Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Giovanni Merici
Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna "Bruno Ubertini"
Erika Scaltriti
Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna "Bruno Ubertini"
Landry Tsoumtsa Meda
Agence Nationale de Sécurité Sanitaire de l'Alimentation, de l'Environnement et du Travail
Pierre-Emmanuel Douarre
Agence Nationale de Sécurité Sanitaire de l'Alimentation, de l'Environnement et du Travail
Sabrina Cadel-Six
Agence Nationale de Sécurité Sanitaire de l'Alimentation, de l'Environnement et du Travail
Riccardo Percudani
University of Parma
Stefano Pongolini
Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna "Bruno Ubertini"
Martina Tambassi
martina.tambassi@izsler.it

Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna "Bruno Ubertini"

Article

Keywords: Salmonella, machine-learning, serovars

Posted Date: October 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11468663/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Non-typhoidal salmonellosis is a major foodborne infection mostly due to the consumption of contaminated animal-derived food. However, the proportions among *Salmonella* serovars causing human disease differ from those isolated from food-producing animals, suggesting differences in virulence. This study investigated whether specific genetic determinants are associated with the hypothesized virulence difference among *Salmonella* serovars. We analysed 1,720 genomes of six serovars, categorized as virulent or attenuated based on their epidemiological data, using machine learning to identify differences in gene repertoires. Discriminatory genes underwent semantic analysis using gene annotation embeddings and genomic localization to identify adjacent genes. Phylogenetic analysis assessed the distribution of the minimal discriminatory gene-set across common *Salmonella* serovars. Machine-learning clearly distinguished virulent from attenuated serovars. Discriminatory genes comprise several virulence genes enriched in virulent serovars. The minimal discriminatory gene-set included *Salmonella* Pathogenicity Island (SPI) 13 and 14 genes, essential to cause gut inflammation and to respond to the immune metabolite itaconate. SPI-13 and SPI-14 genes are absent in the evolutionary branch including attenuated serovars and in one phylogenetically unrelated serovar showing epidemiological patterns consistent with attenuation. Our results support the hypothesis of differential virulence among non-typhoidal *Salmonella* serovars and identify genetic differences potentially associated with this variation.

Introduction

Non-typhoidal salmonellosis is a leading foodborne disease worldwide. While the typhoidal serovars, *Salmonella* Typhi and Paratyphi, are restricted to humans, non-typhoidal *Salmonella* have a broad range of animal hosts and food products derived from infected livestock constitute a major source of human infection. Although more than 2,000 non-typhoidal *Salmonella* serovars have been described, most human infections are attributable to a limited number. Specifically, *S*. Typhimurium, its monophasic variant (4,[5],12:i:-), and *S*. Enteritidis account for approximately 40% of cases in the United States (US) and 80% in the European Union (EU) in 2024.[1–3]

These serovars are also frequently isolated in livestock. While *S*. Typhimurium is detected in all food-producing animals, poultry is the primary reservoir for *S*. Enteritidis and pigs represent the main reservoir for *S*. 4,[5],12:i:-. However, other serovars are also frequently isolated in livestock but contribute only marginally to human disease. In the EU in 2024, *S*. Enteritidis accounted for 38·5% of *Salmonella* isolates in laying hens, followed by *S*. Kentucky (11·6%). Notably, *S*. Enteritidis was responsible for 58·4% of human cases, whereas *S*. Kentucky accounted for only 0·34%.[3] Thus, the ratio between these serovars was less than 4:1 in laying hens but exceeded 150:1 in human infections. Accordingly, in the US in 2024 the ratio between these serovars was almost 1:1 in poultry but exceeded 50:1 in human cases.

S. 4,[5],12:i:- constituted 27·9% of isolates from pigs in the EU in 2024, followed by *S*. Derby (23·4%) and *S*. Rissen (8·1%), with a ratio between *S*. 4,[5],12:i:- and *S*. Derby of approximately 1:1, and around 3:1

between *S*. 4,[5],12:i:- and *S*. Rissen.[3] In humans, however, *S*. 4,[5],12:i:- accounted for 10·4% of cases, compared with 1·2% for *S*. Derby and less than 0·26% for *S*. Rissen, yielding ratios of approximately 8:1 and more than 40:1, respectively.[3] Similarly, in the US the ratio between *S*. 4,[5],12:i:- and *S*. Derby in pigs was about 4:1 vs 10:1 in human cases.[1, 2]

Although multiple factors may contribute to explain these discrepancies in the ratio between serovars in human cases vs animal reservoirs — including different ability to persist along the food production chain — we hypothesized that differences in the repertoire of virulence-associated genetic determinants could be a key determinant. Machine learning (ML) approaches provide valuable opportunities to interrogate this biological complexity.[4] In this study, we applied complementary ML approaches on 1,720 complete genomes retrieved from open-access databases to identify virulence-associated genetic determinants which distinguish *S.* Enteritidis, *S*. Typhimurium, and *S*. 4,[5],12:i:- from *S*. Derby, *S*. Rissen and *S*. Kentucky. These serovars were classified as virulent and attenuated, respectively, based on their relative representation in human infections compared with their animal reservoirs of exposure.

Following gene annotation and identification of candidate determinants, we conducted phylogenetic analyses to reconstruct the evolutionary trajectories of these serovars and their associated genomic features within the global *Salmonella* population.

Identifying such genomic signatures could provide further evidence of differential virulence among non-typhoidal *Salmonella* and establish a genetic framework for its detection. These findings could refine our understanding of virulence at the population level and inform risk assessment and control strategies across the food supply chain.

Methods

Dataset generation

A total of 1,720 genomes from *S*. Typhimurium (297), *S*. 4,[5],12:i:- (297), *S*. Enteritidis (272), *S*. Derby (300), *S*. Rissen (286) and *S*. Kentucky (268) were selected from EnteroBase (Supplementary Table S1) and their whole-genome Multi Locus Sequence Type (wgMLST) allelic profiles were downloaded.[5] Genomes were selected as follows: serovar was predicted by the Achtman 7 gene Multi Locus Sequence Type (MLST) scheme; for Europe and North America, an approximately equal number of genomes per serovar from humans and animals was randomly selected, resulting in about 300 genomes per serovar. For *S*. Typhimurium and 1,4,[5],12:i:- of animal origin, approximately equal numbers of genomes were randomly selected from swine, poultry, and bovine isolates (Supplementary Fig. S1). This selection included several STs per serovar reflecting intra-serovar diversity (Supplementary Table S2).

WgMLST allelic profiles were transformed into a presence/absence matrix of functional genes with a custom script. For each gene, alleles were classified as absent (coded as “0”) when carrying an early stop codon as well as when classified as missing or truncated/fragmented/indeterminable

(coded as '–' or − 1 in the wgMLST original profiles, respectively). Early stop codons were identified by retrieving all alleles for each gene from Enterobase and screening the coding sequences for premature occurrence of any of the three stop codons (TAA, TAG, or TGA) before the expected termination codon. In all other cases, the allele was considered present (coded as "1"). The transformation of the 1,720 wgMLST profiles to a matrix of binary profiles was followed by data curation steps: identical binary profiles were collapsed to a single representative profile, and non-informative genes (i.e., those always present or always absent across all strains) were removed. All data processing was performed in Python 3.9.20 using pandas 2.2.3 and NumPy 2.0.2.[6, 7]

Machine learning analyses

Principal component analysis (PCA), K-means clustering, Random Forest (RF) and Support Vector Machine (SVM) were performed on the presence/absence matrix of functional genes using scikit-learn 1.5.2.[8] K-means clustering was applied using Euclidean distance with n_clusters = 2, n_init = 10, and random_state = 0. Ten independent runs with different random initializations were executed to verify cluster stability, and cluster agreement across runs was quantified using the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). Representative clustering results were visualized by projecting genome coordinates onto the first two principal components.

For supervised analyses, genomes were labelled as "virulent" if belonging to *S*. Typhimurium, *S*. 1,4, [5],12:i:-, *S*. Enteritidis or "attenuated" if belonging to *S*. Derby, *S*. Rissen and *S*. Kentucky. The binary presence/absence matrix was randomly split into a training set (60%) and a test set (40%). For RF, the best-performing configuration was: n_estimators = 300, min_samples_split = 20, min_samples_leaf = 2, max_leaf_nodes = 50, max_features='log2', and max_depth=None. The optimized model was retrained on the full training set (oob_score=True, n_jobs=−1) and evaluated by out-of-bag accuracy, five-fold cross-validation, and test-set metrics including overall accuracy, Cohen's κ, and area under the ROC curve (AUC). ROC curves were computed from predicted class probabilities using roc_curve and AUC from sklearn.metrics and plotted with matplotlib 3.9.2. SVM was trained on the same data and labels as for RF, with probability estimates enabled to compute ROC curves. A randomized search was used to optimize hyperparameters using five-fold cross-validation and random_state = 0. The best configuration (C = 10, kernel='linear', gamma = 1, degree = 2) was retrained and evaluated on the same 60/40 train/test split used for RF. Performance was assessed through accuracy, Cohen's κ, and AUC.

RF was iteratively trained by retaining only genes with non-zero importance scores at each cycle and accuracy = 1. The model was validated using a test panel comprising wgMLST profiles from 1,146 Asian genomes retrieved from EnteroBase (Supplementary Table S3), which were converted into a presence/absence matrix of functional genes as described above.

RF classification performance was assessed using a cross-validation design in which the folds were blocked by serovar. Specifically, all genomes of one serovar were withheld and the model was trained on

the remaining five serovars. Performance was evaluated as the proportion of withheld genomes correctly classified.

Gene-content variation was partitioned by Permutational Multivariate Analysis of Variance (PERMANOVA) based on Jaccard distances, which measures the pairwise dissimilarity between genomes as the proportion of not-shared genes. Serovar, source of isolation and continent of origin (defined as factors) were fitted together with all pairwise interactions, using Type II sums of squares. PERMANOVA significance was assessed empirically by 999 permutations of the labels for each factor. For source and continent, labels were permuted only within each serovar, to prevent inter-serovar differences from confounding the estimated effects of these factors. Accordingly, the minimum attainable p-value was 0.001. Analyses were performed in Python 3 using NumPy, SciPy and scikit-learn.

Semantic and genomic localization analysis of RF discriminative genes

For semantic analysis, functional annotations of the 1,703 RF discriminative genes were retrieved from EnteroBase and embedded using PubMedBERT (microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext). [5, 9] Tokenization and embedding generation were performed in PyTorch 2.2.0 using the transformers 4.41.2 library.[10] Each textual description was tokenized and passed to the pretrained model to obtain sentence embeddings. Specifically, the generated embeddings were clustered using BERTopic 0.16.0 which integrates UMAP (Uniform Manifold Approximation and Projection) and HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) with class-based TF–IDF labeling.[11–13] Dimensionality reduction was performed using UMAP (n_neighbors = 15, min_dist = 0.1, metric = 'cosine'), followed by density-based clustering with HDBSCAN (min_cluster_size = 10, metric = 'euclidean', cluster_selection_method = 'eom'). Both models were initialized with random_state = 42 to ensure reproducibility. [12, 13] Separate analyses were conducted on virulent- and attenuated-enriched genes. This classification was obtained by calculating, for each RF discriminative gene, its presence frequency in virulent and attenuated serovars. The difference in presence frequency ($\Delta f = f_{virulent} - f_{attenuated}$) was then used to identify loci enriched in either virulent or attenuated serovars.

For the genomic localization analysis, BLASTn was performed using BLAST+ v2.12.0.[14] Hits were retained with E-value ≤ 1×10^{-5}, sequence identity ≥ 40%, and coverage ≥ 80%. Analyses were conducted separately for genes enriched in virulent and attenuated serovars, such that genes enriched in virulent serovars were queried against one representative genome of each virulent serovar, and genes enriched in attenuated serovars were queried against one representative genome for each attenuated serovar. Reference genomes were downloaded from NCBI (Supplementary Table S4).[15] DNA parsing and sequence handling were performed through Biopython 1.85.[16] Genes located less than 10 kb apart on the same contig were considered spatially linked. Adjacencies were aggregated across genomes to construct two undirected graphs: a pan-virulent-serovar graph and a pan-attenuated-serovar graph.

Network analyses were conducted in NetworkX 3.5, with edge weights defined as the inverse of the median inter-gene distance w = 1/(1 + distance).[17] Community detection was performed using the Leiden algorithm implemented in leidenalg 0.10.2 and python-igraph 0.11.9, with resolution = 1.0 and random_state = 42.[18]

Leiden communities identified independently for each serovar were subsequently compared across serovars based on gene content. Communities were considered conserved among serovars of the virulent or attenuated group when sharing at least 80% of their genes. For each shared community, the median delta frequency (MΔF) was calculated as the difference between median gene frequency within the community in virulent versus attenuated serovars.

For genes annotated as hypothetical proteins in EnteroBase, corresponding annotations from *S*. Typhimurium SL1344 genome (accession FQ312003.1) were used and literature was consulted for a more detailed assessment of gene function.

Phylogenetic reconstruction

Representative genomes of 57 *Salmonella enterica* serovars were selected by downloading from EnteroBase five genomes corresponding to the most prevalent sequence type (ST) for each serovar and four outgroup genomes were downloaded from NCBI (Supplementary Table S4).[5, 15] Core Single-Nucleotide Polymorphisms (core-SNPs) were identified using kSNP4 v4.1, with the optimal k-mer size (k = 17) determined using KChooser.[19] Maximum-likelihood phylogenetic inference was performed using RAxML-NG under the GTR+ASC_LEWIS model on the core-SNP matrix.[20] The resulting tree was rooted using the four designated outgroup genomes and visualized with iTOL v6.[21]

Results

Virulent and attenuated serovars can be distinguished according to their presence/absence profiles of functional genes

The study dataset was generated by converting the wgMLST allelic profiles of 1,720 genomes from Europe and North America belonging to virulent (*S*. Typhimurium, *S*. 4,[5],12:i:-, *S*. Enteritidis) and attenuated serovars (*S*. Derby, *S*. Rissen and *S*. Kentucky) into a presence/absence matrix of functional genes. After conversion and data curation, the dataset comprised 1,703 unique presence/absence profiles of functional genes and a final set of 9,450 genes. The conversion was done to investigate whether events of gene gain and loss or pseudogenization are involved in the hypothesized virulence differences between virulent and attenuated serovars.

The matrix was first analysed through two unsupervised ML methods, PCA and k-means. The distribution of presence/absence gene profiles along the two main principal components (PC1 and PC2)

largely mirrored the serovar-based clustering, revealing four distinct groups: one corresponding to *S*. Typhimurium and *S*. 1,4,[5],12:i:-, a second to *S*. Rissen, a third to *S*. Enteritidis and a fourth to a mixture of *S*. Kentucky and *S*. Derby (Fig. 1a). Notably, neither source nor geographic origin appeared to influence clustering, as genomes from different hosts and continents were mixed within serovar-specific groups. PC1 scores resulted significantly higher in virulent than in attenuated serovars (Mann–Whitney U test, $p < 0{\cdot}0001$), indicating that these two groups were distinguished based on their presence/absence gene profiles. To confirm this result, k-means clustering was applied with $K = 2$ (Fig. 1b). This algorithm consistently assigned genomes to two clusters that matched the virulent *vs* attenuated partition. Both the Adjusted Rand Index and Normalized Mutual Information were equal to 1.0, confirming that the separation was completely stable and not dependent on initialization.

We next trained two supervised ML models, RF and SVM. Both models perfectly discriminated between virulent and attenuated serovars, achieving 100% performance across all evaluation metrics (Fig. 1c-d). Therefore, the results of both unsupervised and supervised methods demonstrated that virulent and attenuated serovars differ based on their presence/absence profiles of functional genes.

To assess whether this discrimination was influenced by the presence of closely related genomes belonging to the same serovar in both the training and the test set, a cross-validation scheme based on RF was implemented, with folds blocked by serovar. Specifically, six models were trained, each using five serovars and tested on the sixth, which had been excluded from training. The six models correctly classified an average of 97.8% of genomes from the serovars withheld from training (range: 0.868–1.000; Supplementary Fig. S2).

Furthermore, the contribution of serovar, source of isolation and continent of origin to variation in gene presence/absence profiles was quantified by PERMANOVA based on Jaccard distances, which measure the dissimilarity between two genomes as the proportion of not-shared genes. The three factors were fitted together with all their pairwise interactions (serovar × source, serovar × continent and source × continent). Interactions tested whether the association between one factor and variation in gene presence/absence profiles depended on the level of another factor. For example, the serovar × source interaction tested whether differences in gene content between human- and animal-derived genomes vary across serovars. Serovar explained 73.5% of variation in gene presence/absence profiles, whereas source of isolation and continent of origin each explained less than 0.5% (source $R2 = 0.0043$; continent $R2 = 0.0046$), and no interaction exceeded 2.5% ($R2 \leq 0.025$; Supplementary Table S5). Consistently, a RF model using source and continent as the sole predictors, without gene presence/absence information, did not distinguish virulent from attenuated serovars (balanced accuracy 0.497).

Lastly, we compared the genes identified by PCA, RF and SVM models as relevant for distinguishing between virulent and attenuated serovars. We observed that 1,623 genes were consistently identified as discriminative by all models (Supplementary Fig. S3). Furthermore, scatterplots comparing RF gene discriminative scores to those from SVM and PCA analyses (Fig. 1e-f) showed substantial overlap, with highly ranked RF genes concentrated at the extremes of both distributions, while low-importance RF

genes clustered near zero. This result highlights the convergence of RF, SVM, and PC1 results, showing that a subset of the analysed genes plays a key role in distinguishing virulent from attenuated serovars.

Semantic analysis of RF discriminative genes enriched in virulent and in attenuated serovars

We next analysed the functions of the RF discriminative genes. Genes were first classified as enriched in virulent or attenuated serovars by calculating their prevalence in the two groups (Supplementary Table S6). In doing so, 881 genes resulted enriched in the first group and 822 in the other. Then, the gene functions were analysed based on the annotations available in EnteroBase for each wgMLST locus. Textual descriptions were encoded with PubMedBERT to capture semantic relationships among annotations and clustered using BERTopic.[9, 11] This procedure groups genes into clusters according to the similarity of their annotation texts, it labels each cluster using enriched keywords and ranks the clusters by gene count.

BERTopic grouped 793/881 functional genes enriched in virulent serovars into 21 semantically coherent clusters ranging from 125 to 11 included genes (Fig. 2a and Supplementary Table S6). Notably, cluster 2 (“type, transfer, conjugative, fimbrial”) includes several virulence genes located in horizontally acquired loci, such as genes from SPI − 1 (*sipA* and *orgB*), SPI-2 (*ssaD*, *ssaQ*, *ssaU*), SPI-4 (*siiD* and *siiE*), SPI-5 (*pipC*), *spv* locus (*spvC* and *spvR*) and genes belonging to the *lpf*, *pef* and SPI-6 located *saf* fimbrial operons (*lpfE*, *lpfC lpfB*, *pefA*, *pefB pefD, safA* and *safB*).[22–27] Furthermore, cluster 15 (“gifsy1, prophage, gifsy2, preconnector”) encompasses genes belonging to Gifsy-1 and Gifsy-2 prophages, reported to be carriers of several virulence genes.[28]

Regarding the 822 functional genes enriched in attenuated serovars, 754 were assigned to 18 clusters ranging from 222 to 10 included genes (Fig. 2b and Supplementary Table S6). Cluster 8 (“fimbrial, minor, k88, subunit”) comprises genes encoding an expanded repertoire of fimbriae, including stf, stk, tcf, and K88. These genes differ from the fimbrial genes enriched in virulent serovars, indicating that virulent and attenuated serovars have a different repertoire of fimbriae. No other known virulent genes were found enriched in attenuated serovars. Conversely, attenuated serovars resulted enriched in genes encoding mercury, arsenic and copper resistance proteins, included in cluster 4 (“domaincontaining, partial, resistance, domain”).

Taken together, these results revealed that several virulence genes are more frequently present in a functionally intact form in virulent compared to attenuated serovars.

Several discriminative genes co-localize into communities of adjacent genes

As functional analysis revealed that several discriminative genes belong to genomic clusters such as SPIs and prophages, we next inferred gene adjacency networks based on genomic localization to

identify communities of adjacent genes.

The analysis on genes enriched in virulent serovars identified 592/881 genes included in 270 communities of adjacent genes, with 77 in *S*. Enteritidis, 94 in *S*. Typhimurium and 99 in its monophasic variant. Across the 270 communities, 85 (31·48%) consisted of two genes, whereas 185 (68·52%) contained more than two genes, revealing that discriminative genes are mostly arranged in multi-gene clusters (Fig. 3a).

Communities were then compared across virulent serovars based on their gene content, revealing that 39 communities were present in all virulent serovars (Fig. 3b and Supplementary Table S7). Some shared communities include genes identified through the BERTopic analysis, such as those of SPI-1, SPI-2, SPI-4, SPI-5 and the *lpf* operon, and spatial localization enabled a comprehensive analysis of all genes belonging to these virulence clusters. For instance, a shared community includes the already identified SPI-1 genes *sipA* and *orgB* but also *hilC* and *sicP* (Fig. 3c), while the SPI-2 shared cluster includes *ssaD, ssaQ* and *ssaU* together with *ssrA* and *sseF* (Fig. 3d). Other shared communities were found to include additional virulence genes, such as those of the *rfa and rfb* operons, encoding the biosynthesis pathway of lipopolysaccharide core oligosaccharide and O-antigen, respectively, part of SPI-13 and the entire SPI-14.

Communities shared among virulent serovars have different frequencies in virulent compared to attenuated serovars, as measured by MΔF (Supplementary Table S7). Interestingly, some shared communities have MΔF > 90%, indicating that their genes are almost exclusively present in virulent serovars. Among them, we found part of SPI-13, SPI-14, and clusters encoding a phosphoenolpyruvate-dependent sugar phosphotransferase system and the D-galactonate metabolic pathway.[29] Other shared communities, such as SPI-1, SPI-2, SPI-4 and the *rfa* cluster, have MΔF < 10%, indicating a slightly reduced frequency in attenuated compared to virulent serovars.

The same analysis was performed on functional genes enriched in attenuated serovars. In this case, 455/822 genes resulted included in 170 communities of adjacent genes, with 53 in *S*. Kentucky, 60 in *S*. Rissen and 57 in *S*. Derby. Across the 170 communities, 48 (28·24%) consisted of two genes, and 122 (71·76%) contained more than two genes (Fig. 3e and Supplementary Table S7). A total of 20 communities were found in all attenuated serovars (Fig. 3f), with one corresponding to the acr3-type arsenic resistance operon. It has MΔF=-99·9%, indicating that it is almost exclusively present in attenuated serovars. Differently, the other heavy-metal resistance genes identified through the semantic analysis resulted not shared among all attenuated serovars. The other shared communities include genes with unknown function, except for that with galactitol metabolic pathway genes, with a slightly reduced frequency in virulent compared to attenuated serovars (MΔF=-8·4%).

Overall, these results indicate that most genes enriched in both virulent and attenuated serovars are associated with specific genomic clusters that, in virulent serovars, harbour several virulence genes.

A minimum set of 22 genes is sufficient to distinguish between virulent and attenuated serovars

To identify the minimal gene set required to distinguish between virulent and attenuated serovars, we iteratively trained the RF model by retaining only genes with non-zero importance scores at each cycle. The model retained perfect discrimination between virulent and attenuated serovars as the number of genes decreased to 22. To validate the discriminative power of the 22 genes, we used the model exclusively trained on them and tested it on an independent dataset of 1,146 Asian genomes to ensure no overlap with the training genomes. The model achieved accuracy of 0.998, Cohen's kappa of 0·997, and ROC AUC of 1·0 (Supplementary Fig. S4), confirming that these 22 genes robustly distinguish virulent from attenuated serovars.

Analysing their prevalence in the two groups, 14 genes resulted almost always present in virulent but not in attenuated serovars and eight genes showed the opposite trend (Table 1 and Supplementary Fig. S5). Six genes of virulent serovars are located in SPI-13, with three genes belonging to the itaconate degradation operon and one gene encoding the itaconate-responsive transcriptional regulator RipR.[30, 31] The other SPI-13 genes encode a catecholamine-induced arylsulfatase and its maturase.[32] Four other genes found only in virulent serovars belong to SPI-14, with one corresponding to *loiA*, which encodes a SPI-1 transcriptional regulator.[33]

The gene *orgB* was also associated with virulent serovars. However, unlike the other genes that are missing, the *orgB* allele of attenuated serovars contains an early stop codon located in the third-to-last codon of the coding sequence, suggesting that it may not significantly affect the protein function. The remaining five functional genes detected only in virulent serovars have not yet been fully characterized.

Among the genes found only in attenuated serovars, six belong to the already identified acr3-type arsenic resistance operon, including its integrase, whereas two are poorly characterized.[34]

The synteny of genes belonging to SPI-13, SPI-14 and the arsenic resistance operon across serovars was analysed, revealing a strong genomic conservation of the three loci and the surrounding genes within virulent *vs* attenuated serovars (Fig. 4a-c).

Phylogenetic distribution of the 22-genes set across Salmonella enterica serovars

We next examined the phylogeny and the distribution of the 22 genes across the studied serovars, as well as the other *Salmonella enterica subspecies enterica* serovars more frequently isolated from humans in Europe and the United States (Fig. 5).[35] The genomes of the originally identified attenuated serovars belonged to a well-supported clade, together with all serovars that closely matched the attenuated 22-gene profile (average similarity 91·5%), with the exception of *S.* Ohio. Notably, many of these serovars display epidemiological patterns consistent with attenuation, being common in livestock but rarely associated with human disease. Indeed, *S.* Mbandaka is the third most common serovar in broilers (5·5%) and the fifth in laying hens (2·5%), yet it accounts for less than 0·26% of human infections.[3]

Similarly, *S.* Agona, frequently isolated from turkeys (46·1%) and present also in laying hens (1·4%), broilers (0·88%), bovines (0·72%), contributes to only 0.56% of human cases.[3] *S.* Senftenberg, isolated from laying hens (1·1%), broilers (0·95%) and turkeys (0·92%), shows a comparable pattern, being responsible for less than 0·26% of human infections.[3] The remaining serovars in this group as well as S. Ohio are isolated at low rates in livestock and are rarely associated with human salmonellosis.

Serovars sharing the virulent 22-gene profile are distributed across different branches of the tree. While *S.* Typhimurium and *S.* 1,4,[5],12:i:- cluster together, *S.* Enteritidis genomes are in a distinct clade including also *S.* Dublin and *S.* Berta. The other serovars with the virulent 22-genes profile, such as *S.* Paratyphi B, *S.* Java and *S.* Braenderup, belong to distinct branches of the tree. This indicates that serovars with the virulent 22-gene profile are not phylogenetically closely related. Interestingly, if we focus on non-typhoidal serovars, it can be observed that SPI-13 and SPI-14 genes are largely conserved except for the attenuated serovar clade and *S.* Ohio, suggesting independent gene loss in the common ancestor of the attenuated serovar clade and in the S. Ohio lineage. Differently, the acr3-type arsenic resistance operon occurs in several phylogenetically distant serovars, suggesting multiple independent acquisitions.

Table 1

Set of 22 genes that distinguish between virulent and attenuated serovars. For each gene, information on the wgMLST locus, reference genome locus tag (*S*. Typhimurium LT2 for virulent serovars; *S*. Derby CP029486 for attenuated serovars), associated SPI, gene name, protein function, and status are reported.

wgMLST locus	locus tag	SPI	Gene	Function	Association
STMMW_30771	STM3117	13	\	itaconate response	virulent serovars
STMMW_30791	STM3119	13	*ripB*	itaconate response	virulent serovars
STMMW_30801	STM3120	13	*ripC*	itaconate response	virulent serovars
STMMW_30811	STM3121	13	*ripR*	itaconate-responsive transcriptional regulator	virulent serovars
STMMW_30821	STM3122	13	\	arylsulfatase	virulent serovars
STMMW_30831	STM3123	13	\	arylsulfatase maturase	virulent serovars
STMMW_09061	STM0855	14	\	hypothetical electron transfer flavoprotein beta subunit	virulent serovars
STMMW_09081	STM0857	14	\	hypothetical acyl-CoA dehydrogenase	virulent serovars
STMMW_09101	STM0859	14	*loiA*	SPI-1 transcriptional regulator	virulent serovars
STMMW_09111	STM0860	14	\	hypothetical inner membrane protein	virulent serovars
STMMW_28311	STM2869	1	*orgB*	oxygen-regulated invasion protein	virulent serovars
STMMW_09821	STM0971	\		putative cytoplasmic protein	virulent serovars
STMMW_15482	\	\	*IsrF*	IsrF ncRNA	virulent serovars
STMMW_29961	STM3035	\	\	putative periplasmic protein	virulent serovars
T_RS03710	DAD50_11315	\	\	tetratricopeptide repeat protein	attenuated serovars
SPA_RS19020	DAD50_01920	\	\	Crp/Fnr family transcriptional regulator	attenuated serovars

wgMLST locus	locus tag	SPI	Gene	Function	Association
SNSL254_RS11460	DAD50_11970	\	*arsR*	transcriptional regulator	attenuated serovars
SNSL254_RS11455	DAD50_11975	\	*arsC*	arsenate reductase	attenuated serovars
SNSL254_RS11450	DAD50_11980	\	*arsD*	As[III]-responsive regulator	attenuated serovars
SNSL254_RS11445	DAD50_11985	\	*arsA*	arsenical pump-driving ATPase	attenuated serovars
SNSL254_RS11440	DAD50_11990	\	*acr3*	As[III] efflux antiporter	attenuated serovars
SNSL254_RS11435	DAD50_11995	\	\	integrase	attenuated serovars

Discussion

Epidemiological data show that the disproportionate contribution of specific non-typhoidal *Salmonella* serovars to human infections cannot be explained by their different prevalence in livestock reservoirs and the consequent diverse level of human exposure. Additional factors, including differences in survival along the food production chain and virulence-associated traits, may contribute to these epidemiological patterns. In this study, we specifically investigated the contribution of virulence-associated genetic determinants to these differences. To this end, we used ML approaches to evaluate whether *Salmonella* serovars categorized as virulent and attenuated based on their epidemiological data can be distinguished based on their genetic repertoire. Unsupervised ML methods revealed a clear partition of genomes into the two groups and supervised methods showed perfect ability to discriminate between them. **Importantly, this concordance among methods extended beyond the separation of the two groups, including the set of discriminative genes identified by RF, SVM and PCA that were largely overlapping across the three methods.** Together, these results indicate that the observed separation is due to a strong genetic signature. This separation was not an artefact of the validation design: when cross-validation folds were blocked by serovar, the model retained complete rank separation on serovars absent from the training set, indicating that the discriminative signal is shared between the two groups rather than specific to individual serovars. Furthermore, the separation is not attributable to the composition of the dataset, since source of isolation and continent of origin together accounted for less than 1% of gene-content variation, against 73.5% for serovar. This result is consistent with the fact that the three attenuated serovars of the study belong to the same branch in the *Salmonella* evolutionary tree. Therefore, that genetic signal reflects the evolutionary specificity of this branch. Nevertheless, semantic and genomic localization analyses of RF discriminative genes showed that many known virulence determinants are less frequently detected in the attenuated serovars than in virulent serovars**. In particular, some of these genes resulted absent in attenuated serovars, such as the** itaconate degradation cluster contained in **SPI-13 and the whole SPI-14, while o**thers showed a slightly reduced

frequency in attenuated compared to virulent serovars, such as genes of SPI-1, that we analysed in a previous study and found subjected to decay in attenuated serovars.[36]

Looking for a minimum essential set of genes required by RF to distinguish between virulent and attenuated serovars, we identified 22 such genes. This set includes virulence genes present exclusively in virulent serovars, such as the entire gene cluster for itaconate degradation in SPI-13 and genes of SPI-14. Studies using a mouse intestinal infection model have shown that the absence of SPI-14 leads to attenuation of virulence, as loss of the SPI-14 gene *loiA* reduces SPI-1 gene expression required for invasion of the intestinal epithelium and induction of inflammatory responses.[33] In light of that study, the absence of *loiA* in our attenuated serovars could be a major determinant of their reduced pathogenic potential. It is also plausible that the lack of *loiA* affected the observed decay of SPI-1 genes in attenuated serovars, because the altered expression of those genes could have rendered them no longer necessary.

The itaconate degradation gene cluster has been extensively studied because it enables the bacterium to resist itaconate, a compound produced mainly by macrophages to kill internalized bacteria as well as to regulate the immune response to pathogen exposure.[37,38]

The importance of itaconate as a general antimicrobial agent of the innate immune system is substantiated by the production of itaconate-degrading enzymes in many pathogens other than *Salmonella*, such as *Pseudomonas aeruginosa*, *Yersinia pestis* and *Mycobacterium tuberculosis*. Therefore, lack of the itaconate degradation cluster in attenuated serovars likely plays a crucial role in limiting their survival within the host and, consequently, their ability to cause disease.

The 22 gene set also includes the arsenic resistance operon which is present in attenuated but not in virulent serovars. This finding is consistent with the diffusion of these serovars in livestock environment, where heavy metals are used as feed additives or disinfectants.[39]

The phylogenetic tree of *Salmonella* provides clues about the genetic events that may have led to the emergence of the attenuated serovars. The three attenuated serovars of the study are genetically closely related and their evolutionary branch also includes other serovars. These same-branch serovars not only share the SPI-13 and SPI-14 deletion profile of the study serovars but also display a remarkable epidemiological profile of attenuation. At the same time, there is also a phylogenetically unrelated serovar, *S.* Ohio, which has the same deletions and is rarely isolated in livestock as well as from human cases of salmonellosis. Altogether, this evidence is consistent with almost identical deletion events that occurred independently in the two lineages that have led to the Ohio serovar, on one side, and to the other attenuated serovars of the study, on the other side. These genetic events could be explained by the fact that both deleted genomic regions correspond to horizontally acquired elements. While SPI-14 is a single genetic element, SPI-13 is composed of distinct modules such as that including the itaconate degradation operon, that can mobilize independently from the rest of the pathogenicity island because they are flanked by nearly perfect direct repeats.[40] The results indicate that the most common

attenuated serovars likely originated from deletion events affecting SPI-14 and part of SPI-13 which occurred at least twice in the evolutionary history of *Salmonella*.

Overall, the combination of the known virulence role of SPI-13 and SPI-14, the ancestral reconstruction identifying their loss on the branch leading to the attenuated lineage, and the independent occurrence of the same deletions in the phylogenetically unrelated serovar *S.* Ohio, provide a biologically meaningful framework supporting the hypothesis that these genomic losses contributed to the evolution of the attenuated phenotype. Evidently, the potential attenuation of these serovars has not compromised their ecological success, as they are widely distributed in livestock, likely as a result of adaptation to new ecological niches such as intensive animal farming.

This study contributes to our understanding of the genomic features potentially associated with differences in virulence among Salmonella serovars, using a global population genomics approach.

Declarations

Additional Information

The authors declare that there are no conflicts of interest

Funding information

This work was supported by the Italian Ministry of Health through Grant No. PRC2025001 and by the Italian Ministry of University and Research (MUR) under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 3.3, funded by the European Union – NextGenerationEU (DM 117/2023).

Data availability

The wgMLST profiles used in this study are publicly available in Enterobase database and the corresponding genomes are listed in Supplementary Table S1 and Supplementary Table S3. Genomes used for the phylogenetic tree in this study are publicly available in Enterobase database and in the National Center for Biotechnology Information (NCBI) and are listed in Supplementary Table S4. Code, analytical workflows and parameters used for data processing are available in Materials and Methods and at

https://github.com/IZSLER-GenomicEpidemiology/Salmonella_Virulence.

Author contributions

G.M., S.P. and M.T. conceptualised the study. G.M. and M.T. developed the methodology, curated the data and conducted the investigation. G.M. performed the formal analyses, developed the software and prepared the visualisations. All authors validated the results. G.M., M.T., E.S., L.T.M., P.-E.D. and S.C.-S. provided resources. E.S. and M.T. acquired funding and administered the project. E.S., S.C.-S., P.-E.D.,

L.T.M., R.P., S.P. and M.T. supervised the study. G.M., S.P. and M.T. wrote the original manuscript draft. All authors reviewed and edited the manuscript.

References

1. Quarterly Sampling Reports on Salmonella and Campylobacter.
2. Centers for Disease Control and Prevention (CDC). BEAM (Bacteria, Enterics, Ameba, and Mycotics) Dashboard. Atlanta, Georgia: U.S. Department of Health and Human Services. www.cdc.gov/beam/dashboard/index.html.
3. The European Union One Health 2024 Zoonoses Report. *EFSA Journal* **23**, (2025).
4. Munck, N., Njage, P. M. K., Leekitcharoenphon, P., Litrup, E. & Hald, T. Application of Whole-Genome Sequences and Machine Learning in Source Attribution of *Salmonella* Typhimurium. *Risk Analysis* **40**, 1693–1705 (2020).
5. Zhou, Z., Alikhan, N.-F., Mohamed, K., Fan, Y. & Achtman, M. The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. *Genome Res.* **30**, 138–152 (2020).
6. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
7. Mckinney, W. *Data Structures for Statistical Computing in Python*. (2010).
8. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos Pedregosa, Varoquaux, Gramfort et al. Matthieu Perrot. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
9. Gu, Y. *et al.* Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthc.* **3**, (2022).
10. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. http://arxiv.org/abs/1912.01703 (2019).
11. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. http://arxiv.org/abs/2203.05794 (2022).
12. McInnes, L., Healy, J. & Astels, S. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software* **2**, 205 (2017).
13. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).
14. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
15. O'Leary, N. A. *et al.* Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Sci. Data* **11**, 732 (2024).
16. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
17. Hagberg hagberg, A. A., -Los, lanlgov, Schult, D. A. & Swart swart, P. J. *Exploring Network Structure, Dynamics, and Function Using NetworkX*.

http://conference.scipy.org/proceedings/SciPy2008/paper_2 (2008).

18. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
19. Hall, B. G. & Nisbet, J. Building Phylogenetic Trees From Genome Sequences With kSNP4. *Mol. Biol. Evol.* **40**, (2023).
20. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).
21. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* **52**, W78–W82 (2024).
22. Gerlach, R. G., Jäckel, D., Geymeier, N. & Hensel, M. *Salmonella* Pathogenicity Island 4-Mediated Adhesion Is Coregulated with Invasion Genes in *Salmonella enterica*. *Infect. Immun.* **75**, 4697–4709 (2007).
23. Guiney, D. G. & Fierer, J. The Role of the spv Genes in Salmonella Pathogenesis. *Front. Microbiol.* **2**, (2011).
24. Bäumler, A. J., Tsolis, R. M. & Heffron, F. The lpf fimbrial operon mediates adhesion of Salmonella typhimurium to murine Peyer's patches. *Proceedings of the National Academy of Sciences* **93**, 279–283 (1996).
25. Bäumler, A. J. *et al.* The pef fimbrial operon of Salmonella typhimurium mediates adhesion to murine small intestine and is necessary for fluid accumulation in the infant mouse. *Infect. Immun.* **64**, 61–68 (1996).
26. Bäumler, A. J. *et al.* Contribution of horizontal gene transfer and deletion events to development of distinctive patterns of fimbrial operons during evolution of Salmonella serotypes. *J. Bacteriol.* **179**, 317–322 (1997).
27. Lou, L., Zhang, P., Piao, R. & Wang, Y. Salmonella Pathogenicity Island 1 (SPI-1) and Its Complex Regulatory Network. *Front. Cell. Infect. Microbiol.* **9**, (2019).
28. Figueroa-Bossi, N., Uzzau, S., Maloriol, D. & Bossi, L. Variable assortment of prophages provides a transferable repertoire of pathogenic determinants in *Salmonella*. *Mol. Microbiol.* **39**, 260–272 (2001).
29. Reizer, J., Michotey, V., Reizer, A. & Saier, M. H. Novel phosphotransferase system genes revealed by bacterial genome analysis: Unique, putative fructose- and glucoside-specific systems. *Protein Science* **3**, 440–450 (1994).
30. Elder, J. R., Paul, N. C., Burin, R., Guard, J. & Shah, D. H. Genomic organization and role of SPI-13 in nutritional fitness of Salmonella. *International Journal of Medical Microbiology* **308**, 1043–1052 (2018).
31. Hersch, S. J. & Navarre, W. W. The *Salmonella* LysR Family Regulator RipR Activates the SPI-13-Encoded Itaconate Degradation Cluster. *Infect. Immun.* **88**, (2020).

32. Das, S., Sreevidya, V. S., Udvadia, A. J. & Gyaneshwar, P. Dopamine-induced sulfatase and its regulator are required for Salmonella enterica serovar Typhimurium pathogenesis. *Microbiology (N. Y).* **165**, 302–310 (2019).
33. Jiang, L. *et al.* Signal transduction pathway mediated by the novel regulator LoiA for low oxygen tension induced Salmonella Typhimurium invasion. *PLoS Pathog.* **13**, e1006429 (2017).
34. Mourão, J. *et al.* Tolerance to arsenic contaminant among multidrug-resistant and copper-tolerant Salmonella successful clones is associated with diverse operons and genetic contexts. *Environ. Microbiol.* **22**, 2829–2842 (2020).
35. Cherchame, E., Ilango, G., Noël, V. & Cadel-Six, S. Polyphyly in widespread Salmonella enterica serovars and using genomic proximity to choose the best reference genome for bioinformatics analyses. *Front. Public Health* **10**, (2022).
36. Tambassi, M. *et al. Salmonella* pathogenicity Island 1 undergoes decay in serovars adapted to swine and poultry. *Microbiol. Spectr.* **13**, (2025).
37. Duncan, D. & Auclair, K. Itaconate: an antimicrobial metabolite of macrophages. *Can. J. Chem.* **100**, 104–113 (2022).
38. Peace, C. G. & O'Neill, L. A. J. The role of itaconate in host defense and inflammation. *Journal of Clinical Investigation* **132**, (2022).
39. Deng, W. *et al.* Antibiotic Resistance in *Salmonella* from Retail Foods of Animal Origin and Its Association with Disinfectant and Heavy Metal Resistance. *Microbial Drug Resistance* **24**, 782–791 (2018).
40. Espinoza, R. A. *et al.* Differential roles for pathogenicity islands SPI-13 and SPI-8 in the interaction of Salmonella Enteritidis and Salmonella Typhi with murine and human macrophages. *Biol. Res.* **50**, 5 (2017).

Figures

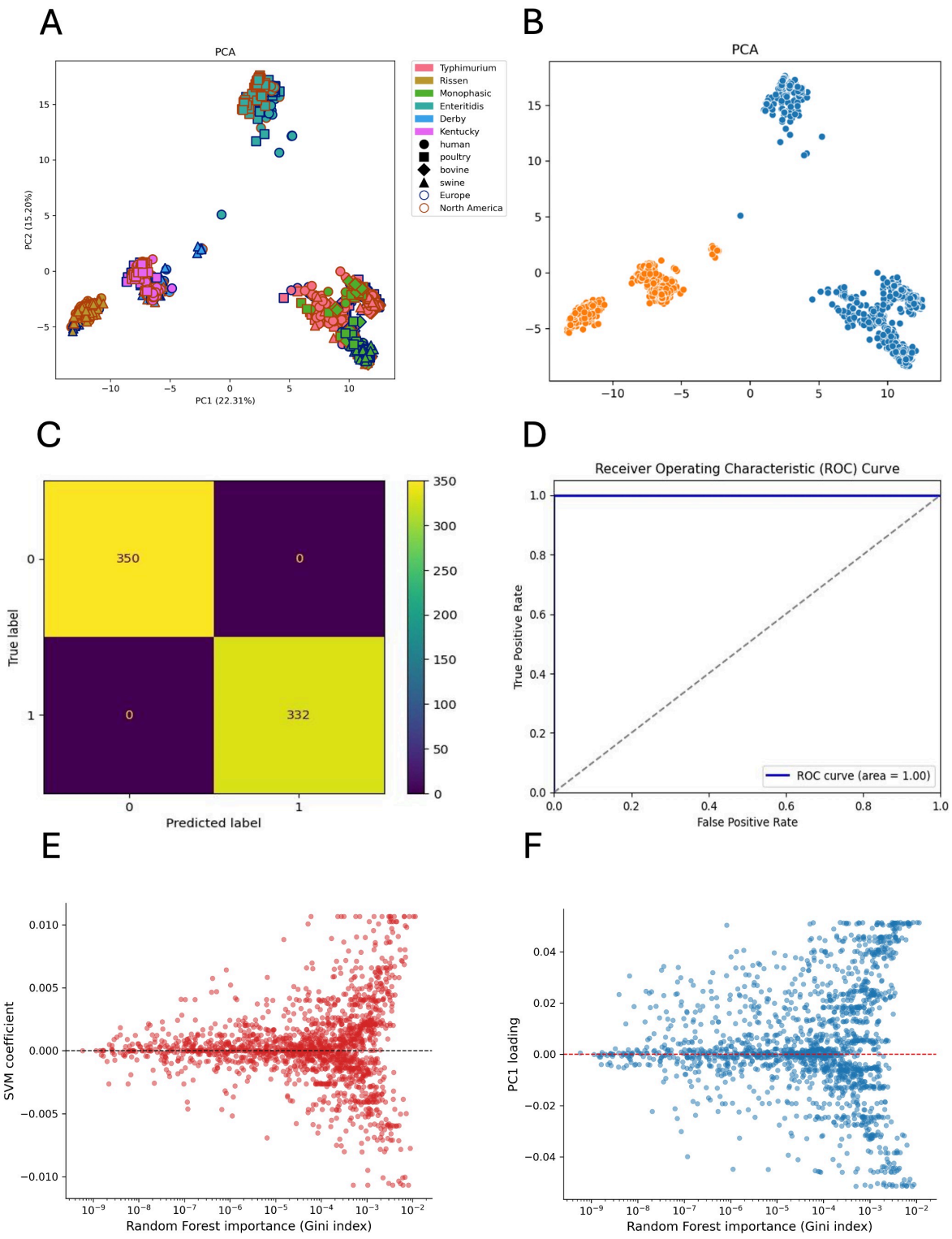
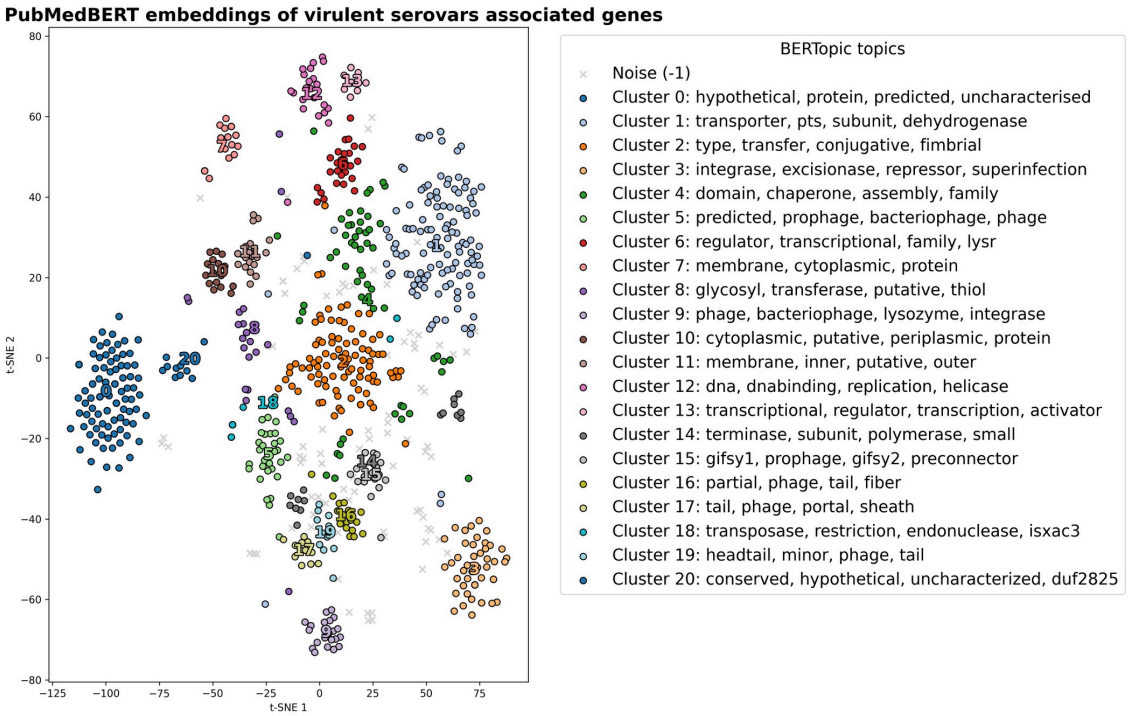


Figure 1

Unsupervised and supervised machine learning discrimination of virulent and attenuated *Salmonella*serovars. (a) PCA of the binary presence/absence matrix showing the main axes of genetic variation among *Salmonella enterica* serovars. (b) K-means clustering (K = 2) projected onto the PCA space, with genomes colored according to cluster assignment. Virulent serovars are shown in blue, while attenuated serovars are shown in orange. (c) Confusion matrix illustrating the performance of both the

RF and SVM models on the independent test set (40% of the total dataset), comparing predicted versus actual classifications. Both models achieved accuracy of 1·0, Cohen's κ of 1·0 and an out-of-bag (OOB) score of 1·0 with a mean accuracy of 1·0 in 5-fold cross-validation. (d) Receiver operating characteristic (ROC) curve of both RF and SVM models evaluated on the independent test set (AUC = 1·0). Scatterplots comparing the discriminative scores derived from RF (corresponding to Random Forest importances, Gini index) with those obtained from SVM (SVM coefficients) (e) and PCA (PC1 loadings) (f).

A



B

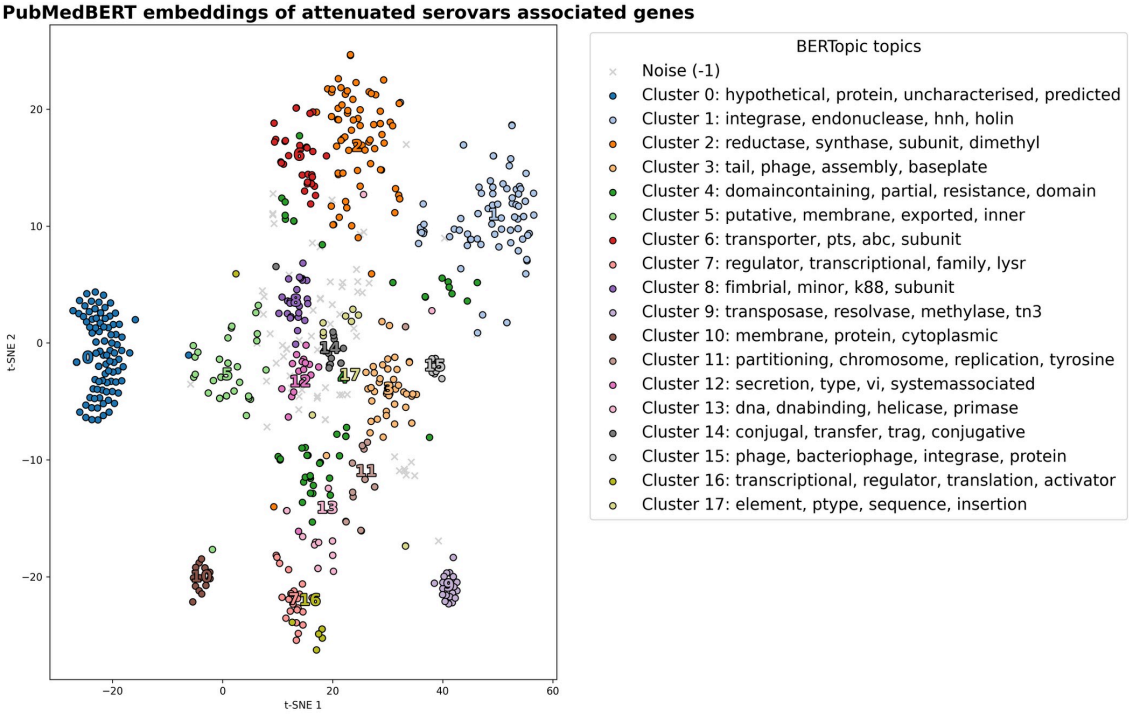


Figure 2

Topic modeling of virulent and attenuated *Salmonella enterica* serovars based on PubMedBERT text embeddings of Enterobase annotations. Two-dimensional t-SNE projection of PubMedBERT embeddings for genes enriched in virulent (a) and attenuated (b) serovars, grouped by BERTopic clusters.

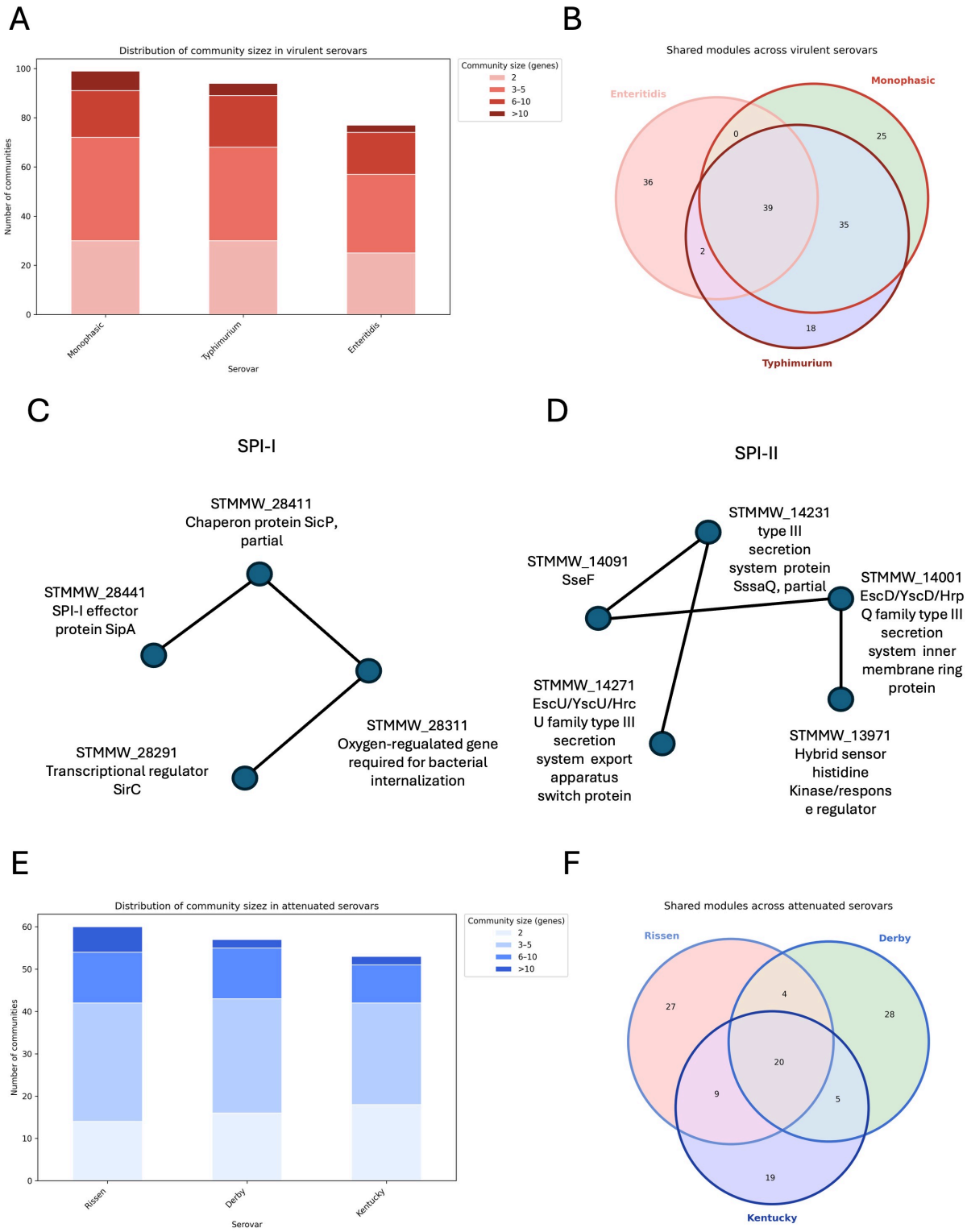


Figure 3

Genomic organization and community structure of discriminative genes across *Salmonella* serovars. (a and e) Distribution of community sizes (number of genes per community) across virulent and attenuated serovars. (b and f) Venn diagram showing shared Leiden communities across virulent and attenuated serovars. Communities were considered identical when sharing at least 80% of their genes. (c and d) Communities including SPI-1 and SPI-2 genes shared among virulent serovars.

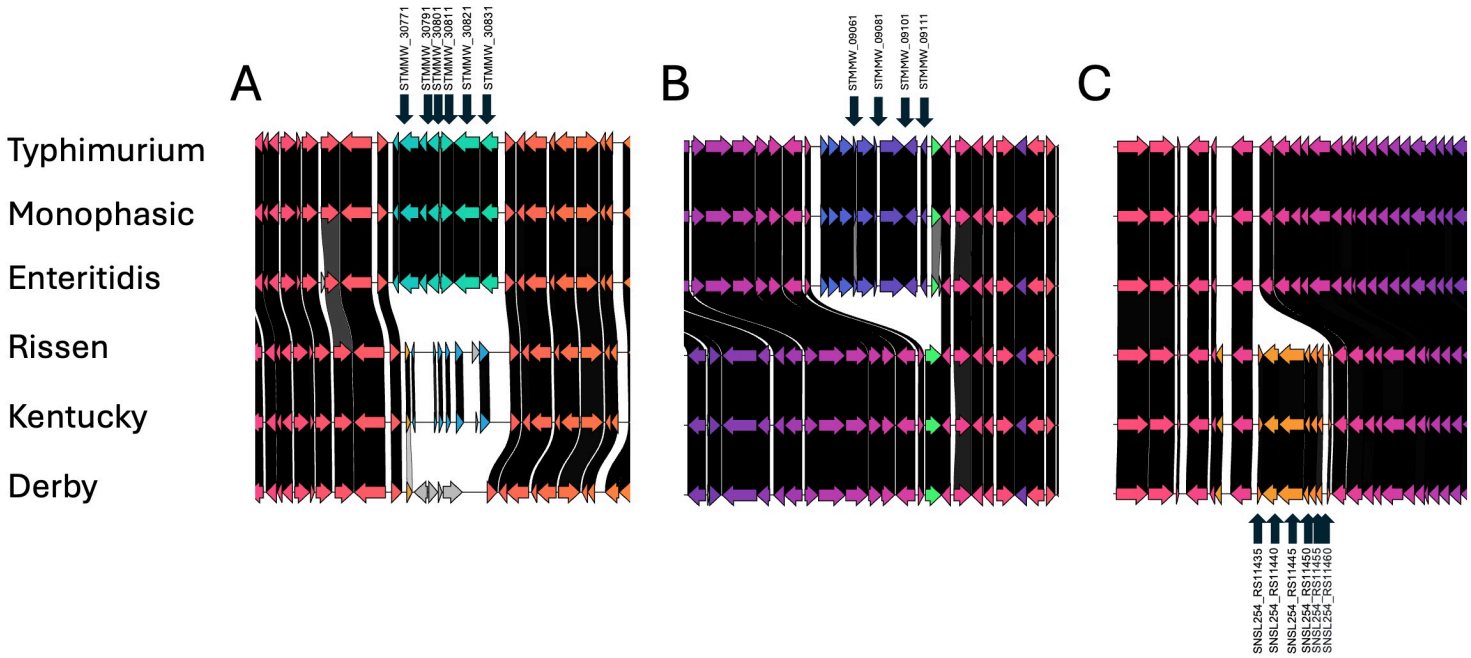


Figure 4

Genomic context of partial SPI-13, SPI-14 the arsenic resistance operon in virulent and attenuated serovars. CDSs are shown as arrows; shared genes are colored, and homologous regions are connected by shaded ribbons proportional to nucleotide identity. SPI-13 (a) and SPI-14 (b) are present only in virulent serovars, whereas the arsenic resistance operon (c) shows the opposite distribution. Operon context was verified in Geneious Prime 2024 and visualized with Clinker 0.0.27.

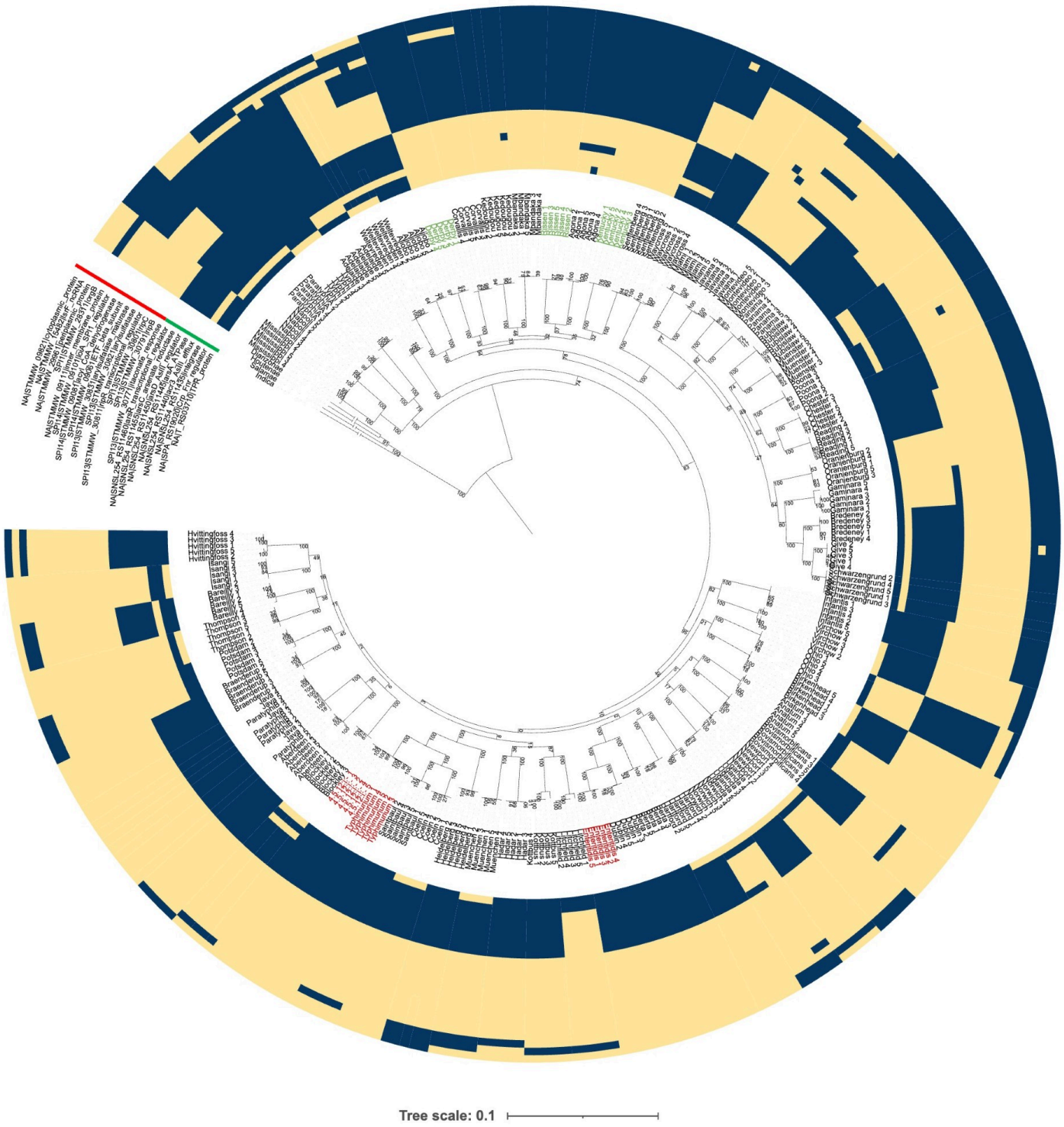


Figure 5

Phylogenetic distribution of the 22-gene set distinguishing virulent and attenuated serovars. Core-SNP maximum-likelihood phylogeny annotated with the 22-gene distribution. Heatmap indicates gene presence (yellow) or absence (blue). Genes of virulent and attenuated serovars are marked by red and green lines, respectively, adjacent to their description. Names of serovars included in the machine learning analysis are shown in green for attenuated serovars and in red for virulent serovars.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTablesscientificreports.xlsx
- SupplementaryFiguresscientificreports.pdf