Validated source attribution methods for NGS data

COMPARE Deliverable 4.4

Version 01, November 2018

Version 01. November 2018
Due Date month: 48
Completed month: 48
Authors:
DTU: Nanna Munck and Tine Hald
UNIBO: Daniel Remondini, Alessandra Merlotti and Frederique Pasquali.
ANSES: L. Guillier
RKI: Niklas Willrich and Anika Schielke

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 643476.

Contents

Introduction

The attribution of sources of foodborne pathogens is a key-issue for public health and risk management, as it estimates the relative contributions of different sources to the human disease burden, which is needed to set priorities for food safety interventions and to measure the impact of such interventions (Pires et al., 2009). Attribution of sources integrates a growing number of diverse methodological approaches and data types. Beyond traditional epidemiological approaches, including observational studies based on outbreak investigations (Pires et al., 2010) and case-control/cohort studies (Fullerton et al., 2012) a number of microbiological approaches to source attribution have been developed in recent years (Mughini-Gras et al., 2018).

The principle of source attribution methods based on microbial subtyping is to partition human cases caused by a given pathogen over a number of putative sources of infection based on the distribution of pathogen subtypes in humans and in sources. These methods therefore rely on the grouping by subtype of pathogen strains from humans and their potential sources, with the subtypes being defined by phenotyping (e.g. antimicrobial resistance, phage typing) and/or genotyping (e.g. *MLST, MLVA*) methods.

Population genetics models can be empirically validated using self-attribution (Kittl et al., 2013; Sheppard et al., 2009), which provides an indication that the attributions are sound. Few comparative studies of models using WGS have been published (Nielsen et al., 2017), it is therefore difficult to recommend the use of one model over another. The results of the self-attribution are thus undoubtedly the best criterion for assessing the relevance of a source attribution model (ANSES, 2017).

Review of existing methods for source attribution using genomic data

Source attribution using STRUCTURE approach

STRUCTURE was developed by Pritchard *et al*. (Pritchard et al., 2000) and is one of the first explicit models examining the genetic structure of microbial populations. This model assumes the existence of *K* (unknown) populations, each of which is characterized by a set of allelic frequencies at each locus. 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.

The principle of the model is to estimate the allelic frequencies in the 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.

STRUCTURE can be applied to different genetic targets (allele type, SNP) as long as 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. The results are mostly presented in graphical form and/or as percent attributions, corresponding to the average of the membership coefficients. The

overall attributions are sometimes accompanied by measures of uncertainty like 95% confidence/credible intervals, although the exact principle of their calculation is not always explicitly mentioned.
The model is available as open-access software and has been first applied to small number of loci (Kovac et al., 2017). More recent examples showed that the method is also relevant for whole-genome sequencing (WGS) data (Nielsen et al., 2017; Thépault et al., 2017; Thépault et al., 2018).

Aberdeen source attribution model (Nielsen et al., 2017)

The Aberdeen approach relies on SNPs or cgMLST. In this method (Nielsen et al., 2017), attribution is based on the similarity between strains to attribute the strains to different sources (i.e. ruminants, poultry, and pigs). A strain is attributed to the reservoir which has the maximum number of similar SNPs. This is simply calculated by summing the number of loci that are identical. Hence, each of the *N* human isolates used in a study are allocated to a source. For example if *n* are allocated to ovine then the attribution score to the ovine source is n/N. Sample size correction is carried out, if the number of strains in sources are different. The sample size of the smallest source is used. The Aberdeen model has been applied to *Listeria monocytogenes* genomic data.

Source attribution using phylogeographic model (Dearlove et al., 2016)

The approach published by Dearlove et al. (2016) combines a phylogenetic reconstruction (of aligned sequence data) under a coalescent model and a method for phylogeography (or discrete traits mapping) proposed by Lemey et al. (2009). This approach permits to estimate the probability of any node in the phylogenetic tree of association to a host and also to obtain the relative rates of transition between host species. The human isolates are set an 'unknown' source population and the posterior probability of association to host can be estimated. This approach has been applied to *Campylobacter (*Dearlove *et al.,* 2016).

Description of methods developed in the context of COMPARE project

DTU, APHA, UNIBO, ANSES and RKI have in the context of the COMPARE project developed new methods that use WGS for source attribution. These methods used identical data sets of *Salmonella* Typhimurium genomic strains collected for this purpose specifically. Data sets from Denmark, Germany, United Kingdom and France were collected. Principles and validation of the newly developed methods are presented below.

Machine-learning approach (DTU)

Principle

Machine learning (ML) is the collective name for mathematical models that learn from data and improves with experience, meaning more data. The models are defined by algorithms capable of recognizing patterns in large and complex datasets making the method useful for analyzing DNA sequence data. The algorithms identify features in the dataset relevant for the purpose of the model (i.e. host-associated genes) enabling the ability to make strong predictions. For the purpose of source attribution, we used the core genome multi-locus sequence types (cgMLST) as input data, i.e. all core genes are used in the analysis and the strains are differentiated by their allelic variations. The cgMLST profiles for each sequence were obtained using the Enterobase scheme in BioNumerics version 7.6 (Applied Maths, Sint-Martens-Latem, Belgium). The core genome of *Salmonella* consist of 3,002 loci with one single locus having several allelic variations, thereby providing a high discriminatory power compared to previous methods used. We applied a supervised classification ML model. The classification is

supervised, because the machine is 'told' from which of the different animal reservoirs (classes) each of the specific isolates from food and animal originates, and the model then identifies those cgMLST loci that are able to differentiate between these reservoirs/sources based on their allelic variation. The number of features (loci) were reduced by applying NearZeroVariance, which excludes loci with no information useful for distinguishing between the different sources. Both a high and a low discriminatory model were developed based on the 104 and the 20 most important loci respectively. The dataset was upsampled to account for unequal sample sizes among the sources. Different machine learning algorithms were tested, namely random forest and logit boost. The ML model was developed using a training dataset consisting of the majority (70%) of the animal and food isolates, see Figure 1. The accuracy of the model was determined from the models' ability to predict the origin of the remaining part (30%), the test dataset, of the animal and food isolates. This approach is also referred to as self-attribution. The model does not compute uncertainty intervals per se, but takes the uncertainties into account by repeating the model building process 10 times using different training and test subsets and applying a 7-fold cross validation for each model version. The performance of the model was reported in a confusion matrix from which an overall accuracy was calculated (Table 1). The model with the most satisfying accuracy was selected as the final model. For the Danish dataset, the logit boost model performed the best.
The probability of each human strain to originate from a specific source was predicted from the final model. The sum of these probabilities within each source equals the total number of human cases attributed per source. Human strains whose source could not be predicted, are referred to an unknown source category.
The ML model includes only domestic and sporadic human cases, i.e. human cases with no or unknown travel history as well as one case from each domestic outbreak. Results obtained from both the high and low discriminatory models based on the Danish dataset are visualized in Figure 2 and elaborated in Table 2. The uncertainty of the results is reflected in the Figure 2, where the probability of each human case to belong to one of the seven sources is illustrated.

Validation

Besides the principle of self-attribution as described above, the ML model can be validated by comparing the results with the results of other source attribution models using the same or very similar datasets.

The Danish dataset used to develop the ML model was a subset of the *Salmonella* samples used for the Danish so-called "Salmonella source account" published in the Annual Reports on Zoonoses in 2013 and 2014 (see e.g. Anonymous, 2015). In 2013 and 2014, the attribution of human salmonellosis cases was based on a mathematical model that compares the number of human cases caused by different *Salmonella* subtypes with the distribution of the same subtypes isolated from various animal-food sources using a Bayesian modelling approach that also accounts for the prevalence of the *Salmonella* subtypes in the different sources and the amount of each food source consumed per year (de Knegt at al., 2016; Hald et al., 2004). The *Salmonella* subtypes were defined by serotyping, Multiple Locus Variable Tandem Repeat Analysis (MLVA) for *S*. Enteritidis and *S*. Typhimurium strains and resistance profiling. To validate the developed ML model, we will compare the attribution results obtained by the Bayesian model to those obtained with the new model. This work will be undertaken during the next two months.

The ML method was, however, applied to prepare the Danish *Salmonella* source account for 2017 data as published in the Annual Report on Zoonoses in Denmark 2017 (Anonymous 2018, 2018). Regrettably, the *Salmonella* strains from 2017 were not MLVA typed, so it was not possible to make a comparison of the two models using the same data set. Still, the results obtained with the ML model was in line with the attribution

results obtained by the Bayesian model applying 2016 data (Anonymous, 2017), which gives creditability to both approaches.

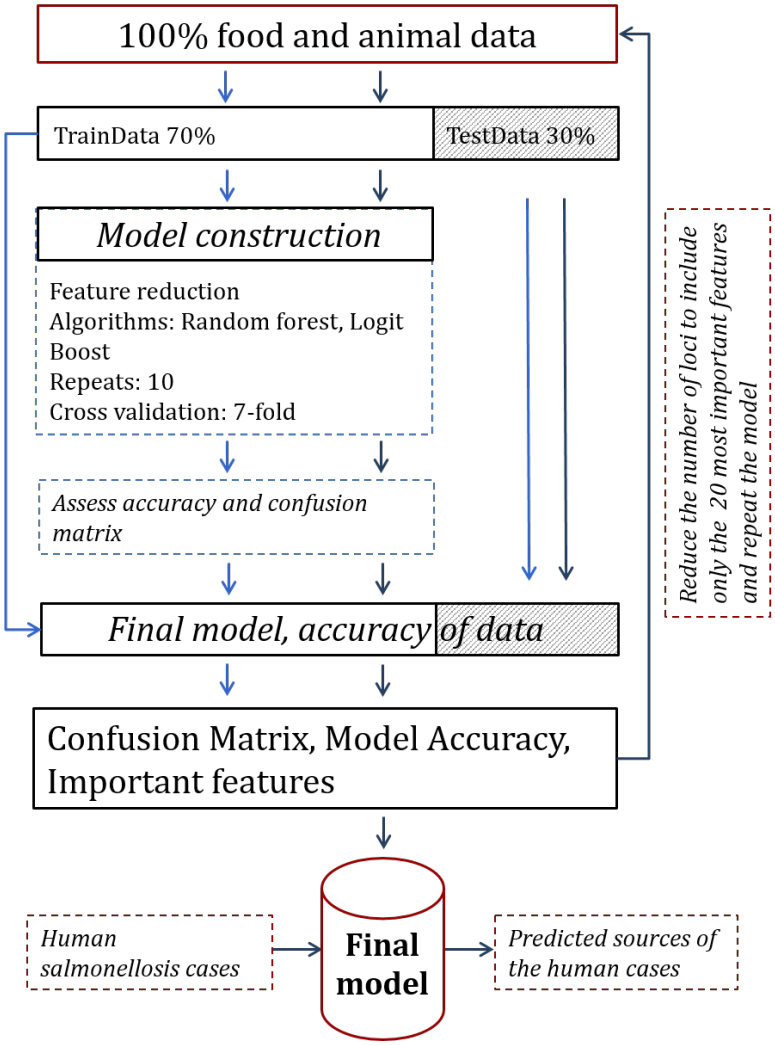

FIGURE 1 CONCEPTUAL MODEL OF THE ML MODEL FLOW.

TABLE 1 CONFUSION MATRIX FROM THE LOW DISCRIMINATORY ML MODEL. ACCURACY: 0.98 (CI 0.97-0.99).

% of 125 samples	Broilers (DK)	Cattle (DK)	Cattle (Import)	Ducks (Import)	Layers (DK)	Pigs (DK)	Pigs (Import)
Broilers (DK)	60,8	0	0	0	0	1,6	0
Cattle (DK)	0	100	0	0	0	0	0
Cattle (Import)	0	0	100	0	0	0	0
Ducks (Import)	0	0	0	100	0	0	0
Layers (DK)	0	0	0	0	100	0	0
Pigs (DK)	4	0	0	0	0	25,6	0
Pigs (Import)	0	0	0	0	0	5,6	38,4
Not predicted	35,2	0	0	0	0	67,2	61,6

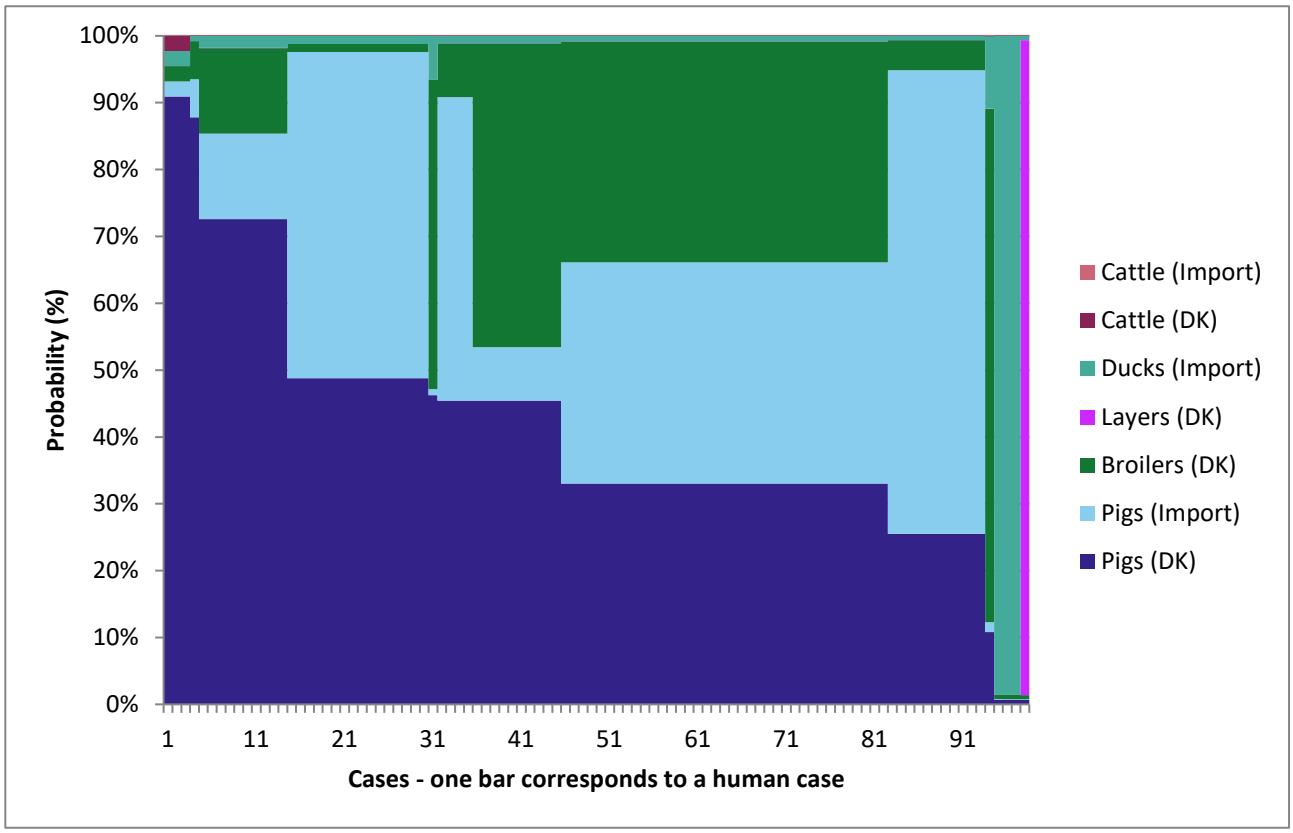

FIGURE 2 RESULTS FROM THE LOW DISCRIMINATORY LOGIT BOOST SA MODEL. ALL 98 PREDICTED HUMAN CASES ARE LINED UP ALONG THE X-AXIS AND THE SOURCE SPECIFIC PROBABILITIES FOR EACH OF THE HUMAN CASES ARE STACKED ALONG THE Y-AXIS. HUMAN CASES ATTRIBUTED TO AN UNKNOWN SOURCE NOT SHOWN.

TABLE 2 RESULTS FROM THE LOW AND HIGH DISCRIMINATORY ML MODELS.

	Discriminatory level	
	Low, 20 loci	High, 104 loci
Number of human cases predicted (%)	98 (69.5)	54 (38)
Accuracy	0.98 (0.97 ; 0.99)	0.98 (0.97 ; 0.99)
Prediction	**n (% of 141 human cases)**	
Broilers (DK)	20 (14.5)	6 (4.4)
Cattle (DK)	0 (0.1)	0 (0)
Cattle (Import)	0 (0)	0 (0)
Ducks (Import)	4 (2.9)	1 (0.7)
Layers (DK)	1 (0.7)	1 (0.7)
Pigs (DK)	41 (28.8)	25 (17.8)
Pigs (Import)	32 (22.5)	21 (14.7)
Travel cases	23 (16.3)	23 (16.3)
Unknowns	20 (14.2)	64 (45.4)

Accessory-genome based approach (ANSES)

Principle

Many bacterial phenotypes are related to the presence or absence of particular genes that are inherited through descent or acquired through lateral gene transfer. The full complement of all genes among a set of strains is referred to as the pan-genome.

The accessory-genome based source attribution method is a two steps method (see Figure 3 and Figure 4). First, the annotation was carried out using Prokka (Seemann, 2014) with default parameters. Prokka uses the assemblies as input and produces GFF3-files, including sequences and annotations. These files are used to extract the pangenome of the host isolates with the software Roary (Page et al., 2015). Finally, an enrichment of gene in different hosts was performed using Scoary (Brynildsrud et al., 2016) by following the instructions provided on https://github.com/AdmiralenOla/Scoary. A file including phenotypic traits, the phylogenic tree based on SNPs and the gene presence/absence matrix from Roary were thus used as input-data to Scoary and a pan-genome wide association study to know which genes were enriched in the different host groups was carried out.

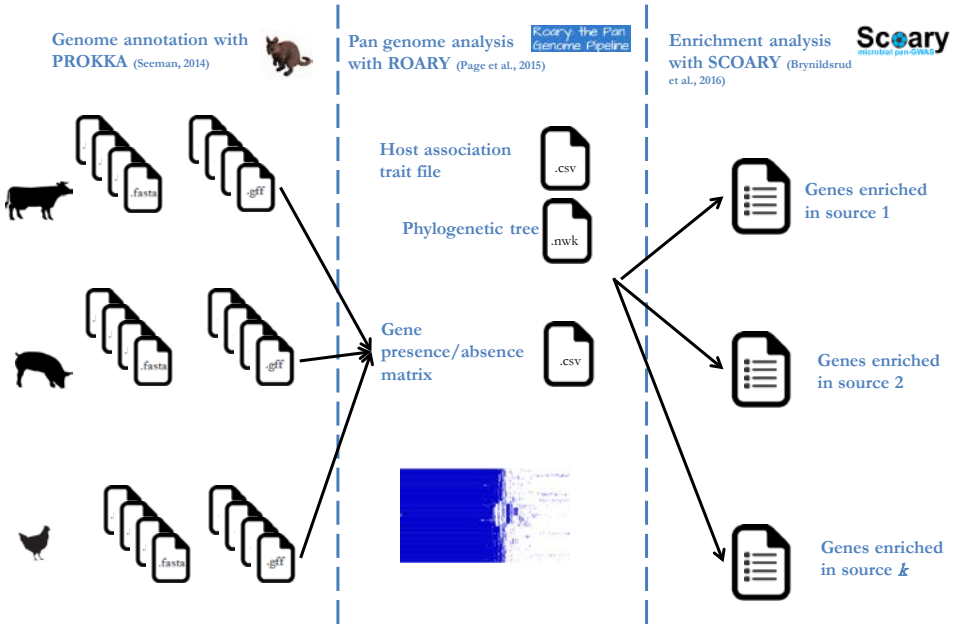

FIGURE 3 FIRST STEP OF THE ACCESSORY-BASED SOURCE ATTRIBUTION METHOD: DETERMINATION OF GENES ENRICHED IN THE DIFFERENT HOSTS.

Then a multinomial logistic model is fitted to the presence/absence of the whole set of genes enriched in the different hosts. Multinomial regression is similar to logistic regression but it is applicable when the response variable is a nominal categorical variable with more than two levels.

The fitted multinomial model can then be used to predict the probabilities of association to the different hosts of a strain based on the presence/absence of genes included in the model.

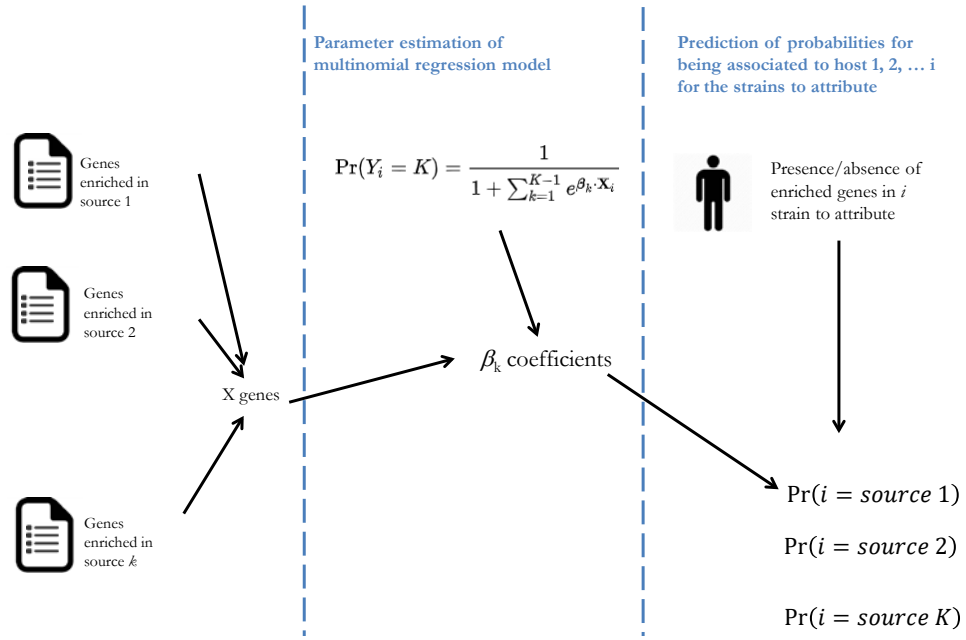

FIGURE 4 SECOND STEP OF THE ACCESSORY-BASED SOURCE ATTRIBUTION METHOD: FITTING OF MULTINOMIAL MODEL AND CALCULATION OF PROBABILITIES OF ASSOCIATION TO HOSTS FOR STRAINS TO ATTRIBUTE.

Validation

When developing models for prediction, it is important to test how well the model performs in predicting the target variable on out of sample observations. Self-Attribution was used to assess the performance of the ABSA method. The "French" dataset gathering different strains from three different sources of *Salmonella* Typhimurium was used for the validation step. The process involves using the model estimates to predict values on the training set. Then, the predicted target variable can be compared to the observed values for each observation. The dataset was split in two parts: 70% for model fitting and 30% for testing.

The accuracy of self-attribution of the ABSA model applied to the French dataset of Salmonella Typhimurium datasets are shown in Figure 5. Accuracy of 80% is achieved (95%CI [55%-92%]), indicating that ABSA model performance are suitable for source attribution.

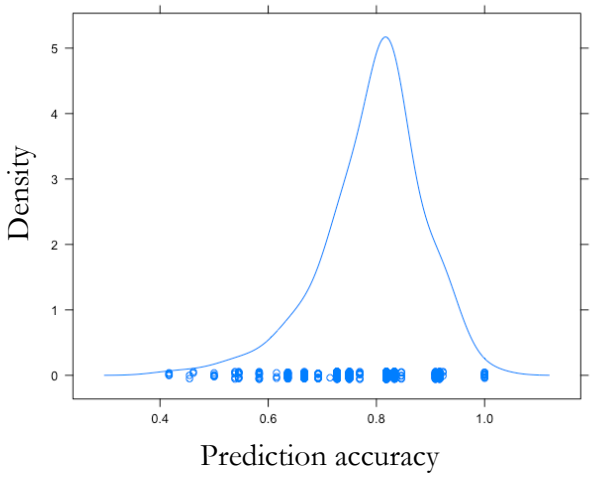

FIGURE 5 PREDICTION ACCURACY VALUES OF THE ABSA MODEL ON THE FRENCH *SALMONELLA* OBTAINED WITH CROSS-VALIDATION.

An asymmetric Island model at cgMLST level (RKI)

Principle

The RKI developed a Source Attribution model applicable for cgMLST data combining a variable selection step based on the Group-LASSO with a Bayesian inference approach based on a generalization of Wright's asymmetric island model as implemented in the *iSource* program available from http://www.danielwilson.me.uk/iSource.html.

First, data on 56 test isolates was set aside for validation purposes and then the rest of the isolates where used as training data.

The first step (variable selection) was used to reduce the very high number of loci in the cgMLST to the ones with possible information content for the question of source attribution. To this end a logistic group LASSO regression as implemented in the R-package grplasso (https://cran.r-project.org/web/packages/grplasso/) was used. For all three possible sources (cattle, pig, chicken) logistic regressions predicting one *vs*. the others were run with the loci as predictors and each time the ~50 coefficients with largest magnitude were selected. Duplicates were removed and one arrived at ~120 selected loci to be used in the next step.

In the second step, the information on the selected loci was used as input to the asymmetric island model as implemented in iSource. The method estimated relative posterior probabilities for each test isolate to originate from the different sources. The output is a matrix of posterior probabilities of each human isolate originating from each of the putative sources. In this model the posterior distribution was simulated using a Markov chain – Monte Carlo approach and model parameters for migration, mutation and recombination rate were simulated by a nested MCMC approach. This model has already been applied to source attribution in the case of *Campylobacter.*
As there is, even on the cgMLST level, little genetic variation for *Salmonella* compared to *Campylobacter*, the information content for source attribution per locus was low. At the same time, the cgMLST consists of too many loci for the model to be directly applicable. This made the variable selection step necessary to select informative loci and apply the method in the context of *Salmonella.*

Validation

The RKI evaluated the performance of the model on the remaining „test“ isolate set aside for this purpose, resulting in ~30% misclassification rate (if one selects the source by largest posterior probability). Rerunning with randomized training/test set gave similar results. The method will be further validated with data from other European partners.

Network based approach (UNIBO)

Principle

In the Network-based approach, distances between genomes are analysed in order to identify the primary source they belong to. Distance matrices based on SNP, cgMLST and wgMLST were considered as weighted networks, in which each node represents a genome and the weight link represents the inverse of the distance between them: $w = 1/d$. Subsequently, a thresholding was applied in order to keep as connected only the closest nodes: $w > t$. The choice of threshold t was made by looking at weight count distributions (see Figure 6 as an example for Danish data set).

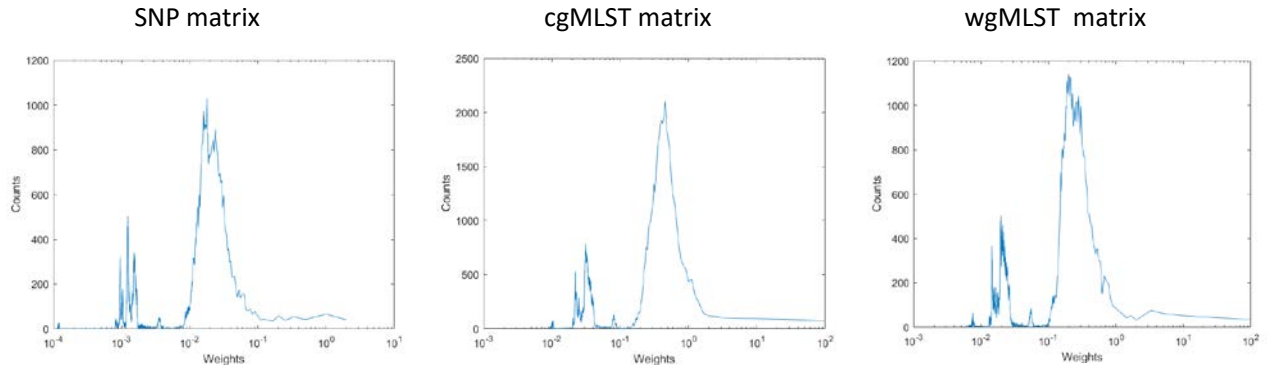

FIGURE 6 WEIGHT COUNT DISTRIBUTION FOR SNP, CGMLST AND WGMLST MATRICES OF THE **DANISH DATA SET.**

Validation

We validated our method by:

- applying the network-based approach **on different measure types for the same dataset**, namely SNP, cgMLST and wgMLST distance matrices (Table 3);
- validating the thresholding procedure obtained on the Danish dataset **onto another dataset**, namely the German data set (see Table 3);

As we can see from results shown in Table 3, we obtained good clustering performances on Danish primary sources, independently on the type of measure used, and different clustering performances on German primary sources, depending on the inclusion of cattle sources or not (note: in the German dataset cattle are 69 out of 190 primary sources, while in the Danish data set they are 2 out of 210).

TABLE 3 CLUSTERING PERFORMANCE MEASURED AS PERCENTAGE OF COHERENT SOURCE CLUSTERING.

Data set	SNP matrix	cgMLST matrix	wgMLST matrix	
Danish	89 %	90 %	90 %	
German	62 %	52 %	52 %	With cattle
German	81 %	80 %	80 %	Without cattle

We will further test our method with the following approaches:

- A further internal crossvalidation can be obtained inside each dataset (DE, DK, FR, UK) considering a 70%-30% crossvalidation for each measure (SNP, cgMLST and wgMLST) repeated multiple times, and then calculating the ROC curve as a function of threshold t;
- Moreover, we plan to perform a comparison between the output of the phylogenetic SNP tree (obtained via CSI phylogeny toolbox) and the network obtained through our thresholding procedure.

Conclusion

In conclusion, we have shown that source attribution methods based on sequence data are applicable and results obtained acceptable.

Application of the described source attribution methods to different datasets will further investigate which methods are best suited for different types of datasets.

References

Anonymous, 2015. Annual Report on Zoonoses in Denmark 2014, National Food Institute, Technical University of Denmark.

Anonymous, 2017. Annual Report on Zoonoses in Denmark 2016, National Food Institute, Technical University of Denmark.

Anonymous, 2018. Annual Report on Zoonoses in Denmark 2017, National Food Institute, Technical University of Denmark.

ANSES, 2017. Attribution des sources des maladies infectieuses d'origine alimentaire. https://www.anses.fr/fr/system/files/BIORISK2015SA0162Ra.pdf.

Brynildsrud, O., Bohlin, J., Scheffer, L., Eldholm, V., 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome biology 17(1), 238.

Dearlove, B.L., Cody, A.J., Pascoe, B., Méric, G., Wilson, D.J., Sheppard, S.K., 2016. Rapid host switching in generalist Campylobacter strains erodes the signal for tracing human infections. The ISME journal 10(3), 721.

de Knegt, L. V., Pires, S. M., Löfström, C., Sørensen, G., Pedersen, K., Torpdahl, M., Nielsen, E. M., Hald, T. (2016). 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. Risk Analysis, 36(3), p.571-588. http://dx.doi.org/10.1111/risa.12483

Hald, T., Vose, D., Wegener, H.C., Koupeev, T., 2004. A Bayesian Approach to Quantify the Contribution of Animal-Food Sources to Human Salmonellosis - Hald - 2004 - Risk Analysis - Wiley Online Library 24.

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

Kittl, S., Heckel, G., Korczak, B.M., Kuhnert, P., 2013. Source attribution of human *Campylobacter* isolates by MLST and Fla-typing and association of genotypes with quinolone resistance. PLoS ONE 8(11).

Kovac, J., Stessl, B., Čadež, N., Gruntar, I., Cimerman, M., Stingl, K., Lušicky, M., Ocepek, M., Wagner, M., Smole Možina, S., 2017. Population structure and attribution of human clinical Campylobacter jejuni isolates from central Europe to livestock and environmental sources. Zoonoses and Public Health 65(1), 51-58.

Lemey, P., Rambaut, A., Drummond, A.J., Suchard, M.A., 2009. Bayesian phylogeography finds its roots. PLoS computational biology 5(9), e1000520.

Mughini-Gras, L., Kooh, P., Augustin, J.-C., David, J., Fravalo, P., Guillier, L., Jourdan-Da Silva, N., Thébault, A., Sanaa, M., Watier, L., 2018. Source attribution of foodborne diseases: potentialities, hurdles and future expectations. Frontiers in microbiology 9, 1983.

Nielsen, E.M., Björkman, J.T., Kiil, K., Grant, K., Dallman, T., Painset, A., Amar, C., Roussel, S., Guillier, L., Félix, B., Rotariu, O., Perez‐Reche, F., Forbes, K., Strachan, N., 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 Supporting Publications 14(2).

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T., Fookes, M., Falush, D., Keane, J.A., Parkhill, J., 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31(22), 3691-3693.

Pires, S.M., Evers, E.G., van Pelt, W., Ayers, T., Scallan, E., Angulo, F.J., Havelaar, A., Hald, T., 2009. Attributing the human disease burden of foodborne infections to specific sources. Foodborne pathogens and disease 6(4), 417-424.

Pires, S.M., Vigre, H., Makela, P., Hald, T., 2010. Using outbreak data for source attribution of human salmonellosis and campylobacteriosis in Europe. Foodborne pathogens and disease 7(11), 1351-1361.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155(2), 945-959.

Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30(14), 2068-2069.

Sheppard, S.K., Dallas, J.F., Strachan, N.J.C., MacRae, M., McCarthy, N.D., Wilson, D.J., Gormley, F.J., Falush, D., Ogden, I.D., Maiden, M.C.J., Forbes, K.J., 2009. *Campylobacter* Genotyping to Determine the Source of Human Infection. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 48(8), 1072-1078.

Thépault, A., Méric, G., Rivoal, K., Pascoe, B., Mageiros, L., Touzain, F., Rose, V., Béven, V., Chemaly, M., Sheppard, S.K., 2017. Genome-wide identification of host-segregating epidemiological markers for source attribution in Campylobacter jejuni. Applied and environmental microbiology, AEM. 03085-03016.

Thépault, A., Rose, V., Quesne, S., Poezevara, T., Béven, V., Hirchaud, E., Touzain, F., Lucas, P., Méric, G., Mageiros, L., Sheppard, S.K., Chemaly, M., Rivoal, K., 2018. Ruminant and chicken: important sources of campylobacteriosis in France despite a variation of source attribution in 2009 and 2015. Scientific Reports 8(1), 9305.

Wilson DJ, et al. Tracing the source of campylobacteriosis. PLoS Genetics. 2008;4:e1000203. doi: 10.1371/journal.pgen.1000203.