Source Attribution of Human Campylobacteriosis Using Whole-Genome Sequencing Data and Network Analysis

Wainaina, Lynda; Merlotti, Alessandra; Remondini, Daniel; Henri, Clementine; Hald, Tine; Njage, Patrick Murigu Kamau

Published in:
Pathogens

Link to article, DOI:
10.3390/pathogens11060645

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

pathogens

MDPI

Article

Source Attribution of Human Campylobacteriosis Using Whole-Genome Sequencing Data and Network Analysis

Lynda Wainaina [1], Alessandra Merlotti [2], Daniel Remondini [2], Clementine Henri [3], Tine Hald [4] and Patrick Murigu Kamau Njage [3,*]

[1] Department of Mathematics, University of Padova, 35121 Padova, Italy; lyndanduta.wainaina@studenti.unipd.it
[2] Department of Physics and Astronomy, University of Bologna, 40126 Bologna, Italy; alessandra.merlotti2@unibo.it (A.M.); daniel.remondini@unibo.it (D.R.)
[3] Research Group for Genomic Epidemiology, National Food Institute, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark; clehen@food.dtu.dk
[4] Research Group for Foodborne Pathogens and Epidemiology, National Food Institute, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark; tiha@food.dtu.dk
* Correspondence: panj@food.dtu.dk

Abstract: *Campylobacter* spp. are a leading and increasing cause of gastrointestinal infections worldwide. Source attribution, which apportions human infection cases to different animal species and food reservoirs, has been instrumental in control- and evidence-based intervention efforts. The rapid increase in whole-genome sequencing data provides an opportunity for higher-resolution source attribution models. Important challenges, including the high dimension and complex structure of WGS data, have inspired concerted research efforts to develop new models. We propose network analysis models as an accurate, high-resolution source attribution approach for the sources of human campylobacteriosis. A weighted network analysis approach was used in this study for source attribution comparing different WGS data inputs. The compared model inputs consisted of cgMLST and wgMLST distance matrices from 717 human and 717 animal isolates from cattle, chickens, dogs, ducks, pigs and turkeys. SNP distance matrices from 720 human and 720 animal isolates were also used. The data were collected from 2015 to 2017 in Denmark, with the animal sources consisting of domestic and imports from 7 European countries. Clusters consisted of network nodes representing respective genomes and links representing distances between genomes. Based on the results, animal sources were the main driving factor for cluster formation, followed by type of species and sampling year. The coherence source clustering (CSC) values based on animal sources were 78%, 81% and 78% for cgMLST, wgMLST and SNP, respectively. The CSC values based on *Campylobacter* species were 78%, 79% and 69% for cgMLST, wgMLST and SNP, respectively. Including human isolates in the network resulted in 88%, 77% and 88% of the total human isolates being clustered with the different animal sources for cgMLST, wgMLST and SNP, respectively. Between 12% and 23% of human isolates were not attributed to any animal source. Most of the human genomes were attributed to chickens from Denmark, with an average attribution percentage of 52.8%, 52.2% and 51.2% for cgMLST, wgMLST and SNP distance matrices respectively, while ducks from Denmark showed the least attribution of 0% for all three distance matrices. The best-performing model was the one using wgMLST distance matrix as input data, which had a CSC value of 81%. Results from our study show that the weighted network-based approach for source attribution is reliable and can be used as an alternative method for source attribution considering the high performance of the model. The model is also robust across the different *Campylobacter* species, animal sources and WGS data types used as input.

Keywords: source attribution; *Campylobacter*; campylobacteriosis; network analysis; whole-genome sequencing; coherence source clustering

Citation: Wainaina, L.; Merlotti, A.; Remondini, D.; Henri, C.; Hald, T.; Njage, P.M.K. Source Attribution of Human Campylobacteriosis Using Whole-Genome Sequencing Data and Network Analysis. *Pathogens* **2022**, *11*, 645. https://doi.org/10.3390/pathogens11060645

Academic Editor: Lawrence S. Young and Bart C. Weimer

Received: 1 April 2022
Accepted: 28 May 2022
Published: 3 June 2022

1. Introduction

Human campylobacteriosis is among the most common zoonotic diseases, caused mainly by the bacteria *Campylobacter jejuni* and *Campylobacter coli*. Campylobacteriosis continues to be a major problem worldwide, including Denmark, which has seen the number of cases rising from 4547 in 2018 to 5389 in 2019. The increase in cases in Denmark was attributed to a large outbreak in chicken meat [1,2]. The main sources of human infection have been attributed to contaminated meat, poultry, water, milk and contact with farm animals [3]. Considering that many human campylobacteriosis cases have been attributed to various animal sources, there is a need to determine the relative contribution of the different exposures from animals to the total number of human cases [4].

Source attribution, which apportions human infection cases to different animal species and food reservoirs, has been instrumental in control- and evidence-based intervention efforts. Several methods for source attribution are available, including the microbial subtyping approach and comparative exposure assessment approach. Microbial subtyping involves characterizing isolates of specific pathogens by phenotypic and genotypic subtyping methods. The principle for this approach involves comparing isolates from different food and animal sources with those from humans. The comparative exposure assessment approach, on the other hand, determines the relative importance of the known transmission routes by estimating the human exposure to the pathogen through each route [5].

The microbial subtyping attribution approach has been proven to be a valuable source attribution method as it assumes that the distribution of subtypes in the collection of microbial isolates for each source used in the attribution exercise is similar to the true distribution of subtypes in each source. There are two main types of microbial subtyping attribution models: the frequency-matched attribution model, which compares human strain types in the sources and population genetic models based on modeling the organism's evolutionary history [6,7]. Previous studies have reported several applications of the microbial subtyping approach including source attribution of human salmonellosis which was developed in Denmark [4,5,8]. The use of multilocus sequence typing (MLST) is another common example of the microbial subtyping approach, which has been used to identify lineages in bacterial populations by indexing the variation present in seven housekeeping genes located in various parts of the chromosome [9]. MLST data have been previously utilized to attribute the sources of human *C. jejuni* infections in New Zealand, as well as *Salmonella* in Denmark, using the Danish Salmonella source account model and the ClonalFrame algorithm [4,9].

Whole-genome sequencing (WGS) has been proven to be the most informative approach for the characterization of bacterial isolates and has been used to analyze multiple bacterial outbreaks, such as tuberculosis, listeriosis and salmonellosis, among others [10–12]. WGS data sets have become increasingly available. However, one of the limitations of WGS data is the complexity in data analysis due to variable gene content and difficulties interpreting obtained results [13]. Despite this, many studies have suggested approaches to overcome the limited discriminatory power of MLST by exploiting WGS data. These approaches can be grouped into methods based on the core genome or whole genome multilocus, termed gene-by-gene approaches and single-nucleotide polymorphism (SNP) detection, which segregate by host [14]. The gene-by-gene approaches assess the diversity of isolates based on alleles found for all wgMLST or cgMLST genes of the species of interest [15] while SNP-based methods distinguish isolates based on SNPs present in the entire genome, including the intergenic regions, potentially offering a higher resolution [16,17].

Different approaches have been used for source attribution using WGS data sets, including machine learning which has previously been applied in source attribution for *Salmonella enterica*, *Escherichia coli* and *C. jejuni* [13,18–20] and to predict the severity or outcome of microbial infections [21–25]. The machine learning approach involves training different algorithms and obtaining the best-performing model while obtaining the attribution probabilities of human isolates to different sources. Network analysis, on the other hand, has recently been demonstrated as an accurate approach for the source at-

tribution of human salmonellosis [4]. Network analysis is based on weighted networks theory, where pairwise distance matrices from source attribution can be visualized as fully connected networks. Nodes in this theory correspond to *Campylobacter* isolates and links correspond to genetic distances. Weaker links imply greater genetic distance between isolates. Network analysis is useful in extracting network communities corresponding to different animal sources, where network communities correspond to groups of vertices with a higher probability of being connected to each other than other members of that group [26].

The probability of a human isolate to be associated with an animal source is computed as the function of the number of links that the human isolate has with other animal isolates. A specific animal source to which human genomes of *Campylobacter* are attributed can also be extracted from the network analysis. Using the network approach, we can identify which structural features of a data set play a fundamental role in determining the internal coherence of clusters [4], such as animal sources, species type and year of origin, etc. We demonstrated the potential of weighted networks for source attribution of human campylobacteriosis using whole-genome sequencing data. We compared the effect of different types of WGS data inputs namely cgMLST, wgMLST and SNP on the accuracy of the weighted network-based source attribution models.

2. Materials and Methods

2.1. Data Set

The data used in this study were collected between May 2015 and March 2017 from *Campylobacter* monitoring projects in Denmark. The data set was composed of 283 *C. jejuni* isolates and 434 other unknown *Campylobacter* species isolated from chickens, cattle, pigs, dogs, ducks and turkeys. The test material used were intestinal content (swabs, stools or appendices) and meat from various products collected either in a slaughterhouse or in the retail trade, originating from Danish or foreign production as well as the production environment. The *Campylobacter* isolates' metadata were obtained from the Danish Food and Veterinary Administration (foedevarestyrelsen) and the sequenced genomes were extracted from the Center for Genetic Epidemiology (Food Institute Section of Computerome). The following information from the databases was used: sample ID, year of collection, country of origin and source (host of the *Campylobacter* isolate).

The human cases data set consisted of isolates received from Statens Serum Institute's surveillance from January 2015 to December 2017. Isolates from humans with known travel history were not included in the data set. Data cleaning was performed to remove duplicates and isolates with incomplete metadata. The input data set for the network analysis consisted of cgMLST and wgMLST distance matrices from 717 food and 717 human isolates and SNPs from 720 food and 720 human isolates. The SNP distance matrix contained more isolates than cgMLST and wgMLST due to more matching isolate identification codes between the SNP data and the source metadata. The population structure was obtained from the phylogenetic analysis, as shown in Figure 1. This indicated that human isolates intermixed with other food and animal isolates, indicating that human *Campylobacter* strains were more likely to have originated from the sources described. Furthermore, sources were also genetically well distributed within the tree. Seawater and vegetables were omitted from the final input data used for network analysis since they were few and would have resulted in unreliable source attribution of human *Campylobacter* cases. All isolates used can be found under the bioproject number set up by the Statens Serum Institute, PRJEB31119 [27].

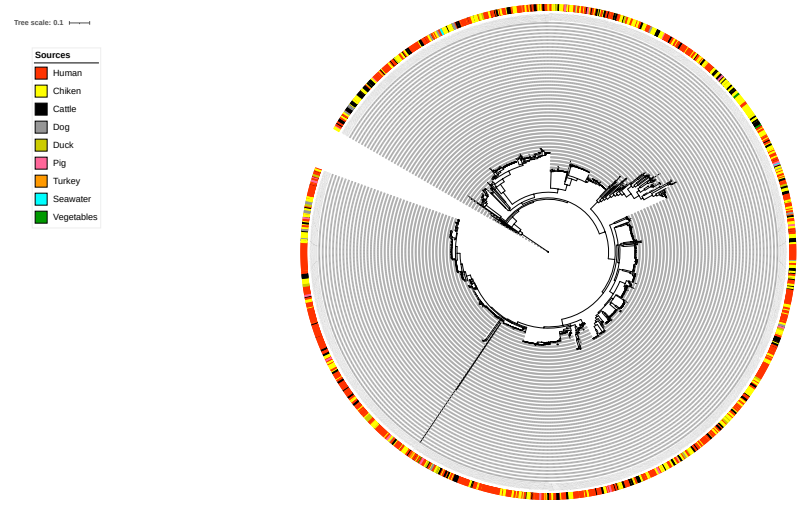

Figure 1. Phylogenetic tree from SNP distance matrix.

2.2. Bioinformatics Analysis

2.2.1. Assemblies

The raw reads were de novo assembled. The procedure was done using the Food QC & Assembly pipeline, which includes assembler SPAdes version 3.9 [28]. The quality of the assembly was assessed using the number of contigs, N50, and the total size of the assembly. Assemblies were scaffold assemblies; genome assemblies with less than 500 contigs were kept in the data set. Eventually, the total size of the assembly was checked to match the expected size for a *C. jejuni* genome, which is between 1.6 to 1.7 million base pairs (Mbp).

2.2.2. cgMLST and wgMLST

Core genome multilocus sequence typing (cgMLST) compares allelic profiles of several loci. CgMLST includes the core genome of *Campylobacter* and contains 1343 genes, as defined by Cody et al. in 2017 [29]. We performed core genome multilocus sequence typing (cgMLST) using the scheme developed by Cody *et al.* [29] available from the Center for Genomic Epidemiology pipeline [30]. Similarly, the wgMLST scheme used in this work includes 1643 genes from the re-annotation of the genome sequence of reference *C. jejuni* genome NCTC 11,168 [31].

2.2.3. SNP

The SNP matrix was built using the CSI phylogeny pipeline accessible from the Center for Genomic Epidemiology [30,32]. The paired-end reads were mapped to the reference genomes using Burrows-Wheeler Aligner (BWA) [33]. The SNP analysis was performed using the reference genome: *C. jejuni* subsp. *jejuni* NCTC 11168 = ATCC 700819 (accession NC 002163.1). SNPs were determined using mpileup commands from SAMTools version 0.1.18 [34,35]. The SNPs were filtered according to five parameters: a minimum distance of 10 bps between each SNP, a minimum of 10x depth and 10 percent of the breadth coverage, the mapping quality was above 30, the SNP quality was higher than 20 and all indels were excluded [28,29,32–34]. For each genome, SNPs were concatenated to a single alignment corresponding to the positions of the reference genome. ItoL version 6 was used for the visualization of the phylogenetic tree, where the number of SNPs between isolates is equivalent to the distance in the tree [36].

All bioinformatic analysis were performed using Danish National Supercomputer for Life Sciences, Computerome 2.0, a local server for a Linux-based command-line system [37].

2.3. Network Analysis

The weighted network approach was used in this study, where the pairwise distance matrix was represented as a network with nodes corresponding to human *Campylobacter*

isolates and links as a function of the pairwise distance. This pairwise distance was calculated as the number of different MLST alleles or SNPs between two isolated sequences. The assumption is that genomes coming from the same source show smaller distances. A fully connected weighted network, with weight calculated as 1/distance assigned to each link, was built in MATLAB [38]. A threshold was applied such that, in the resulting binary network, nodes were connected by an edge if the weight was greater than the threshold. The threshold was applied to remove weaker links with larger genetic distances, and it was chosen to maximize the internal coherence of clusters and minimize the number of isolated nodes. In the resulting binarized network, nodes were linked with an edge only if their weight was greater than the threshold value and clusters identified using the thresholding procedure [4].

The best threshold values were obtained using a 70/30 cross-validation procedure on the animal source data and were chosen in order to maximize the internal coherence of clusters (CSC, Equation (1) [4]) and minimize the number of isolated nodes. The 70/30 cross-validation procedure involved randomly selecting a network training set consisting of 70% of animal origin samples and using this set to obtain a best threshold value. This threshold value was applied to the network constructed using the test set composed of the remaining 30% of the animal samples for the calculation of the CSC as shown in Equation (1). This procedure was repeated 100 times and the most frequent threshold value was selected as the best overall threshold for further use in source clustering. The best threshold was then applied to the full pairwise distance matrix consisting of both animal and human isolates such that the human sources could be attributed to specific animal sources [4]. The best threshold was used to maximize the score function on distance matrices, as shown in Equation (2) [4]. The graphical visualizations of the network were obtained using the MATLAB 'Plot' function with the force-directed graph layout [39].

$$CSC = \frac{\sum_{i=1}^{N_c} TP_i}{\sum_{i=1}^{N_c} T_i} 100 \quad (1)$$

$$Score = \left(1 - \frac{N_{ISO}}{N_{TOT}}\right) CSC \quad (2)$$

N_{TOT} is the total number of nodes in the network, while N_{ISO} is the number of isolated nodes that do not have any links to other nodes. CSC is the coherent source clustering, which measures the algorithm's clustering performance, where TP_i is the number of true positives in the i^{th} cluster (majority of isolates from the same source in the same cluster) and T_i is the total number of nodes inside the i^{th} cluster [4].

3. Results and Discussion

We compared results obtained using cgMLST, wgMLST and SNP distance matrices from the network analysis. Figure 2 shows the distribution of input data which corresponded to Figure 3 indicating the mean percentage attribution probability for the three distance matrices. We observed that chickens from Denmark were the main sources of human campylobacteriosis cases, with a percentage of attribution of 52.84%, 52.17% and 51.22% for cgMLST, wgMLST and SNP, respectively, while ducks from Denmark were the least probable source of infection. These results are in harmony with previous reports showing chicken meat as the main source of campylobacteriosis in Denmark [3]. The best threshold values obtained from the cross-validation were 0.1141, 0.0105 and 1715 for cgMLST, wgMLST and SNP distance matrices, respectively (Table 1). These values were used to maximize the score function from 100 runs of cross-validation.

The network-based method achieved 78%, 81% and 78% coherent source clustering for cgMLST, wgMLST and SNP distance matrices, respectively (Table 1). The results indicated that animal sources were the main factors driving the clustering, followed by type of *Campylobacter* species and finally year of origin (Table 2). Table 3 shows results from adding human isolates to the network, where 88%, 77% and 88% were clustered within the existing

animal network for cgMLST, wgMLST and SNP, respectively, while the remaining isolates were not to linked to *Campylobacter* genomes from any of the animal sources. The algorithm performed reasonably well in source attribution. However, some isolates were wrongly classified. For example, 43 chicken isolates were classified as cattle isolates Table 4. This misclassification is also apparent in Figure 4, where different sources in the main clusters 1, 4 and 5 from network analysis using cgMLST distance matrix as input data cannot be clearly distinguished. A consideration of the country of origin of animal sources showed that regionality affects cluster formation, as seen in Figures 5–7, where most isolates are clustered according to origin.

We noted that despite the class imbalance in the input data (Figure 2), the less abundant sources, such as dogs from Denmark, still had 100% human isolates linked to the sources, as shown in Figures 4, 8 and 9. This is an indication that sample imbalance does not affect source attribution using the network analysis method [4] and that the most consumed animal sources are most likely to cause the majority of *Campylobacter* infections. Class imbalances in other models lead to important patterns in the predictors being associated with the larger classes which results in less predictions for classes with less samples [40]. The best-performing model was the one with the wgMLST distance matrix as the input data, which had a CSC value of 81%. We calculated the confusion matrix for the cgMLST, wgMLST and SNP distance matrices' clustering results (Tables 4–6). The weighted network analysis approach provided quite good results considering the model performance in comparison to other source attribution models [13,18–20] and microbial infection severity and outcome prediction models such as machine learning [21–25].

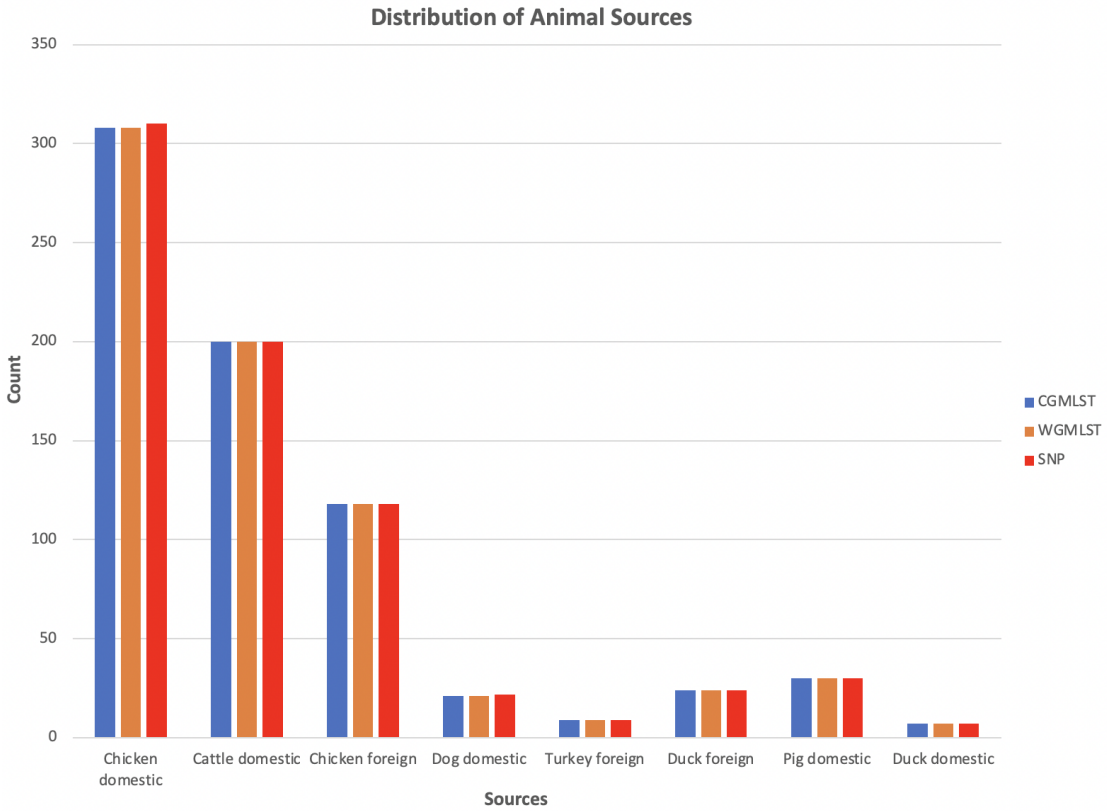

Figure 2. Distribution of animal sources used as input data for network analysis.

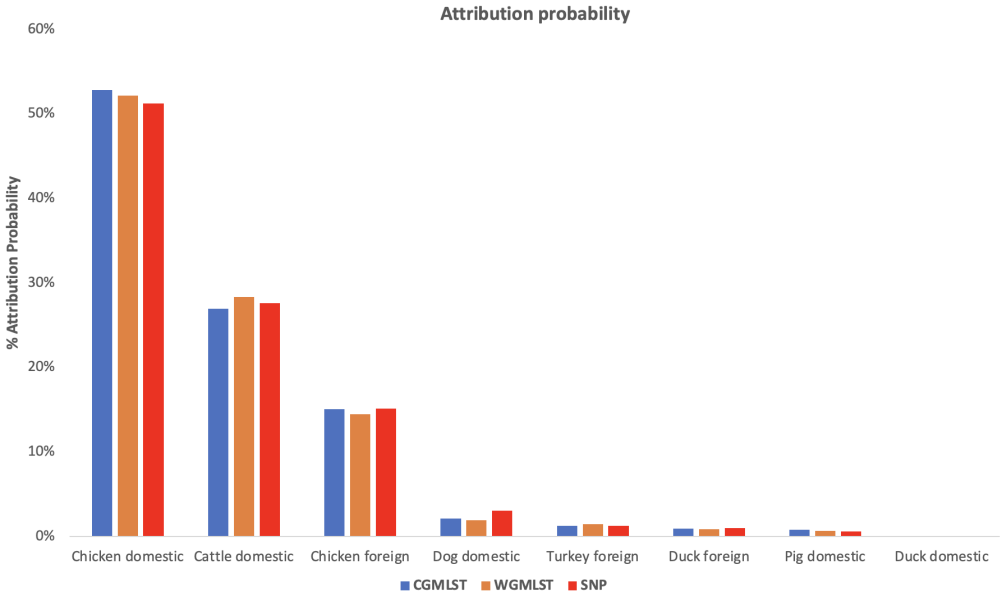

Figure 3. Mean probability (in percentage) of a human isolate to be attributed to a source, calculated for each of the considered pairwise distance matrices (cgMLST, wgMLST and SNP).

Table 1. Best threshold based on the animal of origin for networks based on SNP, cgMLST, and wgMLST distance matrices.

Performance	cgMLST	wgMLST	SNP
Best threshold	0.1141	0.0105	1715

Table 2. Coherent source clustering (CSC) according to country of origin, type of *campylobacter* and year of origin for networks based on SNP, cgMLST and wgMLST distance matrices.

CSC	cgMLST	wgMLST	SNP
Species	78%	79%	69%
Year	61%	64%	67%

Table 3. Number of attributed and not attributed human isolates.

Human Isolates	cgMLST	wgMLST	SNP
Attributed	632	558	633
Not attributed	85	159	86

Table 4. Confusion matrix obtained from source clustering results for cgMLST distance matrix.

True / Pred	Cattle.dk	Chkn.dk	Chkn.for	Dog.dk	Duck.dk	Duck.for	Pig.dk	Turkey.for
Cattle.dk	151	32	1	2	0	1	1	4
Chkn.dk	43	259	30	13	0	4	2	0
Chkn.for	2	20	87	0	0	1	0	2
Dog.dk	0	0	0	5	0	0	0	0
Duck.dk	0	0	0	0	7	0	0	0
Duck.for	0	0	0	0	0	17	0	0
Pig.dk	0	0	0	0	0	0	29	0
Turkey.for	0	0	0	0	0	0	0	4

True—true isolates of the same source; Pred - predicted isolates; Cattle.dk—cattle from Denmark; Chkn.dk—chickens from Denmark; Chkn.for—chickens from foreign countries; Dog.dk—dogs from Denmark; Duck.dk—ducks from Denmark; Duck.for—ducks from foreign countries; Pig.dk—pigs from Denmark; Turkey.for—turkeys from foreign countries.

Table 5. Confusion matrix obtained from source clustering results for wgMLST distance matrix.

True / Pred	Cattle.dk	Chkn.dk	Chkn.for	Dog.dk	Duck.dk	Duck.for	Pig.dk	Turkey.for
Cattle.dk	157	31	2	2	0	1	1	2
Chkn.dk	37	263	24	9	0	2	1	0
Chkn.for	2	14	92	0	0	2	0	2
Dog.dk	0	0	0	9	0	0	0	0
Duck.dk	0	0	0	0	7	0	0	0
Duck.for	0	0	0	0	0	19	0	0
Pig.dk	1	0	0	0	0	0	28	0
Turkey.for	2	1	0	1	0	0	0	5

True—true isolates of the same source; Pred—predicted isolates; Cattle.dk—cattle from Denmark; Chkn.dk—chickens from Denmark; Chkn.for—chickens from foreign countries; Dog.dk—dogs from Denmark; Duck.dk—ducks from Denmark; Duck.for—ducks from foreign countries; Pig.dk—pigs from Denmark; Turkey.for—turkeys from foreign countries.

Table 6. Confusion matrix obtained from source clustering results for SNP distance matrix.

True / Pred	Cattle.dk	Chkn.dk	Chkn.for	Dog.dk	Duck.dk	Duck.for	Pig.dk	Turkey.for
Cattle.dk	158	36	4	4	0	2	1	4
Chkn.dk	39	253	27	13	0	3	1	0
Chkn.for	2	20	89	0	0	1	0	2
Dog.dk	0	0	0	5	0	0	0	0
Duck.dk	0	0	0	0	7	0	0	0
Duck.for	0	0	0	0	0	18	0	0
Pig.dk	1	0	0	0	0	0	28	0
Turkey.for	0	0	0	0	0	0	0	3

True—true isolates of the same source; Pred—predicted isolates; Cattle.dk—cattle from Denmark; Chkn.dk—chickens from Denmark; Chkn.for—chickens from foreign countries; Dog.dk—dogs from Denmark; Duck.dk—ducks from Denmark; Duck.for—ducks from foreign countries; Pig.dk—pigs from Denmark; Turkey.for—turkeys from foreign countries.

The F1 scores calculated from the confusion matrices above were: 75.96%, 79.94% and 74.93% for cgMLST, wgMLST and SNP, respectively. The best-performing model from the F1 score was based on wgMLST distance matrix as input data which is also in agreement to the model's high CSC value of 81%.

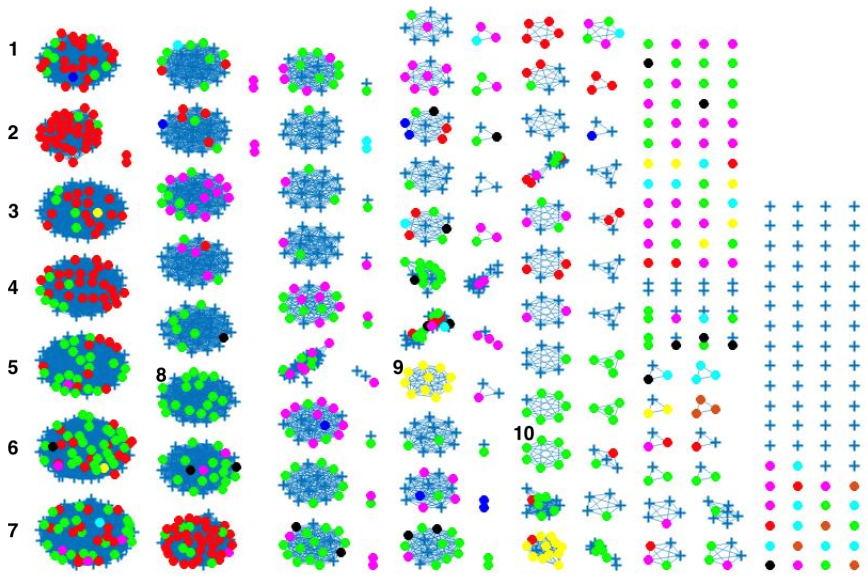

Figure 4. Source clustering results (force-directed graph drawing algorithm) obtained using cgMLST distance matrix as model input. Nodes represent different animal isolates. Cluster number 1–7 (misclassification

of isolates within the cluster), 8–10 (correct classification of isolates within the cluster). Legend: red—cattle from Denmark ; green—chickens from Denmark; magenta—chickens from foreign countries; black—dogs from Denmark ; dark blue—turkeys from foreign countries; yellow—pigs from Denmark; cyan—ducks from foreign countries ; light brown—ducks from Denmark; blue crosses—Human isolates. Foreign (Germany, Netherlands, Italy, France, Poland, UK, Hungary).

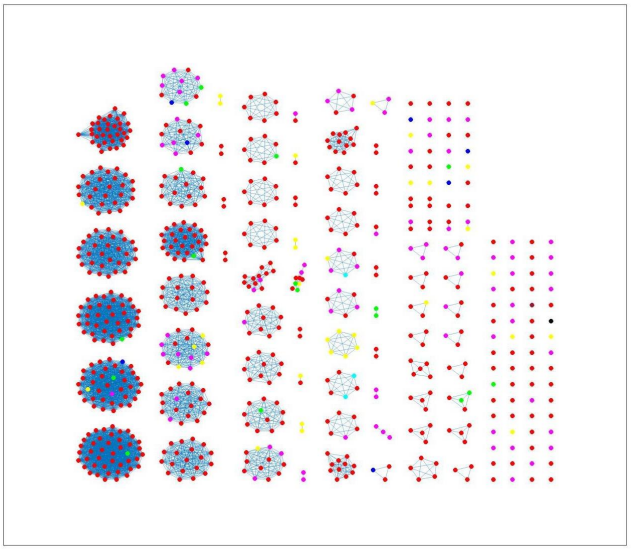

Figure 5. Clustering results (force-directed graph drawing algorithm) obtained using cgMLST distance matrix as model input. Nodes represent country of origin for different animal isolates. Legend: red—Denmark; yellow—Poland; magenta—France; green—Germany; blue—Netherlands; cyan—Italy; black—UK; brown—Hungary.

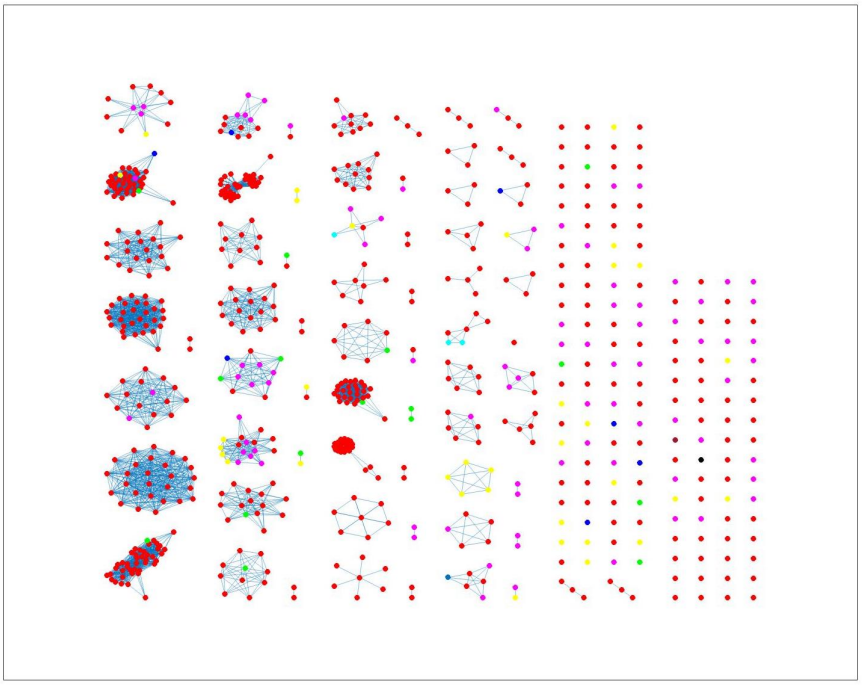

Figure 6. Clustering results (force-directed graph drawing algorithm) obtained using wgMLST distance matrix as model input. Nodes represent country of origin for different animal isolates. Legend: red—Denmark; yellow—Poland; magenta—France; green—Germany; blue—Netherlands; cyan—Italy; black—UK; brown—Hungary.

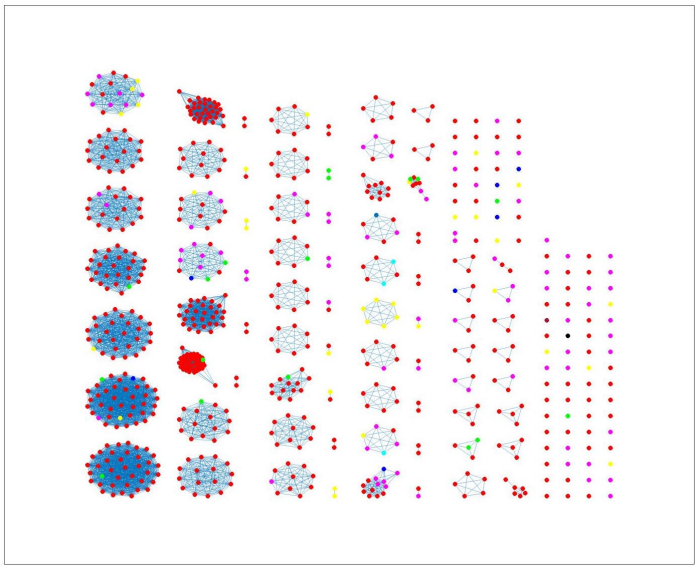

Figure 7. Clustering results (force-directed graph drawing algorithm) obtained using SNP distance matrix as model input. Nodes represent country of origin for different animal isolates. Legend: red—Denmark; yellow—Poland; magenta—France; green—Germany; blue—Netherlands; cyan—Italy; black—UK; brown—Hungary.

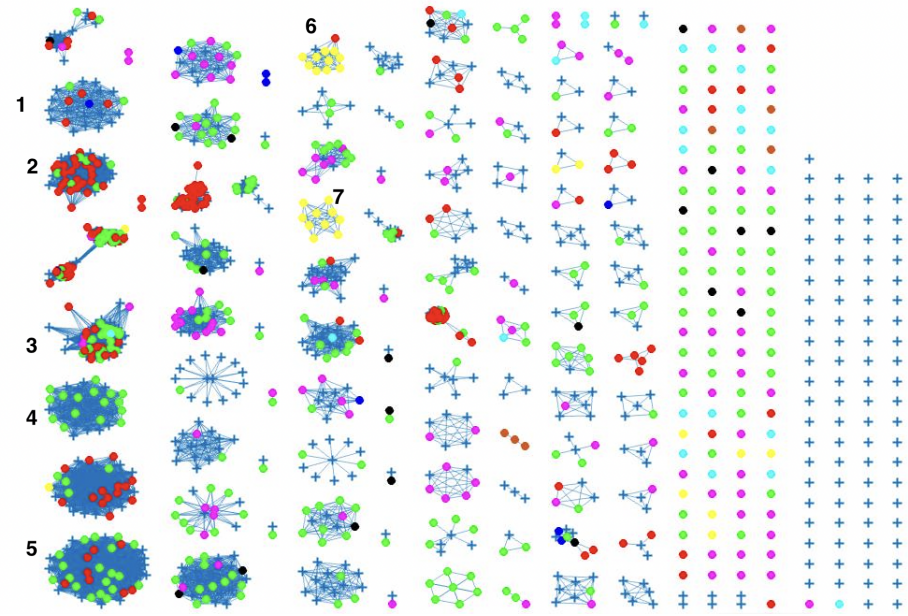

Figure 8. Source clustering results (force-directed graph drawing algorithm) obtained using wgMLST distance matrix as model input. Nodes represent different animal isolates. Cluster number 1–3, 5–6 (misclassification of isolates within the cluster) 4 and 7 (correct classification of isolates within the cluster). Legend: red—cattle from Denmark; green—chickens from Denmark; magenta—chickens from foreign countries; black—dogs from Denmark; dark blue—turkeys from foreign countries; yellow—pigs from Denmark; cyan—ducks from foreign countries; light brown—ducks from Denmark; blue crosses—Human isolates.

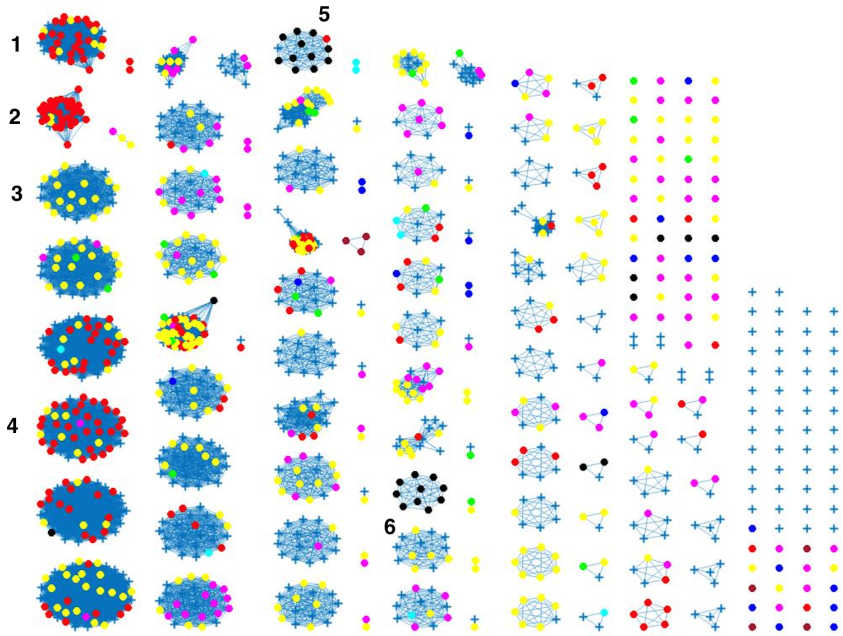

Figure 9. Source clustering results (force-directed graph drawing algorithm) obtained using SNP distance matrix as model input. Nodes represent different animal isolates. Cluster number 1, 2, 4, 5 (misclassification of isolates within the cluster), 3 and 6 (correct classification of isolates within the cluster). Legend: red—cattle from Denmark; yellow—chickens from Denmark; magenta—chickens from foreign countries; green—dogs from Denmark; cyan—turkeys from foreign countries; black—pigs from Denmark ; blue—ducks from foreign countries; light brown—ducks from Denmark ; blue crosses—Human isolates.

Similar clustering results were observed from the network analysis approach as observed above. Figure 8 indicates some confusion in distinguishing between different sources. For example, in cluster 2, there is no proper separation between chickens and cattle from Denmark, which is also observed in Figure 9 (clusters 1, 2, 4). However, a high proportion of the food sources where less isolates were available such as pigs, were attributed to human cases as observed in cluster 6 in Figure 8. The results from the wgMLST distance matrix input data show that the network-based algorithm performs best in clustering considering the high CSC value of 81%. The results in Figures 6 and 7 show that the region of origin of the animal sources has an influence on cluster formation. Considering that most of the animal isolates are from Denmark, the main clusters are dominated by Danish isolates, with some clusters consisting of less abundant isolates such as imports from Poland, as observed in Figures 5–7.

The weighted network-based approach showed high specificity due to the number of links between each human sample and each animal source in all three networks (cgMLST, wgMLST and SNP), as observed in Figures 10–12. We also observed 100% attribution of some human samples to less abundant sources, such as dogs from Denmark (Figures 10–12), an indication that the algorithm used was not influenced by the sample size. Results from the network analysis comparing the three distance matrices as inputs suggested that the model is robust to the changes in the form of WGS used as model input (Figures 4, 8, 9 and Tables 4–6). In addition, since there was a class imbalance between isolates from Denmark and imported isolates, the finding that country of origin influenced cluster formation in this analysis should be further investigated using isolates from different countries or regions.

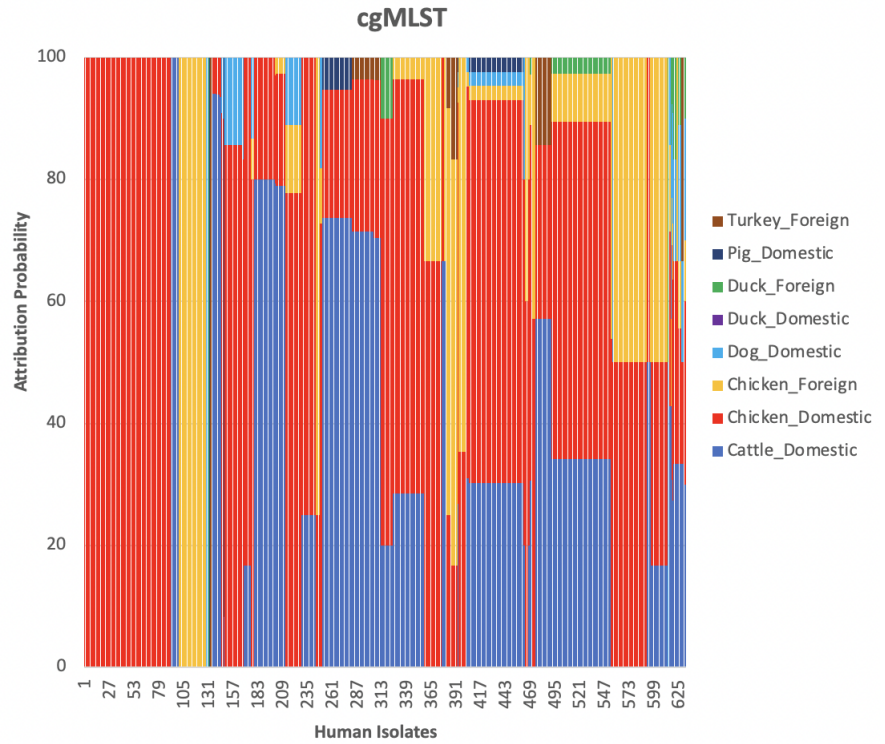

Figure 10. The probability of a human isolate to originate from each source as determined by source attribution analysis using the network-based approach on the cgMLST pairwise distance matrix.

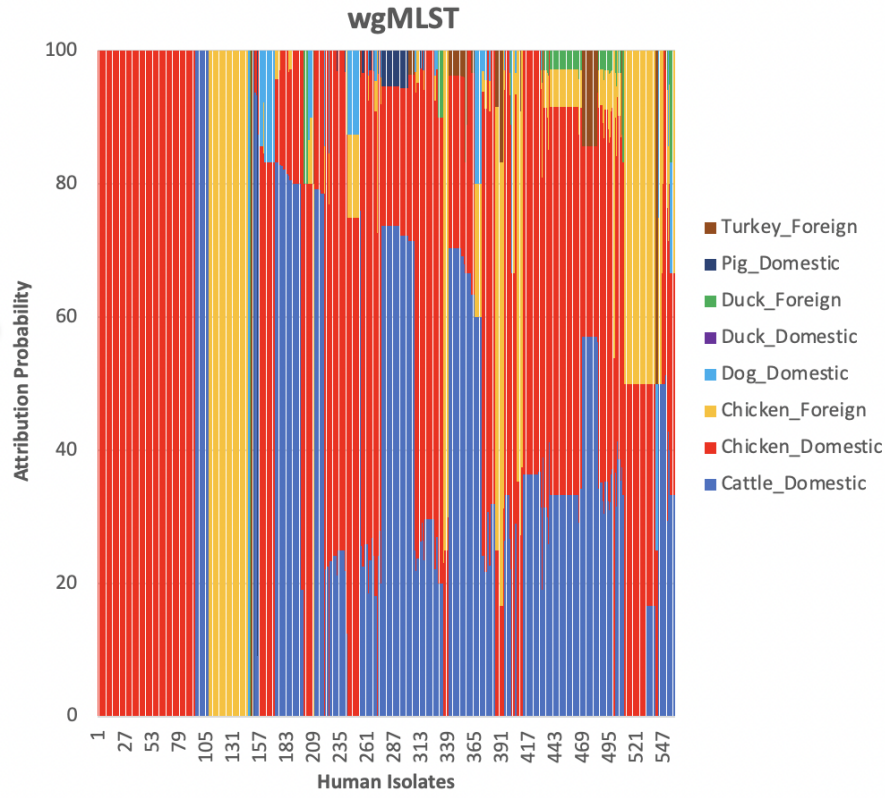

Figure 11. The probability of a human isolate to originate from each source as determined by source attribution analysis using the network-based approach on the wgMLST pairwise distance matrix.

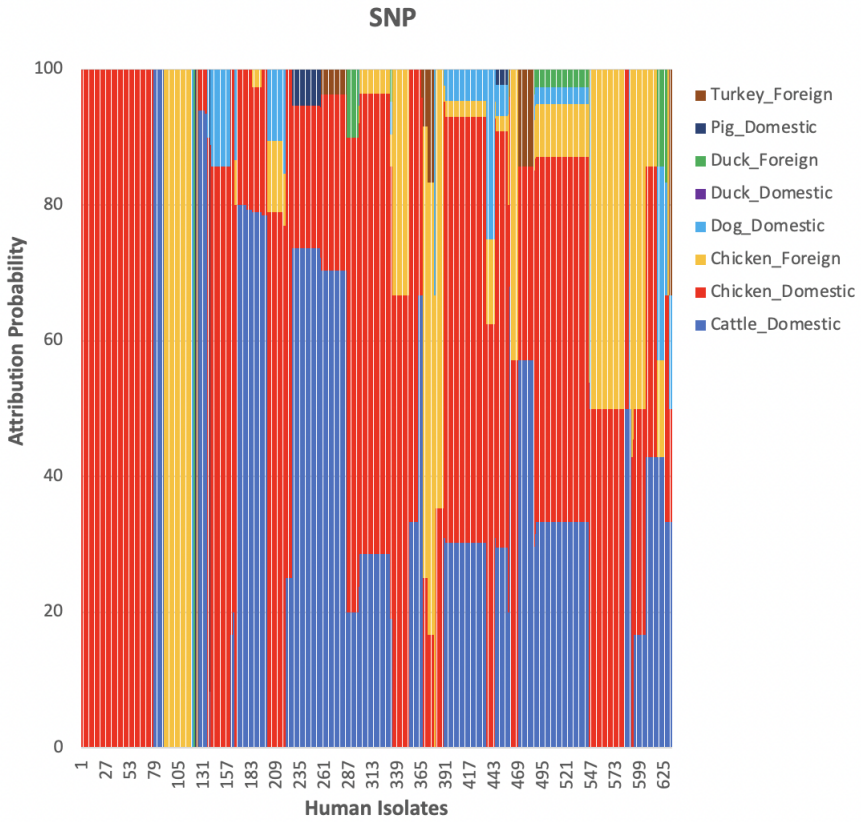

Figure 12. The probability of a human isolate to originate from each source as determined by source attribution analysis using the network-based approach on the SNP pairwise distance matrix.

4. Conclusions

This study aimed at attributing human *Campylobacter* cases to different animal sources using a weighted network-based approach to exploit the potential of WGS data in conducting higher-resolution source attribution. We demonstrated that despite the high intra-species genetic diversity in *Campylobacter* [41], which would result in low discriminatory power in differentiating the different sources [4], the network analysis approach showed good discriminatory power, maximized cluster coherence and reduced the number of human isolates not attributed. The results obtained were robust to the different subtyping data used. However, the wgMLST distance matrix as input data may provide more accurate inputs than cgMLST and SNP, although this results in more human isolates not attributed to any sources. Chickens were the main cause of human *Campylobacter* infections. The analysis based on the country of origin of animal sources indicated that regionality affects cluster formation. Further studies are therefore recommended using data sets from different countries and different potential sources to confirm the reliability of the network-based approach as an alternative for source attribution.

Author Contributions: Conceptualization, P.M.K.N., A.M. and L.W. ; methodology, A.M., L.W., D.R. and P.M.K.N.; software, A.M. and L.W.; validation, A.M., L.W. and P.M.K.N.; formal analysis, L.W., A.M. and P.M.K.N.; data curation, C.H., T.H. and L.W.; writing—original draft preparation, L.W., P.M.K.N. and C.H; writing—review and editing, P.M.K.N. and A.M.; visualization, L.W., A.M. and P.M.K.N.; supervision, P.M.K.N. and T.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used can be found under the bioproject number set up by Statens Serum Institute PRJEB31119 [27].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. *Campylobacter* in Denmark. Available online: https://www.foodsafetynews.com/2020/02/campylobacter-infections-at-record-high-in-denmark/ (accessed on 30 March 2022).
2. Wingstrand, A.; Neimann,J.; Engberg, J.; Nielsen, E.M.; Gerner-Smidt, P.; Wegener, H.C.; Mølba, K. Fresh chicken as main risk factor for campylobacteriosis, Denmark. *Emerg. Infect. Dis.* **2006**, *12*, 280–285. [CrossRef] [PubMed]
3. Sheppard, S.K.; Colles, F.M.; McCARTHY, N.D.; Strachan, N.J.C.; Ogden, I.D.; Forbes, K.J.; Dallas, J.F.; Maiden, M.C.J. Niche segregation and genetic structure of Campylobacter jenuni populations from wild and agricultural host species. *Eur. Pubmed Cent.* **2011**, *20*, 3484–3490.
4. Merlotti, A.; Manfreda, G.; Munck, N.; Hald, T.; Litrup, E.; Nielsen, E.M.; Remondini, D.; Pasquali F. Network Approach to Source Attribution of Salmonella enterica Serovar Typhimurium and Its Monophasic Variant. *Front. Microbiol.* **2020**, *11*, 1205. [CrossRef] [PubMed]
5. Pires, S.M.; Evers, E.E.; Van Pely, W.; Ayers, T.; Scallan, E.; Angulo, F.J.; Havelaar, A.; Hald, T. Attributing the Human Disease Burden of Foodborne Infections to Specific Sources.*Foodborne Pathog. Dis.* **2009**, *6*, 417–424. [CrossRef] [PubMed]
6. Ravel, A.; Hurst, M.; Petrica, N.; David, J.; Mutschall, S.K.; Pintar, K.; Taboada, E.N.;Pollari, F. Source attribution of human campylobacteriosis at the point of exposure by combining comparative exposure assessment and subtype comparison based on comparative genomic fingerprinting. *PLoS ONE* **2017**, *12*, e0183790. [CrossRef]
7. Scientific Opinion of the Panel on Biological Hazards on a request from EFSA on Overview of methods for source attribution for human illness from food borne microbiological hazards. Overview of methods for source attribution for human cases of food borne microbiological hazards. *EFSA J.* **2008**, *764*, 1–43.
8. Hald, T.; Vose, D.; Wegener, H.C.; Koupeev, T. Bayesian approach to quantify the contribution of animal-food sources to human salmonellosis. *Risk Anal.* **2004**, *24*, 251–265. [CrossRef]
9. Dingle, K.E.; Colles, F.M.; Ure, R.; Wagenaar, J.A.; Duim, B.; Bolton, F.J.; Fox, A.J.; Wareing, D.R.A.; Maiden, M.C.J. Molecular characterization of Campylobacter jejuni clones: a rational basis for epidemiological investigations. *Emerg. Infect. Dis.* **2002**, *8*, 949–955. [CrossRef]
10. Inns, T.; Ashton, P.M.; Herrera-Leon, S.; Lighthill, J.; Foulkes, S.; Jombart , T.; Rehman, Y.; Fox, A.; Dallman, T.; Pinna, E.D.E; et al. Prospective use of whole genome sequencing (WGS) detected a multi-country outbreak of Salmonella Enteritidis. *Epidemiol. Infect.* **2017**, *145*, 289–298. [CrossRef]
11. Genestet, C.; Tatai, C.; Berland, J.L.; Claude, J.B.; Westeel, E.; Hodille, E.; Fredenucci, I.; Rasigade, J.P.; Ponsoda, M.; Jacomo, V.; et al. Prospective whole-genome sequencing in tuberculosis outbreak investigation. France, 2017–2018. *Emerg. Infect. Dis.* **2019**, *25*, 589–592. [CrossRef]
12. Schjørring, S.; Lassen, S.G.; Jensen, T.; Moura, A.; Kjeldgaard, J.S.; Müller, L.; Thielke, S.; Leclercq, A.; Maury, M.M.; Tourdjman, M.; et al. Cross-border outbreak of listeriosis caused by cold-smoked salmon, revealed by integrated surveillance and whole genome sequencing (WGS), Denmark and France, 2015 to 2017. *Eurosurveillance* **2017**, *22*, 8–12. [CrossRef] [PubMed]
13. Arning, N.; Sheppard, S.K.; Bayliss, S.; Clifton, D.A.; Wilson, D.J. Machine learning to predict the source of campylobacteriosis using whole genome data. *PLoS Genet.* **2021**, *17*, e1009436. [CrossRef] [PubMed]
14. ECDC. *Expert Opinion on Whole Genome Sequencing for Public Health Surveillance*; ECDC: Stockholm, Sweden; Solna, Sweden, 2016.
15. Maiden, M.C.J.; Rensburg, M.J.J.V.; Bray, J.E.; Earle, S.G.; Ford, S.A.; Jolley, K.A.; McCarthy, N.D. MLST revisited: The gene-by-gene approach to bacterial genomics. *Nat. Rev. Microbiol.* **2013**, *11*, 728–736. [CrossRef]
16. Saltykova, A.; Mattheus, W.; Bertrand, S.; Roosens, N.H.C.; Marchal, K.; De Keersmaecker, S.C.J. Detailed Evaluation of Data Analysis Tools for Subtyping of Bacterial Isolates Based on Whole Genome Sequencing: Neisseria meningitidis as a Proof of Concept. *Front. Microbiol.* **2019**, *10*, 1–3. [CrossRef] [PubMed]
17. Treangen, T.J.; Ondov, B.D.; Koren, S.; Phillippy, A.M. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* **2014**, *15*, 524. [CrossRef]
18. Zhang, S.; Li, S.; Gu, W.D.; Bakker, H.; Boxrud, D.; Taylor, A.; Roe, C.; Driebe, E. Engelthaler, D.M.; Allard, M.; et al. Zoonotic Source Attribution of Salmonella enterica Serotype Typhimurium Using Genomic Surveillance Data, United States. *Emerg. Infect. Dis.* **2019**, *25*, 82–91. [CrossRef]
19. Lupolova, N.; Dallman, T.J.; Holden, N.J.; Gally, D.L. Patchy promiscuity: machine learning applied to predict the host specificity of *Salmonella enterica* and *Escherichia coli*. *Microb. Genom.* **2017**, *3*, e000135. [CrossRef]
20. Munck, N.; Njage, P.M.K.; Leekitcharoenphon, P.; Litrup, E.; Hald, T. Application of Whole-Genome Sequences and Machine Learning in Source Attribution of Salmonella Typhimurium. *Risk Anal.* **2020**, *40*, 1700–1703. [CrossRef]
21. Njage, P.M.K.; Leekitcharoenphon, P.; Hansen, L.T.; Hendriksen, R.S.; Faes, C.; Aerts, M.; Hald, T. Quantitative Microbial Risk Assessment Based on Whole Genome Sequencing Data: Case of Listeria monocytogenes. *Microorganisms* **2020**, *8*, 1772. [CrossRef]

22. Njage, P.M.K.; Henry, C.; Leekitcharoenphon, P.; Roussel, S.; Hendriksen, RS.; Hald, T. Potential of machine learning methods as a tool for predicting risk of illness applying next generation sequencing data: Case of Listeria monocytogenes. *Risk Anal.* **2019**, *39*, 1397–1410.
23. Njage, P.M.K.; Leekitcharoenphon, S.; Hald, T. Machine learning as a tool for microbial risk assessment using next generation sequencing data: predicting clinical outcomes in shigatoxigenic *Escherichia coli*. *Int. J. Food Microbiol.* **2019**, *292*, 72–82. [CrossRef] [PubMed]
24. Tanui C.K.; Karanth S.; Njage P.M.K; Meng J.; Pradhan A.K. Machine learning-based predictive modeling to identify genotypic traits associated with Salmonella enterica disease endpoints in isolates from ground chicken. *LWT* **2022**, *154*, 112701. [CrossRef]
25. Bandoy, D.; Weimer, B.C. Biological Machine Learning Combined with Campylobacter Population Genomics Reveals Virulence Gene Allelic Variants Cause Disease. *Microorganisms* **2020**, *8*, 549. [CrossRef] [PubMed]
26. Santo F.; Darko, H. Community detection in networks: A user guide. *Phys. Rep.* **2016**, *659*, 1–44.
27. Joensen, K.G.; Kiil, K.; Gantzhorn, M.R.; Nauerby, B.; Engberg, J.; Holt, H.M.; Nielsen, H.L.; Petersen, A.M.; Kuhn, K.G.; Sandø, G.; Ethelberg, S.; Nielsen, E.M. Whole−Genome Sequencing to Detect Numerous Campylobacter jejuni Outbreaks and Match Patient Isolates to Sources, Denmark, 2015–2017. *Emerg. Infect. Dis.* **2020**, *26*, 523–532. [CrossRef]
28. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A. A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Prjibelski, A.D.; et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **2012**, *19*, 455–77. [CrossRef]
29. Cody, A.J.; Bray, J.E.; Jolley, K.A.; McCarthy, N.D.; Maiden, M.C.J. Coregenome Multilocus Sequence Typing Scheme for Stable,Comparative Analyses of Campylobacter jejuni and *C. coli* Human Disease Isolates. *J. Clin. Microbiol.* **2017**, *55* 2086–2097. [CrossRef]
30. Center for Genomic Epidemiology. Available online: https://www.genomicepidemiology.org/ (accessed on 31 March 2022).
31. Cody, A.J.; McCarthy, N.D.; van Rensburg, M.J.; Isinkaye, T.; Bentley, S.D.; Parkhill, J.; Dingle, K.E.; Jolley, K.A.; Maiden, M.C.J. Real-time genomic epidemiological evaluation of human Campylobacter isolates by use of whole-genome multilocus sequence typing. *J. Clin. Microbiol.* **2013**, *51*, 2526–2534. [CrossRef]
32. Kaas, R.S.; Leekitcharoenphon, P.; Aarestrup, F.M.; Lund O. Solving the Problem of Comparing Whole Bacterial Genomes across Different Sequencing Platforms. *PLoS ONE* **2014**, *9*, 1–6.
33. Heng L. A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data. *Bioinformatics* **2011**, *27*, 2987–2993.
34. Heng L.; Durbin R. Fast and Accurate Long-Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* **2010**, *26*, 589–595.
35. Heng, L.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin R. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079.
36. Ivica, L.; Bork, P.; Interactive Tree Of Life (iTOL) v6: Recent updates and new developments. *Nucleic Acids Res.* **2019**, *47*, 256–59.
37. Computerome 2.0. Available online: https://www.computerome.dk (accessed on 30 March 2022).
38. MATLABR2021b: https://www.mathworks.com/products/get-matlab.html?s_tid=gn_getml (accessed on 30 March 2022).
39. Fruchterman, T.; Reingold, E. Graph drawing by force-directed placement. *Soft. Prac. Exp.* **1991**, *21 (11)* 1129–1164. [CrossRef]
40. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*, 1st ed.; Springer: New York, NY, USA, **2013**; pp. 415–419.
41. Woodcock, D.J.; Krusche, P.; Strachan, N.J.C.; Forbes, K.J.; Cohan, F.M.; Méric, G.; Sheppard, K.S. Genomic plasticity and rapid host switching can promote the evolution of generalism: a case study in the zoonotic pathogen Campylobacter. *Sci. Rep.* **2017**, *7*, 9650. [CrossRef] [PubMed]