Machine learning to predict the source of campylobacteriosis using whole genome data

Nicolas Arning*[1], Samuel K. Sheppard[2], David A. Clifton[3] and Daniel J. Wilson[1]

[1] Big Data institute, Nuffield Department of Population Health, University of Oxford, Li Ka Shing Centre for Health Information and Discovery, Old Road Campus, Oxford, OX3 7LF, UK

[2] The Milner Centre of Evolution, Department of Biology & Biochemistry, University of Bath, Claverton Down, Bath, BA2 7AZ, UK

[3] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, OX3 7DQ, UK

Abstract

Campylobacteriosis is among the world's most common foodborne illnesses, caused predominantly by the bacterium *Campylobacter jejuni*. Effective interventions require determination of the infection source which is challenging as transmission occurs via multiple sources such as contaminated meat, poultry, and drinking water. Strain variation has allowed source tracking based upon allelic variation in multi-locus sequence typing (MLST) genes allowing isolates from infected individuals to be attributed to specific animal or environmental reservoirs. However, the accuracy of probabilistic attribution models has been limited by the ability to differentiate isolates based upon just 7 MLST genes. Here, we broaden the input data spectrum to include core genome MLST (cgMLST) and whole genome sequences (WGS), and implement multiple machine learning algorithms, allowing more accurate source attribution. We increase attribution accuracy from 64% using the standard iSource population genetic approach to 71% for MLST, 85% for cgMLST and 78% for kmerized WGS data using machine learning. To gain insight beyond the source model prediction, we use Bayesian inference to analyse the relative affinity of *C. jejuni* strains to infect humans and identified potential differences, in source-human transmission ability among clonally related isolates in the most common disease causing lineage (ST-21 clonal complex). Providing generalizable computationally efficient methods, based upon machine learning and population genetics, we provide a scalable approach to global disease surveillance that can continuously incorporate novel samples for source attribution and identify fine-scale variation in transmission potential.

Author summary

C. jejuni are the most common cause of food-borne bacterial gastroenteritis but the relative contribution of different sources are incompletely understood. We traced the origin of human *C. jejuni* infections using machine learning algorithms that compare the DNA sequences of bacteria sampled from infected people, contaminated chickens, cattle, sheep, wild birds and the environment. This approach achieved improvement in accuracy of source attribution by 33% over existing methods that use only a subset of genes within the genome and provided evidence for the relative contribution of different infection sources. Sometimes even very similar bacteria showed differences, demonstrating the value of basing analyses on the entire genome when developing this algorithm that can be used for understanding the global epidemiology and other important bacterial infections.

Introduction

Campylobacter jejuni and *Campylobacter coli* are among the most common causes of gastroenteritis globally and are responsible for approximately nine million annual cases in the European Union (1,2). These zoonotic bacteria are a common commensal constituent of the gut microbiota of bird and animal species (3,4) but cause serious infections in humans. Symptoms include nausea, fever, abdominal pain, and severe diarrhoea, with potential for the development of debilitating, and sometimes fatal, sequelae (5,6). Various infection sources have been identified including animal faeces, contaminated drinking water and especially raw or under-cooked poultry and other meats (7). However, effectively combating disease requires a detailed understanding of the relative contribution of different sources to human infection.

As in many other bacterial species, *Campylobacter* populations represent diverse assemblages of strains (3,8–10). Within this structured population, some lineages are more commonly observed in particular host species (3,4,11). Because of this host association, DNA sequence comparisons of bacteria from human gastroenteritis and potential reservoir populations have potential to reveal the infection source. This has identified contaminated poultry as a major source of human infection (12,13). Based on the body of evidence including DNA sequence analysis (14), targeted interventions have been implemented, including improved biosecurity measures on poultry farms, which have halved recorded campylobacteriosis cases in New Zealand (15,16).

Extending the principal of linking source-sink populations using genotype data, methods have been developed to attribute *C. jejuni* to the likely source based on bacterial gene frequencies in potential reservoir populations (17,18). Among the most common genotyping approaches for *C. jejuni* has been multi-locus sequence typing (MLST) that catalogues DNA sequence variation across seven housekeeping genes that are common to all strains (19,20). Isolates with identical alleles at all loci are assigned to the same sequence type (ST) and those with identical sequences at most or all loci are grouped within the same clonal complex (CC). Using these data, and allele frequencies, it has been possible to probabilistically assign clinical isolates (STs and CCs) to host source using source attribution models such as the asymmetric island model implemented in *iSource* (17) and the Bayesian population assignment model STRUCTURE (18,21). Both methods have been instructive in estimating the relative contribution of a range of domestic and wild animal hosts to human infection, with poultry often identified as the principal source of human campylobacteriosis across different regions and countries (17,18,22–25).

There are two main limitations when using genotype data to for bacterial source attribution. The first is that the ability to attribute is only as good as the degree of genotype segregation. For example, in *C. jejuni* there are host restricted genotypes (3,26) that can be readily attributed to a given host source when observed in human infections, as well as ecological generalists (27,28) that have relatively recently transitioned between hosts and cannot therefore be attributed with confidence (29). While host switching potentially imposes a biological constraint on quantitative attribution models, the second limitation is far more tractable. Specifically, most current source attribution methods are subject to limitations imposed by the underlying data. Reflecting the technology of the time, MLST-based source attribution is based only on a small fraction of the genome (approximately 0.2% for *C. jejuni* (25)) and there is considerable potential for better strain differentiation using current techniques.

The increasing availability of large whole genome sequence (WGS) datasets has greatly enhanced analyses of bacterial population structure and diversity (30). However, exploiting the full information can be challenging due to variable gene content and the complexity of interpreting the short reads produced by next generation sequencing. Notwithstanding this, some studies have attempted to overcome the limited discriminatory power of MLST in attribution studies by screening WGS data to identify elements (SNPs and genes) that segregate by host (31–33). Using these host segregating markers as input data has improved the resolution of existing attribution models, including STRUCTURE, and provided information about potential infection reservoirs and the UK and France. However, using bespoke marker selection approaches with software designed for MLST data does not maximize the potential of WGS data for source attribution.

Here, we present a machine learning approach using WGS data to predict the source of human *C. jejuni* infection. This has two principal advantages over existing techniques. First, building on WGS-based machine learning source attribution approaches applied to *Salmonella enterica* and *Escherichia coli* (34,35), we take an agnostic approach to identify which machine learning tool performs best from a broad range of available algorithms. Second, we use a WGS input capture approach using data types deposited in public databases allowing the analysis of existing MLST, core-genome MLST and WGS datasets and the reuse of data for continuous updatable monitoring in a generalizable framework. Thus, we aimed to overcome limitations of the currently available methods and use the output to investigate the infective potential of *C. jejuni* strains.

Methods

Dataset acquisition

A total of 5,798 *C. jejuni* and *C. coli* genomes isolated from various sources and host species were available on the public database for molecular typing and microbial genome diversity: PubMLST (https://pubmlst.org/) (S1 Table). WGS data corresponded to MLST ST and CC designations as well as core genome (cg) MLST classes. The dataset was divided into training (75%) and testing (25%) sets using phylogeny-aware sorting, wherein all members of one ST were sorted entirely into either training or testing sets (S1 Table). The ST based sorting accounts for the phylogenetic non-independence of samples (36). To allow for sufficient sample sizes per reservoir population (hereafter "class"), only the five most prevalent classes for MLST and cgMLST were used (chicken, cattle, sheep, wild bird and environment). For farm animals the classes "chicken" and "chicken offal or meat" were combined to "chicken" (likewise for sheep and cattle), whilst "environment", "sand" and "river water" were combined into "environment", consistent with previous studies (18,37).

Feature engineering

The allelic profiles of MLST and cgMLST were used directly. To potentially exploit the gradient of separation encoded in the sequences underlying the MLST allelic profiles, we downloaded the underlying allele sequences and encoded the nucleotides as dummy variables and k-mers (k=21) using DSK (38). DSK was also used for encoding the WGS as k-mers. Using k=21 led to a prohibitively large input vector due to the number of unique k-mers found in all genomes (109,675,176). We reduced the number of k-mers by applying a variance threshold where k-mers which were present or absent in more than 99% of the samples were discarded, reducing the numbers of unique k-mers to 7,285,583. Furthermore, we performed feature selection by testing the dependence of the source labels on every individual k-mer using the Chi-Square statistic. To avoid data-leakage we only performed the feature selection using the training data and labels to select the 100,000 k-mers with the highest score.

Algorithm training

All machine learning and deep learning was performed in Python (for a list of all algorithms see Figure 1). The xgboost library (39) was used for the gradient boosting classifiers with all other machine learners implemented in scikit-learn (40). The hyper-parameters for each classifier were chosen using Cartesian grid search on five-fold cross-validation of the training set. The Keras library (https://github.com/keras-team/keras) was used to construct deep learning algorithms aimed at supplying a wide range of commonly used architectures. We found this to work best, empirically, given that there is no principled means of architecture selection for such models. Specifically: (i) A recurrent neural network consisting of a layer with 64 gated recurrent units, a 50% dropout layer and Rectified Linear Unit (ReLU) activation layer; (ii) A 1-dimensional convolutional network with two convolutional layers of kernel size 3 and 5 respectively and 30 filters, both followed by 50% dropout layers and a ReLU layer; (iii) A Long short-term memory network consisting of one LSTM layer with 64 units and a 50% dropout layer; (iv) A Shallow dense network with one dense layer with 64 units followed by a 50% dropout layer and a ReLU activation layer; (v) A Deep dense network with 6 dense layers starting with 128 units and halving units with each successive layer. All individual dense layers are followed by a 50% dropout layer and a ReLU layer.

To all deep learning architectures, we added an output layer comprising a dense layer with soft-max activation with one unit for every class. We encoded the labels as dummy variables and used categorical cross-entropy as a loss function together with the Adam optimiser (41). Cyclical learning rates were used with a maximum learning rate of 0.1 and a minimum learning rate of 0.0001 to overcome local minima. The accuracy on the test set was measured at every epoch and the overall best performing weights were stored as a checkpoint. The data was deployed in batches of 128 samples with every batch randomly undersampled so that each class was represented in equal proportions. The training was run for 500 generations with early stopping after 50 generations.

Algorithm testing

Both machine learning and deep learning were tested on the same 25% test set. The original data were skewed in source composition by ratios which did not necessarily reflect source origin of infection. We therefore used two methods to rebalance the classes in testing. The first test set featured an even distribution of classes, whereas the second undersampled the over-abundant chicken-origin genomes to emulate relative contribution to human disease. We used the ratios predicted by Wilson et al. (12), where *Campylobacter* genomes from chickens were 1.61 times more common than those from cattle. In both methods, rebalancing the classes was achieved by undersampling, which we repeated 200 times with replacement and averaged the accuracy over all iterations whilst also recording the variance. For performance metrics we registered accuracy, precision (positive predictive value), recall (sensitivity), F1, negative predictive value, specificity and speed. Speed was measured relative to other classifiers where a scale was defined with 0 being the slowest classifier and 1 being the quickest and all intermediate values being normalised within these confines. For comparison to previous methods, iSource was applied to the test dataset (17). Having established that XGBoost on cgMLST was the best performing source attribution method, we retrained the classifier with both training and testing data and applied it to all 15,988 human cgMLST samples available on the PubMLST database. The prediction took 892 milliseconds on a Dell OptiPlex 7060 desktop using ten threads on an Intel Core i7-8700 CPU and 16 GB RAM.

Phylogenetic analysis

We defined the generalist index as the number of sources the ST was found in across all isolates in the dataset, which included additional samples for which only MLST data was available (S1 Table). We built a phylogeny of CC21 genomes from both source-associated and human isolates using Neighbour Joining, based on pairwise hamming distances of k-mer presence/absence in the WGS dataset, as described by Hedge and Wilson (42). We used TreeBreaker to infer the evolution of phenotypes across the phylogenetic tree of ST-21 and the most closely related sequence types. The known labels of the source-associated samples were used as phenotypic information for input into TreeBreaker (43) together with the phylogeny of CC21. TreeBreaker was run for 5,500,000 iterations with 500,000 iterations as burn-in and 1000 iterations between sampling. The phylogenetic trees were visualised with Microreact (44) and arranged alongside the results of TreeBreaker in Inkscape.

Results and Discussion

Machine learning outperforms popular attribution models for MLST data

In order to anchor our source attribution performance to previous efforts, we compared results using the machine learning classifiers to source probabilities estimated using the asymmetric island model implemented in iSource, which is based on MLST and the most commonly used source attribution method to date (45). The best performing machine learner on the MLST allelic profile was a random forest (61.9%/68.5% balanced/unbalanced) which performed slightly better than iSource (61%/64%) (Figure 1). Since loci within allelic profiles are deemed either to match or not, and underlying nucleotides sequences are ignored, we investigated whether exploiting the gradient of nucleotide differentiation would lead to better attribution. We used dummy variables and generated k-mers from the sequences underlying the MLST allele labels. The additional feature encodings boosted the top achieving accuracies on MLST to 67.9%/70.7% from dummy variables and 63%/67.5% from k-mers, showing the value of the additional nucleotide-level information.

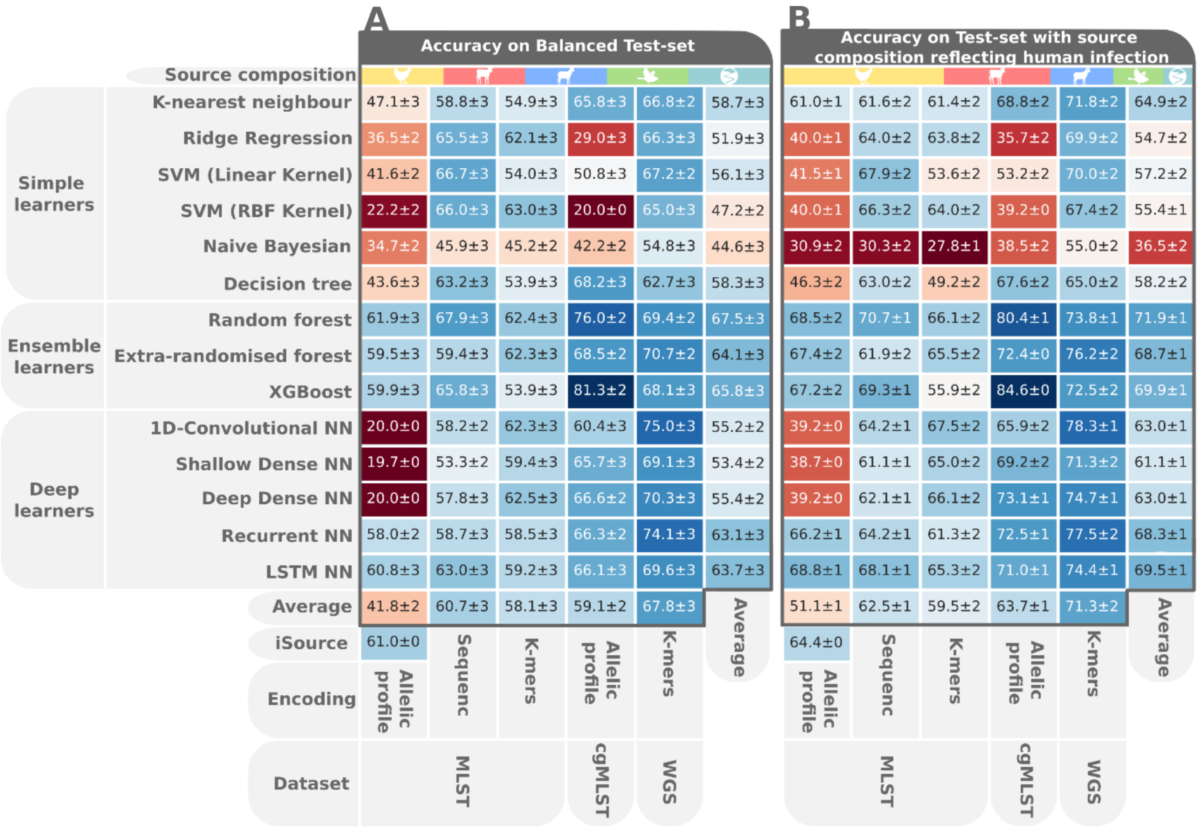


Figure 1: A heatmap showing classifier performance on the class balanced (A) and imbalanced (B) test set. The individual cells are coloured according to the average accuracy on 200 rounds of resampling with replacement with the variance noted next to the average accuracy. The averages of accuracy per classifiers are shown in the rightmost column, whereas the bottom column shows the averages per data type.

Core genome and WGS datasets increase the power of source attribution models

Having established the competitiveness of machine learning approaches for source attribution using MLST data, we turned our attention to whole genome datasets. Gene-by-gene approaches to cataloguing genomic variation in *Campylobacter* (46) and other species are a logical extension of seven-locus MLST in response to the increasing availability of large WGS datasets. Formalizing this approach to derive an approximation of the core genome for *C. jejuni* allowed the implementation of a cgMLST scheme containing 1,343 genes, that are present in the majority (>95%) of *C. jejuni* genomes (47). This has potential to increase the power of attribution models to discriminate the source of *Campylobacter* isolates based on host segregating genetic variation within the genome (37). The strong performance of tree-based ensemble classifiers continued when using cgMLST data where the XGBoost classifier achieved 81.3%/84.6% accuracy, the highest accuracy over all data types and classifiers.

q

Next, we assessed the relative performance of machine learners when applied to k-mers produced from WGS, where the average attribution performance was the highest among all datasets. The best-performing algorithm was a 1-D convolutional neural net (75.0/78.3%), performing better than the top-achieving classifier on MLST but worse than the best classifier on cgMLST despite WGS encoding more genomic information. This may be explained by the feature selection used to limit the input vector to 100,000 k-mers. Beyond comparing classifier performance on different data types, we also wanted to investigate what led to the difference in performance.

The comparison of average accuracy across all data types reveals that with an increase in encoded variation the average performance across all algorithms improves. This is especially apparent in MLST where, although capturing the same 0.3% of the genome in all isolates, the additional variation in the underlying sequences can be leveraged for better performance. When comparing the average accuracy between classifiers we observed that decision-tree based ensemble learners performed well across all datasets, with random forests performing best on average. The excellent performance of ensemble tree learners on genomic data has been reported on genomic data (48–50) and is linked to their ability to handle correlation as well as interaction of features which is an inherent feature of genomic data (50).

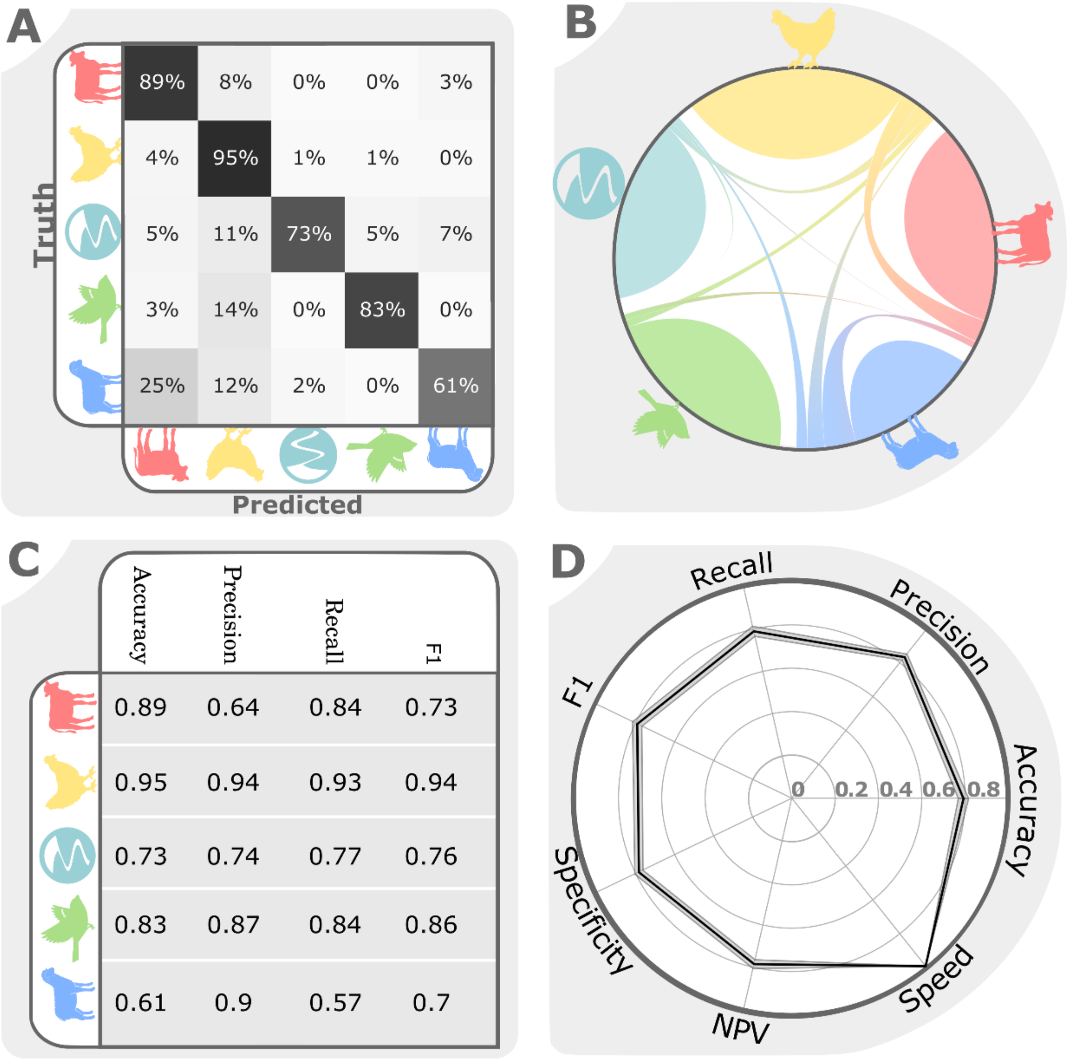
Amongst simple learners the K-nearest neighbour algorithm (KNN) performed best, probably owing to the hereditary nature of the phenotypic trait used as classes here. Host association is inherited both genetically, in the ability to colonise different hosts, and environmentally, in the colocation of parent and offspring cells. These patterns of inheritance result in more closely related sequences being more likely to be associated with the same phenotype. Heritability could explain the success of the KNN algorithm which is based on proximity in hyperdimensional feature space (51), which in our case is genetic similarity which is a proxy for relatedness.

The deep learners generally improved in performance with higher dimensionality of the input data - from MLST to WGS data. Among all deep learning architectures, the RNN and LSTM performed best, which was to be expected as DNA is transcribed, and mRNA translated, sequentially 5’ to 3’. Both RNNs and LSTMs process input data sequentially and input weights are also adjusted sequentially in back-propagation as opposed to the dense or convolutional architectures where input weights are tweaked concurrently. Having investigated trends across all datasets and algorithms we focused on the best-achieving classifier for a more thorough analysis of how classification performance was driven by different factors within the underlying data.

Host transition imposes a biological limit on source attribution models

To better understand the limitations of attribution algorithms we investigated the factors driving misclassification in the different models with different datasets. The XGBoost implementation of gradient boosted decision trees, using the cgMLST dataset, was the overall best-performing classifier in our analyses. Consequently, this was used to investigate attribution performance further (Figure 2).

Among all source populations the most frequent misclassification was found between sheep and cattle, which is a common source of errors in source attribution (17) owing to strongly overlapping gene pools stemming from frequent cross-species transmission that may reflect commonalities in physiological features of the ruminant gastrointestinal tracts (52). We also looked at factors besides source reservoir of the sample, as circumstances like geographical origin of the isolate (56) and the season in which they were sampled (57) have been shown to influence source attribution. We therefore stratified classification accuracy by continent, year, generalist index and *Campylobacter* species using the full non-undersampled Test dataset (Figure 3, S1 Table).



	89%	8%	0%	0%	3%
	4%	95%	1%	1%	0%
	5%	11%	73%	5%	7%
	3%	14%	0%	83%	0%
	25%	12%	2%	0%	61%

	Accuracy	Precision	Recall	F1
	0.89	0.64	0.84	0.73
	0.95	0.94	0.93	0.94
	0.73	0.74	0.77	0.76
	0.83	0.87	0.84	0.86
	0.61	0.9	0.57	0.7

Figure 2: XGBoost Classifier performance on cgMLST: A) Misclassification matrix per source. The

diagonal represents correct classification and off-diagonal fields are misclassifications. The percentages are calculated per row. B) Misclassification matrix as depicted in a flow diagram. C) Classifier performance on the unbalanced test set according to four different metrics per source population. D) Radar plot showing the classifier performance on the unbalanced test by seven metrics averaged over 200 rounds of resampling with replacement. The variation is depicted as a shaded surface underneath the black line representing the average.

Investigating the accuracy of the XGBoost classifier per sample size revealed that the low number of wild bird samples (212 samples; 84% accuracy) did not impede classification performance when compared to more abundant source samples like cattle (716 samples; 84% accuracy) and sheep (584 samples; 57% accuracy), presumably because wild bird STs tend to be atypical compared to the other reservoirs (46). To investigate how the ability to colonise multiple hosts affected performance, we defined a ‘generalist index’ as the number of hosts in which an ST was found across all PubMLST samples (S1 Table). The performance across generalist indices showed that strains restricted to fewer hosts were predicted with higher accuracy. This is likely due to host switching blurring the source-specific genetic signal, as previously reported (29). Consistent with this, 58% of all wild bird samples belonged to STs only found in this niche, compared to 41% in environment, 9% in cattle, 3% in sheep and 32% in chicken. Besides *C. jejuni*, an estimated 10% of campylobacteriosis cases are caused by *Campylobacter coli* (53). Consistent with previous studies, we found improved accuracy over attribution of *C. jejuni*, potentially reflecting more pronounced strain segregation by host (29), as well as a higher proportion of environmental and sheep associated strains in human infection (11,54,55) (Figure 3).

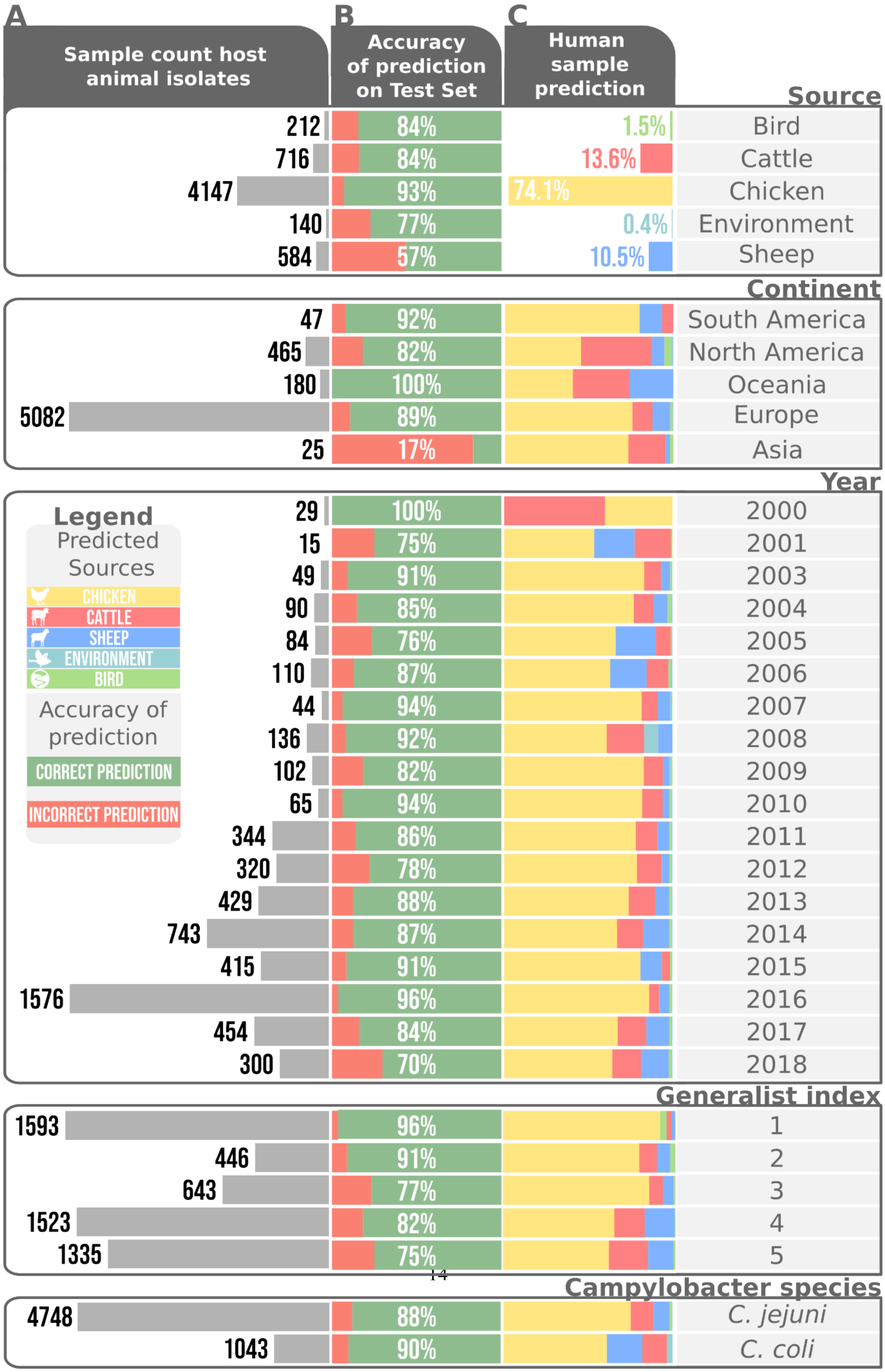
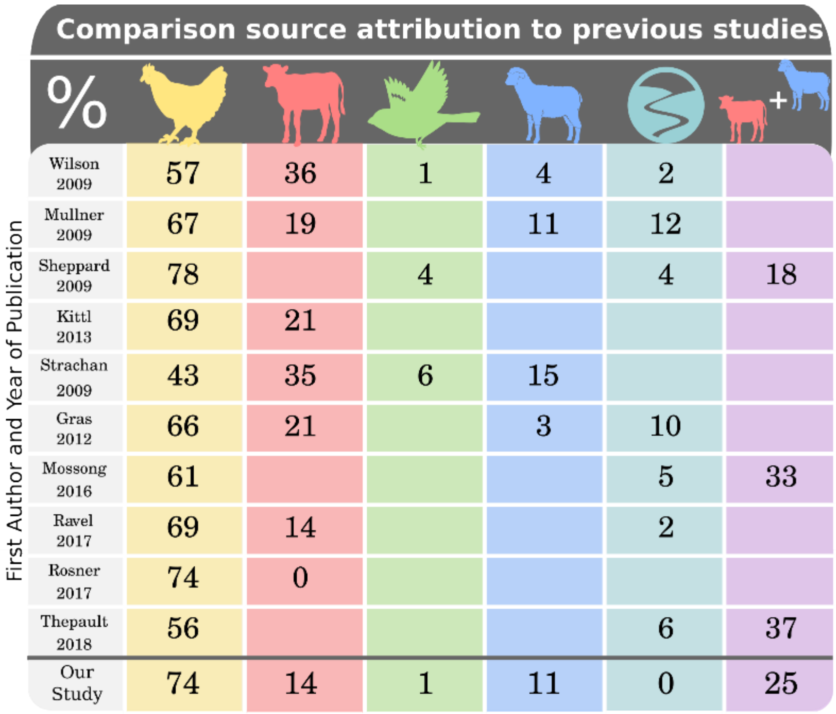

A
Sample count host animal isolates
B
Accuracy of prediction on Test Set
C
Human sample prediction
Source
212 84% 1.5% Bird
716 84% 13.6% Cattle
4147 93% 74.1% Chicken
140 77% 0.4% Environment
584 57% 10.5% Sheep
Continent
47 92% South America
465 82% North America
180 100% Oceania
5082 89% Europe
25 17% Asia
Year
29 100% 2000
15 75% 2001
49 91% 2003
90 85% 2004
84 76% 2005
110 87% 2006
44 94% 2007
136 92% 2008
102 82% 2009
65 94% 2010
344 86% 2011
320 78% 2012
429 88% 2013
743 87% 2014
415 91% 2015
1576 96% 2016
454 84% 2017
300 70% 2018
Legend
Predicted Sources
CHICKEN
CATTLE
SHEEP
ENVIRONMENT
BIRD
Accuracy of prediction
CORRECT PREDICTION
INCORRECT PREDICTION
Generalist index
1593 96% 1
446 91% 2
643 77% 3
1523 82% 4
1335 75% 5
Campylobacter species
4748 88% C. jejuni
1043 90% C. coli

Figure 3: Source attribution per source, continent, year generalist index and *Campylobacter* species. A) Sample sizes across different factors in the imbalanced training set. B) Prediction accuracy on the full test dataset divided by different factors. C) Source attribution stratified into varying factors

Having analysed the classification accuracy within the dataset, the machine learning method was compared to previous source attribution studies (Figure 4). Attribution of cases to chicken was consistent with higher estimates from previous studies, resulting in less attribution to all other sources, with environment identified as the source of just 0.4% of human infections. This differences in our prediction to previous studies could reflect the greater discriminatory power of cgMLST data over MLST.



Comparison source attribution to previous studies

%	🐔	🐄	🐦	🐑	🌊	🐄+🐑
Wilson 2009	57	36	1	4	2	
Mullner 2009	67	19		11	12	
Sheppard 2009	78		4		4	18
Kittl 2013	69	21				
Strachan 2009	43	35	6	15		
Gras 2012	66	21		3	10	
Mossong 2016	61				5	33
Ravel 2017	69	14			2	
Rosner 2017	74	0				
Thepault 2018	56				6	37
Our Study	74	14	1	11	0	25

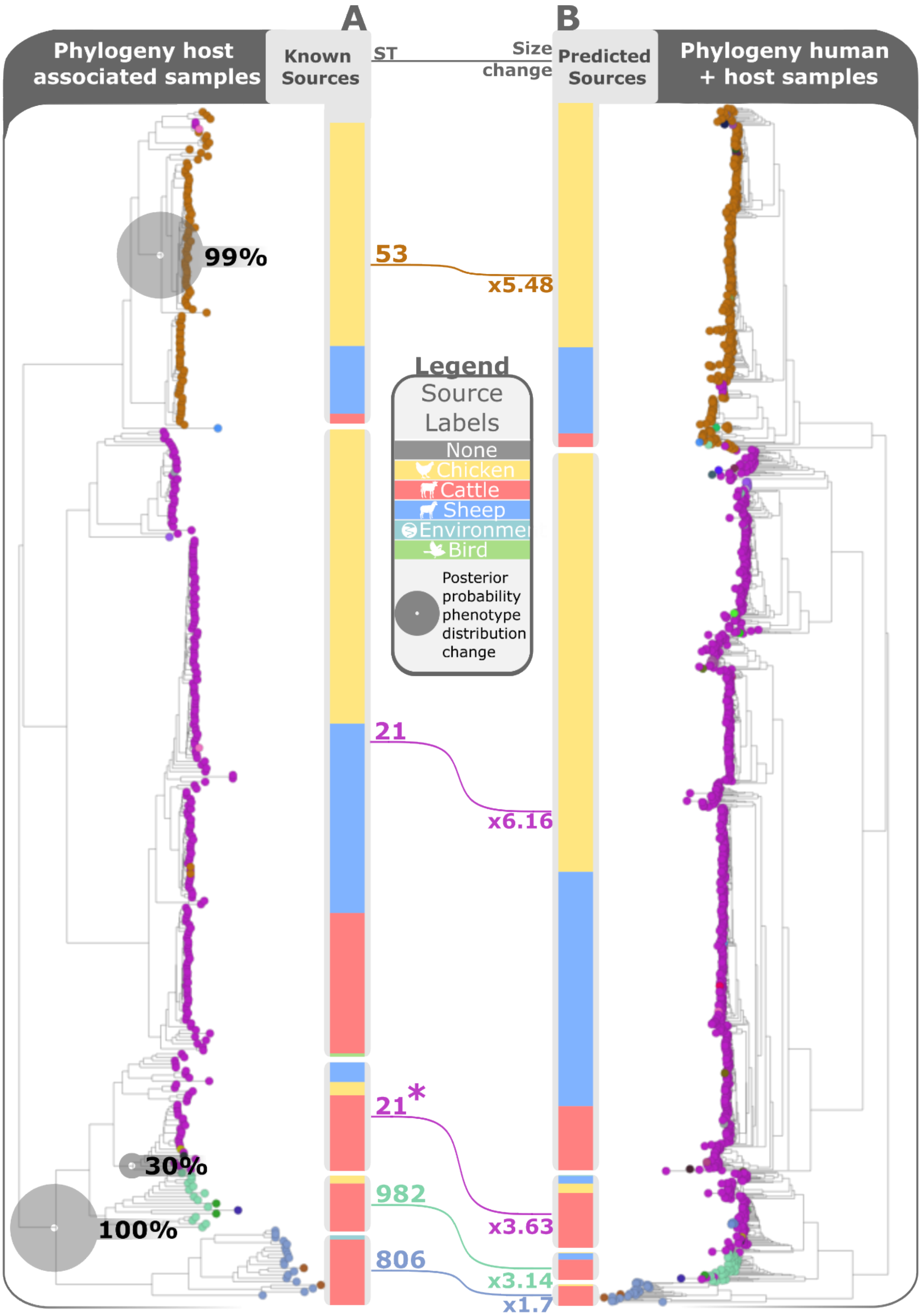
First Author and Year of Publication

Figure 4: Comparison of our source attribution to previously published studies

The fine-grained structure of source attribution can be identified with machine learning

Attribution predictions are inferred from the observed frequencies of genotypes in host reservoirs assessed through sampling. However, the relative source composition observed in sampling does not necessarily correspond to host contributions to human infection as some strains that are found at low frequency in the host could be more infectious to humans. For example, some *C. jejuni* strains increase in relative frequency through different stages of the poultry slaughter and production chain because they have genes that promote survival outside of the host (58). There is also evidence that there is a genetic bottleneck at the point of human infection that promotes colonization by strains that have specific genes conferring human niche tropism (59). Analysis of WGS or cgMLST data can potentially allow for changes in relative frequency and provide finer-grained source attribution, potentially at the level of the individual genome.

To identify evidence of differential host affinities, we applied treeBreaker (43) to trace the evolution of a host association along the phylogeny of CC-21, the most commonly found clonal complex in human infection (27). CC-21 frequently colonizes all host sources analysed in this study and is therefore considered a generalist strain, potentially complicating accurate attribution. TreeBreaker detected a change in host association on a branch that groups together a cattle-associated ST-21 subgroup with the cattle-associated lineages ST-982 and ST-806 (Figure 5A). The source composition in this clade (asterisked in Figure 5A) differed from the rest of CC-21, which were predominantly composed of chicken and sheep isolates. Moreover, the asterisked clade differed in its propensity for transmission to humans. Overall, CC-21 was over-represented among human infections, perhaps reflecting its generalist affinities. Yet the asterisked clade was over-represented only 1.7 to 3.6-fold, compared to 5.5 to 6.2-fold for the rest of CC-21 (Figure 5B).



300

301 **Figure 5:** Phylogeny of clonal complex 21 of host animal associated samples (A) and bar charts showing

302 the known source distribution and human samples (B) alongside the predicted source distribution. The

phylogeny is based on Neighbour joining using hamming distance of the k-mers drawn from WGS. The connecting lines show the increase in frequency of the clades in human samples and the size of the grey circles show the posterior probability of a change in phenotypic distribution along the branches of the tree.

As the host association changed within CC-21, the ability to transmit to humans appears to have changed as well. This in turn induced a change in the source composition of CC-21 sampled from human infections compared to CC-21 sampled from animals. Previous studies analysing source attribution based on MLST would have overlooked these shifts.

Outlook and conclusions

The increasing availability of large pathogen genome datasets, algorithms and resources for analysing them, has created possibilities for investigating the transmission of zoonotic diseases that are incompletely understood. It is clear from the data presented here that tree-based ensemble methods for machine learning classification using bacterial genomic data provide considerable utility for improving the accuracy host source attribution for human campylobacteriosis. Key to the effectiveness of this approach is leveraging the full gradient of genomic differentiation afforded by WGS or cgMLST analysis. Host associated genetic variation can be observed in both core and accessory genes (60) but using these data presents practical considerations. With more computational resources available, it may be possible to analyse all k-mers present in the WGS samples (here 109,675,176 unique kmers) with multiple algorithms accompanied by cross-validation and bootstrap replication.

Beyond simple attribution to host source, resolving the fine-grained structure of genomic signatures of association has considerable potential to account for differences in the relative frequency of sub-lineages in samples taken from reservoir hosts and human disease. This can provide important clues about the propensity of strains to survive outside of the host for long enough to transmit to humans as well as the capacity to colonize the human gut given the opportunity (58,59). This of course leads to questions about the genomic basis of bacterial adaptation, specifically the extent to which 'associated' genetic elements represent adaptations and whether the same genes and alleles enable colonisation of different host animals.

Improving on the approaches described here, better sampling and incremental training of the XGBoost classifier has considerable potential. The classifier's low computational requirements and high prediction speed make it an excellent tool for analysing large genome datasets. Furthermore, by using phylogeny-aware train/test splitting for measuring performance, prediction remains accurate when new genetic variants are introduced because the algorithm can be incrementally trained with new data. This has considerable potential for developing automated and continuous disease surveillance systems to reduce campylobacteriosis that remains one of the most common food-borne illness in the world.

Acknowledgments

N. A. is a recipient of a BBSRC scholarship and thus supported by funding from the Biotechnology and Biological Sciences Research Council (BBSRC) (grant number BB/M011224/1). SKS was supported by Wellcome Trust (088786/C/09/Z) and Medical Research Council (MR/M501608/1 and MR/L015080/1) grants. D. J. W. is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (grant number: 101237/Z/13/B) and by the Robertson Foundation. The research was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

N.A. would like to thank David Eyre, Christophe Fraser and Alexandra Casey for insightful comments.

Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Conflicts of interest

DAC declares grants from GlaxoSmithKline and personal fees from Oxford University Innovation, BioBeats, and Sensyne Health, in areas unrelated to this work

Supporting information

S1 Table. Metadata of all *Campylobacter* isolates used in this study. Contains the accession numbers, year and country of isolation, source label, generalist index, ST, CC ,prediction by our classifier, *Campylobacter species* and whether the samples were used in training or testing.

References

1. The European Union One Health 2018 Zoonoses Report. EFSA J. 2019;17(12):e05926.

2. Kaakoush NO, Castaño-Rodríguez N, Mitchell HM, Man SM. Global Epidemiology of Campylobacter Infection. Clin Microbiol Rev. 2015 Jul;28(3):687–720.

3. Sheppard SK, Colles FM, McCARTHY ND, Strachan NJC, Ogden ID, Forbes KJ, et al. Niche segregation and genetic structure of Campylobacter jejuni populations from wild and agricultural host species. Mol Ecol. 2011;20(16):3484–90.

4. Sheppard SK, Colles F, Richardson J, Cody AJ, Elson R, Lawson A, et al. Host Association of Campylobacter Genotypes Transcends Geographic Variation. Appl Environ Microbiol. 2010 Aug;76(15):5269–77.

5. Nachamkin I, Allos BM, Ho T. Campylobacter Species and Guillain-Barré Syndrome. Clin Microbiol Rev. 1998 Jul;11(3):555–67.

6. Nielsen LN, Sheppard SK, McCarthy ND, Maiden MCJ, Ingmer H, Krogfelt KA. MLST clustering of Campylobacter jejuni isolates from patients with gastroenteritis, reactive arthritis and Guillain–Barré syndrome. J Appl Microbiol. 2010 Feb;108(2):591–9.

7. Altekruse SF, Stern NJ, Fields PI, Swerdlow DL. Campylobacter jejuni—An Emerging Foodborne Pathogen. Emerg Infect Dis. 1999;5(1):28–35.

8. Gilbert MJ, Miller WG, Yee E, Zomer AL, van der Graaf-van Bloois L, Fitzgerald C, et al. Comparative Genomics of Campylobacter fetus from Reptiles and Mammals Reveals Divergent Evolution in Host-Associated Lineages. Genome Biol Evol. 2016 Jul 2;8(6):2006–19.

9. Kirk KF, Méric G, Nielsen HL, Pascoe B, Sheppard SK, Thorlacius-Ussing O, et al. Molecular epidemiology and comparative genomics of Campylobacter concisus strains from saliva, faeces and gut mucosal biopsies in inflammatory bowel disease. Sci Rep. 2018 Jan 30;8(1):1902.

10. Sheppard SK, Dallas JF, Wilson DJ, Strachan NJC, McCarthy ND, Jolley KA, et al. Evolution of an Agriculture-Associated Disease Causing Campylobacter coli Clade: Evidence from National Surveillance Data in Scotland. PLOS ONE. 2010 Dec 15;5(12):e15708.

11. Ogden ID, Dallas JF, MacRae M, Rotariu O, Reay KW, Leitch M, et al. Campylobacter excreted into the environment by animal sources: prevalence, concentration shed, and host association. Foodborne Pathog Dis. 2009 Dec;6(10):1161–70.

12. Institute of Environmental Science and Research Ltd. Notifiable and other diseases in New Zealand: Annual Report 2006. Porirua NZ Inst. 2007;

13. Sheppard SK, Dallas JF, MacRae M, McCarthy ND, Sproston EL, Gormley FJ, et al.

Campylobacter genotypes from food animals, environmental sources and clinical disease in Scotland 2005/6. Int J Food Microbiol. 2009 Aug 31;134(1–2):96–103.

14. Nichols GL, Richardson JF, Sheppard SK, Lane C, Sarran C. Campylobacter epidemiology: a descriptive study reviewing 1 million cases in England and Wales between 1989 and 2011. BMJ Open. 2012 Jan 1;2(4):e001179.

15. Sears A, Baker MG, Wilson N, Marshall J, Muellner P, Campbell DM, et al. Marked Campylobacteriosis Decline after Interventions Aimed at Poultry, New Zealand. Emerg Infect Dis. 2011 Jun;17(6):1007–15.

16. Nohra A, Grinberg A, Marshall JC, Midwinter AC, Collins-Emerson JM, French NP. Shifts in the Molecular Epidemiology of Campylobacter jejuni Infections in a Sentinel Region of New Zealand following Implementation of Food Safety Interventions by the Poultry Industry. Appl Environ Microbiol [Internet]. 2020 Feb 18 [cited 2021 Jan 6];86(5). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7028974/

17. Wilson DJ, Gabriel E, Leatherbarrow AJH, Cheesbrough J, Gee S, Bolton E, et al. Tracing the Source of Campylobacteriosis. PLOS Genet. 2008 Sep;4(9):e1000203.

18. Sheppard SK, Dallas JF, Strachan NJC, MacRae M, McCarthy ND, Wilson DJ, et al. Campylobacter Genotyping to Determine the Source of Human Infection. Clin Infect Dis. 2009 Apr;48(8):1072–8.

19. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, et al. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A. 1998 Mar;95(6):3140–5.

20. Dingle KE, Colles FM, Wareing DR, Ure R, Fox AJ, Bolton FE, et al. Multilocus sequence typing system for Campylobacter jejuni. J Clin Microbiol. 2001 Jan;39(1):14–23.

21. Pritchard JK, Stephens M, Donnelly P. Inference of Population Structure Using Multilocus Genotype Data. Genetics. 2000 Jun;155(2):945–59.

22. Mullner P, Spencer SEF, Wilson DJ, Jones G, Noble AD, Midwinter AC, et al. Assigning the source of human campylobacteriosis in New Zealand: A comparative genetic and epidemiological approach. Infect Genet Evol. 2009 Dec;9(6):1311–9.

23. Boysen L, Rosenquist H, Larsson JT, Nielsen EM, Sørensen G, Nordentoft S, et al. Source attribution of human campylobacteriosis in Denmark. Epidemiol Infect. 2014 Aug;142(8):1599–608.

24. Di Giannatale E, Garofolo G, Alessiani A, Di Donato G, Candeloro L, Vencia W, et al. Tracing Back Clinical Campylobacter jejuni in the Northwest of Italy and Assessing Their Potential Source. Front Microbiol [Internet]. 2016 Jun 13 [cited 2021 Feb 3];7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4904018/

25. Kittl S, Heckel G, Korczak BM, Kuhnert P. Source Attribution of Human Campylobacter Isolates by MLST and Fla-Typing and Association of Genotypes with Quinolone Resistance. PLOS ONE.

2013 Nov;8(11):e81796.

26. Mourkas E, Taylor AJ, Méric G, Bayliss SC, Pascoe B, Mageiros L, et al. Agricultural intensification and the evolution of host specialism in the enteric pathogen Campylobacter jejuni. Proc Natl Acad Sci. 2020 May 19;117(20):11018–28.

27. Sheppard SK, Cheng L, Méric G, Haan CPA de, Llarena A-K, Marttinen P, et al. Cryptic ecology among host generalist Campylobacter jejuni in domestic animals. Mol Ecol. 2014;23(10):2442–51.

28. Woodcock DJ, Krusche P, Strachan NJC, Forbes KJ, Cohan FM, Méric G, et al. Genomic plasticity and rapid host switching can promote the evolution of generalism: a case study in the zoonotic pathogen Campylobacter. Sci Rep. 2017 Aug;7(1):1–13.

29. Dearlove BL, Cody AJ, Pascoe B, Méric G, Wilson DJ, Sheppard SK. Rapid host switching in generalist Campylobacter strains erodes the signal for tracing human infections. ISME J. 2016 Mar;10(3):721–9.

30. Sheppard SK, Guttman DS, Fitzgerald JR. Population genomics of bacterial host adaptation. Nat Rev Genet. 2018 Sep;19(9):549–65.

31. Thépault A, Rose V, Quesne S, Poezevara T, Béven V, Hirchaud E, et al. Ruminant and chicken: important sources of campylobacteriosis in France despite a variation of source attribution in 2009 and 2015. Sci Rep. 2018 Jun;8(1):9305.

32. Jehanne Q, Pascoe B, Bénéjat L, Ducournau A, Buissonnière A, Mourkas E, et al. Genome-Wide Identification of Host-Segregating Single-Nucleotide Polymorphisms for Source Attribution of Clinical Campylobacter coli Isolates. Appl Environ Microbiol [Internet]. 2020 Nov 24 [cited 2021 Feb 3];86(24). Available from: https://aem.asm.org/content/86/24/e01787-20

33. Berthenet E, Thépault A, Chemaly M, Rivoal K, Ducournau A, Buissonnière A, et al. Source attribution of Campylobacter jejuni shows variable importance of chicken and ruminants reservoirs in non-invasive and invasive French clinical isolates. Sci Rep. 2019 May 30;9(1):8098.

34. Zhang S, Li S, Gu W, den Bakker H, Boxrud D, Taylor A, et al. Zoonotic Source Attribution of Salmonella enterica Serotype Typhimurium Using Genomic Surveillance Data, United States. Emerg Infect Dis. 2019;25(1):82–91.

35. Lupolova N, Dallman TJ, Holden NJ, Gally DL. Patchy promiscuity: machine learning applied to predict the host specificity of Salmonella enterica and Escherichia coli. Microb Genomics [Internet]. 2017 Oct [cited 2019 Sep 16];3(10). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5695212/

36. Lees JA, Mai TT, Galardini M, Wheeler NE, Horsfield ST, Parkhill J, et al. Improved Prediction of Bacterial Genotype-Phenotype Associations Using Interpretable Pangenome-Spanning Regressions. mBio [Internet]. 2020 Aug 25 [cited 2021 Feb 3];11(4). Available from: https://mbio.asm.org/content/11/4/e01344-20

37. Thépault A, Méric G, Rivoal K, Pascoe B, Mageiros L, Touzain F, et al. Genome-Wide

Identification of Host-Segregating Epidemiological Markers for Source Attribution in Campylobacter jejuni. Appl Environ Microbiol. 2017 Apr 1;83(7).

38. Rizk G, Lavenier D, Chikhi R. DSK: k-mer counting with very low memory usage. Bioinformatics. 2013 Mar;29(5):652–3.

39. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. New York, NY, USA: ACM; 2016 [cited 2019 Sep 17]. p. 785–94. (KDD '16). Available from: http://doi.acm.org/10.1145/2939672.2939785

40. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.

41. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. ArXiv14126980 Cs [Internet]. 2014 Dec [cited 2019 Sep 17]; Available from: http://arxiv.org/abs/1412.6980

42. Hedge J, Wilson DJ. Bacterial Phylogenetic Reconstruction from Whole Genomes Is Robust to Recombination but Demographic Inference Is Not. mBio [Internet]. 2014 Dec 31 [cited 2020 Nov 18];5(6). Available from: https://mbio.asm.org/content/5/6/e02158-14

43. Ansari MA, Didelot X. Bayesian Inference of the Evolution of a Phenotype Distribution on a Phylogenetic Tree. Genetics. 2016 Sep 1;204(1):89–98.

44. Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microb Genomics. 2016;2(11):e000093.

45. Cody AJ, Maiden MC, Strachan NJ, McCarthy ND. A systematic review of source attribution of human campylobacteriosis using multilocus sequence typing. Eurosurveillance [Internet]. 2019 Oct [cited 2020 Jan 27];24(43). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6820127/

46. Sheppard SK, Jolley KA, Maiden MCJ. A Gene-By-Gene Approach to Bacterial Population Genomics: Whole Genome MLST of Campylobacter. Genes. 2012 Apr 12;3(2):261–77.

47. Cody AJ, Bray JE, Jolley KA, McCarthy ND, Maiden MCJ. Core Genome Multilocus Sequence Typing Scheme for Stable, Comparative Analyses of Campylobacter jejuni and C. coli Human Disease Isolates. J Clin Microbiol. 2017 Jul;55(7):2086–97.

48. Austerlitz F, David O, Schaeffer B, Bleakley K, Olteanu M, Leblois R, et al. DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. BMC Bioinformatics. 2009 Nov;10(14):S10.

49. Deneke C, Rentzsch R, Renard BY. PaPrBaG: A machine learning approach for the detection of novel pathogens from NGS data. Sci Rep. 2017 Jan;7:39194.

50. Chen X, Ishwaran H. Random Forests for Genomic Data Analysis. Genomics. 2012 Jun;99(6):323–9.

51. Kotsiantis SB, Zaharakis ID, Pintelas PE. Machine learning: a review of classification and combining techniques. Artif Intell Rev. 2006 Nov;26(3):159–90.

52. Kwan PSL, Birtles A, Bolton FJ, French NP, Robinson SE, Newbold LS, et al. Longitudinal Study of the Molecular Epidemiology of Campylobacter jejuni in Cattle on Dairy Farms. Appl Environ Microbiol. 2008 Jun;74(12):3626–33.

53. Sheppard SK, Maiden MCJ. The Evolution of Campylobacter jejuni and Campylobacter coli. Cold Spring Harb Perspect Biol [Internet]. 2015 Aug [cited 2019 Sep 3];7(8). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4526750/

54. Roux F, Sproston E, Rotariu O, MacRae M, Sheppard SK, Bessell P, et al. Elucidating the Aetiology of Human Campylobacter coli Infections. PLoS ONE [Internet]. 2013 May [cited 2020 Feb 14];8(5). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3667194/

55. Strachan NJC, Gormley FJ, Rotariu O, Ogden ID, Miller G, Dunn GM, et al. Attribution of Campylobacter Infections in Northeast Scotland to Specific Sources by Use of Multilocus Sequence Typing. J Infect Dis. 2009 Apr;199(8):1205–8.

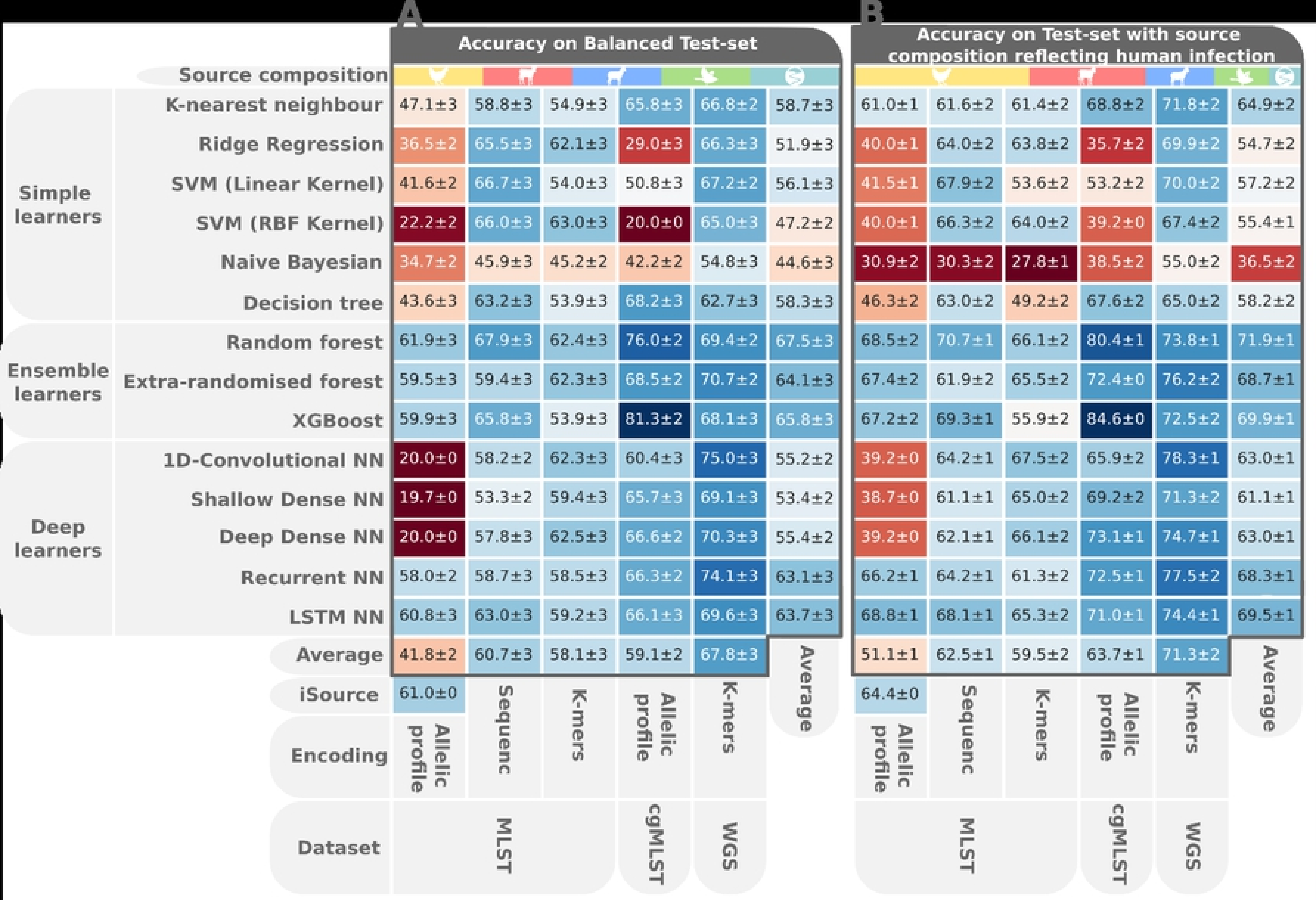
56. Pérez-Reche FJ, Rotariu O, Lopes BS, Forbes KJ, Strachan NJC. Mining whole genome sequence data to efficiently attribute individuals to source populations. Sci Rep. 2020 Jul 22;10(1):12124.

57. STRACHAN NJC, ROTARIU O, SMITH-PALMER A, COWDEN J, SHEPPARD SK, O'BRIEN SJ, et al. Identifying the seasonal origins of human campylobacteriosis. Epidemiol Infect. 2013 Jun;141(6):1267–75.

58. Yahara K, Méric G, Taylor AJ, Vries SPW de, Murray S, Pascoe B, et al. Genome-wide association of functional traits linked with Campylobacter jejuni survival from farm to fork. Environ Microbiol. 2017;19(1):361–80.

59. Méric G, McNally A, Pessia A, Mourkas E, Pascoe B, Mageiros L, et al. Convergent Amino Acid Signatures in Polyphyletic Campylobacter jejuni Subpopulations Suggest Human Niche Tropism. Genome Biol Evol. 2018 Mar 1;10(3):763–74.

60. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, et al. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. Proc Natl Acad Sci. 2013 Jul;110(29):11923–7.



A

Accuracy on Balanced Test-set

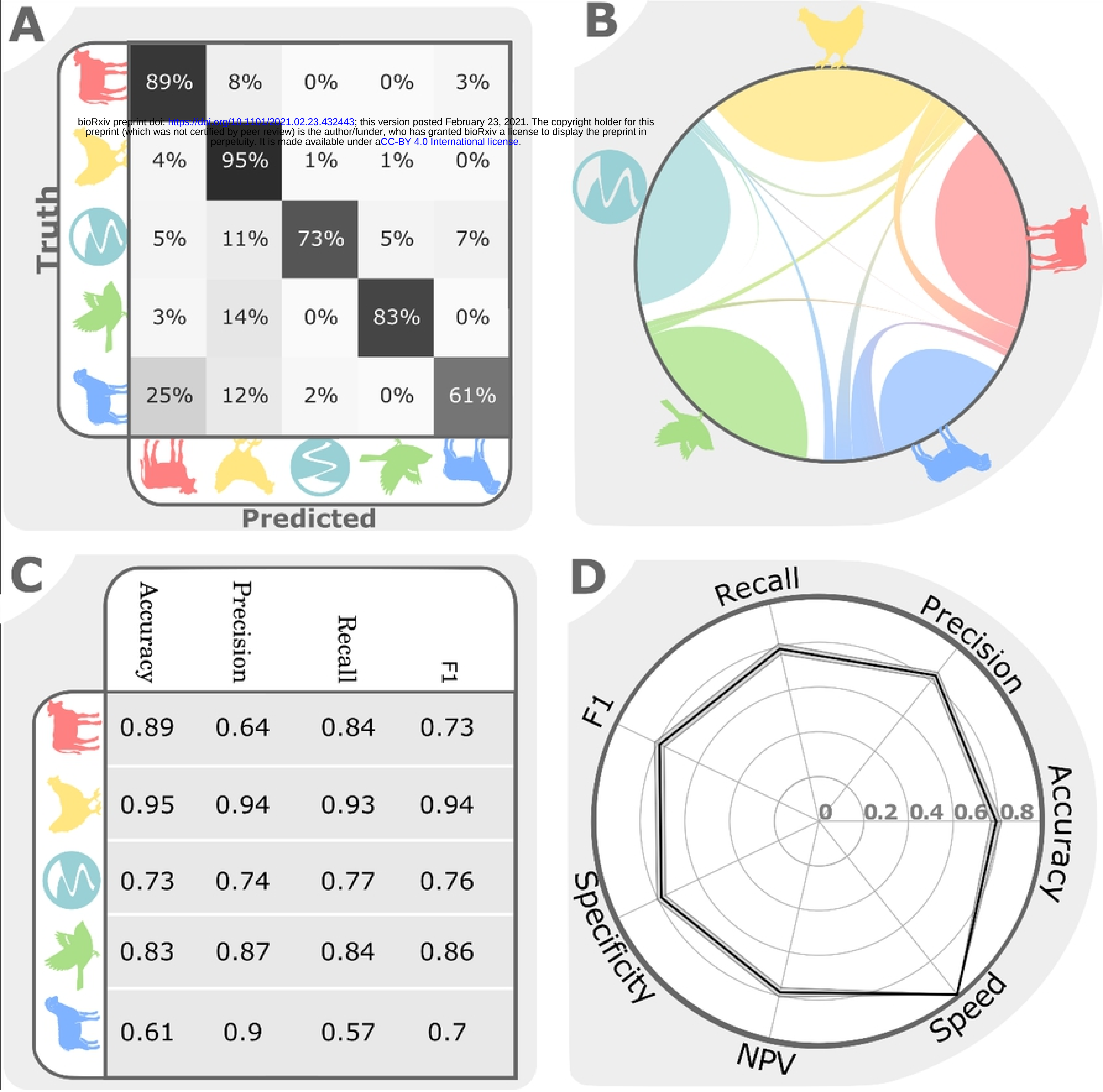
		Allelic profile	Sequenc	K-mers	Allelic profile	K-mers	Average
Dataset		MLST			cgMLST	WGS	
Simple learners	K-nearest neighbour	47.1±3	58.8±3	54.9±3	65.8±3	66.8±2	58.7±3
	Ridge Regression	36.5±2	65.5±3	62.1±3	29.0±3	66.3±3	51.9±3
	SVM (Linear Kernel)	41.6±2	66.7±3	54.0±3	50.8±3	67.2±2	56.1±3
	SVM (RBF Kernel)	22.2±2	66.0±3	63.0±3	20.0±0	65.0±3	47.2±2
	Naive Bayesian	34.7±2	45.9±3	45.2±2	42.2±2	54.8±3	44.6±3
	Decision tree	43.6±3	63.2±3	53.9±3	68.2±3	62.7±3	58.3±3
Ensemble learners	Random forest	61.9±3	67.9±3	62.4±3	76.0±2	69.4±2	67.5±3
	Extra-randomised forest	59.5±3	59.4±3	62.3±3	68.5±2	70.7±2	64.1±3
	XGBoost	59.9±3	65.8±3	53.9±3	81.3±2	68.1±3	65.8±3
Deep learners	1D-Convolutional NN	20.0±0	58.2±2	62.3±3	60.4±3	75.0±3	55.2±2
	Shallow Dense NN	19.7±0	53.3±2	59.4±3	65.7±3	69.1±3	53.4±2
	Deep Dense NN	20.0±0	57.8±3	62.5±3	66.6±2	70.3±3	55.4±2
	Recurrent NN	58.0±2	58.7±3	58.5±3	66.3±2	74.1±3	63.1±3
	LSTM NN	60.8±3	63.0±3	59.2±3	66.1±3	69.6±3	63.7±3
	Average	41.8±2	60.7±3	58.1±3	59.1±2	67.8±3	
	iSource	61.0±0					

B

Accuracy on Test-set with source composition reflecting human infection

		Allelic profile	Sequenc	K-mers	Allelic profile	K-mers	Average
Dataset		MLST			cgMLST	WGS	
Simple learners	K-nearest neighbour	61.0±1	61.6±2	61.4±2	68.8±2	71.8±2	64.9±2
	Ridge Regression	40.0±1	64.0±2	63.8±2	35.7±2	69.9±2	54.7±2
	SVM (Linear Kernel)	41.5±1	67.9±2	53.6±2	53.2±2	70.0±2	57.2±2
	SVM (RBF Kernel)	40.0±1	66.3±2	64.0±2	39.2±0	67.4±2	55.4±1
	Naive Bayesian	30.9±2	30.3±2	27.8±1	38.5±2	55.0±2	36.5±2
	Decision tree	46.3±2	63.0±2	49.2±2	67.6±2	65.0±2	58.2±2
Ensemble learners	Random forest	68.5±2	70.7±1	66.1±2	80.4±1	73.8±1	71.9±1
	Extra-randomised forest	67.4±2	61.9±2	65.5±2	72.4±0	76.2±2	68.7±1
	XGBoost	67.2±2	69.3±1	55.9±2	84.6±0	72.5±2	69.9±1
Deep learners	1D-Convolutional NN	39.2±0	64.2±1	67.5±2	65.9±2	78.3±1	63.0±1
	Shallow Dense NN	38.7±0	61.1±1	65.0±2	69.2±2	71.3±2	61.1±1
	Deep Dense NN	39.2±0	62.1±1	66.1±2	73.1±1	74.7±1	63.0±1
	Recurrent NN	66.2±1	64.2±1	61.3±2	72.5±1	77.5±2	68.3±1
	LSTM NN	68.8±1	68.1±1	65.3±2	71.0±1	74.4±1	69.5±1
	Average	51.1±1	62.5±1	59.5±2	63.7±1	71.3±2	
	iSource	64.4±0					

Figure 1



Truth \ Predicted	🐄	🐔	〰	🐦	🐕
🐄	89%	8%	0%	0%	3%
🐔	4%	95%	1%	1%	0%
〰	5%	11%	73%	5%	7%
🐦	3%	14%	0%	83%	0%
🐕	25%	12%	2%	0%	61%

	Accuracy	Precision	Recall	F1
🐄	0.89	0.64	0.84	0.73
🐔	0.95	0.94	0.93	0.94
〰	0.73	0.74	0.77	0.76
🐦	0.83	0.87	0.84	0.86
🐕	0.61	0.9	0.57	0.7

Figure 2

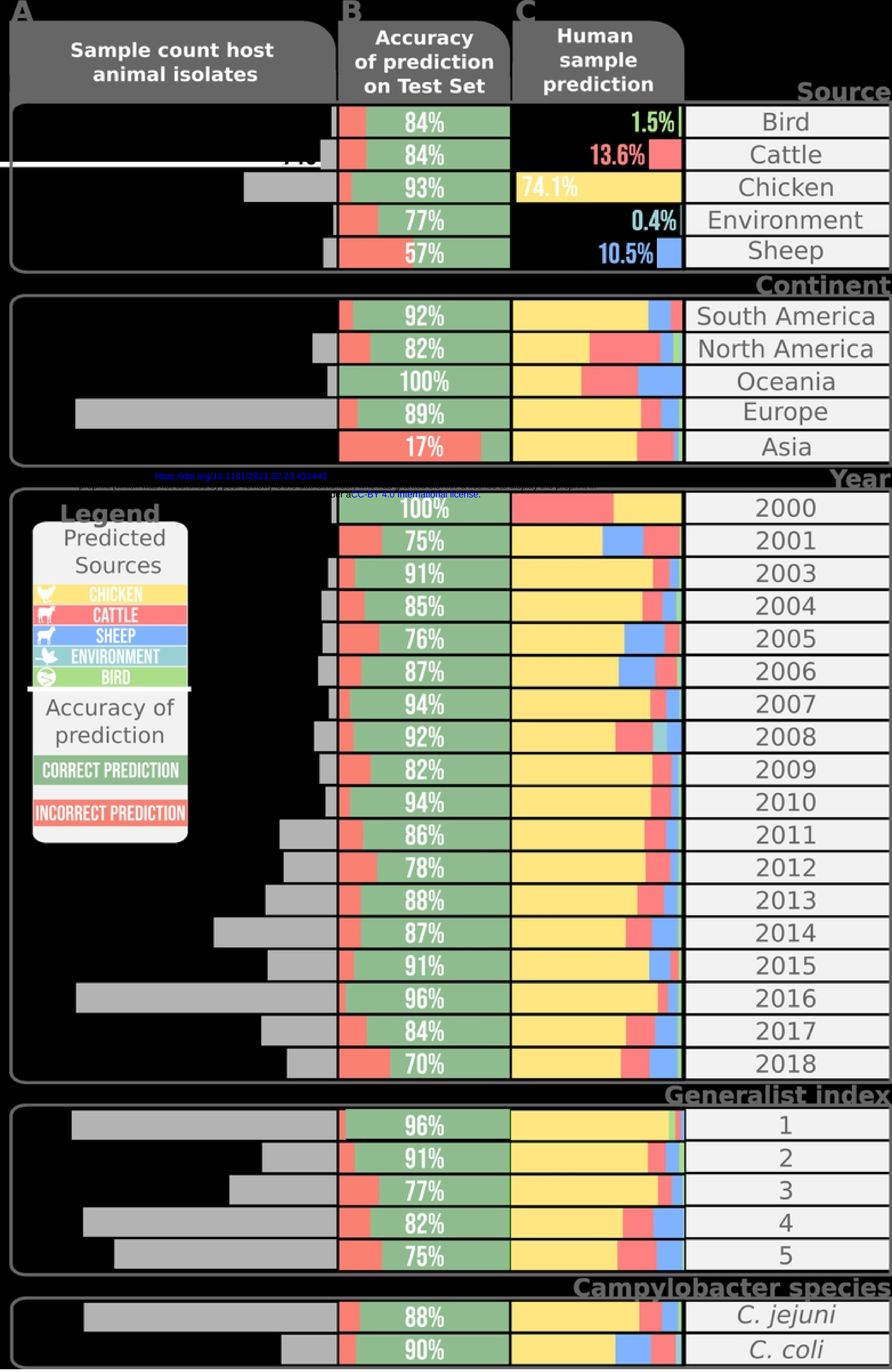
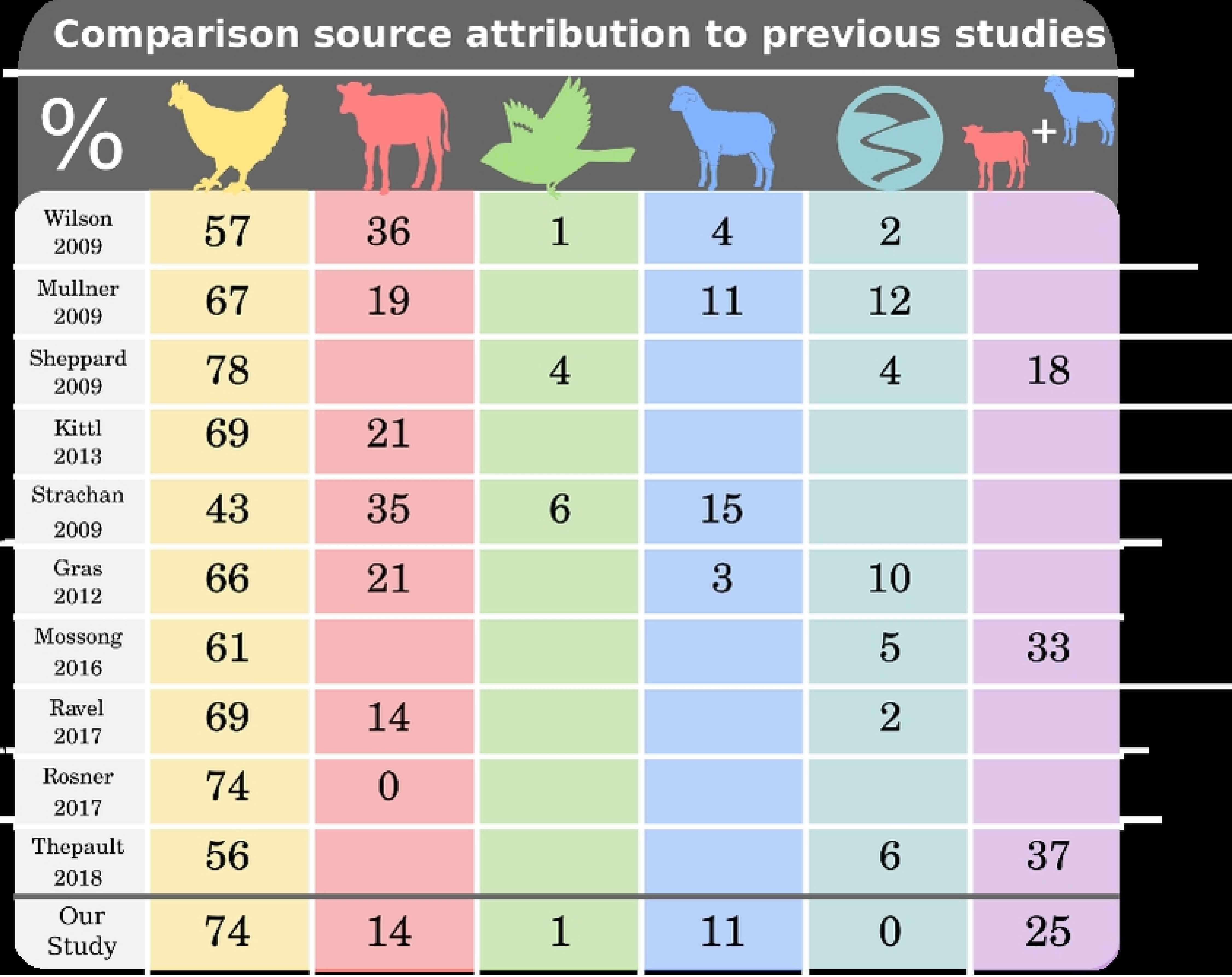

A
Sample count host animal isolates
B
Accuracy of prediction on Test Set
C
Human sample prediction
Source
84%
1.5%
Bird
84%
13.6%
Cattle
93%
74.1%
Chicken
77%
0.4%
Environment
57%
10.5%
Sheep
Continent
92%
South America
82%
North America
100%
Oceania
89%
Europe
17%
Asia
Year
100%
2000
75%
2001
91%
2003
85%
2004
76%
2005
87%
2006
94%
2007
92%
2008
82%
2009
94%
2010
86%
2011
78%
2012
88%
2013
87%
2014
91%
2015
96%
2016
84%
2017
70%
2018
Generalist index
96%
1
91%
2
77%
3
82%
4
75%
5
Campylobacter species
88%
C. jejuni
90%
C. coli
Legend
Predicted Sources
CHICKEN
CATTLE
SHEEP
ENVIRONMENT
BIRD
Accuracy of prediction
CORRECT PREDICTION
INCORRECT PREDICTION

Figure 3



Comparison source attribution to previous studies

%	🐔	🐄	🐦	🐑	🌊	🐄+🐑
Wilson 2009	57	36	1	4	2	
Mullner 2009	67	19		11	12	
Sheppard 2009	78		4		4	18
Kittl 2013	69	21				
Strachan 2009	43	35	6	15		
Gras 2012	66	21		3	10	
Mossong 2016	61				5	33
Ravel 2017	69	14			2	
Rosner 2017	74	0				
Thepault 2018	56				6	37
Our Study	74	14	1	11	0	25

Figure 4

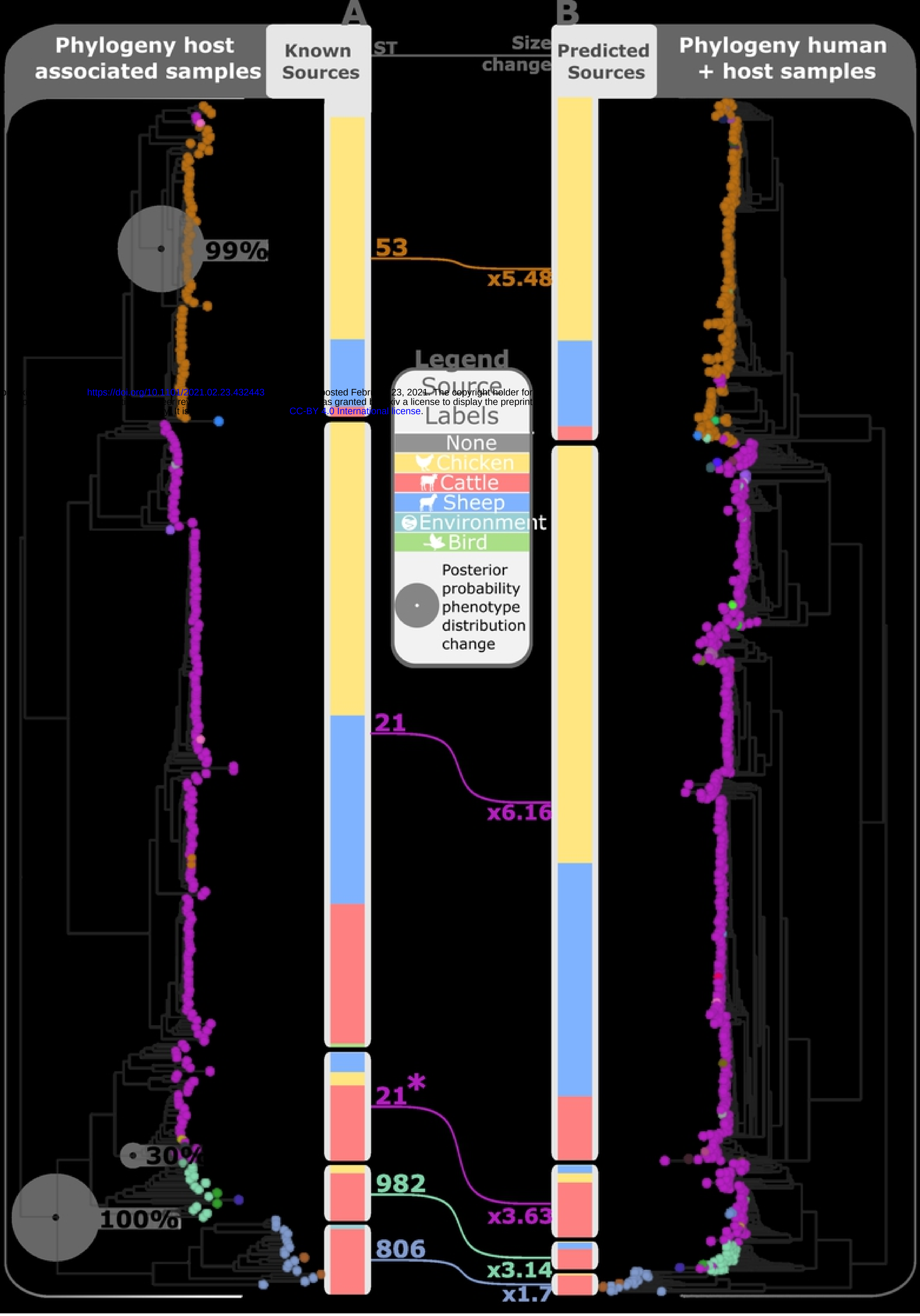

A
B
Phylogeny host associated samples
Known Sources
ST
Size change
Predicted Sources
Phylogeny human + host samples
99%
53
x5.48
Legend
Source Labels
None
Chicken
Cattle
Sheep
Environment
Bird
Posterior probability phenotype distribution change
21
x6.16
21*
x3.63
30%
100%
982
x3.14
806
x1.7

Figure 5