



# Is the Evolution of *Salmonella enterica* subsp. *enterica* Linked to Restriction-Modification Systems?

Louise Roer,<sup>a</sup> Rene S. Hendriksen,<sup>a</sup> Pimlapas Leekitcharoenphon,<sup>a</sup>  
Oksana Lukjancenko,<sup>a</sup>  Rolf Sommer Kaas,<sup>a</sup> Henrik Hasman,<sup>a,b</sup>  
Frank M. Aarestrup<sup>a</sup>

Division for Epidemiology and Microbial Genomics, National Food Institute, Technical University of Denmark, Kongens Lyngby, Denmark<sup>a</sup>; Section for Microbiology and Diagnostics, Statens Serum Institut, Copenhagen, Denmark<sup>b</sup>

**ABSTRACT** *Salmonella enterica* subsp. *enterica* bacteria are highly diverse food-borne pathogens that are subdivided into more than 1,500 serovars. The diversity is believed to result from mutational evolution, as well as intra- and interspecies recombination that potentially could be influenced by restriction-modification (RM) systems. The aim of this study was to investigate whether RM systems were linked to the evolution of *Salmonella enterica* subsp. *enterica*. The study included 221 *Salmonella enterica* genomes, of which 68 were *de novo* sequenced and 153 were public available genomes from ENA. The data set covered 97 different serovars of *Salmonella enterica* subsp. *enterica* and an additional five genomes from four other *Salmonella* subspecies as an outgroup for constructing the phylogenetic trees. The phylogenetic trees were constructed based on multiple alignment of core genes, as well as the presence or absence of pangenes. The topology of the trees was compared to the presence of RM systems, antimicrobial resistance (AMR) genes, *Salmonella* pathogenicity islands (SPIs), and plasmid replicons. We did not observe any correlation between evolution and the RM systems in *S. enterica* subsp. *enterica*. However, sublineage correlations and serovar-specific patterns were observed. Additionally, we conclude that plasmid replicons, SPIs, and AMR were all better correlated to serovars than to RM systems. This study suggests a limited influence of RM systems on the evolution of *Salmonella enterica* subsp. *enterica*, which could be due to the conjugational mode of horizontal gene transfer in *Salmonella*. Thus, we conclude that other factors must be involved in shaping the evolution of bacteria.

**IMPORTANCE** The evolution of bacterial pathogens, their plasticity and ability to rapidly change and adapt to new surroundings are crucial for understanding the epidemiology and public health. With the application of genomics, it became clear that horizontal gene transfer played a key role in evolution. To understand the evolution and diversification of pathogens, we need to understand the processes that drive the horizontal gene transfer. Restriction-modification systems are thought to cause rearrangements within the chromosome, as well as act as a barrier to horizontal gene transfer. However, here we show that the correlation between restriction-modification systems and evolution in other bacterial species does not apply to *Salmonella enterica* subsp. *enterica*. In summary, from this work, we conclude that other mechanisms might be involved in controlling and shaping the evolution of *Salmonella enterica* subsp. *enterica*.

**KEYWORDS:** restriction-modification systems, evolution, *Salmonella* phylogenetic analysis, next-generation sequencing, whole-genome sequencing

Received 4 February 2016 Accepted 11 May 2016 Published 21 June 2016

**Citation** Roer L, Hendriksen RS, Leekitcharoenphon P, Lukjancenko O, Kaas RS, Hasman H, Aarestrup FM. 2016. Is the evolution of *Salmonella enterica* subsp. *enterica* linked to restriction-modification systems? *mSystems* 1(3): e00009-16. doi:10.1128/mSystems.00009-16.

**Editor** Jonathan Eisen, UC Davis Genome Center

**Copyright** © 2016 Roer et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Louise Roer, [louise.roer@yahoo.dk](mailto:louise.roer@yahoo.dk).

The *Salmonella* genus is highly diverse even though it is comprised of only two species, *Salmonella enterica* and *Salmonella bongori*. The species *S. enterica* contains six subspecies, of which the highly diverse members of *S. enterica* subsp. *enterica* can be subdivided into more than 1,500 serovars. This subspecies is a common cause of diseases in humans and domestic animals (1, 2) and one of the leading causes of foodborne illness worldwide (3).

Recombination between genomes is thought to be a major driver in evolution (4) and to contribute to the diversity within the *Salmonella* genus (5, 6). It was suggested that in *Neisseria meningitidis*, the phylogeny is associated with the content of restriction-modification systems (RM systems) (7). Furthermore, rearrangements of genomes caused by RM systems are described as factors that could influence the evolution of pathogens (4). In addition to their role in rearrangements, RM systems are also considered to be a barrier for horizontal gene transfer (HGT) between bacteria, thus serving as an immune defense system for uptake of foreign DNA (8–10). The contribution, to our knowledge, has never been quantified, and thus, we have recently shown that for conjugational transfer between isogenic *Escherichia coli* isolates, the barriers of RM systems are not absolute (11).

RM systems are comprised of a restriction enzyme (RE) and a cognate methyltransferase. The restriction enzyme recognizes and digests foreign incoming DNA, whereas the methyltransferase performs methylation of the bacterium's own DNA to protect itself from degradation by the cognate restriction enzyme (12, 13). This enables the bacterium to distinguish between its own (methylated) DNA and incoming nonmethylated DNA.

The RM systems are divided into four types (I to IV), based on their protein complexes, the subunit composition, and the functionality of the system (14). The type I systems are complexes of three gene products: *hsdR* (R, restriction), *hsdM* (M, methylation), and *hsdS* (S, sequence specificity). This type cleaves nonmethylated DNA randomly at a remote distance from the recognition sequence determined by the specificity subunit. The protein complex of all three gene products,  $R_2M_2S$  (two subunits of R and M and one subunit of S) must be established prior to restriction, whereas a complex of only the *hsdM* and *hsdS* proteins ( $M_2S$ ) is needed for methylation of the DNA (15). Type II systems are only made up of methyltransferases and restriction enzymes, where the function and composition of the M and R products varies depending on the subtype of the system. Type II systems modulate (cleave and/or methylate) unmethylated DNA at specific recognition sites, making them suitable as molecular biological tools to cut DNA for cloning or other analysis where only a piece of DNA is needed (12, 15). Type III systems, consisting of the gene products Res and Mod, hemimethylate the DNA and cleave DNA at specific sites 25 to 27 bp downstream from the recognition sequence (16), whereas type IV, compared to types I to III, does not encode a methyltransferase and only cleaves methylated DNA (12, 15).

In this study, we elucidated the potential association between RM systems and the phylogeny of *S. enterica* subsp. *enterica* serovars. We tested the hypothesis that RM systems might be linked to the evolution of *S. enterica* subsp. *enterica* and thereby be responsible for the diversification of the species. The most effective source of variation within the genome is caused by HGT (17), transferring, e.g., antimicrobial resistance (AMR) genes between bacteria. In *Salmonella*, the *Salmonella* pathogenicity islands (SPIs) are believed to be acquired by horizontal gene transfer and to have an effect on the structure of the genome (18, 19). Thus, we also elucidated the content of plasmids, AMR genes, and SPIs in correlation to the RM systems and the phylogeny of the species to examine their effect on the evolution.

## RESULTS

**Genomes.** A total of 221 *Salmonella* genomes were included in the analysis. One hundred fifty-three genomes previously described by Timme et al. (1) were retrieved from the European Nucleotide Archive; their accession numbers are listed in Table S1 in the supplemental material. This collection was merged with 68 genomes sequenced

as part of the 100K Foodborne Pathogen Genome Project (<http://100kgenome.vetmed.ucdavis.edu/>, NCBI BioProject accession number [PRJNA186441](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA186441); individual accession numbers are listed in Table S1). The final collection consisted of 216 *S. enterica* subsp. *enterica* genomes and five genomes of other *S. enterica* subspecies. The 221 *Salmonella* genomes are summarized in the supplemental material (see Table S1).

**Characterization of restriction-modification systems.** To characterize the RM systems of the 221 genomes, a whole-genome-sequencing (WGS) analysis was performed using the newly developed tool Restriction-ModificationFinder 1.0. Partial systems were completed by individual BLAST analysis of up- and downstream sequences against the REBASE database (20). In total, we identified 113 putative RM systems, including 58 type I RM systems (TI), of which 43 had unknown recognition sequences, 23 type II RM systems (TII), 2 type III RM systems (TIII), and 30 type IV RM systems (TIV). In addition, numerous methyltransferases outside the RM systems were identified, including type I, type II, and type III methyltransferases (see Table S2 in the supplemental material).

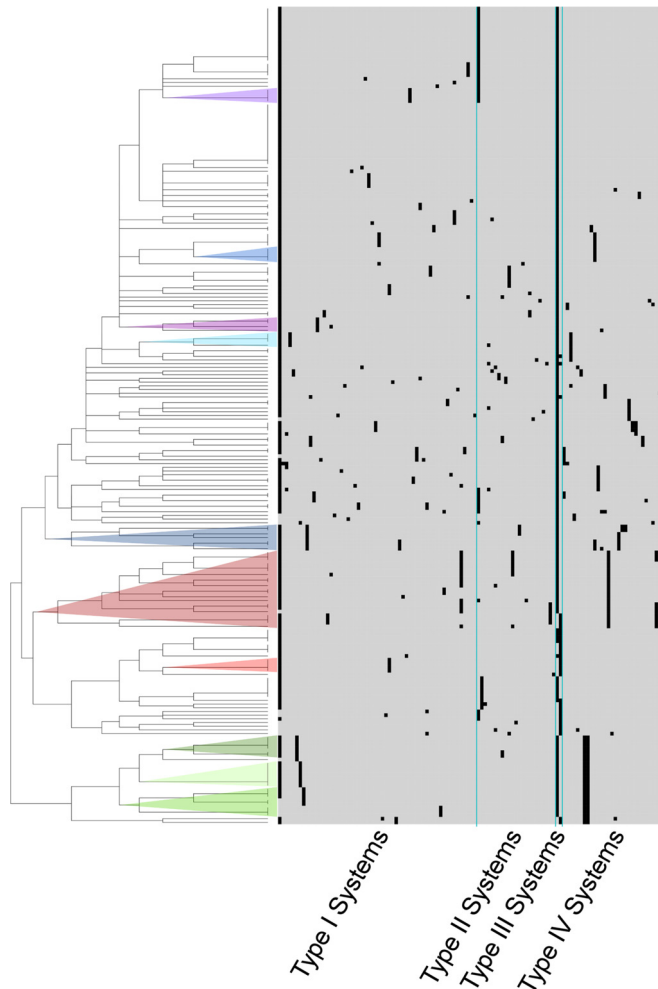
The additional methyltransferases (without associated restriction enzymes) identified were not included in the analysis, as the barrier for HGT is thought to be caused by the cleavage from REs.

One of the genomes only contained one RM system, while the other genomes contained between two and seven systems. All genomes contained a type III RM system, one of which was shared by 198 genomes. The type I system TI-1 was shared by 203 of the genomes, and 37 RM systems were specific to a single genome. The remaining systems were shared by 2 to 38 genomes. The distribution of the RM systems is illustrated in the presence/absence matrix in Fig. 1 and presented in detail in Fig. S1. The analysis revealed very diverse content of RM systems, and in assessing the highest level of discrimination in the cladogram, 120 distinct clusters were formed, with 77 clusters containing a single genome.

**The salmonella pan- and core genomes.** The pan- and core genomes were estimated based on the 221 *S. enterica* genomes. The progression of the pan- and core genomes as increasing numbers of the *S. enterica* genomes were added to the analysis is shown in Fig. S2 in the supplemental material. In analyzing the pangenome, consisting of any gene families found, the gene families increased gradually with the one-by-one addition of the different *S. enterica* subsp. *enterica* serovars, compared to a distinct increase in the number of gene families with the inclusion of *S. enterica* subsp. *diarizonae*. In contrast, the number of conserved gene families across the *S. enterica* subsp. *enterica* genomes in the core genome analysis was relatively consistent; however, the number of core families dropped when other *S. enterica* subspecies were introduced into the analysis. The final analysis comprising all 221 *S. enterica* genomes contained 16,375 gene families in the pangenome (for a representative genome, see Text S1 in the supplemental material) and 2,138 gene families in the core genome (for a representative genome, see Text S2). Analyzing the total number of gene families in the pangenome, each *S. enterica* genome contributed, on average, 65 new gene families, increasing the diversity within the *S. enterica* species.

**The link between evolution and restriction-modification systems.** To study the genomic evolution of *S. enterica* subsp. *enterica*, differences within the core genes were examined for all 221 genomes, and the results are illustrated by the phylogenetic core genome tree in Fig. 2. This evolution of *S. enterica* subsp. *enterica* is formed not only by the differences in the genes shared by the genomes but also by the loss of genes leading to differences in gene content. Figure 3 displays the pangenome tree, based on the presence and absence of genes across all the genomes included in the study.

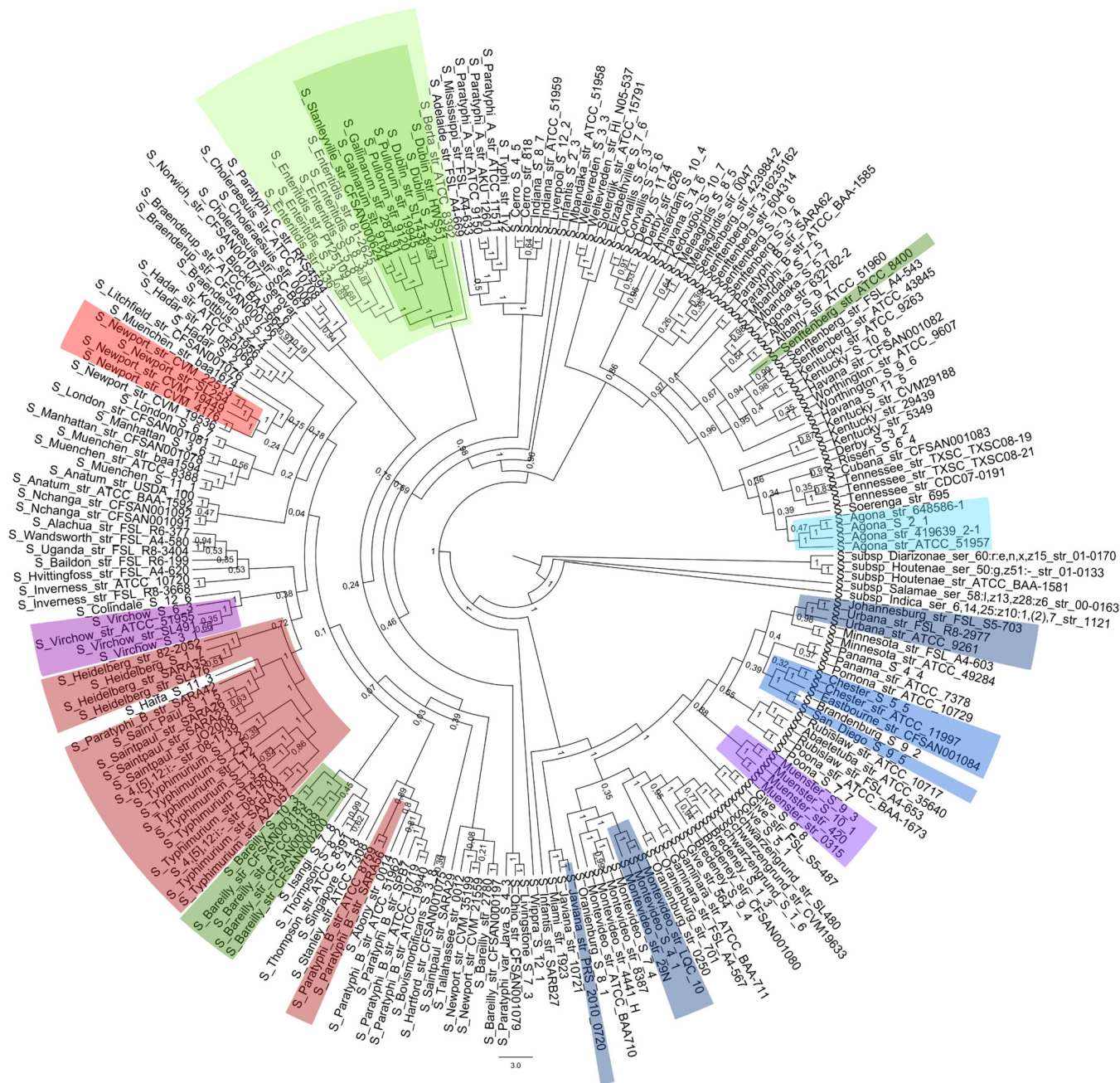
For both the core and pangenome trees, the high bootstrap values of 1 near the root of the trees reflected a good phylogenetic representation of the data of the *Salmonella* genus, whereas the low bootstrap values found in some of the branches indicated difficulties determining the definite arrangement within the branch. These



**FIG 1** Presence or absence matrix of the 113 restriction-modification systems. In the matrix, each row represents one genome analyzed, and each column represents one of 113 RM systems. The cladogram is a hierarchical clustering of the genomes based on the Euclidean algorithm. Convergence of genomes with highly similar RM system contents and discrete phylogenetic clades on the core and pangenome trees is indicated by the colors in the cladogram.

difficulties could be caused by different possibilities, such as the genomes being highly similar in their core or pangenome content, high levels of gene transfer or recombination, or convergent evolution. However, both the core and pangenome trees separated the serovars from each other.

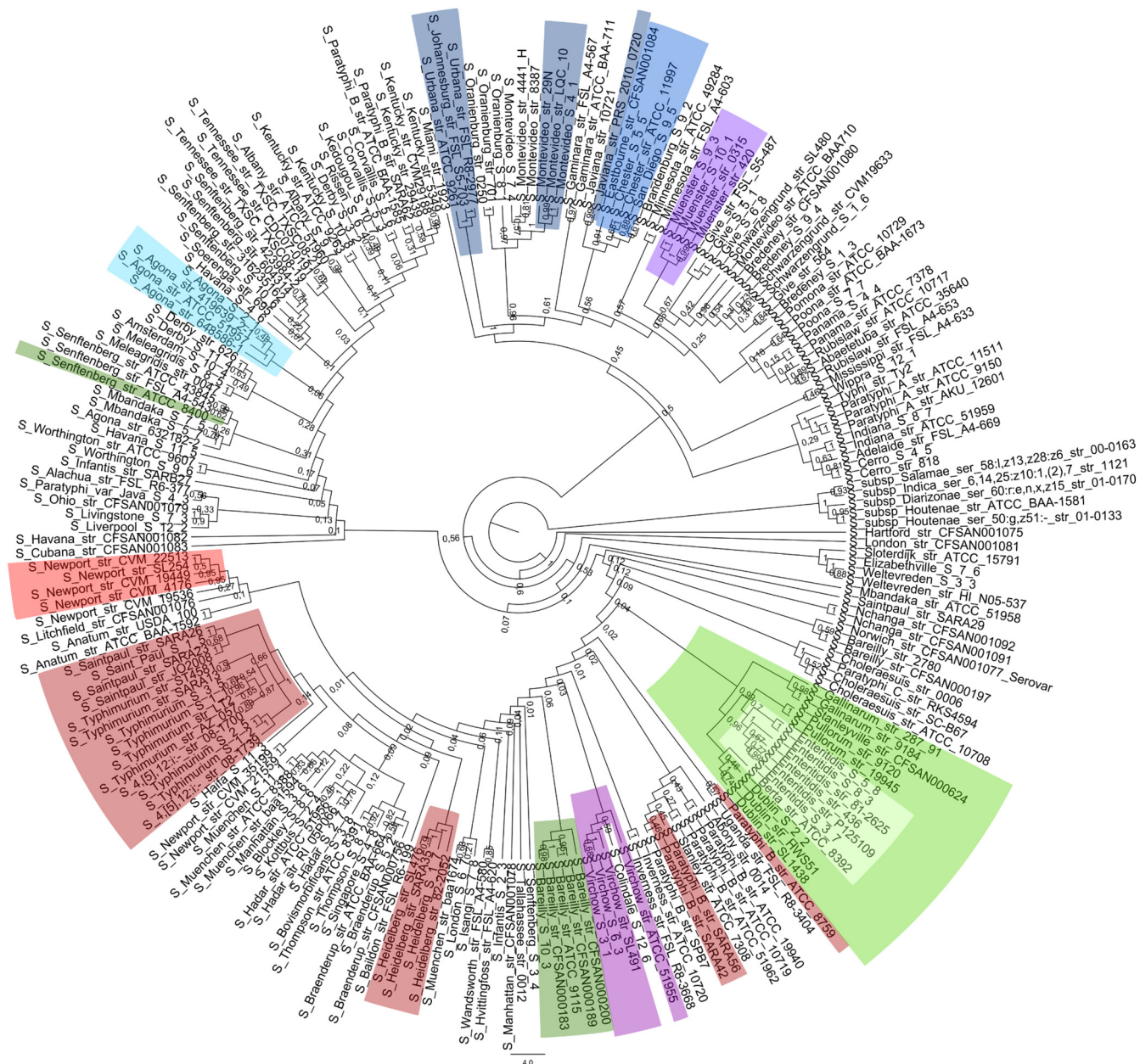
To examine the hypothesis of RM systems being linked to the evolution of *S. enterica* subsp. *enterica*, parallel analyses were performed to identify groups of genomes with highly similar RM patterns that also formed discrete phylogenetic clusters on the core and pangenome trees. Genomes forming distinct clusters by their RM systems are indicated by different colors in the matrix of RM systems (Fig. 1) and in the core and pangenome trees (Fig. 2 and 3). A few small clusters with almost identical RM content, partly clustering together in the core and pangenome trees, were observed. In most cases, the genomes within a cluster of the same color belonged to the same serovar, as for *S. enterica* subsp. *enterica* serovar Bareilly or Agona (dark green and light blue). However, three larger clades, with 21, 8, and 7 genomes, were identified in the RM matrix. Of the 21 genomes in the first clade, containing *S. enterica* subsp. *enterica* serovars Typhimurium, Saint Paul, Paratyphi B, and Heidelberg, 90% of the genomes were located together in one distinct cluster in the core genome tree, and the remaining 10% were also located together. For the pangenome tree, the 21 genomes



**FIG 2** Concatenated core genome tree of *Salmonella enterica* subspecies *enterica* serovars constructed on 1,072 core gene clusters. The phylogenetic tree is constructed on core genes and represented as a cladogram. Discrete phylogenetic clades with highly identical RM system contents are indicated by the colors defined in Fig. 1.

were identified at 4 different locations, with a distribution of 67%, 19%, 9%, and 5% of the genomes, respectively. The two larger clades of 8 and 7 genomes were located adjacent to each other in the RM matrix, the first containing *S. enterica* subsp. *enterica* serovars Stanleyville, Gallinarum, Pullorum, and Dublin and the second containing serovars Enteritidis and Berta. When comparing the two RM clades to the core and pangenome trees, the two clades clustered together with all 100% of the genomes in both trees, indicating evolutionary relatedness.

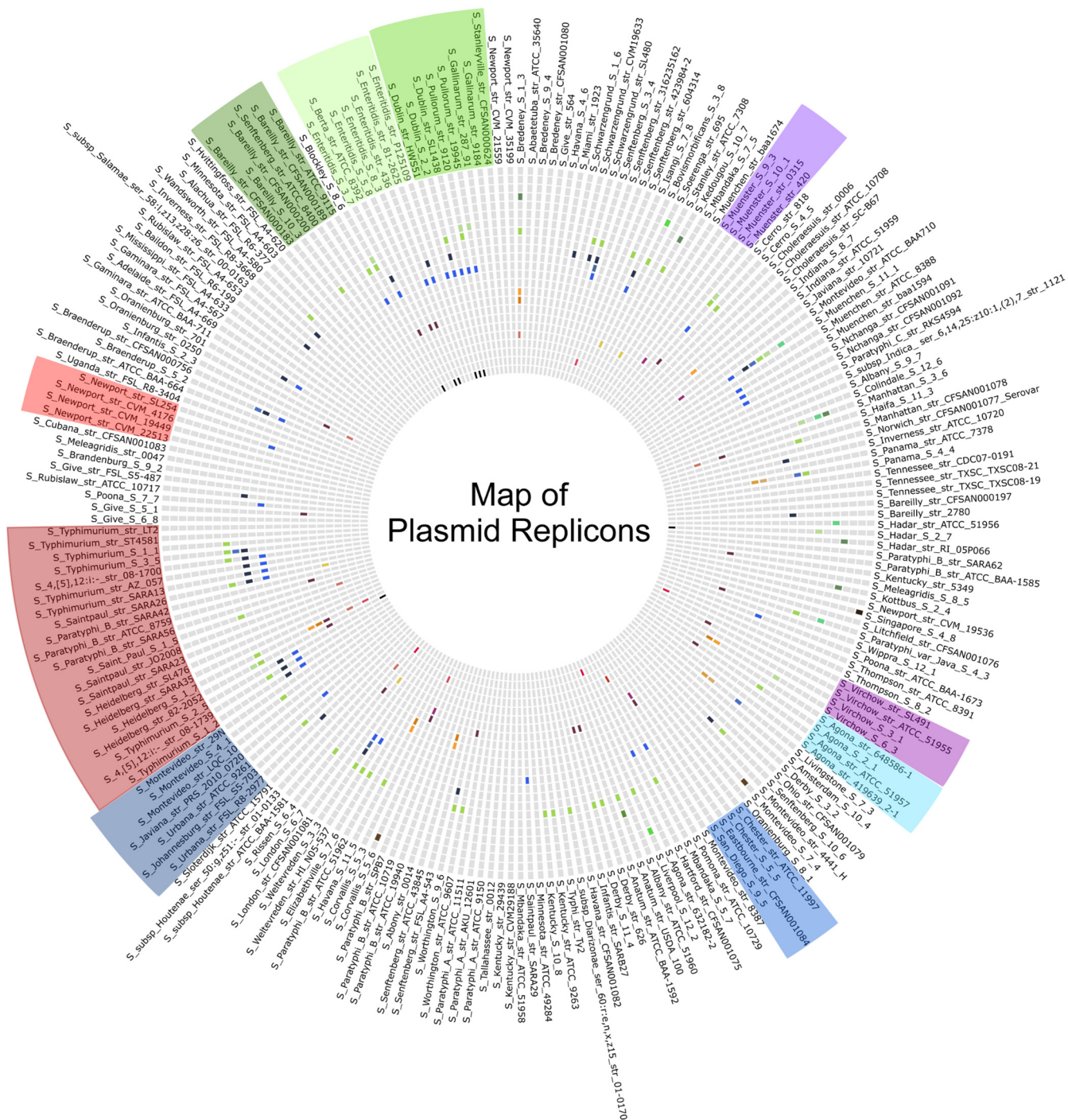
As all the genomes contained a type III system, it was hypothesized to be the first introduced into *S. enterica* subsp. *enterica*. The core and pangenome trees were compared with the phylogenetic relationship between the type III systems (see Fig. S4



**FIG 3** Pangenome tree of *Salmonella enterica* subspecies *enterica* serovars with RM systems indicated by colors. The phylogenetic tree is constructed from the presence/absence matrix of genes across genomes and represented as a cladogram. The colors represent the different groups of RM systems defined in Fig. 1.

in the supplemental material). From the visual inspection and comparison of the trees, we found a common distinct clade in the phylogeny of the type III systems (marked in Fig. S4) where 48 of the 49 genomes were located together in the core genome tree, whereas on the pangenome tree, 47 of the 49 genomes were located together in a distinct clade. No other correlations were observed between the type III systems and the phylogeny of *S. enterica* subsp. *enterica*.

**Plasmid replicons, antimicrobial resistance, and pathogenicity islands in *Salmonella enterica*.** All 221 *Salmonella* genomes were analyzed for their content of plasmid replicons by using the Center for Genomic Epidemiology (CGE) Web tool PlasmidFinder 1.2, a tool proven capable of detecting 100% of the previously characterized and fully sequenced plasmids applied to the analysis (24 plasmids), in addition to detecting a broad variety of plasmid replicons among a collection of *S. enterica*



**FIG 4** Map of plasmid replicons in *Salmonella enterica*. The genomes are ordered according to their RM systems and colored as in Fig. 1. The circles in the map represent the replicons, from the outside to the middle, *incA/C*, *incA/C2*, ColMGD2, Col156, Col8282, ColE10, ColpVC, ColRNAI, *incFIA*, *incFIB*, *incFIC*, *incFII*, *incHI1A*, *incHI1B*, *incHI2*, *incHI2A*, *incI1*, *incI2*, *incN*, *incP*, p0111, *incQ1*, *incR*, *incX1*, and *incX4*. In the plasmid replicon map, light gray indicates the absence of a replicon in the given genome, and the colors indicate the presence of a specific replicon.

subsp. *enterica* serovar Typhimurium isolates (21). Out of the 221 genomes tested in this study, 118 did not contain any replicons present in the PlasmidFinder database. In the remaining 103 genomes, 40 different plasmid replicons were identified, with each genome containing up to seven different replicons.

Assessing the replicons in comparison with the RM systems observed in the *S. enterica* subsp. *enterica* serovars, no visual similarity was observed (Fig. 4), and RM systems

and plasmid replicons were never observed on the same contigs. Evaluating the plasmid replicons for correlation to serovars, no correlation was observed between the serovars and the quantity of replicons; the 11 genomes with the highest replicon content represented 10 different serovars. However, for multiple isolates with the same serovar, common replicons were observed, such as with the *incFII* and *incX1* replicons present in all of the *S. enterica* subsp. *enterica* serovar Dublin genomes. Additionally, *incFII* was observed in 50% of the *S. enterica* subsp. *enterica* serovar Enteritidis genomes and, together with the *incFIB* replicon, in 50% of the *S. enterica* subsp. *enterica* serovar Typhimurium and both 4,[5],12:i: genomes. Interestingly, even though the numbers of replicons are not equal in identical serovars, the visual inspections imply an association between replicons and serovars and no association between replicons and RM systems.

AMR genes were found in 220 of the genomes, varying from 1 to 19 different genes per genome, with *aac(6′)-Iy* present as the only resistance gene in 140 of the genomes. The correlation of AMR genes and RM systems was examined (see Fig. S3 in the supplemental material); however, no visual conjunction was observed. Despite this potential bias, the assessment of correlation between AMR genes and plasmid replicons revealed different resistance genes located on the same contigs as plasmid replicons. Few genomes contained more than four plasmid replicons, and most contained between 10 and 16 resistance genes, correlating high resistance to the number of replicons present in the genomes.

The presence of SPIs was assessed in all 221 *Salmonella* genomes by utilizing the newly developed Web tool SPI-Finder 1.0, and the results are visualized in Fig. 5. SPIs were found in all 221 genomes, and the number of SPIs in each genome varied from 1 to 14 islands/genes of islands.

The comparison of the SPIs and the content of RM systems revealed no clear association (Fig. 5). However, in associating the SPIs to individual *Salmonella* serovars, such as *S. enterica* subsp. *enterica* serovars Typhimurium, Paratyphi A, Choleraesuis, Heidelberg, and Saintpaul, association was apparent within the specific serovars.

## DISCUSSION

For decades, RM systems have been recognized for their ability to act as “immune systems” for bacteria, helping to determine whether or not foreign DNA was established in the cell. For *N. meningitidis*, the impact of RM systems in evolution was recently elucidated, and the purpose of the current study was to clarify whether a similar association could be identified in *S. enterica* subsp. *enterica*, by investigating a large subset of different whole-genome-sequenced *S. enterica* subsp. *enterica* serovars and an outgroup of five genomes from other subspecies. However, it was not possible to detect any significant association between RM systems and the overall evolution of *Salmonella enterica* subsp. *enterica*, even though we did observe genomes from the same RM clades in discrete phylogenetic clusters of both the core and pangenome trees.

The *Neisseria* genus, including *N. meningitidis*, is known to serve as a paradigm for natural transformation, where genetic exchange happens frequently due to their persistent competence independent of the phase of their life cycle (22). As RM systems have been recognized as barriers for transformation in multiple species (23–25), the link between a naturally transformable species and the RM systems, as shown for *N. meningitidis*, seems reasonable. However, transformation in *Salmonella* is unlikely, as they are not naturally transformable, and the transfer of genetic material happens mainly through conjugation (26). In addition, a recent study performed in *E. coli* (11) indicated that the barriers imposed by the RM systems in conjugational plasmid transfer were not absolute, which could explain the lack of significant association in our study when trying to associate RM systems to the plasmid replicons, AMR genes, SPIs, and their influence on the evolution.

The core and pangenome trees were constructed with a method previously described by Leekitcharoenphon et al. (27), where 73 genomes were evaluated. They found a core genome of 2,882 genes and a cognate pangenome of 10,581 genes. In our





analysis will not increase the gene pool (28). This also supports the knowledge of *S. enterica* subsp. *enterica* being a highly diverse bacterial species.

Vasu and Nagaraja (4) recently described how changes in the specificity or the acquisition of new RM systems could alter the strains genetically from the original clonal population, as the methyltransferase modifies the genome of its new host and the RE prevents genetic exchange between closely related strains. Thus, mutations accumulate in the “new” strain, leading to genetic diversity. Assessing the RM systems in the 221 genomes, we identified 113 RM systems and numerous individual methyltransferases, with each genome harboring 1 to 7 RM systems. We found type III systems in all of the genomes, with 198 of the 221 genomes analyzed sharing a recognition sequence; the recognition sequences for the remaining isolates are yet unknown. Taking this result into consideration, it is plausible that the type III system was the first RM system introduced into *S. enterica* subsp. *enterica*, with subsequently a greater diversity following the later acquisition of new RM systems as described by Vasu and Nagaraja (4). However, the same RM systems could also have been introduced into different branches at different time points, resulting in identical RM system clades across the trees. To assess this hypothesis, a phylogenetic relationship of the type III systems was constructed and compared with the evolution depicted in the core and pangenome trees. We observed that a distinct clade on the type III system tree also recurred as distinct clades on both the core and pangenome trees, with high bootstrap values.

As the function probably is the same for all the type III systems found here (including the systems without recognition sequences assigned as yet), and as a type III system is found in all the genomes, we do not believe that the type III systems has contributed to the evolution, but the results could suggest that the type III system was the first introduced.

When further assessing the pan- and core genome trees according to all the RM systems present, there are indications of some clustering of genomes with similar RM systems, i.e., the cluster of *S. enterica* subsp. *enterica* serovars Enteritidis and Dublin, as well as the red RM system clade consisting of serovars Typhimurium and Heidelberg located together in both trees. However, the influence is not significant, indicating that the evolution could be driven by several factors. For instance, a previous study compared 28 *Salmonella enterica* isolates and provided evidence that clustered regularly interspaced short palindromic repeats (CRISPR) mediated sublineage evolution (29). Other drivers in evolution are host and environmental adaptations, which besides gene acquisition can be caused by gene loss and deletion, gene duplication, and changes within genes by, e.g., mutations (17, 30).

In the study on *N. meningitidis*, the researchers investigated the association of RM systems, homologous recombination, and the phylogenetic network (7). The main study was performed on 20 genomes, covering 5 serogroups out of the recorded 13 serogroups for *N. meningitidis* (31, 32). Budroni et al. found that genomes from the same clonal complex (CC) were located together in phylogenetic clades based on their core genes. In addition, the clades could be associated with the RM systems identified (7). In our study, we investigated 217 genomes of *S. enterica* subsp. *enterica*, including 97 different serovars, along with an outgroup of five genomes from four other sub-species. Considering the highly diverse data set investigated in this study compared to the one used in the study of *N. meningitidis*, we observed small trends of sublineage association of RM systems and evolution. This could indicate that even though the data set investigated in this study was comprehensive, more genomes of each serovar should be included to cover the complete picture of the influence of RM systems in the evolution of *S. enterica* subsp. *enterica*. With the current speed in WGS, this might be realistic in the nearest future. Thus, even in the ideal data scenario, the lack of association is very likely due to the incomplete barrier of RM systems in conjugation.

As for all database-dependent approaches, the methods are only capable of detecting and reporting records present in the database explored. Our analysis for detecting the RM systems was limited to the current knowledge presented in the REBASE (33),

where the recognition sequences of various type I-specificity subunits were not yet determined. Thus, it is likely that some strains have acquired RM systems with identical recognition patterns; however, this is presently unknown. This is why all of the database-dependent analyses should be interpreted with care.

The plasmid replicons which potentially could have an effect on the bacterial diversity due to horizontal gene transfer were also identified, but no clear correlation between the RM systems and the content of harbored plasmid replicons was observed. However, this approach might be complicated by the fact that plasmids are transferable and affected by factors such as fitness costs, selective pressure (34–37), and again, the fact that the RM barrier is not absolute (11). Thus, the analysis performed on the plasmid replicons illustrates the current status at the time of isolation, in contrast to what plasmids potentially could be acquired. AMR can be encoded by genes located on transferable plasmids; this potentially could reflect the promiscuity of the genomes reflected in current time, which could explain the lack of association between AMR genes and RM systems. This might also be explained by the possibility that the data set is biased with respect to selective criteria, e.g., susceptibility to antimicrobial agents, plasmid content, or virulence (SPI) of the isolates. However, the content of the RM systems is not believed to be affected by the possible biases.

The mechanism behind the acquisition of SPIs is horizontal gene transfer (18, 19). Nevertheless, the maintenance of SPIs within the genomes is considered stable (38) and is therefore a good measure of the barriers of RM systems compared to plasmid replicons and AMR genes, which can easily be lost if they do not confer any beneficial traits to the host. Despite this speculation, an influence of RM systems on the distribution of the SPIs was not supported by our analysis—on the contrary, there were indications of some SPIs being serovar specific, which corresponds to previous findings (38).

In conclusion, recombination and rearrangement events caused by RM systems are, in several cases, described as driving factors for evolution, contributing to the diversity within a species (4, 39–44). However, high recombination between two distantly related lineages of *S. enterica* is exceptional (6, 45), thus explaining the difficulties of linking the RM systems to the evolution of *S. enterica* subsp. *enterica*. Thus, recombination occurs within and between closely related serovars (6).

In this study, we showed that RM systems could not be linked to the evolution of *S. enterica* subsp. *enterica*, very likely due to the incomplete barriers of RM systems in conjugation. However, we observed closely related serovars with identical RM systems, i.e., *S. enterica* subsp. *enterica* serovars Dublin and Enteritidis, suggesting that to elaborate further on the hypothesis of RM systems being involved in the evolution of *Salmonella enterica* subsp. *enterica*, either a collection of closely related serovars or a more comprehensive data set with multiple representatives from each serovar could be assessed to expand on the hypothesis that the evolution of subgroups of *S. enterica* subsp. *enterica* RM systems could have stronger links between their genomic evolution and the presence of RM systems compared to the lack of association for the entire subspecies *enterica*.

## MATERIALS AND METHODS

**Salmonella strains.** From an in-house strain collection at the Technical University of Denmark, National Food Institute (DTU FOOD), a subcollection of 68 *S. enterica* subsp. *enterica* isolates with global origins and a focus on multidrug resistance was submitted to the 100K Food Pathogen Genome Project (<http://100kgenome.vetmed.ucdavis.edu/>, NCBI BioProject accession number [PRJNA186441](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA186441)) for WGS. Subsequently, the genomes from that project were merged with a genomic collection consisting of 105 *Salmonella* strains mainly originating from the American Type Culture Collection, often pansusceptible, which were sequenced by the U.S. Food and Drug Administration Center for Food Safety and Applied Nutrition (FDA-CFSAN) and U.S. Department of Agriculture (USDA) (1), and with another 48 publicly available *Salmonella* genomes retrieved from the European Nucleotide Archive for inclusion in this study. The final data set of 216 *S. enterica* subsp. *enterica* genomes was constructed with a focus on diversity and included a total of 97 different *Salmonella* serovars. Additionally, five genomes of four other subspecies were included in the data set to form an outgroup. This data set might have a built-in bias with respect to the selective criteria, e.g., susceptibility to antimicrobial agents. Full genomic information is shown in Table S1 in the supplemental material.

**Whole-genome sequencing.** Nutrient agar sticks containing the 68 in-house *Salmonella* isolates were dispatched to the School of Veterinary Medicine, UC Davis, CA, USA, in relation to the 100K Foodborne Pathogen Genome Project. The genomic DNAs (gDNAs) were extracted from overnight cultures using KAPA enzyme lysis buffer and the DNeasy blood and tissue kit from Qiagen. The gDNAs were subsequently fragmented to average sizes of 200 to 450 bp using the Diogenode Bioruptor NGS or Covaris E220. Sequencing libraries were prepared using the KAPA high-throughput (HTP) library preparation kit for Illumina platforms. Briefly, the fragmented gDNAs were adenylated, end paired, and ligated to NEXTflex-96 DNA bar code-indexed sequencing adaptors (Bioo Scientific). Following the ligation, indexed double-stranded DNAs (dsDNAs) were selected by size by using AmPure beads and then amplified by PCR. The indexed sequencing libraries were pooled and sequenced on an Illumina HiSeq 2000 with PE100 plus index read. The raw reads of the 68 sequenced genomes, received from UC Davis, were assembled using the Assembler 1.0 pipeline from the CGE, available on <http://cge.cbs.dtu.dk/services/all.php>, which is based on the Velvet algorithms for *de novo* short-read assembly. A complete list of genomic sequence data is available in Table S1 in the supplemental material.

**Construction of core and pangenome plot and pan- and core genome trees.** Open reading frames (ORFs) were predicted on the contigs by using the Prodigal software (46). The same gene predictor was subsequently used to eliminate biases in annotation quality and to standardize the genes found in all genomes (47).

The predicted genes were translated into amino acid sequences and aligned all-against-all using BLASTP (48). Two genes were determined as a gene pair if the alignment length was at least 50% of the longest sequence and more than 50% of the aligned sequences were reported as identical (the 50/50 rule). As a member of a gene pair can be a member of other gene pairs by this method, all gene pair-sharing members were subsequently combined into a gene family, which ensures that each gene will belong exclusively to only one gene family (28, 47, 49–53).

**Pan- and core genome plot.** The core and pangenome plot was constructed by comparing the gene families from all genomes. The pangenome was constructed from the union of the genes from the genomes under consideration, while the core genome was built from the intersection of gene families shared by every genome under analysis (27, 28).

**Pangenome tree.** The pangenome tree was reconstructed from a matrix consisting of gene families (rows) and genomes (columns). In the matrix, the absence and presence of genes across the genomes were represented by 0's and 1's, respectively. The genomes were clustered using hierarchical clustering of the relative Manhattan distance between genomes, and the bootstrap values were calculated to represent the confidence level of branches (27, 54).

**Core genome tree.** The core genes were aligned in a BLAST-like manner, using BLAT version 35 (55), to the predicted genes of each genome. The genes found in all genomes were then aligned using MUSCLE version 3.8.31 (56) and concatenated to a single alignment. Five hundred resamples of the alignment were created with Seqboot version 3.67 (part of the PHYLIP package [57]).

A gene was considered identified according to the 50/50 rule. DNADist (57) was employed to calculate the genomic distances from the initial alignment, as well as each of the 500 resamples. FastMe (58) was used to calculate trees from the distance matrices. The tree from the original alignment was compared to the 500 trees created from the resamples by using CompareToBootstrap (59). The final tree was visualized with FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>).

**Construction of RM-Finder and SPI-Finder.** To be able to analyze the genomes for their content of RM systems and SPIs, two publically available online tools were developed. Both tools, Restriction-ModificationFinder (<https://cge.cbs.dtu.dk/services/Restriction-ModificationFinder/>) and SPI-Finder (<https://cge.cbs.dtu.dk/services/SPIFinder/>), were built on a BLAST-based methodology for detection of genes from customized databases, originally developed by Zankari et al. for *in silico* detection of acquired resistance genes (60). The tools were developed to process both preassembled genomes and data of raw reads from different sequencing platforms, with user selection parameters of minimum percent identity (%ID) and minimum length. The default settings were chosen as minimum %ID at 95%, to avoid noise and fragments of the genes, and a minimum length of 60%, to be able to detect genes in the start or end of contigs from bad assemblies.

The database behind RM-Finder originates from the authoritative source REBASE (20, 33) and includes type I to IV restriction genes, methyltransferases, and specificity units. The database is categorized into two groups, one only including genes with confirmed functions and one where putative genes are included. The RM-Finder database is updated monthly.

The *Salmonella enterica* records from the PATHogenicity Island DataBase (PAIDB) served as inspiration for the customized database behind SPI-Finder (61, 62).

Extensive explanations of the output can be accessed on separate tabs at the Web sites for the online tools, <https://cge.cbs.dtu.dk/services/Restriction-ModificationFinder/> and <https://cge.cbs.dtu.dk/services/SPIFinder/>.

**Identification of RMS genes, plasmid replicons, SPIs, and antimicrobial resistance.** To analyze the content of RM systems in the 221 genomes, all ORFs were submitted to Restriction-ModificationFinder version 1.1, available on the CGE website (<https://cge.cbs.dtu.dk/services/Restriction-ModificationFinder/>), to identify restriction (R), methyltransferase (M), and specificity (S) genes. Subsequently, the R, M, and S genes identified were individually inspected to form RM systems, and putative systems were assigned when all genes required were present and adjacent on the contig, even if truncated or frame shifted. For systems with unknown specificity, systems were assigned according to the specificity subunit present. However, incomplete systems were investigated for truncated genes. Additionally, contigs with incomplete systems were inspected by BLAST against REBASE, and putative

genes for completion were revealed. Based on the predicted recognition sequences, the systems were merged and named according to the type of system (see Table S2 in the supplemental material). An RM system presence/absence matrix was constructed in R version 2.14 (<http://cran.r-project.org/bin/windows/base/old/2.14.0/>) with hierarchical clustering and Euclidean distance (Fig. 1; see also Fig. S1).

A phylogeny of the type III RM systems was created by extracting the ORFs for the methyltransferases and the restriction enzymes from the genomes, and the tree was constructed as described for the core genome tree and visualized by FigTree (see Fig. S4 in the supplemental material).

The 221 draft genomes were analyzed for their content of plasmid replicons, pathogenicity islands, and antimicrobial resistance genes by using the CGE Web tools PlasmidFinder 1.1 (21), SPIFinder 1.0 (<https://cge.cbs.dtu.dk/services/SPIFinder/>), and ResFinder 2.0 (60) with %ID of 80.00 and minimum length at 60.00%. The content of plasmid replicons and pathogenicity islands was interpreted by Circos plots (<http://circos.ca/>). For the antimicrobial resistance genes, a presence/absence matrix was constructed as described for the RM systems (see Fig. S2 in the supplemental material).

**Data availability.** Data underlying the findings are shared on figshare (<https://dx.doi.org/10.6084/m9.figshare.c.3247351>) (63), as follows: gene clusters of the 16,375 *Salmonella* pangenes for all genomes (Pan-gene\_clusters.tar.gz), gene clusters of the 2,138 *Salmonella* core genes for all genomes (Core-gene\_clusters.tar.gz), concatenated alignment for the core genome tree (Concatenated\_alignment\_core-genome\_tree.fa), raw tree file for the core genome tree (Core-genome\_tree.newick), raw tree file for the pangenome tree (Pan-genome\_tree.newick), output information from Restriction-ModificationFinder for each of the 221 genomes (RMFinder\_output.xlsx), concatenated alignment of type III RM systems (Concatenated\_alignment\_typelll\_RMsystems.fa), and raw tree file for type III RM systems (Type\_III\_tree.newick). The 68 *Salmonella* genomes sequenced as part of the 100K Foodborne Pathogen Genome Project (<http://100kgenome.vetmed.ucdavis.edu/>) are available under NCBI BioProject accession number PRJNA186441; individual accession numbers are listed in Table S1.

## SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at <http://dx.doi.org/10.1128/mSystems.00009-16>.

Figure S1, TIF file, 2.6 MB.

Figure S2, TIF file, 1.8 MB.

Figure S3, TIF file, 2.6 MB.

Figure S4, TIF file, 2.5 MB.

Table S1, XLSX file, 0.1 MB.

Table S2, XLSX file, 0.1 MB.

Text S1, TXT file, 6.3 MB.

Text S2, TXT file, 1.1 MB.

## ACKNOWLEDGMENT

This work was supported by the grant 11-106571 from the Danish Council for Independent Research—Technology and Production Sciences.

## FUNDING INFORMATION

This work, including the efforts of Louise Roer, was funded by Technology and Production Sciences (11-106571).

## REFERENCES

1. Timme RE, Pettengill JB, Allard MW, Strain E, Barrangou R, Wehnes C, Van Kessel JS, Karns JS, Musser SM, Brown EW. 2013. Phylogenetic diversity of the enteric pathogen *Salmonella enterica* subsp. *enterica* inferred from genome-wide reference-free SNP characters. *Genome Biol Evol* 5:2109–2123. <http://dx.doi.org/10.1093/gbe/evt159>.
2. Porwollik S, Boyd EF, Choy C, Cheng P, Florea L, Proctor E, McClelland M. 2004. Characterization of *Salmonella enterica* subspecies I genovars by use of microarrays. *J Bacteriol* 186:5883–5898. <http://dx.doi.org/10.1128/JB.186.17.5883-5898.2004>.
3. Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson M-A, Roy SL, Jones JL, Griffin PM. 2011. Foodborne illness acquired in the United States—major pathogens. *Emerg Infect Dis* 17:7–15. <http://dx.doi.org/10.3201/eid1701.P11101>.
4. Vasu K, Nagaraja V. 2013. Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol Mol Biol Rev* 77:53–72. <http://dx.doi.org/10.1128/MMBR.00044-12>.
5. Octavia S, Lan R. 2006. Frequent recombination and low level of clonality within *Salmonella enterica* subspecies I. *Microbiology* 152: 1099–1108. <http://dx.doi.org/10.1099/mic.0.28486-0>.
6. Didelot X, Bowden R, Street T, Golubchik T, Spencer C, McVean G, Sangal V, Anjum MF, Achtman M, Falush D, Donnelly P. 2011. Recombination and population structure in *Salmonella enterica*. *PLoS Genet* 7:e1002191. <http://dx.doi.org/10.1371/journal.pgen.1002191>.
7. Budroni S, Siena E, Dunning Hotopp JC, Seib KL, Serruto D, Nofroni C, Comanducci M, Riley DR, Daugherty SC, Angiuoli SV, Covacci A, Pizza M, Rappuoli R, Moxon ER, Tettelin H, Medini D. 2011. Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination. *Proc Natl Acad Sci U S A* 108:4494–4499. <http://dx.doi.org/10.1073/pnas.1019751108>.
8. Tock MR, Dryden DT. 2005. The biology of restriction and anti-restriction. *Curr Opin Microbiol* 8:466–472. <http://dx.doi.org/10.1016/j.mib.2005.06.003>.
9. Vasu K, Nagamalleswari E, Nagaraja V. 2012. Promiscuous restriction is a cellular defense strategy that confers fitness advantage to bacteria. *Proc Natl Acad Sci U S A* 109:E1287–E1293. <http://dx.doi.org/10.1073/pnas.1119226109>.
10. Veiga H, Pinho MG. 2009. Inactivation of the Saul type I restriction-modification system is not sufficient to generate *Staphylococcus aureus* strains capable of efficiently accepting foreign DNA. *Appl Environ Microbiol* 75:3034–3038. <http://dx.doi.org/10.1128/AEM.01862-08>.

11. Roer L, Møller Aarestrup F, Hasman H. 2015. The EcoKI type I restriction-modification system in *Escherichia coli* affects but is not an absolute barrier for conjugation. *J Bacteriol* **197**:337–342. <http://dx.doi.org/10.1128/JB.02418-14>.
12. Williams RJ. 2003. Restriction endonucleases: classification, properties, and applications. *Mol Biotechnol* **23**:225–243. <http://dx.doi.org/10.1385/MB:23:3:225>.
13. Murray NE. 2000. Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle). *Microbiol Mol Biol Rev* **64**:412–434. <http://dx.doi.org/10.1128/MMBR.64.2.412-434.2000>.
14. Bickle TA, Krüger DH. 1993. Biology of DNA restriction. *Microbiol Rev* **57**:434–450.
15. Roberts RJ, Belfort M, Bestor T, Bhagwat AS, Bickle TA, Bitinaite J, Blumenthal RM, Degtyarev SKh, Dryden DT, Dybvig K, Firman K, Gromova ES, Gumpert RI, Halford SE, Hattman S, Heitman J, Hornby DP, Janulaitis A, Jeltsch A, Josephsen J, Kiss A, Kleenhammer TR, Kobayashi I, Kong H, Krüger DH, Lacks S, Marinus MG, Miyahara M, Morgan RD, Murray NE, Nagaraja V, Piekarowicz A, Pingoud A, Raleigh E, Rao DN, Reich N, Repin VE, Selker EU, Shaw P-C, Stein DC, Stoddard BL, Szybalski W, Trautner TA, Van Etten JL, Vitor JMB, Wilson GG, Xu S. 2003. A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res* **31**:1805–1812. <http://dx.doi.org/10.1093/nar/gkg274>.
16. Rao DN, Dryden DT, Bheemanaik S. 2014. Type III restriction-modification enzymes: a historical perspective. *Nucleic Acids Res* **42**:45–55. <http://dx.doi.org/10.1093/nar/gkt616>.
17. Pallen MJ, Wren BW. 2007. Bacterial pathogenomics. *Nature* **449**:835–842. <http://dx.doi.org/10.1038/nature06248>.
18. Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW. 2009. Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiol Rev* **33**:376–393. <http://dx.doi.org/10.1111/j.1574-6976.2008.00136.x>.
19. Garcia-Vallvé S, Romeu A, Palau J. 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res* **10**:1719–1725. <http://dx.doi.org/10.1101/gr.130000>.
20. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE. <http://rebase.neb.com>.
21. Carattoli A, Zankari E, García-Fernández A, Larsen MV, Lund O, Villa L, Møller Aarestrup F, Hasman H. 2014. *In silico* detection and typing of plasmids using plasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* **58**:3895–3903. <http://dx.doi.org/10.1128/AAC.02412-14>.
22. Hamilton HL, Dillard JP. 2006. Natural transformation of *Neisseria gonorrhoeae*: from DNA donation to homologous recombination. *Mol Microbiol* **59**:376–385. <http://dx.doi.org/10.1111/j.1365-2958.2005.04964.x>.
23. Moser DP, Zarka D, Kallas T. 1993. Characterization of a restriction barrier and electrotransformation of the cyanobacterium *Nostoc PCC 7121*. *Arch Microbiol* **160**:229–237.
24. Ando T, Xu Q, Torres M, Kusugami K, Israel DA, Blaser MJ. 2000. Restriction-modification system differences in *Helicobacter pylori* are a barrier to interstrain plasmid transfer. *Mol Microbiol* **37**:1052–1065. <http://dx.doi.org/10.1046/j.1365-2958.2000.02049.x>.
25. Monk IR, Shah IM, Xu M, Tan MW, Foster TJ. 2012. Transforming the untransformable: application of direct transformation to manipulate genetically *Staphylococcus aureus* and *Staphylococcus epidermidis*. *mBio* **3**:e00277-11. <http://dx.doi.org/10.1128/mBio.00277-11>.
26. Ferguson GC, Heinemann JA, Kennedy MA. 2002. Gene transfer between *Salmonella enterica* serovar *Typhimurium* inside epithelial cells. *J Bacteriol* **184**:2235–2242. <http://dx.doi.org/10.1128/JB.184.8.2235-2242.2002>.
27. Leekitcharoenphon P, Lukjancenko O, Friis C, Aarestrup FM, Ussery DW. 2012. Genomic variation in *Salmonella enterica* core genes for epidemiological typing. *BMC Genomics* **13**:88. <http://dx.doi.org/10.1186/1471-2164-13-88>.
28. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angioli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proc Natl Acad Sci U S A* **102**:13950–13955. <http://dx.doi.org/10.1073/pnas.0506758102>.
29. Fricke WF, Mammel MK, McDermott PF, Tartera C, White DG, LeClerc JE, Ravel J, Cebula TA. 2011. Comparative genomics of 28 *Salmonella enterica* isolates: evidence for CRISPR-mediated adaptive sublineage evolution. *J Bacteriol* **193**:3556–3568. <http://dx.doi.org/10.1128/JB.00297-11>.
30. Fraser-Liggett CM. 2005. Insights on biology and evolution from microbial genome sequencing. *Genome Res* **15**:1603–1610. <http://dx.doi.org/10.1101/gr.3724205>.
31. Roush SW, Baldy LM. 2012. Manual for the surveillance of vaccine-preventable diseases. Centers for Disease Control and Prevention, Atlanta, GA. <http://www.cdc.gov/vaccines/pubs/surv-manual/>.
32. Rosenstein NE, Perkins BA, Stephens DS, Popovic T, Hughes JM. 2001. Meningococcal disease. *N Engl J Med* **344**:1378–1388. <http://dx.doi.org/10.1056/NEJM200105033441807>.
33. Roberts RJ, Vincze T, Posfai J, Macelis D. 2010. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res* **38**:D234–D236. <http://dx.doi.org/10.1093/nar/gkp874>.
34. Dahlberg C, Chao L. 2003. Amelioration of the cost of conjugative plasmid carriage in *Escherichia coli* K12. *Genetics* **165**:1641–1649.
35. Bergstrom CT, Lipsitch M, Levin BR. 2000. Natural selection, infectious transfer and the existence conditions for bacterial plasmids. *Genetics* **155**:1505–1519.
36. Platt TG, Bever JD, Fuqua C. 2012. A cooperative virulence plasmid imposes a high fitness cost under conditions that induce pathogenesis. *Proc Biol Sci* **279**:1691–1699. <http://dx.doi.org/10.1098/rspb.2011.2002>.
37. Müller K, Aabo S, Birk T, Mordhorst H, Bjarnadóttir B, Agersø Y. 2012. Survival and growth of epidemically successful and nonsuccessful *Salmonella enterica* clones after freezing and dehydration. *J Food Prot* **75**:456–464. <http://dx.doi.org/10.4315/0362-028X.JFP-11-167>.
38. Hensel M. 2004. Evolution of pathogenicity islands of *Salmonella enterica*. *Int J Med Microbiol* **294**:95–102. <http://dx.doi.org/10.1016/j.ijmm.2004.06.025>.
39. Dressler D, Potter H. 1982. Molecular mechanisms in genetic recombination. *Annu Rev Biochem* **51**:727–761. <http://dx.doi.org/10.1146/annurev.bi.51.070182.003455>.
40. Milkman R, Raleigh EA, McKane M, Cryderman D, Bilodeau P, McWeeny K. 1999. Molecular evolution of the *Escherichia coli* chromosome. V. Recombination patterns among strains of diverse origin. *Genetics* **153**:539–554.
41. King G, Murray NE. 1994. Restriction enzymes in cells, not eppendorfs. *Trends Microbiol* **2**:465–469. [http://dx.doi.org/10.1016/0966-842X\(94\)90649-1](http://dx.doi.org/10.1016/0966-842X(94)90649-1).
42. Chang S, Cohen SN. 1977. In vivo site-specific genetic recombination promoted by the EcoRI restriction endonuclease. *Proc Natl Acad Sci U S A* **74**:4811–4815. <http://dx.doi.org/10.1073/pnas.74.11.4811>.
43. Arber W. 2000. Genetic variation: molecular mechanisms and impact on microbial evolution. *FEMS Microbiol Rev* **24**:1–7. <http://dx.doi.org/10.1111/j.1574-6976.2000.tb00529.x>.
44. Furuta Y, Abe K, Kobayashi I. 2010. Genome comparison and context analysis reveals putative mobile forms of restriction-modification systems and related rearrangements. *Nucleic Acids Res* **38**:2428–2443. <http://dx.doi.org/10.1093/nar/gkp1226>.
45. Didelot X, Achtman M, Parkhill J, Thomson NR, Falush D. 2007. A bimodal pattern of relatedness between the *Salmonella* Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination? *Genome Res* **17**:61–68. <http://dx.doi.org/10.1101/gr.5512906>.
46. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**:119. <http://dx.doi.org/10.1186/1471-2105-11-119>.
47. Jacobsen A, Hendriksen RS, Aarestrup FM, Ussery DW, Friis C. 2011. The *Salmonella enterica* pan-genome. *Microb Ecol* **62**:487–504. <http://dx.doi.org/10.1007/s00248-011-9880-1>.
48. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**:403–410. [http://dx.doi.org/10.1016/S0022-2836\(05\)80360-2](http://dx.doi.org/10.1016/S0022-2836(05)80360-2).
49. Lukjancenko O, Wassenaar TM, Ussery DW. 2010. Comparison of 61 sequenced *Escherichia coli* genomes. *Microb Ecol* **60**:708–720. <http://dx.doi.org/10.1007/s00248-010-9717-3>.
50. Kaas RS, Friis C, Ussery DW, Aarestrup FM. 2012. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse

- Escherichia coli* genomes. *BMC Genomics* **13**:577. <http://dx.doi.org/10.1186/1471-2164-13-577>.
51. **Friis C, Wassenaar TM, Javed MA, Snipen L, Lagesen K, Hallin PF, Newell DG, Toszeghy M, Ridley A, Manning G, Ussery DW.** 2010. Genomic characterization of *Campylobacter jejuni* strain M1. *PLoS One* **5**:e12253. <http://dx.doi.org/10.1371/journal.pone.0012253>.
  52. **Vesth T, Wassenaar TM, Hallin PF, Snipen L, Lagesen K, Ussery DW.** 2010. On the origins of a *Vibrio* species. *Microb Ecol* **59**:1–13. <http://dx.doi.org/10.1007/s00248-009-9596-7>.
  53. **Lukjancenko O, Ussery DW, Wassenaar TM.** 2012. Comparative genomics of *Bifidobacterium*, *Lactobacillus* and related probiotic genera. *Microb Ecol* **63**:651–673. <http://dx.doi.org/10.1007/s00248-011-9948-y>.
  54. **Snipen L, Ussery DW.** 2010. Standard operating procedure for computing pangenome trees. *Stand Genomic Sci* **2**:135–141. <http://dx.doi.org/10.4056/sigs.38923>.
  55. **Kent WJ.** 2002. BLAT—the BLAST-Like Alignment Tool. *Genome Res* **12**:656–664. <http://dx.doi.org/10.1101/gr.229202>.
  56. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**:1792–1797. <http://dx.doi.org/10.1093/nar/gkh340>.
  57. **Felsenstein J.** 1989. PHYLIP—phylogeny inference package 3.6. *Cladistics* **5**:164–166.
  58. **Desper R, Gascuel O.** 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J Comput Biol* **9**:687–705. <http://dx.doi.org/10.1089/106652702761034136>.
  59. **Price MN, Dehal PS, Arkin AP.** 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* **5**:e9490. <http://dx.doi.org/10.1371/journal.pone.0009490>.
  60. **Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV.** 2012. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* **67**:2640–2644. <http://dx.doi.org/10.1093/jac/dks261>.
  61. **Yoon SH, Park Y-K, Lee S, Choi D, Oh TK, Hur C-G, Kim JF.** 2007. Towards pathogenomics: a web-based resource for pathogenicity islands. *Nucleic Acids Res* **35**:D395–D400. <http://dx.doi.org/10.1093/nar/gkl790>.
  62. **Yoon SH, Park Y-K, Kim JF.** 2015. PAIDB v2.0: exploration and analysis of pathogenicity and resistance islands. *Nucleic Acids Res* **43**:D624–D630. <http://dx.doi.org/10.1093/nar/gku985>.
  63. **Roer L, Hendriksen RS, Leekitcharoenphon P, Lukjancenko O, Kaas RS, Hasman H, Aarestrup FM.** 2016. Data collection for “Is the evolution of *Salmonella enterica* subspecies *enterica* linked to restriction-modification systems?” figshare. <http://dx.doi.org/10.6084/m9.figshare.c.3247351.v1>.