

Virulence diversity in non-typhoidal *Salmonella*: a population genomics study

Giovanni Merici^{a,b}, Erika Scaltriti^a, Landry Tsoumts Meda^c, Pierre-Emmanuel Douarre^c, Sabrina Cadel-Six^c, Riccardo Percudani^b, Stefano Pongolini^a, Martina Tambassi^{a,d*}

^a Center for Genomic Epidemiology – Research and Education Department, Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia-Romagna, 43126 Parma, Italy

^b Department of Chemistry, Life Sciences and Environmental Sustainability, University of Parma, Parco Area delle Scienze 11/a, 43124 Parma, Italy

^c Salmonella and Listeria Unit (SEL), ANSES, Laboratory for Food Safety, Maisons-Alfort, France

^d Department of Surgical, Medical, Dental and Morphological Sciences with Interest in Transplant, Oncological and Regenerative Medicine, University of Modena and Reggio Emilia, 41124 Modena, Italy

This file contains:

Supplementary Fig. S1

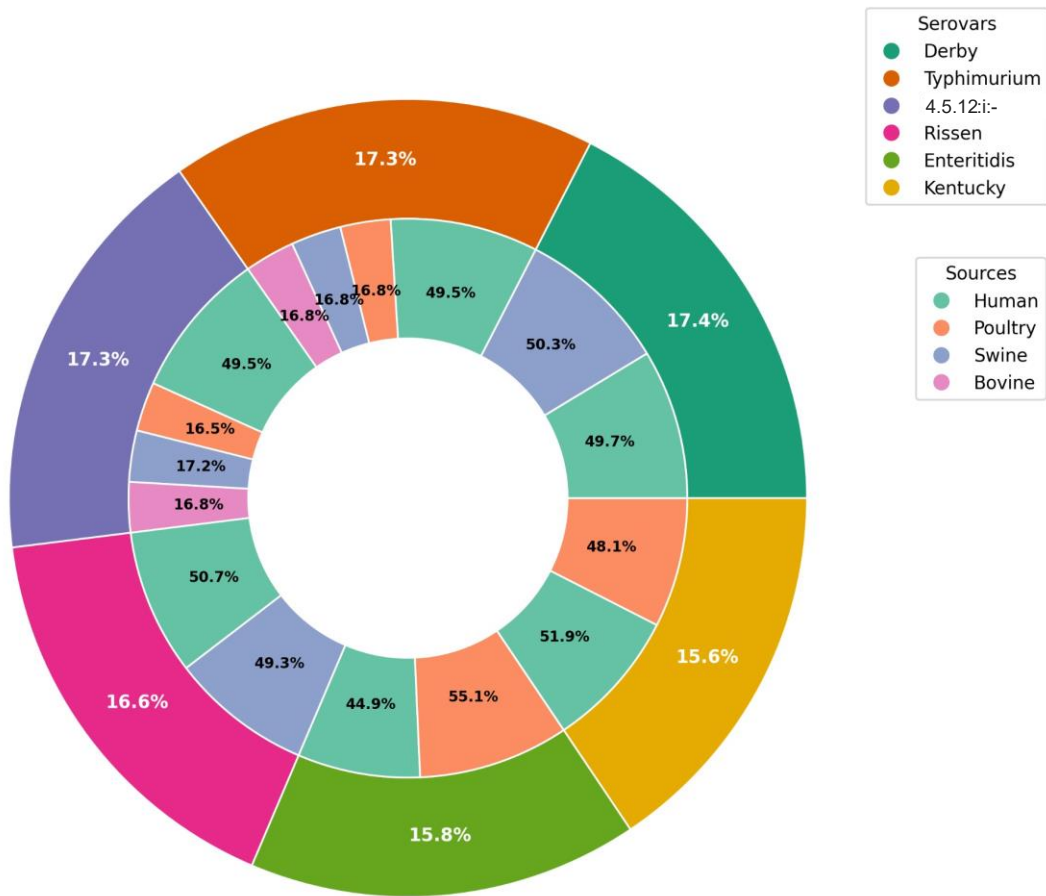
Supplementary Fig. S2

Supplementary Fig. S3

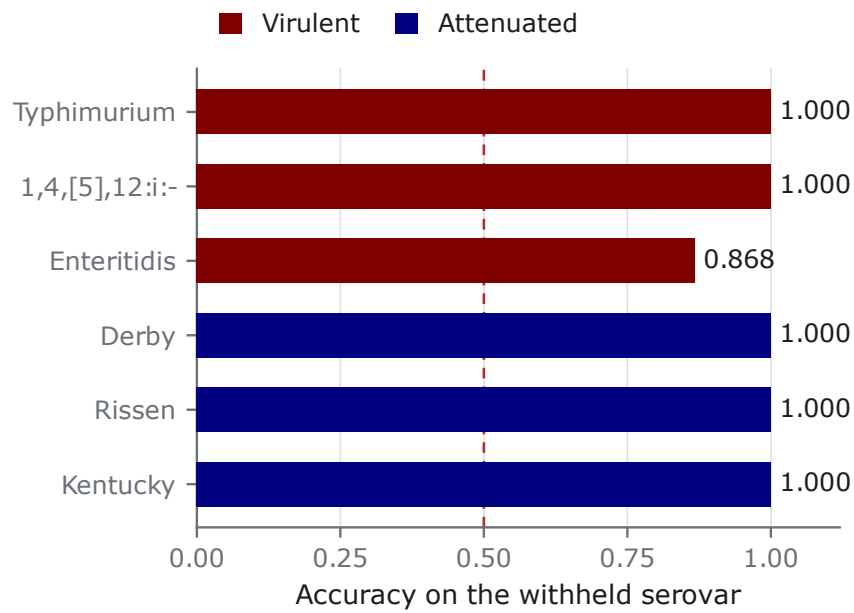
Supplementary Fig. S4

Supplementary Fig. S5

Serovar and Source Distribution

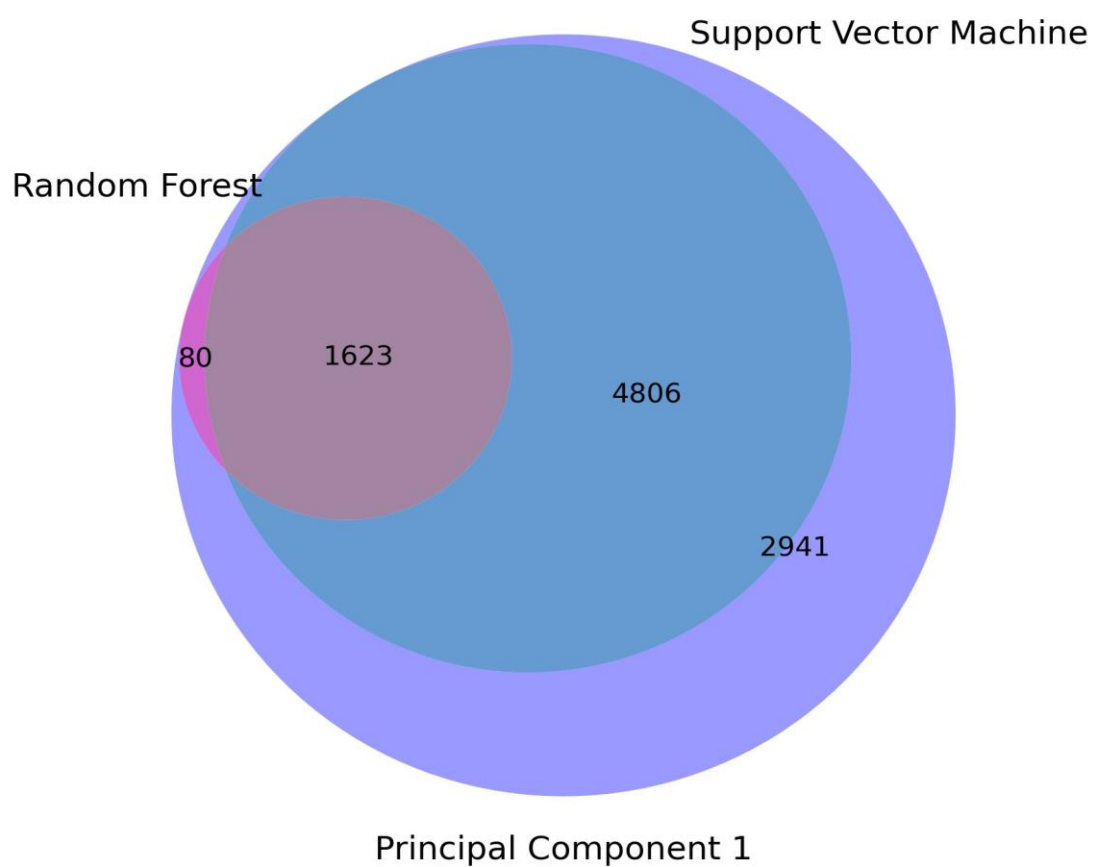


Supplementary Fig. S1: Serovar and source distribution in the dataset. Donut chart shows the proportional representation of the six *Salmonella enterica* serovars under study (outer ring) and their main isolation sources (inner ring).



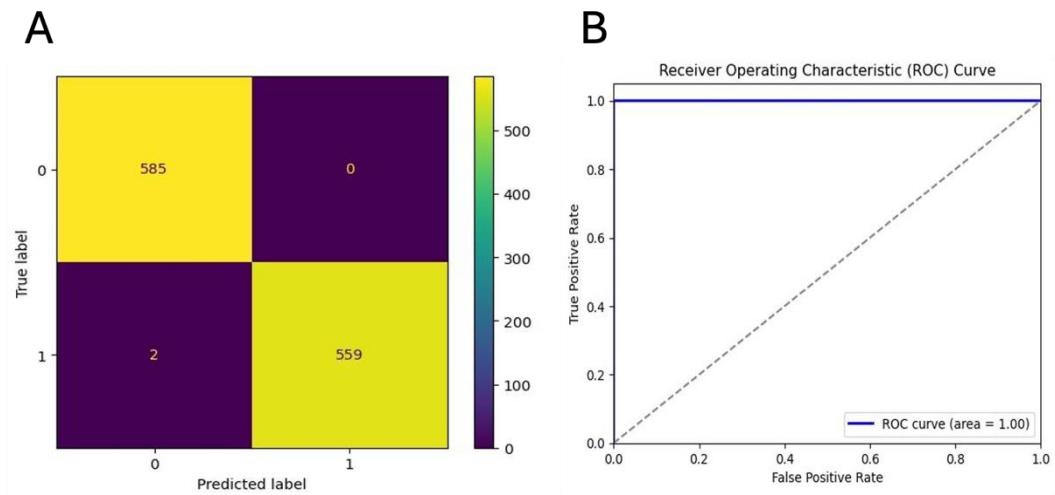
Supplementary Fig. S2: Blocked cross-validation of the Random Forest classifier. Accuracy on each withheld serovar under the leave-one-serovar-out design, in which all genomes of one serovar were withheld and the model was trained on the remaining five. Each fold contains a single class, so accuracy corresponds to the recall for that serovar. The dashed line marks chance (0.5).

Overlap of discriminative genes

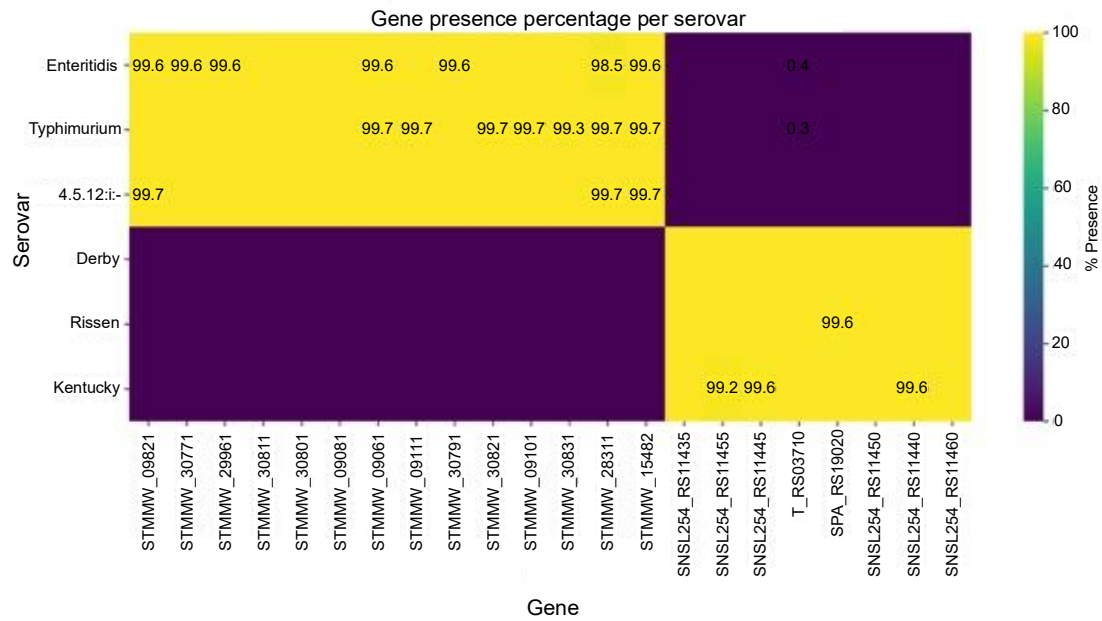


30

31 **Supplementary Fig. S3:** Venn diagram showing overlap among genes assigned non-zero weights by Random
32 Forest, SVM, and PCA (PC1 loadings).



Supplementary Fig. S4: Validation of the 22-gene predictive model on an independent dataset. (A) Confusion matrix showing true versus predicted class labels for genomes from virulent (label 1) and attenuated (label 0) *Salmonella enterica* strains. (B) Receiver Operating Characteristic (ROC) curve for the same model, showing perfect discrimination between classes (AUC = 1.00).



Supplementary Fig. S5: Percentage of gene presence of the 22 genes. Heatmap showing the percentage of presence of the 22 genes identified by RF classifier in the six serovars. Values are expressed as percentages; cells without numerical labels indicate either complete presence (100%) or complete absence (0%).