurrent exposure to multiple sources and a variety of biases complicate epidemiological studies (*Fullerton et al., 2012*). Some examples of biases include recall bias, information bias, and selection bias. Selection bias refers to the differential selection between cases and controls, where they should have the same characteristics. Whereas, information bias refers to mistakes in the measurements of associations. Recall bias occurs when experiences cannot be recalled accurately or details are omitted. To prevent recall bias the recall period often equals the maximum duration of the incubation of the disease. In addition, these epidemiological studies suffer from a variety of other confounding effects. Partially, these effects can be accounted for, raising statistical power, by controlling, e.g. age, gender, health-related issues and geographical area (*Fullerton et al., 2012*). Nevertheless, these approaches are subject to certain biases. This may warrant combining epidemiological approaches with molecular/microbiological approaches in order to improve source attribution.

2.2. Microbiological & molecular methods

The molecular methods often include data gained through phenotypic expression patterns of pathogens used to attribute a source to the sampled pathogen (*van Belkum et al., 2007*). Whereas, the microbiological methods can roughly be subdivided in population genetic models and frequency-matching models (*Mughini-gras et al., 2019*). Frequency-matching models determine the putative source in a probabilistic manner by accounting for exposure, prevalence of pathogen subtypes in sources, and frequency of cases by these subtypes (*Mughini-gras et al., 2019*). On the other hand, the population genetic models are based on various types of genotype data discussed in the genotyping section. This genotype data is then used for source attribution. Either directly through comparison of the samples to the source samples (*Foley et al., 2006*) or through the use of Bayesian models (*Wilson et al., 2008*), machine learning approaches (*Arning et al., 2021, Lupolova et al., 2019 & Munck et al., 2020*), phylogenetic trees (*Gymoese et al., 2017 & Henri et al., 2017*), and network-like approaches (*Merlotti et al., 2020*) to name a few.

2.2.1. Phenotyping

Phenotyping includes methods such as phage-typing (*Barco et al., 2013*), serotyping (*Grimont & Weill., 2007*), antimicrobial resistance profiling (*Barco et al., 2013*), and biotyping (*Mughini-gras et al., 2019*) . Phenotyping is based on the assumption that similar organisms group together, as a result of the expression of their genotype (*van Belkum et al., 2007*). For example, antimicrobial resistance profiling tests a pathogen's or strain's susceptibility to antimicrobial agents, whereas biotyping differentiates based on reactions to biochemical tests (*Barco et al., 2013*). These methodologies used to dominate the field, but decreased sequencing costs and improvement in sequencing sensitivity out-phased these now classical techniques.

Here only phage-typing and serotyping are discussed in detail, since to a large extent these used to dominate the field.

Phage-typing

Phage-typing attempts to separate strains into phage-types based on the ability of a phage to lyse said strains (*Barco et al., 2013*). This ability depends both on the serovar specific receptors present on the bacterial surface and the molecular characteristics of the phage (*Ferrari et al., 2017*).

Phage-typing keeps being persistently used into the recent years, despite limitations attached to the method. Firstly, there is a lack of available phages, with only over 300 phage types being used to discriminate (*Ferrari et al., 2017*). Secondly, changes in phage types, such as serotypes, called phage conversion, can occur. Underlying reasons for phage conversion include changes to lipopolysaccharide, a component of the outer wall of gram-negative bacteria, loss/acquisition of plasmids, and the expression of so-called temperate phages (*Barco et al., 2013 & Ferrari et al., 2017*). Temperate phages progress through lysogenic lifecycle, thus will not lyse. Thirdly, more common phage-types cannot be discriminated between (*Hopkins et al., 2011*). Lastly, as compared more recent genotyping methods capable of differentiating serovars showing little phenotypic variation, phage-typing has relatively low discriminatory power and differentiates poorly among isolates of different animal sources (*Merlotti et al., 2020*). However, there do exist specific strains that are considered clones by a genotyping method called PFGE, but differ in the amount of lysis reactions (*Gorman and Adley., 2004*).

Serotyping

Serotyping is the classification among bacteria or virus species based on the expression of their surface antigens (*Barco et al., 2013*). The method uses agglutination, clumping of bacteria or viruses with specific serotypes. Serotyping is possible due to certain serotypes being associated with geographical regions or particular hosts (*Ranieri et al., 2013*). Traditional serotyping considers antigenic variability in flagellar and somatic antigens (*Barco et al., 2013*).

There are several drawbacks associated with serotyping. For starters, not all (sub)species can be distinguished amongst. Therefore, serotyping often lacks the sensitivity to discriminate food-borne illnesses or infer phylogenetic relationships (*Ranieri et al., 2013*). Some studies attempted to supplement serotyping with new serotypes (*Wattiau et al., 2011*). However, even if resolution is improved, the antisera used is expensive and the procedure takes at least three days to complete (*Ranieri et al., 2013*). As such, serotyping is slowly being phased out by cheaper genotyping techniques, which can reveal the genetic fundaments underlying phenotypic expression.

2.2.2. Genotyping

There are multiple different genotyping approaches such as PFGE (*Swaminathan et al., 2001*), MLVA (*Lindstedt et al., 2003*), ribotyping (*Grimont & Grimont., 1986*) and MLST (*Maiden et al., 1998*). Additionally, the advent of whole genome sequencing allowed complete sets of genes to be considered, and consequently resulted in whole-genome MLST (wgMLST) and core genome MLST (cgMLST) (*Sheppard et al., 2012* & *Alikhan et al., 2018*). Genotyping allows for the comparison in overlap between different sources and human samples, based on the presence or absences of a core set of genes/SNPs/sequences. Genotyping can, however, also be used in conjunction with more complicated models and techniques to derive source attribution from more complicated patterns. For example, machine learning and Bayesian models may be applied, phylogenetic trees can be constructed, or networks can be created. Additionally, genotypes can be used to infer hamming distances, allele frequencies, or other frequency-based genetic distances. To illustrate, k-mer frequencies can be used to discern differences between isolates, through which source attribution may occur by means of the aforementioned techniques.

All by all, a variety of different elements within the genome can be considered, such as tandem repeat sequences, ribosomal RNA, or antimicrobial resistance genes. Equally, CRISPR-cassettes are an option, since the spacers undergo acquisition and removal of obsolete ones through time, since a set amount of spacers is present within the genome in order to retain a sufficient relative fitness. As such, inference of phylogeny and strain differentiation can be observed through sequencing of the CRISPR system (*Ferrari et al., 2017*).

Restriction fragments length polymorphism (RFLP) (*Swaminathan et al., 2001*) and ribotyping (*Grimont & Grimont., 1986*) are some of the first genotyping techniques. Both are PCR-based techniques, based on the amplification of genome fragments through the use of restriction enzymes. The selection of restriction enzymes decides the size and number of fragments, which can then be separated on an electrophoresis gel. The selection of restriction enzymes influences the discriminatory power of the techniques. Ribotyping specifically uses rRNA information. Other techniques exist which focus on different kinds of fragments (*Wang et al., 2011 & Ferrari et al., 2017*). RFLP is often used in conjunction with Pulsed-Field Gel Electrophoresis (PFGE), a technique fragmenting the genome using partially off-target restriction enzymes combined with electrophoresis (*Swaminathan et al., 2001*). PFGE results in approximately 15 fragments of varying sizes in accordance with specified strains. Nowadays these PCR-based genotyping techniques are largely supplanted by better discriminating genomics-based techniques, such as MLST (*Ribot et al., 2019*).

Here the most prominent genotyping techniques are discussed, but other techniques, such as fla-sequence typing (*Meinersmann et al., 1997*) and comparative genomic fingerprinting (*Taboada et al., 2012*), exist as well.

MLVA

Multi-locus variable number tandem repeat analysis (MLVA) (*Lindstedt et al., 2003*) looks at the variable number tandem repeats, which are various regions with nucleotide repeats both in the coding and non-coding DNA of bacterial genomes (*Sabat et al., 2013*). These regions may vary from a couple to over a hundred base pairs in length (Lindstedt et al., 2003). The number and composition of these tandem repeat sequences of different lengths at different loci in the genome is represented as a profile (*Mughini-gras et al., 2018*). In short, the bacterial sequence gets amplified through PCR and through gel electrophoresis or capillary electrophoresis the size of the fragments containing tandem repeat sequences is determined (*Ferrari et al., 2017*). Based thereon, MLVA can differentiate epidemic strains from endemic ones, since profile changes occur rapidly (*Mughini-gras et al., 2018*).

Dually, this comes with a downside. Since regardless of the stability of the tandem repeats, loci changes may occur; e.g. recombinations and mutations (*Hopkins et al., 2007*). As well as, multiple changes may occur at a single locus, which erases traces of common ancestry between MLVA profiles (*Wuyts et al., 2013*). For salmonella typhimurium certain loci were identified as the hotbed of variations and measures were taken to account for these characteristics, a study for example allowed a single variant in those loci to be present (*Schjørring et al., 2016*). This may be due to their location in prophages or plasmids being relatively unstable or absent (*Litrup et al., 2010*). Therefore, understanding the evolutionary dynamics of the changes in repeat sequence and their relationship to the genome is fundamental in performing adequate MLVA analysis (*Ferrari et al., 2017*). Resultantly, MLVA should be considered only in well characterized bacterial strains.

MLVA is used in outbreak studies and surveillance, since compared to PFGE it boasts a higher discriminatory power (*EFSA, 2013*). However, as compared to next-generation sequencing data MLVA may be lacking. In a study by *Octavia et al., 2015*, MLVA showed a lack of resolution as compared to whole genome sequencing, since whilst the MLVA types were identical a definite difference in one or more SNPs was observed. Therefore, similar MLVA profiles may be observed in cases that may be epidemiologically unrelated (Octavia et al., 2015b). Moreover, in another study *Octavia et al., 2015b* found that earlier detection with less required cases per cluster was possible with whole genome data, since linked cases are easier identified.

However, whole genome sequencing often employs short-read sequencing, which makes the reconstruction of the entirety of the repeats nigh impossible (*Munck et al., 2020*). The continued improvements to long-read sequencing (*De Roeck et al., 2019*) should make MLVA possible and affordable, either alone or in combination with short-read sequencing.

Multi-locus sequence typing

A better known sequencing-based source attribution method is multi-locus sequence typing (MLST). MLST depends on schematic differences in a group of chosen loci in the genome of bacterial species. Depending on the loci the sequence type (ST) and clonal complex (CC) can be established (*Maiden et al., 1998*). As such, the number and diversity of known ST determines the source attribution capabilities of MLST. For example, There is a 7-gene MLST scheme for *Salmonella* capable of determining species level, but when determining subspecies or serotype level the technique may lack in resolution (*Foley et al., 2006*). The reason being that these genes are housekeeping genes, which only slowly accumulate nucleotide changes. As such, MLST is able to reconstruct evolutionary histories, but is less capable of determining differences among strains when researching an outbreak (*Foley et al., 2006*). Nevertheless, genomes showing greater plasticity also present complications, since spurious links are inferred between isolates due to higher recombination rates (*Lees et al., 2019*). However, a benefit to MLST is that besides the direct comparison of attained STs to known STs for the characterization of the strains a variety of different techniques can be used to analyze the data, such as phylogenetic trees, machine learning, and frequency-matching models. To this effect distances between isolates or allele frequencies can be calculated.

Whole-genome sequencing has opened up new avenues for genotyping, and especially for MLST. These come in the form of whole-genome MLST (wgMLST) and core-genome MLST (cgMLST). Both cgMLST and wgMLST compare a number of hundreds to thousands of variable sequences or genes (*Sheppard et al., 2012* & *Alikhan et al., 2018*), but they differ in what genes they consider. cgMLST considers the core genes, which are those sequences common to every isolate of the samples, whereas wgMLST in addition considers the accessory genes.

Recent efforts recommend wgMLST as the de facto epidemiological tool for source attribution (*Ribot et al., 2019*). However, this oversimplifies the efforts required and present complications of wgMLST.

One problem concerns standardization similar to the MLST schemes for both wgMLST and cgMLST. Bacteria that often recombine have a fast growing number of orthologs, which grows with the number of sequenced genomes. Resulting in a continuously growing wgMLST scheme necessitating continuous re-annotation (*Alikhan et al., 2018*). Hence requiring a high computational demand to create and work with a standardized, well working, reference frame. Routine surveillance with WGS would be required (*Franz et al., 2016*). Whereas, simply not updating the schemes will result in a quick outdating of said scheme. Therefore, there are arguments to be made for the use of the computationally less expensive cgMLST (*Alikhan et al., 2018*). cgMLST on the other hand, drops in resolution if not all scheme sequences are identified (*Lees et al., 2019*).

To overcome this problem associated with wgMLST and to a lesser extent cgMLST practices were introduced to identify the more important sequence and gene content differences between strains; e.g. based on partitioning of nucleotide k-mers into clusters (*Lees et al., 2019*), through means of Bayesian analysis to form clusters (*Cheng et al., 2013*), or with STRUCTURE discussed later in this review (*Pritchard et al., 2000*). Many of these approaches to selecting loci however lack scalability to whole genome levels (*Sheppard et al., 2012 & Lees et al., 2019*). Also, working with alleles rather than nucleotide sequences can somewhat alleviate this problem (*Sheppard et al., 2012*). However,

this may cloud the evolutionary dynamics involved in the process of differentiation, whereas nucleotide usage shows similarities between alleles (*Alikhan et al., 2018*).

The incorporation of more loci into the MLST scheme may in the long run also facilitate focus on differences in the likelihood and speed of nucleotide changes throughout genomes and their differential importance in regards to source attribution.

2.3. Source attribution tools

2.3.1. Phylogenetic trees

Phylogenetic trees are often applied in surveillance (*Henri et al., 2017 & Hudson et al., 2021*) and outbreak investigation research (*Chiaverini et al., 2021 & Gymoese et al., 2017*). Phylogenetic trees are constructed from genotyping data, from which between strains proximities are calculated. Thereafter, clustering is performed based on a cut-off of a certain number of nodes from the root node (*Mughini-gras et al., 2019*). There are different metrics to assess similarity between clusters. The downside of using phylogenetic trees for source attribution is that recombination events make a phylogeny approach more difficult for non-clonal bacterial species, since they may lengthen branch length or eviscerate recognizable ancestry. Whilst, this may not necessarily be troublesome for outbreak analysis if the time frame is short, i.e. for pathogens with fairly small generation time (Didelot et al., 2014), on longer timescales phylogenetic trees become unreliable (*Hall et al., 2015*). Here generation time refers to the time between the host becoming infected and causing subsequent infections.

More recent models attempted to overcome this problem by incorporating the mutational process and epidemiological information (*Hall et al., 2015 & Ypma et al., 2013*), through for example a molecular clock analysis (*Didelot et al., 2014*). However, attaching a mutational process to the phylogeny requires certain considerations. Firstly, multiple lineages of bacteria undergoing multiple mutation events should be allowed within a given host (*Ypma et al., 2013*). Secondly, branching times in phylogeny and transmission events do not necessarily coincide (*Didelot et al., 2014*). Another point needs to be considered as well; phylogenetic and transmission trees are separate entities and should be treated as such. This means that phylogenetic trees cannot reveal the sequential transmission events (*Didelot et al., 2014*). Recently attempts were made to reconcile the transmission and phylogenetic tree (*Ypma et al., 2013 & Hall et al., 2015*). Since this review is interested in the end nodes of transmission and not necessarily the route, transmission trees are outside the scope of this review. However, methodologies considering the route of transmission, inherently, also perform source attribution and therefore these methods are discussed shortly below.

Ypma et al., 2013 attempted to integrate the phylogenetic tree with the transmission tree, by linking up individual phylogenies within the host in accordance to the transmission tree. Nevertheless, this created a two step process. As a result, certain assumptions are made concerning the population structure of host or pathogens. For example, pathogens are limited to a particular host thus forgoing free mixing of the isolates. Since the research was constrained by a fixed prior phylogeny, *Hall et al., 2015* attempted to perform both phylogeny inference and node labeling simultaneously. As such, *Hall et al., 2015* adopted a Bayesian Markov Chain Monte Carlo (MCMC) approach to perform these processes concurrently, since the probability spaces calculated are highly dimensional and consequently mathematically complicated. Improvements can still be made; ideally each infected host should be seen as an individual entity, rather than a member of a group of hosts between which free mixing can occur (*Hall et al., 2015*).

2.3.2. Frequency-matching models

Source attribution can be performed using frequency-matching models. These models probabilistically infer the likeliest source for each isolate through comparison of weighted subtype frequencies. The weights applied may include prevalence in sources and exposure to those sources (*Mughini-gras et al., 2019*). Frequency-matching models output a number or percentage of cases attributable to each source together with a confidence interval or, in the case of a Bayesian model, a credibility interval (*Mughini-gras et al., 2019*). Two of the most prevalent frequency-matching models are the Hald model (*Hald et al., 2004*), the Dutch model (van Pelt et al., 1999) and the modified versions of these (Mullner et al., 2009b & Mughini-gras et al., 2014).

Dutch equation:

λ_{ij}: expected number of cases of subtype (i) from a source (j)

The Dutch model provides source attribution through calculation of the number of cases of a subtype (i) from a certain source (j) proportional to the total number of observed cases from said subtype (o_i). This proportion is then multiplied by the probability that the subtype comes from the source (*van Pelt et al., 1999*). The model does, however, not consider the capacity of subtypes and sources to cause infection (*Mughini-gras et al., 2019*). For foodborne diseases the assumption that each source has an equal impact is lifted through the introduction of weights to food consumption and probabilities of consumption of raw or undercooked food by the population (*Mughini-gras et al., 2014*). Additionally, the inclusion of subtype-specific prevalences mitigates the equal impact of different subtypes (*Mughini-gras et al., 2018*). The Hald model, on the other hand, is mainly designed for source attribution of *Salmonella* in a foodborne context. The model applies the same data type to a Bayesian framework using MCMC simulations, deciding stochastically the number of human cases belonging to a source (*Hald et al., 2004*). The Hald model's largest drawback is the number of parameters taken into consideration. Multiple adaptations attempt to reduce the number of parameters. For example, by fixing parameters in multiple quantified *Salmonella enterica* serotypes (*David et al., 2013*). The modified Hald model enlarges the models applicability by allowing non-food sources and pathogens besides *Salmonella*, as well as dealing with data uncertainty (*Mullner et al., 2009*).

The biggest shortcoming of both the Hald and Dutch model is the inability to attribute all subtypes found to a source, which consequently means that if a subtype is only present in the human samples will not be attributed to a source. This results in a fraction of 'unknown source'. Additionally, attribution of a subtype to a source does not necessarily mean it is the only source they are present in. The other, related source, may be absent (*Mughini-gras et al., 2019*). *David et al., 2013b* included sources though of lesser importance, which led to reassignment of a quarter of the cases. This highlights the importance for these models to include as many sources as possible.

2.3.3. Bayesian models based on genotyping data

AIM

The asymmetrical island model (AIM) is a Bayesian model of source attribution, which attempts to model DNA sequence evolution of bacteria by estimating horizontal gene transfer (HGT), recombination and de novo mutation. Furthermore, the model tries to explain zoonotic transmission (*Wilson et al., 2008*). To do so, the model considers the different sources as singular islands, between which migration might occur after a certain number of generations, representing gene flow. Given the allelic frequencies of loci in the source populations, estimates of mutation, recombination and migration rates can be attempted (*Mughini-gras et al., 2019*). Afterwards, human cases can be assigned to a single source population on a probabilistic basis (*Wilson et al., 2008*).

The major benefit of this model is the consideration of genetic relationships among isolates. By considering mutations, recombination and HGT, AIM takes advantage of the highly discriminate subtyping methods. The high discriminatory power of these subtyping methods can in an overlap comparison between subtypes merely obscure the actual source of a strain (*Mullner et al., 2009 & Barco et al., 2015*). Similarly, by considering the genetic relationship among isolates the model may be better able to predict the sources of highly recombinant strains, such as *C. jejuni*. However, if a strain was never before found in any source it is attributed to the source with the highest recombination rate. Another shortcoming of AIM is AIM's attempt to attribute all strains to a source, even if the actual source of the strain was not included in the model. Lastly, AIM assumes humans to be the recipient of the strain, which creates directionality. The direction of this directionality may be wrong (*Mäesaar et al., 2020*).

STRUCTURE

STRUCTURE uses Bayesian inference to determine allelic frequencies of populations, representing sources, at different loci. Isolates are then assigned to populations probabilistically, reflecting uncertainty (*Pritchard et al., 2000*). This is based on the allele frequencies at different loci and the allelic profiles. Admixing is possible, if specified, allowing assignment to multiple populations.

The main drawbacks of STRUCTURE are related to the optimal number of populations, thus equaling problems in defining optimal source separation. Like AIM, if a source is absent STRUCTURE may wrongly assign an isolate to another source.

2.3.4. Network approaches

The relation between the isolates derived from humans and the sources can be perceived as a bipartite system, where the connection is represented by, for example, the genetic relatedness as a proxy for the weight of the link. Therefore, a network based approach to source attribution is a distinct possibility. Such a network by evaluating the clustering coherence of reservoirs and driving forces of parameters in cluster formation (such as metadata on reservoir, country of origin and year) should give an overview of subtype and source associations. A study by *Merlotti et al., 2020* constructed a weighted network using pairwise SNP, cgMLST and wgMLST matrices for source attribution of *S. enterica* Typhimurium. Herein, nodes represent isolates and links between nodes as pairwise distances calculated as the number of different nucleotides or number of different alleles between two isolates. This builds on the underlying assumption that if two genomes are derived from the same source this should show in a smaller distance value.

The study found that after animal sources, the geographical region was a major driver of source attribution, which again highlights the importance of geographical segregation in source attribution. However, additional parameters can also be applied.

A benefit of a network approach is that isolates from unknown origins can be differentiated between and assigned to various putative unknown sources. However, on the other hand a worry with a network approach for source attribution is the identification of too many clusters. The high discriminatory power of genomic subtyping can lead to formation of too many clusters with a large fraction of isolates being unattributable. Though this fear seemed unfounded in *Merlotti et al., 2020*'s study, pathogens with higher interspecies genetic diversity, or without clear metadata and unclear source separation to separate clusters for benchmarking, may complicate analysis.

Nevertheless, source attribution through a network approach is a relatively unexplored avenue of research. Besides *Merlotti et al., 2020* further research needs to be applied to evaluate network approaches to perform source attribution.

2.3.5. Unsupervised machine learning

Unsupervised machine learning (ML), in contrast to supervised ML, does not use labels for the isolates used in the training. Common examples of unsupervised ML include various clustering techniques and principal component analysis. Though principal component analysis belongs to the dimension reduction techniques, both clustering and dimension reduction attempt to group the data. The number of features are thereby compressed during dimension reduction, whilst in clustering they are not (*Lupolova et al., 2019*).

Unsupervised ML does not per se require labels for the isolates used during training. However, patterns within the data may be difficult to associate with a source cluster. Additional metadata helps determine the capability of the algorithm to assign isolates to the sources. This is to say you determine the population structure a priori. Similarly spatio-temporal data may be used. However, when lacking metadata altogether there are methods for approximating the optimal number of clusters either beforehand (*Charrad et al., 2014*) or after the fact (*Arbelaitz et al., 2013*). Another benefit to these clustering techniques is, similarly to network approaches, the direct visualization of clusters and by extent cluster separation, if labels are provided.

Lupolova et al., 2019 looked into the use of a variety of ML algorithms for source attribution of *S. enterica* Typhimurium, including unsupervised ML. According to the study, assigning the number of clusters prior based on metadata produced better results. However, when four clusters based on the number of hosts were decided (avian, bovine, human and swine); the underlying structure of the clustering was not entirely related to the host. After varying the number of clusters *Lupolova et al., 2019* inferred that clusters of mixed host origins contain generalist strains, since gene content differences should be less. Additionally, the study found a group of isolates switching clusters between runs. These singular isolates jumping clusters between runs may indicate a missing source or rare strains captured only once in the sampling. All by all, complete separation of host populations was not achieved.

2.3.6. Supervised machine learning

The main difference between supervised and unsupervised machine learning is the designation of classes. The model is trained with training data of already characterized data whose source is known. Since the model will know whether correct predictions were made based on the labels of the samples, the model is able to learn. The benefit of supervised machine learning techniques is that they can learn to recognize various patterns in the data, thus leading to a clearer separation of sources. Using supervised ML, *Lupolova et al., 2019* was able to classify most samples to their sources, which was not achieved with unsupervised ML. Source attribution through supervised machine learning can be done through a multitude of machine learning algorithms, e.g.; random forest (RF) (*Munck et al., 2020 & Wheeler et al., 2018*), neural networks (NN) (*Lupolova et al., 2019*) or support vector machines (SVM) (*Lupolova et al., 2017 & 2019, Arning et al., 2021*).

There are apparent limitations to supervised ML though. Firstly, characterized data must exist beforehand for training of the model. Secondly, the number of samples needed for training can be substantial. Thirdly, an ever-present worry in supervised ML is overfitting. Overfitting refers to the too precise alignment of cut-offs to the data points within the training set. Different metrics exist, depending on the balance of the dataset, for evaluation of the performance of such an algorithm and to prevent overfitting.

When supervised ML is applied correctly, it provides a powerful tool to uncover patterns of genes/mutations associated with a particular source. For example, RF allows for viewing of the most important features. Other techniques, such as SVM and NN do not readily reveal important features. However, the biological relevance of various entangled, possibly polygenic, traits may be rather vague. Furthermore, supervised ML allows a probabilistic classification, providing further information on potentially difficult to classify strains, such as those with broad host ranges. In other words it has the potential to predict zoonotic potential (*Lupolova et al., 2017*). The dimensional space may be substantial, however feature selection can be done prior with for example dimension reduction (*Lupolova et al., 2019*), dissimilarity analysis (*Duarte et al., 2021*), the boruta function (*Munck et al., 2020*), and random forest allows viewing of the most important features for selection making.

Discussion and conclusion

There are many ways to approach source attribution. Here we discussed the various epidemiological, Phenotyping, genotyping, and techniques based on genotyping and/or sequencing data, such as phylogenetic trees, frequency-matching models, network approaches, and machine learning. Furthermore, there exist various other miscellaneous source attribution approaches not further elaborated upon, and combinations of methods can be applied. Each of the approaches have their benefits and drawbacks pertaining to the associated biological problems, such as the nature and behavior of the pathogens and sources. This again brings us to our main question. To rehash; what methods or combinations of methods are best able to overcome the problems and limitations inherent in source attribution to perform accurate and precise attribution to the correct sources?

To determine the most relevant method the biological problems associated with the pathogen strains are to be considered first. The most interesting pathogens in relation to source attribution are those with broad host ranges, since those lacking these will not often need re-evaluation after the initial determination of the source. Simultaneously, pathogens with broad host ranges are more problematic to attribute to a specific source, due the potential lack of specific phylogenetic associations. Additionally, host switching can cause errors in the metadata annotation and erases genetic signatures (*Dearlove et al., 2015*). On top of this, different strains of a pathogen species may behave differently. Therefore, a technique's ability to generalize to different pathogen strains is key.

Classically phenotyping methods were the method of choice. However, in many cases they do not provide the same resolution as genotyping techniques, except for very specific cases, such as differences in the number of lysis reactions of certain sub-strains (*Garman and Adley., 2004*). Even so, higher levels of resolution may give light to finer patterns, like specific SNPs. The development especially in genotyping, specifically wgMLST and cgMLST, may make this possible. Additionally, WGS data allows several different genetic signatures to be used concurrently, like tandem repeats and point mutations. Consequently, genotyping techniques are generally preferred.

Whole-genome sequencing has the capability to overcome problems associated with the lack of genomic signatures in source attribution, but also raises new questions. Such as; what are the commonalities and differences between pathogens adapting to a single host and those adapting broad host ranges, such as patterns of gene degradation (*Wheeler et al., 2018*)? And more fundamentally; to what extent are genetic signals associated with adaptations? The last question feeds back into the need to standardize and create a continuously updated reference frame for wgMLST, where growing number of samples necessitate removal and re-annotation. Temporal differences between strain pathogens may relate to either, but research needs to show how genetic changes accommodate host changes over time.

The lack of spatio-temporal data makes investigating their effects on source attribution challenging. Links between geographical distance between pathogen populations are established (*Lupolova et al., 2019*). *Griekspoor et al., 2013* saw less of a geographical pattern in birds, but this may relate to ranging behavior of birds compared to confined broiler chickens or cattle. Furthermore, as opposed to free mixing populations, pathogens within a host cannot readily interact with pathogens in another host of the same host species (*Hall et al., 2015*). The spatial effect may thus be linked to both exposure to hosts and transmissibility of the pathogen. Concerning the temporal aspect a

tangible link was made between genomic time-scale dissimilarities and the pathogen's genome plasticity. Even though temporal data will benefit source attribution allowing evaluation of contributions over time and impacts of interventions (*Mughini-gras et al., 2019*), a base level understanding of the effect of pathogen behavior through time on source attribution is required. Figuring out the 'sweet spot' of time past in between subsequent data acquisitions for which direct ancestry of different serovars can be determined is necessary. For each pathogen these questions should be handled before spatio-temporal data is fully incorporated.

Despite the availability of multitudes of sequencing data these are often poorly annotated. Additionally, not every study will have the capital to perform a plurality of sequencing sessions for an indiscriminate amount of potential sources. Therefore, in such cases this needs to be considered when choosing an attribution method. For example, AIM performs as well as the Hald or Dutch model, but requires a larger amount of samples. Moreover, introduction of a standardized way of annotating metadata would allow easier re-use of data and separation of clusters in clustering approaches based on additional information.

We need to acknowledge that within this review we were not able to evaluate all pathogen and source characteristics. Certain pathogen behaviors were left unconsidered, like pathogen interactions with one another, human-to-human transmission, and the effects of niche occupation. Neither was the related topic of attributing anti-microbial resistance to a putative source taken into consideration, or were biases implicit in source representation imbalances discussed.

The method preferable for analyzing genotyping data may not always be apparent. Besides the quality and completeness of the data, the characteristics of the pathogen and designated sources in the epidemiological context play a major role in the suitability of a technique. The various approaches all have their upsides and downsides. However, many of these techniques can be used in consortium with one another and various studies have compared the performance of different techniques with one another (*Lupolova et al., 2019 & Arning et al., 2021*). Clustering approaches for instances are potentially hindered by high levels of resolution, causing the creation of too many clusters. Phylogenetic trees may on their own not be the most suitable approach, since non-clonal species eviscerate recognizable ancestry. Moreover, the phylogenetic trees cannot fully omit the transmission tree. ML, and especially supervised ML may offer valuable insights into patterns of information within the data, which to the naked eye are hidden. These features may then be used by other approaches to improve source attribution. Frequency-matching models may offer an intuitive approach to source attribution, but like supervised ML they have a predefined number of sources, which causes them to erroneously attribute all strains to a source. Since the inclusion of missing sources can cause a large shift in source attribution (*David et al., 2013b*), clustering, phylogenetic or network approaches should be used as well. Network approaches in particular are a relatively unexplored avenue of research in the field of source attribution and should be evaluated in terms of usefulness. More generally, case-control studies can be combined with microbial subtyping approaches to infer sources and transmission pathways.

We hypothesized that all by all, no standard approach to all source attribution tasks exists, and there probably will not be a general fit-all approach since pathogen behavior differs. As such, decisions on a case-by-case basis need to be made considering behavior and available data. For each differently behaving bacteria a standardized protocol needs to be developed. This protocol likely consists of an

array of different WGS genotyping approaches used in conjunction with one another, in order to approximate their ability to perform source attribution. For example, for highly recombinant strains phylogenetic trees may not offer the best solution. Increasingly the existing gaps in knowledge are bridged allowing more sophisticated approaches to be taken. The use of spatio-temporal information for instance allows temporal and geographical trends to be considered. Despite this, much research still needs to be performed in regards to evaluating and optimizing decision making for different pathogens and for considering sources.

References

Alikhan N-F., Zhou Z., Sergeant M.J., Achtman M. A genomic overview of the population structure of Salmonella. PLoS Genetics. 2018. 14(4), e1007261. https://dx.doi.org/10.1371/journal.pgen.1007261

Allos B.M. *Campylobacter* jejuni infections: update on emerging issues and trends. Clinical Infectious Diseases. 2001. 32(8), 1201-1206. https://doi.org/10.1086/319760

Arbelaitz O., Gurrutxaga I., Muguerza J., Pérez J.M., Perona I. An extensive comparative study of cluster validity indices. Pattern Recognition. 2013. 46(1), 243-256. https://doi.org/10.1016/j.patcog.2012.07.021

Arning N., Sheppard S.K., Bayliss S., Clifton D.A., Wilson D.J. Machine learning to predict the source of campylobacteriosis using whole genome data. PLoS Genetics. 2021. 17(10), e1009436. https://doi.org/10.1371/journal.pgen.1009436

Barco L., Barrucci F., Olsen J.E., Ricci A. *Salmonella* source attribution based on microbial subtyping. International Journal of Food Microbiology. 2013. 163(2-3), 193-203. https://doi.org/10.1016/j.ijfoodmicro.2013.03.05

Barco L., Barrucci F., Cortini E., Ramon E., Olsen J.E., Luzzi I., Lettini A.A., Ricci A. Ascertaining the relationship between *Salmonella* Typhimurium and *Salmonella* 4,[5],12:i:- by MLVA and inferring the sources of human salmonellosis due to the two serovars in Italy. Frontiers in Microbiology. 2015. 6(301). https://doi.org/10.3389/fmicb.2015.00301

Van Belkum A., Tassios P.T., Dijkshoorn L., Haeggman S., Cookson B., Fry N.K., Fussing V., Green J., Feil E., Gerner-Smidt P., Brisse S., Streulens M. Guidelines for the validation and application of typing methods for use in bacterial epidemiology. Clinical Microbiology and Infection. 2007. 13(3), 1-46. https://doi.org/10.1111/j.1469-0691.2007.01786.x

Charrad M., Ghazzali N., Boiteau V., Niknafs A. NbClust: An R package for determining the relevant number of clusters in a data set. Journal of Statistical Software. 2014. 61(6). https://doi.org/10.18637/jss.v061.i06

Cheng L., Conner T.R., Sirén J., Aanensen D.M., Corander J. Hierarchical and spatially explicity clustering of DNA sequences with BAPS software. Molecular Biology and Evolution. 2013. 30(5), 1224-1228. https://doi.org/10.1093/molbev/mst028

Chiaverini A., Guidi F., Torresi M., Acciari V.A., Centorotola G., Cornacchia A., Centorame P., Marfoglia C., Blasi G., Domenico M.D., Migliorati G., Roussel S., Pomilio F., Sevellec Y. Phylogenetic analysis and genome-wide association study applied to an Italian *Listeria monocytogenes* outbreak. Frontiers in Microbiology. 2021. 12, 750065. https://doi.org/10.3389/fmicb.2021.750065

David J.M., Guillemot D., Bemrah N. Thébault A., Brisabois A., Chemaly M., Weill F-X., Sanders P., Watier L. The Bayesian microbial subtyping attribution model: robustness to prior information and a proposition. Risk Analysis. 2013. 33(3), 397-408. https://doi.org/10.1111/j.1539-6924.2012.01877.x

David J.M., Sanders P., Bemrah N., Granier S.A., Denis M., Weill F-X., Guillemot D., Watier L. Attribution of the French human salmonellosis cases to the main food-sources according to the type of surveillance data. Preventive Veterinary Medicine. 2013b. 110(1), 12-27. https://doi.org/10.1016/j.prevetmed.2013.02.002

Dearlove B., Cody A.J., Pascoe B., Méric G., Wilson D.J., Sheppard S.K. Rapid host switching in generalist Campylobacter strains erodes the signal for tracing human infection. International Society for Microbial Ecology. 2016. 10(3), 721-729. doi: 10.1038/ismej.2015.149

Deutsche Presse-Agentur, Agence France-Presse. Listeria-tainted sausage deaths in Germany lead to calls for better consumer protection. Deutsche Welle. 2019. https://p.dw.com/p/3QmIZ

Didelot X., Gardy J., Colijn C. Bayesian inference of infectious disease transmission from whole-genome sequence data. Molecular Biology and Evolution. 2014. 31(7), 1869-1879. https://doi.org/10.1093/molbev/msu121

Duarte A.S.R., Röder T., van Gompel L., Petersen T.N., Hansen R.B., Hansen I.M., Bossers A., Aarestrup F.M., Wagenaar J.A., Hald T. Metagenomics-based approach to source-attribution of antimicrobial resistance determinants – identification of reservoir resistome signatures. Frontiers in Microbiology. 2021. 11, 601407. https://doi.org/10.3389/fmicb.2020.601407

EFSA. Scientific opinion on the evaluation of molecular typing methods for major foodborne microbiological hazards and their use for attribution modeling, outbreak investigation and scanning surveillance: part 1. EFSA Journal. 2013. 11(12), 3502. https://doi.org/10.2903/j.efsa.2013.3502

Excoffier L., Smouse P.E., Quattro J.M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics. 1992. 131(2), 479-491. https://doi.org/10.1093/genetics/131.2.479

Ferrari R.G., Panzenhagen P.H.N., Conte-Junior C.A. Phenotypic and genotypic eligible methods for *Salmonella Typhimurium* source tracking. Frontiers in Microbiology. 8, 2587. https://doi.org/10.3389/fmicb.2017.02587

Fitzgerald J.R. Livestock-associated *Staphylococcus* aureus: origin, evolution and public health threat. Trends in Microbiology. 2012. 20(4), 192-198. https://doi.org/10.1016/j.tim.2012.01.006

Foley S.L., White D.G., McDermott P.F., Walker R.D., Rhodes B., Fedorka-Cray P.J., Simjee S., Zhao S. Comparison of subtyping methods for differentiating *Salmonella enterica* serovar Typhimurium

isolates obtained from food animal sources. Journal of Clinical Microbiology. 2006. 44(10), 3569-3577. https://doi.org/10.1128/jcm.00745-06

Franz E., Mughini-Gras L., Dallman T. Significance of whole genome sequencing for surveillance, source attribution and microbial risk assessment of foodborne pathogens. Current Opinion in Food Science. 2016. 8, 74-79. http://dx.doi.org/10.1016/j.cofs.2016.04.004

Fullerton K.E., Scallan E., Kirk M.D., Mahon B.E., Angulo F.J., de Valk H., van Pelt W., Gauci C., Hauri A.M., Majowicz S., O'Brien S.J. Case-control studies of sporadic enteric infections: a review and discussion of studies conducted internationally from 1990 to 2009. Foodborne Pathogens and Disease. 2012. 9(4), 281-292. https://dx.doi.org/10.1089/fpd.2011.1065

Gorman R., Adley C.C. Characterization of *Salmonella enterica* serotype Typhimurium isolates from human, food, and animal sources in the republic of Ireland. Journal of Clinical Microbiology. 2004. 42(5), 2314-2316. https://doi.org/10.1128/JCM.42.5.2314-2316.2004

Griekspoor P., Colles F.M., McCarthy N.D., Hansbro P.M., Ashhurst-Smith C., Olsen B., Hasselquist D., Maiden M.C.J., Waldenström J. Marked host specificity and lack of phylogeographic population structure of *Campylobacter jejuni* in wild birds. Molecular Ecology. 2013. 22, 1463-1472. https://doi-org.proxy.library.uu.nl/10.1111/mec.12144

Grimont F., Grimont P.A., Ribosomal ribonucleic acid gene restriction patterns as potential taxonomic tools. Annales de l'Institut Pasteur. Microbiology. 1986. 137B(2), 165-175. https://doi.org/10.1016/s0769-2609(86)80105-3

Grimont P.A.D., Weill F.X. Antigenic formulae of *Salmonella* serovars. WHO Collaborating Centre for Reference and Research on Salmonella. 2007. http://www.pasteur.fr/ip/portal/action/WebdriveActionEvent/oid/01s-000036-089

Gripp E., Hlahla D., Didelot X., Kops F., Maurischat S., Tedin K., Alter T., Ellerbroek L., Schreiber K., Schomburg D., Janssen T., Bartholomäus P., Hofreuter D., Woltemate S., Uhr M., Brenneke B., Grüning P., Gerlach G., Wieler L., Suerbaum S., Josenhans C. Closely related Campylobacter jejuni strains from different sources reveal a generalist rather than a specialist lifestyle. BMC Genomics. 2011. 12, 584. https://doi.org/10.1186/1471-2164-12-584

Gymoese P., Sørensen G., Litrup E., Olsen J.E., Nielsen E.M., Torpdahl M. Investigation of outbreaks of *Salmonella enterica* serovar Typhimurium and its monophasic variants using whole-genome sequencing, Denmark. Emerging Infectious Diseases. 2017. 23(10), 1631-1639. https://dx.doi.org/10.3201/eid2310.161248

Hald T., Vose D., Wegener H.C., Koupeev T. A Bayesian approach to quantify the contribution of animal-food sources to human salmonellosis. 2004. Risk Analysis. 24(1), 255-269. https://doi.org/10.1111/j.0272-4332.2004.00427.x

Hall M., Woolhouse M., Rambaut A. Epidemic reconstruction in a phylogenetics framework: transmission trees as partitions of the node set. PLoS Computational Biology. 2015. 11(12), e1004613. https://doi.org/10.1371/journal.pcbi.1004613

Henri C., Leekitcharoenphon P., Carleton H.A., Radomski N., Kaas R.S., Mariet J-F., Felten A., Aarestrup F.M., Smidt P.G., Roussel S., Guillier L., Mistou M-Y., Hendriksen R.S. An assessment of different genomic approaches for inferring phylogeny of *Listeria monocytogenes*. Frontiers in Microbiology. 2017. 8, 2351. https://doi.org/10.3389/fmicb.2017.02351

Hoffmann S., Maculloch B., Batz M. Economic burden of major foodborne illnesses acquired in the United States. 2015. https://www.ers.usda.gov/webdocs/publications/43984/52807_eib140.pdf?v=5219.4

Hopkins K.L., Maguire C., Best E., Liebana E., Threlfall E.J. Stability of multiple-locus variable-number tandem repeats in *Salmonella enterica* serovar Typhimurium. Journal of Clinical Microbiology. 2007. 45(9), 3058-3061. https://doi.org/10.1128/JCM.00715-07

Hopkins K.L., Peters T.M., de Pinna E., Wain J. Standardisation of multilocus variable-number tandem-repeat analysis (MLVA) for subtyping of *Salmonella enterica* serovar Enteritidis. Euro Surveillance. 2011. 16(32), 19942. https://doi.org/10.2807/ese.16.32.19942-en

Hudson L.K., Andershock W.E., Yan R., Golwalkar M., M'ikanatha N.M., Nachamkin I., Thomas L.S., Moore C., Qian X., Steece R., Garman K.N., Dunn J.R., Kovac J., Denes T.G. Phylogenetic analysis reveals source attribution of patterns for *Campylobacter* spp. In Tennessee and Pennsylvania. Miroorganisms. 2021. 9(11), 2300. https://doi.org/10.3390/microorganisms9112300

Lees J.A., Harris S.R., Tonkin-Hill G., Gladstone R.A., Lo S.W., Weiser J.N., Corander J., Bentley S.D., Croucher N.J. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Research. 2019. 29(2), 304-316. https://dx.doi.org/10.1101/gr.241455.118

Lindstedt B-A., Heir E., Gjernes E., Kapperud G. DNA fingerprinting of *Salmonella enterica* subsp. enterica serovar typhimurium with emphasis on phage type DT104 based on variable number of tandem repeat loci. Journal of Clinical Microbiology. 2003. 41(4), 1469-1479. https://doi.org/10.1128/jcm.41.4.1469-1479.2003

Litrup E., Christensen H., Nordentoft S., Nielsen E.M., Davies R.H., Helmuth R., Bisgaard M. Use of multiple-locus variable-number tandem-repeats analysis (MLVA) typing to characterize *Salmonella* Typhimurium DT41 broiler breeder infections. Journal of Applied Microbiology. 2010. 109(6), 2032-2038. https://doi.org/10.1111/j.1365-2672.2010.04833.x

Lupolova N., Dallman T.J., Holden N.J., Gally D.L. Patchy promiscuity: machine learning applied to predict the host specificity of *Salmonella enterica* and *Escherichia coli*. Microbial Genomics. 2017. 3(10), e000135. https://doi.org/10.1099/mgen.0.000135

Lupolova N., Lycett S.J., Gally D.L. A guide to machine learning for bacterial host attribution using genome sequence data. Microbial Genomics. 2019. 5(12), e000317. https://doi.org/10.1099/mgen.0.000317

Mäesaar M., Tedersoo T., Meremäe K., Roasto M. The source attribution analysis revealed the prevalent role of poultry over cattle and wild birds in human campylobacteriosis cases in the Baltic States. PLoS ONE. 2020. 15(7), e0235841. https://doi.org/10.1371/journal.pone.0235841

Maiden M.C., Bygraves J.A., Feil E., Morelli G., Russell J.E., Urwin R., Zhang Q., Zhou J., Zurth K., Caugant D.A., Feavers I.M., Achtman M., Spratt B.G. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proceedings of the National Academy of Sciences of the United States of America. 1998. 95(6), 3140-3145. https://doi.org/10.1073/pnas.95.6.3140

Meinersmann R.J., Helsel L.O., Fields P.I., Hiett K.L. Discrimination of *Campylobacter jejuni* isolates by fla gene sequencing. Journal of Clinical Microbiology. 1997. 35(11), 2810-2814. https://doi.org/10.1128/jcm.35.11.2810-2814.1997

Merlotti A., Manfreda G., Munck N., Hald T., Litrup E. Nielsen E.M., Remondini D., Pasquali F. Network approach to source attribution of *Salmonella enterica* serovar Typhimurium and its monophasic variant. Frontiers in Microbiology. 2020. 11, 1205. https://doi.org/10.3389/fmicb.2020.01205

Mughini-Gras L., Barrucci F., Smid J.H., Graziani C., Luzzi I., Ricci A., Barco L., Rosmini R., Havellar A.H., van Pelt W., Busani L. Attribution of human *Salmonella* infections to animal and food sources in Italy (2002-2010): adaptations of the Dutch and modified Hald source attribution models. Epidemiology and Infection. 2014. 142(5), 1070-1082. https://doi.org/10.1017/s0950268813001829

Mughini-Gras L., Franz E., van Pelt W. New paradigms for *Salmonella* source attribution based on microbial subtyping. Food Microbiology. 2018. 71, 60-67. https://doi.org/10.1016/j.fm.2017.03.002

Mughini-Gras L., Kooh P., Fravalo P., Augustin J-C., Guillier L., David J., Thébault A., Carlin F., Leclercq A., Jourdan-Da-Silva N., Pavio N., Villena I., Sanaa M., Watier L. Critical orientation in the jungle of currently available methods and types of data for source attribution of foodborne diseases. Frontiers in Microbiology. 2019. 10, 2578. https://doi.org/10.3389/fmicb.2019.02578

Mullner P., Spencer S.E.F., Wilson D.J., Jones G., Noble A.D., Midwinter A.C. Assigning the source of human campylobacteriosis in New Zealand: a comparative genetic and epidemiological approach. Infection, Genetics and Evolution. 2009. 9(6), 1311-1319. https://doi.org/10.1016/j.meegid.2009.09.003

Mullner P., Jones G., Noble A., Spencer S.E.F., Hathaway S., French N.P. Source attribution of food-borne zoonoses in New Zealand: a modified Hald model. Risk Analysis. 2009b. 29(7), 970-984. https://doi.org/10.1111/j.1539-6924.2009.01224.x

Munck N., Njage P.M.K., Leekitcharoenphon P., Litrup E., Hald T. Application of whole-genome sequencing and machine learning in source attribution of *Salmonella* Typhimurium. Risk Analysis. 2020. 40, 1693-1705. https://doi-org.proxy.library.uu.nl/10.1111/risa.13510

Octavia S., Wang Q., Tanaka M.M., Sintchenko V., Lan R. Genomic variability of serial human isolates of *Salmonella enterica* serovar Typhimurium associated with prolonged carriage. Journal of Clinical Microbiology. 2015. 53(11), 3507-3514. https://doi.org/10.1128/JCM.01733-15

Octavia S., Wang Q., Tanaka M.M., Kaur S., Sintchenko V., Lan R. Delineating community outbreaks of *Salmonella enterica* serovar Typhimurium by use of whole-genome sequencing: insights into genomic

variability within an outbreak. Journal of Clinical Microbiology. 2015. 53(4), 1063-1071. https://doi.org/10.1128/JCM.03235-14

Palma F., Manfreda G., Silva M., Parisi A., Barker D.O.R., Taboada E.N., Pasquali F., Rossi M. Genome-wide identification of geographical segregated genetic markers in *Salmonella enterica* serovar Typhimurium variant 4,[5],12;i;. Scientific Reports. 2018. 8(1), 15251. https://doi.org/10.1038/s41598-018-33266-5

van Pelt W., van de Giessen A., van Leeuwen W., Wannet W., Henken A., Evers E.G. Oorsprong, omvang en kosten van humane salmonellose. Deel 1. Oorsprong van humane salmonellose met betrekking to varken, rund, kip, ei en overige bronnen. Infectieziekten Bulletin. 1999. 10, 240-243.

Pritchard J.K., Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000. 155(2), 945-959. https://doi.org/10.1093/genetics/155.2.945

Ranieri L.M., Shi C., Switt A.I.M., den Bakker H.C., Wiedmann M. Comparison of typing methods with a new procedure based on sequence characterization for *Salmonella* serovar prediction. Journal of Clinical Microbiology. 2013. 51(6), 1786-1797. https://dx.doi.org/10.1128/JCM.03201-12

Ribot E.M., Freeman M., Hise K.B., Gerner-Smidt P. PulseNet: entering the age of next-generation sequencing. Foodborne Pathogens and Disease. 2019. 16(7), 451-456. https://dx.doi.org/10.1089/fpd.2019.2634

De Roeck A., De Coster W., Bossaerts L., Cacace R., De Pooter T., Van Dongen J., D'Hert S., De Rijk P., Strazisar M., Van Broeckhoven C., Sleegers K. NanoSatellite: accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION. Genome Biology. 2019. 20(239). https://doi.org/10.1186/s13059-019-1856-3

Sabat A.J., Budimir A., Nashev D., Sá-Leão R., van Dijl J.M., Laurent F., Grundmann H., Friedrich A.W., ESCMID Study Group of Epidemiological Markers. Overview of molecular typing methods for outbreak detection and epidemiological surveillance. Euro Surveillance. 2013. 18(4), 20380. https://doi.org/10.2807/ese.18.04.20380-en

Schjørring S., Niskanen T., Torpdahl M., Björkman J.T., Nielsen E.M. Evaluation of molecular typing of foodborne pathogens in European reference laboratories from 2012 to 2013. Euro surveillance. 2016. 21(50), 30429. https://doi.org/10.2807/1560-7917.ES.2016.21.50.30429

Sheppard S.K., Jolley K.A., Maiden M.C.J. A gene-by-gene approach to bacterial population genomics: whole genome MLST of Campylobacter. Genes. 2012. 3(2), 261-277. https://doi.org/10.3390/genes3020261

Smid J.H., Mughini-Gras L., de Boer A.G., French N.P., Havelaar A.H., Wagenaar J.A., van Pelt W. Practicalities of using non-local or non-recent multilocus sequence typing data for source attribution in space and time of human campylobacteriosis. PLoS ONE. 2013. 8(2), e55029. https://doi.org/10.1371/journal.pone.0055029

Swaminathan B., Barrett T.J., Hunter S.B., Tauxe R.V., CDC PulseNet Task Force. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. Emerging Infectious Diseases. 2001. 7(3), 382-389. https://doi.org/10.3201/eid0703.010303

Taboada E.N., Ross S.L., Mutschall S.K., Mackinnon J.M., Roberts M.J., Buchanan C.J., Kruczkiewicz P., Jokinen C.C., Thomas J.E., Nash J.H.E., Gannon V.P.J., Marshall B., Pollari F., Clark C.G. Development and validation of a comparative genomic fingerprinting method for high-resolution genotyping of *Campylobacter jejuni*. Journal of Clinical Microbiology. 2012. 50(3), 788-797. https://doi.org/10.1128/jcm.00669-11

Wagenaar J.A. Campylobacter: animal reservoirs, human infections, and options for control. Zoonoses-infections affecting humans and animals: focus on public health aspects. 2015. ISBN: 9789401794572, 9789401794565.

Wang B., Wang C., McKean J.D., Logue C.M., Gebreyes W.A., Tivendale K.A., O'Connor A.M. *Salmonella enterica* in swine production: assessing the association between amplified fragment length polymorphism and epidemiological units of concern. Applied and Environmental Microbiology. 2011. 77(22), 8080-8087. https://doi.org/10.1128/aem.00064-11

Wattiau P., Boland C., Bertrand S. Methodologies for *Salmonella enteric* subsp. *enterica* subtyping: gold standards and alternatives. Applied and Environmental Microbiology. 2011. 77(22), 7877-7885. https://dx.doi.org/10.1128/AEM.05527-11

Wheeler N.E., Gardner P.P., Barquist L. Machine learning identifies signatures of host adaptation in the bacterial pathogen Salmonella enteric. PLoS Genetics. 2018. 14(5), e1007333. https://doi.org/10.1371/journal.pgen.1007333

Wilson D.J., Gabriel E., Leatherbarrow A.J.H., Cheesbrough J., Gee S., Bolton E., Fox A., Fearnhead P., Hart C.A., Diggle P.J. Tracing the source of Camplyobacteriosis. PLoS Genetics. 2008. 4(9), e1000203. https://doi.org/10.1371/journal.pgen.1000203

Woodcock D.J., Krusche P., Strachan N.J.C., Forbes K.J., Cohan F.M., Méric G., Sheppard S.K. Genomic plasticity and rapid host switching can promote the evolution of generalism: a case study in the zoonotic pathogen *Campylobacter*. Scientific Report. 2017. 7, 9650. https://doi-org.proxy.library.uu.nl/10.1038/s41598-017-09483-9

Wuyts V., Mattheus W., de Bex, G.D.L., Wildemauwe C., Roosens N.H.C., Marchal K., De Keersmaecker S.C.J., Bertrand S. MLVA as a tool for public health surveillance of human *Salmonella* Typhimurium: prospective study in Belgium and evaluation of MLVA loci stability. PLoS ONE. 2013. 8(12), e84055. https://doi.org/10.1371/journal.pone.0084055

Ypma R.J.F., van Ballegooijen W.M., Wallinga J. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. Genetics. 2013. 195(3), 1055-1062. https://doi.org/10.1534/genetics.113.154856

Zhang S., Li S., Gu W., den Bakker H., Boxrud D., Taylor A., Roe C., Driebe E., Engelthaler D.M., Allard M., Brown E., McDermott P., Zhoa S., Bruce B.B., Trees E., Fields P.I., Deng X. Salmonella enteric

serotype Typhimurium using genomic surveillance data, United States. Emerging Infectious Diseases. 2019. 25(1), 82-91. https://doi.org/10.3201/eid2501.180835