Hierarchical machine learning predicts geographical origin of *Salmonella* within four minutes of sequencing

Sion C. Bayliss[1]*, Rebecca K. Locke[2,3], Claire Jenkins[4], Marie Anne Chattaway[4], Timothy J. Dallman[5] and Lauren A. Cowley[2]

[1]Bristol Veterinary School, University of Bristol, Langford, Bristol, BS40 5DU, UK, s.bayliss@bristol.ac.uk

[2]Milner Centre for Evolution, Life Sciences Department, University of Bath, BA2 7AY, UK

[3]Genomic Laboratory Hub (GLH), Addenbrooke's Hospital, Cambridge University Hospitals NHS Foundation Trust, CB2 0QQ, UK

[4]Gastrointestinal Reference Services, UK Health Security Agency, Colindale, NW9 5EQ, UK

[5]Institute for Risk Assessment Sciences (IRAS), Utrecht University, 3508 TD Utrecht, Netherlands

*Corresponding Author

Abstract

Salmonella enterica serovar Enteritidis is one of the most frequent causes of Salmonellosis globally and is commonly transmitted from animals to humans by the consumption of contaminated foodstuffs. Herein, we detail the development and application of a hierarchical machine learning model to rapidly identify and trace the geographical source of *S.* Enteritidis infections from whole genome sequencing data. 2,313 *S*. Enteritidis genomes collected by the UKHSA between 2014-2019 were used to train a 'local classifier per node' hierarchical classifier to attribute isolates to 4 continents, 11 sub-regions and 38 countries (53 classes). Highest classification accuracy was achieved at the continental level followed by the sub-regional and country levels (macro F1: 0.954, 0.718, 0.661 respectively). A number of countries commonly visited by UK travellers were predicted with high accuracy (hF1: >0.9). Longitudinal analysis and validation with publicly accessible international samples indicated that predictions were robust to prospective external datasets. The hierarchical machine learning framework provides granular geographical source prediction directly from sequencing reads in <4 minutes per sample, facilitating rapid outbreak resolution and real-time genomic epidemiology.

Keywords: genomics, machine learning, epidemiology, public health, gastroenteritis

Introduction

Diarrhoeal disease is the most common illnesses caused by contaminated food, with 550 million people falling ill each year, including 220 million children under the age of 5 years (WHO 2022). High disease-burden foodborne pathogens represent a major public health concern, necessitating real-time epidemiological monitoring and follow-up. Outbreak investigations are often confounded by the complexity of the international food-trade networks which distributes zoonotic food-borne pathogens across the globe (Gould et al. 2017). Understanding the contributing factors, whether they be environmental, geographical or zoonotic is critical for designing public health intervention strategies to combat and prevent food-borne pathogen outbreaks (Pires et al. 2009).

In many developed nations, *Salmonella enterica* serovar Enteritidis is the most common cause of diarrhoeal disease (UKHSA 2021) and represents a significant economic and healthcare burden (Daniel et al. 2020). In the UK, nationwide vaccination and monitoring programmes have been responsible for a precipitous drop in detectable *Salmonella* from local animal products and a concurrent drop in human infection rates (Surveillance, Zoonoses, Epidemiology and Risk Food and Farming Group 2007; Tam et al. 2012). However, recent studies have identified imported foodstuffs (McLauchlin et al. 2019; Somorin, Odeyemi, and Ateba 2021) and foreign travel (PHE 2017) as more pertinent *Salmonella* infection risks.

Whole genome sequencing (WGS) has become a powerful tool for untangling complex networks of pathogen dissemination by providing high-resolution sub-typing for transmission tracing as well as detailed information on antimicrobial and virulence status (Allard et al. 2018). Since 2014, the UK Health Security Agency (UKHSA) has routinely applied WGS to all clinically identified cases of *Salmonella* in the UK alongside collecting detailed metadata, such as patient recent foreign travel (Ashton et al. 2016). This programme has been instrumental in identifying various international outbreaks, such as the largest known salmonellosis outbreak in Europe where, between 2015 and 2018 *S.* Enteritidis-contaminated eggs resulted in 1,209 reported cases across 16 countries (Dallman et al. 2016; Pijnacker et al. 2019). This approach to inferring geographical source from pathogen genomes has historically required phylogenetic population structure analysis which requires significant bioinformatic skills, is computationally expensive and scales poorly with the increasingly vast collections of bacterial genomes available for analysis (Cowley et al. 2016, Dallman et al. 2016). Furthermore, due to the resources involved, this type of investigation is often only undertaken in exceptional cases representing a threat to public health.

In order to promote rapid outbreak responses, novel methods are required to translate the increasing volumes of pathogen genomes generated by surveillance programmes into immediate and actionable information for epidemiologists. *S.* Enteritidis represents a desirable initial target for such tools due to its large public health burden. WGS provides high-resolution information about pathogen strain relatedness and, by association, contains contextualised information on geographical or host origin. In the case of *S.* Enteritidis, which has a population structure observably stratified by geographical source (Li et al. 2021, Feasey et al. 2016), there is potential for genomics to provide key information for successful epidemiological follow-up, namely the likely country of origin of an infectious strain, and an opportunity to rapidly enact intervention strategies. Herein we present the first implementation of a hierarchical machine learning (hML) classifier for geographical source attribution of *S.* Enteritidis genomes. We applied this model to the UKHSA's large and uniquely detailed genomic database of clinical *S.* Enteritidis isolates collected between January 2014-April 2019. Using these data, we have generated a fully automated analysis pipeline with which to monitor imported cases of *S.* Enteritidis. The hML model was structured to provide a granular and multi-level prediction of the geographical origin of an *S.* Enteritidis genome and can do so in under 4 minutes directly from raw sequencing reads.

Results

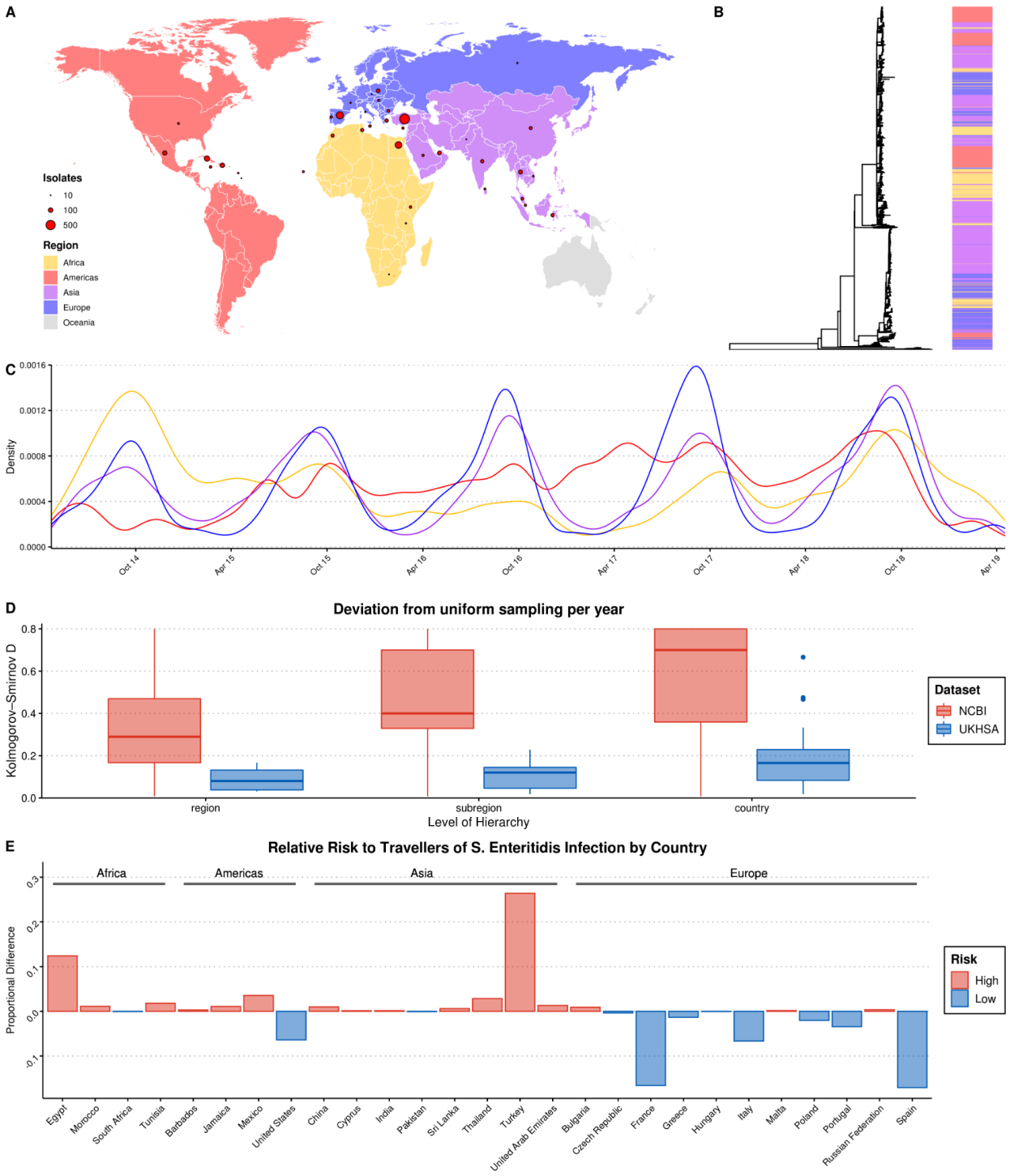


Figure 1. Summary of *S.* Enteritidis isolates collected by the UKHSA from UK clinical patients who recently reported foreign travel between 2014-2019. A) Geographical distribution of 2313 *S.* Enteritis isolates by reported foreign travel. Variably sized points

represent the number of samples per country. The map is coloured by region (Africa: yellow, Americas: red, Asia: purple, Europe: blue). **B)** Maximum likelihood phylogenetic tree of 2313 *S.* Enteritidis isolates with bar coloured by region of origin. **C)** Kernel density plot indicating sampling density per region through time. **D)** Comparison of the consistency of sampling effort of the UKHSA to all publicly available *S.* Enteritidis isolates on NCBI for the same period. Isolates were resampled to control for variable sample number per year and compared to a uniform distribution using the Kolmogorov-Smirnov D statistic (NCBI = red, UKHSA = blue). Higher values indicate greater deviation from a uniform distribution. **E)** Relative risk per country of acquiring *S.* Enteritidis infection when travelling. A risk score was generated by dividing the proportion of UKHSA clinical isolates per country by the proportion of all UK travellers travelling to that country as recorded by the Office of National Statistics (ONS). Only countries present in both datasets were used to calculate proportions.

The UKHSA genomic surveillance programme consistently samples *S.* Enteritidis associated with international travel through time

Broad and unbiased surveillance by the UKHSA of all clinically reported *S.* Enteritidis cases in the UK coupled with returning traveller data has provided a large genomic dataset representative of *S.* Enteritidis infections in the UK between 2014-2019 **(Figure 1)**. This consisted of 10,223 isolates, of which 3,434 had matched recent reported travel data collected as a part of the UKHSA's 'enhanced surveillance' programme.

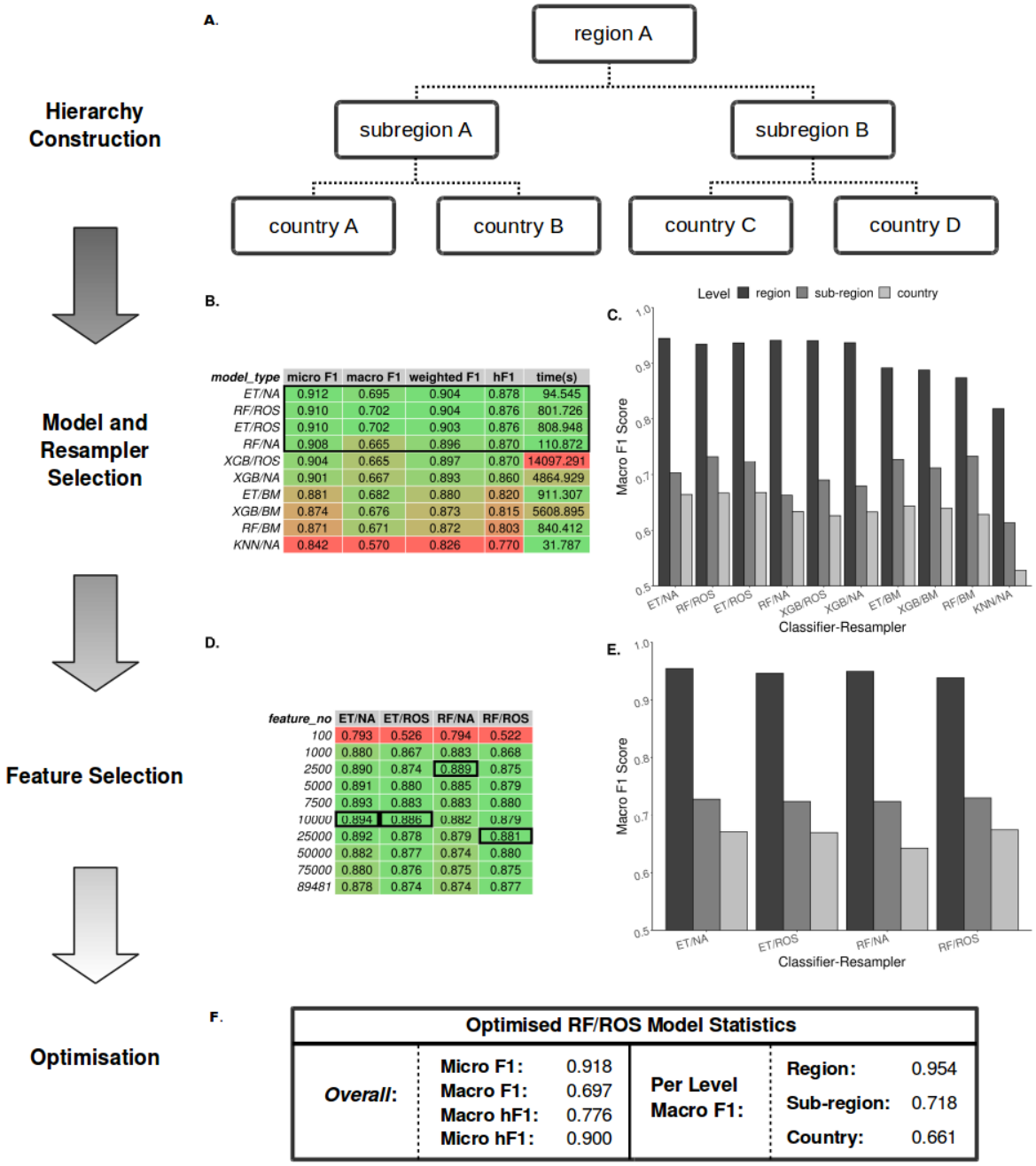
Recent travel was reported from 122 countries across 5 continents. A source of potential bias, common to bacterial genomics analysis, is the over-representation of clonally related isolates due to their increased prevalence during outbreaks (Feil and Spratt 2001; Ebel et al. 2016). A single, random representative isolate per country was selected for each clone, defined as a 5-SNP cluster by SNP Address, in order to reduce the influence of highly related clonal outbreaks on the resulting ML model (Dallman et al. 2018). The 5-SNP cutoff is routinely used as the definition of an outbreak by the UKHSA for genomic disease surveillance (Chattaway et al. 2019). After quality-filtering, downsampling and removal of low incidence countries (<10 isolates), 2,313 genomes from 38 countries from 4 continents were included in the final dataset for ML (**Figure 1A**).

Grouping these countries by geographic region and subregion using the UN M49 Standard for Regional Codes provided a framework for a granular geographical analysis and an established hierarchy comprised of countries/subregions/regions (Statistics Division of the United Nations Secretariat 2020). Phylogenetic analysis indicated that the dataset displayed a strong phylogeographical signal, with large clusters of isolates from geographically related countries clustering together (**Figure 1B**). An interactive maximum likelihood phylogenetic tree is available

in Microreact at https://microreact.org/project/kQEhcTy4ohcqN9bjcPUWLw-ukhsasenteritidishml (Argimón et al. 2016). *S.* Enteritidis infection rates in Europe and Asia were highly seasonal, with significantly increased infection during the summer months (**Figure 1C**). This was less pronounced in travel to/from the Americas and Africa.

An analysis of the consistency of sampling effort over time identified that some notable countries in the dataset were comprised of samples collected predominantly during a single year, such as Sri Lanka, Tunisia and Dominica, and others were missing data from one or more years (**Figure S1**). However, a comparison of the UKHSA collection to a dataset of location/date matched genomes from the NCBI Pathogen Detection Database identified that the UKHSA dataset represented a significantly more consistent sampling effort which was less influenced by sporadic outbreaks (**Figure 1D**).

The countries with the highest number of travel associated *S.* Enteritidis cases were Turkey (804), Spain (357), Egypt (343), Cuba (190) and the Dominican Republic (117) (**Figure 1A**). Controlling for the volume of UK travellers as recorded by the Office of National Statistics over a matched time period allowed for the identification of countries with disproportionate low- or high-risk of *S.* Enteritidis infection after travelling (**Figure 1E**)(Office of National Statistics 2020). Travellers to Turkey and Egypt were at a disproportionately higher risk of *S.* Enteritidis infection during this period. Conversely, travellers to France and Spain, two of the more popular travel destinations for UK travellers, were very low-risk. This highlights that the dataset was a product of the volume of UK-travel combined with a variable risk of infection per country. Consequently, there was a large degree of class imbalance (i.e. different number of isolates per country) in the dataset used for ML, in addition to low/absent sampling coverage of some parts of the globe. However, when considering larger geographical groupings (i.e. subregion/region) these imbalances were less pronounced or absent.



model_type	micro F1	macro F1	weighted F1	hF1	time(s)
ET/NA	0.912	0.695	0.904	0.878	94.545
RF/ROS	0.910	0.702	0.904	0.876	801.726
ET/ROS	0.910	0.702	0.903	0.876	808.948
RF/NA	0.908	0.665	0.896	0.870	110.872
XGB/ROS	0.904	0.665	0.897	0.870	14097.291
XGB/NA	0.901	0.667	0.893	0.860	4864.929
ET/BM	0.881	0.682	0.880	0.820	911.307
XGB/BM	0.874	0.676	0.873	0.815	5608.895
RF/BM	0.871	0.671	0.872	0.803	840.412
KNN/NA	0.842	0.570	0.826	0.770	31.787

feature_no	ET/NA	ET/ROS	RF/NA	RF/ROS
100	0.793	0.526	0.794	0.522
1000	0.880	0.867	0.883	0.868
2500	0.890	0.874	0.889	0.875
5000	0.891	0.880	0.885	0.879
7500	0.893	0.883	0.883	0.880
10000	0.894	0.886	0.882	0.879
25000	0.892	0.878	0.879	0.881
50000	0.882	0.877	0.874	0.880
75000	0.880	0.876	0.875	0.875
89481	0.878	0.874	0.874	0.877

Figure 2. Summary statistics showing model and resampling scheme selection, feature selection and optimisation. A) Example schematic of a geographical hierarchy based upon the UN M49 Standard for regional codes. **B)** Table of summary statistics for the ten top-performing co-optimised model and resampling methods from a cohort of 36 combinations, sorted by hF1 (high-low). Training time reported in final column in seconds. A black box indicates the top four models used for feature selection. **C)** Grouped bar chart comparing macro F1 per hierarchical level for the ten top-performing model/resampled combinations. **D)** Table of summary statistics for random forest feature selection applied to the four top-performing co-optimised model and resampling methods. Black boxes indicate the optimal number of features

per combination. **E)** Grouped bar chart comparing macro F1 per hierarchical level for the four top-performing co-optimised model and resampling methods after feature selection optimisation. **F)** Summary statistics for the final optimised Random Forest - Random Oversampler model (25,000 features selected). ML Model abbreviations: Random Forest (RF), XGBoost (XGB), Extra Trees (ET), K-Nearest Neighbours (KNN). Resampler abbreviations: No Resampling (NA), Random Undersampler (RUS), Random Oversampler (ROS), Balancing Mean (BM), Hierarchical Mean (HM).

A novel hierarchical model provides real-time geographical source attribution prediction directly from sequencing reads in under four minutes

Taking advantage of the hierarchical structure of geographical data, we designed a multi-level hML classifier following a “Local Classifier per Node” framework (Silla and Freitas 2011). This was made up of 15 individual multi-class classifiers, one per node (1 root, 4 regional, 11 sub-regional). In total, 53 individual classes (4 regions, 11 sub-regions and 38 countries) were predictable by the model. Sample classification was performed using a top-down approach, where samples are first classified at the root node into ‘regions’ (e.g. Africa) then, if the predicted probability is greater than a minimum threshold value (0.5), samples are passed to the appropriate regional node to be classified into a nested subset of ‘subregions’ (e.g. Northern Africa) and finally passed to a subregional node where they are classified into a nested subset of ‘countries’ (e.g. Egypt) (**Figure 2A**). Using this scheme no samples could be attributed to a class which was not a predicted class of a previous classifier (i.e. Africa→Southern Asia→France is not possible, but Africa→Northern Africa→Egypt is). Sample classification was exclusive, disallowing multiple classifications on the same hierarchical level for a single sample.

For rapid and minimal sample processing and to provide end-to-end sample classification directly from the sequencer, the model was trained on filtered unitig patterns (presence-absence) generated from quality controlled genomic short-read data files. Each local classifier per node was trained using only data pertinent to that node (e.g. a local subregion classifier was trained only on the data from its constituent countries). This end-to-end process, from FASTQ to sample prediction, is available on GitHub (https://github.com/SionBayliss/HierarchicalML).

Due to the imbalanced nature of the real-world surveillance dataset, it was necessary to test a range of classifier and resampler algorithms before selecting the top performing models (**Figure 2B-C**). The top 4 models subsequently underwent feature selection (**Figure 2D-E**) followed by parameter optimisation using the TPOT genetic algorithm (**Figure 2F**). The optimised hML model produced a more accurate classification of the test dataset than a ‘flat’ classifier applied to a similarly pre-processed dataset (macro F1: 0.61) (**Table S1**). Based on these comparisons, the most desirable assessment metrics overall (i.e. high macro F1 at the country level, **Figure 2E**) were from feature selection by a Random Forest (RF) classifier (25,000 unitig patterns)

before random oversampling to correct for class imbalance and final classification using an optimised RF model. Assuming <100x read coverage, the entire pipeline takes ~3.5 min to classify a novel samples using the pre-optimised model.

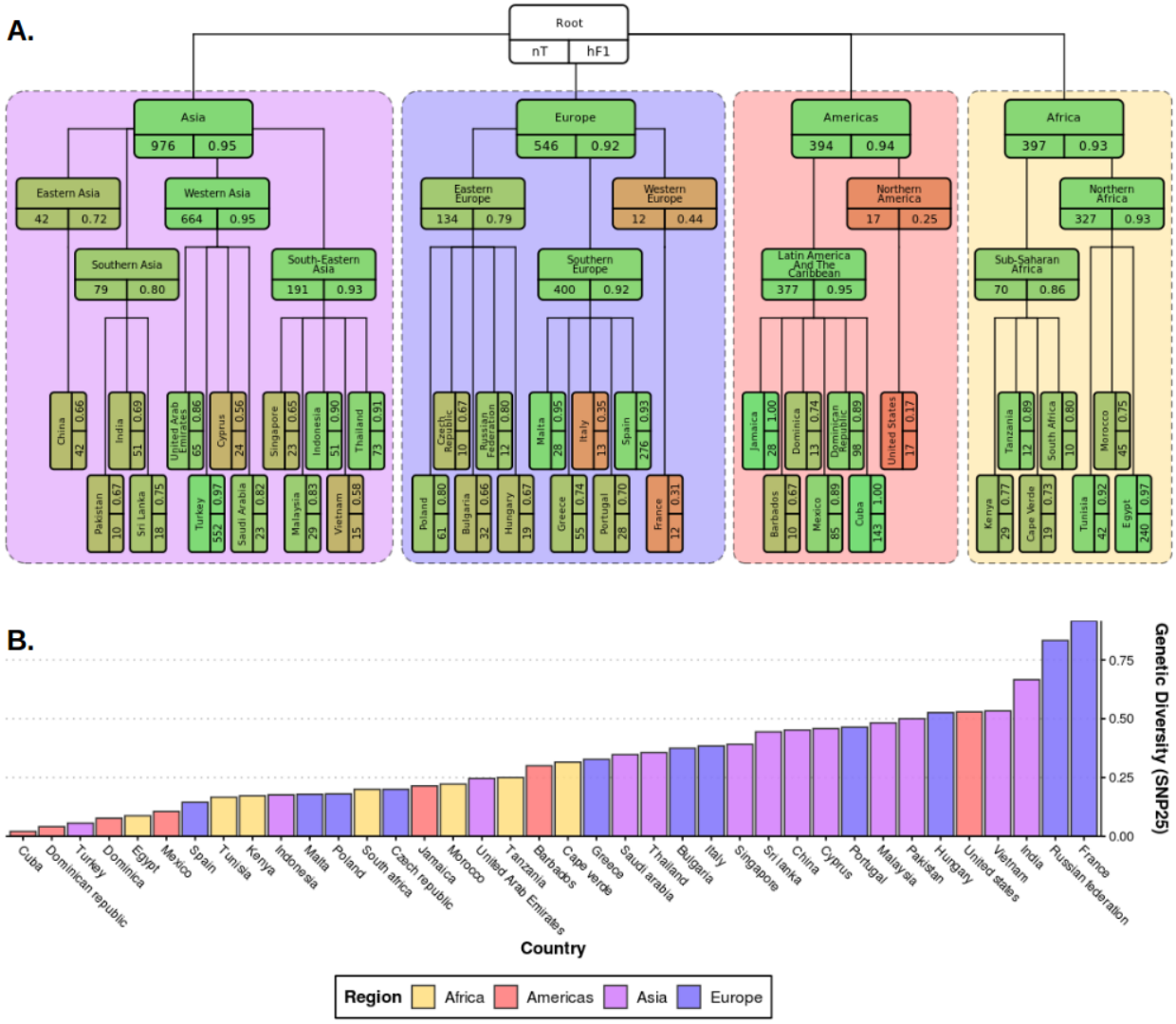


Figure 3. Plots summarising test results from hML model and genetic diversity of dataset. A) Diagrammatic representation of classification metrics of the hML classifier applied to the test dataset. Links between classes/nodes in the hierarchy are indicated by connecting lines. Boxes represent individual classes in the ML model and are coloured by their hierarchical F1 (hF1) scores. The top panel of each class box displays the class label, the bottom left panel indicates the total number of samples for that class before the train/test split (75%/25%) and the bottom right panel shows the class hF1 score calculated from the test dataset. Classes within individual 'regions' (continents) were contained in a coloured background panel. **B)** Bar plot of genomic diversity per country. Genomic diversity was estimated as the number of 25 SNP single linkage clusters divided by the total numbers of samples per class. Panels (A) and bars (B) were coloured according to region (Africa: yellow, Americas: red, Asia: purple, Europe: blue).

Granular predictions are provided at a regional, subregional and individual country level

Classification metrics were highest at the regional level (macro F1: 0.954), but less discriminatory at the sub-regional (macro F1: 0.718) and country levels (macro F1: 0.661) (**Figure 3A**). There was a moderate degree of variation in predictive accuracy between different geographical locations. Africa was the most consistent accurately classified region, with all country and subregional classes presenting a hF1 of >0.7. In The Americas, Latin American and Caribbean countries showed very high classification metrics (hF1: >0.8), whereas samples for the United States were consistently misclassified. All Asian classes were classified at moderate to high accuracy (hF1: 0.58-0.95). Europe was generally well classified (hF1: >0.66), but contained two classes, France and Italy, which were classified poorly (hF1: ~0.3). Further scrutiny of poorly performing classes showed a correlation between lack of training data and lower prediction accuracy (**Figure S2**). This did not fully explain the observed results as some countries with a similarly low number of samples to poorly predicted countries (e.g. Czech Republic, Pakistan) showed moderate classification accuracy. An analysis of genetic diversity within classes indicated that at least two of the most poorly classified countries, France and the United States, displayed both low numbers of samples and high genetic diversity (**Figure 3B, Figure S3**). Samples from recent travel to France were particularly diverse, arising from multiple highly diverse clades of *S.* Enteritidis. A range of countries which were commonly visited by UK travellers, such as Cuba, Egypt, Indonesia, Jamaica, Malta, Spain, Thailand, Tunisia and Turkey, were were well predicted (hF1: >0.9).

In summary, these results suggest that the optimised model can attribute the geographical source of *S.* Enteritidis isolates with high confidence at a regional level whilst also providing more nuanced and granular predictions for a range of countries regularly visited by UK travellers with a very high degree of accuracy.

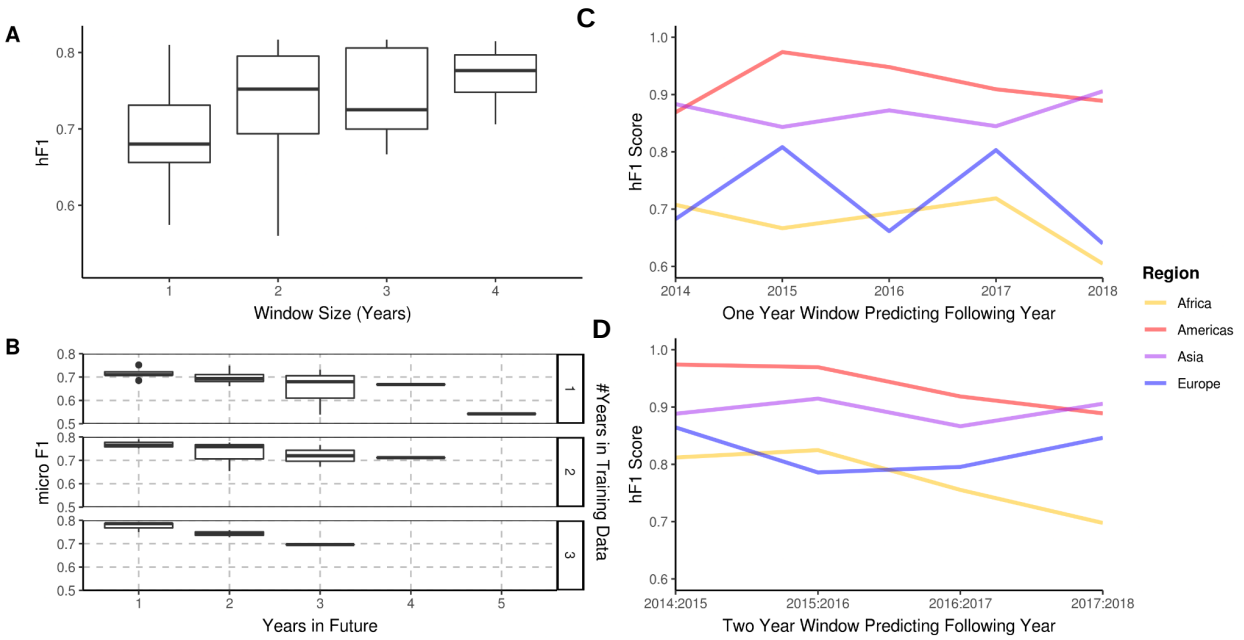


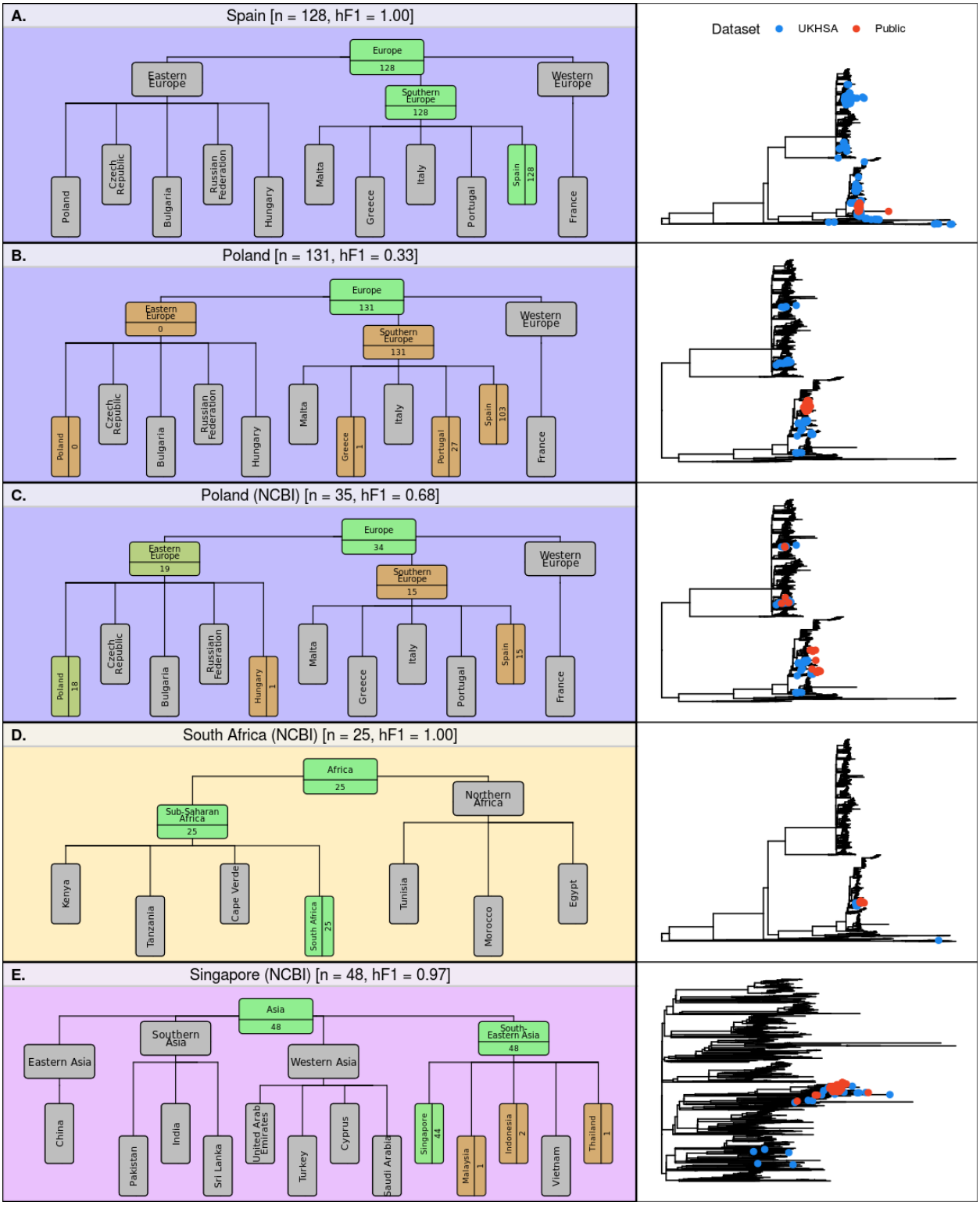
Figure 4. A longitudinal analysis of the predictive performance of hML models on 2313 *S. enteritidis* samples. **A)** The hF1 scores of hML RF models trained on a subset of samples from a 1-4 year moving window predicting the following year. **B)** The micro F1 scores of hML RF models trained on a subset of samples from a 1-3 year moving window predicting 1-5 years into the future. **C)** Regional hF1 scores of hML RF models trained on one-year sampling windows predicting the following year. **D)** Regional hF1 scores of hierarchical models trained on two-year sample windows predicting the following year.

Models demonstrate durability to future predictions with two years previous training data proving sufficient signal for accurate subsequent year predictions

Bacterial population lineage composition is not expected to remain static through time, therefore predictive models based on genomic data require periodic retraining. To understand the amount of data required for accurate prospective prediction, we compared the outcomes of four yearly window sizes (1, 2, 3 and 4 years) for the prediction of subsequent years (**Figure 4A**). Predictive accuracy and consistency of prediction of the subsequent year improved on increasing window size, with hF1 beginning to asymptote after two years. The largest improvement was observed between 1-2 years' worth of data. A minor decrease in predictive accuracy (micro F1: ~0.05 per year) was observed for each additional year into the future (**Figure 4B**). A regional breakdown of the hF1 score indicated that a one-year window varied in predictive accuracy per region per year (**Figure 4C**), but that a two-year window provided more consistent and accurate predictions (**Figure 4D**). These results indicate that, should this model be instituted for ML-enhanced genomic surveillance, it would require retraining on the two

256 previous years samples each year to provide an optimal trade-off between predictive accuracy
257 and training time.

258



259

Figure 5. Hierarchical classification summaries for five additional validation datasets. **A)** 128 samples from an international outbreak which originated in Spain in 2015 (Inns et al. 2017) **B)** 131 samples from a large-scale international outbreak which originated from Polish eggs between 2015-2018 (Pijnacker et al. 2019). **C)** 35 samples from Poland uploaded to the NCBI database between 2014-2019. **D)** 25 samples from South Africa uploaded to the NCBI database between 2014-2019. **E)** 48 samples from Singapore uploaded to the NCBI database between 2014-2019. The number of samples assigned per class are indicated for classes relevant to the query dataset. Class boxes are coloured by the proportion of correctly/incorrectly classified samples (correct: green, incorrect: red). Right-hand panel for A-E displays phylogenetic tree indicating where validation data (red) and training data (blue) cluster for that class.

The optimised hML model provides accurate predictions in 4/5 validation datasets

The hML model was further validated by application to a series of external, independent datasets (Table S4). The initial datasets were from two UK-based, well-characterised and epidemiologically-traced imported food outbreaks. Sample redundancy between validation and training datasets was removed before comparison. A 2015 outbreak epidemiologically-traced to eggs imported from Spain (Inns et al. 2017) was 100% correctly attributed to a Spanish origin (**Figure 5A**). A multi-country outbreak originating from Polish egg farms (Pijnacker et al. 2019), comprising of two distinct lineages each differing by 5 or fewer SNPs, was correctly attributed to a European origin (131/131 cases), but subsequently misattributed to a Southern Europe (131/131 cases), Spanish origin (103/130) (**Figure 5B**). This complex outbreak was a particularly difficult test case for the model, as it had been continuously causing cases in 16 European countries for several years (2015-2018). The outbreak cases were also phylogenetically distinct from those associated with travel to Poland in the UKHSA dataset (**Figure 5B**).

The model was further tested on datasets extracted from public databases. Three countries from three different regions were identified from the NCBI database as having acceptable sample numbers (>20), falling within the timeframe of the current model (2014-2019) and having been sampled from a country included in the current model hierarchy (South Africa, Singapore and Poland). The Polish dataset, sampled from poultry, was attributed to a European origin with high accuracy (34/35, 97.1%), 19 of these were subsequently attributed to an East European origin (19/35, 54.3%) or which 18 were attributed to Poland (18/35, 51.4%) (**Figure 5C**). The remaining 15 samples were misclassified as having a Spanish origin (15/35, 42.9%). The South African dataset of clinical cases, was correctly attributed to a South African origin with 100% accuracy (25/25) (**Figure 5D**). The Singaporean dataset of primarily human cases was correctly attributed to South-East Asia (48/48) with 91.7% of samples being correctly attributed to a

Singaporean origin (44/48) and 4 samples being attributed to Indonesia (2), Malaysia (1) and Thailand (1) (**Figure 5E**).

Application of the hML model to independent validation datasets indicated that that the model provides highly accurate and granular predictions for single-country outbreaks which occur over short time-frames. A large-scale, long-term, multi-country outbreak was poorly predicted by the model, most likely because many outbreak-associated isolates were mislabelled, but present in the training dataset (i.e. not labelled as Poland). However, application of the hML model to a range of other independent validation datasets provided highly accurate and granular predictions.

Discussion

Outbreaks caused by foodborne pathogens, where rapid response times are essential for effective interventions, represent an epidemiological challenge for public health bodies as they arise from complex interconnected global supply chains which include many potential sources of infection. Our optimised hML model generates accurate predictions of the geographical origin of *S.* Enteritidis genomes directly from raw read data in under 4 minutes per sample. The output of this pipeline is a predicted probability per hierarchical level (i.e. region, sub-region and country levels) allowing for granular source attribution alongside an assessment of confidence in the prediction. The observed classification accuracy of the model was high, varying across both hierarchical levels and individual classes (**Figures 2-3**). Accuracy was highest at the regional level (macro F1: 0.954), which contained the highest number of samples for training and the lowest level of class imbalance, before dropping at the subregional (macro F1: 0.718) and country levels (macro F1: 0.661), both of which contained generally lower and more variable numbers of samples, increasing class imbalance (**Figure 3**). It should be noted that these macro values, being an average of the F1 score across all classes, were strongly influenced by outliers representing a handful of classes with poor classification accuracy (e.g. United States, France, US, Italy). Variation in classification accuracy was negatively associated with both low sample number and increasing within-class genetic diversity (**Figure S2-3**), although this did not fully explain the variation in predictive accuracy observed in the model, suggesting other complex factors may be involved. Although variation in predictive accuracy was observed, a number of commonly visited countries were predicted with extremely high predictive accuracy (e.g. Cuba, Egypt, Indonesia, Jamaica, Malta, Spain, Thailand, Tunisia and Turkey) which would be of high utility for UK epidemiologists tracing foodborne outbreaks .

Selection and optimisation of the ML classifier and resampler methods, as well as pre-classification feature selection was performed in a stepwise manner to assess the impact of each step on the resulting model. Ensemble classifiers outperformed other classifiers (**Figure 2**, **Table S2**). Interestingly, considering the large variation observed in some nodes of the hierarchy, class imbalance was not found to have a large impact on the resulting model. Two of the top four best-performing classifier/resampler combinations included no resampling. These were both ensemble classifiers, suggesting that these were better able to account for the high dimensionality, high collinearity and variable levels of class imbalance than other classifier models. The resampling itself was observed to differentially impact on the various hierarchical levels (**Figure 2**). The model required a sample to be classified at a higher hierarchical level before passed to the next nested level (e.g. classified as European before being classes as West/East/South European) which biased model selection towards favouring classifiers which improved regional and subregional predictions. It should be noted that each classifier/resampler/feature selection step was applied uniformly across the hierarchy prior to parameter optimization, to allow for an assessment of individual steps on the resulting model. An attractive, although less informative, approach for future retraining would be to optimise all steps (classifier/resampler/feature selection) on a per node basis using an AutoML method, such as the genetic algorithm used in the current work to optimise hyperparameters in the final hML RF model (Olson et al. 2016). Validation using external datasets indicated that the model robustly predicted the majority of novel samples, but was less effective for complex multi-country, multi-year outbreaks (**Figure 5**).

We acknowledge that this study has a number of important limitations. Whilst we have presented evidence that the hML model can robustly predict geographical source for *S.* Enteritidis, the COVID-19 pandemic has resulted in a vast disruption to international travel over the past two years. The UKHSA is already seeing a return to pre-pandemic levels of imported infection and expect the pattern of seasonality of imported *S.* Enteritidis infections to return. However, there may be unquantified impacts on the global food network which could result in a reduction in the prediction accuracy of the current model and would likely require retraining on post-pandemic data. Furthermore, the dataset, collected as part of the UKHSA's national genomic surveillance program, is a product of two factors; the countries to which UK residents commonly travel and the variable likelihood of *S.* Enteritidis infection whilst in these countries (**Figure 1**). Both factors influence the geographical distribution of sampling locations present in the dataset as well as the number of available samples per country. For example, of the 122 countries present in the initial dataset only 38 (31%) were present in sufficient quantities over

the surveillance period to include in the model. The second factor influencing the composition of the dataset is the variable risk of infection posed by the countries present in the dataset. Some commonly visited countries posed only minor risk of infection (e.g. France) whilst others posed a disproportionately high risk of infection on visiting (e.g. Turkey) (**Figure 1E**). Due to the complex nature of both the global food supply chain and pathogen transmission dynamics it is likely that these factors will cause misattribution in some cases; for instance, UK-travellers might travel to country A and consume contaminated foodstuffs imported from country B which was absent from the training data, and the model would then 'correctly' assign future cases to country A. One might also imagine that country C, which is only rarely visited by UK-travellers, produces a contaminated foodstuff and exports it to country D, which is often visited by UK-travellers, causing the model to misattribute a sample from country C to D. In these cases the model would provide useful information on the origin of infection, but would not identify the true reservoir of the pathogen. Evidence of this form of misattribution has been presented in the Polish eggs outbreak dataset used for validation **(Figure 5B)**. This outbreak represents a particularly difficult attribution problem as many of the contaminated eggs were distributed to 18 different EU countries during the sampling period used to train the model – allowing for the inclusion of potentially erroneous sample labels in the training data and confounding the discriminatory signal. This outbreak represents an important limitation of the current model and illustrates how difficult multi-country outbreaks or long-established food network contamination pathways can confound accurate classification.

We have presented a conceptually simple hML model which is able to successfully predict the geographical source of infection for a wide range of popular destinations frequented by UK travellers. However, the basis of this work was a single dataset which, although exceptional in breadth of sampling and metadata, compositionally reflects locations routinely visited by UK travellers. A number of destinations were present in the dataset for which there is little to no data present in public databases during a matching time period (e.g. Malta). This demonstrates the additional utility provided by the collection of recent travel information as part of a national surveillance programme to generate consistent geographical coverage for disease-sampling over a much larger geographical area. One could imagine that coordination between public health bodies with complementary citizen travel preferences would allow for a small network of national genomic surveillance initiatives to provide global coverage of gastrointestinal diseases without the costly necessity of instituting a comprehensive global surveillance initiatives. If this was coupled with a similar hML model to the one detailed here, then rapid and precise global predictions of the source of gastrointestinal disease outbreaks could be achieved.

This study provides a framework for future hML applications in the area of pathogen genomics. Other geographically stratified problems which might benefit from a hierarchical approach include antimicrobial resistance in *Escherichia coli* (Ingle et al. 2018), transmission of *Staphylococcus aureus* in hospital networks (Donker et al. 2017) and application to other serovars of *Salmonella enterica* (Lupolova et al. 2017). The model presented herein is conceptually simple and does not incorporate temporal information, allow for multiple labels to be assigned to samples or incorporate animal host/food source information, which may improve prediction of multi-country outbreaks or provide additional information to further enhance epidemiological follow-up (Lupolova et al. 2017; Zhang et al. 2019; Lupolova, Lycett, and Gally 2019). This work represents the first application of hML for the automation of genomics-based geographical source attribution. The rapidity of predictions from raw sequencing data should greatly enhance epidemiologists' ability to trace the source of gastrointestinal outbreaks by condensing complex genetic data into understandable and actionable outputs.

Methods

Genome Collection and Processing

The initial dataset consisted of 10,223 *S.* Enteritidis isolates collected and sequenced by UKHSA between 2014-2019 as a part of their routine disease monitoring programme. Raw read data was downloaded from the Short Read Archive (Bioproject: PRJNA248792) **(Table S3)**. Reads were filtered using Trimmomatic 0.39, residual Illumina adapter sequences were removed, the first and last 3 base pairs in a read were trimmed, before a sliding window quality trimming of 4 bases with a quality threshold of 20 was applied and reads of less than 36 bp (after trimming) were removed (ILLUMINACLIP:PE_All.fasta:2:30:10:2:keepBothReads SLIDINGWINDOW:4:20 LEADING:3 TRAILING:3 MINLEN:36) (Bolger, Lohse, and Usadel 2014). The coverage of the resulting file was estimated against the size of *S.* Enteritidis reference genome P12510 (Thomson et al. 2008) and downsampled to ~100x coverage using in-house scripts. Musket 1.1 was applied for k-mer spectrum read correction using a k-mer size of 31 bp. Unitigs were generated directly from filtered reads using bcalm 2.2.3 with a k-mer size of 31 bp and a minimum k-mer abundance of 6 (estimated from data) (Chikhi, Limasset, and Medvedev 2016).

Reads were mapped to *S.* Enteritidis reference genome P12510 and variants were called using SNIPPY 4.6 with a minimum coverage of 10x and a mapping quality of 60 (Seemann n.d.). A

reduced alignment of variant sites was generated using snp-sites (Page et al. 2016) and passed to IQ-Tree for phylogenetic reconstruction using a GTR+I+G substitution model and ultra-fast bootstrapping (1000 bootstraps) (Nguyen et al. 2015). An interactive maximum likelihood phylogenetic tree is available in Microreact at https://microreact.org/project/kQEhcTy4ohcqN9bjcPUWLw-ukhsasenteritidishml.

Isolate Selection for Machine Learning

Of the initial 10,223 isolates, 3,434 had matched recent reported travel data, in the form of 'country' of recent travel (within 28 days of generating symptoms), provided by the clinical laboratories on submission of the isolate. These isolates were selected as potential candidates for ML model construction and testing. The 3434 initial isolates were subsequently filtered to identify samples that had both consistent metadata and good sequence quality. The criteria for inclusion included: a) clear and uncontradictory recent travel metadata, unrecognised or nonexistent locations (e.g. Yugoslavia) were removed; b) reads that were not genetically distant from the majority of isolates in the collection as measured by MASH distance; c) reads that had >28x coverage of *S.* Enteritidis reference genome P12510; d) samples that did not have a total unitig length greater than 5,250,000 bp; e) k-mers created from the reads that did not have a singleton frequency of >0.5. A total of 220 samples were excluded due to these criteria.

Further dataset reduction was performed to remove genetically identical, or near-identical, isolates from the collection prior to processing for ML. SNP Address was used as a proxy for genetic relatedness and a single example of each SNP Address (SNP5) was randomly selected per country for further analysis. Samples were assigned to region and subregion based upon the UN M49 Standard for regional codes (https://unstats.un.org/unsd/methodology/m49/overview) [Accessed: 30 Nov 2021]. Recent travel information which represented historically distinct subregions or autonomous subregions of a larger country grouping (e.g. Hong Kong), were considered a part of the larger entity for the purpose of classification. Any country with less than ten representative samples after filtering were excluded from further analysis. The final sample collection after filtering contained 2313 samples (**Table S3**).

The relative risk of acquiring *S.* Enteritidis infection when travelling was estimated using the ratio of the proportion of UKHSA clinical isolates reported as having recently travelled to each country over the proportion of all UK travellers travelling to that country as recorded by the Office of National Statistics (Office of National Statistics 2020). The data used for this comparison was date (year) and location (country) matched before proportions were calculated.

The NCBI Pathogen Genome database was interrogated to identify *S.* Enteritidis isolates collected in countries present in the hML model over a matching time period [Accessed: 18 Nov 2021]. 2019 was excluded as UKHSA samples were only available for a relatively short duration of the year (Jan-April). Variation in isolate number collected each year was controlled for by resampling with replacement (1000 samples). The resulting yearly sampling data was compared to a uniform distribution using the Kolmogorov-Smirnov D statistic to quantify deviation from a consistent sampling scheme for both the UKHSA and NBCI reference dataset.

Unitig Processing for ML

A coloured De Bruijn graph was constructed for the complete genome collection by passing the unitigs for individual samples to unitig caller v1.2.0 (https://github.com/johnlees/unitig-caller). This identified 426,647 unique unitigs present in the 2313 genome collection. The resulting De Brujin graph was then queried to establish the presence/absence of each unitig on a per sample basis. Unitigs present/absent in only a single sample were removed. Unitigs which co-occurred in the same subset of samples (i.e. perfect correlation) were clustered into a single feature 'pattern' for input into feature selection and ML model building algorithms. This reduced the input from 426,647 unitig features into 94,865 pattern features. The dataset was split into 75%-25% train-test ratio stratified by country for downstream applications.

Hierarchical Classifier Design

A class-prediction top-down hierarchical classifier framework was developed to be compatible with any scikit-learn classifier which generates a per-sample predicted probability value. The framework followed a Local Classifier per Node (LCN) approach as detailed in Silla and Freitas 2011 (Silla and Freitas 2011), wherein a multi-class classifier was fitted at each node of the hierarchy to differentiate between the various child classes of that node. The geographical hierarchy used as the basis of the classifier was constructed using the three levels of geographical labels, region->subregion->country, assigned by UN M49 Standard for regional codes. At each node samples were relabelled according to the current hierarchical level (e.g. Cuba and the United States would be labelled as Americas at the root/regional node) before being passed to the classifier. Only samples in classes relevant to the current hierarchical node classifier were included in the training data for each classifier (i.e. samples from countries which were not a part of the region/subregion being trained were excluded).

Hierarchical Classification Strategy

The hierarchical classifier framework allowed for flexible assignment of samples to a single unambiguous class in the hierarchy (i.e. a single region/sub-region/country or unclassified). Samples were first classified at the root (regional) node. A classification was assigned if the predicted probabilities adhered to the threshold criteria, that the maximum probability value at that node exceeds a threshold value (0.5). Only then would the sample be passed to the next node in the hierarchy. This is not permissive of multiple, conflicting classifications.

A range of hierarchical and non-hierarchical statistics were applied to aid model assessment. Standard non-hierarchical statistics were generated for each individual classifier/node (precision, recall, micro/macro/weighted F1) as well as their hierarchical analogues (hP, hR, hF1). These were calculated as described by Kiritchenko et al 2005 (Kiritchenko, Matwin, and Famili 2005):

$$hP = \frac{\Sigma i \left| \hat{T}\, i \right|}{\Sigma i \left| \hat{P}\, i \right|} \qquad hR = \frac{\Sigma i \left| \hat{T}\, i \right|}{\Sigma i \left| \hat{T}\, i \right|} \qquad hF1 = \frac{2 * hP * hR}{hP + hR}$$

where $\hat{P}\, i$ is the set of predicted classes consisting of the most specific class (i.e. the lowest level of the hierarchy) predicted for test example *i* plus all parent classes and $\hat{T}\, i$ is the set of classes consisting of the most specific true class of test example *i* and all its ancestor classes. Summations were calculated over all test samples. Macro hF1 was calculated by taking the average F1 of all classes of interest.

Feature Selection, Resampling, Model Testing and Optimisation

During model selection various combinations of classifier, resampler and feature selection models were applied to the test-train dataset to assess their suitability for model building. These were assessed based on a combination of non-hierarchical statistics including overall micro/macro F1 and micro/macro F1 per hierarchical level. The implemented classifier models included K-Nearest Neighbours, Support Vector, Random Forest, Gaussian Naive Bayes, XGBoost, Extra Trees. All classifier models were implemented using scikit-learn using a set seed value and default parameters with the exception of Random Forest, Extra Trees and XGBoost classifiers which were run with n_estimators = 1000 and SVC which was run with probability = True.

The implemented resampling schemes included downsampling (smallest class), upsampling (largest class), resampling to the mean count of all classes and hierarchically aware implementation for the previously described samplers. Hierarchically aware resampling was developed using in-house scripts to iteratively apply a resampler to each of the lowest levels of the hierarchy (country) before passing the resampled data to higher levels in the hierarchy for further resampling.

An all-vs-all comparison of classifier vs resampler models was used to identify the most suitable combinations for further optimisation (**Table S2**). In all cases a fixed seed value was used for comparison of models and resamplers. The top 4 combinations of classifier-resampler were selected for feature selection comparison. The implemented feature selection method was Random Forest using varying numbers of patterns as training data (**Figure 2D**).

The final classifier-resampler-selection combination was passed to a genetic algorithm framework (TPOT) to identify an approximation of optimal parameters from a wide range of possible combinations (sklearn.ensemble.RandomForestClassifier': 'n_estimators': [100, 500, 1000], 'criterion': ['gini', 'entropy'], 'max_features': np.array([0.05, 0.1 , 0.15, 0.2 , 0.25, 0.3 , 0.35, 0.4 , 0.45, 0.5 , 0.55, 0.6 , 0.65, 0.7 , 0.75, 0.8 , 0.85, 0.9 , 0.95, 1.]), 'min_samples_split': range(2, 21), 'min_samples_leaf': range(1, 21), 'bootstrap': [True, False]) (Olson et al. 2016). The TPOT genetic algorithm used the macro F1 score per node as the optimisation metric, was run for 100 generations with a population size of 50 and stratified 3-fold cross validation of the input database and was stopped if no model improvement was found for 10 generations. A ‘flat’ model was also trained and tested in the same manner for comparison, whereby a multiclass Random Forest classifier was provided with a randomly oversampled dataset which only included ‘country’ class labels (i.e. region/subregion were ignored) (**Table S1**).

Validation Dataset Collection and Processing

Various public datasets were used as additional validation data including outbreaks described in previous publications (Inns et al. 2017; Pijnacker et al. 2019) and samples from the 38 countries included in this study identified from the NCBI Pathogen Genome database [Accessed: 18 Nov 2021]. In the case of samples taken from previous publications, accession numbers were identified from these manuscripts, samples downloaded and passed through the genome and unitig processing pipelines described above. Additionally, NCBI Pathogen Genome database metadata was downloaded [Accessed: 18 Nov 2021] and filtered to return only samples which had publicly accessible read files, country metadata from the 38 countries and were collected between 2014-2018. Three representative countries were chosen to trial the model on (Poland,

South Africa and Singapore). Read data was downloaded and passed through the unitig processing pipeline as described above. The presence of unitigs generated from the UKHSA collection which formed the basis to the hML model was ascertained using unitig-counter (https://github.com/johnlees/unitig-counter) (Jaillard et al. 2018). The unitig features were then converted into the patterns generated from the UKHSA collection as described above.

Data Availability

The final optimised hierarchical model as well as a pipeline for pre-processing raw read data to unitigs/patterns for input is available from https://github.com/SionBayliss/HierarchicalML with a short description and tutorial for ease of use. This end-to-end process, from FASTQ to prediction, is open access and available to users. Short read sequencing data is available from the Short Read Archive (Bioproject: PRJNA248792).

Acknowledgements

We would like to acknowledge both Dr Harry Thorpe and Dr Nicola Coyle who have both previously contributed to the development of scripts which underlie the unitig processing pipeline. This work was funded by an Academy of Medical Sciences Springboard grant (SBF005\1089). CJ, TD and MAC are affiliated to the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Gastrointestinal Infections and Genomics and Enabling Data at University of Liverpool and University of Warwick respectively in partnership with the UK Health Security Agency (UKHSA). CJ and MAC are based at UKHSA. The views expressed are those of the author(s) and not necessarily those of the NIHR, the Department of Health and Social Care or the UK Health Security Agency.

References

Allard, M. W., Bell, R., Ferreira, C. M., Gonzalez-Escalona, N., Hoffmann, M., Muruvanda, T., Ottesen, A., Ramachandran, P., Reed, E., Sharma, S., Stevens, E., Timme, R., Zheng, J., & Brown, E. W. (2018). Genomics of foodborne pathogens for microbial food safety. *Current Opinion in Biotechnology*, *49*, 224–229.

Argimón, S., Abudahab, K., Goater, R. J. E., Fedosejev, A., Bhai, J., Glasner, C., Feil, E. J., Holden, M. T. G., Yeats, C. A., Grundmann, H., Spratt, B. G., & Aanensen, D. M. (2016). Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microbial Genomics*, *2*(11), e000093.

Ashton, P. M., Nair, S., Peters, T. M., Bale, J. A., Powell, D. G., Painset, A., Tewolde, R., Schaefer, U., Jenkins, C., Dallman, T. J., de Pinna, E. M., Grant, K. A., & Salmonella Whole

589 Genome Sequencing Implementation Group. (2016). Identification of *Salmonella* for public
590 health surveillance using whole genome sequencing. *PeerJ*, *4*, e1752.

591 Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina
592 sequence data. *Bioinformatics*, *30*(15), 2114–2120.

593 Brown, E., Dessai, U., McGarry, S., & Gerner-Smidt, P. (2019). Use of Whole-Genome
594 Sequencing for Food Safety and Public Health in the United States. *Foodborne Pathogens and*
595 *Disease*, *16*(7), 441–450.

596 Chattaway, M. A., Dallman, T. J., Larkin, L., Nair, S., McCormick, J., Mikhail, A., Hartman, H.,
597 Godbole, G., Powell, D., Day, M., Smith, R., & Grant, K. (2019). The Transformation of
598 Reference Microbiology Methods and Surveillance for *Salmonella* With the Use of Whole
599 Genome Sequencing in England and Wales. *Frontiers in Public Health*, *7*, 317.

600 Chikhi, R., Limasset, A., & Medvedev, P. (2016). Compacting de Bruijn graphs from sequencing
601 data quickly and in low memory. *Bioinformatics* , *32*(12), 201–208.

602 Cowley, L. A., Dallman, T. J., Fitzgerald, S., Irvine, N., Rooney, P. J., McAteer, S. P., Day, M.,
603 Perry, N. T., Bono, J. L., Jenkins, C., & Gally, D. L. (2016). Short-term evolution of Shiga toxin-
604 producing *Escherichia coli* O157:H7 between two food-borne outbreaks. *Microbial Genomics*,
605 *2*(9), e000084.

606 Dallman, T., Inns, T., Jombart, T., Ashton, P., Loman, N., Chatt, C., Messelhaeusser, U., Rabsch,
607 W., Simon, S., Nikisins, S., Bernard, H., le Hello, S., Jourdan da-Silva, N., Kornschober, C.,
608 Mossong, J., Hawkey, P., de Pinna, E., Grant, K., & Cleary, P. (2016). Phylogenetic structure of
609 European *Salmonella* Enteritidis outbreak correlates with national and international egg
610 distribution network. *Microbial Genomics*, *2*(8), e000070.

611 Daniel, N., Casadevall, N., Sun, P., Sugden, D., & Aldin, V. (2020). *The Burden of Foodborne*
612 *Disease in the UK 2018*. Food Standards Agency.

613 Donker, T., Reuter, S., Scriberras, J., Reynolds, R., Brown, N. M., Török, M. E., James, R.,
614 Network, E. O. E. M. R., Aanensen, D. M., Bentley, S. D., Holden, M. T. G., Parkhill, J., Spratt,
615 B. G., Peacock, S. J., Feil, E. J., & Grundmann, H. (2017). Population genetic structuring of
616 methicillin-resistant *Staphylococcus aureus* clone EMRSA-15 within UK reflects patient referral
617 patterns. *Microbial Genomics*, *3*(7), e000113.

618 Ebel, E. D., Williams, M. S., Cole, D., Travis, C. C., Klontz, K. C., Golden, N. J., & Hoekstra, R.
619 M. (2016). Comparing Characteristics of Sporadic and Outbreak-Associated Foodborne
620 Illnesses, United States, 2004-2011. *Emerging Infectious Diseases*, *22*(7), 1193–1200.

621 Feil, E. J., & Spratt, B. G. (2001). Recombination and the Population Structures of Bacterial
622 Pathogens. *Annual Reviews in Microbiology, 55,* 561-90.

623 Gould, L. H., Kline, J., Monahan, C., & Vierk, K. (2017). Outbreaks of Disease Associated with
624 Food Imported into the United States, 1996-20141. *Emerging Infectious Diseases*, *23*(3), 525–
625 528.

626 Ingle, D. J., Levine, M. M., Kotloff, K. L., Holt, K. E., & Robins-Browne, R. M. (2018). Dynamics
627 of antimicrobial resistance in intestinal Escherichia coli from children in community settings in
628 South Asia and sub-Saharan Africa. *Nature Microbiology*, *3*(9), 1063–1073.

629 Inns, T., Ashton, P. M., Herrera-Leon, S., Lighthill, J., Foulkes, S., Jombart, T., Rehman, Y., Fox,
630 A., Dallman, T., DE Pinna, E., Browning, L., Coia, J. E., Edeghere, O., & Vivancos, R. (2017).
631 Prospective use of whole genome sequencing (WGS) detected a multi-country outbreak of
632 *Salmonella* Enteritidis. *Epidemiology and Infection*, *145*(2), 289–298.

Jaillard, M., Lima, L., Tournoud, M., Mahé, P., van Belkum, A., Lacroix, V., & Jacob, L. (2018). A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS Genetics*, *14*(11), e1007758.

Kiritchenko, S., Matwin, S., & Famili, F. (2005, January 1). Functional Annotation of Genes Using Hierarchical Text Categorization. *BioLINK SIG: Linking Literature, Information and Knowledge for Biology, a Joint Meeting of The ISMB BioLINK Special Interest Group on Text Data Mining and The ACL Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*.

Li, S., He, Y., Mann, D. A., & Deng, X. (2021). Global spread of *Salmonella* Enteritidis via centralized sourcing and international trade of poultry breeding stocks. *Nature Communications*, *12*(1), 5109.

Lupolova, N., Dallman, T. J., Holden, N. J., & Gally, D. L. (2017). Patchy promiscuity: machine learning applied to predict the host specificity of Salmonella enterica and Escherichia coli. *Microbial Genomics*, *3*(10), e000135.

Lupolova, N., Lycett, S. J., & Gally, D. L. (2019). A guide to machine learning for bacterial host attribution using genome sequence data. *Microbial Genomics*, *5*(12).

McLauchlin, J., Aird, H., Andrews, N., Chattaway, M., de Pinna, E., Elviss, N., Jørgensen, F., Larkin, L., & Willis, C. (2019). Public health risks associated with *Salmonella* contamination of imported edible betel leaves: Analysis of results from England, 2011-2017. *International Journal of Food Microbiology*, *298*, 1–10.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274.

Office of National Statistics. (2020). *Travel trends estimates: UK residents' visits abroad: 2009-2019*.
https://www.ons.gov.uk/peoplepopulationandcommunity/leisureandtourism/datasets/ukresidentsvisitsabroad

Olson, R. S., Bartley, N., Urbanowicz, R. J., & Moore, J. H. (2016). Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science. *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, 485–492.

Page, A. J., Taylor, B., Delaney, A. J., Soares, J., Seemann, T., Keane, J. A., & Harris, S. R. (2016). SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial Genomics*, *2*(4), e000056.

PHE. (2017). *Travel-associated non typhoidal* Salmonella *infection in England, Wales and Northern Ireland: 2014*. PHE.

Pijnacker, R., Dallman, T. J., Tijsma, A. S. L., Hawkins, G., Larkin, L., Kotila, S. M., Amore, G., Amato, E., Suzuki, P. M., Denayer, S., Klamer, S., Pászti, J., McCormick, J., Hartman, H., Hughes, G. J., Brandal, L. C. T., Brown, D., Mossong, J., Jernberg, C., … International Outbreak Investigation Team. (2019). An international outbreak of *Salmonella enterica* serotype Enteritidis linked to eggs from Poland: a microbiological and epidemiological study. *The Lancet Infectious Diseases*, *19*(7), 778–786.

Pires, S. M., Evers, E. G., van Pelt, W., Ayers, T., Scallan, E., Angulo, F. J., Havelaar, A., Hald, T., & Med-Vet-Net Workpackage 28 Working Group. (2009). Attributing the human disease burden of foodborne infections to specific sources. *Foodborne Pathogens and Disease*, *6*(4), 417–424.

678 Seemann, T. (n.d.). *snippy*. Github. Retrieved November 15, 2018, from
679 https://github.com/tseemann/snippy

680 Silla, C. N., & Freitas, A. A. (2011). A survey of hierarchical classification across different
681 application domains. *Data Mining and Knowledge Discovery*, *22*(1), 31–72.

682 Somorin, Y. M., Odeyemi, O. A., & Ateba, C. N. (2021). Salmonella is the most common
683 foodborne pathogen in African food exports to the European Union: Analysis of the Rapid Alert
684 System for Food and Feed (1999–2019). *Food Control*, *123*, 107849.

685 Statistics Division of the United Nations Secretariat. (2020). *Standard country or area codes for*
686 *statistical use (M49)*. United Nations. https://unstats.un.org/unsd/methodology/m49/

687 Surveillance, Zoonoses, Epidemiology and Risk Food and Farming Group. (2007). *UK National*
688 *Control Programme for Salmonella in Layers (gallus gallus)*. DEFRA.

689 Tam, C. C., Rodrigues, L. C., Viviani, L., Dodds, J. P., Evans, M. R., Hunter, P. R., Gray, J. J.,
690 Letley, L. H., Rait, G., Tompkins, D. S., O'Brien, S. J., & IID2 Study Executive Committee.
691 (2012). Longitudinal study of infectious intestinal disease in the UK (IID2 study): incidence in the
692 community and presenting to general practice. *Gut*, *61*(1), 69–77.

693 Thomson, N. R., Clayton, D. J., Windhorst, D., Vernikos, G., Davidson, S., Churcher, C., Quail,
694 M. A., Stevens, M., Jones, M. A., Watson, M., Barron, A., Layton, A., Pickard, D., Kingsley, R. A.,
695 Bignell, A., Clark, L., Harris, B., Ormond, D., Abdellah, Z., … Parkhill, J. (2008). Comparative
696 genome analysis of *Salmonella* Enteritidis PT4 and *Salmonella* Gallinarum 287/91 provides
697 insights into evolutionary and host adaptation pathways. *Genome Research*, *18*(10), 1624–
698 1637.

699 UKHSA. (2021). *Non-typhoidal* Salmonella *data 2010 to 2019*.
700 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/fi
701 le/1026208/salmonella-annual-report-2019.pdf

702 WHO. (2022). *Factsheet: Non-typhoidal* Salmonella. https://www.who.int/news-room/fact-
703 sheets/detail/salmonella-(non-typhoidal)

704 Zhang, S., Li, S., Gu, W., den Bakker, H., Boxrud, D., Taylor, A., Roe, C., Driebe, E.,
705 Engelthaler, D. M., Allard, M., Brown, E., McDermott, P., Zhao, S., Bruce, B. B., Trees, E.,
706 Fields, P. I., & Deng, X. (2019). Zoonotic source attribution of *Salmonella* enterica serotype
707 Typhimurium using genomic surveillance data, United States. *Emerging Infectious Diseases*,
708 *25*(1), 82–91.