AB_SA: Tracing the source of bacterial strains based on accessory genes. Application to *Salmonella* Typhimurium environmental strains

October 31, 2019

Author names and affiliations

Laurent Guillier¹, Michèle Gourmelon², Solen Lozach², Sabrina Cadel-Six¹, Marie-Léone Vignaud¹, Nanna Munck³, Tine Hald³, Federica Palma¹

¹Anses, University of Paris-EST, Anses, Laboratory for food safety, Maisons-Alfort, France
²Ifremer, RBE, SGMM, Health, Environment and Microbiology Laboratory, Plouzané, France
³Research Group for Genomic Epidemiology, National Food Institute, Technical University of Denmark, Kgs. Lyngby, Denmark

Correspondance

Laurent Guillier laurent.guillier@anses.fr

Abstract

The partitioning of pathogenic strains isolated in environmental or human cases to their original source is challenging. The pathogens usually colonize multiple animal hosts, including livestock, which contaminate food-producing and environment (e.g. soil and water), posing additional public health burden and major challenges in the identification of the source. Genomic data opens new opportunities for the development of statistical models aiming to infer the likely source of pathogen contamination. Here, we propose a computationally fast and efficient multinomial logistic regression (MLR) source attribution classifier to predict the animal source of bacterial isolates based on “source-enriched” loci extracted from the accessory-genome profiles of a pangenomic dataset. Depending on the accuracy of the model’s self-attribution step, the modeler selects the number of candidate accessory genes that better fit the model for calculating the likelihood of (source) category membership. The accessory genes-based source attribution (AB_SA) method was applied on a dataset of strains of *Salmonella* Typhimurium and its monophasic variants (S. 1,4,[5],12:i:-). The model was trained on 69 strains with known animal source categories (i.e., poultry, ruminant, and pig). The AB_SA method helped to identify eight genes as predictors among the 2,802 accessory genes. The self-attribution accuracy was 80%. The AB_SA model was then able to classify
25 over 29 S. Typhimurium and S. 1,4,[5],12:i:- isolates collected from the environment (considered as unknown source) into a specific category (i.e., animal source), with more than 85% of probability. The AB_SA method herein described provides a user-friendly and valuable tool to perform source attribution studies in few steps. AB_SA is written in R and freely available at https://github.com/lguillier/AB_SA.

**Keywords**
Source attribution, *Salmonella* Typhimurium, environmental contamination, pangenome-wide enrichment analysis, multinomial logistic regression

**Author Notes**
All supporting data, code, and protocols have been provided within the article and through supplementary data files.
Supplementary material is available with the online version of this article.

**Abbreviations**
AB_SA, accessory-based source attribution; MLR, multinomial logistic regression; SNPs, single nucleotide polymorphisms; GFF, general feature format; AIC, Akaike information criteria.

**Data Summary**
1. The AB_SA model is written in R, open-source and freely available Github under the GNU GPLv3 licence (https://github.com/lguillier/AB_SA).
2. All sequencing reads used to generate the assemblies analyzed in this study have been deposited in the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena) under project number PRJEB16326. Genome metadata and ENA run accession ID for all the assemblies are reported in the supplementary material.

**Impact Statement**
This article describes AB_SA (“Accessory-Based Source Attribution method”), a novel approach for source attribution based on “source enriched” accessory genomics data and unsupervised multinomial logistic regression. We demonstrate that the AB_SA method enables the animal source prediction of large-scale datasets of bacterial populations through rapid and easy identification of source predictors from the non-core genomic regions. Herein, AB_SA correctly self-attribute the animal source of a set of S. Typhimurium and S. 1,4,[5],12:i:- isolates and further classifies the 84% of strains contaminating natural environments in the pig category (with high probability ranging between ~85 and ~99%).
Introduction

Tracing the origin of pathogenic microbial strains associated with human diseases or contamination of environmental settings is crucial for identifying targets for interventions in the food production chain from farm-to-fork. The process of estimating the probability that human cases or environmental contamination cases arise from putative sources of infection (i.e. animal reservoir, food product, and environmental) can be referred to as source attribution. A variety of methodological approaches has been developed for source attribution of foodborne pathogens: epidemiological approaches (e.g., outbreak data analysis, case-control/cohort studies), microbial subtyping methods, comparative exposure assessment, intervention studies, and expert elicitations (1; 2).

Some of the source attribution methods based on microbial sub-typing specifically consider genotypic data (3; 1). Genetic variations in microorganisms are the result of different evolutionary forces. These can be prompted by either neutral processes (genetic drift) or adaptive processes, such as the emergence of a competitively advantageous mutation in a given environment. Most bacterial populations are structured, i.e., their entirety does not form a genetically homogeneous unit, but rather consists of several distinct lineages or sub-lineages that are totally or partially isolated from one another. Factors such as geographical isolation, combined with random phenomena of genetic drift and sometimes with local adaptation, drives genetic differentiation.

When analysing genetic data for a given microorganism based on targets like a certain number and type of alleles, microsatellites, or single nucleotide polymorphisms (SNPs), the objective is often to detect whether these microorganisms are structured into different subpopulations, and if so, to identify the number of clusters, the strains composing them, and possible recombination events. Historical methods for studying the genetic relatedness of microbial populations are based on the reconstruction of phylogenetic trees from a matrix of proximities for each pair of strains, with these genetic proximities being typically calculated using the methods proposed by (4) or (5). Once the tree is built, it can be ‘cut’ at a certain point (e.g., after three levels of nodes from the root) to define the different clusters of strains (more or less equivalent to sub-types). Visually exploring the composition of the clusters (i.e., isolates from different backgrounds) provides a general overview for inferring sources and transmission. However, this approach has been rarely applied in source attribution as inference by phylogeny relies upon the robustness of the tree built on the genetic diversity between isolates, and strains to attribute and strains from sources are usually phylogenetically intermixed (6). Indeed, closely related strains can be found in multiple hosts challenging the association to a specific source by phylogenetic clustering (7). A particular case showing the utility of phylogenetic methods in the attribution of human salmonellosis to specific sources (e.g., turkey), by using epidemiological and genomic data, has been reported through the investigation of *Salmonella* Derby genetic diversity (8).

A much different approach relies on the assumption that genetic data (e.g., frequency of different allele numbers at a locus) can be explained by a probabilistic model whose parameters are unknown. Comparing genetic data (frequencies) among different strain populations make it possible to establish a link between them, e.g., between strains from human cases and different sources. Two structured population genetics models that are currently widely used for source attribution of foodborne diseases are the so-called STRUCTURE approach (9) and the Asymmetric Island Model (AIM) (10). These two models are based on different principles of genetic structuring of microbial populations, but the overall attribution approach is similar. These approaches have been successfully applied for source attribution of human sporadic...
strains for *Campylobacter* spp. (11; 12), *Salmonella* spp. and *Listeria monocytogenes* (13).

Machine learning approaches are gaining interest in identifying the underlying genetic features associated with traits of microbial pathogens (14) and their use is also discussed for tracing the origin of an outbreak (15). For source attribution, recent studies consider this approach for predicting the source of sporadic human cases (16; 17; 18). In particular, (17) applied a Random Forest classifier for genomic source prediction. The authors revealed that 50 key genetic features were sufficient for robust source prediction of strains from *Salmonella* Typhimurium serovar. Most of the genetic features were accessory genes.

Among other supervised classification techniques, multinomial logistic regression (MLR) is a multi-class classification procedure. It is an extension of binary logistic regression allowing for more than two outcomes events. MLR has proven to provide a pertinent framework to carry out association analyses across multiple phenotypic traits (19; 20) and in foodborne outbreak investigations as a rule-out tool (21). Complex phenotypes, such as host adaptation in specific niches, have been linked to the presence of genes and genetic elements in some strains but not in others (referred as the “accessory genome”), mainly driven by horizontal DNA transfer (22; 7; 23). Association analysis on a pangenome scale may relate patterns of genotypic variation (e.g., the differential composition of accessory genes of multi-host lineages) to specific zoonotic niches. The objective of this study was therefore to study the performance of multinomial logistic regression in source attribution. In particular, the method was used for predicting to which animal reservoir will environmental strains of *Salmonella* Typhimurium (S. Typhimurium) and its monophasic variants (S. 1,4,[5],12:i:-) be attributed, given the variable set of source-enriched genes from the strains’ accessory genome.

**Materials and Methods**

**Preliminary genomic analysis**

High-quality assemblies of 98 bacterial isolates (see Application to *Salmonella* Typhimurium and its monophasic variant genomes dataset” below) were generated by DTU FoodQCPipeline (https://bitbucket.org/RolfKaas/foodqcpipeline). In short, the FoodQCPipeline trimmed the raw reads using bbduk2. Reads were then de novo assembled using SPAdes v3.11.05 (24) in the last step of the pipeline. FastQC v0.11.5 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was applied in multiple steps of reads processing (e.g., before and after trimming), generating a quality control report for each sample. The quality of the de novo assemblies was finally assessed using Quast v4.5 (25). Besides, the maximum-likelihood phylogenetic reconstruction of the 98 genomes dataset was built on single-nucleotide polymorphisms (SNPs) identified in the core genome alignments to assess the applicability of the dataset in (26). The annotated tree shows that environmental isolates (i.e. unknown source) were intermixed with potential sources and the dataset is therefore eligible for source attribution.
Accessory Based Source attribution (AB-SA) method

The Accessory Based Source attribution (AB-SA) method is based on genomics data. The method is a two-step process. First, the accessory genes enriched in the different sources are calculated (see “Preparation of input data for multinomial logistic model pangenome analysis” below). Then, a multinomial logistic regression is developed to predict the probability of animal sources membership on environmental isolates based on “source-enriched” accessory genes.

Multinomial regression is thus used to explain the relationship between one nominal dependent variable (with more than two levels), that is the source, and one or more independent variables, i.e. the enriched genes. For a source attribution situation with K sources, the multinomial regression model estimates k \((k = K - 1)\) equations:

\[
\text{Prob}(\text{Source} = 1) = \frac{\exp(\beta_{10} + \beta_{11}\text{Gene1} + \beta_{12}\text{Gene2} + \ldots + \beta_{1N}\text{GeneN})}{1+\exp(\beta_{10} + \beta_{11}\text{Gene1} + \beta_{12}\text{Gene2} + \ldots + \beta_{1N}\text{GeneN})} \quad (1)
\]

\[
\text{Prob}(\text{Source} = 2) = \frac{\exp(\beta_{20} + \beta_{21}\text{Gene1} + \beta_{22}\text{Gene2} + \ldots + \beta_{2N}\text{GeneN})}{1+\exp(\beta_{20} + \beta_{21}\text{Gene1} + \beta_{22}\text{Gene2} + \ldots + \beta_{2N}\text{GeneN})} \quad (2)
\]

\[\ldots\]

\[
\text{Prob}(\text{Source} = k) = \frac{\exp(\beta_{0k} + \beta_{1k}\text{Gene1} + \beta_{2k}\text{Gene2} + \ldots + \beta_{Nk}\text{GeneN})}{1+\exp(\beta_{0k} + \beta_{1k}\text{Gene1} + \beta_{2k}\text{Gene2} + \ldots + \beta_{Nk}\text{GeneN})} \quad (3)
\]

For the final source, the probability of association is derived from the K-1 other equations:

\[
\text{Prob}(\text{Source} = K) = 1 - (\text{Prob}(\text{Source} = 1) + \text{Prob}(\text{Source} = 2) + \ldots + \text{Prob}(\text{Source} = k)) \quad (4)
\]

Preparation of input data for multinomial logistic model pangenome analysis

The genomes assemblies were annotated in GFF3 format by using Prokka (Seemann, 2014). Roary (27) was used to determine the pangenome of the whole dataset of strains from the annotated assemblies. The dataset included both strains with known (i.e., animal reservoir) and unknown sources (i.e., environmental strains to attribute). To know which genes were enriched in each host groups, Scoary was used (28). Scoary takes as input the gene\_presence\_absence.csv file from Roary and a traits file reporting the source associated with each strain. Notice that the traits file is restricted to strains with known sources. The --no\_pairwise and --collapse options from Scoary were applied to determine the genes that are enriched in each source. The --no\_pairwise flag is used for enrichment analysis to avoid pairwise comparisons and perform population structure-naive calculation. The --collapse flag was used to merge genes that present the same pattern of distribution in the sources (a single gene, the first of each merged group of genes is then taken into account by the AB-SA method). Besides, the naive p-value is used to show the genes most overrepresented in a specific source. A naive threshold p-value of 0.01 was used to establish the list of potential genes of interest for the attribution of the source, sorted by strength of association per trait. To reduce or increase the number of genes to be considered by the AB-SA method, the user can modify the threshold p-value.

The CreateInputMNL function from the AB-SA method creates the input files for multinomial logistic model. It takes as arguments: the trait file used as input of Scoary and the gene\_presence\_absence .Rtab from Roary, and the number of enriched genes to be taken in each source. The function returns two files: a
file for fitting multinomial logistic model on data originating from source, and a file used for determining
the probability of association the source of unknown strains (sporadic human strains or environmental
strains).

Training and testing of the multinomial logistic model

Association of sources and genes are then carried out by the multinomial model built on a combination
of functions from the “caret”, “e1071” and “ROCR” R packages (29).

A split-sample approach was applied for training and testing the model using the dataset with known
sources to select model tuning values and estimate the model performance through resampling. The
dataset with known sources was randomly partitioned into complementary subsets in a ratio of 70:30
meaning that 70% of data will be used for model training and the remaining 30% for model testing
(evaluating model performance) by (K-fold and bootstrapping) resampling method (30). The resampling
method creates modified datasets of samples from the training sets and a model is fit to each resampled
dataset to predict the corresponding set of hold-out samples. The aggregation of the results of each
hold-out sample set is then used to estimate the resampling performance for finally assessing the more
appropriate combination of tuning parameters to consider for the final model refit on the entire dataset.
The training and testing are carried by MNLTrainTest function in AB_SA. It takes as inputs the output
file from CreateInputMNL function, the partition percentage (70:30 as default value) as well as the
number of bootstrap.

Assessment of the model’s performance

Trained models were assessed through accuracy metrics to select the optimal model. To avoid overfitting,
the assessment of fitted logistic multinomial models was carried out with regularization. The Akaike
Information Criteria (AIC) was used as a statistical measure of fit to penalize the number of parameters
(predictor variables) included in the multinomial logistic model, helping to identify too complex models
that tend to memorize training data. The AB_SA MNLTrainTest function returns both the density plot
showing the performance estimates of accuracy and the AIC value.

Prediction of strains with unknown source

Based on the accuracy of the trained model and the AIC values, the appropriate number of genes
to include in the multinomial logistic model is selected by the modeler. The full set of strains with
known sources is fitted with MNLfit function for these genes and samples with unknown source are then
predicted using the AB_SA MNLpredict function, which returns a data frame with probability values
for each animal source.
Application to *Salmonella* Typhimurium and its monophasic variant genomes dataset

The dataset used to explore the relationship between genes of the accessory genome and the animal host is constituted by 98 strains belonging to *S*. Typhimurium and its monophasic variant and collected between 2010-2015. The dataset is fully described elsewhere (26) and reported in the Supplementary Material. The dataset is constituted of strains from known sources and strains with an unknown animal source (i.e. strains to attribute). The set of isolates with a known source comprises strains isolated from pigs (n=49), poultry (layers, broilers, turkeys, and ducks) (n=14) and ruminants (cattle, sheep, and goat) (n=6). For the strains to attribute, 29 strains of *Salmonella* were isolated from the environment (e.g., fresh or brackish water and soil) from both ANSES and IFREMER with the collaboration of the University of Caen (France). The IFREMER’s environmental isolates originate from a research project (31). They were isolated from freshwater (n=12) in Brittany and brackish water in Normandy (n=3). ANSES environmental isolates were isolated from soils (n=3) and freshwater (n=10). One strain isolated in crustacean was also associated with this environmental dataset (26).

Results

Pangenome-wide enrichment analysis of the *Salmonella* dataset

In total, 98 *S*. Typhimurium and monophasic variant strains were used in this study as input for implementing a multinomial logistic regression model of source attribution. Over 98 strains, 29 were isolated from environmental (i.e., water, soil samples, and a crab isolate) while the remaining 69 from animal sources (i.e., pigs, poultry, and ruminants). The whole dataset of genomes was used to extract the pangenome while only the genomes from animal sources were used to score accessory genes as enriched in each animal source. Over 6,988 genes composing the pangenome, 40% (n=2,802) represented the accessory genome (present in <99% of strains) and were then considered for the enrichment of genes in the animal sources. With a naive p-value <0.006, 10 genes were retained as enriched in the selected animal sources (Fig. 1, available in the Supplementary Material). However, a cluster of four correlated genes was collapsed into a merged unit. Only the first gene of the merged unit together with nine additional “source-enriched” genes will be considered as predictors by the AB-SA model. Most of the candidates’ genes are specifically enriched in a specific animal source (i.e. ruminants (n=3), pigs (n=2) and poultry (n=2)), while the remaining (n=3) are enriched in multiple sources (e.g., pigs and poultry).

Feeding the MLR model with the source-enriched genes, a maximum number of genes to consider for predicting the source is arbitrarily selected. In order to select the optimal set of predictors, different numbers of genes (from 1 to 5) were tested, and for each case, accuracy and AIC were assessed (Table 1).
Assessment of multinomial logistic model to predict the strains with known origins

For further performing an accurate animal source prediction, it is necessary to select the genes set that better discriminates pigs, poultry and ruminant-related genomes. When testing the ability to classify strains with known source by randomly selecting 70% and 30% of genomes for training and testing respectively, accuracy ranged from 0.71 to 0.82 according to the different genes included in the model (Table 1). Yet confidence intervals of the accuracies are large, and they could be considered as equivalent. AIC values help to distinguish the best model among the tested ones. In this study, the model including a total of eight genes as predictors (Table 1) was found to be the best model (with the lowest AIC value). This set of “best” predictive genes is therefore used by the AB-SA model to classify genomes with unknown source. The relative importance of each predictor in estimating the model is calculated (Fig. 2). This statistical measure relates to the weight of each predictor in making a prediction, not whether or not the prediction is accurate. Fig. 2 presents the values of fitted $\beta_k$ parameters. It shows that some genes have a higher weight than others. For example, group 6195 presence is strongly associated with bovine. In the same way, group 852 represents the highest coefficients for poultry.

Prediction of the origin of environmental strains

Fig. 3 shows the probabilities of the different environmental strains to be associated with one of the three sources. Six strains (i.e., 9, 12, 14, 25, 28, and 29) have a very high membership probability, that is, superior to 99%, to one of the three sources. The majority of the strains (n=19) has a high probability, ranging between 85-95%, of being associated to pig sources. For the four remaining strains (i.e., 5, 7, 11, and 24), the probability to be associated with a specific source is lower than 80% (e.g., ranging from ~39% to ~77%). Fig. 4 shows the global attribution of the whole dataset of environmental strains to the three animal sources. For this dataset, pig source was found to be the main source of the strains isolated in the environment.
Discussion

Multinomial model for (WGS)-based source attribution

The environment is not a natural reservoir of *Salmonella* Typhimurium. This study thus focused on attribution of 29 *Salmonella* strains isolated from the environment (i.e. river and brackish water, soil, and crab) to potential sources using accessory based genes. The animal sources were grouped into three major categories (i.e., ruminants, poultry, and pigs) constituted by a dataset of 69 isolates of *Salmonella* Typhimurium and its monophasic variants. Genetic factors predictive of animal sources were identified in the accessory genome of isolates from this dataset through an innovative workflow based on enrichment and multinomial logistic analysis. After the selection of source-enriched genes as predictors, a maximum likelihood estimation is used by the AB_SA model to assess the probability of belonging to each animal source (categorical membership) for a given environmental isolate. The MLR-based models are less sensitive to data assumptions (e.g., normality, linearity and homogeneity of data) and have the advantage of limitation of overfitting, which is a common pitfall in machine learning approaches (32). It occurs when a complex model is trained on too few data points and becomes specific to the training data. As AB_SA returns both accuracy and AIC value (a penalty for model complexity), overfitting is prevented.

In addition, it is flexible in the choice of the level of significance (through threshold value for p-value) of enriched genes and in the number of candidate predictors per source to feed the multinomial logistic model. The MLR model also provides a measure of the weight of each predictor (i.e., gene) although the interpretation of the coefficients is not immediate. The interpretation might be further complicated by not having a single model, but as many models as the number of sources minus one.

Consistent with previously published source attribution studies, the overall self-attribution accuracy of this study was 75-80% (17; 18; 12).

Number of genes predictive of animal source

The enrichment step of the AB_SA method allowed us to select the most relevant genes for modeling their effects with the multinomial logistic model. Within the 2,802 accessory genes, eight genes were finally selected in the best model for the prediction of the origin of environmental strains. The more predictors used in the training phase improved the model performance. However, using more than eight genes as predictors returns a worse AIC value in this dataset since the use of more data training leads to over-fitting in regression. This study confirms that high dimensional input data for source attribution models do not guarantee high performance. For example, (13) didn’t observe for *L. monocytogenes* an important improvement of accuracy of source attribution models when passing from 7-locus MLST to cgMLST.

Source prediction of strains using accessory genes as predictors

In this study, eight genomic factors making up the ancillary genomes of the observed *Salmonella* dataset were selected as predictive of animal sources (Table 2, sequences available in the Supplementary Material).
by the AB\_SA method. Although the presence of several hypothetical proteins, the selected genes had different putative functions including structural function (e.g., cfiA gene involved in membrane protein catalysis for ATP synthesis, transport and motility) (33) as well as DNA packaging and lysis (e.g., DNA braking-rejoining protein, lysozyme, prophage tail assembly protein) (Table 2).

Interestingly, the majority (n=7/8) of these predictors were located in mobile genetic elements regions such as putative prophages and plasmid. The presence of unique accessory genomic biomarkers within prophage regions has been recently shown to contribute to enhancing the ability of tracing-back the origin of epidemic clones of S. Typhimurium and its monophasic variant on a geographical scale (34). Here, half of the animal source genomic markers identified in the assessment of the probability of correct self-attrition were prophage-related genes, suggesting that prophage elements might play a crucial role also in source-tracking studies. As observed by Zhang et al. (17), plasmid and prophage-related loci may constitute highly informative predictors of livestock sources of S. Typhimurium and therefore contribute to the optimization of source attribution models for surveillance or outbreak investigations.

The importance of accessory genes in host adaptation is not limited to Salmonella. Recent studies focused on the pangenome of Campylobacter jejuni showed that host-segregating genomic factors located in the accessory genomes constitute epidemiological markers for source attribution (12; 35).

Source of the Salmonella spp. environmental strains

Salmonella strains are frequently detected in surface water samples, e.g. 30.1% of samples from three French coastal catchments (31), 43% of samples from Georgia (USA) (36) and 23% of samples from Canada (37). Among Salmonella serovars, Salmonella Typhimurium and its monophasic variants were isolated more frequently in some sites than others (31). Pigs, poultry, and ruminants (e.g., cattle) constitute a relevant source of these pathogens acting as natural Salmonella reservoirs and without showing any symptoms while shedding into the environment (38). In this study, pig source was found to be the most likely source of contamination of most of the S. Typhimurium and S. 1,4,[5],12:i:- strains isolated in the environment, confirming the results from several sources tracking studies that pointed out some associations between environmental Salmonella spp. strains and pig production (39; 37). This is also consistent with the fact that pigs are asymptomatic carriers of S. Typhimurium generating a major reservoir for Salmonella contaminating the environment around the primary production sites (40; 41).

Although waterborne outbreaks of Salmonella have been rarely reported, a secondary waterborne outbreak has been linked to S. 1,4,[5],12:i:- and the leaching of animal faecal matter into groundwater destined for human consumption in a rural community (42). In addition, anthropological activities such as fertilization with animal manure and irrigation with water contaminated by livestock and/or wildlife faecal matter seem to be likely involved in Salmonella outbreaks and consumption of fresh produce (43; 44). Gaining insights into the animal sources of strains contaminating natural environments is therefore of great importance for providing evidence to support targeted interventions and policy development to reduce the public health risk.
Author statements

Funding information
This work was supported by the project ‘COllaborative Management Platform for detection and Analyses of (Re-) emerging and foodborne outbreaks in Europe’ (COMPARE) which has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 643476.

Author contributions
LG, MG, NM, TH, and FP conceived the study. SCS contributed to the selection of the data. SL and MLV performed the pre-sequencing steps. LG conceptualized algorithms. LG and FP implemented scripts and drafted the manuscript. All authors commented and approved the final manuscript, take public responsibility for appropriate portions of the content and agree to be accountable for all aspects of the work in terms of accuracy or integrity.

Conflict of interest
The authors declare that there are no conflicts of interest.

Ethical approval
This study only included previously sequenced and published data. No new samples were collected for this study.

References
1. Mughini-Gras L, Franz E, Pelt W van. New paradigms for Salmonella source attribution based on microbial subtyping. Food Microbiology. May 2018;71:60–7.
2. Authority EFS. Overview of methods for source attribution for human illness from food-borne microbiological hazards-Scientific Opinion of the Panel on Biological Hazards. Efsa Journal. 2008;6(7):764.
3. Barco L, Barrucci F, Olsen JE, Ricci A. Salmonella source attribution based on microbial subtyping. International Journal of Food Microbiology. May 2013;163(2-3):193–203.
4. Nei M, Tajima F, Tateno Y. Accuracy of estimated phylogenetic trees from molecular data. Journal of molecular evolution. 1983;19(2):153–70.
5. Reynolds J, Weir BS, Cockerham CC. Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics. 1983;105(3):767–79.
6. Mather AE, Vaughan TG, French NP. Molecular Approaches to Understanding Transmission and Source Attribution in Nontyphoidal Salmonella and Their Application in Africa.. Clin Infect Dis. 2015;61 Suppl 4:S259–65.
7. Sheppard SK, Guttman DS, Fitzgerald JR. Population genomics of bacterial host adaptation.. Nat Rev Genet. 2018;19:549–65.
8. Sévellec Y, Felten A, Radomski N, Granier SA, Le HS, Petrovska L, et al. Genetic Diversity of Salmonella Derby from the Poultry Sector in Europe. Pathogens. 2019;8.

9. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945–59.

10. Wilson DJ, Gabriel E, Leatherbarrow AJH, Cheesbrough J, Gee S, Bolton E, et al. Tracing the Source of Campylobacteriosis. Guttmann DS, editor. PLoS Genetics. September 2008;4(9):e1000203.

11. Liao S-J, Marshall J, Hazelton ML, French NP. Extending statistical models for source attribution of zoonotic diseases: a study of campylobacteriosis. Journal of The Royal Society Interface. January 2019;16(150):20180534.

12. Thépault A, Méric G, Rivoal K, Pascoe B, Mageiros L, Touzain F, et al. Genome-Wide Identification of Host-Segregating Epidemiological Markers for Source Attribution in Campylobacter jejuni. Applied and Environmental Microbiology. January 2017;83(7).

13. Nielsen EM, Björkman JT, Kil K, Grant K, Dallman T, Painset A, et al. Closing gaps for performing a risk assessment on Listeria monocytogenes in ready-to-eat (RTE) foods: activity 3, the comparison of isolates from different compartments along the food chain, and from humans using whole genome sequencing (WGS) analysis. Elsa Supporting publication. 2017:EN-1151:170.

14. Njage PMK, Henri C, Leekitcharoenphon P, Mistou M-Y, Hendriksen RS, Hald T. Machine Learning Methods as a Tool for Predicting Risk of Illness Applying Next-Generation Sequencing Data. Risk Analysis. November 2019;39(6):1397-1413.

15. Wheeler NE. Tracing outbreaks with machine learning. Nature Reviews Microbiology. February 2019;17(5):269–.

16. Munck NSM, Njage PMK, Litrup E, Hald T. Trends and sources in human salmonellosis [Internet]. http://orbit.dtu.dk/files/154732546/Rapport_Annual_Report_on_Zoonoses_2017.pdf; 2018. Available at: http://orbit.dtu.dk/files/154732546/Rapport_Annual_Report_on_Zoonoses_2017.pdf

17. Zhang S, Li S, Gu W, Bakker H den, Boxrud D, Taylor A, et al. Zoonotic Source Attribution of Salmonella enterica Serotype Typhimurium Using Genomic Surveillance Data United States. Emerging Infectious Diseases. 2019;25(1).

18. Lupolova N, Dallman TJ, Holden NJ, Gally DL. Patchy promiscuity: machine learning applied to predict the host specificity of Salmonella enterica and Escherichia coli. Microbial Genomics. October 2017;3(10).

19. Jostins L, McVean G. Trinculo: Bayesian and frequentist multinomial logistic regression for genome-wide association studies of multi-category phenotypes. Bioinformatics [Internet]. February 2016;32(12):1898–900. Available at: https://doi.org/10.1093%2Fbioinformatics%2Fbtw075

20. Serra A, Galdi P, Tagliaferri R. Machine learning for bioinformatics and neuroimaging. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery [Internet]. February 2018;8(5):e1248. Available at: https://doi.org/10.1002%2Fwidm.1248

21. White A, Cronquist A, Bedrick EJ, Scallan E. Food Source Prediction of Shiga Toxin-Producing Escherichia coli Outbreaks Using Demographic and Outbreak Characteristics, United States, 1998-2014.
22. Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. Curr Opin Microbiol. 2015;23:148–54.

23. Richardson EJ, Bacigalupe R, Harrison EM, Weinert LA, Lycett S, Vriel M, et al. Gene exchange drives the ecological success of a multi-host bacterial pathogen. Nat Ecol Evol. 2018;2:1468–78.

24. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology [Internet]. 2012;19(5):455–77. Available at: https://doi.org/10.1089%2Fcmb.2012.0021

25. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics [Internet]. 2013;29(8):1072–5. Available at: https://doi.org/10.1093%2Fbioinformatics%2Fbtt086

26. Munck N, Leekitcharoenphon P, Littrup E, Kaas R, Schielke A, Guillier L, et al. Four European Salmonella Typhimurium datasets collected to develop WGS-based source attribution methods. Scientific Data. in revision;

27. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics [Internet]. 2015;31(22):3691–3. Available at: https://doi.org/10.1093%2Fbioinformatics%2Fbtv421

28. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol. November 2016;17:238.

29. Kuhn M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software [Internet]. 2008;28(5). Available at: https://doi.org/10.18637%2Fjss.v028.i05

30. Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society Series B (Methodological). 1974;36(2):111–47.

31. Rincé A, Balière C, Hervio-Heath D, Cozien J, Lozach S, Parnaudeau S, et al. Occurrence of Bacterial Pathogens and Human Noroviruses in Shellfish-Harvesting Areas and Their Catchments in France. Front Microbiol. 2018;9:2443.

32. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nature Reviews Genetics [Internet]. 2015;16(6):321–32. Available at: https://doi.org/10.1038%2Fnrg3920

33. Granjon T, Marini O, Aouchi Y, Dahidin P, Bucet R, Marcillat O, et al. Structure-Function Relations in Oxaloacetate Decarboxylase Complex. Fluorescence and Infrared Approaches to Monitor Oxomalonate and Na+ Binding Effect. PLoS ONE. 2010;5(6):e10935.

34. Palma F, Manfreda G, Silva M, Parisi A, Barker DOR, Taboada EN, et al. Genome-wide identification of geographical segregated genetic markers in Salmonella enterica serovar Typhimurium variant 4,[5],12:i:-. Scientific Reports [Internet]. October 2018;8(1). Available at: https://doi.org/10.1038%2Fs41598-018-33266-5

35. Buchanan CJ, Webb AL, Mutschall SK, Kruczkiewicz P, Barker DOR, Hetman BM, et al. A Genome-Wide Association Study to Identify Diagnostic Markers for Human Pathogenic Campylobacter jejuni Strains. Frontiers in Microbiology [Internet]. 2017;8. Available at: https://doi.org/10.3389%2Ffmicb.2017.01224
36. Vereen Jr E, Lowrance RR, Jenkins MB, Adams P, Rajeev S, Lipp EK. Landscape and seasonal factors influence Salmonella and Campylobacter prevalence in a rural mixed use watershed. Water research. 2013;47(16):6075–85.

37. Flockhart L, Pintar K, Cook A, McEwen S, Friendship R, Kelton D, et al. Distribution of Salmonella in Humans Production Animal Operations and a Watershed in a FoodNet Canada Sentinel Site. Zoonoses and Public Health. 2016;64(1):41–52.

38. Wray C, Wray A. Salmonella in domestic animals [Internet]. Wray C, Wray A, editors. CABI; 2000. Available at: https://doi.org/10.1079%2F9780851992617.0000

39. Patchanee P, Molla B, White N, Line DE, Gebreyes WA. Tracking salmonella contamination in various watersheds and phenotypic and genotypic diversity. Foodborne Pathog Dis. 2010;7:1113–20.

40. Botteldoorn N, Heyndrickx M, Rijpens N, Grijspeerdt K, Herman L. Salmonella on pig carcasses: positive pigs and cross contamination in the slaughterhouse. Journal of Applied Microbiology. 2003;95(5):891–903.

41. Bonardi S. Salmonella in the pork production chain and its impact on human health in the European Union. Epidemiology and Infection. 2017;145(8):1513–26.

42. Kozlica J, Claudet AL, Solomon D, Dunn JR, Carpenter LR. Waterborne outbreak of Salmonella I 4,[5],12:i:-. Foodborne Pathogens and Disease. 2010;7(11):1431–3.

43. Doyle MP, Erickson MC. Summer meeting 2007 – the problems with fresh produce: an overview. Journal of Applied Microbiology [Internet]. 2008;105(2):317–30. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2672.2008.03746.x

44. Nygård K, Lassen J, Vold L, Andersson Y, Fisher I, Löfdahl S, et al. Outbreak of Salmonella Thompson Infections Linked to Imported Rucola Lettuce. Foodborne Pathogens and Disease. 2008;5(2):165–73.
Tables

Table 1: Tested multinomial logistic models with the accuracy obtained with the training and the AIC values

| No genes/source | group_176 | group_763 | group_1926 | ymdB_2 | ylcG_2 | group_252 | group_6195 | group_160 | group_158 | Accuracy  | AIC      |
|-----------------|-----------|-----------|------------|--------|--------|-----------|------------|-----------|-----------|-----------|----------|
| 1               | X         |          | X          |        |        |           |            | X         |           | 0.82      | [0.67,0.90] | 92.7     |
| 2               | X         |          | X          |        |        | X         |            | X         |           | 0.82      | [0.58,0.92] | 89.5     |
| 3               | X         |          | X          |        |        | X         |            | X         |           | 0.75      | [0.64,0.91] | 91.2     |
| 4               | X         |          | X          |        |        | X         | X          | X         |           | 0.74      | [0.55,0.90] | 81.2     |
| 5               | X         |          | X          |        |        | X         | X          | X         |           | 0.71      | [0.36,0.90] | 83.6     |

Table 2: Predictors of animal sources. Number of isolates harboring the predictors with relative percentage of isolates from the different animal sources and environment and gene annotation from nucleotide and amino acid sequences obtain with Prokka and KEGG, respectively.

| Predictors | Total isolates | Pig isolates | Poultry isolates | Ruminant isolates | Environmental isolates | Prokka Annotation | KEGG protein homology |
|------------|----------------|--------------|------------------|-------------------|------------------------|-------------------|----------------------|
| group_176  | 57             | 0.58         | 0.05             | 0.04              | 0.33                   | Hypothetical protein | Putative Gifsy-2 prophage protein/DNA breaking-rejoining protein |
| group_763  | 12             | 0.33         | 0.58             | 0.08              | 0                      | rrrD - Lysozyme    | Bacteriophage Lysozyme |
| ymdB_2     | 12             | 0.17         | 0.25             | 0.33              | 0                      | Hypothetical protein | Gifsy-1 prophage tail assembly-like protein |
| group_1926 | 6              | 0            | 0.67             | 0.25              | 0.33                   | Hypothetical protein | Gifsy-2 prophage protein/DNA breaking-rejoining protein |
| group_252  | 5              | 0            | 0.6             | 0                 | 0.4                    | Hypothetical protein | Uncharacterized protein from E.coli plasmid pKM13516 |
| group_6195 | 3              | 0            | 0                | 0.67              | 0.33                   | cfaA - 2 carboxylase large subunit | Oxalacetate decarboxylase (Na+ extruding) subunit alpha |
| group_160  | 8              | 0.38         | 0                | 0.38              | 0.25                   | Hypothetical protein | Uncharacterized protein |
| group_158  | 15             | 0.33         | 0.2              | 0.27              | 0.2                    | Hypothetical protein | Uncharacterized protein |
Figure Captions

Figure 1. Number of enriched genes for the three animal sources

Figure 2. Estimates of the parameters of the multinomial logistic model. Ruminants _FR was the reference condition.

Figure 3. Individual source attribution of the 29 environmental strains of Salmonella Typhimurium to the three animal sources according to the ABSA method carried out on eight genes of the accessory genomes.

Figure 4. Global attribution of the environmental Salmonella Typhimurium strains to the three sources.
Figures

Figure 1: Number of enriched genes for the three animal sources

Figure 2: Estimates of the parameters of the multinomial logistic model. Ruminants _FR was the reference condition.
Figure 3: Individual source attribution of the 29 environmental strains of *Salmonella* Typhimurium to the three animal sources according to the ABSA method carried out on eight genes of the accessory genomes.

Figure 4: Global attribution of the environmental *Salmonella* Typhimurium strains to the three sources.