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

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

[1]Milner Centre for Evolution, Department of Biology & Biochemistry, University of Bath, UK
[2]Genomic Laboratory Hub (GLH), Addenbrooke's Hospital, Cambridge University Hospitals NHS Foundation Trust, UK
[3]Gastrointestinal Reference Services, UK Health Security Agency, Colindale, UK
[4]Institute for Risk Assessment Sciences (IRAS), Utrecht University, 3508 TD Utrecht, Netherlands

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. Rapid geographical source attribution of suspect food vehicles facilitates outbreak management. In this study, 2,313 *S.* Enteritidis genomes collected by the UKHSA between 2014-2019 were used to train a hierarchical machine learning classifier to predict geographical origin of isolates for 38 countries. 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). Longitudinal analysis and validation with publicly accessible international samples indicated that predictions were robust to prospective external datasets. This 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.

1 UKHSA have the first comprehensive genomic surveillance dataset to consistently sample phylogeographic population structure of *S.* Enteritidis

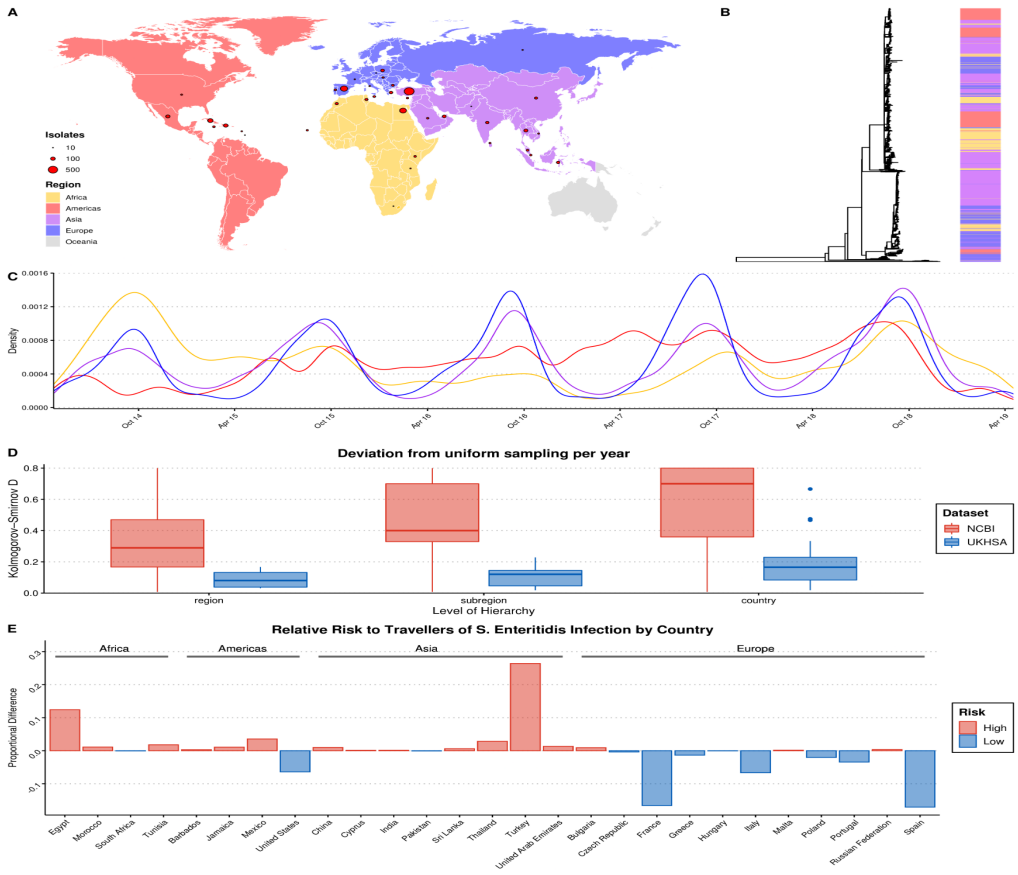


Figure 1. Summary of *S.* Enteritidis isolates collected by the UKHSA from UK clinical patients who recently reported foreign travel between 2014-2019. A) Geographical distribution of 2313 *S.* Enteritis isolates by reported foreign travel. Points are variably sized to represent the number of samples 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) (proportion of UKHSA/proportion of ONS). Only ONS data from a matched set of destinations was used to calculate proportions.

Broad and unbiased surveillance of all reported *S.* Enteritidis cases in the UK between 2014-April 2019 coupled with returning traveller data has provided a large genomic dataset representative of infections acquired by UK tourists during global travel. The total surveillance dataset (2014-April 2019) 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.

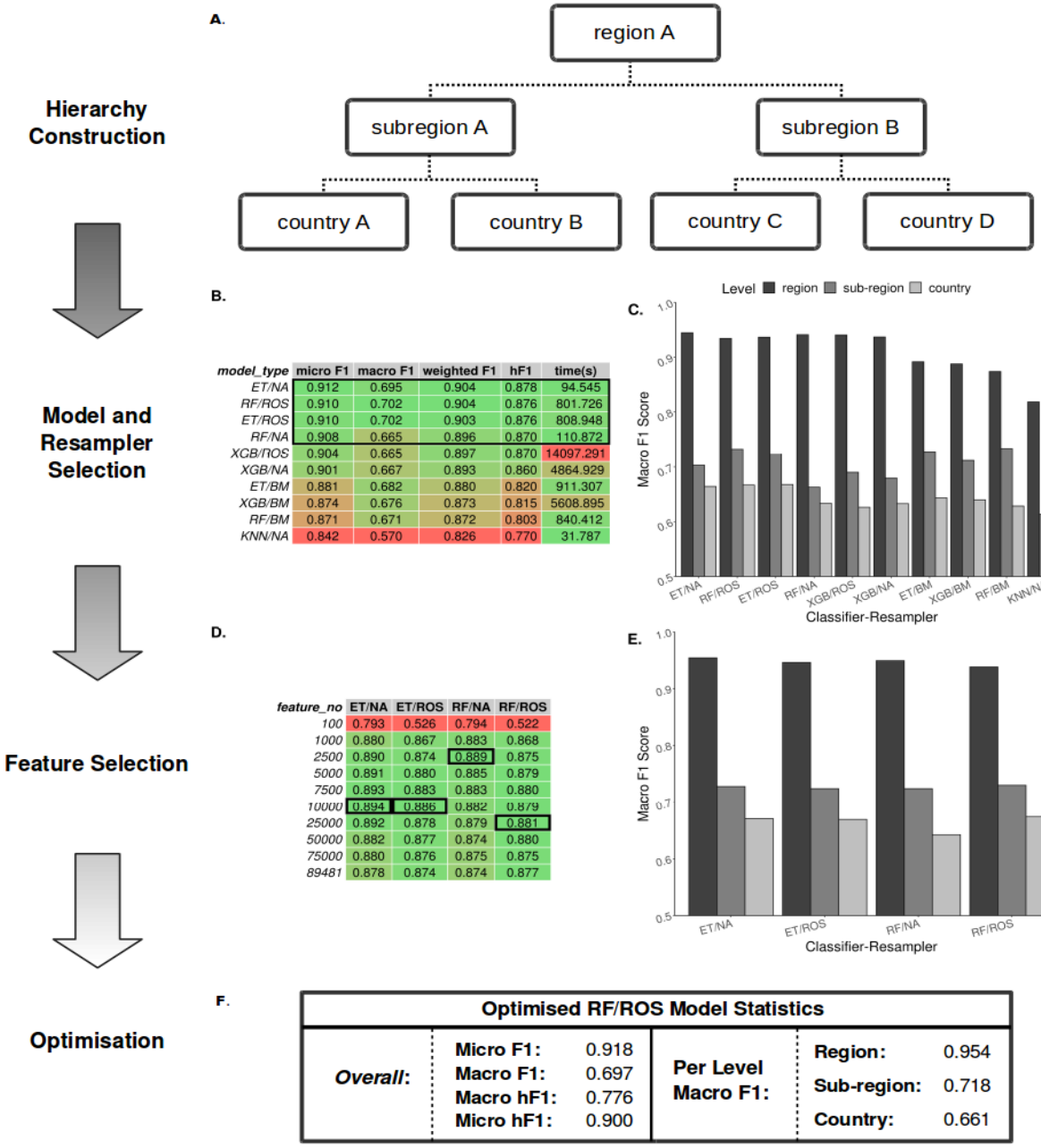
Recent travel was reported from 122 countries and 5 continents. A source of potential bias, common to bacterial genomics analysis, is the overrepresentation of clonally related isolates due to increased prevalence during outbreaks (Feil and Spratt 2003; Ebel et al. 2016). In order to reduce the influence of highly related clonal outbreaks on the ML model a single, random representative isolate per country was selected for each clone, defined as a 5 SNP cluster identified by SNP Address (Dallman et al. 2018). This definition is routinely used by the UKHSA for genomic disease surveillance (Chattaway et al. 2019). After quality filtering, down sampling of outbreaks and removal of low incidence countries represented by less than 10 isolates, 2,313 genomes from 38 countries and 4 continents were included in the final input dataset for machine learning (**Figure 1A**).

Grouping these countries by geographic region and subregion using the UN M49 Standard for regional codes provided a framework for more granular analysis and a simplistic hierarchical framework (i.e. countries were contained in subregions which were contained in regions/continents)(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). As expected, *S.* Enteritidis infection rates were highly seasonal, with significantly increased infection during the summer months to Europe and Asia, but a less pronounced seasonal impact on travel to/from the Americas and Africa (**Figure 1C**).

An analysis of the consistency of sampling per country per year indicated 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. However, contrasting the consistency of the UKHSA dataset with location and date matched publicly available *S.* Enteritidis genomes from the NCBI indicated that the sampling consistency of the UKHSA dataset is significantly less influenced by sporadic outbreaks than the public dataset and represented a more consistent sample (**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**). If we control for the volume of UK-travel to that destination during a matched time period it becomes evident that Turkey and Egypt had a disproportionate risk of *S.* Enteritidis infection (**Figure 1E**)(Office of National Statistics 2020). Conversely, France and Spain, two of the most popular travel destinations for UK travellers, had a low risk of infection. Consequently, there was a large degree of class imbalance (different number of isolates per country) in the resulting dataset. When considering larger geographical groupings, such as region/continent and subregion, the imbalance is less pronounced.

2 A novel hierarchical model provides real-time geographic source attribution prediction directly from sequencing reads within four minutes



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. 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). 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).

Taking advantage of the clear hierarchical structure in geographical data, we designed a multi-level hierarchical machine learning (hML) classification model following the "Local Classifier per Node" framework. This is made up of 15 individual classifiers, one per node (1 root, 4 regional, 11 sub-regional). In total, 53 individual classes (4 regions, 11 sub-regions and 38 countries) are predictable by the model. Sample classification progresses through a series of classifiers starting from the root, which attributes 'region', followed by 'subregion' and finally 'country' classification (**Figure 2A**). Classification is performed in a top-down approach, where samples are classified first in the region node and, if the predicted probability is greater than a minimum threshold value (>0.5), the sample is then passed deeper into the hierarchy (i.e. a sample cannot be attributed to a sub-region which is not part of a previously predicted region). Sample classification is exclusive, disallowing multiple classifications on the same hierarchical level for a single sample.

To streamline results for users without bioinformatics skills, the model was trained on raw genomic short-read data files (FASTQ), the most commonly adopted datatype in public health surveillance, to provide an end-to-end sample classification directly from a sequencing machine. Raw reads were quality filtered and trimmed, before being converted into Unitigs (presence/absence) as the input datatype for machine learning**.** Each local classifier per node was trained using only pertinent data for that local node (e.g. a local subregion classifier was trained only on the data from countries which comprised it). This end-to-end process, from FASTQ to prediction, is open access 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). Based on these comparisons, the most desirable assessment metrics overall (i.e. high macro F1 at the country level, **Figure 2E**) were found to be produced by passing the top 25,000 patterns from a Random Forest (RF) classifier (feature selection) to a subsequent RF classifier and a random oversampler applied to the class imbalance (RF/ROS). Individual sample classification takes approximately 3.5 mins (depending on file size) from raw read data, through unitig processing, pattern matching, the model and finally to classification output.

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

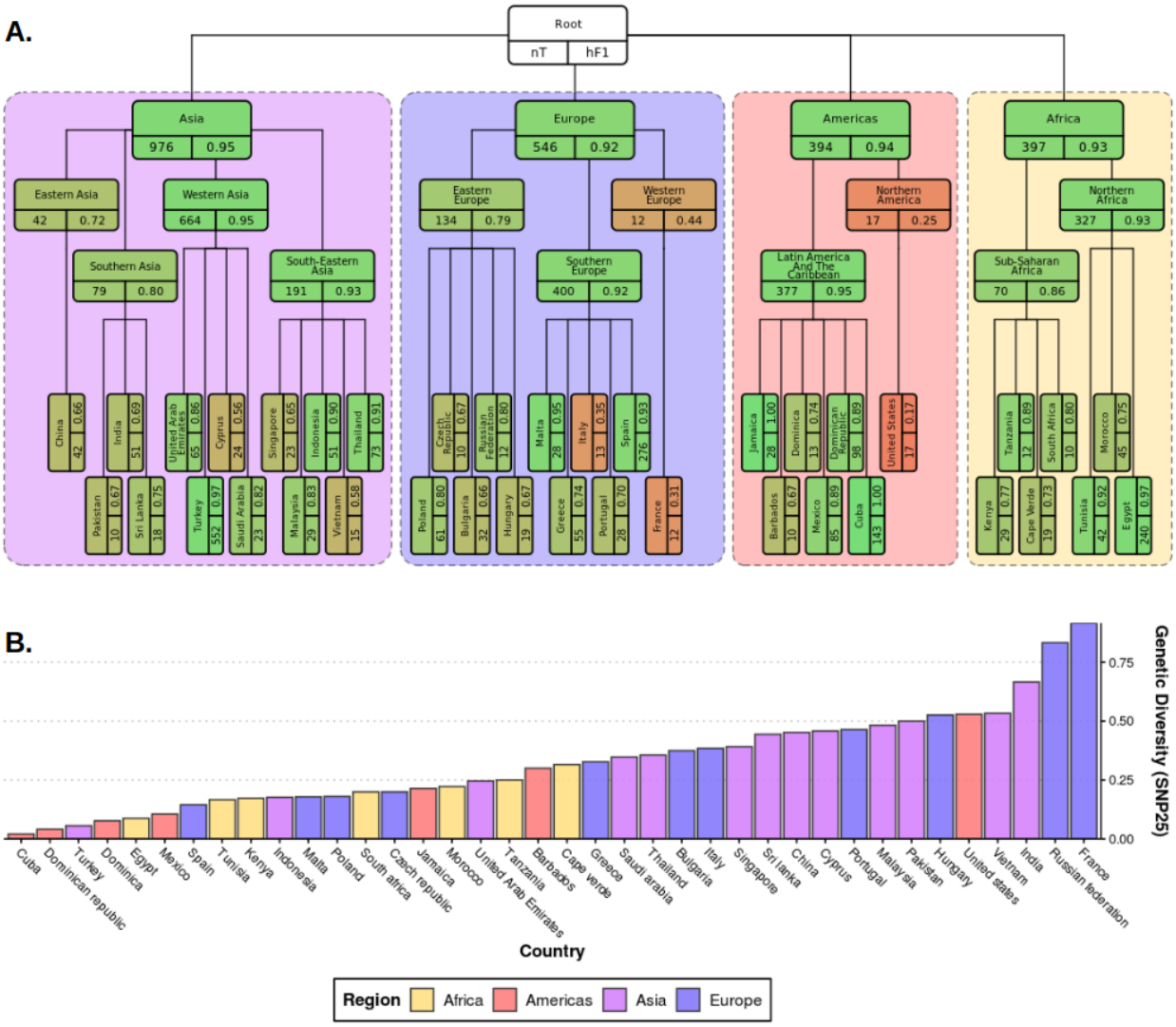


Figure 3. Plots summarising test results from hML model and genetic diversity of dataset. A) Diagrammatic representation of classification metrics of the hML classifier applied to the test dataset. Links between classes/nodes in the hierarchy are indicated by connecting lines. Boxes represent an individual class in the 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).

Classification metrics were highest at the regional level (macro F1: 0.954), less discriminatory at the sub-regional level (macroF1: 0.718), and finally the country level (macroF1: 0.661). Further scrutiny of poorer perfoming countries clearly showed that much less training data was available which resulted in lower prediction accuracy (**Figure 3A**). For example, all African classes showed very high classification metrics (hF1: >0.7), whereas Europe had two classes, France and Italy, which were classified less well (hF1: ~0.3). Similarly, many Latin American and Caribbean countries showed very high classification metrics (hF1: >0.8), whereas samples for the United States were consistently misclassified. There was a clear correlation between number of available samples and classification accuracy. However, some countries with a low number of samples (e.g. Czech Republic, Pakistan)

showed moderate classification accuracy whilst having similar numbers of samples to the most poorly predicted classes (Italy, France, United States). An investigation of genetic diversity indicated that at least two of the most poorly classified classes, France and the United States, had both low sample numbers and very high genetic diversity (**Figure 3B**). The French isolates were particularly diverse, arising from multiple highly diverse clades.

Overall, ~95% of tested samples were predicted into the correct region (macro F1 0.954) and 33/38 country classes predicted correctly >60% of the time (macro F1 0.661, **Figure 3A**). Although the model performed very differently at the country level (hF1 range 0.17-1.00), it consistently predicted region with a very high degree of accuracy. This suggests that the model can be broadly applied to classifying samples with high confidence at the regional level whilst also successfully predicting a range of countries regularly visited by UK travellers with greater than 90% accuracy (e.g. Cuba, Egypt, Indonesia, Jamaica, Malta, Spain, Thailand, Tunisia and Turkey).

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

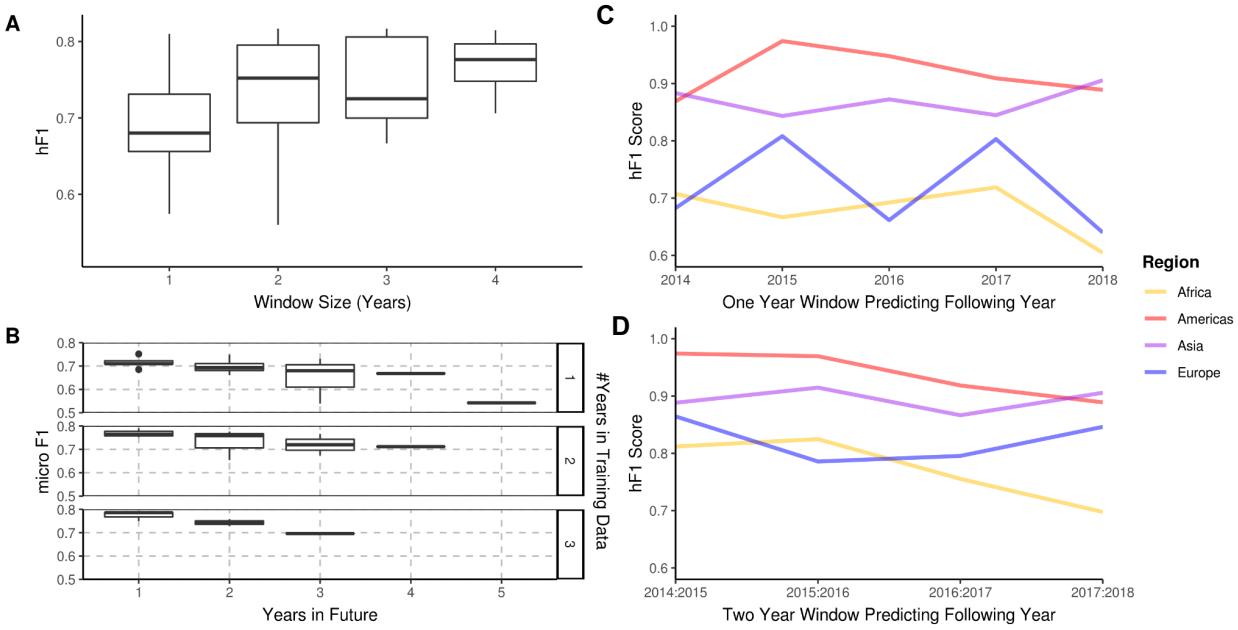


Figure 4. Plots summarising longitudinal analysis of the predictive accuracy of hierarchical models on 2313 *S.* Enteritidis samples. **A)** hF1 scores for hierarchical models trained on data using 1-5 year training window sizes predicting the following year. **B)** micro F1 scores for hierarchical models trained on data using 1-3 year training window sizes predicting 1-5 years into the future. **C)** hF1 scores of hierarchical models trained on one-year sample windows predicting the following year stratified by region. **D)** hF1 scores of hierarchical models trained on two-year sample windows predicting the following year stratified by region.

Bacterial population lineage composition is not expected to remain static through time, therefore predictive models based on genomic data will require periodic retraining. To understand how much data is required for accurate prospective prediction, we compared the outcomes of four yearly window sizes (1, 2, 3 and 4 years) to predict the subsequent years (**Figure 4A**). Predictive accuracy and consistency of prediction of the subsequent year improved on increasing window size, with the largest improvement observed between one and two years' worth of data. A minor decrease in predictive accuracy of ~0.5 micro F1 was observed for each additional year into the future (**Figure 4B**). A breakdown of the hF1 per region indicates that a one-year window varied in predictive accuracy per class per year (**Figure 4C**) but that a two-year model was more consistent and has good predictive accuracy per class per window (**Figure 4D**). An optimal window of data collection for machine learning, balancing predictive

accuracy against the cost of surveillance, would require model retraining each year on the data from the previous two years.

5 External previously unseen datasets validate the model's accuracy in four out of five tests

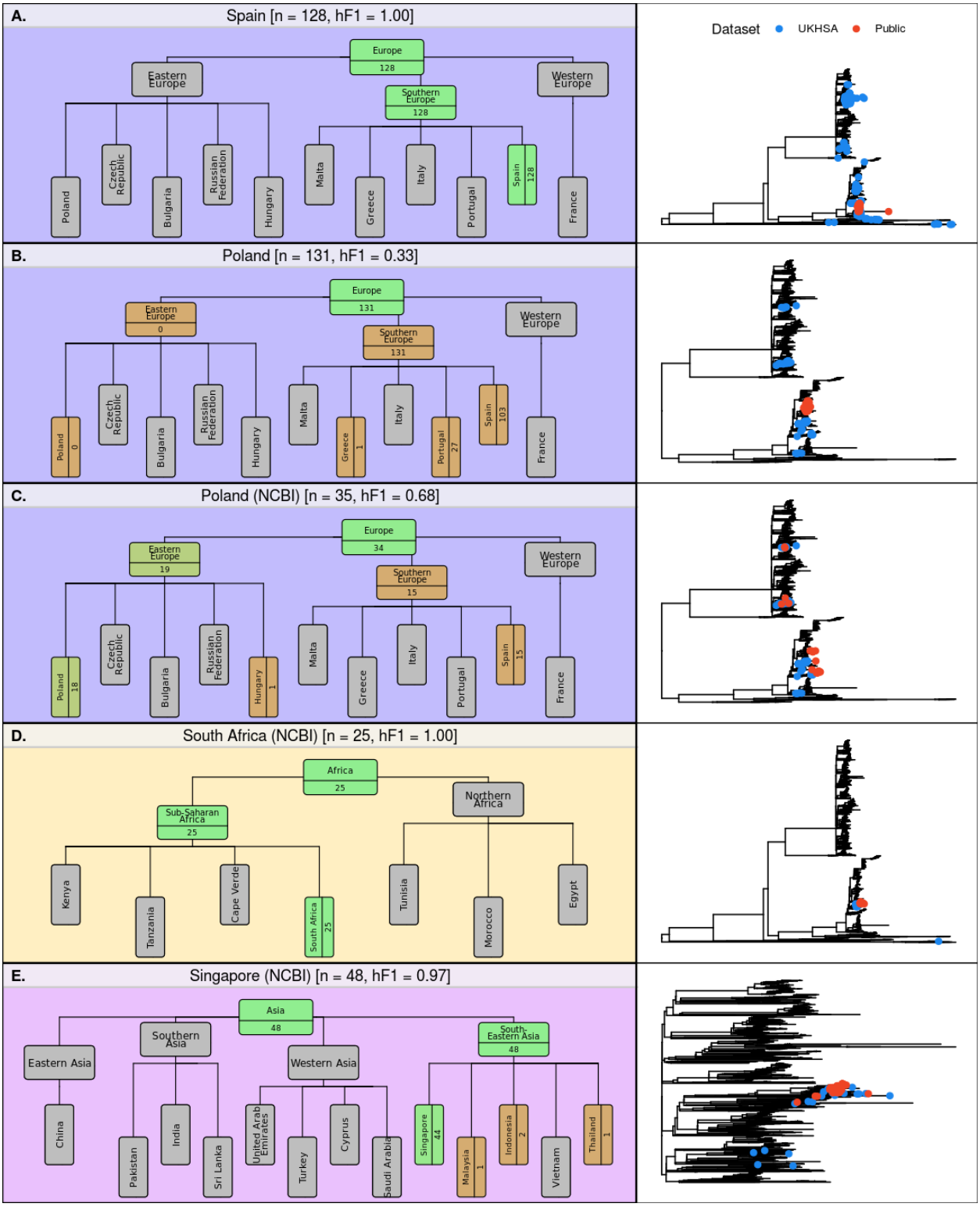


Figure 5. Hierarchical classification summaries for five additional validation datasets. **A)** 128 samples from an international outbreak originating in Spain in 2015 (Inns et al. 2017) **B)** 131 samples from a large-scale international outbreak originating from Polish eggs between 2015-2018 (Pijnacker et al. 2019). **C)** 35 samples uploaded by Poland to the NCBI database between 2014-2019. **D)** 25 samples uploaded by South Africa to the NCBI database between 2014-2019. **E)** 48 samples uploaded by Singapore to the NCBI database between 2014-2019. The number of samples assigned per relevant class is indicated. 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.

We tested whether the model was able to accurately attribute the geographical origins of two epidemiologically traced and well-characterised UK-imported food outbreaks. The first outbreak, identified as arising from eggs imported from Spain (Inns et al. 2017), was 100% successful at attributing all 128 UK cases to a Spanish origin (**Figure 5A**). The second, epidemiologically-traced to a multi country outbreak of eggs originating from Poland (Pijnacker et al. 2019), comprising two distinct lineages each differing by 5 or fewer SNPs, was primarily attributed to a European origin (131/131 cases), subsequently misattributed to Southern Europe (131/131 cases) and majority misattributed to Spain (103/130) (**Figure 5B**). This complex outbreak was particularly difficult for the model to attribute as it had been continuously causing cases in 16 European countries for several years (2015-2018), confusing the signal. Outbreak cases were also phylogenetically distinct from those associated with travel to Poland in the UKHSA dataset (**Figure 5B**). Although, the model works well for more singular source attribution at a single country source, outbreaks that are long standing and associated with multiple countries are likely to have a confused signal in terms of labels and phylogeography.

The model was further tested on datasets extracted from public databases and uploaded by their country of sampling. Three representative countries from three different regions were identified from the NCBI database as having moderate sample numbers (>20), falling within the timeframe of the current model (2014-2019) and having arisen from a country included in the current model hierarchy (South Africa, Singapore and Poland). The Polish dataset was attributed to a European origin with high accuracy (34/35, 97.1%), 19 of these were subsequently correctly attributed to an East European origin (19/35, 54.3%) or which 18 were correctly classified as 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 was correctly attributed to a South African origin with complete accuracy (25/25, 100%) (**Figure 5D**). The Singaporean dataset was correctly attributed at a South-East Asian level (48/48) with 91.7% of samples being correctly attributed to a Singaporean origin (44/48) and 4 samples being misattributed to Indonesia (2), Malaysia (1) and Thailand (1) (**Figure 5E**).

Conclusion. This hierarchical classifier will inform epidemiologists of transmission history through geographical source attribution. Information about the infection provided quickly through raw sequencing reads gives epidemiologists the opportunity to choose the best intervention immediately, eliminating the costly and slow current methods in which population structure analysis is required. This is the first time that ML models have been developed to automate genomics-based geographical source attribution, greatly enhancing our ability to condense complex genetic data into actionable information for epidemiologists which can be used for rapid answers of where infections have come from within minutes of sequencing.

Argimón, Silvia, Khalil Abudahab, Richard J. E. Goater, Artemij Fedosejev, Jyothish Bhai, Corinna Glasner, Edward J. Feil, et al. 2016. "Microreact: Visualizing and Sharing Data for Genomic Epidemiology and Phylogeography." *Microbial Genomics* 2 (11): e000093.

Chattaway, Marie Anne, Timothy J. Dallman, Lesley Larkin, Satheesh Nair, Jacquelyn McCormick, Amy Mikhail, Hassan Hartman, et al. 2019. "The Transformation of Reference Microbiology Methods and Surveillance for Salmonella With the Use of Whole Genome Sequencing in England and Wales." *Frontiers in Public Health* 7 (November): 317.

Dallman, Tim, Thomas Inns, Thibaut Jombart, Philip Ashton, Nicolas Loman, Carol Chatt, Ute Messelhaeusser, et al. 2016. "Phylogenetic Structure of European Salmonella Enteritidis Outbreak Correlates with National and International Egg Distribution Network." *Microbial Genomics* 2 (8): e000070.

Ebel, Eric D., Michael S. Williams, Dana Cole, Curtis C. Travis, Karl C. Klontz, Neal J. Golden, and Robert M. Hoekstra. 2016. "Comparing Characteristics of Sporadic and Outbreak-Associated Foodborne Illnesses, United States, 2004-2011." *Emerging Infectious Diseases* 22 (7): 1193–1200.

Feil, Edward J., and Brian G. Spratt. 2003. "Recombination and the Population Structures of Bacterial Pathogens," November.

https://doi.org/10.1146/annurev.micro.55.1.561

Inns, T., P. M. Ashton, S. Herrera-Leon, J. Lighthill, S. Foulkes, T. Jombart, Y. Rehman, et al. 2017. "Prospective Use of Whole Genome Sequencing (WGS) Detected a Multi-Country Outbreak of Salmonella Enteritidis." *Epidemiology and Infection* 145 (2): 289–98.

Pijnacker, Roan, Timothy J. Dallman, Aloys S. L. Tijsma, Gillian Hawkins, Lesley Larkin, Saara M. Kotila, Giusi Amore, et al. 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): 778–86.