ECML PKDD 2022

Machine Learning for Microbial Genomics,
ECML PKDD 2022

Machine learning for rapid geographical source attribution of *Salmonella* Enteritidis infections

Dr Sion C. Bayliss

Veterinary and Bacterial Genome Bioinformatics Fellow

University of Bristol

UK Health Security Agency

University of BRISTOL

UNIVERSITY OF BATH

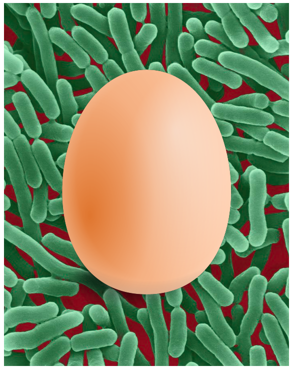
The Academy *of* Medical Sciences

Salmonella enterica subspecies Enteritidis

Salmonella is a bacterial pathogen which causes diarrhea, fever, abdominal cramps, and, in severe cases, hospitalisation.

UK Health Security Agency

The UK has ~8,500 *Salmonella* cases annually, of which ~2,500 are *Salmonella enterica* subspecies Enteritidis.



S. Enteritidis infection is associated with consumption of contaminated foodstuffs, particularly poultry meat and eggs.

National monitoring and vaccination programmes have greatly reduced salmonellosis associated with local food production.

Many clinical *S.* Enteritidis cases identified in the UK are thought to be associated with foreign travel or consumption of imported foodstuffs.

Rapid geographical source attribution of an infecting strain will allow targeted epidemiological follow-up and rapid outbreak resolution.

Source attribution to improve outbreak response

Clinical *S.* Enteritidis Genomes

+

‘Recent Reported Travel’

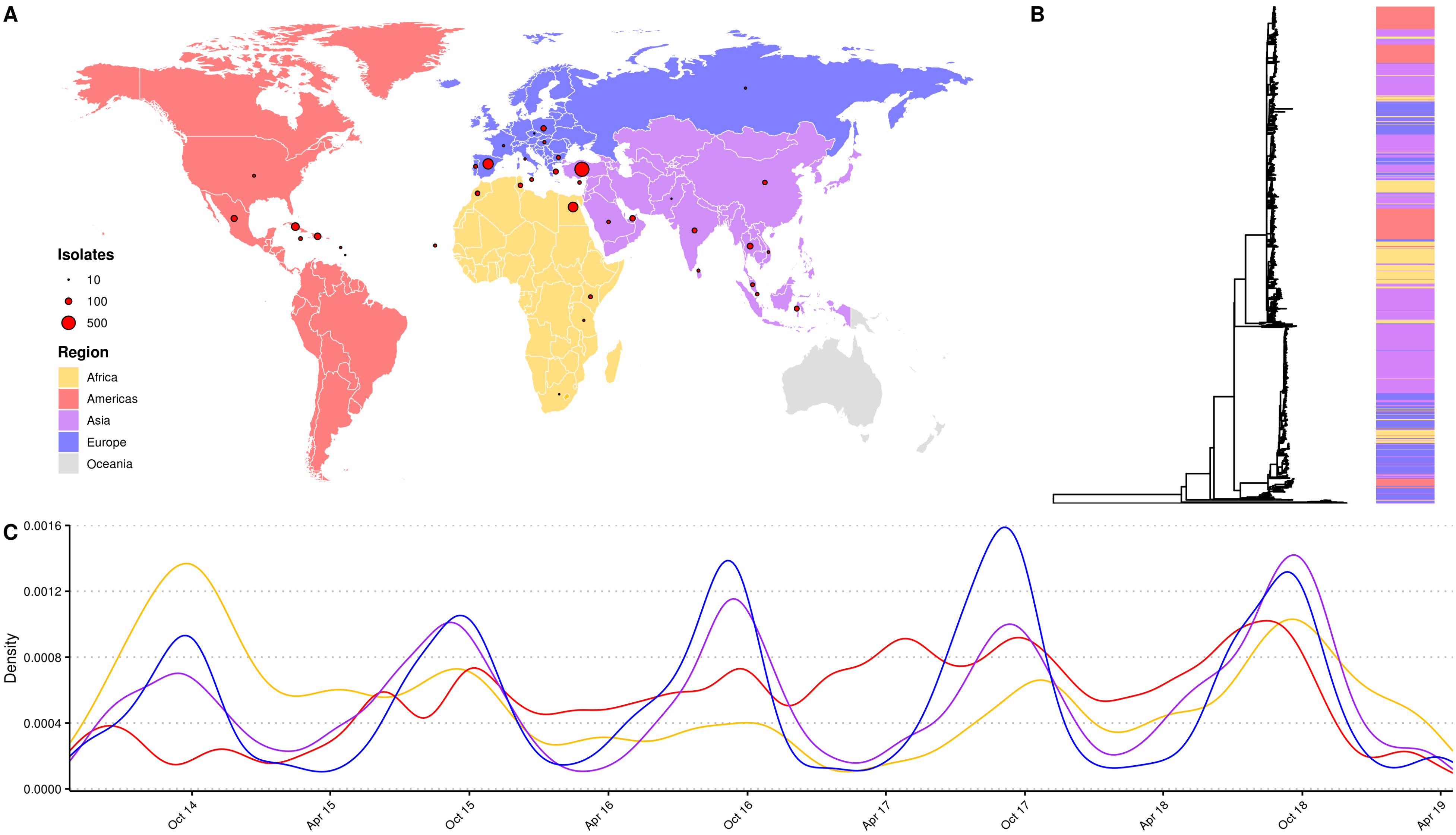
~3,000
2014-2019

Project Aims:

1/ Accurately prediction the geographical origin of *S.* Entertidis infections

2/ Rapidly provide granular information for epidemiologists

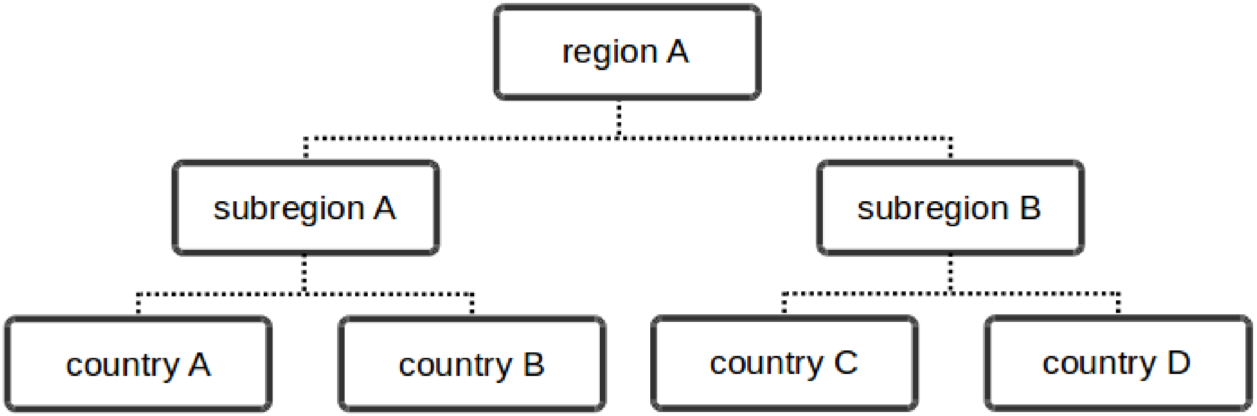
Enteritidis has a strong phylogeographical signal



Hierarchical Classification using ML

Hierarchical classification is a useful tool for problems which have a discrete class hierarchy, in this case Region → Sub-region → Country.

A Local Classifier per Node (LCN) approach was adopted[1]



region A

subregion A

subregion B

country A

country B

country C

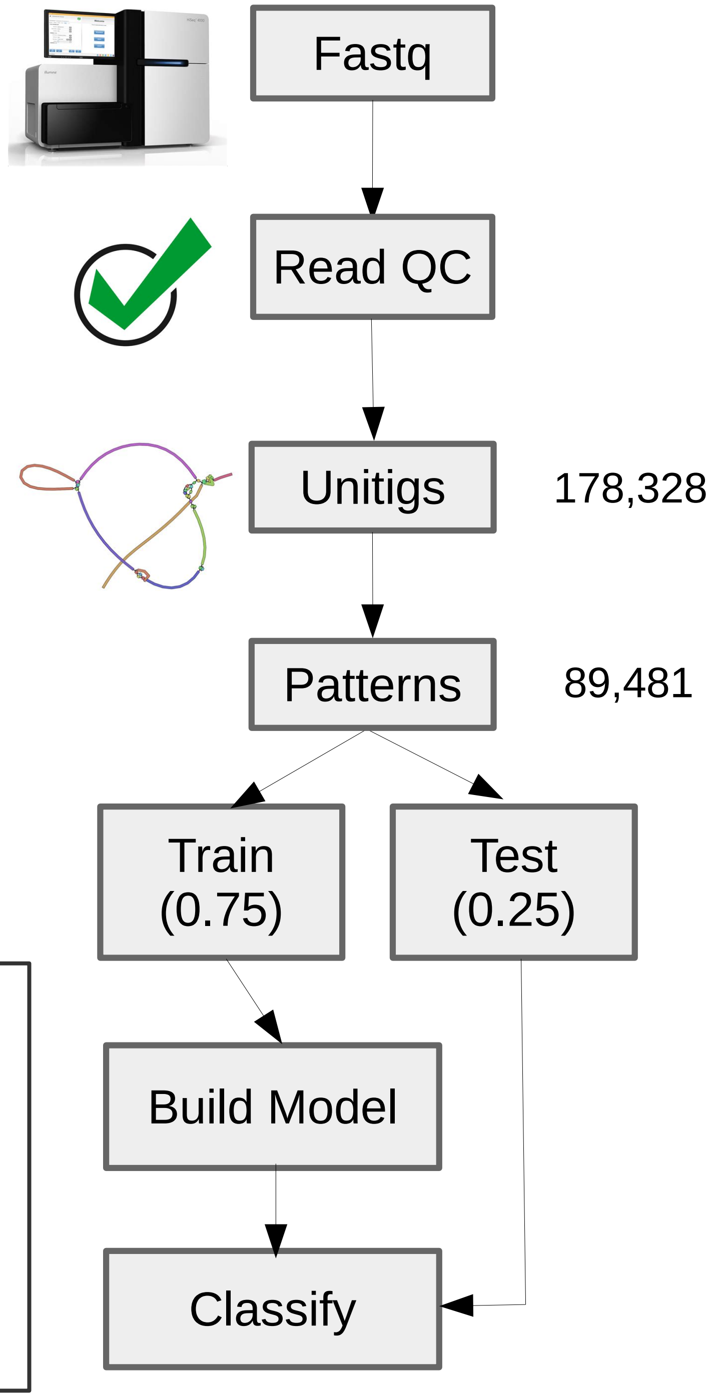
country D

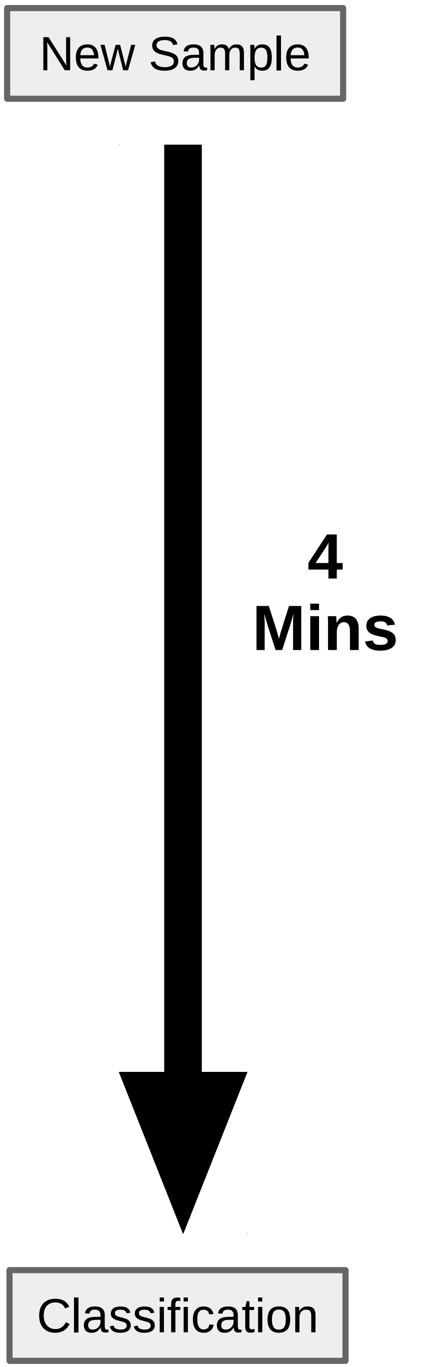
This generates a separate probability score per level of the hierarchy:

Europe	Southern Europe	Spain
0.9	0.7	0.3
✓	✓	✗

Silla, C. N., & Freitas, A. A. (2011). Data Mining and Knowledge Discovery, 22(1), 31–72.

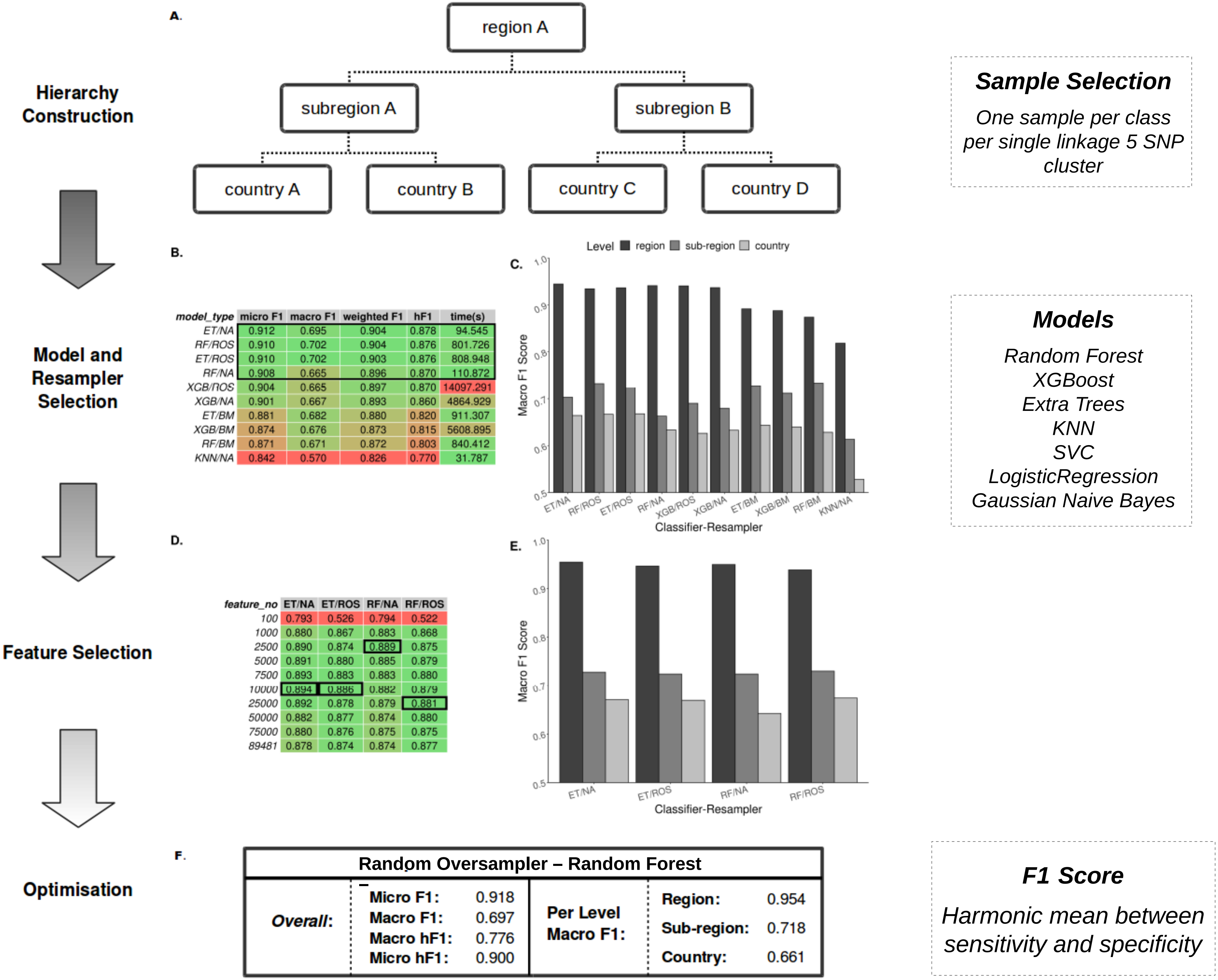
Input Data





Unitig

A sequence representing the strings resulting from compaction of k-mers along maximal paths with non-branching nodes in a *de Brujin* graph



Hierarchy Construction

A.

Sample Selection

One sample per class per single linkage 5 SNP cluster

Model and Resampler Selection

B.

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

C.

Models

Random Forest
XGBoost
Extra Trees
KNN
SVC
LogisticRegression
Gaussian Naive Bayes

Feature Selection

D.

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

E.

Optimisation

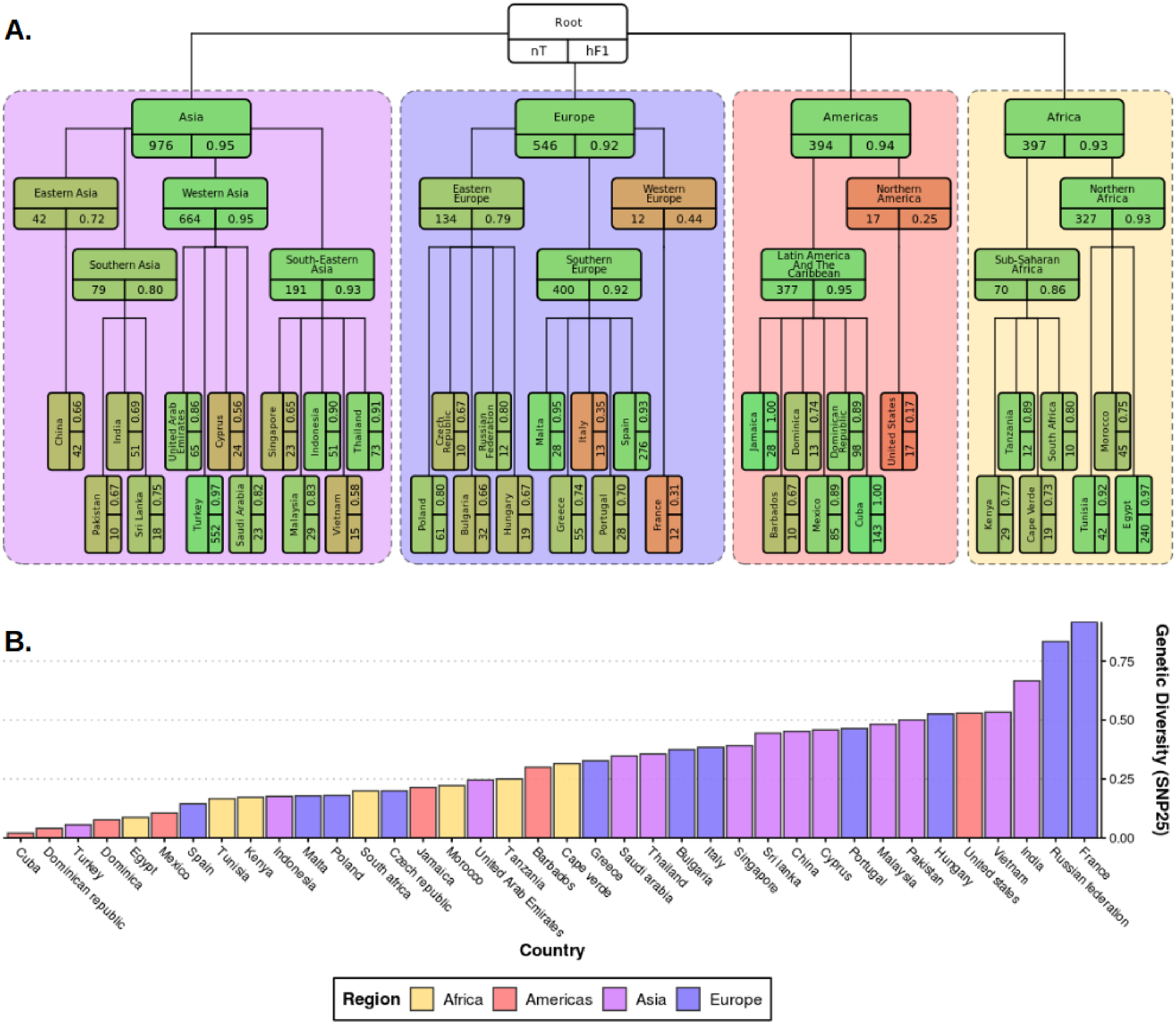
F.

Random Oversampler – Random Forest

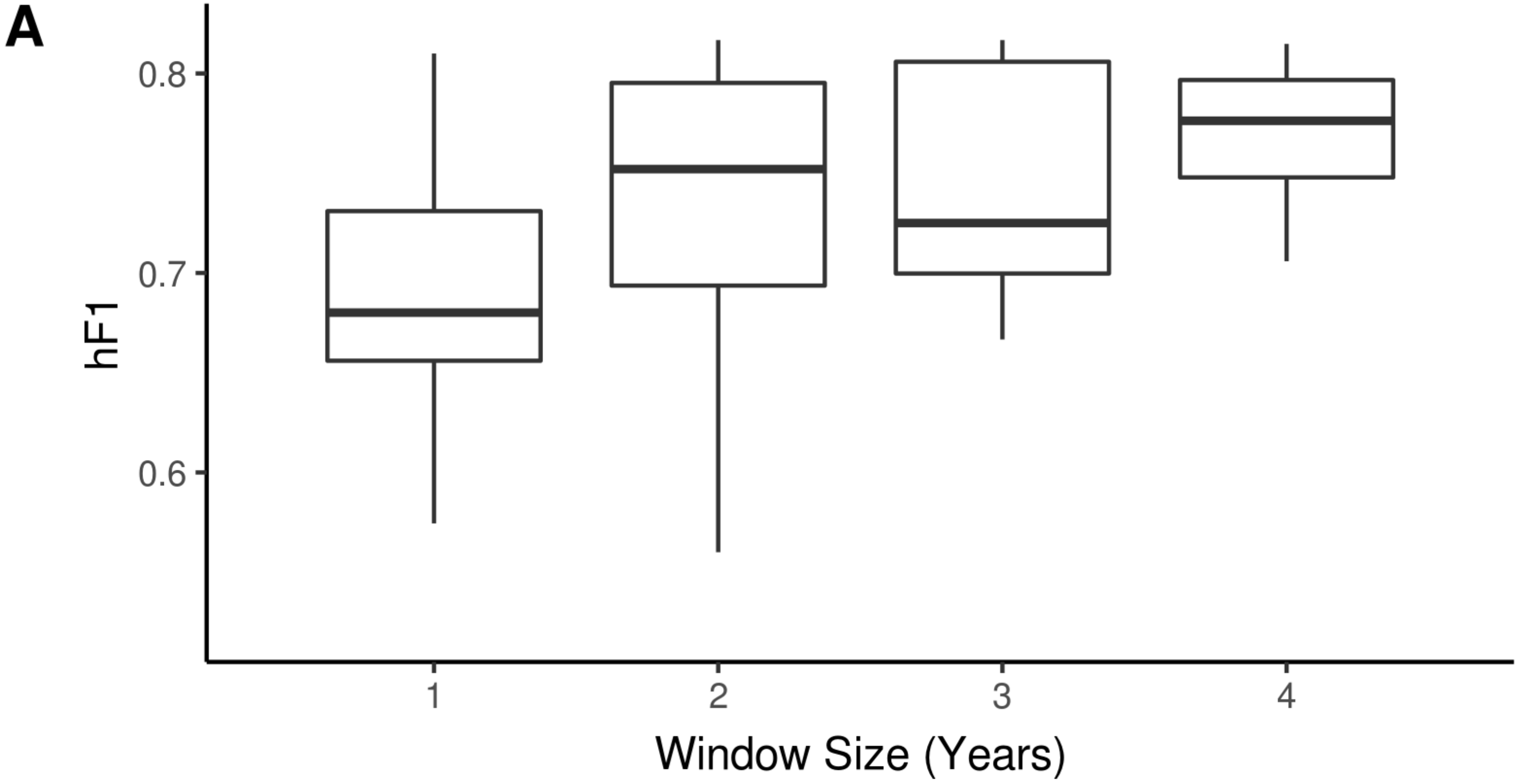
***Overall*:**	**Micro F1:**	0.918	**Per Level Macro F1:**	**Region:**	0.954
	Macro F1:	0.697		**Sub-region:**	0.718
	Macro hF1:	0.776		**Country:**	0.661
	Micro hF1:	0.900			

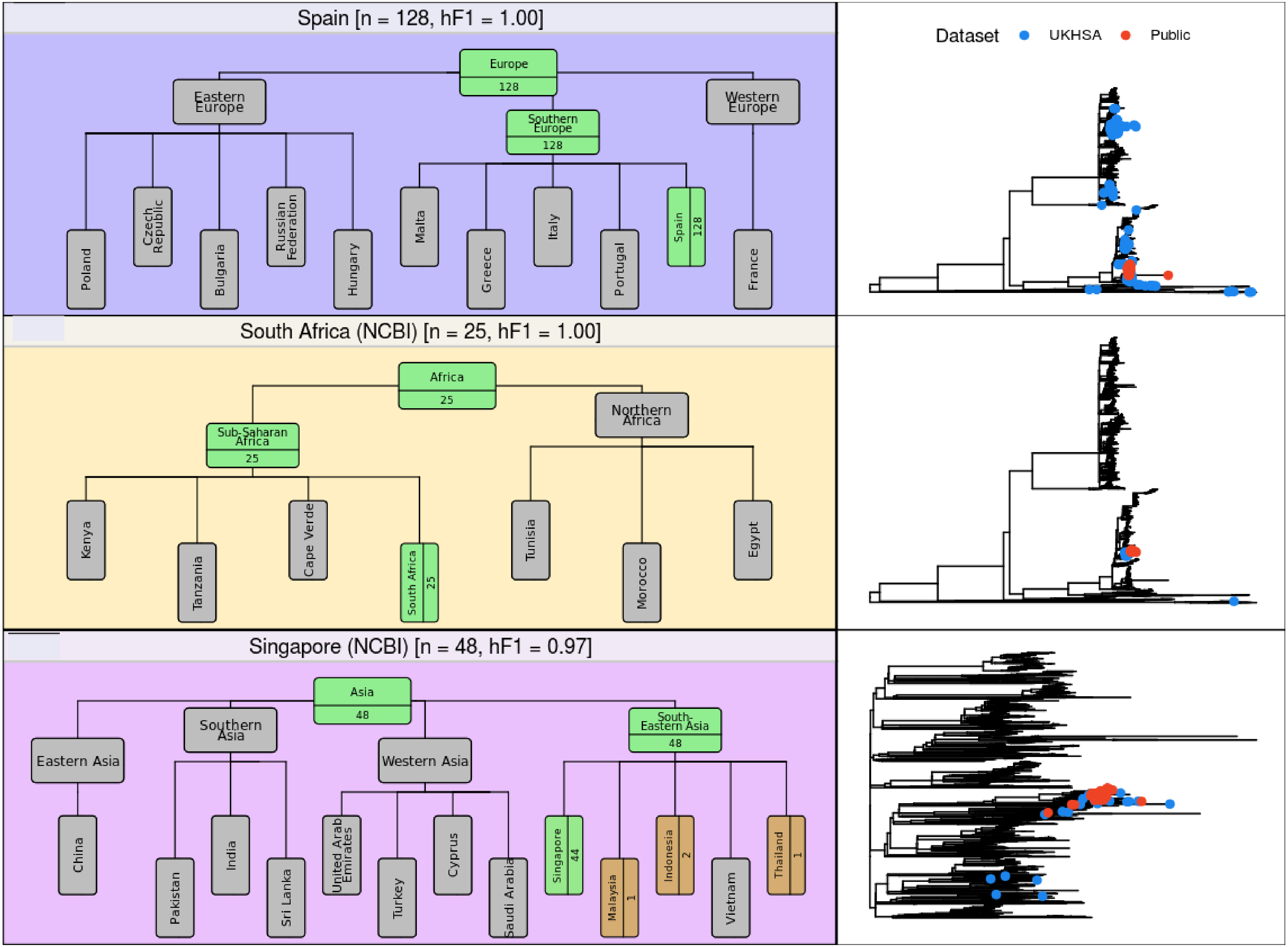
F1 Score

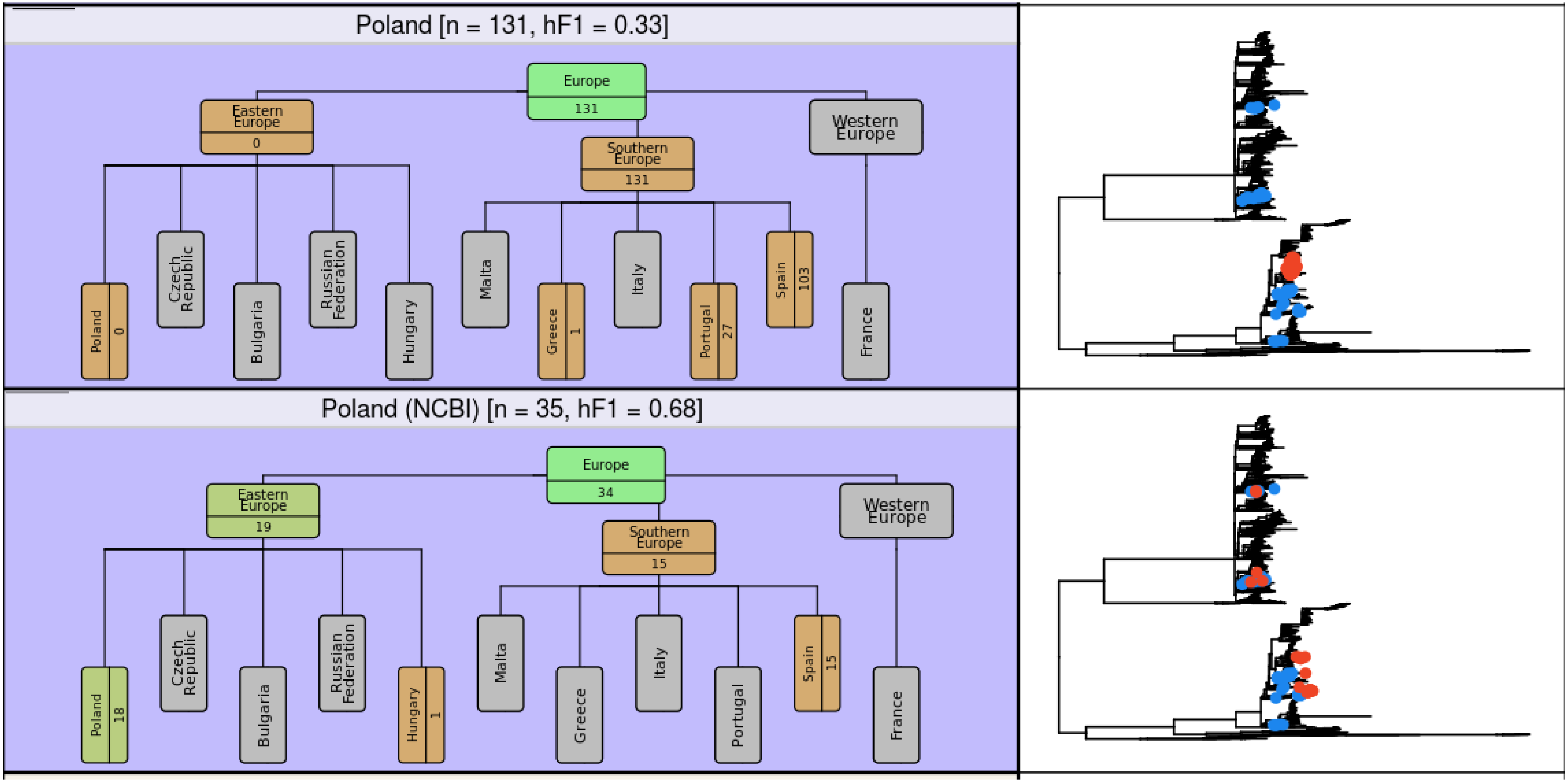
Harmonic mean between sensitivity and specificity


A.
Root
nT
hF1
Asia
976
0.95
Eastern Asia
42
0.72
Western Asia
664
0.95
Southern Asia
79
0.80
South-Eastern Asia
191
0.93
China
42
0.66
India
51
0.69
United Arab Emirates
65
0.86
Cyprus
24
0.56
Singapore
23
0.65
Indonesia
51
0.90
Thailand
73
0.91
Pakistan
10
0.67
Sri Lanka
18
0.75
Turkey
552
0.97
Saudi Arabia
23
0.82
Malaysia
29
0.83
Vietnam
15
0.58
Europe
546
0.92
Eastern Europe
134
0.79
Western Europe
12
0.44
Southern Europe
400
0.92
Czech Republic
10
0.67
Russian Federation
12
0.80
Malta
28
0.95
Italy
13
0.35
Spain
276
0.93
Poland
61
0.80
Bulgaria
32
0.66
Hungary
19
0.67
Greece
55
0.74
Portugal
28
0.70
France
12
0.31
Americas
394
0.94
Northern America
17
0.25
Latin America And The Caribbean
377
0.95
Jamaica
28
1.00
Dominica
13
0.74
Dominican Republic
98
0.89
United States
17
0.17
Barbados
10
0.67
Mexico
85
0.89
Cuba
143
1.00
Africa
397
0.93
Northern Africa
327
0.93
Sub-Saharan Africa
70
0.86
Tanzania
12
0.89
South Africa
10
0.80
Morocco
45
0.75
Kenya
29
0.77
Cape Verde
19
0.73
Tunisia
42
0.92
Egypt
240
0.97
B.
Genetic Diversity (SNP25)
0.00
0.25
0.50
0.75
Country
Cuba
Dominican republic
Turkey
Dominica
Egypt
Mexico
Spain
Tunisia
Kenya
Indonesia
Malta
Poland
South africa
Czech republic
Jamaica
Morocco
United Arab Emirates
Tanzania
Barbados
Cape verde
Greece
Saudi arabia
Thailand
Bulgaria
Italy
Singapore
Sri lanka
China
Cyprus
Portugal
Malaysia
Pakistan
Hungary
United states
Vietnam
India
Russian federation
France
Region
Africa
Americas
Asia
Europe

Robust Subsequent Year Prediction




Spain [n = 128, hF1 = 1.00]
Europe
128
Eastern Europe
Southern Europe
128
Western Europe
Poland
Czech Republic
Bulgaria
Russian Federation
Hungary
Malta
Greece
Italy
Portugal
Spain
128
France
Dataset
UKHSA
Public
South Africa (NCBI) [n = 25, hF1 = 1.00]
Africa
25
Sub-Saharan Africa
25
Northern Africa
Kenya
Tanzania
Cape Verde
South Africa
25
Tunisia
Morocco
Egypt
Singapore (NCBI) [n = 48, hF1 = 0.97]
Asia
48
Eastern Asia
Southern Asia
Western Asia
South-Eastern Asia
48
China
Pakistan
India
Sri Lanka
United Arab Emirates
Turkey
Cyprus
Saudi Arabia
Singapore
44
Malaysia
1
Indonesia
2
Vietnam
Thailand
1



Poland [n = 131, hF1 = 0.33]

Poland (NCBI) [n = 35, hF1 = 0.68]

Summary

Benefits

- Hierarchical classifiers can accurately and granularly predict geographical source of *S.* Enteritidis.
- From raw data to classification in <4 minutes.
- Models are robust to both temporal perturbations and novel data.

Limitations

- The model can only predict countries for which we have data, new datasets are needed.
- The do not always identify the ultimate source of the sample (i.e. food supply chains are complex).

Acknowledgements

UNIVERSITY OF BATH

UK Health Security Agency

Lauren Cowley (PI)

Marie Anne Chattaway

Claire Jenkins

Becky Locke
(Addenbrooks)

Timothy Dallman
(University of Utrecht)

University of BRISTOL

The Academy *of* Medical Sciences

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

Sion C. Bayliss, Rebecca K. Locke, Claire Jenkins, Marie Anne Chattaway,
Timothy J. Dallman, Lauren A. Cowley

doi: https://doi.org/10.1101/2022.08.23.22279111