



OPEN

## Structure classification of the proteins from *Salmonella enterica* pangenome revealed novel potential pathogenicity islands

Kirill E. Medvedev<sup>1,✉</sup>, Jing Zhang<sup>1,2</sup>, R. Dustin Schaeffer<sup>1</sup>, Lisa N. Kinch<sup>3,4</sup>, Qian Cong<sup>1,2</sup> & Nick V. Grishin<sup>1,5</sup>

*Salmonella enterica* is a pathogenic bacterium known for causing severe typhoid fever in humans, making it important to study due to its potential health risks and significant impact on public health. This study provides evolutionary classification of proteins from *Salmonella enterica* pangenome. We classified 17,238 domains from 13,147 proteins from 79,758 *Salmonella enterica* strains and studied in detail domains of 272 proteins from 14 characterized *Salmonella* pathogenicity islands (SPIs). Among SPIs-related proteins, 90 proteins function in the secretion machinery. 41% domains of SPI proteins have no previous sequence annotation. By comparing clinical and environmental isolates, we identified 3682 proteins that are overrepresented in clinical group that we consider as potentially pathogenic. Among domains of potentially pathogenic proteins only 50% domains were annotated by sequence methods previously. Moreover, 36% (1330 out of 3682) of potentially pathogenic proteins cannot be classified into Evolutionary Classification of Protein Domains database (ECOD). Among classified domains of potentially pathogenic proteins the most populated homology groups include helix-turn-helix (HTH), Immunoglobulin-related, and P-loop domains-related. Functional analysis revealed overrepresentation of these protein in biological processes related to viral entry into host cell, antibiotic biosynthesis, DNA metabolism and conformation change, and underrepresentation in translational processes. Analysis of the potentially pathogenic proteins indicates that they form 119 clusters or novel potential pathogenicity islands (NPPIs) within the *Salmonella* genome, suggesting their potential contribution to the bacterium's virulence. One of the NPPIs revealed significant overrepresentation of potentially pathogenic proteins. Overall, our analysis revealed that identified potentially pathogenic proteins are poorly studied.

Salmonellae are gram-negative bacteria that are members of the family *Enterobacteriaceae*. According to the most recent nomenclatural system, there are currently two recognized species: *Salmonella enterica* and *S. bongori*<sup>1</sup>. *S. enterica* comprises six subspecies and accounts for roughly 2000 of the 2600 serovars in the genus<sup>2</sup>. *Salmonella* species can be categorized into typhoid and non-typhoid based on their capacity to induce particular pathologies. Non-typhoid serovars (Typhimurium, Enteritidis, etc.) can cause gastroenteritis and infect humans and other animals<sup>3</sup>. These serovars are primarily spread through animal-derived foods like beef, pork, poultry, and raw eggs that may be contaminated<sup>4</sup> and usually cause mild symptoms that do not require antibiotic treatment<sup>3</sup>. Typhoid serovars (*Typhi*, *Sendai*, *Paratyphi*) have a limited host range, being highly adapted to humans and having only higher primates and humans as their hosts<sup>4</sup>. These serovars lead to the development of typhoid fever and are associated with various symptoms, including high fever, diarrhea, vomiting, headaches, and, in severe cases, death<sup>3</sup>. It is estimated that there are over 20 million cases of typhoid fever annually, leading to more than 200,000 deaths<sup>5</sup>. These serovars can be transmitted through various means, including water, milk, raw vegetables, seafood, and eggs that have been contaminated. For example, *S. Typhi* and *S. Paratyphi* do not

<sup>1</sup>Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA. <sup>2</sup>Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA. <sup>3</sup>Department of Molecular Biology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA. <sup>4</sup>Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA. <sup>5</sup>Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA. ✉email: Kirill.Medvedev@UTSouthwestern.edu

have animal reservoirs apart from higher primates, their presence suggests contamination resulting from poor hygiene practices during the preparation and handling of food and water<sup>6</sup>, and is still a major health problem, especially in the developing world with poor sanitation<sup>7</sup>. Moreover, it was shown that *S. Typhi* infection increases the risk of gallbladder cancer development<sup>8</sup>.

Bacterial secretion systems are intricate molecular machinery found in various bacteria that facilitate the transport of proteins, toxins, and other molecules across the bacterial cell envelope<sup>9</sup>. These systems are crucial for the survival, virulence, and adaptation of bacteria in various environments. There are multiple types of bacterial secretion systems, and they can be broadly categorized into several classes, such as T1SS, T2SS, T3SS, etc.<sup>9</sup>. These systems play a crucial role in bacterial interactions with their environment, including interactions with host cells and other bacteria<sup>9</sup>. Some bacterial secretion systems are associated with virulence and contribute to the bacterial pathogenicity<sup>10</sup>. The proteins responsible for the functioning of bacterial secretion systems are typically found within pathogenicity islands, along with other proteins that enhance the virulence of the bacteria<sup>9</sup>. Pathogenicity islands are a distinct category of genomic islands acquired by microorganisms through horizontal gene transfer<sup>11</sup>. These islands are substantial genomic regions, ranging from 10 to 200 kilobases in size, that become integrated into the genome of pathogenic bacteria and are absent in non-pathogenic bacteria of the same or closely related species<sup>11</sup>. As of today, there are 17 known *Salmonella* pathogenicity islands, some of which are poorly studied<sup>12</sup>. Understanding secretion systems and the function of pathogenicity-related proteins, as well as the identification of other proteins that contribute to bacterial virulence is essential to develop targeted strategies for combating bacterial infections and diseases.

Recent advancements in protein structure prediction have demonstrated that deep learning approaches like AlphaFold can attain atomic-resolution outcomes, even for such rapidly evolving proteins as bacterial proteins<sup>13</sup>. The structure generated by these structure prediction methods can offer insights into distant homologs that may have lost their sequence signals of homology, thereby providing clues to potential protein functions<sup>14,15</sup>. The Evolutionary Classification Of protein Domains (ECOD) database<sup>16,17</sup> provides a valuable resource for analyzing predicted protein structures by placing them in an evolutionary context with experimentally determined protein structures from the Protein Data Bank. ECOD stands as a hierarchical evolutionary classification system that distinguishes itself from other structure-based domain classifications by primarily grouping domains based on homology rather than topology. A significant characteristic of ECOD is its focus on distant homology, culminating in a comprehensive catalog of evolutionary relationships among the categorized domains<sup>16</sup>.

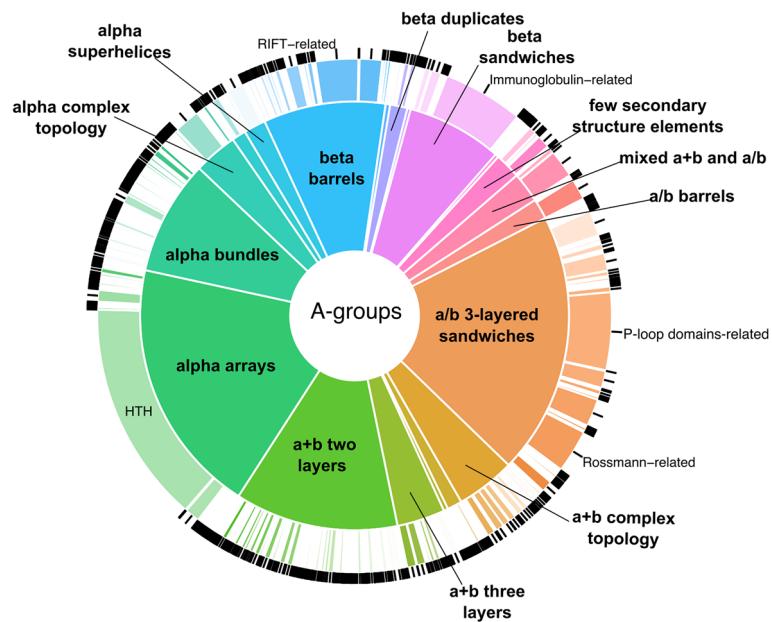
Here we provide an evolutionary classification of protein domains from *Salmonella enterica* pangenome. Using Domain Parser for AlphaFold Models (DPAM) we identified 25,233 domains from 13,147 proteins (protein clusters' representatives). We evaluated domains of 272 proteins from 14 SPIs and studied their ECOD group distribution based on their relation to the secretion machinery. Moreover, we identified 3,682 potentially pathogenic proteins using a comparison of clinical and environmental isolates. The analysis of these proteins indicates that they form 119 clusters or islands within the *Salmonella* genome, suggesting their potential contribution to the bacterium's virulence. Functional analysis revealed that these proteins are significantly overrepresented in biological processes related to host cell invasion. ECOD hierarchy of classified *Salmonella enterica* domains is available online: [http://prodata.swmed.edu/ecod/index\\_salm.php](http://prodata.swmed.edu/ecod/index_salm.php).

## Results and discussion

### Evolutionary classification of *Salmonella enterica* AlphaFold models

Domains parsing and classification AlphaFold (AF) models of proteins from *Salmonella enterica* pangenome into Evolutionary Classification Of protein Domains (ECOD)<sup>16,17</sup> hierarchical groups were conducted using Domain Parser for AlphaFold Models (DPAM)<sup>18</sup> recently developed by our group. Overall, DPAM identified 25,233 domains from 13,147 proteins, which are representatives of protein clusters (see "Materials and methods"). 17,238 (68%) domains have been assigned to the existing groups in the ECOD hierarchy. The unassigned set consists of 7995 domains with a low DPAM probability (< 0.85). These domains can either be categorized within the current ECOD hierarchy through manual assessment or may represent new folds that have not been documented in ECOD. ECOD hierarchy of classified *Salmonella enterica* domains is available as a public database, which also provides a downloadable format of the data ([http://prodata.swmed.edu/ecod/index\\_salm.php](http://prodata.swmed.edu/ecod/index_salm.php)).

Figure 1 illustrates the distribution of architecture (A-groups) and homology (H-groups) ECOD levels for assigned domains from *Salmonella enterica* pangenome. The most prevalent architecture is a/b three-layered sandwiches, constituting 20% of *Salmonella* domains. This percentage is higher than the fraction (18%) found among classified ECOD domains from experimentally determined structures in the Protein Data Bank (Supplementary Table S1). This A-group includes domains from the proteins that contain Rossmann-like motifs whose function is mostly related to metabolism<sup>19,20</sup>. The most populated H-group (P-loop domains-related) within this architecture includes ATPases that are part of bacterial secretion systems and play an important role during the injection of bacterial effector proteins<sup>21</sup>. The second most populated A-group, alpha arrays, is the most overrepresented in comparison to ECOD domains from experimental structures. The *Salmonella* set fraction of this architecture (18.7%) is significantly higher than that of the PDB set (9.1%) (Supplementary Table S1). Helix-turn-helix (HTH) domains comprise the majority of alpha arrays A-group (Fig. 1). In bacteria, proteins containing HTH motifs often serve as transcription factors<sup>22</sup>. The "few secondary structure elements" and "extended segments" A-groups are the most significantly underrepresented. These groups contain domains with small and flexible structures that are difficult to classify, contain a limited secondary structure, and might include signal peptides<sup>23</sup>.



**Figure 1.** *Salmonella enterica* proteins statistics. Distribution of ECOD architecture (inner pie chart) and homology (outer pie chart) groups for assigned domains of AF models from *Salmonella enterica* pangenome. The size of the sections in the pie chart corresponds to the number of assigned domains belonging to specific A- and H-groups. The names of A-groups are shown in bold. HTH stands for helix-turn-helix H-group.

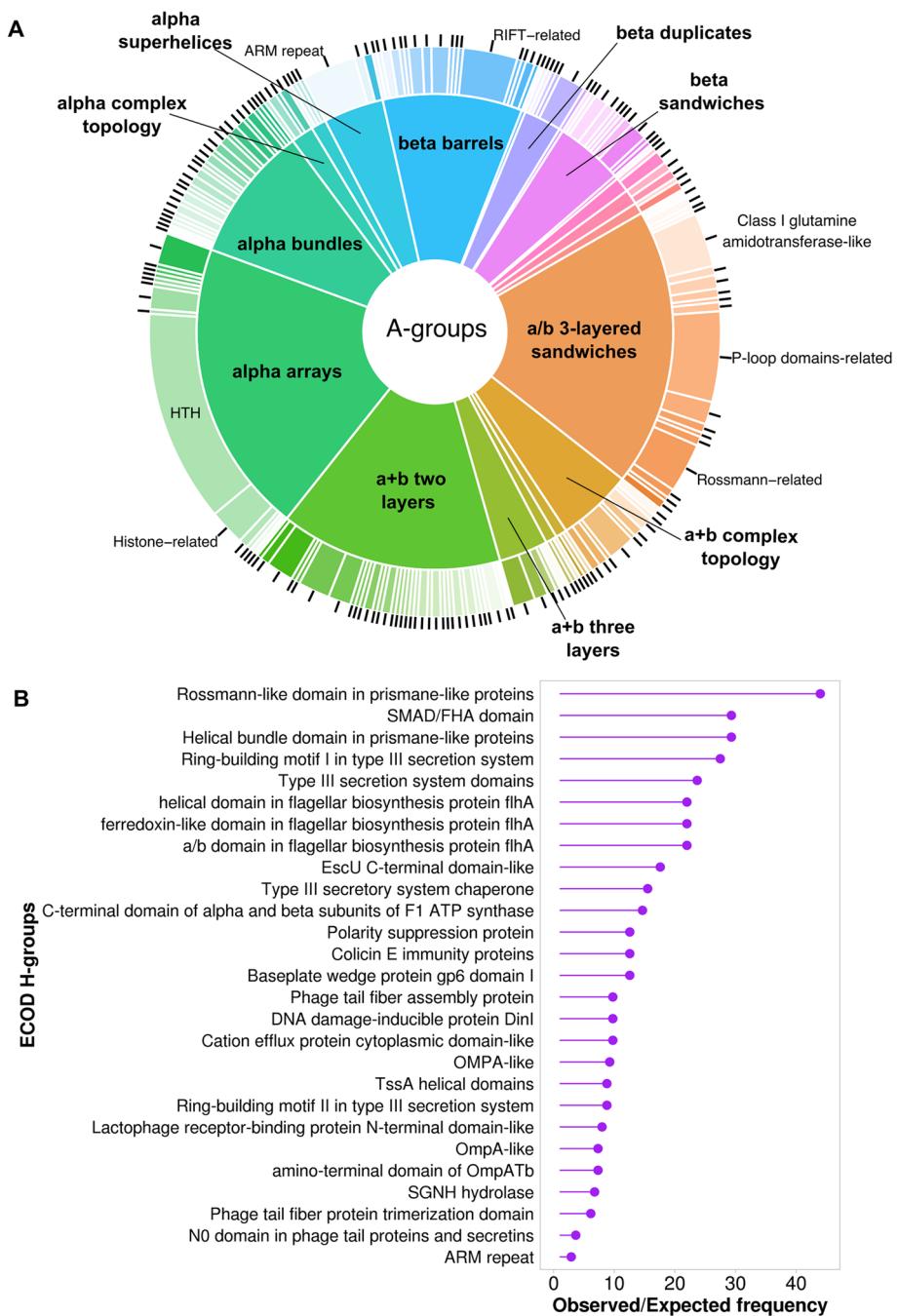
### Domain classification of *Salmonella* pathogenicity islands (SPIs) related proteins

Using published studies, we created a list of protein-coding genes that are part of *Salmonella* pathogenicity islands (SPIs). Overall, 272 protein-coding genes from SPI-1 to SPI-14 were identified (Supplementary Table S2). SPIs-related proteins were mapped to the representatives of protein clusters retrieved from *Salmonella enterica* pangenome (see “Materials and methods”). 28 out of 272 (10%) SPIs-related proteins have experimentally determined 3D protein structure and additionally, 62% (168 out of 272) have homologous proteins among known 3D structures in RCSB Protein Data Bank<sup>24</sup> based on sequence similarity identified by BLAST<sup>25</sup> (E-value < 0.001) (Supplementary Table S2). Thus, 75 out of 272 (28%) of SPIs-related proteins do not have experimentally determined 3D structures and have no close homologs in the RCSB Protein Data Bank. Using DPAM, 223 out of 272 (82%) SPI-related proteins have been classified into ECOD hierarchy up to topology level (T-group). The major reasons for the unclassified 18% of SPIs-related proteins are the high fraction of disordered regions and protein length exceeding 2000 amino acids.

Figure 2 shows the distribution of architecture (A-groups) and homology (H-groups) levels (Fig. 2A) and significantly overrepresented H-groups among SPIs-related proteins in comparison to the *Salmonella* pangenome (Fig. 2B). 41% domains of SPI proteins have no previous sequence annotation and have been determined using AF structural data. The top 3 most populated ECOD A-groups among SPIs related proteins include alpha arrays, a/b three-layered sandwiches and a + b two layers (Fig. 2A). The largest fraction of alpha arrays constitutes HTH H-group. Domains containing helix-turn-helix motifs (HTH) often play an important role in transcription factors in prokaryotes<sup>22,26</sup>. The most populated SPIs-related H-groups from the a/b three-layered sandwich architecture are P-loop domains-related, Rossmann-related, and Class I glutamine amidotransferase-like. Domains from these H-groups contain minimal Rossmann-like motif (RLM). As we showed previously Rossmann folds are present in ancient domains that have often undergone significant divergence, resulting in a wide array of functions, the execution of various enzymatic reactions, and might constitute up to 30% of the bacterial proteome<sup>19,20</sup>. Among the significantly overrepresented SPIs-related H-groups, protein domains associated with the type III secretion system (T3SS) are particularly notable (Ring-building motif I in type III secretion system, Type III secretion system domains, Type III secretory system chaperone, etc.) (Fig. 2B). In *Salmonella* T3SS is encoded by SPI-1 and SPI-2, which are the most conserved among *Salmonella* serovars<sup>27,28</sup>.

Based on published studies we divided SPIs-related proteins into three groups: involved in secretion machinery, not involved in secretion machinery, and unknown function (Supplementary Table S2).

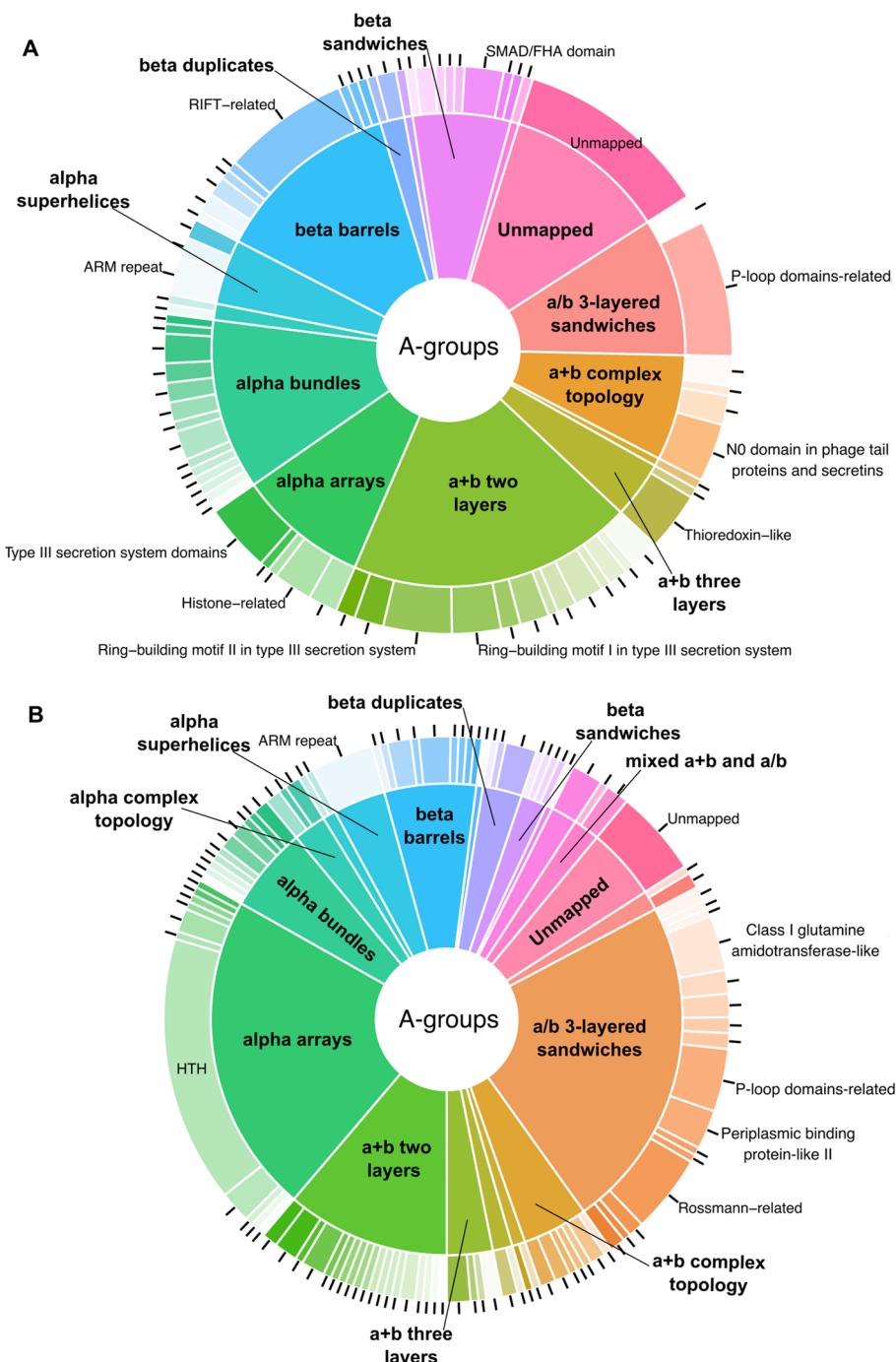
The distribution of ECOD groups for proteins involved in the secretion machinery revealed that A-groups a + b two layers, beta barrels, and alpha bundles have the largest fraction (Fig. 3A). The most populated H-groups from a + b two layers architecture are Ring-building motifs I and II in T3SS. Lipoprotein PrgK (UniProt ID: P41786) forms the ring of the injectisome needle complex and contains both these domains<sup>29</sup>. Surface presentation of antigens protein SpaO (UniProt ID: Q56022) adopts two RIFT-related domains from beta barrels architecture. This protein was shown to be the part of sorting platform of T3SS that is responsible for the sequential delivery of the secreted proteins<sup>30,31</sup>. Intracellular multiplication protein F (IcmF) is an inner membrane protein that forms the core of the type VI secretion system (T6SS)<sup>32</sup>. This protein is expressed from SPI-6<sup>33</sup> and contains



**Figure 2.** SPIs-related proteins statistics. (A) Distribution of ECOD A- (inner pie chart) and H-groups (outer pie chart) for SPIs-related proteins. (B) The ratio of observed and expected frequencies of SPIs-related protein domains defines significantly overrepresented (ratio > 1) ECOD H-groups.

several structural domains, one of which adopts alpha bundles architecture and is a part of Type VI secretion protein IcmF helical domain ECOD H-group. The distribution of ECOD groups for proteins not involved in the secretion machinery (Fig. 3B) revealed a similarity to the overall distribution of ECOD groups for SPIs-related proteins (Fig. 2A) with a/b three-layered sandwiches, alpha arrays and a + b two layers being the most populated architectures. The A-group a/b three-layered sandwiches also stand out for proteins related to the secretion machinery (Fig. 3A), as this architecture is predominantly represented by a single H-group (P-loop domains-related).

SPI proteins that are associated with the secretion machinery and contain P-loop domains serve as ATPases (for example invasion protein InvC, EC: 7.4.2.8, UniProt ID: P0A1B9). These ATPases are integral components of the T3SS and are used to inject bacterial effector proteins into eukaryotic host cells<sup>21</sup>.



**Figure 3.** Comparison of SPIs-related proteins involved in secretion machinery functioning vs all rest with known function. (A) Distribution of ECOD A- (inner pie chart) and H-groups (outer pie chart) for SPIs-related proteins involved in secretion machinery. (B) Distribution of ECOD A- (inner pie chart) and H-groups (outer pie chart) for SPIs-related proteins with known functions not involved in secretion machinery.

Among all SPIs related proteins only 78% (212 out of 272) contain mapping to the Pfam database<sup>34</sup> (Supplementary Table S2). Our analysis using HHpred<sup>35</sup> showed that some of the 22% unmapped proteins contain domains that have confident hits in the Pfam database. For example, secretion system apparatus protein SsaP (UniProt: P0A239), which is part of the T3SS secretion machinery, showed a similarity with two Pfam families: PF02120—Flagellar hook-length control protein FliK (Probability: 97.88%, E-value: 2.0E-4) and PF09483—Type III secretion protein (HpaP) (Probability: 97.84%, E-value: 9.5E-5). The inner membrane protein FidL (UniProt: H9L432) showed similarity to PF15941—FidL-like putative membrane protein (Probability: 99.78%, E-value: 1.8E-18) and pilus assembly protein PilP (UniProt: Q9RHF2) showed similarity to PF17456—Toxin-coregulated pilus protein S (Probability: 99.79%, E-value: 3.5E-18). Among all SPIs-related proteins, 36% (99 out of 272)

don't have Gene Ontology (GO) assignments according to UniProt KB (Supplementary Table S2) and their function is often unclear. For example, the Inner membrane protein (UniProt: H9L4B1, gene: orf319), which is located within SPI-2<sup>27</sup>, has no GO assignment and its function is vague. The N-terminal domain of H9L4B1 adopts a Flavodoxin-like structure that belongs to the Class I glutamine amidotransferase-like homology group (Fig. 4A). Structure searches against Protein Data Bank using FoldSeek<sup>36</sup> revealed [4Fe-4S]-dependent thiouracil desulfidase TudS from *E. coli* (PDB ID: 6ZW9) as the most similar (Fig. 4B). This enzyme catalyzes desulfuration of thiouracil to uracil and harbors a [4Fe-4S] cluster bound by three cysteines colored in magenta (Fig. 4B)<sup>37</sup>. H9L4B1 contains three cysteines at the same positions as TudS and one additional cysteine at the C-terminal part of the third beta strand (Fig. 4A). This suggests that H9L4B1 is also capable of binding iron-sulfur clusters and might have potential enzymatic activity. Moreover, both these domains are members of the Pfam family PF04463 (2-thiouracil desulfurase). Another example is the ParB family protein (UniProt: Q8Z1M1) whose function is also unclear. ParB structure includes three domains, two of which have been classified by DPAM. The N-terminal domain, shown in light blue (Fig. 4C), adopts a + b complex topology and is affiliated with the ParB/Sulfiredoxin homology group. The middle domains, shown in light orange, adopt an alpha arrays architecture and belong to the Histone-related H-group. For the C-terminal domain, depicted in red (Fig. 4C), DPAM did not identify a confident hit in ECOD. FoldSeek search against AFDB using the C-terminal domain of Q8Z1M1 showed many bacterial proteins with this domain. For example, Glutamate-1-semialdehyde 2,1-aminomutase from *Nocardia brasiliensis* (UniProt: K0EV93) (Fig. 4D). This protein contains two domains: C-terminal which is similar to the C-terminal domain of ParB, and N-terminal, which belongs to Pfam family Aminotransferase class-III (PF00202). Several known experimental protein structures belong to PF00202, one of them is dialkylglycine decarboxylase from *Burkholderia cepacia* (PDB: 1D7R) (Fig. 4E). The C-terminal domain of this protein belongs to C-terminal domain in some PLP-dependent transferases ECOD H-group. Therefore, C-terminal domains of ParB, glutamate-1-semialdehyde 2,1-aminomutase, and dialkyl glycine decarboxylase are homologous, however, our classification revealed topological differences—decarboxylase contains an additional beta hairpin shown in magenta (Fig. 4F-H).

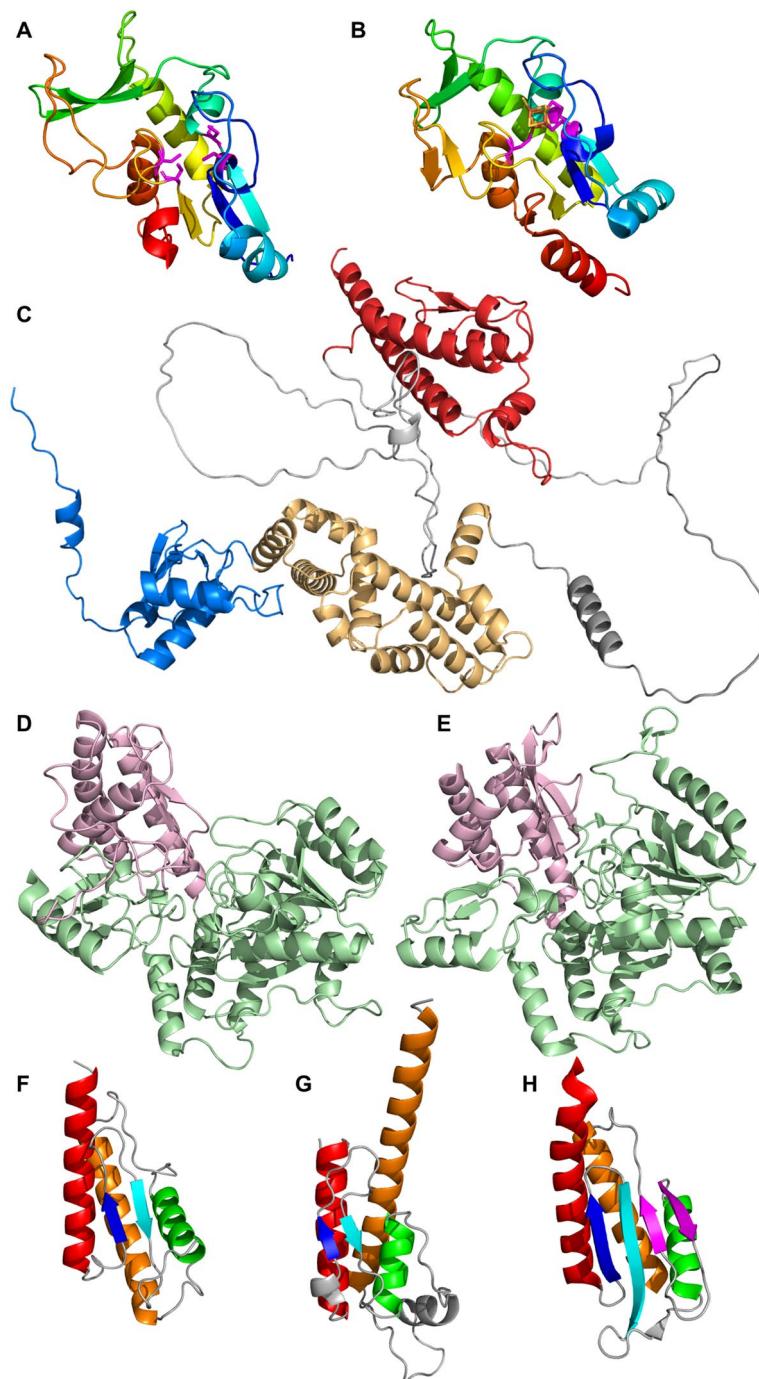
#### Potentially pathogenic proteins revealed similarity in ECOD group distribution with SPIs-related proteins

We defined potentially pathogenic proteins in *Salmonella enterica* pangenome using a comparison of patients and all other types of isolates (food and environment). The least conserved proteins between these two groups were considered potentially pathogenic. The identified proteins were clustered to remove redundancy (see “[Materials and methods](#)”) which resulted in 3,682 clusters of potentially pathogenic proteins. Representative proteins from these clusters were used for further analysis. To verify the enrichment of known virulence-related proteins among potentially pathogenic proteins we ran BLAST<sup>25</sup> against the virulence factor database (VFDB)<sup>38</sup> and calculated: (1) observed frequency of virulence-related proteins in the set of identified proteins and (2) expected frequency of virulence-related proteins in *Salmonella enterica* pangenome. The ratio of observed and expected frequencies was 1.282 with a Chi-Square test P-value < 0.00001. It means that virulence-related proteins are significantly overrepresented in the set of identified potentially pathogenic proteins in comparison to the whole pangenome.

We mapped potentially pathogenic proteins to the UniProt database and retrieved Gene Ontology (GO) annotations. Only 45% of these proteins have GO annotations. We calculated over and underrepresentation of biological process categories from the GO Metagenomic subset for potentially pathogenic proteins. Our analysis revealed that these proteins are overrepresented in biological processes related to viral entry into host cell, antibiotic biosynthesis, DNA metabolism and conformation change, and underrepresentation in translational processes (Fig. 5).

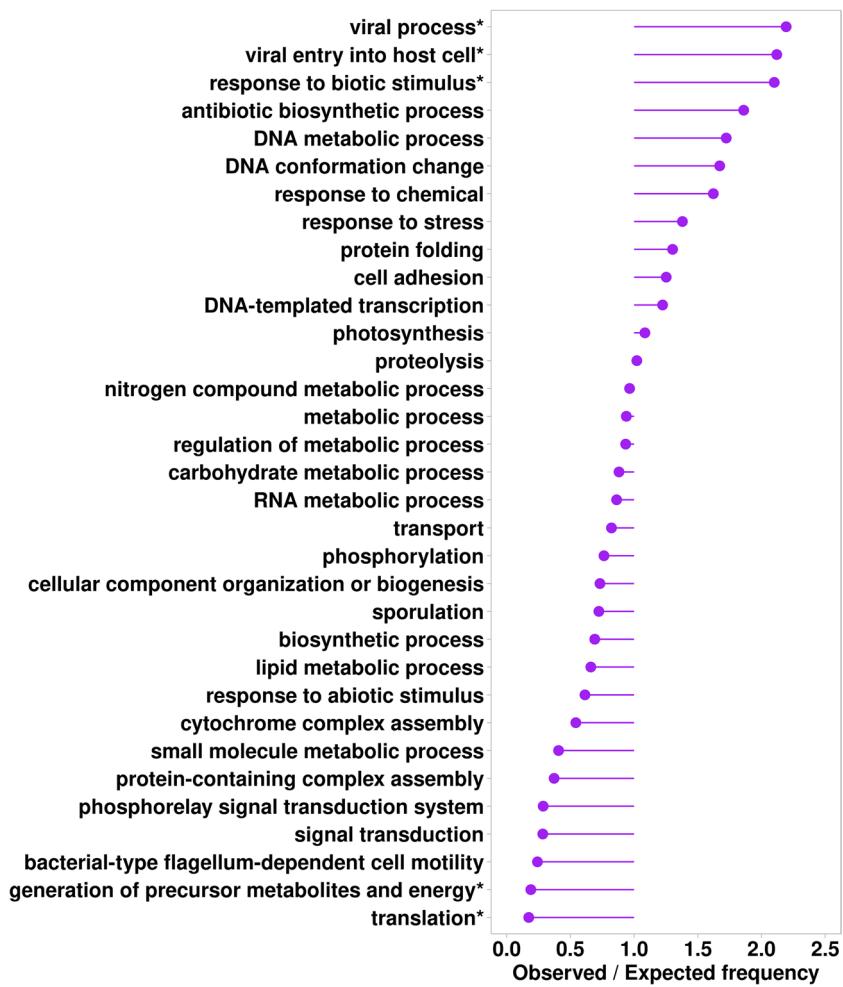
Distribution of ECOD architecture and homology groups of potentially pathogenic proteins (Fig. 6A) showed significant similarity with one of the SPIs-related proteins (Fig. 2A). Interestingly, 50% of these proteins' domains have not been annotated previously using sequence-based approaches in the InterPro database<sup>39</sup>. In both cases the dominating A-groups are alpha arrays, a/b three-layered sandwiches, a + b two layers, alpha bundles, and beta barrels. We consider this similarity as further confirmation of the virulence association of identified potentially pathogenic proteins. The most notable difference in the distribution of ECOD groups between SPIs-related and potentially pathogenic proteins is the significantly higher proportion of Immunoglobulin-related (beta sandwiches architecture) and Restriction endonuclease-like (a/b three-layered sandwiches architecture) H-groups in potential pathogenic proteins. These two H-groups are among significantly overrepresented groups in potential pathogenic proteins in comparison to the whole pangenome (Fig. 6B). Restriction endonuclease-like homology group is represented by single ECOD T-group (ECOD ID: 2008.1.1). Restriction endonucleases or restrictionases are bacterial and archaeal enzymes that cleave foreign DNA (for example bacteriophages) into fragments as a defense mechanism against invading viruses<sup>40</sup>. Immunoglobulin-related homology group is represented by three T-groups in our dataset. T-group Immunoglobulin/Fibronectin type III (ECOD ID: 11.1.1) includes domains from such *Salmonella* proteins as alpha-2-macroglobulin (PDB: 4U48), which is a plasma membrane protein and protects the bacterial cell from host peptidases<sup>41</sup>. The second T-group Prealbumin-like (ECOD ID: 11.1.4) includes such proteins with experimentally defined structures as lipoprotein TssJ from *Pseudomonas aeruginosa* (PDB: 3ZHN) which is an outer-membrane-associated lipoprotein and is crucial for the function of T6SS<sup>42</sup>. Finally, the third T-group Common fold of diphtheria toxin (ECOD ID: 11.1.5) includes *Salmonella* fimbrial protein SefD with experimentally defined structure (PDB: 3UIY). It was shown that SefD is required for uptake or survival in macrophages after infection<sup>43</sup>. Therefore, the examples discussed above collectively represent groups of proteins that can contribute to *Salmonella* pathogenicity.

For representatives of each cluster from the whole *Salmonella enterica* pangenome we calculated the average pair-wise distance and using Multidimensional Scaling (MDS) we converted the pair-wise distance matrix to



**Figure 4.** Representative domain structures of SPIs-related proteins. (A) N-terminal domain of Inner membrane protein (UniProt: H9L4B1, gene: orf319). Cysteines are colored magenta. (B) [4Fe-4S]-dependent thiouracil desulfidase TudS from *E. coli* (PDB: 6ZW9). Cysteines are colored magenta. (C) ParB family protein structure (UniProt: Q8Z1M1). (D) Glutamate-1-semialdehyde 2,1-aminomutase from *Nocardia brasiliensis* (UniProt: K0EV93). The N-terminal domain is shown in light green, C-terminal in light pink. (E) Dialkylglycine decarboxylase from *Burkholderia cepacia* (PDB: 1D7R). The N-terminal domain is shown in light green, C-terminal in light pink. (F) C-terminal domain of Glutamate-1-semialdehyde 2,1-aminomutase (UniProt: K0EV93). (G) C-terminal domain of ParB family protein (UniProt: Q8Z1M1). (H) C-terminal domain of dialkylglycine decarboxylase (PDB: 1D7R).

one dimension. By sorting representatives by MDS value we obtained the list of protein-coding genes, in which adjacent genes should also be adjacent in the genome. Using chosen criteria (see “Materials and methods”)



**Figure 5.** The ratio of observed and expected frequencies of biological processes from the GO Metagenomics subset defines over (ratio > 1) and under (ratio < 1) represented biological process categories for potentially pathogenic proteins. Asterisks denote significant values according to the chi-square test ( $P < 0.0001$ ).

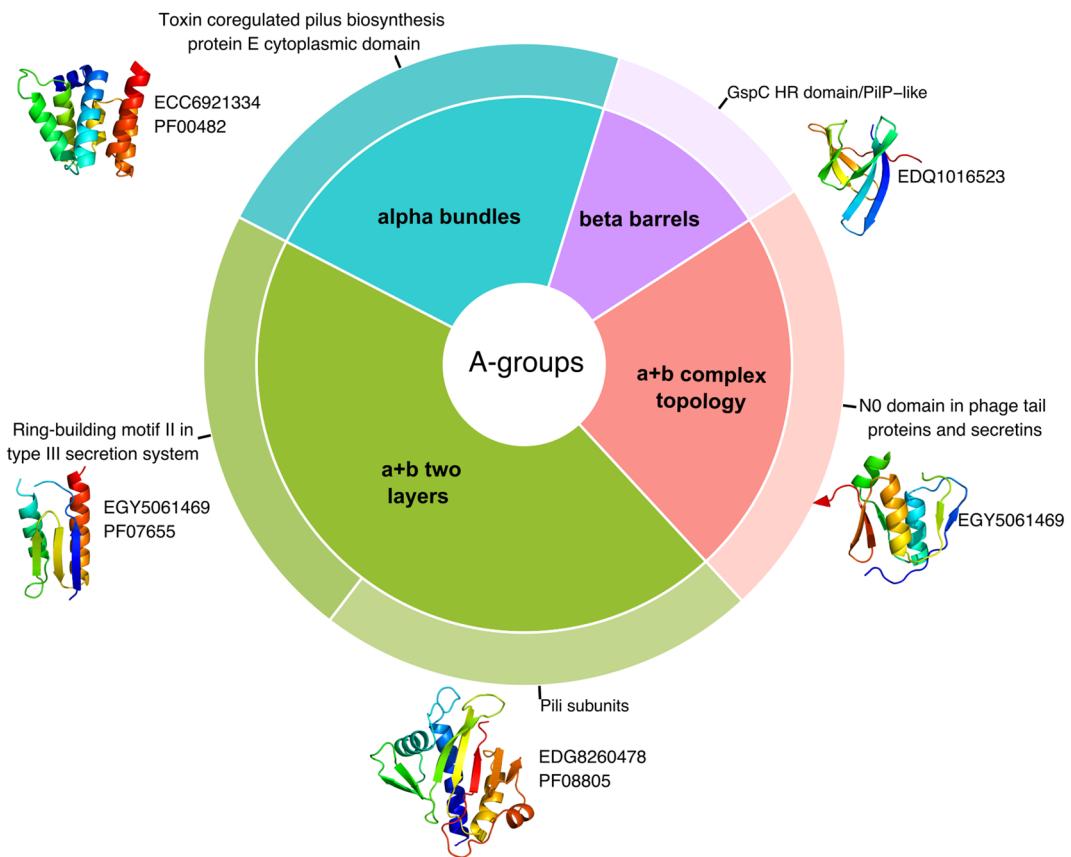
we identified 119 novel potential pathogenicity islands (NPPIs) (Supplementary File 1). One of the NPPIs (NPPI99) revealed significant overrepresentation (ratio of observed and expected frequencies = 6.15; Chi-Square P-value = 3.77E-08) of potentially pathogenic proteins with confident hits within VFDB. Distribution of ECOD groups of potentially pathogenic proteins within NPPI99 revealed the dominance of homology groups related to T3SS and pilus elements (Fig. 7).

However, the majority of the identified NPPIs do not feature many proteins with confident hits within VFDB and include a significant number of proteins with unknown functions. An example of this is NPPI7, which contains 50% (9 out of 18) of potentially pathogenic proteins with unclear functions (Supplementary Table S3). The distribution of ECOD categories within NPPI7 revealed the prevalence of alpha arrays, beta barrels, and a + b two-layer architectures (Fig. 8A). The most populated A-group alpha arrays are represented by a single H-group—HTH. Domains containing helix-turn-helix motif (HTH) are known to be part of proteins that often function as transcription factors<sup>22</sup>. Indeed, one of the proteins from NPPI7 that consist of one HTH domain is a helix-turn-helix transcriptional regulator (NCBI: ECO0783223.1) (Fig. 8B). Hypothetical protein K5A74\_004162 with unknown function (NCBI: EIB4898420.1) adopts two HTH domains (Fig. 8C). The best HHpred hit against Pfam database (Probability: 97.79%, E-value: 0.000026) revealed Domain of unknown function (DUF1870, PF08965) for both domains, which is part of HTH Pfam clan. BLAST run against UniProt KB revealed the set of nearly identical bacterial proteins XRE family transcriptional regulator (Identity: 97.5%, E-value: 1.9e-105). Therefore, the hypothetical protein K5A74\_004162 is likely a DNA-binding protein involved in transcription regulation. Hypothetical protein AAK88\_22130 (NCBI: ECO0783211.1) adopts a + b two layers topology and is assigned to Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase H-group (Fig. 8D). Our analysis showed that this protein belongs to Pfam family DUF5983 (Family of unknown function, PF19419). Its closest homolog identified by DPAM is Toxoflavin Lyase (TflA) from the bacteria *Paenibacillus polymyxa* (PDB: 3PKX). TflA is responsible for the degradation of the toxin toxoflavin<sup>44</sup>. The residues that take part in toxoflavin binding are shown in magenta (Fig. 8E). Hypothetical protein AAK88\_22130 contains four residues (two tryptophans, one arginine, and one aspartic acid) that might coordinate a small molecule in a similar position



**Figure 6.** Potentially pathogenic proteins statistics. **(A)** Distribution of ECOD A- (inner pie chart) and H-groups (outer pie chart) for potentially pathogenic proteins. **(B)** The ratio of observed and expected frequencies of potentially pathogenic protein domains defines significantly overrepresented (ratio > 1) and underrepresented (ratio < 1) ECOD H-groups.

to how TfIA coordinates toxoflavin (Fig. 8D, E). Hypothetical protein QD83\_004322 (NCBI: EDV2320831.1) adopts beta barrels architecture and is assigned to the SH3 H-group. This protein includes one beta barrels domain and a single alpha helix at the N-terminal end (Fig. 8F). The closest homolog identified by DPAM is bacterial translation initiation factor 5A (IF-5A, PDB: 1KKB) (Fig. 8G). IF-5A plays a role in the initial stage of peptide bond formation during translation<sup>45</sup>. HHpred run against the Pfam database for the hypothetical protein QD83\_004322 revealed a significant hit (Probability: 96.95%, E-value: 0.00031) only for N-terminal helix–lysis protein (PF02402). This family includes a signal peptide motif and a lipid attachment site. The beta-barrel domain of QD83\_004322 does not have significant hits in the Pfam database. BLAST run against UniProt KB revealed the set of bacterial lipoproteins with 98% sequence identity (E-value: 3.4e-91; UniProt:



**Figure 7.** Novel potential pathogenicity island (NPPI99) proteins statistics. Distribution of ECOD A- (inner pie chart) and H-groups (outer pie chart) for potentially pathogenic proteins within NPPI99. NCBI and Pfam IDs are shown near domain structures.

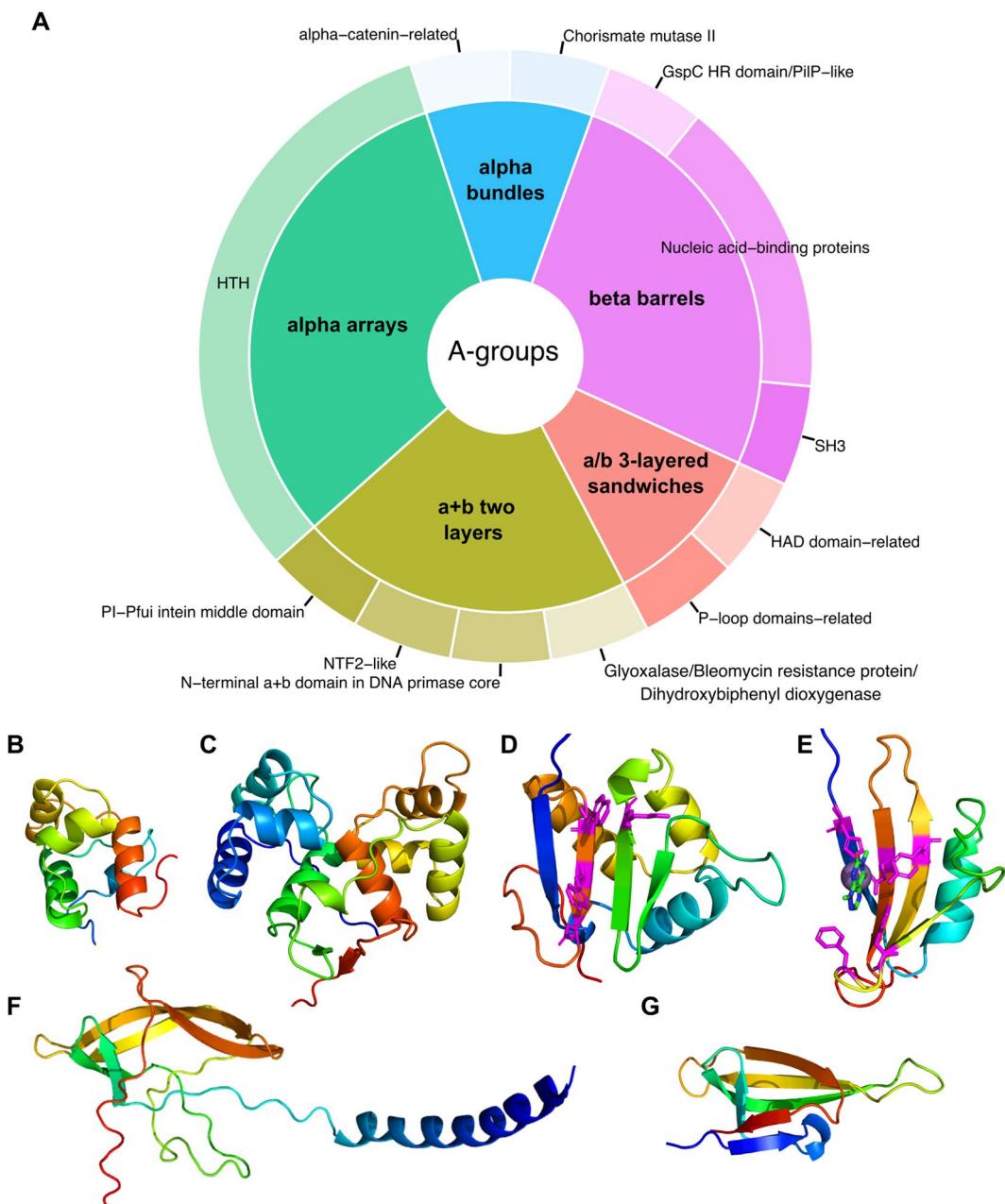
A0A626QDK6). Thus, the hypothetical protein QD83\_004322 is likely a lipoprotein with a lipid attachment site as defined by the Pfam family.

Antibiotics stand out as one of the most effective, successful, and cost-effective classes of drugs for treating bacterial diseases<sup>46</sup>. However, the unregulated usage of antibiotics has resulted in the emergence of multidrug-resistant pathogens, prompting an immediate necessity to devise new therapies. Significant advancements have been made in understanding various mechanisms underlying antibiotic action and the primary mechanisms through which bacteria can develop resistance to the inhibitory or lethal effects of antibiotics<sup>47</sup>. One of the approaches to mitigate the proliferation of multidrug-resistant strains and reduce the spread of resistance determinants involves targeting virulence factors with small molecules<sup>9</sup>. Understanding the structural insights of virulence-promoting proteins is crucial for the development of new targeted therapies using small molecules. In this study, we identified novel potentially pathogenic proteins and delineated their structural characteristics, some of which may serve as promising targets for treating *Salmonella* infection.

## Conclusion

In this study we provided evolutionary classification of protein domains from *Salmonella enterica* pangenome. Our classification includes 17,238 domains from 13,147 proteins from 79,758 *Salmonella enterica* strains. Overall, the most prevalent architecture groups among assigned domains from *Salmonella* pangenome are a/b three-layered sandwiches and alpha arrays. Detailed analysis of 272 proteins from 14 characterized *Salmonella* SPIs revealed that 75 out of 272 (28%) of SPIs-related proteins do not have experimentally determined 3D structures and have no close homologs in the RCSB Protein Data Bank. Moreover, 41% domains of SPI proteins have no previous sequence annotation and have been determined using AlphaFold structural data. The top 3 most populated ECOD A-groups among SPIs related proteins also include alpha arrays, a/b three-layered sandwiches and a + b two layers. We also investigated the subgroup of SPIs-related proteins that are involved in the secretion machinery. The distribution of ECOD groups for proteins involved in the secretion machinery revealed that A-groups a + b two layers, beta barrels, and alpha bundles have the largest fraction.

By comparison of patients and non-patients isolates we defined 3,682 clusters of potentially pathogenic proteins. Functional analysis using Gene Ontology showed that these proteins are overrepresented in biological processes related to viral entry into host cell, antibiotic biosynthesis, DNA metabolism and conformation change, and underrepresentation in translational processes. Distribution of ECOD architecture and homology groups of potentially pathogenic proteins showed significant similarity with one of the SPIs-related proteins. Moreover, 50%



**Figure 8.** Novel potential pathogenicity island (NPPI7) proteins statistics. (A) Distribution of ECOD A- (inner pie chart) and H-groups (outer pie chart) for potentially pathogenic proteins within NPPI7. (B) AF model of a helix-turn-helix transcriptional regulator (NCBI: ECO0783223.1). (C) AF model of hypothetical protein K5A74\_004162 (NCBI: EIB4898420.1). (D) AF model of hypothetical protein AAK88\_22130 (NCBI: ECO0783211.1). Residues that potentially can coordinate small molecules are shown in sticks and colored magenta. (E) Domain of toxoflavin lyase from *Paenibacillus polymyxa* (PDB: 3PKX). Residues that bind toxoflavin are colored in magenta. (F) AF model of hypothetical protein QD83\_004322 (NCBI: EDV2320831.1). (G) Structure of translation initiation factor 5A (PDB: 1BKB).

of assigned domains of potentially pathogenic proteins have not been annotated previously using sequence-based approaches. Using the average pair-wise distance for genes corresponded to potentially pathogenic proteins, we identified 119 novel potential pathogenicity islands (NPPIs). While the majority of NPPIs include a significant number of proteins with unknown functions, NPPI99 showed significant overrepresentation of potentially pathogenic proteins. Distribution of ECOD groups of potentially pathogenic proteins within NPPI99 revealed the prevalence of homology groups related to T3SS and pilus elements. The identification of potentially pathogenic proteins and their structural insights may pave the way for the development of novel treatments for *Salmonella* infection.

## Materials and methods

### Identification of potentially pathogenic proteins in *Salmonella* pangenome

We obtained metadata for 450,571 *Salmonella enterica* genomes from NCBI, most of which lack specific details such as isolation\_source, host, host\_health\_state, and host\_disease. 79,758 out of 450,571 genomes have gene annotation and can be confidently categorized based on their metadata into three groups: 14,903 from diseased patients (disease); 55,931 from food, food packages, food process facilities, and other food-related facilities or items (food); 8924 from the environment such as sewage, pond, and soil (environment). We hypothesized that proteins predominantly found in patient isolates but less so in food and environment might be virulence-associated. Thus, we initially clustered proteins encoded by 79,758 *Salmonella enterica* strains using easy-cluster in mmseq2 with identity 50% and coverage 0.8 and aim to identify protein clusters enriched with proteins from the disease group versus the food and environment ones. Recognizing the presence of redundant isolates could skew the enrichment analysis, we intended to remove proteomes highly similar. The similarity of two proteomes was assessed by tallying the proteins in one proteome that coexisted in the same clusters as proteins from the other proteome and then calculating its ratio by dividing it by the total number of proteins in the proteome. Several cutoffs, namely 0.999, 0.998, 0.995, 0.99, 0.98, 0.97, 0.96, and 0.95, were used to remove redundancy. At each cutoff, after removing the redundant proteomes, we identified clusters enriched with proteins from the disease group against the food and environment groups respectively using a Fisher exact test, with a significance level of Holm-Bonferroni adjusted  $p < 0.05$ .

### Domain parsing and classification of the *Salmonella* proteins

We obtained 54,797 clusters from 79,758 proteomes by easy-cluster in mmseq2 with identity 50% and coverage 0.8, of which 11,126 clusters contained proteins from no less than 80 proteomes which is around 0.1% of total proteomes. For obtaining structures of protein clusters, we searched against the AlphaFold Database (AFDB)<sup>48</sup> using the representative sequence of each cluster selected by mmseq2 by default with BLASTP in Diamond 2.0.13<sup>49</sup>. 9246 clusters with representative sequences aligned with the AFDB, meeting the criteria of at least 90% identity, above 85% coverage, and a length difference of no more than 31 amino acids. For the 1880 clusters without matches in the AFDB, we generated their structures using AlphaFold<sup>13</sup>, using alignments from HHsearch<sup>50</sup> with parameters -Z 100000 -B 100000. These structures were subsequently segmented and classified into ECOD domains via the domain parser for AlphaFold models (DPAM) with default settings described in<sup>18,51</sup>. Briefly, DPAM relies on ECOD hits found by HHsuite and Dali<sup>50,52</sup>, as well as inter-residue distances and predicted aligned errors (PAE) of AF models to partition structures into compact domains. Each parsed domain was assigned to the ECOD hierarchy by finding an ECOD parent domain showing the highest DPAM probability for being in the same ECOD topology group (T-group) as the query domain. The DPAM probability was calculated based on sequence and structure similarities evaluated by HHsuite and Dali and the consensus between them. Additionally, domains detected by the DPAM pipeline were further evaluated by the number of secondary structure elements (residues predicted as H, G, or I by DSSP<sup>53</sup> were considered a helix, and residues predicted as E or B by DSSP are considered as a strand) and the completeness of a domain. A domain with less than 3 secondary structure elements and without confident (HHsuite probability and coverage 80%) ECOD hits by HHsuite is considered a “simple topology”. A domain that is too short compared to its ECOD parent domain (residues aligned) or the median length of the ECOD T-group (length) it is assigned to is considered a “partial domain”. Domains passing these filters are regarded as “low confidence” and cannot be automatically assigned if their DPAM probabilities are below 0.85, while the rest are considered “good domains” with confident assignments.

### Calculation of over and underrepresentation of ECOD H-groups for SPIs and potentially pathogenic proteins

The list of genes that are located within the following *Salmonella* pathogenicity islands (SPIs) was retrieved from the published studies: SPI-1<sup>27</sup>, SPI-2<sup>27</sup>, SPI-3<sup>54</sup>, SPI-4<sup>55</sup>, SPI-5<sup>56</sup>, SPI-6<sup>33</sup>, SPI-7<sup>57</sup>, SPI-8<sup>58</sup>, SPI-9<sup>59</sup>, SPI-10<sup>60</sup>, SPI-11<sup>61</sup>, SPI-12<sup>62</sup>, SPI-13<sup>63</sup>, SPI-14<sup>63</sup>. Over and underrepresentation of SPIs and potentially pathogenic proteins in ECOD homology groups was calculated as the ratio of observed and expected frequencies. The observed frequency for each ECOD H-group of SPI protein domains was calculated as a ratio of the number of the SPI protein domains assigned to this H-group over the sum of all SPI protein domains assigned to any H-group. The observed frequency for each ECOD H-group of potentially pathogenic protein domains was calculated as a ratio of the number of the potentially pathogenic protein domains assigned to this H-group within the *Salmonella* pangenome over the sum of all potentially pathogenic protein domains assigned to any H-group within the *Salmonella* pangenome. The expected frequency for each ECOD H-group was calculated as the ratio of the total number of domains assigned to this H-group in *Salmonella* pangenome over the total number of domains assigned to any H-group in *Salmonella* pangenome. The significance of the representation was checked using the chi-squared test. P-value  $< 0.01$  was considered significant. Statistical analysis was conducted using the R package, v4.2.1.

### Functional analysis of potentially pathogenic proteins

Gene Ontology (GO)<sup>64</sup> biological processes (BPs) information was retrieved for each protein from UniProt KB<sup>65</sup>. BP GO terms were mapped to GO terms from a Metagenomics slim subset. Overall there are 42 top-level BPs in GO Metagenomics slim subset. One protein can take part in several BPs. Over and underrepresentation of potentially pathogenic proteins in BPs were calculated as the ratio of observed and expected frequencies. The observed frequency in each BP was calculated as a ratio of the total number of the potentially pathogenic proteins in a particular BP over the sum of all potentially pathogenic proteins mapped to any BP. The expected frequency

in each BP was calculated as the ratio of the total number of *Salmonella* proteins found for each particular BP to the total amount of proteins mapped to any BP. The significance of overrepresentation was checked using the Chi-square test (P-value < 0.0001 is considered significant).

### Identification of Novel Potential Pathogenicity Islands (NPPIs)

To get the average distance between protein clusters, we calculated the pair-wise distance of proteins in each assembly by counting the minimum number of protein-coding genes between target proteins. The average pair-wise distance of protein clusters was calculated and Multidimensional Scaling (MDS)<sup>66</sup> was used to convert the pair-wise distance matrix to one-dimensional protein localization. Using MDS values for each protein-coding gene from *Salmonella* pangenome we sorted obtained list by MDS values. Genes that are adjacent in this list should also be adjacent in the genome. We used this sorted list to identify clusters of potentially pathogenic proteins that could form novel potential pathogenicity islands (NPPIs) based on the following criteria: (a) NPPI should contain no less than 5 potentially pathogenic proteins; (b) NPPI should contain no more than 5 nonpathogenic proteins in a row. These criteria enabled us to identify approximately two hundred NPPIs. Verification of NPPIs was conducted using the virulence factor database (VFDB)<sup>38</sup>. For each potentially pathogenic protein, we ran BLAST<sup>25</sup> against the set A of VFDB (4232 proteins). We verified the NPPIs by calculating the over and underrepresentation of potentially pathogenic proteins with significant hits in VFDB within each NPPI, as the ratio of observed to expected frequencies. Any BLAST hit with an E-value < 0.01 was considered significant. The observed frequency for each NPPI was calculated as a ratio of the number of the potentially pathogenic proteins in this NPPI with a significant hit in VFDB over the sum of potentially pathogenic proteins in any NPPI with a significant hit in VFDB. The expected frequency for each NPPI was calculated as the ratio of the total number of potentially pathogenic proteins in this NPPI over the total number of potentially pathogenic proteins in any NPPI. The significance of the representation was checked using the chi-squared test. P-value < 0.01 was considered significant.

### Data availability

All generated data are included in this manuscript and are available online: [http://prodata.swmed.edu/ecod/index\\_salm.php](http://prodata.swmed.edu/ecod/index_salm.php).

Received: 20 February 2024; Accepted: 30 April 2024

Published online: 28 May 2024

### References

1. Tindall, B. J., Grimont, P. A. D., Garrity, G. M. & Euzéby, J. P. Nomenclature and taxonomy of the genus *Salmonella*. *Int. J. Syst. Evol. Microbiol.* **55**, 521–524. <https://doi.org/10.1099/ijts.0.63580-0> (2005).
2. LeLievre, V., Besnard, A., Schlusselhuber, M., Desmarests, N. & Dalmasso, M. Phages for biocontrol in foods: What opportunities for *Salmonella* sp. control along the dairy food chain?. *Food Microbiol.* **78**, 89–98. <https://doi.org/10.1016/j.fm.2018.10.009> (2019).
3. Ferrari, R. G. *et al.* Worldwide epidemiology of salmonella serovars in animal-based foods: A meta-analysis. *Appl. Environ. Microbiol.* **85**, 14. <https://doi.org/10.1128/AEM.00591-19> (2019).
4. Gal-Mor, O., Boyle, E. C. & Grassl, G. A. Same species, different diseases: How and why typhoidal and non-typhoidal *Salmonella enterica* serovars differ. *Front. Microbiol.* **5**, 391. <https://doi.org/10.3389/fmicb.2014.00391> (2014).
5. Johnson, R. *et al.* Comparison of *Salmonella enterica* serovars typhi and typhimurium reveals typhoidal serovar-specific responses to bile. *Infect. Immun.* **86**, 3. <https://doi.org/10.1128/IAI.00490-17> (2018).
6. Newell, D. G. *et al.* Food-borne diseases—the challenges of 20 years ago still persist while new ones continue to emerge. *Int. J. Food Microbiol.* **139**(Suppl. 1), S3–15. <https://doi.org/10.1016/j.ijfoodmicro.2010.01.021> (2010).
7. Wain, J., Hendriksen, R. S., Mikoleit, M. L., Keddy, K. H. & Ochiai, R. L. Typhoid fever. *Lancet* **385**, 1136–1145. [https://doi.org/10.1016/S0140-6736\(13\)62708-7](https://doi.org/10.1016/S0140-6736(13)62708-7) (2015).
8. Upadhayay, A., Pal, D. & Kumar, A. *Salmonella typhi* induced oncogenesis in gallbladder cancer: Co-relation and progression. *Adv. Cancer Biol. Metastasis* **4**, 100032. <https://doi.org/10.1016/j.adcanc.2022.100032> (2022).
9. Costa, T. R. *et al.* Secretion systems in Gram-negative bacteria: structural and mechanistic insights. *Nat. Rev. Microbiol.* **13**, 343–359. <https://doi.org/10.1038/nrmicro3456> (2015).
10. Pallen, M. J. & Wren, B. W. Bacterial pathogenomics. *Nature* **449**, 835–842. <https://doi.org/10.1038/nature06248> (2007).
11. Juhas, M. *et al.* Genomic islands: Tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiol. Rev.* **33**, 376–393. <https://doi.org/10.1111/j.1574-6976.2008.00136.x> (2009).
12. Vernikos, G. S. & Parkhill, J. Interpolated variable order motifs for identification of horizontally acquired DNA: Revisiting the *Salmonella* pathogenicity islands. *Bioinformatics* **22**, 2196–2203. <https://doi.org/10.1093/bioinformatics/btl369> (2006).
13. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589. <https://doi.org/10.1038/s41586-021-03819-2> (2021).
14. Medvedev, K. E., Schaeffer, R. D., Chen, K. S. & Grishin, N. V. Pan-cancer structurome reveals overrepresentation of beta sandwiches and underrepresentation of alpha helical domains. *Sci. Rep.* **13**, 11988. <https://doi.org/10.1038/s41598-023-39273-5> (2023).
15. Schaeffer, R. D. *et al.* ECOD domain classification of 48 whole proteomes from AlphaFold Structure Database using DPAM2. *PLoS Comput. Biol.* **20**, e1011586. <https://doi.org/10.1371/journal.pcbi.1011586> (2024).
16. Cheng, H. *et al.* ECOD: An evolutionary classification of protein domains. *PLoS Comput. Biol.* **10**, e1003926. <https://doi.org/10.1371/journal.pcbi.1003926> (2014).
17. Schaeffer, R. D. *et al.* ECOD: identification of distant homology among multidomain and transmembrane domain proteins. *BMC Mol. Cell Biol.* **20**, 18. <https://doi.org/10.1186/s12860-019-0204-5> (2019).
18. Zhang, J., Schaeffer, R. D., Durham, J., Cong, Q. & Grishin, N. V. DPAM: A domain parser for AlphaFold models. *Protein Sci.* **32**, e4548. <https://doi.org/10.1002/pro.4548> (2023).
19. Medvedev, K. E., Kinch, L. N., Schaeffer, R. D. & Grishin, N. V. Functional analysis of Rossmann-like domains reveals convergent evolution of topology and reaction pathways. *PLoS Comput. Biol.* **15**, e1007569. <https://doi.org/10.1371/journal.pcbi.1007569> (2019).
20. Medvedev, K. E., Kinch, L. N., Dustin Schaeffer, R., Pei, J. & Grishin, N. V. A fifth of the protein world: Rossmann-like proteins as an evolutionarily successful structural unit. *J. Mol. Biol.* **433**, 166788. <https://doi.org/10.1016/j.jmb.2020.166788> (2021).

21. Kato, J., Lefebre, M. & Galan, J. E. Structural features reminiscent of ATP-driven protein translocases are essential for the function of a type III secretion-associated ATPase. *J. Bacteriol.* **197**, 3007–3014. <https://doi.org/10.1128/JB.00434-15> (2015).
22. Browning, D. F., Butala, M. & Busby, S. J. W. Bacterial transcription factors: Regulation by pick “N” Mix. *J. Mol. Biol.* **431**, 4067–4077. <https://doi.org/10.1016/j.jmb.2019.04.011> (2019).
23. Owji, H., Nezafat, N., Negahdaripour, M., Hajiebrahimi, A. & Ghasemi, Y. A comprehensive review of signal peptides: Structure, roles, and applications. *Eur. J. Cell Biol.* **97**, 422–441. <https://doi.org/10.1016/j.ejcb.2018.06.003> (2018).
24. Berman, H. M. *et al.* The protein data bank. *Nucleic Acids Res.* **28**, 235–242. <https://doi.org/10.1093/nar/28.1.235> (2000).
25. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinf.* **10**, 421. <https://doi.org/10.1186/1471-2105-10-421> (2009).
26. Madan-Babu, M. & Teichmann, S. A. Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res.* **31**, 1234–1244. <https://doi.org/10.1093/nar/gkg210> (2003).
27. Hansen-Westerv, I. & Hensel, M. *Salmonella* pathogenicity islands encoding type III secretion systems. *Microbes Infect.* **3**, 549–559. [https://doi.org/10.1016/s1286-4579\(01\)01411-3](https://doi.org/10.1016/s1286-4579(01)01411-3) (2001).
28. Lou, L., Zhang, P., Piao, R. & Wang, Y. *Salmonella* pathogenicity Island 1 (SPI-1) and its complex regulatory network. *Front. Cell Infect. Microbiol.* **9**, 270. <https://doi.org/10.3389/fcimb.2019.00270> (2019).
29. Miletic, S. *et al.* Substrate-engaged type III secretion system structures reveal gating mechanism for unfolded protein translocation. *Nat. Commun.* **12**, 1546. <https://doi.org/10.1038/s41467-021-21143-1> (2021).
30. Lara-Tejero, M., Kato, J., Wagner, S., Liu, X. & Galan, J. E. A sorting platform determines the order of protein secretion in bacterial type III systems. *Science* **331**, 1188–1191. <https://doi.org/10.1126/science.1201476> (2011).
31. Notti, R. Q., Bhattacharya, S., Lilic, M. & Stebbins, C. E. A common assembly module in injectisome and flagellar type III secretion sorting platforms. *Nat. Commun.* **6**, 7125. <https://doi.org/10.1038/ncomms8125> (2015).
32. Bonemann, G., Pietrosiuk, A. & Mogk, A. Tubules and donuts: A type VI secretion story. *Mol. Microbiol.* **76**, 815–821. <https://doi.org/10.1111/j.1365-2958.2010.07171.x> (2010).
33. Blondel, C. J., Jimenez, J. C., Contreras, I. & Santiviago, C. A. Comparative genomic analysis uncovers 3 novel loci encoding type six secretion systems differentially distributed in *Salmonella* serotypes. *BMC Genom.* **10**, 354. <https://doi.org/10.1186/1471-2164-10-354> (2009).
34. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419. <https://doi.org/10.1093/nar/gkaa913> (2021).
35. Gabler, F. *et al.* Protein sequence analysis using the MPI bioinformatics toolkit. *Curr. Protoc. Bioinf.* **72**, e108. <https://doi.org/10.1002/cpb1.108> (2020).
36. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* <https://doi.org/10.1038/s41587-023-01773-0> (2023).
37. Zhou, J. *et al.* Structural evidence for a [4Fe–5S] intermediate in the non-redox desulfurization of thiouracil. *Angew. Chem. Int. Ed. Engl.* **60**, 424–431. <https://doi.org/10.1002/anie.202011211> (2021).
38. Liu, B., Zheng, D., Zhou, S., Chen, L. & Yang, J. VFDB 2022: A general classification scheme for bacterial virulence factors. *Nucleic Acids Res.* **50**, D912–D917. <https://doi.org/10.1093/nar/gkab1107> (2022).
39. Paysan-Lafosse, T. *et al.* InterPro in 2022. *Nucleic Acids Res.* **51**, D418–D427. <https://doi.org/10.1093/nar/gkac993> (2023).
40. Rao, D. N., Dryden, D. T. & Bheemanaik, S. Type III restriction-modification enzymes: A historical perspective. *Nucleic Acids Res.* **42**, 45–55. <https://doi.org/10.1093/nar/gkt616> (2014).
41. Wong, S. G. & Dessen, A. Structure of a bacterial alpha-2-macroglobulin reveals mimicry of eukaryotic innate immunity. *Nat. Commun.* **5**, 4917. <https://doi.org/10.1038/ncomms5917> (2014).
42. Robb, C. S., Assmus, M., Nano, F. E. & Boraston, A. B. Structure of the T6SS lipoprotein TssJ1 from *Pseudomonas aeruginosa*. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **69**, 607–610. <https://doi.org/10.1107/S1744309113012220> (2013).
43. Edwards, R. A., Schifferli, D. M. & Maloy, S. R. A role for *Salmonella* fimbriae in intraperitoneal infections. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 1258–1262. <https://doi.org/10.1073/pnas.97.3.1258> (2000).
44. Fenwick, M. K., Philmus, B., Begley, T. P. & Earlick, S. E. Toxoflavin lyase requires a novel 1-His-2-carboxylate facial triad. *Biochemistry* **50**, 1091–1100. <https://doi.org/10.1021/bi0101741v> (2011).
45. Peat, T. S., Newman, J., Waldo, G. S., Berendzen, J. & Terwilliger, T. C. Structure of translation initiation factor 5A from *Pyrobaculum aerophilum* at 1.75 Å resolution. *Structure* **6**, 1207–1214. [https://doi.org/10.1016/s0969-2126\(98\)00120-8](https://doi.org/10.1016/s0969-2126(98)00120-8) (1998).
46. Diard, M. *et al.* Antibiotic treatment selects for cooperative virulence of *Salmonella typhimurium*. *Curr. Biol.* **24**, 2000–2005. <https://doi.org/10.1016/j.cub.2014.07.028> (2014).
47. Darby, E. M. *et al.* Molecular mechanisms of antibiotic resistance revisited. *Nat. Rev. Microbiol.* **21**, 280–295. <https://doi.org/10.1038/s41579-022-00820-y> (2023).
48. Varadi, M. *et al.* AlphaFold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444. <https://doi.org/10.1093/nar/gkab1061> (2022).
49. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60. <https://doi.org/10.1038/nmeth.3176> (2015).
50. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinf.* **20**, 473. <https://doi.org/10.1186/s12859-019-3019-7> (2019).
51. Schaeffer, R. D. *et al.* Classification of domains in predicted structures of the human proteome. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2214069120. <https://doi.org/10.1073/pnas.2214069120> (2023).
52. Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138. <https://doi.org/10.1006/jmbi.1993.1489> (1993).
53. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637. <https://doi.org/10.1002/bip.360221211> (1983).
54. Amavvisit, P., Lightfoot, D., Browning, G. E. & Markham, P. F. Variation between pathogenic serovars within *Salmonella* pathogenicity islands. *J. Bacteriol.* **185**, 3624–3635. <https://doi.org/10.1128/JB.185.12.3624-3635.2003> (2003).
55. Gerlach, R. G. *et al.* *Salmonella* pathogenicity Island 4 encodes a giant non-fimbrial adhesin and the cognate type 1 secretion system. *Cell Microbiol.* **9**, 1834–1850. <https://doi.org/10.1111/j.1462-5822.2007.00919.x> (2007).
56. Knodler, L. A. *et al.* *Salmonella* effectors within a single pathogenicity island are differentially expressed and translocated by separate type III secretion systems. *Mol. Microbiol.* **43**, 1089–1103. <https://doi.org/10.1046/j.1365-2958.2002.02820.x> (2002).
57. Pickard, D. *et al.* Composition, acquisition, and distribution of the Vi exopolysaccharide-encoding *Salmonella enterica* pathogenicity island SPI-7. *J. Bacteriol.* **185**, 5055–5065. <https://doi.org/10.1128/JB.185.17.5055-5065.2003> (2003).
58. 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. <https://doi.org/10.1186/s40659-017-0109-8> (2017).
59. Velasquez, J. C. *et al.* SPI-9 of *Salmonella enterica* serovar Typhi is constituted by an operon positively regulated by RpoS and contributes to adherence to epithelial cells in culture. *Microbiol. (Read.)* **162**, 1367–1378. <https://doi.org/10.1099/mic.0.000319> (2016).
60. Bishop, A. L. *et al.* Analysis of the hypervariable region of the *Salmonella enterica* genome associated with tRNA(leuX). *J. Bacteriol.* **187**, 2469–2482. <https://doi.org/10.1128/JB.187.7.2469-2482.2005> (2005).
61. Chiu, C. H. *et al.* The genome sequence of *Salmonella enterica* serovar Choleraesuis, a highly invasive and resistant zoonotic pathogen. *Nucleic Acids Res.* **33**, 1690–1698. <https://doi.org/10.1093/nar/gki297> (2005).

62. Tomljenovic-Berube, A. M. *et al.* Mapping and regulation of genes within *Salmonella* pathogenicity island 12 that contribute to *in vivo* fitness of *Salmonella enterica* Serovar Typhimurium. *Infect. Immun.* **81**, 2394–2404. <https://doi.org/10.1128/IAI.00067-13> (2013).
63. Shah, D. H. *et al.* Identification of *Salmonella gallinarum* virulence genes in a chicken infection model using PCR-based signature-tagged mutagenesis. *Microbiol. (Read.)* **151**, 3957–3968. <https://doi.org/10.1099/mic.0.28126-0> (2005).
64. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology, The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29. <https://doi.org/10.1038/75556> (2000).
65. UniProt, C. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531. <https://doi.org/10.1093/nar/gkac1052> (2023).
66. Kruskal, J. B. & Wish, M. *Multidimensional Scaling* (SAGE Publications Inc., 1978).

### Author contributions

Kirill E. Medvedev: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Visualization, Writing - Original Draft, Project administration. Jing Zhang: Methodology, Software, Investigation, Data Curation, Writing - Original Draft. R. Dustin Schaeffer: Methodology, Software, Formal analysis, Funding acquisition, Writing - Review & Editing. Lisa N. Kinch: Investigation, Validation, Formal analysis. Qian Cong: Conceptualization, Investigation, Resources, Funding acquisition, Writing - Review & Editing. Nick V. Grishin: Conceptualization, Resources, Funding acquisition, Writing - Review & Editing.

### Funding

The study is supported by grants from the National Institute of General Medical Sciences of the National Institutes of Health GM127390 (to N.V.G.), GM147367 (to R.D.S), the Welch Foundation I-1505 (to N.V.G.), the National Science Foundation DBI 2224128 (to N.V.G.). Q.C. is a Southwestern Medical Foundation-endowed scholar. This research is partly supported by grant I-2095-20220331 to Q.C. from Welch Foundation.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-60991-x>.

**Correspondence** and requests for materials should be addressed to K.E.M.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024