# Report on the critical and quantitative evaluation of existing and novel source attribution methods

A critical evaluation of models available to perform source attribution is a useful tool to inform choices when initiating a source attribution sampling effort or planning to analyse existing surveillance data to estimate an attribution of human disease to reservoirs of pathogens. The statistical models described to perform source attribution is a constantly growing list which began with a Bayesian method to estimate the relative contribution of human cases of Salmonellosis by matching the frequency of serotypes in animal and food reservoirs of the pathogen to those found in human cases of disease. This method was adapted in several publications to improve its estimation by adding covariates to the estimation and was applied to other pathogens and typing methods. These models that compare the frequency of categorical types of a pathogen in reservoirs with those found in humans will be referred to as “frequency matching” approaches in the present document and represent the most common approach to source attribution. A second class of models utilize the information that may be available from typing methods that measure multiple characters of an isolate and therefore allow for the estimation of a phylogenetic similarity among pathogens and subsequent cluster analysis. These methods have been applied primarily to Campylobacter typed by MLST and cgMLST and are referred to as “population genetics” source attribution models. A third class of model has been introduced in recent years that employs machine learning methods to estimate source attribution. These models employ various input data and may utilize a phylogenetic approach or an approach that compares the frequency of types in reservoirs and humans. These methods have been applied to WGS data with various subsequent classification methods and are referred to at “machine learning” source attribution models. In the current report a method known as “TRACE” for standardized model documentation and evaluation is employed to classify each model in a list of known publications produced by European experts in the field of source attribution. The method outlines a set of parameters describing model evaluation, data quality and methodological transparency that should be communicated in a publication that describes a statistical model (Grimm et al., 2014). We categorized each of the known models by these criteria an our own classifier for source attribution models described above to give the reader a clear overview of what models are available for source attribution today and to support the choice of which model to use.

## The TRACE framework

The parameters included in the descriptions in the document include the citation and class of model as well as the following TRACE framework details (taken verbatim from Grimm et al., 2014):

### Problem formulation

” The decision-making context in which the model will be used; the types of model clients or stakeholders addressed; a precise specification of the question(s) that should be answered with the model, including a specification of necessary model outputs; and a statement of the domain of applicability of the model, including the extent of acceptable extrapolations.”

### Model description

“The model, i.e. a detailed written model description. For individual/agent-based and other simulation models, the ODD protocol is recommended as standard format. For complex submodels, include concise explanations of the underlying rationale. Model users should learn what the model is, how it works, and what guided its design.”

### Data evaluation

“The quality and sources of numerical and qualitative data used to parametrize the model, both directly and inversely via calibration, and of the observed patterns that were used to design the overall model structure. This critical evaluation will allow model users to assess the scope and the uncertainty of the data and knowledge on which the model is based.”

### Conceptual model evaluation

The simplifying assumptions underlying a model’s design, both with regard to empirical knowledge and general, basic principles. This critical evaluation allows model users to understand that model design was not *ad hoc* but based on carefully scrutinized considerations.

### Implementation verification

“(1)Whether the computer code implementing the model has been thoroughly tested for programming errors, (2) whether the implemented model performs as indicated by the model description, and (3) how the software has been designed and documented to provide necessary usability tools (interfaces, automation of experiments, etc.) and to facilitate future installation, modification, and maintenance.”

### Model output verification

“(1)How well model output matches observations and (2) how much calibration and effects of environmental drivers were involved in obtaining good fits of model output and data.”

### Model analysis

“(1) How sensitive model output is to changes in model parameters (sensitivity analysis), and (2) how well the emergence of model output has been understood.”

### Model output corroboration

“How model predictions compare to independent data and patterns that were not used, and preferably not even known, while the model was developed, parametrized, and verified. By documenting model output corroboration, model users learn about evidence which, in addition to model output verification, indicates that the model is structurally realistic so that its predictions can be trusted to some degree.”

Not all TRACE parameters could be extracted from the original articles describing each model, therefore, some models only have some of the above parameters included.

## List of models and TRACE evaluation

### 1. Original Danish (Hald) model

#### Citation

A Bayesian approach to quantify the contribution of animal-food sources to human salmonellosis. Hald T, Vose D, Wegener HC, Koupeev T Risk Anal. 2004 Feb; 24(1):255-69. 00

#### Model class

Frequency matching

#### Problem formulation

To estimate directly the relative contributions (attributable proportions) of different potential sources of zoonotic infections to the human cases in settings with intensive public health surveillance

#### Model description

Bayesian frequentist model that calculates the number of domestic and sporadic cases caused by different Salmonella subtypes as a function of the prevalence of these types in the animal-food sources and the amount of food source consumed, and the differences between subtype and food sources with respect to cause infection (multiparameter prior).

#### Data evaluation

Danish animal and human surveillance used - robust and representative data.

#### Implementation verification

Model applied and adapted in multiple countries since 2004.

#### Model analysis

Model dependent on data quality - representativeness of surveillance data and discrimination power. If data are sub-optimal, model would not converge. Otherwise, data scarcity is reflected in wide uncertainty intervals.

#### Model output corroboration

Results coherent with other epidemiological evidence.

#### Other citations using the model

* The attribution of human infections with antimicrobial resistant Salmonella bacteria in Denmark to sources of animal origin. Hald T, Lo Fo Wong DM, Aarestrup FM Foodborne Pathog Dis. 2007 Fall; 4(3):313-26.
* Attribution of human Listeria monocytogenes infections in England and Wales to ready-to-eat food sources placed on the market: adaptation of the Hald Salmonella source attribution model. Little CL, Pires SM, Gillespie IA, Grant K, Nichols GL Foodborne Pathog Dis. 2010 Jul; 7(7):749-56.
* Assessing the differences in public health impact of salmonella subtypes using a bayesian microbial subtyping approach for source attribution. Pires SM, Hald T Foodborne Pathog Dis. 2010 Feb; 7(2):143-51.
* Application of Bayesian techniques to model the burden of human salmonellosis attributable to U.S. food commodities at the point of processing: adaptation of a Danish model. Guo C, Hoekstra RM, Schroeder CM, Pires SM, Ong KL, Hartnett E, Naugle A, Harman J, Bennett P, Cieslak P, Scallan E, Rose B, Holt KG, Kissler B, Mbandi E, Roodsari R, Angulo FJ, Cole D Foodborne Pathog Dis. 2011 Apr; 8(4):509-16.
* Source attribution of human Salmonella cases in Sweden. Wahlström H, Andersson Y, Plym-Forshell L, Pires SM Epidemiol Infect. 2011 Aug; 139(8):1246-53.
* Application of Molecular Typing Results in Source Attribution Models: The Case of Multiple Locus Variable Number Tandem Repeat Analysis (MLVA) of Salmonella Isolates Obtained from Integrated Surveillance in Denmark. de Knegt LV, Pires SM, Löfström C, Sørensen G, Pedersen K, Torpdahl M, Nielsen EM, Hald T Risk Anal. 2016 Mar; 36(3):571-88.

### 2. Original Dutch model

#### Citation

van Pelt W., van de Giessen A., van Leeuwen W., Wannet W., Henken A., Evers E. G., et al. (1999). Oorsprong, omvang en kosten van humane salmonellose. Deel 1. Oorsprong van humane salmonellose met betrekking tot varken, rund, kip, ei en overige bronnen. Infect. Bull. 10 240–243.

#### Model class

Frequency matching

#### Problem formulation

To estimate directly the relative contributions (attributable proportions) of different potential sources of zoonotic infections to the human cases in settings with intensive public health surveillance

#### Model description

It is a simple frequentist model that allows for direct (proportional) source attribution of zoonotic pathogens. It is based on both the within-source and between-source distributions of microbial isolates belonging to the different subtypes.

#### Data evaluation

The model was originally developed for the annual report on Salmonella source attribution based on serotyping data derived from of nationwide public health surveillance in the Netherlands (of high quality in terms of coverage, sources surveilled and number of isolates typed). Given its simplicity, the model has broad applicability (in time, region, target population, subtyping method and pathogen).

#### Conceptual model evaluation

The number of human cases of a given subtype attributable to a given source is proportional to the observed number of human cases of that subtype multiplied by the probability for that subtype of coming from that source. This probability is directly estimated from the frequencies of the subtypes in the sources. This model does not take into account differences among sources (e.g., exposure) or among subtypes (e.g., pathogenicity) in their ability to cause infection, thereby assuming an equal impact of the different sources and subtypes on the human population. The model does not allow for the attribution of subtypes identified in humans but not in sources. Human cases with subtypes present in sources will be attributed to these sources even if they are actually linked to other sources not considered in the model.

#### Implementation verification

The model is not computationally demanding and is implemented in different formats and can even been run in a spreadsheet. The model has been applied multiple times since 1999.

#### Model output verification

Being a frequentist model for direct attribution, output estimates that match the data are always transparently produced. No fit is needed (only propagation of uncertainty)

#### Model analysis

Model is highly sensitive to the quality of the data and additional parameters being included as weighting factors to account for differences among sources and types to cause human disease

#### Model output corroboration

The model has been applied on data different from the Dutch ones (New Zealand and France, for instance), to provide ‘baseline’ estimates for comparison with other (new) models

#### Other citations using the model

* Salmonella source attribution based on microbial subtyping: does including data on food consumption matter? Mughini-Gras L, van Pelt W Int J Food Microbiol. 2014 Nov 17; 191():109-15.
* The Bayesian microbial subtyping attribution model: robustness to prior information and a proposition. David JM, Guillemot D, Bemrah N, Thébault A, Brisabois A, Chemaly M, Weill FX, Sanders P, Watier L Risk Anal. 2013 Mar; 33(3):397-408
* Mullner P, Jones G, Noble A, Spencer SE, Hathaway S, French NP. Source attribution of food-borne zoonoses in New Zealand: a modified Hald model.
* Mullner P, Jones G, Noble A, Spencer SE, Hathaway S, French NP Risk Anal. 2009 Jul; 29(7):970-84.

### 3. Modified Dutch model

#### Citation

Risk factors for human salmonellosis originating from pigs, cattle, broiler chickens and egg laying hens: a combined case-control and source attribution analysis. Mughini-Gras L, Enserink R, Friesema I, Heck M, van Duynhoven Y, van Pelt W PLoS One. 2014; 9(2):e87933.

#### Model class

Frequency matching

#### Problem formulation

To estimate directly the relative contributions (attributable proportions) of different potential sources of zoonotic infections to the human cases in settings with intensive public health surveillance

#### Model description

It is a modified version of the original Dutch model that still allows for direct (proportional) source attribution of zoonotic pathogens based on both the within-source and between-source distributions of microbial isolates belonging to the different subtypes, but it also accounts for differential impact of the sources on human cases as a function of prevalence and exposure (i.e. food consumption)

#### Data evaluation

The model has been applied to Salmonella, STEC and Listeria source attribution based on both serotyping and genotyping data derived from nationwide public health surveillance systems with high quality in terms of coverage, sources surveilled and number of isolates typed. For full implementation it requires additional data on prevalence and food consumption habits. But it remains a relatively simple model with broad applicability (in time, region, target population, subtyping method and pathogen).

#### Conceptual model evaluation

The number of human cases of a given subtype attributable to a given source is proportional to the observed number of human cases of that subtype multiplied by the probability for that subtype of originating from that source. This probability is estimated directly from the frequencies of the subtypes in the sources, weighted by the overall prevalence of the pathogen in the sources and the exposure of the human population thereto (i.e. the food consumption weights, a function of the per-capita food consumption per source and the probability for these foods to be consumed raw/undercooked by the population). This model, therefore, does not take into account differences among sources subtypes (e.g., pathogenicity) like the original model. but it lifts the assumption of equal impact of the different sources on the human population by introducing parameters that reflect the ability of the sources to act as a vehicle for the pathogen. The model does not allow for the attribution of subtypes identified in humans but not in sources. Human cases with subtypes present in sources will be attributed to these sources even if they are actually linked to other sources not considered in the model.

#### Implementation verification

The model is not computationally demanding and is implemented in different formats and can even been run in a spreadsheet. The model has been applied multiple times since 2014.

#### Model output verification

Being still a frequentist model for direct attribution, output estimates that match the data are always transparently produced. No fit is needed (only propagation of uncertainty). The modifications done have been mostly based on the modified Hald model - New Zealand version. Notably, of the frequentist framework of the original Dutch model for direct attribution, its modified version go toward a more stochastic framework based on Monte Carlo simulation.

#### Model analysis

Model is highly sensitive to the quality of the data and additional parameters being included as weighting factors to account for differences among sources and types to cause human disease

#### Model output corroboration

The model has been applied on data different from the Dutch ones (Italy, for instance), to provide ‘baseline’ estimates for comparison with other (new) models

#### Other citations using the model

* Attribution of human Salmonella infections to animal and food sources in Italy (2002-2010): adaptations of the Dutch and modified Hald source attribution models. Mughini-Gras L, Barrucci F, Smid JH, Graziani C, Luzzi I, Ricci A, Barco L, Rosmini R, Havelaar AH, VAN Pelt W, Busani L Epidemiol Infect. 2014 May; 142(5):1070-82.
* Attribution of human infections with Shiga toxin-producing Escherichia coli (STEC) to livestock sources and identification of source-specific risk factors, The Netherlands (2010-2014). Mughini-Gras L, van Pelt W, van der Voort M, Heck M, Friesema I, Franz E Zoonoses Public Health. 2018 Feb; 65(1):e8-e22.
* Salmonella source attribution based on microbial subtyping: does including data on food consumption matter? Mughini-Gras L, van Pelt W Int J Food Microbiol. 2014 Nov 17; 191():109-15.
* Tracing the sources of human salmonellosis: a multi-model comparison of phenotyping and genotyping methods. Mughini-Gras L, Smid J, Enserink R, Franz E, Schouls L, Heck M, van Pelt W Infect Genet Evol. 2014 Dec; 28():251-60.
* Increase in reptile-associated human salmonellosis and shift toward adulthood in the age groups at risk, the Netherlands, 1985 to 2014. Mughini-Gras L, Heck M, van Pelt W Euro Surveill. 2016 Aug 25; 21(34):.
* Attribution of Listeria monocytogenes human infections to food and animal sources in Northern Italy. Filipello V, Mughini-Gras L, Gallina S, Vitale N, Mannelli A, Pontello M, Decastelli L, Allard MW, Brown EW, Lomonaco S. Food Microbiol. 2020 Aug;89:103433. doi: 10.1016/j.fm.2020.103433. Epub 2020 Jan 20.

### 4. Aberdeen model

#### Citation

Nielsen E. M., Björkman J. T., Kiil K., Grant K., Dallman T., Painset A., et al. (2017). 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. EFSA Support. Publ. 14:1151E–n/a.

#### Model class

Frequency matching

#### Problem formulation

Estimate the proportion of human cases of disease among several reservoirs. Maybe be applied to any typing method.

#### Model description

The proportion of human cases attributable to each source is estimated by the overlap between a source and the human isolate types expressed as a percentage.

#### Implementation verification

No verifiable implementation of the model was provided

#### Model output corroboration

Model output was compared to STRUCTURE and Dutch models which has similar results.

### 5. Modified Danish (Hald) model - New Zealand (Mullner’s) adaptation

#### Citation

* Mullner P, Jones G, Noble A, Spencer SE, Hathaway S, French NP. Source attribution of food-borne zoonoses in New Zealand: a modified Hald model.
* Mullner P, Jones G, Noble A, Spencer SE, Hathaway S, French NP Risk Anal. 2009 Jul; 29(7):970-84.

#### Model class

Frequency matching

#### Problem formulation

To estimate directly the relative contributions (attributable proportions) of different potential sources of zoonotic infections to the human cases in settings with intensive public health surveillance

#### Model description

Bayesian model modifying Hald et al. (2004). Modifications: methodology for incorporating uncertainty in the prevalence parameters; using a hierarchical model to set subtype-dependent factors (qi) as random effects from a non-informative a priori distribution where only the first two moments (hyper-parameters) of the distribution are estimated; and exponential priors to set source-specific parameters; approaches developed to include potentially pathogenic subtypes and to avoid food consumption weights when considering environmental sources; splitting data into time periods (using time as a third dimension).

#### Data evaluation

Data for campylobacteriosis source attribution were generated by a sentinel surveillance study for C. jejuni in the Manawatu region of New Zealand conducted between 2005 and 2008. For Salmonella, nationwide data included subtyping information for human cases, which was of good quality, but it was poor for the sources. Thus, in absence of large-scale high-quality surveillance data, information from different studies was used to estimate the prevalence of the different Salmonella types in the food sources. These included data provided by routine surveillance as well as small-scale surveys.

#### Implementation verification

Model has been applied successfully in other countries.

#### Model output verification

Modifications to the original Hald model have allowed for implementation of the model in countries with less robust surveillance data. Results coherent with other epidemiological evidence.

#### Model analysis

Modifications to the original Hald model have allowed for implementation of the model in countries with less robust surveillance data.

#### Model output corroboration

Results coherent with other epidemiological evidence.

#### Other citations using the model

* Attribution of human Salmonella infections to animal and food sources in Italy (2002-2010): adaptations of the Dutch and modified Hald source attribution models. Mughini-Gras L, Barrucci F, Smid JH, Graziani C, Luzzi I, Ricci A, Barco L, Rosmini R, Havelaar AH, VAN Pelt W, Busani L Epidemiol Infect. 2014 May; 142(5):1070-82.
* Attribution of human infections with Shiga toxin-producing Escherichia coli (STEC) to livestock sources and identification of source-specific risk factors, The Netherlands (2010-2014). Mughiniq-Gras L, van Pelt W, van der Voort M, Heck M, Friesema I, Franz E Zoonoses Public Health. 2018 Feb; 65(1):e8-e22.
* Salmonella source attribution based on microbial subtyping: does including data on food consumption matter? Mughini-Gras L, van Pelt W Int J Food Microbiol. 2014 Nov 17; 191():109-15.
* sourceR: Classification and source attribution of infectious agents among heterogeneous populations. Miller P, Marshall J, French N, Jewell C PLoS Comput Biol. 2017 May; 13(5):e1005564.v
* Tracing the sources of human salmonellosis: a multi-model comparison of phenotyping and genotyping methods. Mughini-Gras L, Smid J, Enserink R, Franz E, Schouls L, Heck M, van Pelt W Infect Genet Evol. 2014 Dec; 28():251-60.
* Attributable sources of community-acquired carriage of Escherichia coli containing β-lactam antibiotic resistance genes: a population-based modelling study. Mughini-Gras L, Dorado-García A, van Duijkeren E, van den Bunt G, Dierikx CM, Bonten MJM, Bootsma MCJ, Schmitt H, Hald T, Evers EG, de Koeijer A, van Pelt W, Franz E, Mevius DJ, Heederik DJJ; ESBL Attribution Consortium. Lancet Planet Health. 2019 Aug;3(8):e357-e369. doi: 10.1016/S2542-5196(19)30130-5

### 6. Modified Danish (Hald) model - France (David’s) adaptation

#### Citation

The Bayesian microbial subtyping attribution model: robustness to prior information and a proposition. David JM, Guillemot D, Bemrah N, Thébault A, Brisabois A, Chemaly M, Weill FX, Sanders P, Watier L Risk Anal. 2013 Mar; 33(3):397-408

#### Model class

Frequency matching

#### Problem formulation

To estimate directly the relative contributions (attributable proportions) of different potential sources of zoonotic infections to the human cases in settings with intensive public health surveillance

#### Model description

Adaptation of Hald model where the subtype dependent factors for unique types were set as percentage of human cases per point of prevalence in the source. This parameterization requires the availability of at least as many specific types not corresponding to Enteritidis and Typhimurium as there are different sources.

#### Data evaluation

Data on prevalence in sources were collected through an active, regulation-based surveillance system that produces representative prevalence data (as ideally required for the approach) and a passive system relying on voluntary laboratories that produces data not meeting the standards set by Hald et al. (2004) but covering a broader range of sources.

#### Conceptual model evaluation

The model estimates appeared to be robust to a deviation from the required quality of the data, due to a passive design of the sources data collection (but with a national coverage ensured), and to the consequent use of proportions instead of prevalences for the contamination indicator, but sensitive to the number of sources included.

#### Model output verification

Results were consistent with other epidemiological evidence

#### Other citations using the model

Attribution of the French human Salmonellosis cases to the main food-sources according to the type of surveillance data. David JM, Sanders P, Bemrah N, Granier SA, Denis M, Weill FX, Guillemot D, Watier L Prev Vet Med. 2013 May 15; 110(1):12-27.

### 7. Modified Danish (Hald) model - Germany (Sharp’s) adaptation

#### Citation

Sharp et al. 2014. BMTW, 127(11-12):464-77

#### Model class

Frequency matching

### 8. Modified Danish (Hald) model - Denmark (De Knegt’s) multi-country adaptation

#### Citation

Attributing foodborne salmonellosis in humans to animal reservoirs in the European Union using a multi-country stochastic model. DE Knegt LV, Pires SM, Hald T Epidemiol Infect. 2015 Apr; 143(6):1175-86.

#### Model class

Frequency matching

#### Problem formulation

To estimate directly the relative contributions (attributable proportions) of different potential sources of zoonotic infections to the human cases in settings with intensive public health surveillance

#### Model description

Adaptation of the Hald model to be able to include data from multiple countries, as well as origin of the source of infection (account for trade of foods between countries).

#### Data evaluation

Model was successfully applied with subtyping data with low discriminatory power (serotyping). Possible because it used data from multiple countries, while anchoring the subtype-dependent factor as unidimensional.

### 9. Modified Danish (Hald) model - Finland (Ranta’s) time-series adaptation

#### Citation

Bayesian temporal source attribution of foodborne zoonoses: Campylobacter in Finland and Norway. Ranta J, Matjushin D, Virtanen T, Kuusi M, Viljugrein H, Hofshagen M, Hakkinen M Risk Anal. 2011 Jul; 31(7):1156-71.

#### Model class

Frequency matching

#### Problem formulation

To estimate the relative contributions (attributable proportions) of different potential sources of zoonotic infections to the human cases based on (sub)typing data derived from source surveillance

#### Model description

It is a fully Bayesian source attribution model based on the Hald model for the analysis of time series data that also include a model fit assessment. The model takes into consideration the uncertain prevalence of bacteria per source over a long time span, and includes a novel structure for investigating the total share of unmonitored sources for which no data exist.

#### Data evaluation

The model has been applied so far on campylobacteriosis in Finland (2004–2007) and Norway (2001–2006), as monthly observed time series. But the Bayesian probability model was developed for a general purpose source attribution based on statistical information retrieved from the time series data of both human incidence and food source surveillance

#### Conceptual model evaluation

The model is derived from the Bayes theorem. Previous statistical source attribution approaches are here advanced (1) by explicit modelling of the cases not associated with any of the sources under surveillance over time, (2) by modelling uncertain prevalence in a food source by bacteria type over time, and (3) by implementing formal model fit assessment using posterior predictive discrepancy functions.

#### Implementation verification

The model is described explicitly and implemented in WinBUGS, although the code is not publicly available. The model has not been implemented besides the original publication.

#### Model output verification

The model includes a fit assessment for judgement0 of model performance with respect to relevant quantities of interest. This is relevant because the model aims at a synthesis of several incomplete information sources under significant uncertainty of explanatory variables.

#### Model analysis

The model fit assessment step allows to determine the number of change points at which model fit becomes poor and MCMC simulations began to stick at singular change points with poor mixing. This is not only to provide a “quality stamp” but to understand better how the model works and where it might be improved. It is a tool to diagnose problems and to check model performance in a sensitivity analysis.

#### Model output corroboration

The model has not been implemented besides the original publication.

### 10. Modified Danish (Hald) model - Finland (Mikkela’s) adaptation

#### Citation

A Modular Bayesian Salmonella Source Attribution Model for Sparse Data. Mikkelä A, Ranta J, Tuominen P Risk Anal. 2019 Aug; 39(8):1796-1811.

#### Model class

Frequency matching

#### Problem formulation

The model estimates the proportion of human salmonellosis associated which each reservoir included in the model using a Poisson regression approach where parameters are estimated with a modular Bayesian approach.

#### Model description

The model utilizes serotype observations from humans and salmonella reservoirs over time as well as estimates of exposure of humans to each of these reservoirs. The estimation of the relative contribution of each reservoir is done in a modular three stage Bayesian estimation. A full Bayesian model was considered to be unidentifiable and hence the approach to first estimate the exposure, then the subtype distribution and the use the human cases in an epidemiological model was used.

#### Data evaluation

Data sparsity was evaluated to determine the level of aggregation to be used to maintain reliable results. The fit of the model residual variance was assessed after fitting the model to the data. The exclusion of outbreak cases from the data was done in order to avoid the effects of many identical isolates in humans that could bias the results. Most source attribution models focus on sporadic0 cases.

#### Conceptual model evaluation

The model requires data on typing and exposure estimates. A parameter for each source and serotype and the exposure level is estimates in a 3 stage Bayesian estimation. The approach was evaluated conceptually0 by comparing it to previous approaches that used a full Bayesian estimation and found the parameters to be unidentifiable since multiple combinations of the parameters could results in the same estimate.

#### Implementation verification

Code is not supplied with the publication. The mathematical model is well documented in the paper and its supplementary material and it therefore possible to use for a reimplementation. However, there is an apparent problem with the supplementary material not including several character in an equation that could make the method not reproducible. The model was implemented in the OpenBUGS software which is no longer developed and it dependant on a ‘cut’ function implemented in that language that would need to be available in another language.

#### Model output verification

The authors describe that the impact of the choice of prior and the inclusion and exclusion of subtype specific parameters were both investigated with respect to their impact on the posterior. Neither had a substantial impact. The proportions of subtypes predicted by the model was also compared to the observed data and found to visually fit well. No, data was left out of the estimation to test model fit.

#### Model analysis

Impact of prior and inclusions of parameters were assessed in a sensitivity analysis. The results indicated little influence.

#### Model output corroboration

The results of the source attribution estimation were discussed to highlight0 which features of the input data resulted in the posterior estimates. An estimation of unknown sources on the number of human cases was completed and this was considered to be less reliable result.

### 11. SourceR (New Zealand)

#### Citation

sourceR: Classification and source attribution of infectious agents among heterogeneous populations. Miller P, Marshall J, French N, Jewell C PLoS Comput Biol. 2017 May; 13(5):e1005564

#### Model class

Frequency matching

#### Problem formulation

To estimate the relative contributions (attributable fractions) of different potential sources of zoonotic infections to the human cases in settings with intensive public health surveillance, feasible for various pathogens and sources.

#### Model description

The model builds upon, and unites, the original Hald and modified Hald (Mullner et al., 2009) models. The model is flexible, fully joint (as it combined the different outcomes in a single model), and does not rely on previous approximations and assumptions.

#### Data evaluation

sourceR is demonstrated using *Campylobacter jejuni* isolate data collected in New Zealand between 2005 and 2008. The results were qualitatively similar to results of previous studies using the same dataset.

#### Conceptual model evaluation

Mixing and a posteriori correlations are significantly decreased in comparison to the modified Hald model. Moreover, a Bayesian non-parametric model (Dirichlet process) is used to inform strain-dependent clustering effects, allowing for the identification of strain clusters with similar virulence, pathogenicity, and survivability. This is a significant enhancement, allowing for identifiability improvement over the previous models (i.e., fixing some parameters a priori or modelling0 the type effects hierarchically as random effects). The model also incorporates uncertainty in the prevalence, but it does so by fitting a fully joint model rather than a two-step model like in the modified Hald model. This has the advantage of allowing the human cases to influence the uncertainty in the source data and to preserve the restriction on the sum of the prevalences for each source.

#### Implementation verification

The model is fully checked and implemented/released as R package (sourceR)

#### Model output verification

Model fit and convergence is usually assessed visually using trace and autocorrelation plots.

#### Model analysis

Like the Modified Hald model, the new model incorporates uncertainty in the prevalence matrix into the model, however, it does this by fitting a fully joint model rather than a 2 step model. This has the advantage of allowing the human cases to influence the uncertainty in the source data and preserves the restriction on the sum of the prevalences for each source. The sourceR package implements this model to enable straightforward attribution of cases of zoonotic infection to putative sources of infection.

#### Model output corroboration

The model has not been implemented besides the original publication.

#### Other citations using the model

Also used Australian paper

### 12. Dirichlet model - New Zealand model with covariates

#### Citation

Extending statistical models for source attribution of zoonotic diseases: a study of campylobacteriosis. Liao SJ, Marshall J, Hazelton ML, French NP J R Soc Interface. 2019 Jan 31; 16(150):20180534

#### Model class

Frequency matching

#### Problem formulation

It is a source attribution model that jointly analyses the epidemiological data and genetic information of microbial isolates derived from cases and their putative sources of infection

#### Model description

The model estimates the attribution probability for human cases for each source, conditional on the extent to which each case resides in a rural compared to urban environment. The model incorporates genetic data and evolutionary processes alongside a newly developed genetic-free model.

#### Data evaluation

The model was developed using campylobacteriosis data from the surveillance sentinel in the Manawatu region of New Zealand from March 2005 to December 2014.

#### Conceptual model evaluation

This model is based of the asymmetric Island model, enhancing our understanding of the operation of this model and facilitating model checking. The model is extended in a Bayesian context to incorporate covariates, exploring the effect of human case rurality on attribution results via a linear trend on the logit scale or with separate categories, and performing model comparison. The Dirichlet model differs from the asymmetric Island model, in that it does not model pathogenic evolution, opting instead to infer the sampling distribution of genotypes directly from the observed count data.

#### Implementation verification

All codes for fitting the genotype distributions, estimating attribution probabilities and producing figures, along with the dataset are available on GitHub

#### Model output verification

Robustness analyses were performed as part of the validation of the model, which provided satisfactory results, as the model was able to capture known changes in the epidemiology of the pathogen in a region. The model also provides similar data to the asymmetric island model

#### Model analysis

Both robustness and sensitivity analyses were performed as part of the validation of the model, which provided satisfactory results.

#### Model output corroboration

The model has not been implemented besides the original publication.

### 13. STRUCTURE

#### Citation

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

#### Model class

Population genetics

#### Problem formulation

One of the first explicit models examining the genetic structure of microbial populations to estimate the probabilities that a given strain originates from a given population, i.e. source

#### Model description

This model assumes the existence of K (unknown) populations, each of which is characterized by a set of allelic frequencies at each genetic locus considered. In the simplest model without admixture, each strain is attributed to a single population. The probabilities that a strain belongs to the other populations reflect the attribution uncertainty. In the model with admixture, each locus of a strain is attributed to a population: a strain can therefore be attributed jointly to several populations.

#### Data evaluation

STRUCTURE can be applied to different genetic targets, such as microsatellites, MLST, RFLPs, AFLP, SNPs, etc. provided that this information is available for different loci. It is also necessary that common alleles show some intrinsic diversity. In theory, this approach could work with information limited to two loci and two variants per locus. In practice, however, data often derive from subtyping methods that consider several loci and variants. The model is mainly applied to Campylobacter, and to Listeria to a lesser extent

#### Conceptual model evaluation

The principle of the model is to estimate the allelic frequencies in different populations and their admixtures using Bayesian inference. Tracing the sources of human cases is a particular case of this model without admixture of the source strains, that is, the strains can only belong to one of the K populations, each of which corresponds to a specific source. The allelic frequencies at each locus are characterized for each of the K populations and the strains to be attributed are established from frequencies of characteristic allelic numbers at each locus.

#### Implementation verification

The model is available as an open-access software and its main advantage is that it allows for several loci to be considered

#### Model output verification

The results are mostly presented in graphical form, also for individual strains, or as percent attributions, corresponding to the average of the membership coefficients

#### Model analysis

The main recognized limitation of this model is related to the definition of the optimal number of populations (K). However, in the specific case of source attribution, K corresponds to the number of sources, and because of non-admixture, a strain can only originate from one source.

#### Model output corroboration

The model has been applied on different data types, pathogens and countries

#### Other citations using the model

* Attribution of Listeria monocytogenes human infections to food and animal sources in Northern Italy. Filipello V, Mughini-Gras L, Gallina S, Vitale N, Mannelli A, Pontello M, Decastelli L, Allard MW, Brown EW, Lomonaco S. Food Microbiol. 2020 Aug;89:103433. doi: 10.1016/j.fm.2020.103433. Epub 2020 Jan 20.
* Campylobacter genotyping to determine the source of human infection. Sheppard SK, Dallas JF, Strachan NJ, MacRae M, McCarthy ND, Wilson DJ, Gormley FJ, Falush D, Ogden ID, Maiden MC, Forbes KJ Clin Infect Dis. 2009 Apr 15; 48(8):1072-8.
* Attribution of Campylobacter infections in northeast Scotland to specific sources by use of multilocus sequence typing. Strachan NJ, Gormley FJ, Rotariu O, Ogden ID, Miller G, Dunn GM, Sheppard SK, Dallas JF, Reid TM, Howie H, Maiden MC, Forbes KJ J Infect Dis. 2009 Apr 15; 199(8):1205-8.
* Source attribution of human Campylobacter isolates by MLST and fla-typing and association of genotypes with quinolone resistance. Kittl S, Heckel G, Korczak BM, Kuhnert P PLoS One. 2013; 8(11):e81796.
* Elucidating the aetiology of human Campylobacter coli infections. Roux F, Sproston E, Rotariu O, Macrae M, Sheppard SK, Bessell P, Smith-Palmer A, Cowden J, Maiden MC, Forbes KJ, Strachan NJ PLoS One. 2013; 8(5):e64504.
* Molecular epidemiology of Campylobacter jejuni in a geographically isolated country with a uniquely structured poultry industry. Müllner P, Collins-Emerson JM, Midwinter AC, Carter P, Spencer SE, van der Logt P, Hathaway S, French NP Appl Environ Microbiol. 2010 Apr; 76(7):2145-54.
* Nielsen E. M., Björkman J. T., Kiil K., Grant K., Dallman T., Painset A., et al. (2017). 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. EFSA Support. Publ. 14:1151E–n/a

### 14. Asymmetric island model

#### Citation

Tracing the source of campylobacteriosis. Wilson DJ, Gabriel E, Leatherbarrow AJ, Cheesbrough J, Gee S, Bolton E, Fox A, Fearnhead P, Hart CA, Diggle PJ PLoS Genet. 2008 Sep 26; 4(9):e1000203.

#### Model class

Population genetics

#### Problem formulation

Estimate the relative contribution of various reservoirs of 7 allele MLST typed *Campylobacter jejuni* to similarly typed human cases on Campylobacteriosis. The outcome is the interpretable at the population level as the estimate proportion of human cases as a result of each reservoir or at the individual human case level as the probability of each reservoir having been the source of the specific case. The estimation of the human cases unattributable to any of the included sources is done.

#### Model description

Bayesian estimation of the proportion of human cases attributable to reservoir with a symmetrical Dirichlet prior in which all reservoirs are considered equal. The posterior is estimated using a genetic model of DNA sequence evolution to estimate the source of each human isolate based on its type. The reservoir isolates are used to estimate mutation, recombination and migration rates between the reservoirs. These parameters along with the typing results from humans and reservoirs are then used to estimate the proportion of human cases that each reservoir contributes.

#### Data evaluation

The input data to the model is checked for excess of MLST types unique to humans. Since and assumption of the model is that the major sources of human disease are included as reservoirs in the model input data, an excess of unique types in humans could be an indication that a source remains unsampled and hence the results could be biased. AMOVA is also used to asses if reservoirs are sufficiently0 differentiated to be included in the model as different putative sources.

#### Conceptual model evaluation

The conceptual model is not justified *per se* but represents a comprehensive approach to source attribution by including the possibility of migration between reservoirs and evolution of the bacteria by both recombination and mutation.

#### Implementation verification

The code to run this model written in C++ are supplied in an openly available repository that may be compiled from source and the author also supplies a binary for convenience.

#### Model output verification

Model fit is evaluated using empirical cross-validation involving a leave-one-out approach to determine if predictions are stable to remove of individual observations and if these observations can be accurately predicted with the model. Formal model fit assessment is not done and no calculation of the residual variance is done.

#### Model analysis

Model parameters are entirely estimated from the data and the model is sensitive to data included in the analysis. The authors state that an implicit assumption is that all potential sources of disease in humans are sampled and included in the analysis. The intensity of sampling of sources included in the analysis has been done and it is unclear what appropriate sampling of a reservoir is to avoid bias.

#### Model output corroboration

Empirical cross validation is used and the assessment of excess unique bacterial types in humans. No formal approach is used to corroborate results with other findings.

#### Other citations using the model

* sourceR: Classification and source attribution of infectious agents among heterogeneous populations. Miller P, Marshall J, French N, Jewell C PLoS Comput Biol. 2017 May; 13(5):e1005564.v
* Assigning the source of human campylobacteriosis in New Zealand: a comparative genetic and epidemiological approach. Mullner P, Spencer SE, Wilson DJ, Jones G, Noble AD, Midwinter AC, Collins-Emerson JM, Carter P, Hathaway S, French NP Infect Genet Evol. 2009 Dec; 9(6):1311-9.
* Tracing the sources of human salmonellosis: a multi-model comparison of phenotyping and genotyping methods. Mughini-Gras L, Smid J, Enserink R, Franz E, Schouls L, Heck M, van Pelt W Infect Genet Evol. 2014 Dec; 28():251-60.

### 15. Paper on STEC source attribution

#### Citation

Im, Hanhyeok, Seung-Ho Hwang, Byoung Sik, and Sang Ho. 2021. “Pathogenic Potential Assessment of the Shiga Toxin – Producing Escherichia Coli by a Source Attribution – Considered Machine Learning Model” 118 (20): 1–9. https://doi.org/10.1073/pnas.2018877118.

#### Model class

Machine learning

#### Problem formulation

Various ML models tested to develop the best ML model that can predict the pathogenicity/pathogenic potential of Shiga toxin–producing Escherichia coli (STEC) isolates using their WGS data.

#### Model description

Unsupervised ML algorithms: phylogenetic tree analysis, principal component analysis (PCA), and Gaussian mixture model (GMM) and supervised ML algorithms: Gaussian Naive Bayes, decision trees (DTs), random forest (RF), and support vector machine (SVM) evaluated.

#### Data evaluation

The WGS data and metadata of 3,303 STEC isolates retrieved from the GenBank database at the National Center for Biotechnology Information (NCBI) and grouped in clinical ( 2,292) and environmental ( 354 ) isolates absed on their metadata.

#### Conceptual model evaluation

The Matthews correlation coefficient (MCC) and the area under the receiver operating characteristic curve (AUROC) used to compare the discrimination performances of the supervised ML models.

#### Implementation verification

Decision Function Values applied to the SVM model to calculate and classify each STEC isolate into either pathogenic or nonpathogenic group.

#### Model output verification

Permutation importance analyses of the input dataset to identify genes important for the evaluation of the best performing SVM model and subsequent clusters based on the Spearman rank–order correlation. The permutation importance analyses repeated 10 times for each gene or cluster and decrease of the MCC score used as a weight value.

#### Model analysis

The SVM model successfully predicted the pathogenic potential of the STEC isolates or which the serotype information is not available and assessed the pathogenic potential of the input dataset isolates carrying distinct virulence gene combinations.

#### Model output corroboration

The SVM model correctly predicted the STEC isolates with the history of outbreaks to carry high pathogenic potential.

#### Other citations using the model

No citations yet.

### 16. Munck model - Machine learning

#### Citation

Munck, Nanna, Patrick Murigu Kamau Njage, Pimlapas Leekitcharoenphon, Eva Litrup, and Tine Hald. 2020. “Application of Whole-Genome Sequences and Machine Learning in Source Attribution of Salmonella Typhimurium.” Risk Analysis 40 (9): 1693–1705. https://doi.org/10.1111/risa.13510.

#### Model class

Machine learning

#### Problem formulation

Estimate the proportion of human cases attributable to several reservoirs of cgMLST typed salmonella by machine learning. Both the logit boost and random forest models were applied.

#### Model description

The approach was to utilize food and animal origin cgMLST typed isoaltes to parameterize two machine learning models to predict the origin of the types, Logit boost and random forest. These models were then used to predict the origin of the observations of cgMLST types in the human data.

#### Data evaluation

The data were considered to be appropriate for use in the model because there was no unique “human-only” cluster identified in a phylogenetic analysis of the data. The interspersing of types from humans and the included reservoirs was seen as evidence that the origin of the human cases was contained in the included sources and that the data was appropriate for modeling.

#### Conceptual model evaluation

Two models were compared, a random forest and logit boost. The logit boost approach was selected after a cross-validation of the model results with data left out of the parametrization. The logit boost approach results in 0.93 accuracy to predict the origin of a cgMLST type.

#### Implementation verification

The methodology is described in the paper and accompanied by a non-reproducible R script further detailing the model. Contact with the authors would be required to reproduce the work primarily due to lack of supplied sample data to run the supplied code. However, the tools used to perform the machine learning analysis are available as open source R packages that are peer reviewed.

#### Model output verification

The model results were compared to the predictions from the Hald model although only performed without the consumption parameters. Since it is common to scale the estimates of the Hald model by known consumption estimates it is not clear if this model performs better than a model that included more input parameters such as consumption.

#### Model analysis

The cross validation approach indicated that the model was stable to resampling the input data.

#### Model output corroboration

The results of the model were compared to those obtained by another previously published approach for source attribution (Hald model).

### 17. Guillier - Machine learning model

#### Citation

Guillier, Laurent, Michèle Gourmelon, Solen Lozach, Sabrina Cadel-Six, Marie Léone Vignaud, Nanna Munck, Tine Hald, and Federica Palma. 2020. “AB\_SA: Accessory Genes-Based Source Attribution – Tracing the Source of Salmonella Enterica Typhimurium Environmental Strains.” Microbial Genomics 6 (7): 1–10. https://doi.org/10.1099/mgen.0.000366.

#### Model class

Machine learning

#### Problem formulation

AB\_SA (Accessory genes-Based Source Attribution) method seeks to estimate the relative contributions (attributable fractions) of different potential sources of zoonotic infections to the human cases or to strains sampled from the environment in settings with intensive public health surveillance, feasible for various pathogens and sources.

#### Model description

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.

#### Data evaluation

Technical University of Denmark (DTU) FoodQCPipeline generated high-quality assemblies of 98 French S. enterica Typhimurium and *S.* enterica 1,4,[5],12:i:- isolates collected in 2010–2015 - 49 of the isolates had a known source: pigs (n=49), poultry (layer chickens, broiler chickens, turkeys and ducks) (n=14), ruminants (cattle, sheep and goats) (n=6), and 29 of the isolates required attribution as they were isolated from the environment (e.g. fresh or brackish water and soil).

#### Conceptual model evaluation

AB\_SA is a MLR-based model and hence is not very sensitive to such data assumptions as normality, linearity and homogeneity of data, which allows to avoid overfitting. Additionally, as AB\_SA returns both accuracy and AIC value (a penalty for model complexity), overfitting is prevented. Moreover, the AB\_SA method also returns balanced accuracies for each of the sources. Balanced accuracy values provide more detailed and appropriate information than global accuracy, especially for unbalanced datasets. AB\_SA is flexible in the choice of the level of significance (through a threshold value for the P value) of enriched genes and in the number of candidate predictors per source to feed the multinomial logistic model. This MLR model also provides a measure of the weight of each predictor (i.e. gene) for each source category.

#### Implementation verification

The model is fully checked and implemented/released as R package (AB\_SA) and the documentation is clear and easy to follow (https://github.com/lguillier/AB\_SA).

#### Model output verification

The overall self-attribution accuracy, as well as the source balanced accuracies of this study, were 67–90%, i.e. similar to what is commonly obtained with supervised machine learning models. The enrichment step of the AB\_SA method that has to be performed by the user select the most relevant genes for modelling their effects with the multinomial logistic model is of very high importance and will influence the model outputs. AIC values allow to identify the model with the optimum number of genes from among the tested models.

#### Model analysis

Model is highly sensitive to the quality of the genomic assemblies of the analysed isolates and hence on the quality of the whole genome sequencing data. Identification and selection of source-enriched genes is critical to the performance0 of the AB\_SA method.

#### Model output corroboration

Model implementation has not published as of yet other than in the original publication. Work in progress, where the performance of AB\_SA method was directly compared to the performance of a random forest model suggested that AB\_SA performed worse than RF when attributing sources to the same test set of animal isolates.

#### Other citations using the model

None that explicitly applied the method to perform source attribution of bacterial isolates but the AB\_SA paper is cited in a couple of recent review papers.

#### Citation

Zhang, Shaokang, Shaoting Li, Weidong Gu, Henk den Bakker, Dave Boxrud, Angie Taylor, Chandler Roe, et al. 2019. “Zoonotic Source Attribution of Salmonella Enterica Serotype Typhimurium Using Genomic Surveillance Data, United States.” Emerging Infectious Diseases 25 (1): 82–91. https://doi.org/10.3201/eid2501.180835.

#### Model class

Machine learning

### 18. Lupolova model

#### Citation

Lupolova, Nadejda, Samantha J. Lycett, and David L. Gally. 2019. “A Guide to Machine Learning for Bacterial Host Attribution Using Genome Sequence Data.” Microbial Genomics 5 (12). https://doi.org/10.1099/mgen.0.000317.

#### Model class

Machine learning

#### Problem formulation

To estimate the relative contributions (attributable fractions) of different potential sources of zoonotic infections to the human cases in settings with intensive public health surveillance, feasible for various pathogens and sources.

#### Model description

Multiple unsupervised (kmeans, hierichical agglomerative/divise, latent dirichlet allocation) and supervised models (SVM, RF, NN) evaluated.

#### Data evaluation

1203 S . enterica serovar Typhimurium genome sequences previously analysed in Lupolova et al. (2017). Data split into 4 hosts (avian, porcine, bovine, human). Discriminatory protein variants extracted from pan-genome used as either direct features or euclidean distances calculated based on their presence or absence.

#### Conceptual model evaluation

The models estimate the proportion of each assigned to each cluster (unsupervised) or the accuracy imputed from cross-validation (supervised). Attribution of human isolates to the other three sources was not performed i this study but was in the authors previous work.

#### Implementation verification

The R code to run the different models and their evaluation is available as supplementary. The features used (PVs) are available in the data bibliography.

#### Model output verification

Models were assessed based on cluster composition, Silhouettes index and concordance with phylogeny. With supervised learning methods further assessed based on cross-fold validation accuracy. All methods perform as expected from previous applications and publications.

#### Model analysis

Models trained on discriminatory PVs are inherently linked to the data and labels of the input data set and therefore their applicability to unseen data is unclear.

#### Model output corroboration

Results consistent with previous implemenations

#### Other citations using the model

* Lupolova N, Dallman TJ, Matthews L, Bono JL, Gally DL. Support vector machine applied to predict the zoonotic potential of E. coli O157 cattle isolates. Proc Natl Acad Sci USA. 2016;113:11312–11317. doi: 10.1073/pnas.1606567113. [PMC free article] [PubMed] [CrossRef] [Google Scholar]
* 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. 2017;3:e000135. doi: 10.1099/mgen.0.000135. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

### 19. Wheeler model

#### Citation

Wheeler, Nicole, Paul Gardner, and Lars Barquist. 2017. “Machine Learning Identifies Signatures of Host Adaptation in the Bacterial Pathogen Salmonella Enterica.” Machine Learning Identifies Signatures of Host Adaptation in the Bacterial Pathogen Salmonella Enterica, 204669. https://doi.org/10.1101/204669.

#### Model class

Machine learning

#### Problem formulation

Classification of Salmonella serovars based on lifestyle as either invasive, host-adapted, extra-intestinal or intestinal serovars.

#### Model description

A random forest classifier using DeltaBS functional variant calling to separate intestinal Salmonella serovars from host-adapted, extra-intestinal serovars to identify signatures of mutational burden consistent with adaptation to an invasive lifestyle. For each genome, the functional significance of sequence variation within protein coding genes was quantified using the DeltaBS metric. Following scoring, a bootstrap sampling of genomes were used to train each decision tree. For each node in the tree, a random subset of genes were sampled, and the most informative gene from this set was chosen to split the data. For each node in the tree, the predictive utility of the selected gene (variable importance) was tested by calculating how well the gene separates the samples according to phenotype. Subsequently, a random forest classifier was employed to identify the genes which were most informative of phenotype when viewed collectively.

#### Data evaluation

The model was trained on high quality genomes for 13 well-characterised Salmonella enterica serovars that were retrieved from the NCBI database that were characterised as “gastrointestinal” or “extra-intestinal”, whereas the validation set comprised ten of the same Salmonella serovars. The model on a set of 6,438 orthologous genes that were reduce to 196 genes after several rounds of feature selection.

#### Conceptual model evaluation

Accuracy of the model was assessed using out-of-bag accuracy. This out-of-bag (OOB) measure of accuracy gives us an indication of how well each decision tree in the forest performs at predicting phenotype in a serovar it has never encountered before, using information on DeltaBS differences collected from other serovars.

#### Implementation verification

All genome sequence data are publicly available, and accessions are provided in the appropriate Supplemental Tables. Code and data for reproducing this analysis, performing an equivalent analysis using new data, and assessing the invasiveness index of other Salmonella strains is publicly available at http://www.github.com/UCanCompBio/invasive\_salmonella. The model itself was run in a well known R package “randomForest”.

#### Model output verification

The final model used that 196 of the original 6438 genes for prediction achieved perfect classification accuracy on an independent set of ten genomes of the same serovars as the training data. Additionally, the random forest classifier that we used produces interpretable lists of genes involved in this adaptation, which agreed with results in the literature attained through manual curation of pseudogenes.

#### Model analysis

Perfect OOB classifications were only achieved by the fifth iteration of the model. i.e. after purging the vast majority of the 6438 orthologous genes that were used as model features for the initial model iteration. The need for iterative improvement of the model came from difficulty in correctly classifying the reference strains for serovars Enteritidis and Dublin. This is reflective of their relatively recent divergence and niche adaptation compared to other serovars in the study. S. Gallinarum was classified much more readily than S. Enteritidis and S. Dublin, despite being closely related to both serovars, perhaps due to its host restriction.

#### Model output corroboration

Model indicated invasiveness index obtained for several African strains of novel lineages of *S.* Typhimurium and *S.* Enteritidis was in agreement with what has been published in the literature.

#### Other citations using the model

* Sandholt, A.K.S., Neimanis, A., Roos, A. et al. Genomic signatures of host adaptation in group B Salmonella enterica ST416/ST417 from harbour porpoises. Vet Res 52, 134 (2021). https://doi.org/10.1186/s13567-021-01001-0
* Shen, Z.Y.; Koh, X.P.; Yu, Y.P.; Lau, S.C.K. Genetic Variation and Preliminary Indications of Divergent Niche Adaptation in Cryptic Clade II of Escherichia. Microorganisms 2020, 8, 1713. https://doi.org/10.3390/microorganisms8111713

### 20. Arning model

#### Citation

Arning, Nicolas, Samuel K Sheppard, David A Clifton, and Daniel J Wilson. 2021. “Machine Learning to Predict the Source of Campylobacteriosis Using Whole Genome Data.” BioRxiv, 2021.02.23.432443. https://doi.org/10.1101/2021.02.23.432443.

#### Model class

Machine learning

#### Problem formulation

Estimate the proportion of cases of human campylobacteriosis from various reservoirs. The method relies on cgMLST typing..

#### Model description

An XGboost algorythm on cgMLST was trained on to classify isolates from PubMLST.

### 21. Hudson - Phylogeny model

#### Citation

Hudson et al, 2021

#### Model class

Population genetic. This method could be considered to be a subclass: phylogenetic analyses.

#### Problem formulation

To compare clinical and non-clinical Campylobacter populations from Tennessee (TN) and Pennsylvania (PA), use phylogenetic relatedness to assess source attribution patterns, and identify potential outbreak clusters.

#### Model description

Phylogenetic analyses conducted to high quality assemblies to categorize isolates into species groups and determine the population structure of each species. Analyses conducted with KSNP3. The output core SNP matrix FASTA files used to create phylogenetic trees with evolutionary distances calculated and evolutionary history inferred using the neighbour-joining method. The cluster identification analysis was performed with each of the potential clusters using hqSNP. individually using the CFSAN SNP pipeline

#### Data evaluation

A total of 3080 isolates used in the study, including downloaded sequences from NCBI SRA and 110 clinical isolates collected from patients sequenced for the study. Only assemblies that met high assembly criteria and were clustering into species groups used in the study

#### Conceptual model evaluation

Optimum k-mer value was determined for each run using the included Kchooser utility; optimum values determined and used for analyses were 19 and 21. All isolates were also subtyped using the appropriate MLST schemes on PubMLST

#### Implementation verification

Identified livestock source attribution trends for individual C. jejuni and C. coli clades based on phylogenetic clustering. Additionally, a large number of potential outbreak clusters for C. jejuni, which may indicate that more C. jejuni illnesses share a common source than previously thought.

#### Model output verification

Only genomic relatedness was used for cluster identification in this study; epidemiological data are needed to confirm these clusters.

#### Model analysis

Other researchers have also found that when using WGS methods, the number of isolates that form clusters is higher than previously thought. This may be due to the increased discriminatory power provided by WGS and the resulting analyses compared to previously used methods (e.g., PFGE). Host association discussed.

#### Model output corroboration

C. jejuni and C. coli clonal complexes (CC) associated with different livestock or environmental sources were compared with the model outputs.

#### Other citations using the model

No citations yet.