



<http://www.diva-portal.org>

This is the published version of a paper published in *Forensic Science International*.

Citation for the original published paper (version of record):

Lindgren, P., Myrtenäs, K., Forsman, M., Johansson, A., Stenberg, P. et al. (2019)
A likelihood ratio-based approach for improved source attribution in microbiological
forensic investigations

Forensic Science International, 302: 109869

<https://doi.org/10.1016/j.forsciint.2019.06.027>

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-165758>



A likelihood ratio-based approach for improved source attribution in microbiological forensic investigations

Petter Lindgren^a, Kerstin Myrtennäs^a, Mats Forsman^a, Anders Johansson^b,
Per Stenberg^{a,c}, Anders Nordgaard^d, Jon Ahlinder^{a,*}

^a Department of Biological Agents, Division of CBRN Defence and Security, Swedish Defence Research Agency (FOI), SE-901 82 Umeå, Sweden

^b Department of Clinical Microbiology and Molecular Infection Medicine Sweden (MIMS), Umeå University, SE-901 87 Umeå, Sweden

^c Department of Ecology and Environmental Science (EMG), Umeå University, SE-901 87 Umeå, Sweden

^d Swedish Police, Swedish National Forensic Centre (NFC), SE-581 94 Linköping, Sweden

ARTICLE INFO

Article history:
Available online 2 July 2019

Keywords:
Microbial source tracking
Bayes factor
Hypothesis assessment
Listeria monocytogenes
Francisella tularensis
Likelihood ratio

ABSTRACT

A common objective in microbial forensic investigations is to identify the origin of a recovered pathogenic bacterium by DNA sequencing. However, there is currently no consensus about how degrees of belief in such origin hypotheses should be quantified, interpreted, and communicated to wider audiences. To fill this gap, we have developed a concept based on calculating probabilistic evidential values for microbial forensic hypotheses. The likelihood-ratio method underpinning this concept is widely used in other forensic fields, such as human DNA matching, where results are readily interpretable and have been successfully communicated in juridical hearings. The concept was applied to two case scenarios of interest in microbial forensics: (1) identifying source cultures among series of very similar cultures generated by parallel serial passage of the Tier 1 pathogen *Francisella tularensis*, and (2) finding the production facilities of strains isolated in a real disease outbreak caused by the human pathogen *Listeria monocytogenes*. Evidence values for the studied hypotheses were computed based on signatures derived from whole genome sequencing data, including deep-sequenced low-frequency variants and structural variants such as duplications and deletions acquired during serial passages. In the *F. tularensis* case study, we were able to correctly assign fictive evidence samples to the correct culture batches of origin on the basis of structural variant data. By setting up relevant hypotheses and using data on cultivated batch sources to define the reference populations under each hypothesis, evidential values could be calculated. The results show that extremely similar strains can be separated on the basis of amplified mutational patterns identified by high-throughput sequencing. In the *L. monocytogenes* scenario, analyses of whole genome sequence data conclusively assigned the clinical samples to specific sources of origin, and conclusions were formulated to facilitate communication of the findings. Taken together, these findings demonstrate the potential of using bacterial whole genome sequencing data, including data on both low frequency SNP signatures and structural variants, to calculate evidence values that facilitate interpretation and communication of the results. The concept could be applied in diverse scenarios, including both epidemiological and forensic source tracking of bacterial infectious disease outbreaks.

© 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

The main objective in microbial forensic science is to provide evidence that can be used in legal proceedings by characterizing a

pathogen responsible for an epidemic outbreak. This involves tracing the pathogen to its source, i.e. a production facility, a geographic area or a suspect person (e.g. [1–4]). The ease of get access to pathogenic bacteria, performing mass cultivation and spread it through food, water and air, for example, makes such alternative appealing in a bioterror perspective. As most pathogenic bacterial species, or close relatives, naturally exists in the local habitat, source tracking is difficult as the background can provide inflated rate of false positive detections. To draw reliable conclusions when comparing forensic and epidemiological samples [2], a unified statistical framework is needed, and a

* Corresponding author at: CBRN Defence and Security, Swedish Defence Research Agency (FOI), SE-901 82 Umeå, Sweden.

E-mail addresses: petter.lindgren@foi.se (P. Lindgren), kerstin.myrtennas@foi.se (K. Myrtennäs), mats.forsman@foi.se (M. Forsman), anders.f.johansson@umu.se (A. Johansson), per.stenberg@foi.se (P. Stenberg), anders.nordgaard@liu.se (A. Nordgaard), jon.ahlinder@foi.se (J. Ahlinder).

key issue in this respect is how best to quantify and express one's degree of belief in competing propositions regarding the available evidence.

Unfortunately, no such framework has yet become widely accepted in the microbial forensic science community. One approach, known as phylogenetic forensic science [5,6], relies on the inference of phylogenetic trees based on sampled sources and evidence sequences. The position of an evidence sequence in the tree is then used to determine its (most probable) source and ancestry. Having established a tree, a forensic experiment must then determine whether the evidence is conclusive or inconclusive on the basis of existing knowledge about the organism, its possible transmission routes, its niches, and the source environment under investigation. González-Candelas et al. [7] utilized phylogenetic trees to reconstruct the transmission of hepatitis C in a well-known forensic case involving an anesthetist accused of deliberately infecting patients in Valencia, Spain. Although the phylogenetic forensic science approach has been very successful in court hearings on microbiological cases, it lacks an underlying probabilistic model for source assignment and so cannot be used to quantitatively determine how strongly one can believe that a strain originates from a candidate source. As such, it is somewhat arbitrary and its outcome depends on the forensic investigator's knowledge and judgement.

A framework based on likelihood ratios has become established in forensic analyses of human DNA sequences for criminal investigations, satisfying the requirement for a statistical framework with well-understood and transparent assumptions, and has become a widely accepted basis for interpreting DNA matches [8–10]. This framework relies on the concept of a “random match probability”, i.e. the probability of obtaining a match with some member of the population other than the accused person. This “random match probability” is inferred by considering a validated human population genetic database with a panel of reference markers based on short-tandem-repeats in linkage equilibrium (e.g. [11]). This allows forensic experts to make a quantitative statement about the strength of the support that the genetic evidence provides for given hypotheses. In addition, there is a well-established approach for communicating the results of these analyses in legal proceedings, with defined interpretations of quantitative results derived using the likelihood ratio method [12,13]. No such conventions exist for the analysis of microbial DNA, so it would be desirable to develop a similar framework based on the likelihood ratio method.

Unfortunately, this is a challenging task because unlike the human genome, the evidence material in microbial forensic science (i.e. microbial populations) evolves constantly and rapidly. Bacteria reproduce asexually by cell division, so every daughter cell would be genetically indistinguishable from its ancestor were it not for diverse mechanisms that introduce genetic variation in bacterial populations [14]. These mechanisms can cause unrelated organisms to contain shared sequences. Microbial species, and even groups of organisms within a microbial species, differ in their capacity for genetic variation. Some species such as the human pathogens *Helicobacter pylori* [15] and *Campylobacter jejuni* [16] are very prone to shuffling parts of their genome or integrating genetic material from other organisms into their own genomes, while other species or lineages within species do so rarely, such as the Tier 1 pathogen *F. tularensis* [17]. Species of the latter type are described as clonal, and their genetic content changes only minimally over generations. Many of the most pathogenic bacterial species have clonal (or even monomorphic) population structures [18].

Traditional population genetic analyses conducted in clinical microbiology and investigations into infectious diseases are

based on consensus-level genetic data for the studied isolates. These data are used to establish a genotype by methods such as genome sequencing of large genomic regions (which has recently become feasible at an affordable price and in a realistic time frame) [14,19]. This data generation provide details on the most likely consensus-level mutations present in a sample compared to a reference sample. Recently, however, high throughput sequencing (HTS) has enabled the sequencing of entire bacterial populations within a sample, with sufficient sequencing coverage to permit the detection and quantification of low frequency mutations at an unprecedented level of detail [20,21]. In a forensic setting, this makes it possible to distinguish between possible sources of a sample recovered from a crime scene by identifying signatures of low frequency mutations.

In conclusion, to create a microbial forensic framework, it will be necessary to build on approaches that have proven successful in other areas of forensic science and modify them to suit the unique properties of bacterial pathogens in terms of relevant hypotheses and population genetic structures. We therefore sought to develop an approach for calculating likelihood ratio-based evidential values by adapting known statistical methods to microbial genetic data. The methods were applied in two case studies examining the pathogenic bacteria *L. monocytogenes* (using consensus-level sequences) and *F. tularensis* (using population genetic information on low frequency mutations). We demonstrate that hypotheses can be evaluated by calculating evidential values that can be reformulated into verbal conclusions, enabling efficient legal communication of the results of forensic microbial DNA analysis.

2. Material and methods

2.1. Methods of evidence evaluation

2.1.1. Likelihood ratio method

The likelihood ratio method is the primary tool for evaluating DNA profiling evidence in modern human crime investigations. One of its advantages over other methods is that the evidence is analyzed both in the context of the hypothesis forwarded (usually) by the prosecution and in the context of a relevant alternative to that hypothesis. As such, it quantifies the strength of the evidence for or against each hypothesis in the debated context. Another benefit is that it allows for several independent kinds of evidence to be integrated into a single evidence value. A challenge when using this method is to formulate the competing hypotheses in an appropriate way to ensure that the calculated likelihood ratios are relevant to the question at hand. It must also allow for the individual likelihoods to be calculated or estimated accurately, either analytically or by estimation based on the available data. A common formulation of hypotheses in criminal forensic investigations involving analysis of human DNA is:

H_m : The recovered DNA profile originates from the suspect

H_a : The recovered DNA originates from someone who is not the suspect or a close relative of the suspect.

H_m is referred to as the main hypothesis and H_a as the alternative hypothesis. The likelihood ratio (LR) is then the ratio of the probability of the DNA evidence (*Data*) given the main hypothesis to the probability of *Data* given the alternative hypothesis:

$$LR = \frac{P(Data|H_m)}{P(Data|H_a)} \quad (1)$$

When the hypotheses H_m and H_a are simple, i.e. when each hypothesis represents a single explanation of the *Data*, the

likelihood ratio acts as a bridge between the prior and posterior odds in Bayes' theorem expressed in odds form:

$$\frac{P(H_m|Data)}{P(H_a|Data)} = LR \times \frac{P(H_m)}{P(H_a)} \quad (2)$$

where, $(P(H_m|Data)/P(H_a|Data))$ are the posterior odds and $(P(H_m)/P(H_a))$ are the prior odds. According to the theorem, the prior odds of the main hypothesis are updated to posterior odds in light of the evidence by multiplying the prior odds by the likelihood ratio. The value of the likelihood ratio thus represents the strength of the forensic evidence and determines whether the prior odds are updated in a way that strengthens (if the likelihood ratio is above 1) or weakens (if likelihood ratio is below 1) the evidence for the main hypothesis.

Both the hypotheses formulated in expression (1) fulfil the requirement that one and only one source is pointed out in each hypothesis (i.e. simple hypotheses). In other cases, where the alternative hypothesis comprise of several sources (i.e. a composite hypothesis), the evidence value would be calculated according to the following ratio:

$$P(Data|H_m) / \sum_{i=1}^{N_a} P(Data|H_{a,i})P(H_{a,i}|H_a), \quad (3)$$

where, N_a is the number of sources that H_a comprises and $H_{a,i}$ is the sub-hypothesis that the origin of the recovered evidence is source i ($i = 1, \dots, N_a$). The probability $P(H_{a,i}|H_a)$ is referred to as the relative prior probability that the origin is source i given that it is one of the sources comprised by H_a . Hereafter, we will only consider simple hypotheses.

2.1.2. Statistical models for data

If the scale of *Data* is continuous, the likelihood ratio is the ratio of the probability density function (PDF) valid under H_m evaluated at *Data* to the PDF valid under H_a evaluated at *Data*, i.e.

$$LR = \frac{f(Data|H_m)}{f(Data|H_a)}. \quad (4)$$

Therefore, to calculate the likelihood ratio directly we need to know or estimate the PDFs under the two hypotheses. This may be relatively straightforward, for example where there are well-known and empirically validated probability distributions describing the distribution of the gathered data under different scenarios. In other cases, the shape of the PDFs will be less well-known, necessitating estimation using simulated or real data. This may be done by normal approximation or kernel density estimation (KDE) [22]. If the data are approximately normally distributed, the normal approximation is a good choice for the estimation. If data are not normally distributed and cannot be transformed to normality, the non-parametric KDE method is more suitable. KDE can be described as a "smoothing histogram" where each data observation contributes to the histogram as a distribution of any form instead of as a uniformly distributed variable. KDE is often applied to univariate data in forensic science (e.g. [23,24]) and is also nowadays commonly extended for use with multivariate data (e.g. [25]).

KDE is a reliable method in cases where the training data (consisting of simulated or real observations) for the studied hypotheses features many observations and few variables (e.g. low-dimensional data). However, if there are many variables (high dimensional data) and/or a limited number of data observations for the hypotheses, KDE becomes more complex and less robust, often because of a lack of observations in the tails of the distribution.

To overcome the problem of overly high data dimensionality, variable reduction methods can be used. One way is to fit graphical

models to the data, taking into account the dependency structure between the different variables. Identifying loose or negligible dependencies between (groups of) variables makes it possible to apply lower-dimensional models to such groups. These low-dimensional models can then be combined to form a full model. This method has been used in forensic science investigations involving glass fragment data [26,27]. Another way to reduce the number of variables is to move from a feature-based method in which each feature (variable) add one dimension to the probability density function to a score-based method in which likelihood ratios are calculated based on the PDFs of the variables' scores. One such score-based approach relevant to genetic analyses involves estimating PDFs based on genetic distances rather than individual markers. This is the approach we used to overcome difficulties encountered in the *F. tularensis* case study. Another example of a score based method is the use of multivariate discriminant analysis to obtain a linear combination of the original variables in the form of a decision value, which can then be used to approximate the PDFs under the hypotheses of interest [28].

An alternative way of dealing with high dimensional data is to assign equal prior probabilities to the hypotheses in question and use classification methods (e.g. random forest, support vector machines or logistic regression) to calculate the posterior odds. The equal prior probabilities assumption then allows the likelihood ratio to be calculated backwards using Bayes' theorem (i.e. obtaining the likelihood ratio as the ratio of the posterior odds to the prior odds).

2.2. Case scenarios

2.2.1. Scenario 1

The disease listeriosis is caused by the human pathogen *L. monocytogenes*. An outbreak of this disease in the US reportedly originated from contaminated ice cream [29]. Two production facilities were believed to be the source of the outbreak. As reported by Chen et al. [29], genomes originating from these facilities formed two clusters on a single branch, belonging to serogroup 2b and genetic lineage 1. In a fictive forensic case, production facility 1 was accused of being the origin of the recovered clinical isolates. Two hypotheses were then considered for each patient:

H_m : the source of the isolate from the patient was production facility 1,

H_a : the source of the isolate from the patient was not production facility 1.

To perform the *L. monocytogenes* case study analyses, whole genome sequence reads were downloaded from <ftp://ftp.sra.ebi.ac.uk/>. These genomes are available under Genbank Bioproject accession PRJNA215355, and were originally obtained by whole genome sequencing performed according to the guidelines provided by the Centers for Disease Control and Prevention (<https://www.cdc.gov/amd/>). The paired-end reads were trimmed using Trimmomatic [30]. Assemblies were created using ABySS [31]. The Pilon software package [32] was used to improve the variant detection probability. The strain SRR1917440 was used as a reference in ProgressiveMauve [33] to create an alignment of the genomes in the population. In total 40,548 SNPs were used to create consensus level-sequences of all isolates. Calculations of pair-wise genetic distances in the SNP profiles were based on the Hamming metric as implemented in DiStats [34]. Inference of a phylogenetic tree was performed using BEAST2 [35] with the following non-default settings and priors: a GTR substitution model, the site heterogeneity model was gamma (+G) with a proportion of invariant sites (+I), and the tree prior was a Yule model. Gamma distributions with default values were assigned as priors for the substitution rates, and a strict clock model was assumed for the substitution rate. The GTR + G + I model has been found to outperform (simpler) alternative models [36]. The MCMC chain length was set to 100,000,000 with a thinning of every 10th iteration and a burnin of

10,000,000 iterations. The inferred phylogenetic tree was plotted using ggtree [37]. Kernel smoothing was used to produce probability density distributions under the main and alternative propositions in the same way as in the *F. tularensis* case but with a bandwidth of 0.0005. The results of the evidence calculations for both scenarios were interpreted based on a previously proposed ordinal scale [13]. For further details of the analysis, including the scripts used to generate the results, please visit <https://github.com/FOI-Bioinformatics/MicrobialForensicScience>.

2.2.2. Scenario 2

Consider a fictive situation in which *F. tularensis* has been deliberately spread, e.g. as a powder in letters or in the drinking water system, causing a disease outbreak from which the source strain has been recovered. Preliminary epidemiological investigations demonstrate that there is no plausible natural process by which the patients could have been infected by the recovered strains. Genomic data on the source strain is acquired by sequencing at a relatively low sequence depth of 30×, and a comparison to published genomes reveals it to be identical to strain SCHU-S4 [38]. SCHU-S4 is a well-known highly pathogenic strain that has been distributed to laboratories all over the world since its parental strain SCHU was isolated in 1941 in Ohio, US [38]. For simplicity, we assume that only two laboratories (A and B) keep samples of the outbreak strain as lab stock cultures. Batch cultures of the strain are gathered from the two laboratories. Assume further that one of the laboratory batch cultures, batch culture A, is the suspected culture, while the other culture, batch culture B, acts as a reference sample. The hypotheses to be investigated are:

H_m : Culture batch A is the source of the attack sample

H_a : The genetic variance of the SNP profiles was Culture batch A is not the source of the attack sample.

To calculate reliable likelihood-ratio based evidential values, reference populations under the two competing hypotheses are needed. For this scenario, we used data from an earlier study [39] in which two separate single colonies of *F. tularensis* strain SCHU-S4 FSC237 $\Delta clpB$ were serially propagated over three transfers to form two culture batches. These culture batches were then further propagated over a further 23 transfers in 14 separate parallels, seven for each batch culture. Cultivation was performed in flasks containing Chamberlain's medium using inocula containing 10^6 CFU/mL for each transfer. The serial passages of the batches are represented schematically in Fig. 2. Sequencing was performed using a HiSeq 2000 instrument at depths of 500× for the initial isolates; 15,000× for the two batch cultures; 2500× for transfers 2, 4 and 6; and 500× for transfers 8 and 24. We chose to base our analysis on the sequencing data for the samples with the highest sequencing depth, i.e. the culture batches and transfers 2, 4, and 6.

To visualize the genetic variation between the samples, non-metric multidimensional scaling (nMDS) was performed using the function metaMDS available in the R package vegan [40]. The Euclidean distance between the SNP profiles of the deep sequenced *F. tularensis* populations was chosen as the distance metric. The resulting ordination was visualized using ggplot2 [41]. Probability density functions for the competing hypotheses H_m and H_a were generated by kernel density estimation and based on pairwise Euclidean distances between samples originating from the lab A batch culture and between samples originating from different batch cultures, respectively. To characterize the samples' genetic profiles, we used pairwise Euclidean distances based on both SNP frequencies and the binary classification (present/absent) of the structural variant mutations. Kernel density estimation was performed using the geom_density function in the R package ggplot2 with normally distributed kernels. For structural variant analysis and for SNP analyses after two serial transfers, the default bandwidth "nrd0" was used [22]. A bandwidth of 2 was used for

SNP analyses of samples collected after four and six serial transfers. The genetic variance of the SNP profiles was calculated based on observed SNP frequencies as:

$$V_G = \sum_{i=1}^{12713} p_i(1 - p_i), \quad (4)$$

Assuming all SNPs are independent and p_i is the observed frequency of the polymorphism at the i :th position (there were 12,713 segregating positions in total) (see Ref. [42], adjusted for a haploid species).

The reads were mapped to reference AJ749949.2 using bowtie2 [43] and SNPs were basecalled using varsScan version 2.4.2 [44]. The coverage for each position in each sample was checked using the depth base tool in sambamba [45]. To include positions with a depth of zero, the option -c 0 was applied. The duplications were found by visual inspection of coverage plots over the genome. To facilitate visualization of the coverage figures, only every 100th position is plotted. To verify the mutations, and to determine their approximate positions, a python script was written to go through the coverage for each base in each genome using a sliding window of size 100 (script available at [https://github.com/FOI\[HYPHEN\]Bioinformatics/MicrobialForensics](https://github.com/FOI[HYPHEN]Bioinformatics/MicrobialForensics)). Duplications were defined as areas with 100 or more consecutive sliding windows in which the mean coverage was >1.2 times the median coverage over the entire genome. A deletion was defined as an area with 100 or more consecutive sliding windows in which the mean coverage was <0.2 times the median coverage over the entire genome. Three samples were randomly selected as putative attack samples, one representing each transfer (sample ids: A1-2, A2-4, and A6-6), and were excluded from the estimation of the corresponding PDFs for evidence value calculations.

3. Results

3.1. Tracing the source of a listeriosis outbreak to ice cream production facilities

We performed consensus-level re-analyses of 168 genomes originating from the two production facilities, the environment, the studied outbreak, and earlier outbreaks (see supplementary material in [29] for isolate names). Eight of the 168 genomes originated from the hospitalized patients, and our objective was to compute evidential values for the sources of each of these genomes. A preliminary assessment was conducted by performing a phylogenetic analysis based on called SNPs. Genomes originating from the two facilities were clearly separated, with some sequences originating from other sources being placed in between the two facility clades (Fig. 1A). The sequences obtained from each patient were placed well within the respective facility clades, indicating a distinct source origin for each sequence. The major branches of the tree had high support, with posterior probabilities close or equal to one (results not shown).

The population defined to form probability density functions (PDFs) under H_m and H_a consisted of all strains originating from facilities 1 and 2. All strains of other origins (i.e. environmental strains and those isolated during earlier disease outbreaks) were excluded from the evidence calculations because those origins were not included in the investigated hypotheses. A score-based approach was used to create PDFs of genetic distances within and between production facilities via kernel smoothing such that $f(data|H_m)$ reflects the distribution of pair-wise genetic distances within facility 1 and $f(data|H_a)$ is the distribution of distances between sequences from the two facilities. The distributions under H_m and H_a were clearly separated (Fig. 1B), with means and

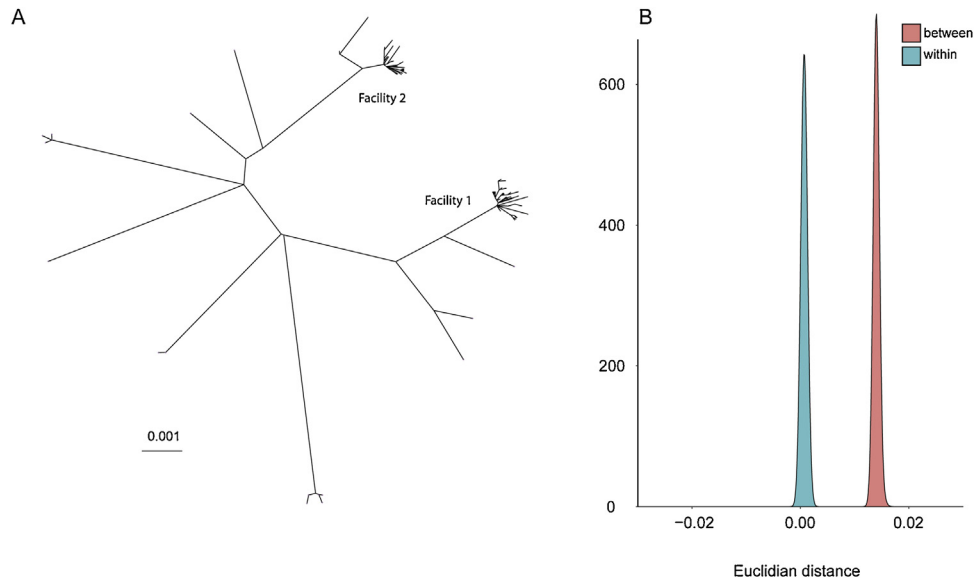


Fig. 1. (A) Inferred unrooted phylogeny of the analyzed *L. monocytogenes* samples based on the maximum clade credibility. Turquoise and green nodes correspond to samples originating from facilities 1 and 2, respectively. (B) Probability density functions for the genetic distance within and between facilities, under H_m and H_a respectively, based on the Hamming genetic distances between samples.

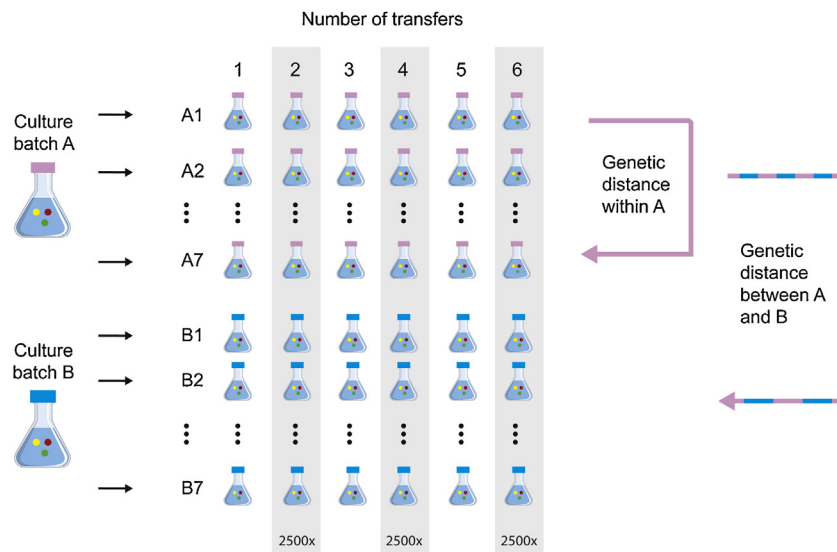


Fig. 2. Cultivation of culture batches to obtain data for the competing hypotheses. Culture batches A and B were propagated in seven serial passages, each involving six transfers. Samples were collected after transfers 2, 4, and 6 for shotgun sequencing with a sequence depth of 2500 \times . Genetic distances within samples originating from culture batch A and between samples originating from different culture batches were calculated to generate reference data for the competing hypotheses.

variances (the latter are given in parentheses) of 0.00070 (0.00018) and 0.01402 (0.00024), respectively.

To calculate evidential values, the genetic distances from sequences associated with facility 1 were calculated for the sequences obtained from each patient, and the densities under H_m and H_a were compared according to Eq. (4). For each distance, a likelihood ratio was obtained from the PDFs and the average likelihood ratio for each patient were reported. All the calculated evidential values were strongly supportive or contradictory, depending on the origin of each isolate (Table 1), and the assignments of clinical samples to their respective sources were fully consistent with the findings of Chen et al. [29]. Using the previously recommended terminology for communicating forensic evidence values based on likelihood ratios to individuals not

trained in statistics [13], the results of the examination extremely strongly support the conclusion that facility 1 was the source of the patient strains, and the possibility that these results would be obtained if an alternative hypothesis were true can be excluded in practice. For strains originating from facility 2, the results of the examination extremely strongly support the conclusion that facility 1 was not the origin of the disease-causing strains. Because only strains from two facilities were included, the results extremely strongly support the conclusion that facility 2 was the source of the strains. The between facilities genetic distance distribution (from which the PDF under H_a is estimated) would more closely resemble the true distribution, if new isolates sampled in additional production facilities would have been included in the analysis.

Table 1
Evidential values and corresponding conclusions for the clinical isolates based on the Hammering genetic distance.

Patient	Isolate	Evidential value	Conclusion
1	SRR1193828	1.25E + 16	The result of the examination extremely strongly support the hypothesis that production facility 1 is the source of strain.
2	SRR1217489	1.05E + 16	The result of the examination extremely strongly support the hypothesis that production facility 1 is the source of strain.
3	SRR1695810	1.72E−17	The result of the examination extremely strongly support the hypothesis that production facility 2 is the source of strain.
4	SRR1745441	1.44E + 16	The result of the examination extremely strongly support the hypothesis that production facility 1 is the source of strain.
5	SRR1975179	1.44E + 16	The result of the examination extremely strongly support the hypothesis that production facility 1 is the source of strain.
6	SRR1996267	2.04E−17	The result of the examination extremely strongly support the hypothesis that production facility 2 is the source of strain.
7	SRR1996268	2.34E−17	The result of the examination extremely strongly support the hypothesis that production facility 2 is the source of strain.
8	SRR2047270	2.34E−17	The result of the examination extremely strongly support the hypothesis that production facility 2 is the source of strain.

To summarize, our re-analysis produced a classification of the eight *L. monocytogenes* clinical isolates in full agreement with that reported previously. However, our analysis has the added value of providing probabilistic assessments of the strength of the classifications and their evidential value.

3.2. Cultivating possible sources of an attack sample enables calculation of an evidential value

For this scenario, we used data from an earlier study [39] in which two bacterial batch samples were serially propagated over

26 passages and shotgun sequenced. Dwibedi et al. [39] report that the batch cultures (batch cultures A and B in this scenario) were derived from two single colonies of strain SCHU-S4 FSC237 Δ clpB. The relationships between the cultivated populations were visualized by performing non-metric multidimensional scaling (nMDS) based on genetic distance and the frequency of single nucleotide polymorphisms (SNPs). This did not reveal any clear trend differentiating the batch cultures, suggesting that analyses based on SNP frequency and genetic distance alone will not discriminate effectively between batches or cultures having undergone different numbers of transfers (Fig. 3A). Populations

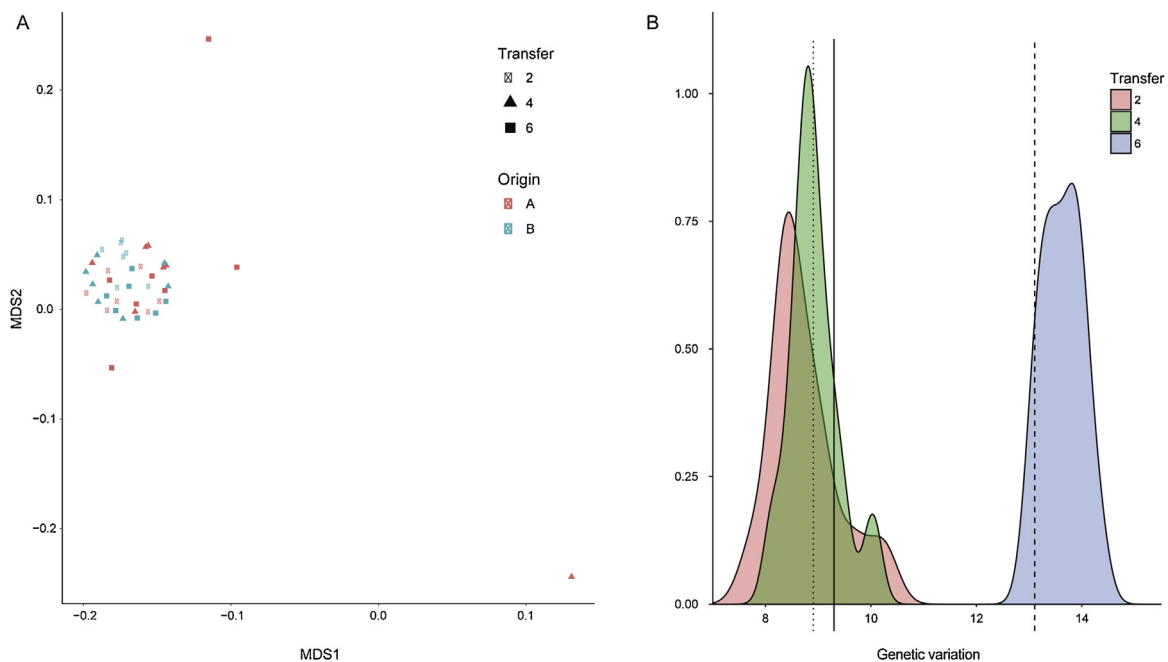


Fig. 3. (A) Non-metric multidimensional scaling of pair-wise genetic distances between samples originating from batch populations A or B after different numbers of transfers. (B) Estimated probability density functions of the genetic variation in SNP frequencies for different numbers of cultivated transfers. The solid, dotted and dashed vertical lines indicate the genetic variation in the attack populations after two, four, and six transfers, respectively.

that had undergone four or six transfers were generally further separated than those that had undergone only two transfers, but most of the populations clustered in the same region of the plot irrespective of their batch origin.

One sample from each sequenced transfer (i.e. transfer two, four, and six) was selected to represent the fictive attack populations (i.e. the evidential material). To perform evidence calculations, one also requires data on a representative reference population. This part of the investigative phase, i.e. the selection of a representative reference population, is highlighted here because in a real forensic case, the analyst will not know how the attack population was amplified and this uncertainty (in the assignment of the reference set) will affect the accuracy with which the evidence is evaluated. To define the reference population for comparison to the attack sample, we calculated the genetic variance of the samples at each transfer. Genetic variance was selected as the metric for comparison because variation should be generated during the serial passages: the greater the number of transfers, the more variation should be present in the population. PDFs of the variance in SNP frequencies were estimated for samples that had undergone two, four and six number of transfers. These PDFs were then used to estimate the likelihood that the genetic variation of the attack populations would be observed under each distribution to identify the reference population with the highest likelihood score. The estimated PDF for the samples that had undergone six transfers was clearly differentiated from those for the samples that had undergone only two and four passages: the means (and variances in parentheses) of the genetic variation for the two-, four-, and six-passage samples were 8.74 (0.43), 8.92 (0.23) and 13.63 (0.14), respectively. The PDFs for the samples that had undergone two and four passages overlapped extensively, suggesting that the choice of reference between two and four transfers is of little importance (Fig. 3B). However, the pdf of the six-transfer population was well separated from the other pdfs. The genetic variation in the attack populations from the two-, four-, and six-transfer samples was 9.30, 8.91, and 13.11, respectively.

The next step was to estimate PDFs for the pair-wise genetic distances based on SNP frequencies in the reference populations after two, four and six transfers for samples originating from batch A (H_m) and between batches A and B (H_a). The resulting distributions could not be separated under the competing hypotheses because their PDFs overlapped very extensively (Fig. 4). In fact, for samples that had undergone six transfers, the PDFs for distances within batch A extended further to the right

(indicating greater genetic distances) than that for distances between batches. The SNP data thus lacked discriminative signatures, indicating that alternative genetic data would be needed to obtain conclusive evidence.

As an alternative genetic signature, the coverage of sequence reads aligned to the reference genome might provide additional discriminative details as described in Dwibedi et al. [39]. A sudden coverage increase or decrease in a specific regions suggesting a duplication or deletion event [46]. Here we are using three structural variants showing an altered distributions for samples originating from different batches (Fig. 3). Two large duplications (approximately 102 kb and 360 kb, respectively) that existed uniquely in samples originating from batch A, and one deletion (approximately 0.2 kb) only existed only in samples originating from batch B. Most samples from batch A exhibited both duplications after passage 2; the remainder exhibited just one. However, one or both of the duplications was lost during subsequent passages in some cases. All samples from batch B exhibited the deletion. The presence/absence of these structural variants in each serial passage after different numbers of transfers is summarized in Table 2, and the corresponding coverage distributions are exemplified in Fig. 5. PDFs for the pair-wise genetic distances (based on the presence/absence of these mutations) between samples originating from within batch A and between the batches were generated by kernel density estimation and are shown in Fig. 6.

Pair-wise genetic distances between the attack samples and the corresponding reference populations from batch A were calculated. Each pair-wise distance was associated with a likelihood ratio (i.e. the ratio of the values of the PDFs at the given pair-wise distance). The average likelihood ratios for the attack samples from transfers two, four and six were $1.12e + 17$, $4.30e + 7$ and $1.02e + 17$, respectively. Table 3 presents the resulting evidential values and descriptions of the conclusions using the terminology previously developed for communicating statistical results in court and in criminal investigations [13]. These fluctuations in obtained LR-score reflects the uncertainty in the analysis, as the number of populations included were relatively few, as were the number of mutation events behind the estimated PDFs. As the kernel density estimation method can produce unreliable densities for regions covered with few data points [22], it is important to assess the sampling procedure to estimate the PDFs.

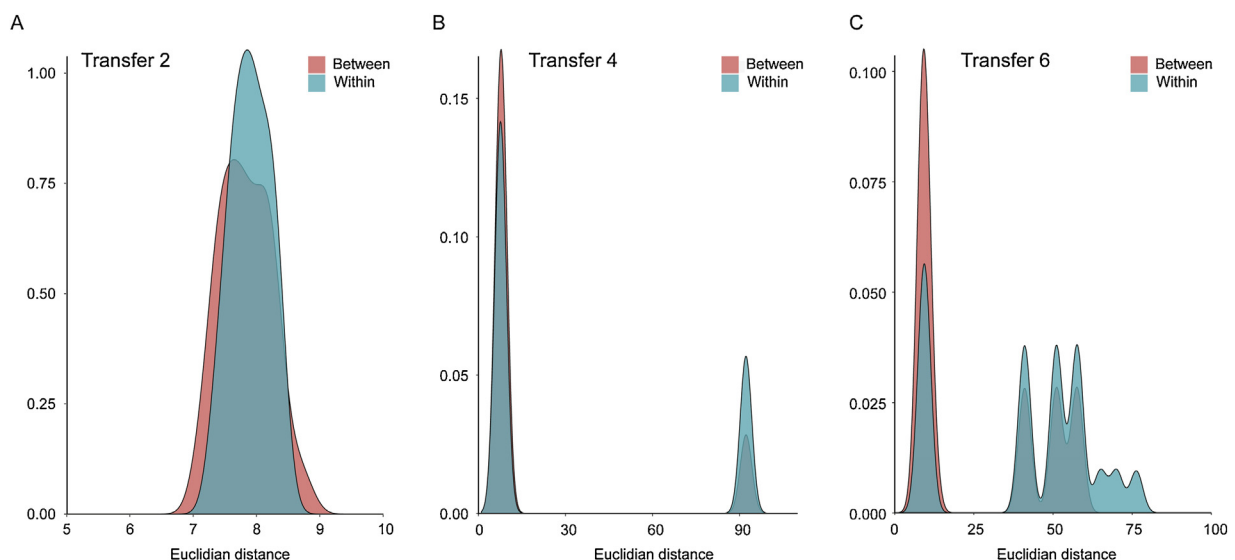


Fig. 4. Estimated probability density functions (PDFs) of pairwise Euclidean distances (based on SNP frequencies) between samples originating within culture batch A and between samples from culture batches A and B, respectively (A) PDFs of samples after 2 transfers. (B) PDFs of samples after 4 transfers. (C) PDFs of samples after 6 transfers.

Table 2
Binary classification (presence/absence, denoted by +/-) of structural variant mutations based on per base coverage for all samples considered in the study. The abbreviation "tr" stands for "transfers". The approximate positions of the mutations with respect to the reference are 352.1 kb–454.1 kb, 1408.3 kb–1767.8 kb, and 920.0 kb–920.2 kb for duplications 1 and 2 and the deletion, respectively.

Serial passage	Duplication 1			Duplication 2			Deletion		
	2 tr	4 tr	6 tr	2 tr	4 tr	6 tr	2 tr	4 tr	6 tr
A1	+	+	+	+	-	-	-	-	-
A2	-	-	-	+	+	+	-	-	-
A3	+	+	+	+	+	-	-	-	-
A4	+	+	+	+	-	-	-	-	-
A5	+	+	+	+	-	-	-	-	-
A6	+	+	+	+	-	-	-	-	-
A7	-	+	+	+	+	-	-	-	-
B1	-	-	-	-	-	-	+	+	+
B2	-	-	-	-	-	-	-	+	+
B3	-	-	-	-	-	-	-	+	+
B4	-	-	-	-	-	-	-	+	+
B5	-	-	-	-	-	-	-	+	+
B6	-	-	-	-	-	-	+	+	+
B7	-	-	-	-	-	-	+	+	+

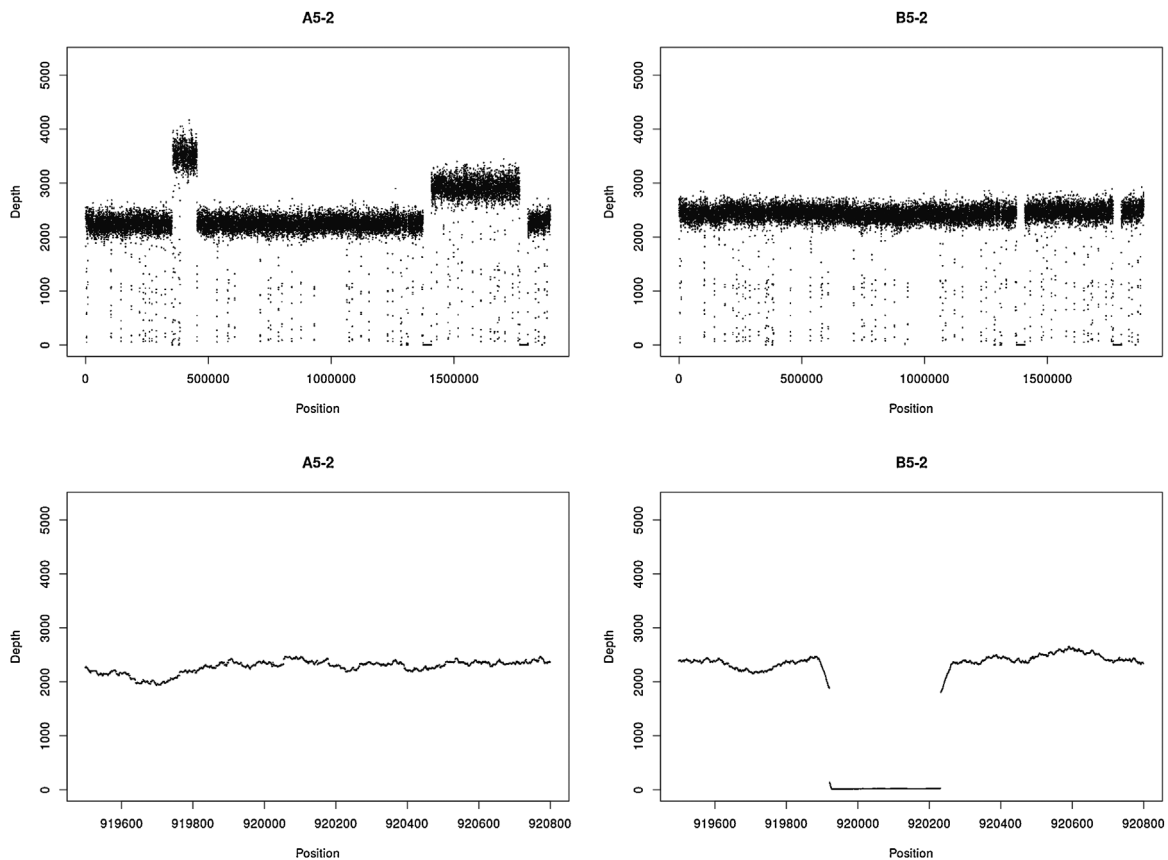


Fig. 5. Typical coverage plots and signals for structural variants. Top row: Data for a sample from culture batch A (A5-2), showing two peaks corresponding to duplication events, and a sample from culture batch B (B5-2) that lacks the duplications. Sporadic drops in coverage are due to repetitive IS-elements distributed across the genome. Bottom row: Expansion of the region around 920 kb in the upper plots showing the decline in coverage corresponding to the deletion event in sample B5-2. No such event is seen in sample A5-2.

To summarize, we were able to define an alternative signature, based on coverage data, for the samples to discriminate between two competing hypotheses regarding batch origin.

4. Discussion

Interest in the field of microbial forensic science has increased dramatically in recent decades. Like other fields in forensic science, the goal of microbial forensic science is to

deliver data/information that are objective, understandable, and informative with respect to stated hypotheses. Unfortunately, current methods for attribution in microbial forensic science lack a probabilistic foundation that can be used to provide quantifiable results. To address this limitation, we propose a method based on the probabilistic quantities known as likelihood ratios, which can be used to quantify degrees of belief in competing hypotheses regarding the source attribution of microbial samples.

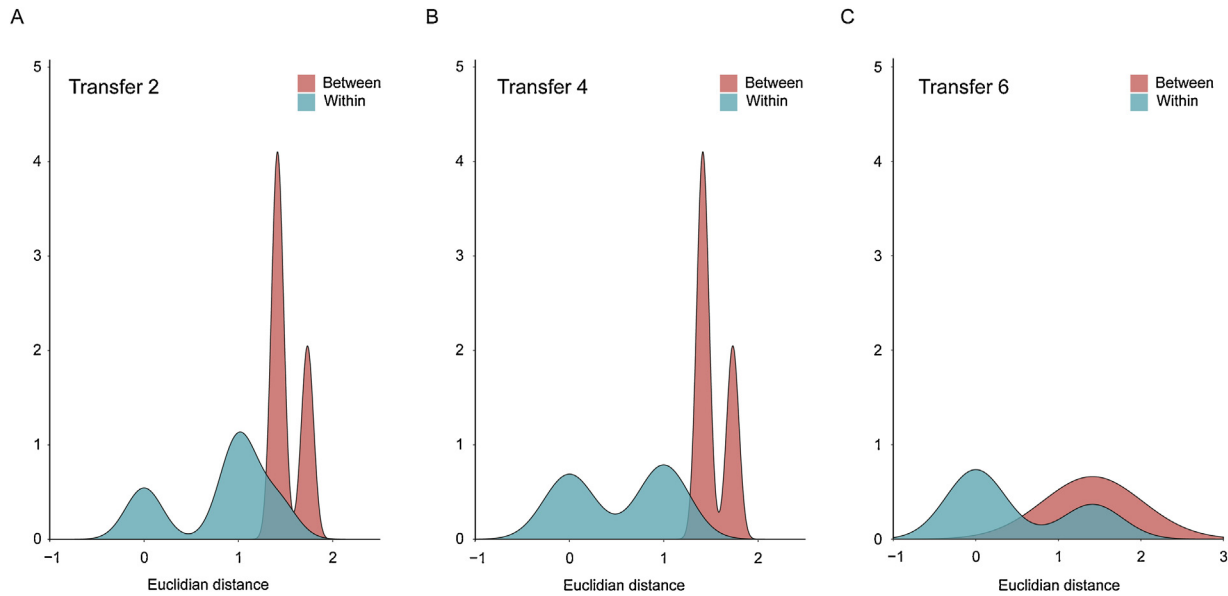


Fig. 6. Estimated probability density functions of pairwise Euclidean distances for structural variants between samples from culture batch A and between samples from culture batches A and B, respectively: (A) PDFs for attack sample A1-2 and the four-transfer reference populations. (B) PDFs for attack sample A2-4 and the four-transfer reference populations. (C) PDFs for attack sample A6-6 and the six-transfer reference populations.

Table 3

Evidential values and corresponding conclusions for the fictive attack populations based on structural variants.

Attack population	Reference population	Evidential value	Conclusion
2	4	1,12E + 17	The result of the examination extremely strongly support the hypothesis that batch culture A is the source of the attack sample.
4	4	5,30E + 07	The result of the examination extremely strongly support the hypothesis that batch culture A is the source of the attack sample.
6	6	1,46E + 01	The result of the examination support to some extent the hypothesis that batch culture A is the source of the attack sample.

Our method was applied in two case scenarios: a food poisoning outbreak of listeriosis caused by the human pathogen *Listeria monocytogenes*, and a batch culture attribution scenario involving the tularemia-causing bacterium *F. tularensis*. In both scenarios, we demonstrated how to set up and evaluate hypotheses regarding sample origin by calculating evidential values, and present conclusions using previously developed terminology for communicating forensic results to people without statistical training [13], which has not been done previously in the context of microbial forensic science.

In the *F. tularensis* case study, we relied on signatures derived from ultra-deep sequencing data, which is crucial when investigating highly monomorphic species because resolution at the consensus level is limited. SNPs are typically the markers of choice when characterizing genomes. However, in this case, an analysis based on SNP frequencies would not enable confident source identification because the vast majority of the SNPs were sample-specific rather than batch-specific. Instead, SNP profiles were used to quantify the accumulated genetic variation over time to identify appropriate reference populations, increasing the accuracy of the evidence evaluation. The main mutations that separated the batch

cultures were structural variants detected in the sequence coverage analysis. These variants are difficult to detect with variant calling algorithms but were identified in the genome coverage profiles (whose construction was enabled by the use of high-throughput sequencing data). It should be noted that in cases involving large-scale cultivation of pathogens for production purposes, the method proposed herein requires cultivation of possible batch culture sources to create representative reference populations relevant to the forensic task at hand. These reference populations are needed to successfully evaluate hypotheses using the likelihood ratio method. Our findings of structural variants generated during laboratory cultivation and previous reports on similar genetic changes in culture experiments using other bacteria suggest that investigating structural genetic variants is a fruitful approach to detect laboratory-induced mutations [39,46,47].

We also demonstrated the use of the likelihood ratio method in an analysis of a real foodborne disease outbreak caused by the pathogen *Listeria monocytogenes* [29]. Chen et al. [29] performed comparative genomic analyses of strains collected at the facilities suspected to be the sources of the infection and the strains that infected the patients. The consensus sequences of the outbreak-associated isolates were then placed in a phylogenetic tree to determine whether they were most likely to have originated from facility one or facility two based on their clustering with the two sets of facility-derived isolates. In addition, a set of SNPs were reported to discriminate between the two facility lineages, indicating adaptation of the isolates to their local environments. Our analysis complements and expands upon the source-tracking analysis of Chen et al. [29] by quantifying the strength of the evidence for the origin of the outbreak strains. Source tracking in this case was relatively straightforward because of the substantial genetic differences between strains from the two facilities, and one could argue that the phylogenetic approach was sufficient for source tracing. However, when such analyses are conducted in the context of a forensic investigation, it is very desirable to be able to quantify the degree of belief in the hypotheses and to be able to communicate that strength of belief in a way that is easily understood by participants in legal proceedings.

Many similar phylogenetic investigations into disease outbreaks have been reported in the literature. For example, Whaley et al. [48] identified the most probable source origin of a *Neisseria meningitidis* outbreak in the U.S. by linking isolates to well-known reference strains of known origin. The *L. monocytogenes* case study discussed here demonstrates that the method presented in this paper could be generally useful in epidemiological investigations in which there is a need to differentiate between genome sequences from suspected sources.

The greater the amount of data available for the different hypotheses under consideration, the better the estimated probability density functions and the more accurate the calculated evidential values. In the case study on *F. tularensis*, seven parallel serial passages were performed for each lab batch. This is probably too few samples to obtain good estimated density functions and calculate robust evidential values. However, our objective was to assess the merits of the likelihood ratio method in microbial forensic science rather than to produce robust evidential values. In a real forensic investigation, we recommend performing at least 100 parallel serial passages for each suspected source. Additionally, further studies are needed to assess the robustness of the evidential values in different scenarios and its dependence on the amount of available data.

The investigations presented in this paper adhered to the standard four-step procedure for conducting forensic investigations, which entails (1) collecting evidence material, (2) examining the evidence, (3) analyzing the evidence, and (4) ensuring that the results of the analysis are reliable and reporting them transparently. Unlike earlier microbial forensic works, we suggest that the whole process of forensic microbial genomic analysis should be based on hypothesis testing using likelihood ratio methods to calculate evidence values, and that these results should be expressed using a Scale of conclusions [13] to enable objective communication of the results of DNA analyses during legal proceedings. The common approach of using phylogenetic methods to connect outbreak isolates to a putative source (i.e. the phylogenetic forensic method), as used in the *L. monocytogenes* case, does not enable quantification of confidence in source attribution; it only permits a candidate source origin to be accepted or excluded, without accounting for the uncertainty in the assignment. Importantly, the likelihood ratio approach should be generally applicable in diverse scenarios – for example, cases in which a recovered evidence isolate may be naturally present in the local environment, or where there is uncertainty about whether a disease outbreak has malicious or natural origins.

5. Conclusions

We have demonstrated the usefulness of the likelihood ratio approach for the evaluation of source attribution in microbial forensic cases by calculating likelihood ratios for two case scenarios: one in which high throughput sequencing data were used to distinguish between extremely similar batch cultures of *F. tularensis*, and one in which publicly available sequencing data were used to identify the source of a *L. monocytogenes* outbreak. The main purpose of the method is to facilitate the interpretation of evidence in legal proceedings. However, it may also have applications in contexts outside criminal investigations, such as in source tracing of waterborne contaminants or during food poisoning outbreaks. To become generally accepted, the concept will have to be refined and adjusted to account for the peculiarities of different organisms and scenarios. We therefore hope that the results presented here will inspire other researchers to develop further applications of the likelihood ratio method in microbial forensic science.

CRedit authorship contribution statement

Petter Lindgren: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Kerstin Myrtenäs:** Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Funding acquisition. **Mats Forsman:** Conceptualization, Resources, Writing – review & editing, Funding acquisition. **Anders Johansson:** Conceptualization, Resources, Writing – review & editing, Funding acquisition. **Per Stenberg:** Conceptualization, Writing – review & editing. **Anders Nordgaard:** Conceptualization, Methodology, Writing – review & editing. **Jon Ahlander:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration, Funding acquisition.

Acknowledgement

We would like to acknowledge our colleagues at FOI Umeå, Pär Larsson, Chinmay Dwivedi (Umeå University), Arthur Wolterink and Annabel Bolck (Netherlands Forensic Institute) for fruitful discussions. In addition, we would like to thank the editor and three reviewers for comments that improved the manuscript. This project was funded by the Swedish Ministry of Foreign Affairs, project number A4952 and the Swedish Ministry of Defence, project A4040.

References

- [1] S.E. Schmedes, A. Sajantila, B. Budowle, Expansion of microbial forensics, *J. Clin. Microbiol.* 54 (2016) 1964–1974, doi:<http://dx.doi.org/10.1128/JCM.00046-16>.
- [2] B. Budowle, S.E. Schmedes, M.S. Ascher, R.M. Atlas, J.P. Burans, R. Chakraborty, J.L. Dunn, C.M. Fraser, D.R. Franz, T.J. Leighton, S.A. Morse, R.S. Murch, J. Ravel, D.L. Rock, T.R. Slezak, S.P. Velsko, A.C. Walsh, R.A. Walters, Toward a system of microbial forensics: from sample collection to interpretation of evidence, *Appl. Environ. Microbiol.* 71 (2005) 2209–2213, doi:<http://dx.doi.org/10.1128/AEM.71.5.2209-2213.2005>.
- [3] T. de Oliveira, O.G. Pybus, A. Rambaut, M. Salemi, S. Cassol, M. Ciccozzi, G. Rezza, G.C. Gattinara, R. D'Arrigo, M. Amicosante, L. Perrin, V. Colizzi, C.F. Perno, Benghazi study group, molecular epidemiology: HIV-1 and HCV sequences from libyan outbreak, *Nature* 444 (2006) 836–837, doi:<http://dx.doi.org/10.1038/444836a>.
- [4] B. Budowle, N.D. Connell, A. Bielecka-Oder, R.R. Colwell, C.R. Corbett, J. Fletcher, M. Forsman, D.R. Kadavy, A. Markotic, S. a Morse, R.S. Murch, A. Sajantila, S.E. Schmedes, K.L. Ternus, S.D. Turner, S. Minot, Validation of high throughput sequencing and microbial forensics applications, *Investig. Genet.* 5 (2014) 9, doi:<http://dx.doi.org/10.1186/2041-2223-5-9>.
- [5] E.M. Volz, S.D.W. Frost, Inferring the source of transmission with phylogenetic data, *PLoS Comput. Biol.* 9 (2013), doi:<http://dx.doi.org/10.1371/journal.pcbi.1003397>.
- [6] M.R. Wilson, S.C. Weaver, R.A. Winegar, Legal, technical, and interpretational considerations in the forensic analysis of viruses, *J. Forensic Sci.* 58 (2013) 344–357, doi:<http://dx.doi.org/10.1111/1556-4029.12065>.
- [7] F. González-Candelas, M.A. Bracho, B. Wróbel, A. Moya, Molecular evolution in court: analysis of a large hepatitis C virus outbreak from an evolving source, *BMC Biol.* 11 (2013) 76, doi:<http://dx.doi.org/10.1186/1741-7007-11-76>.
- [8] D.H. Kaye, DNA evidence: probability, population genetics, and the courts, *Harv. J. Law Technol.* 7 (1993) 101–172, doi:<http://dx.doi.org/10.1163/156855207780860246>.
- [9] I.J. Wilson, M.E. Weale, D.J. Balding, Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities, *J. R. Stat. Soc. Ser. A Stat. Soc.* 166 (2003) 155–188, doi:<http://dx.doi.org/10.1111/1467-985X.00264>.
- [10] F. Taroni, S. Bozza, A. Biedermann, P. Garbolino, C. Aitken, *Data Analysis in Forensic Science: A Bayesian Decision Perspective*, John Wiley & Sons, 2010, doi:<http://dx.doi.org/10.1002/9780470665084>.
- [11] S. Gittelson, T.R. Moretti, A.J. Onorato, B. Budowle, B.S. Weir, J. Buckleton, The factor of 10 in forensic DNA match probabilities, *Forensic Sci. Int. Genet.* 28 (2017) 178–187, doi:<http://dx.doi.org/10.1016/j.fsigen.2017.02.007>.
- [12] R.E. Kass, A.E. Raftery, Bayes factors, *J. Am. Stat. Assoc.* 90 (1995) 773–795, doi:<http://dx.doi.org/10.1038/ejhg.2010.17>.
- [13] A. Nordgaard, R. Ansell, W. Drotz, L. Jaeger, Scale of conclusions for the value of evidence, *Law Probab. Risk* 11 (2012) 1–24, doi:<http://dx.doi.org/10.1093/lpr/mgr020>.

- [14] D.A. Robinson, D. Falush, E.J. Feil, *Bacterial Population Genetics in Infectious Disease*, John Wiley & Sons, 2010, doi:<http://dx.doi.org/10.1002/9780470600122>.
- [15] G. Morelli, X. Didelot, B. Kusecek, S. Schwarz, C. Bahlawane, D. Falush, S. Suerbaum, M. Achtman, Microevolution of *Helicobacter pylori* during prolonged infection of single hosts and within families, *PLoS Genet.* 6 (2010) e1001036, doi:<http://dx.doi.org/10.1371/journal.pgen.1001036>.
- [16] D.J. Wilson, E. Gabriel, A.J.H. Leatherbarrow, J. Cheesbrough, S. Gee, E. Bolton, A. Fox, C.A. Hart, P.J. Diggle, P. Fearhead, Rapid evolution and the importance of recombination to the gastroenteric pathogen *Campylobacter jejuni*, *Mol. Biol. Evol.* 26 (2009) 385–397, doi:<http://dx.doi.org/10.1093/molbev/msn264>.
- [17] A. Sjödin, K. Svensson, C. Ohrman, J. Ahlinder, P. Lindgren, S. Duodo, J. Hnath, J. P. Burans, A. Johansson, D.J. Colquhoun, P. Larsson, M. Forsman, Genome characterisation of the genus *Francisella* reveals insight into similar evolutionary paths in pathogens of mammals and fish, *BMC Genomics* 13 (2012) 268, doi:<http://dx.doi.org/10.1186/1471-2164-13-268>.
- [18] M. Achtman, Insights from genomic comparisons of genetically monomorphic bacterial pathogens, *Philos. Trans. R. Soc. B Biol. Sci.* 367 (2012) 860–867, doi:<http://dx.doi.org/10.1098/rstb.2011.0303>.
- [19] F. Tagini, G. Greub, Bacterial genome sequencing in clinical microbiology: a pathogen-oriented review, *Eur. J. Clin. Microbiol. Infect. Dis.* 36 (2017) 2007–2020, doi:<http://dx.doi.org/10.1007/s10096-017-3024-6>.
- [20] M.A. DePristo, E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, A.A. Philippakis, G. del Angel, M.A. Rivas, M. Hanna, A. McKenna, T.J. Fennell, A.M. Kernytsky, A.Y. Sivachenko, K. Cibulskis, S.B. Gabriel, D. Altshuler, M.J. Daly, A framework for variation discovery and genotyping using next-generation DNA sequencing data, *Nat. Genet.* 43 (2011) 491–498, doi:<http://dx.doi.org/10.1038/ng.806>.
- [21] H. Chen-Harris, M.K. Borucki, C. Torres, T.R. Slezak, J.E. Allen, Ultra-deep mutant spectrum profiling: improving sequencing accuracy using overlapping read pairs, *BMC Genomics* 14 (2013) 1–13, doi:<http://dx.doi.org/10.1186/1471-2164-14-96>.
- [22] B. Silverman, Density estimation for statistics and data analysis, *Monogr. Stat. Appl. Probab.* (1986) 1–22.
- [23] C.G.G. Aitken, Statistical discriminant analysis in forensic science, *J. Forensic Sci. Soc.* 26 (1986) 237–247, doi:[http://dx.doi.org/10.1016/S0015-7368\(86\)72490-0](http://dx.doi.org/10.1016/S0015-7368(86)72490-0).
- [24] D.A. Berry, I.W. Evett, R. Pinchin, Statistical inference in crime investigations using deoxyribonucleic acid profiling, *Appl. Stat.* 41 (1992) 499, doi:<http://dx.doi.org/10.2307/2348086>.
- [25] C.G.G. Aitken, D. Lucy, Evaluation of trace evidence in the form of multivariate data, *J. R. Stat. Soc. Ser. C Appl. Stat.* 53 (2004) 109–122, doi:<http://dx.doi.org/10.1046/j.0035-9254.2003.05271.x/full>.
- [26] G. Zadora, Classification of glass fragments based on elemental composition and refractive index, *J. Forensic Sci.* 54 (2009) 49–59, doi:<http://dx.doi.org/10.1111/j.1556-4029.2008.00905.x>.
- [27] G. Zadora, T. Neocleous, Likelihood ratio model for classification of forensic evidence, *Anal. Chim. Acta* 642 (2009) 266–278, doi:<http://dx.doi.org/10.1016/j.aca.2008.12.013>.
- [28] J. Ahlinder, A. Nordgaard, S.W. Lindström, Chemometrics comes to court: evidence evaluation of chem-bio threat agent attacks, *J. Chemom.* 29 (2015) 267–276, doi:<http://dx.doi.org/10.1002/cem.2699>.
- [29] Y. Chen, Y. Luo, P. Curry, R. Timme, D. Melka, M. Doyle, M. Parish, T.S. Hammack, M.W. Allard, E.W. Brown, E.A. Strain, Assessing the genome level diversity of *Listeria monocytogenes* from contaminated ice cream and environmental samples linked to a listeriosis outbreak in the United States, *PLoS One* 12 (2017) 1–19, doi:<http://dx.doi.org/10.1371/journal.pone.0171389>.
- [30] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for illumina sequence data, *Bioinformatics* 30 (2014) 2114–2120, doi:<http://dx.doi.org/10.1093/bioinformatics/btu170>.
- [31] J.T. Simpson, K. Wong, S.D. Jackman, J.E. Schein, S.J.M.M. Jones, I. Birol, ABySS: a parallel assembler for short read sequence data, *Genome Res.* 19 (2009) 1117–1123, doi:<http://dx.doi.org/10.1101/gr.089532.108>.
- [32] B.J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C.A. Cuomo, Q. Zeng, J. Wortman, S.K. Young, A.M. Earl, Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement, *PLoS One* 9 (2014), doi:<http://dx.doi.org/10.1371/journal.pone.0112963>.
- [33] A.E. Darling, B. Mau, N.T. Perna, progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement, *PLoS One* 5 (2010) e11147.
- [34] J.J. Astrin, H. Höfer, J. Spelda, J. Holstein, S. Bayer, L. Hendrich, B.A. Huber, K.H. Kielhorn, H.J. Krammer, M. Lemke, J.C. Monje, J. Moriniere, B. Rulik, M. Petersen, H. Janssen, C. Muster, Towards a DNA barcode reference database for spiders and harvestmen of Germany, *PLoS One* 11 (2016) 1–24, doi:<http://dx.doi.org/10.1371/journal.pone.0162624>.
- [35] R. Bouckaert, J. Heled, D. Kühnert, T. Vaughan, C.H. Wu, D. Xie, M.A. Suchard, A. Rambaut, A.J. Drummond, BEAST 2: a software platform for bayesian evolutionary analysis, *PLoS Comput. Biol.* 10 (2014) 1–6, doi:<http://dx.doi.org/10.1371/journal.pcbi.1003537>.
- [36] J.G. Sumner, P.D. Jarvis, J. Fernández-Sánchez, B.T. Kaine, M.D. Woodhams, B.R. Holland, Is the general time-reversible model bad for molecular phylogenetics? *Syst. Biol.* 61 (2012) 1069–1074, doi:<http://dx.doi.org/10.1093/sysbio/sys042>.
- [37] G. Yu, D.K. Smith, H. Zhu, Y. Guan, T.T.Y. Lam, Ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data, *Methods Ecol. Evol.* 8 (2017) 28–36, doi:<http://dx.doi.org/10.1111/2041-210X.12628>.
- [38] W. Hesselbrock, L. Foshay, The morphology of bacterium *tularensis*, *J. Bacteriol.* 49 (1945) 209–231.
- [39] C. Dwibedi, P. Lindgren, K. Myrtenäs, M. Granberg, E. Lundmark, C. Öhrman, A. Sjödin, P. Stenberg, J. Ahlinder, M. Forsman, P. Larsson, A. Johansson, Biological amplification of low frequency mutations for bacterial source attribution, in: C. Dwibedi (Ed.), *Francisella tularensis: Persistence, Dissemination and Source Attribution: a Theoretical and Computational Approach* [Dissertation], Umeå University, 2019 URN: urn:nbn:se:umu:diva-156138.
- [40] J. Oksanen, F.G. Blanchet, R. Kindt, P. Legendre, P. Minchin, G. Simpson, P. Solymos, M.H.H. Stevens, H. Wagner, Vegan: Community Ecology Package. R Package Version 2.0-7 Online publication, (2013) .
- [41] H. Wickham, ggplot2, Springer New York, New York, NY, 2009, doi:<http://dx.doi.org/10.1007/978-0-387-98141-3>.
- [42] J. Hallander, P. Waldmann, The effect of non-additive genetic interactions on selection in multi-locus genetic models, *Heredity (Edinb)* 98 (2007) 349–359, doi:<http://dx.doi.org/10.1038/sj.hdy.6800946>.
- [43] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, *Nat. Methods* 9 (2012) 357–359, doi:<http://dx.doi.org/10.1038/nmeth.1923>.
- [44] D.C. Koboldt, Q. Zhang, D.E. Larson, D. Shen, M.D. McLellan, L. Lin, C.A. Miller, E. R. Mardis, L. Ding, R.K. Wilson, VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing, *Genome Res.* 22 (2012) 568–576, doi:<http://dx.doi.org/10.1101/gr.129684.111>.
- [45] A. Tarasov, A.J. Vilella, E. Cuppen, I.J. Nijman, P. Prins, Sambamba: fast processing of NGS alignment formats, *Bioinformatics* 31 (2015) 2032–2034, doi:<http://dx.doi.org/10.1093/bioinformatics/btv098>.
- [46] L. Sandegren, D.I. Andersson, Bacterial gene amplification: implications for the evolution of antibiotic resistance, *Nat. Rev. Microbiol.* 7 (2009) 578–588, doi:<http://dx.doi.org/10.1038/nrmicro2174>.
- [47] B.H. Good, M.J. McDonald, J.E. Barrick, R.E. Lenski, M.M. Desai, The dynamics of molecular evolution over 60,000 generations, *Nature* 551 (2017) 45–50, doi:<http://dx.doi.org/10.1038/nature24287>.
- [48] M.J. Whaley, S.J. Joseph, A.C. Retchless, C.B. Kretz, A. Blain, F. Hu, H.-Y. Chang, S. A. Mbaeyi, J.R. MacNeil, T.D. Read, X. Wang, Whole genome sequencing for investigations of meningococcal outbreaks in the United States: a retrospective analysis, *Sci. Rep.* 8 (2018) 15803, doi:<http://dx.doi.org/10.1038/s41598-018-33622-5>.