BMC Genomics

Similar yet different: phylogenomic analysis to delineate *Salmonella* and *Citrobacter* species boundaries

Ana Victoria C. Pilar[1†], Nicholas Petronella[2†], Forest M. Dussault[2], Adrian J. Verster[2], Sadjia Bekal[3], Roger C. Levesque[4], Lawrence Goodridge[5,6] and Sandeep Tamber[1*]

Abstract

Background: *Salmonella enterica* is a leading cause of foodborne illness worldwide resulting in considerable public health and economic costs. Testing for the presence of this pathogen in food is often hampered by the presence of background microflora that may present as *Salmonella* (false positives). False positive isolates belonging to the genus *Citrobacter* can be difficult to distinguish from *Salmonella* due to similarities in their genetics, cell surface antigens, and other phenotypes. In order to understand the genetic basis of these similarities, a comparative genomic approach was used to define the pan-, core, accessory, and unique coding sequences of a representative population of *Salmonella* and *Citrobacter* strains.

Results: Analysis of the genomic content of 58 *S. enterica* strains and 37 *Citrobacter* strains revealed the presence of 31,130 and 1540 coding sequences within the pan- and core genome of this population. Amino acid sequences unique to either *Salmonella* ($n = 1112$) or *Citrobacter* ($n = 195$) were identified and revealed potential niche-specific adaptations. Phylogenetic network analysis of the protein families encoded by the pan-genome indicated that genetic exchange between *Salmonella* and *Citrobacter* may have led to the acquisition of similar traits and also diversification within the genera.

Conclusions: Core genome analysis suggests that the *Salmonella enterica* and *Citrobacter* populations investigated here share a common evolutionary history. Comparative analysis of the core and pan-genomes was able to define the genetic features that distinguish *Salmonella* from *Citrobacter* and highlight niche specific adaptations.

Keywords: *Citrobacter*, *Salmonella*, Genomics, Core genome, Pan-genome, Network analysis

Background

Salmonella enterica subspecies *enterica* is one of the leading causes of foodborne illnesses in the world. In 2010, there were an estimated 153 million cases of illness worldwide due to the presence of this pathogen [1]. The majority of salmonellosis cases are associated with the ingestion of contaminated meat and poultry products. However, during the past decade, an increasing number of outbreaks in Canada and the United States have been associated with contaminated produce [2–4].

Salmonella can be isolated from many environmental niches including soil, water, and the gastrointestinal systems of animals. The genus consists of two species, *S. enterica* and *S. bongori*. The majority of human infections are attributed to *S. enterica*, which comprises 6 subspecies with over 2500 serological variants (serovars) that are characterized by distinct antigenic profiles [5, 6].

* Correspondence: sandeep.tamber@canada.ca
†Ana Victoria C. Pilar and Nicholas Petronella contributed equally to this work.
[1]Bureau of Microbial Hazards, Health Canada, Ottawa, Ontario, Canada
Full list of author information is available at the end of the article

Pilar *et al. BMC Genomics*   (2020) 21:377

Page 2 of 13

Members of different serovars may exhibit phenotypic differences with respect to pathogenicity, host restriction, resistance to antibiotics, and metabolism. However, these differences can also exist among strains of the same serovar [7–9]. The evolutionary history of *Salmonella*, as constructed from single nucleotide polymorphism (SNP) matrix analysis of 156 genomes, depicts groups of serovars that have monophyletic and polyphyletic lineages [10].

Accurate and precise detection of *Salmonella* in foods is fundamental to ensuring a safe and adequate food supply. Diagnostic tests lacking sensitivity can lead to false negative results through a failure to detect the pathogen. Conversely, results lacking specificity can lead to false positive results through the misidentification of a non-pathogenic organism as a pathogen. During the detection of *Salmonella* in food, false positive results frequently arise due to the presence of *Citrobacter* [11, 12]. Members of this genus resemble *Salmonella* more than any other genera in the family *Enterobacteriaceae* [13, 14]. The two genera possess similar metabolic and antigenic properties. Furthermore, given the diverse nature of both genera, atypical strains may lack typical diagnostic features or may have acquired novel ones that confound their precise identification [1, 13, 15, 16].

As with *Salmonella*, *Citrobacter* is frequently isolated from soil, water, and the digestive tract of animals [17]. Werkman and Gillen originally described members of this genus as intermediate forms between the genera *Escherichia* and *Aerobacter* and defined seven species of *Citrobacter* based on carbohydrate fermentation and gelatin liquefaction [18]. The inclusion of additional intermediate strains bearing similarities to other bacterial species significantly increased the complexity of the genus [19]. DNA hybridization analysis of 116 strains of *Citrobacter* defined 11 genomospecies within the genus [15]. Subsequent analysis of the 16S rRNA sequence and multilocus sequence typing (MLST), indicated three phylogenetic groupings for *Citrobacter*: I (*C. freundii*, *C. youngae*, *C. braakii*, *C. werkmanii*, *C. gillenii*, and *C. murliniae*); II (*C. amalonaticus*, *C. farmeri*, *C. sedlakii*, and *C. rodentium*); and III (*C. koseri*) with groups II and III being more closely related to *Salmonella* and other *Enterobacteriaceae* than group I [16].

Phylogenetic analysis of *Enterobacteriaceae* family members illustrate the complex relationship between closely related genera. The polyphyletic origins of *Citrobacter* have been shown by comparing the sequences of four *Citrobacter* species with those of several *Salmonella*, *Escherichia*, *Klebsiella*, *Enterobacter*, and *Shigella* strains [20]. Additional work by Retchless and Lawrence demonstrated the phylogenetic incongruence or conflicting topologies in the clade containing *Escherichia coli*, *Salmonella,* and *Citrobacter* [21]. This incongruence

may be attributed to recombination between the ancestral taxa of these organisms and the gradual acquisition of recombination barriers. Different regions of the genomes, having different rates of recombination, may undergo genetic isolation due to sequence divergence and ecological adaptation. The gradual formation of barriers to recombination may eventually give rise to heterogeneous subpopulations with orthologous genes of differing lineages and less clear-cut species definitions [21–23]. The ambiguous relationships between *Salmonella*, *Citrobacter,* and *E. coli* present a practical issue in the area of food safety testing and medical diagnosis, as accurate species identification is critical for confirming the presence of pathogens.

To gain insights into the genetic features delineating *Salmonella* and *Citrobacter*, the genomic features of a set of 58 *S. enterica* and 37 *Citrobacter* strains were investigated. This strain collection was chosen to represent the diversity of these two bacterial groups that might be expected on Canadian produce and included five *Citrobacter* strains that were falsely identified as *Salmonella* during initial testing. The objectives of the analysis were to define the core and pan genome of this population of isolates and to identify regions contributing to their diversity that may have impacted the evolution of *Salmonella* and *Citrobacter*.

Results

Characterization of *Citrobacter* strains falsely identified as *Salmonella*

During routine testing of lettuce samples, an industrial producer isolated five presumptive strains testing positive for *Salmonella* using a commercially available rapid identification method. Upon further investigation, these strains were determined to be *Citrobacter* and provided for us to use in this study. Morphologically, these five strains (S646, S647, S648, S1284, and S1285) produced colonies indicative of *Citrobacter* on the following selective/differential agar media; xylose lysine deoxycholate, brilliant green sulfa, triple sugar iron, and lysine iron agars. The five strains were all positive for glucose and sucrose fermentation, gas production, hydrogen sulfide production, and negative for lysine decarboxylation. All strains were negative for agglutination with the *Salmonella* O antiserum poly A-I + vi. However, the five strains produced positive reactions to the *Salmonella* Latex Test kit (Table S1). Variable results were also noted with tests based on biochemical and metabolic profiles (Table S1). API20E provided identification profiles for four of the five strains as *C. freundii* (90.8% identification accuracy, T = 0.65) while strain S1284 did not have a valid identification profile. The closest identified taxon for S1284 was *C. braakii* with 74.4% identity and a T value of 0.49. VITEK 2 identified S1284 as *C. sedlakii* with an 86%

Pilar *et al. BMC Genomics*   (2020) 21:377

Page 3 of 13

match but was unable to provide an identity for strains S646, S647, S648, and S1285. These same strains were only identified to the genus level as *Citrobacter* (scores ranging from 2.1–2.2) using the MALDI Biotyper whereas strain S1284 was identified as *C. braakii* (score = 2.5).

A maximum likelihood phylogenetic tree based on the 16S rRNA sequence of the five false positive *Citrobacter* strains, seven unspeciated *Citrobacter* produce isolates and members of the family *Enterobacteriaceae* (Table S2 and Figure S1) demonstrated that members of the *Citrobacter* genus did not form a distinct clade. Strain S1284 was in a clade with *C. werkmanii* NBRC 105721 while strains S646, S647, S648, and S1285 appeared to be closely related to *C. koseri* ATCC BAA895. The unspeciated produce isolates were in separate nodes on the tree (Groups 1 and 2) and clustered with *C. braakii* and *C. freundii*. The tree was characterized by low nodal support, providing limited taxonomic resolution between *Citrobacter*, *E. coli*, and *Salmonella*.

The average nucleotide identity (ANI) based on pair-wise genome sequence comparison of the five *Citrobacter* strains with *Salmonella* and *Citrobacter* reference genomes (Table S3) was between 84 and 86% for strains S646, S647, S648 and S1285, which precluded the determination of their species identity. *Citrobacter* strain S1284 had an ANI of 94%, which was just below the threshold value for genetic similarity to *C. braakii*.

Genomic analysis of *Salmonella* and *Citrobacter* pan-genome

A comparative analysis of the 58 *Salmonella* and 37 *Citrobacter* genomes was performed to gain insight into the genetic differences between these two groups of bacteria. The 58 *S. enterica* ssp. *enterica* strains chosen for this study encompass the diversity of the subspecies and include strains that may be associated with fresh produce in Canada (Table 1). Of the 38 serovars selected, 19 included representatives of serovars frequently associated with clinical salmonellosis in Canada. The five most commonly reported Canadian serovars were represented by multiple strains; Enteritidis (*n* = 5), Heidelberg (*n* = 4), Javiana (*n* = 4), Newport (*n* = 5), and Typhimurium (*n* = 7). Serovars Typhi, Paratyphi A, and Gallinarum are host-restricted and the remaining 16 serovars (Table 1, grey box) are rarely associated with human disease. The 37 *Citrobacter* strains chosen for this study included human clinical isolates and food isolates including five strains that were falsely identified as *Salmonella* (Table 1, blue box). The genome size of the individual *S. enterica* and *Citrobacter* spp. strains ranged from 4 to 5 Mb and contained 4000 to 5000 genes (Table S4).

The pan-genome of all 95 strains consisted of 31,130 protein-coding genes. Reciprocal comparison of

Table 1 List of strains and genome sequences from the *Salmonella enterica* Foodborne Syst-OMICS (SalFoS) database that were used in this study. Highlighted strains indicate false positive *Salmonella* strains (blue) and rare *Salmonella* serovars (grey)

Strain name	SalFoS ID	Isolation source
Citrobacter amalonaticus	S63	Human
Citrobacter amalonaticus	S68	Human
Citrobacter braakii	S58	Unknown
Citrobacter braakii	S64	Human
Citrobacter braakii	S85	Human
Citrobacter braakii	S86	Human
Citrobacter farmeri	S80	Human
Citrobacter freundii	S20	Fresh produce
Citrobacter freundii	S61	Human
Citrobacter freundii	S66	Human
Citrobacter freundii	S67	Human
Citrobacter freundii	S83	Human
Citrobacter freundii	S89	Human
Citrobacter koseri	S65	Human
Citrobacter koseri	S79	Human
Citrobacter werkmanii	S78	Human
Citrobacter sp.	S60	Human
Citrobacter sp.	S77	Human
Citrobacter sp.	S81	Human
Citrobacter sp.	S91	Human
Citrobacter sp.	S93	Human
Citrobacter sp.	S94	Human
Citrobacter sp.	S95	Human
Citrobacter sp.	S96	Human
Citrobacter sp.	S97	Human
Citrobacter sp.	S646	Fresh produce
Citrobacter sp.	S647	Fresh produce
Citrobacter sp.	S648	Fresh produce
Citrobacter sp.	S1284	Fresh produce
Citrobacter sp.	S1285	Fresh produce
Citrobacter sp.	S649	Fresh produce
Citrobacter sp.	S1278	Fresh produce
Citrobacter sp.	S1279	Fresh produce
Citrobacter sp.	S1280	Fresh produce
Citrobacter sp.	S1281	Fresh produce
Citrobacter sp.	S1282	Fresh produce
Citrobacter sp.	S1283	Fresh produce
S. enterica ser. Anatum	S234	Spices
S. enterica ser. Braenderup	S208	Poultry
S. enterica ser. Brandenburg	S291	Seafood
Salmonella ser. Derby	S228	Seafood
S. enterica ser. Enteritidis	S186	Poultry
S. enterica ser. Enteritidis	S187	Fresh produce
S. enterica ser. Enteritidis	S424	Human
S. enterica ser. Enteritidis	S422	Human
S. enterica ser. Enteritidis	S420	Human
S. enterica ser. Hadar	S219	Poultry
S. enterica ser. Heidelberg	S190	Human
S. enterica ser. Heidelberg	S191	Nuts and seeds
S. enterica ser. Heidelberg	S370	Human
S. enterica ser. Heidelberg	S431	Human
S. enterica ser. Infantis	S198	Pasta
S. enterica ser. Javiana	S200	Human
S. enterica ser. Javiana	S201	Dairy
S. enterica ser. Javiana	S202	Seafood
S. enterica ser. Javiana	S203	Seafood
S. enterica ser. Kentucky	S245	Poultry
S. enterica ser. Litchfield	S272	Human
S. enterica ser. Montevideo	S239	Spices
S. enterica ser. Newport	S195	Nuts and seeds
S. enterica ser. Newport	S196	Fresh produce
S. enterica ser. Newport	S197	Frog
S. enterica ser. Newport	S566	Environmental
S. enterica ser. Newport	S2	Human
S. enterica ser. Oranienburg	S216	Dog
S. enterica ser. Poona	S307	Fresh produce
S. enterica ser. Saintpaul	S204	Nuts and seeds
S. enterica ser. Senftenberg	S271	Nuts and seeds
S. enterica ser. Tennessee	S343	Poultry
S. enterica ser. Typhimurium	S188	Poultry
S. enterica ser. Typhimurium	S189	Chocolate
S. enterica ser. Typhimurium	S415	Human
S. enterica ser. Typhimurium	S441	Environmental
S. enterica ser. Typhimurium	S373	Human
S. enterica ser. Typhimurium	S418	Human
S. enterica ser. Typhimurium	S437	Fresh produce
S. enterica ser. Paratyphi A	S399	Human
S. enterica ser. Paratyphi	S404	Human
S. enterica ser. Gallinarum	S173	Poultry
S. enterica ser. Amager	S25	Unknown
S. enterica ser. Chingola	S32	Seaweed
S. enterica ser. Daytona	S341	Seafood
S. enterica ser. Duesseldorf	S35	Poultry
S. enterica ser. Elisabethville	S36	Reptile
S. enterica ser. Fresno	S38	Animal feed
S. enterica ser. Idikan	S41	Nuts and seeds
S. enterica ser. Luciana	S43	Fresh produce
S. enterica ser. Ohio	S315	Nuts and seeds
S. enterica ser. Orientalis	S45	Nuts and seeds
S. enterica ser. Pasing	S46	Chocolate
S. enterica ser. Solt	S47	Nuts and seeds
S. enterica ser. Tado	S48	Animal feed
S. enterica ser. Taiping	S49	Seafood
S. enterica ser. Tyresoe	S51	Seafood
S. enterica ser. Weston	S54	Seafood

orthologous coding sequences in both genera identified 1540 core gene sequences, encoded by approximately a third of each isolate's genome (Table S5). These results were supported by a rarefaction analysis that demonstrated that the number of shared gene families decreased as more genomes were included in the analysis and reached a plateau at approximately 1500 genes (Fig. 1). Analysis of the orthologous relationships between the pan-genome revealed the presence of 195 *Citrobacter* gene sequences that did not have orthologues in *Salmonella* (i.e. *Citrobacter* unique gene sequences), while the corresponding number of gene sequences unique to *Salmonella* was 1112 (Table S5).
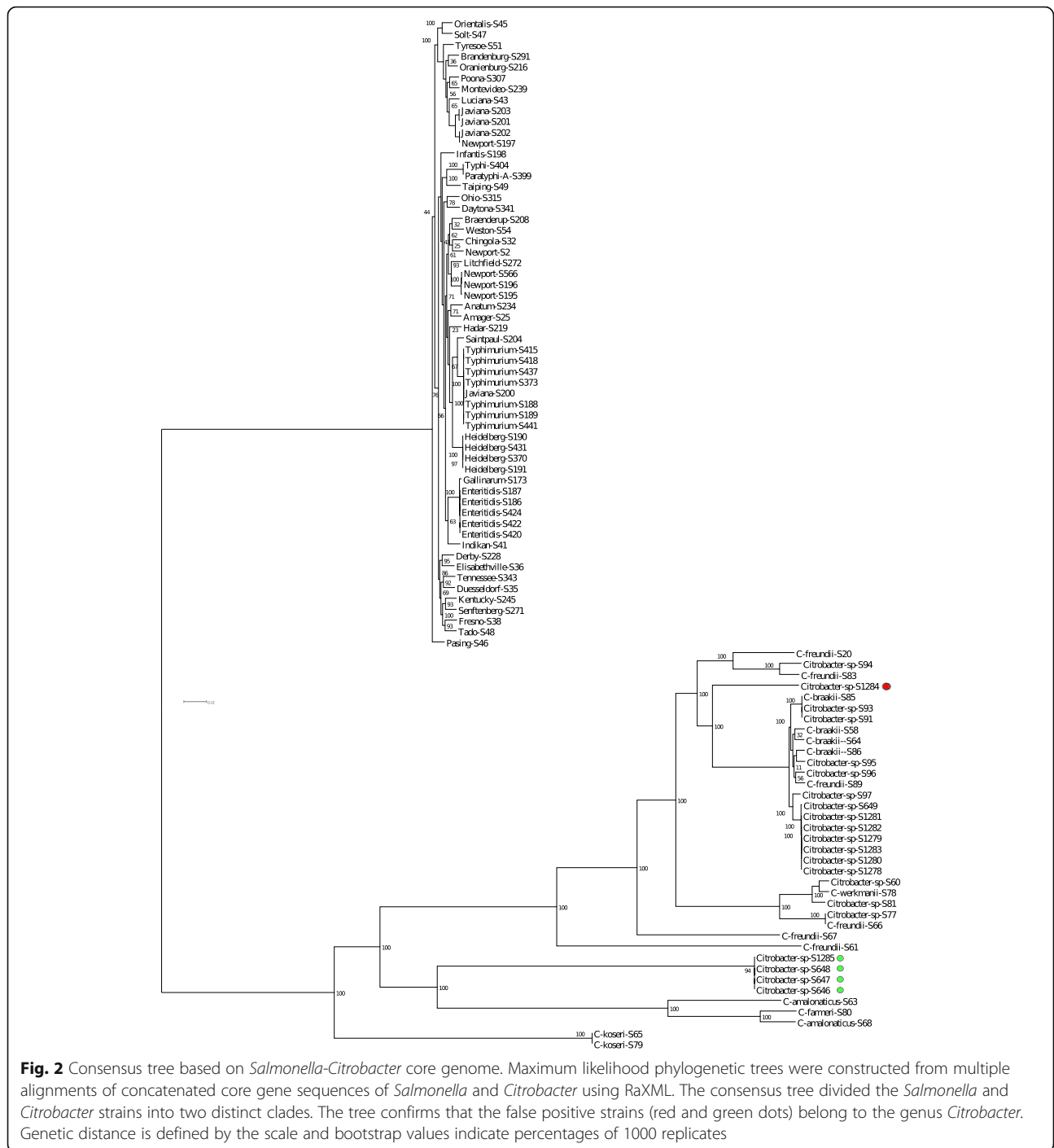
A maximum likelihood phylogenetic tree based on the genomic variations in the amino aicd sequences of the *Salmonella-Citrobacter* core genome effectively divided the 95 strains into one of two clades (Fig. 2). The topology of the *Salmonella* taxon differed from that of *Citrobacter*, showing low bootstrap values in some of the internal nodes and shorter genetic distances. The *Salmonella* strains grouped into 3 major clades with the first clade consisting of Javiana and 8 other serovars. A second diverse clade of strains contained four serovars rarely reported in Canada (Tado, Fresno, Duesseldorf, and Elisabethville), as well as representatives from serovars Kentucky, Senftenberg, and Tennessee. The largest clade resolved the two typhoidal strains from the rest of non-typhoidal *Salmonella* (NTS) strains, with the rare serovar Taiping being most related to Typhi and Paratyphi A. Within the NTS clade, terminal nodes with the

highest support included the clade consisting of the five Enteritidis and one Gallinarum strain, a result consistent with previous studies, as well as clades containing Typhimurium and Heidelberg [24, 25]. Two strains, Newport (S197) and Javiana (S200) were found in clades distinct from those containing other Newport or Javiana strains. The serogroup designation of these two strains were confirmed according to the Kauffman-White serotyping scheme [26]. Also included in the NTS clade were the rare serovars Weston and Chingola that shared an ancestor with strains belonging to serovars Newport and Litchfield. Polyphyletic lineages included serovars Javiana and Newport whose respective members were found in other clades. The core genes of serovar Pasing were the most divergent of the 58 *Salmonella* strains and formed the outgroup for this population.

Overall, the topology of the *Citrobacter* group bore resemblance to previous 16S rRNA and MLST-based trees that grouped 11 *Citrobacter* genomospecies into three previously described species complexes [15, 16]. This portion of the tree had well-supported nodes with the population falling into one of three major clades. The first clade contained the two *C. koseri* strains. The second clade contained a group of *C. amalonaticus* and *C. farmeri* strains that were closely related to four of the five strains falsely identified as *Salmonella* (S646, S647, S648, and S1285) (Fig. 2, green dots). Strain S1284 (red dot) was located in the third clade. It branched off from a larger group containing *C. braakii*, *C. freundii*, and a number of unspeciated strains. Several *C. freundii* strains
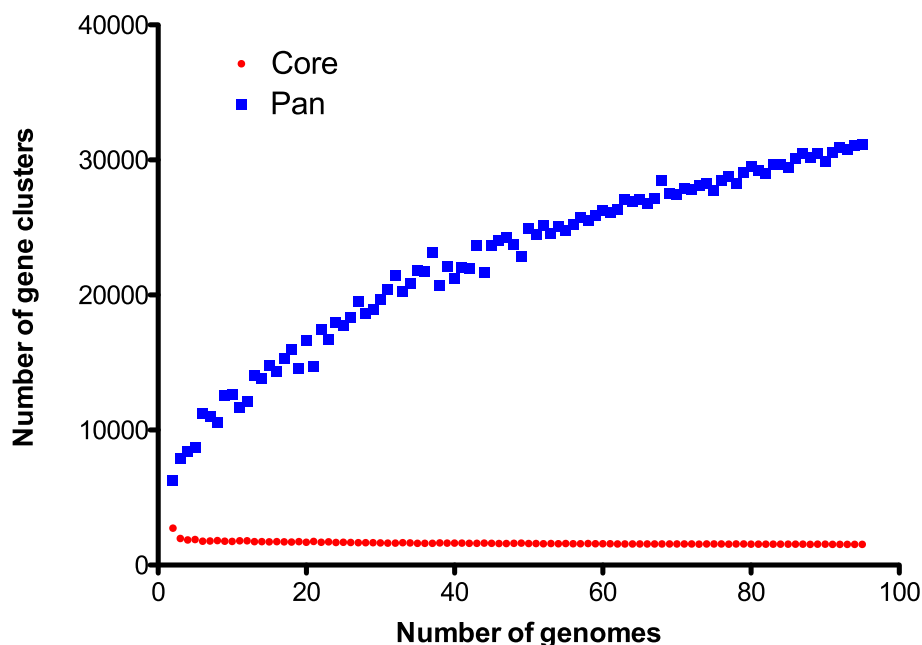
Fig. 1 Estimation of the *Salmonella-Citrobacter* pan- and core genome size. Rarefaction analysis demonstrates that the average number of gene clusters in the pan-genome increases while the conserved core gene clusters reaches a plateau as more genomes are analyzed

Pilar *et al. BMC Genomics*   (2020) 21:377

Page 5 of 13

Fig. 2 Consensus tree based on *Salmonella-Citrobacter* core genome. Maximum likelihood phylogenetic trees were constructed from multiple alignments of concatenated core gene sequences of *Salmonella* and *Citrobacter* using RaXML. The consensus tree divided the *Salmonella* and *Citrobacter* strains into two distinct clades. The tree confirms that the false positive strains (red and green dots) belong to the genus *Citrobacter*. Genetic distance is defined by the scale and bootstrap values indicate percentages of 1000 replicates

were found in different clades, confirming the polyphyletic nature of this species.

The heterogeneity and polyphyletic origin of *Citrobacter* species precluded the accurate taxonomic classification of the unspeciated strains in Fig. 2. Therefore, the core genomes of the *Citrobacter* strains in this study were compared to those of 160

Citrobacter strains from the RefSeq database (Table S6). A maximum likelihood phylogenetic tree based on the variations in the amino acid sequences of the core genome resolved the population into twelve clades (Fig. 3a-l). With a few exceptions, one species was predominant in each clade. This grouping pattern enabled the identification of many of the

Pilar et al. BMC Genomics   (2020) 21:377

Page 6 of 13
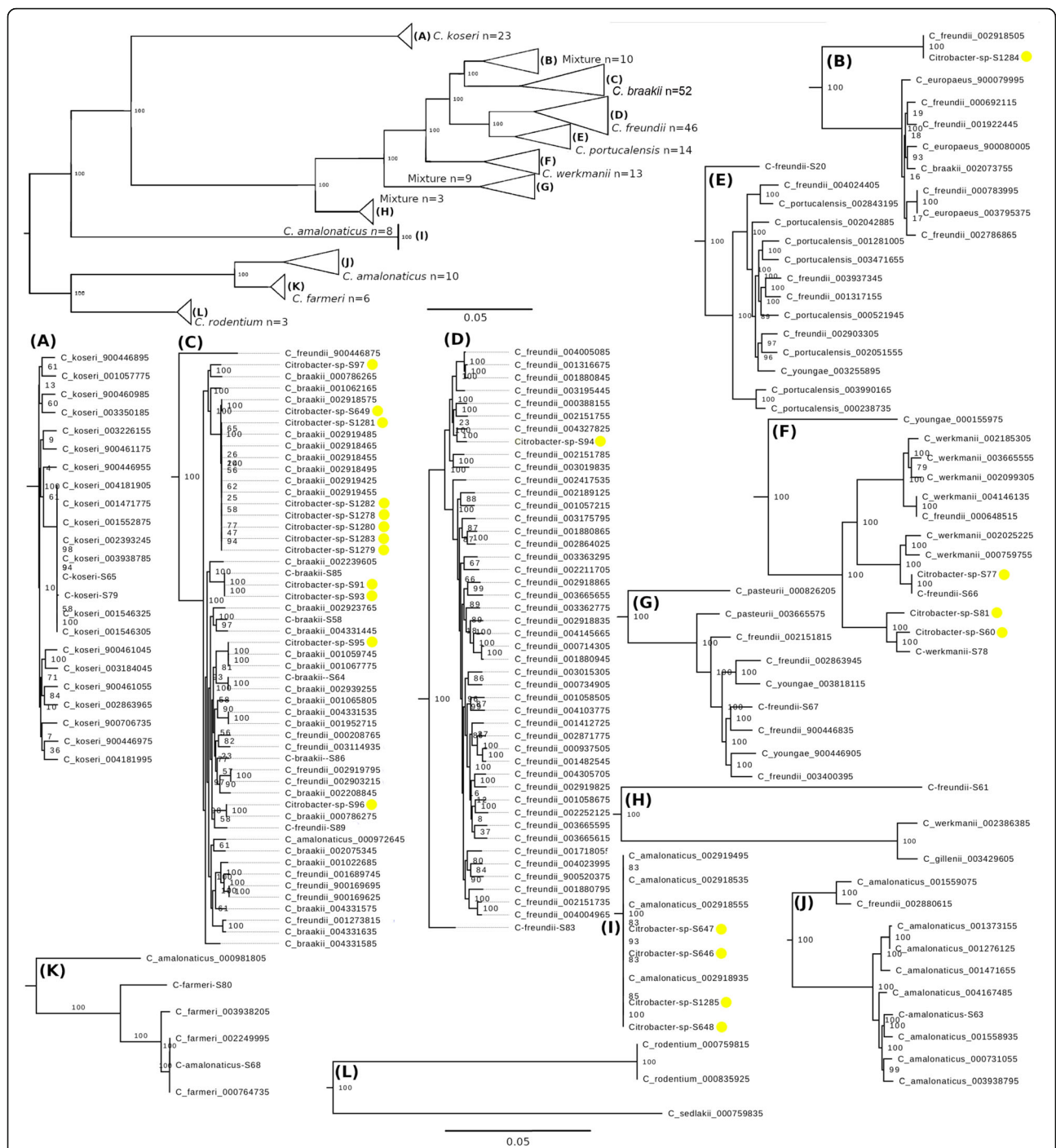
Fig. 3 *Citrobacter* core genome tree. A maximum likelihood phylogenetic tree constructed from the concatenated core amino acid sequences of 197 *Citrobacter* strains was used to determine the taxonomic designation of the unspeciated *Citrobacter* strains in this study (yellow dots). Genome distance estimation was performed using MASH and the core genome was analyzed using centreseq [27]. genes. Genetic distance is defined by the scale and bootstrap values indicate percentages of 1000 replicates. Triangles denote collapsed nodes with *n* representing the number of strains found in each triangle. For each triangle, the letters represent the group of related strains found in the branches of the node

unspeciated *Citrobacter* strains (Fig. 3, yellow dots). Eleven of the twenty-one unspeciated strains clustered with *C. braakii* strains (clade C). However, the species designation for strain S96 remained unclear.

Although this strain was found in the same cluster, where the majority of the strains were *C. braakii*, its proximity to several *C. freundii* and one *C. amalona-ticus* obscured its identification. Four of the five

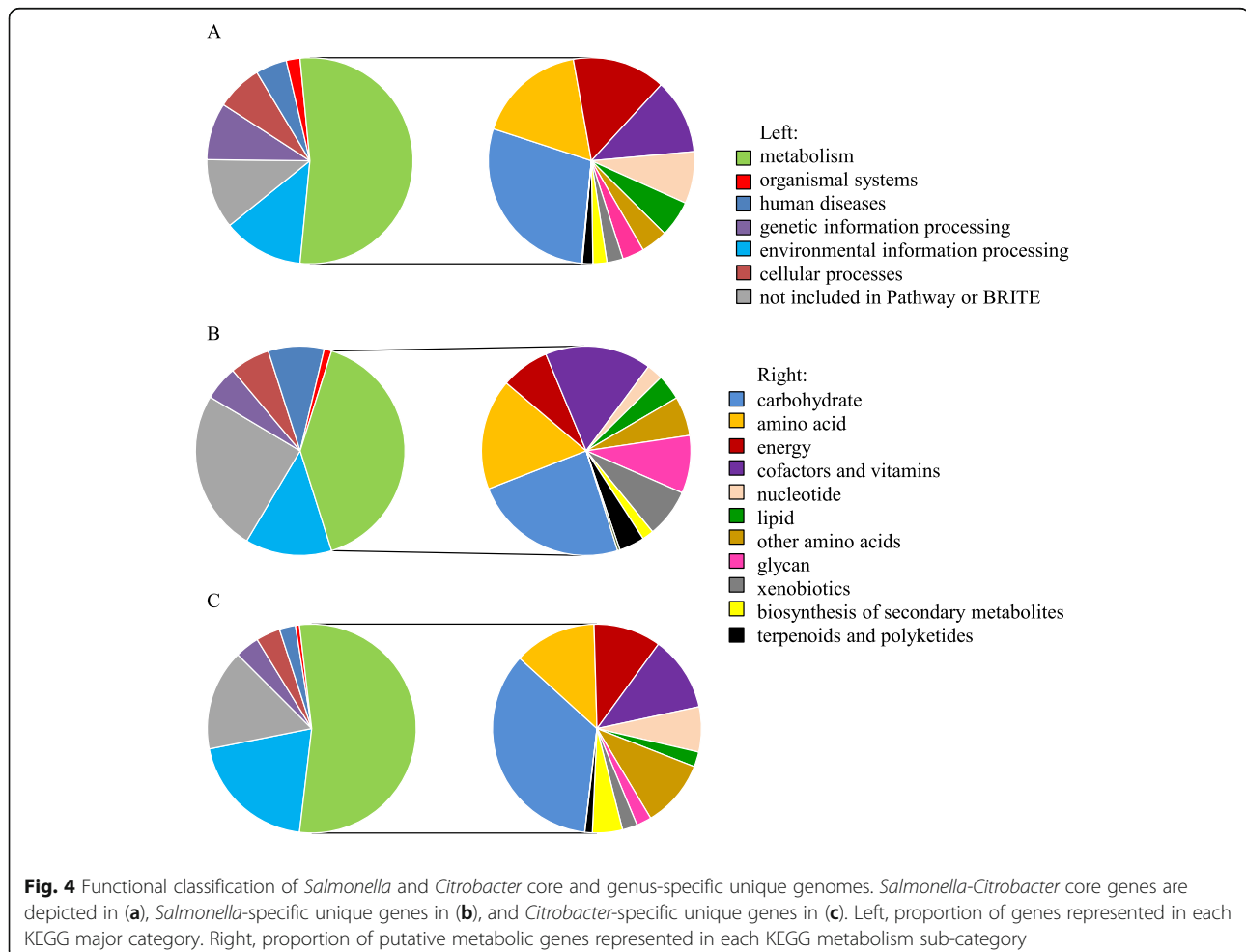Pilar *et al. BMC Genomics*   (2020) 21:377

Page 7 of 13

Citrobacter strains that were falsely identified as *Salmonella* were found in a cluster with several *C. amalonaticus* strains (clade I). The remaining false positive strain, S1284, was found in a clade containing a mixture of several *C. freundii* strains, three *C. europaeus* strains, and one *C. braakii* strain (clade B). However, within this clade, S1284 clustered with a *C. freundii* strain suggesting that the two are closely related.

Functional analysis of *Salmonella-Citrobacter* core genes

The functional classification of the following three sets of gene sequences were analyzed (Table S7); the *Salmonella* and *Citrobacter* core gene sequences, the *Salmonella*-specific unique gene sequences, and *Citrobacter*-specific unique gene sequences. The core gene sequences had 1525 predicted functions in 176 KEGG pathways. A significant number of these sequences, approximating 11% of the predicted core gene functions were designated as either poorly characterized or could not be classified into a KEGG pathway or BRITE

hierarchy. For the unique gene sequences specific to either *Salmonella* or *Citrobacter*, the respective totals were 521 and 135 predicted roles in 117 and 60 pathways. Poorly characterized genes made up a respective 28 and 16% of the *Salmonella* specific and *Citrobacter* specific unique gene sequences.

The majority of the core gene sequences were predicted to encode proteins with roles in the major KEGG categories of metabolism (60%), environmental information processing (14%), and genetic information processing (10%) (Fig. 4). Of the metabolic functions, the three major categories represented were for the metabolism of carbohydrates (17%), amino acids (10%), and energy (9%). In comparison, the major KEGG categories represented in the *Salmonella*-specific unique gene set were metabolism (54%), environmental information processing (18%), and human diseases (12%). Of the metabolic processes, the three major categories were carbohydrate metabolism (13%), amino acid metabolism (9%), and the metabolism of cofactors and vitamin (9%). In contrast, the majority of the functions unique to *Citrobacter* were

Fig. 4 Functional classification of *Salmonella* and *Citrobacter* core and genus-specific unique genomes. *Salmonella-Citrobacter* core genes are depicted in (**a**), *Salmonella*-specific unique genes in (**b**), and *Citrobacter*-specific unique genes in (**c**). Left, proportion of genes represented in each KEGG major category. Right, proportion of putative metabolic genes represented in each KEGG metabolism sub-category
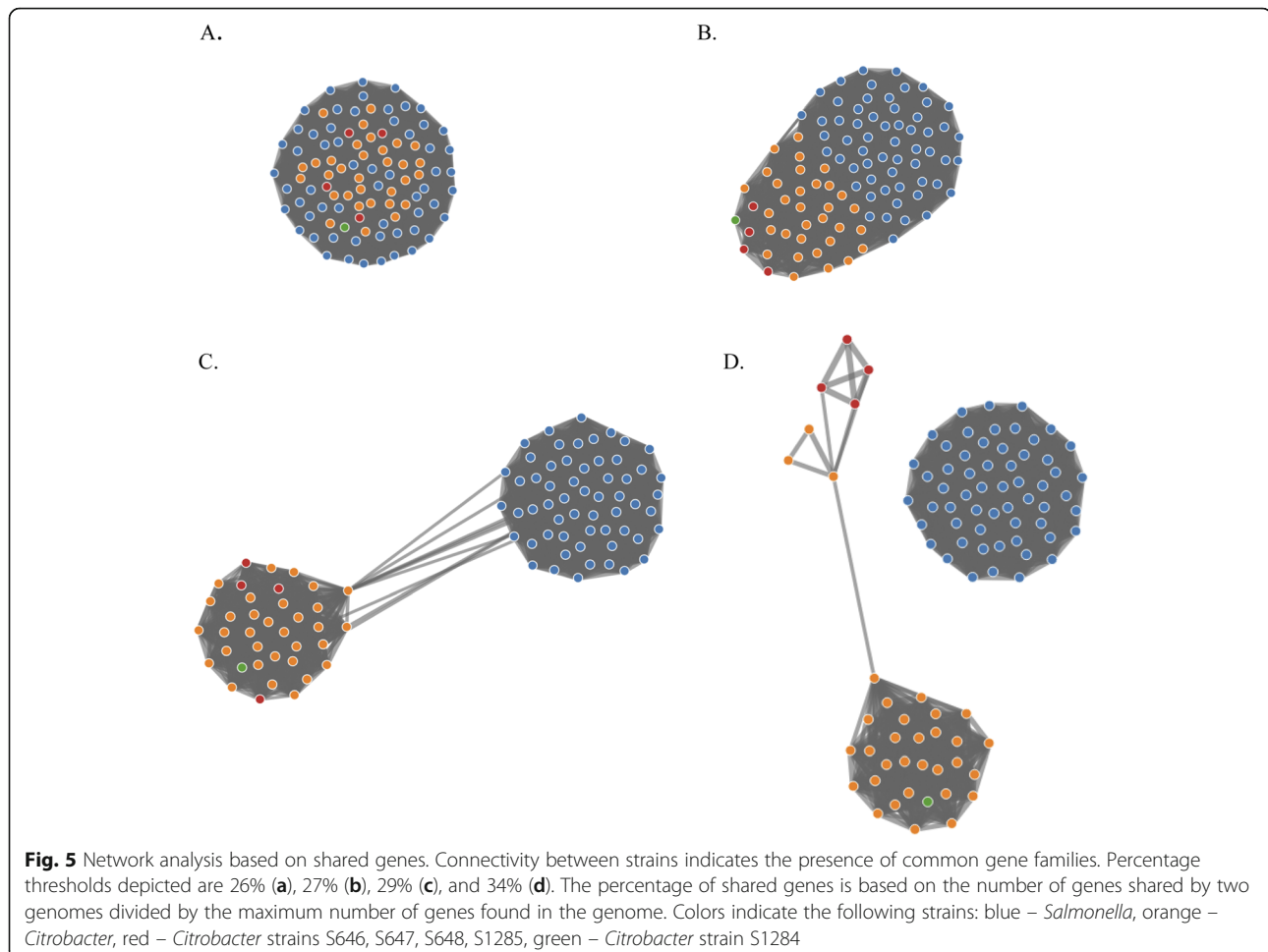
predicted to occur in two major categories; metabolism (64%) and environmental information processing (24%). As with *Salmonella*, the three major categories within metabolism were carbohydrate (22%), amino acid (8%), and cofactors and vitamin metabolism (7%).

The *Salmonella*-specific gene sequences were proportionately higher in some global metabolic pathways such two-component system, ABC transporters, lipopolysaccharide biosynthesis, porphyrin and chlorophyll metabolism, and sulfur metabolism. In addition, 20 *Salmonella*-specific sequences were classified under 12 pathways that were not present in the core and *Citrobacter*-specific gene sequence sets (Table S7, orange), which include geraniol degradation, naphthalene degradation, bacterial infectious disease, and pathways related to organismal systems and human diseases. Predicted metabolic activities of *Citrobacter*-specific gene functions had higher representation in pathways involved with ABC transporters, sulfur metabolism, and glyoxylate and dicarboxylate metabolism and two sequences were uniquely categorized under the atrazine degradation pathway (Table S7, orange).

Network analysis

Network analysis of the *Salmonella-Citrobacter* pangenome provided a complementary view of the evolutionary relationships between the population of strains. Figure 5 illustrates the genetic network of the *Salmonella* and *Citrobacter* strains based on the proportion of protein families shared between the two groups of bacteria. At 26% shared gene families, all of the strains, regardless of their taxonomic designation, formed one community of interconnected members (Fig. 5a). This recapitulates the results of the *Salmonella-Citrobacter* comparative genome analysis, wherein the core gene sequences made up ~ 30% of individual genomes. Increasing the proportion of shared sequences by 1% depicted a community with all the *Salmonella* (blue) strains on one end of the network while the *Citrobacter* strains (orange, red, and green) formed a gradient of decreasing interconnections with *Salmonella* (Fig. 5b). A threshold of 29% shared gene sequences (Fig. 5c) separated *Salmonella* and *Citrobacter* into two communities that remained connected due to the similarities of several *Salmonella* strains with individual *Citrobacter* strains. As

Fig. 5 Network analysis based on shared genes. Connectivity between strains indicates the presence of common gene families. Percentage thresholds depicted are 26% (**a**), 27% (**b**), 29% (**c**), and 34% (**d**). The percentage of shared genes is based on the number of genes shared by two genomes divided by the maximum number of genes found in the genome. Colors indicate the following strains: blue – *Salmonella*, orange – *Citrobacter*, red – *Citrobacter* strains S646, S647, S648, S1285, green – *Citrobacter* strain S1284

the percent threshold was increased to 34%, two distinct communities became apparent and shared no homologues (Fig. 5d). However, the network also depicted the separation of four of the *Citrobacter* strains that were falsely identified as *Salmonella* (red) from the rest of the *Citrobacter* network. However, the *Citrobacter* strain S1284 (green) that was also falsely identified as *Salmonella* maintained many connections within the larger *Citrobacter* community. In contrast, all of the *Salmonella* strains maintained one network due to their genetic similarity.

Discussion

Analysis of the core and pan-genome of a population of *Salmonella* and *Citrobacter* strains offered a reliable delineation of the fuzzy taxonomic boundaries between these closely related bacteria. As the most stable portion of the pan-genome, the core genome encodes characteristics common to all members of a species [28, 29]. Therefore, species identities are maintained despite considerable genomic flux that may occur between and within species. A recent study on the genomic structure of the *Rickettsiales* defined criteria to define genus and species assignments based on core genome alignments, with alignments ≥10% of the average input genome length and ≥ 96.8% identity respectively [30]. Here we show that despite the stringent parameters that were used to identify orthologous genes, approximately one third of each genome in this study was composed of core genes suggesting relationships between genera belonging to the family *Enterobacteriaceae* may be closer than in other bacterial families. A consensus tree based on sequence variations of the amino acid sequences encoded by the *Salmonella-Citrobacter* core genes was able to resolve the two bacterial species into two distinct clades and provided sufficient delineation to infer the taxonomic affiliation of the *Citrobacter* strains. Analysis of 37 *Citrobacter* strains proved insufficient in determining the taxonomic assignment of many of the unspeciated strains in this study. Adding more *Citrobacter* genomes to the analysis greatly increased the resolving power of the analysis. However, the identity of one strain, S96, remained ambiguous due to the uncertainty surrounding the identity of its closest homologues. Of the *Citrobacter* genomes investigated, the ones belonging to *C. freundii* appeared to be the most divergent, appearing in seven of the 12 clades. Given the complexity of *Citrobacter* taxonomy, it may be possible that some strains in the RefSeq database were classified erroneously. However, this observation also calls into question the utility of *C. freundii* as a species designation since based on our analysis, it appears to be a collection of highly disparate strains.

According to the results of the rarefaction analysis, the core genome defined in this study is believed to reflect the diversity of *Salmonella* and *Citrobacter* strains that may be associated with food commodities, specifically fresh produce in Canada. Many *Salmonella-Citrobacter* core gene sequences were predicted to have roles in central cellular processes including DNA replication, transcription, translation, cell division, and key metabolic pathways such as glycolysis and the tricarboxylic acid cycle. In contrast, the functional predictions of some genus-specific unique gene sequences were associated with activities required to thrive in a specific niche, such as specialized metabolic pathways. A limitation of the current study is that the KEGG database is primarily intended to catalogue eukaryotic (human) cellular functions. Thus, many bacterial gene functions are not represented. Just over half of the 1112 sequences identified as unique to our collection of *Salmonella* was mapped to cellular pathways. Given the redundancy expected when the activity of one gene makes contributions to multiple pathways, this number is reduced further. Continued study of bacterial gene functions along with the creation and continued curation of functional classification databases for bacterial genes would enable the more comprehensive and precise predictions of bacterial gene functions.

The stability of the core genome makes it an indispensable tool for defining the genetic integrity of bacterial populations. However, it is ill suited for investigating evolutionary relationships between closely related strains. A consensus tree based on the core genes of *Salmonella* and *Citrobacter* was not able to resolve the basal relationships between *Salmonella* serovars. There was greater support for the evolutionary relationships observed between many of the *Citrobacter* strains in our study, presumably due to the greater diversity of that population. However, many highly related strains were poorly resolved including the false positive strains and other isolates from fresh produce. It is possible that the overall strain diversity reported here is low as it is biased towards strains that were isolated through the context of food safety testing and outbreak investigations. In the absence of alternative sampling plans, the discriminatory power of phylogenetic trees can be increased by reducing the stringency in defining the core genome, or by the inclusion of intergenic or accessory regions. In this regard, public health laboratories have been successfully applying gene-by-gene comparison methods such as whole genome MLST (wgMLST) to outbreak investigations [31–33].

Accessory and singleton gene sequences represent the most dynamic regions of the genomes. They encode mobile elements, as well as many small proteins with putative or hypothetical functions. These elements evolve

quickly and are believed to confer niche-specific selective advantages, as they are continually being lost and gained among strains of multiple lineages [34, 35]. Phylogenetic network analysis provides a visual representation of events that drive bacterial evolution, such as the loss and gain of genes within the pan-genome. These events occur via recombination and horizontal gene transfer (HGT) and are not captured in a phylogenetic tree due to the non-tree-like nature of these processes [36]. Visualization of network connections can be altered by varying the threshold for the number of protein families shared by pairs of genomes within the population providing insight into the evolution of individual species. Reconstruction of networks based on prokaryotic genomes revealed that closely related taxa are strongly interconnected and form a distinct phylogenetic community that promotes gene sharing [37]. Moreover, closely related species that are found in the same environment, such as *Salmonella* and *Citrobacter*, are more interconnected and genetically similar [38]. Genetic exchange between members of the community network, through recombination or HGT, can contribute to the acquisition of traits enabling survival in fluctuating environments and potentially giving rise to strains with atypical biochemical, phenotypic, and antigenic properties [20, 39]. The food environment presents a hostile landscape for microorganisms due to injury or stress caused by food preservation methods and intrinsic factors such as pH, temperature, and competition with other microflora [40]. These conditions induce adaptive responses for survival, such as stress tolerance and/or expanded metabolic capabilities [41]. Our results indicate that genetic exchange between *Salmonella* and *Citrobacter* could have contributed to their similarities as well as the diversification of their respective genera.

Conclusions

Bacterial species determination was initially based on phenotypic properties. However, phenotypic test results are often discrepant. Phenotypic characterization of the five *Citrobacter* strains that were misidentified as *Salmonella* provided variable results and were unable to unambiguously determine the species identity. Bacterial classification based on mass-spectrometry can complement standard phenotypic and biochemical approaches. However, the non-genomic nature of this method limits the phylogenetic investigation of closely related species or strains. This method is also limited by the comprehensiveness of the database used to compare spectra, which to date are heavily biased towards medically relevant pathogens [42]. Classification using 16S rRNA sequences and ANI provided limited taxonomic resolution and thus, less confidence in species identification [43]. The limited variability in the conserved region of the 16S rRNA genes in closely related enteric bacteria results in a decreased ability to resolve relationships below the family level [35]. Core and pan-genome analysis, however, were more informative in delineating *Salmonella* and *Citrobacter*. Both methods provided complementary but congruent results. The core genome analysis of *Citrobacter* highlighted the challenges associated with species designation, particularly in regards to *C. freundii*, which did not constitute a discrete phylogenetic group. Further taxonomic inquiry is needed to clarify the lineage and typing of its members. Our results showed that four of the five *Citrobacter* strains falsely identified as *Salmonella* were phylogenetically similar and the prevalence of these strains in fresh produce and other food products deserves further investigation to help improve the detection of *Salmonella*. The insights gained from the analyses of our study can be used to develop robust molecular assays capable of rapid identification and discrimination of *Salmonella* from *Citrobacter* or other closely related non-pathogens.

Methods

Bacterial strains, genome sequencing, and assembly

Table 1 provides a list of the 37 *Citrobacter* spp. and 58 *S. enterica* subsp. *enterica* strains included in the study. The 58 *S. enterica enterica* strains in this study were chosen to encompass the diversity of the subspecies and represent serovars frequently associated with outbreaks as well as serovars isolated from food and rarely associated with human disease (grey box) [2, 3]. Of the 38 serovars selected, 19 included representatives of serovars frequently associated with clinical salmonellosis in Canada. The five most commonly reported Canadian serovars were represented by multiple strains; Enteritidis ($n = 5$), Heidelberg ($n = 4$), Javiana ($n = 4$), Newport ($n = 5$), and Typhimurium ($n = 7$). The serovars Typhi, Paratyphi A, and Gallinarum are host-restricted while the remaining 16 serovars (Table 1, grey box) are rarely associated with human disease.

Six *Citrobacter* species (*C. amalonaticus*, $n = 2$; *C. braakii*, $n = 5$; *C. farmeri*, $n = 1$; *C. freundii*, $n = 6$, *C. koseri*, $n = 2$; and *C. werkmanii*, $n = 1$) were represented among the 37 *Citrobacter* strains sequenced in this study. Twenty-one strains did not have a species designation and five of these strains were isolated from fresh produce as part of a producer's food safety testing program and originally misidentified as *Salmonella* using a commercial immunoassay test (S646, S647, S648, S1284, S1285, Table 1, blue box).

All strains were sequenced at the Plateforme d'Analyses Génomiques of the Institute for Integrative and Systems Biology (IBIS), Université Laval, Quebec, Canada using the procedures described by the *Salmonella* Syst-OMICS consortium [44]. Raw sequence reads

were downloaded from the SalFos database and de novo genome assembly was performed using SKESA (v 2.3.0) [45]. QUAST was used to assess the quality of assembled genomes [46], while gene calling and annotation were done using the Prokka software v1.12 [47]. All nucleotide sequence data and additional information on the strains are available from the *Salmonella* Foodborne Syst-OMICS database (SalFoS) (https://salfos.ibis.ulaval.ca/).

Phenotypic analysis of false positive isolates

The five *Citrobacter* strains that were falsely identified as *Salmonella* were cultured on tryptic soy agar (BD Difco, NJ, USA) overnight at 35 °C. Isolates were grown in xylose lysine deoxycholate agar (XLD), brilliant green sulfa agar (BGS), triple sugar iron agar (TSI), and lysine iron agar (LIA) for 24 h at 35 °C to determine colony morphologies indicative of *S. enterica* or *Citrobacter*. Serological reactions were assessed using *Salmonella* O antiserum poly A-I + vi (BD Difco, NJ, USA) and the OXOID *Salmonella* Latex Test (Thermo Fisher, Hampshire, UK) following the manufacturer's protocols. *S. enterica* ser. Enteritidis S187 and *C. freundii* ATCC8090 was used as a positive control. Biochemical characterization was carried out using API20E (bioMérieux, Marcy l'Étoile, France) and the VITEK 2 automated microbial identification system (version 07.01, bioMérieux, Marcy l'Étoile, France) using the GN identification card and following the manufacturer's instructions.

The strains were further characterized using a Biotyper matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometer (Bruker Daltonik, Bremen, Germany) following a standard extraction procedure. Briefly, fresh colonies were applied onto a Biotarget plate in duplicate, allowed to dry, and overlaid with freshly prepared α-cyano-4-hydroxy-cinnamic acid (HCCA) matrix. Mass spectra were acquired and the strains were classified using the MALDI Biotyper 3.1 software and Biotyper taxonomy database version 6 (6903 entries, BDAL, Bruker Daltonik). *S. enterica* ser. Enteritidis S187, *C. freundii* ATCC 8090 and *E. coli* ATCC 25922 were used as controls. The software compares the peak profile of a sample's mass spectrum with a reference spectrum in the database and calculates an arbitrary unit score value from 0 to 3 according to a proprietary algorithm. High scores denote similarity between the sample and reference spectra. Classification criteria indicate genus and probable species-level identification for scores of 2.000 to 2.299 and highly probable species identification for values of 2.300 to 3.000.

Comparative genomic analysis

Annotated *Citrobacter* and *Salmonella* protein assemblies were clustered and filtered for redundancy using MMseqs2 [48]. An in-house script called centreseq was used to parse the data [27]. After removal of duplicate genes, the protein sequences were concatenated and re-clustered using the MMseqs2 linclust algorithm. Cut-off values for percent amino acid identity and alignment length were both set at 90%. The complete set of gene sequences identified in all of the genomes was defined as the pangenome. The core genome represented sequences that were present in all of the analyzed genomes and the accessory genome was the set of sequences that were present in a sub-set of at least two genomes. Gene sequences that were only present in one genome were termed singletons and gene sequences that were exclusively found in either *Salmonella* or *Citrobacter* with no orthologue in the other genus were defined as genus-specific unique gene sequences.

A rarefaction analysis was performed to estimate the size of the core genome of *Salmonella* and *Citrobacter*. The analysis was achieved by randomly selecting subsets (ranging 2–95 genomes) of the input samples and calculating the core- and pan- counts. The core- and pan-contents of each subset were calculated five times and the results averaged to create one data point. Curve fitting was done using nonlinear regression (one phase decay) in Graphpad Prism version 5.0b for Mac OS X (Graphpad Software, San Diego, California, USA). The point of the curve at which additional number of genes incorporated into the core genome relative to the number of analyzed genomes begins to plateau indicates the value needed for complete core genome representation for all included genomes.

Phylogenetic trees based on variations in the amino acid sequences encoded by the core genome were constructed using the tree module of centreseq [27]. Essentially, the core genome amino acid sequences were extracted and aligned using MUSCLE [49]. The alignment file was concatenated and used to construct a maximum likelihood phylogenetic tree using RAxML v8.2.9 [50] with 1000 bootstrap iterations. The numbers on the internal nodes of the tree represent the proportion of individual trees that are congruent with the consensus tree and indicate support for the separation of the taxa at that particular node.

Citrobacter sequences used to construct the *Citrobacter* specific phylogenetic tree were downloaded from RefSeq in June 2019 (*n* = 326) [51]. Genome distance estimation was performed using MASH [52]. Strains with low distance scores and in which *Citrobacter* was not the best hit were removed from the dataset (*n* = 18). Of the remaining RefSeq genomes, 160 sequences were chosen for further analysis to represent the 13 *Citrobacter* species depicted on the tree (Table S6).

Functional activities were assigned to the predicted amino acid sequences of the *Salmonella-Citrobacter* core

Pilar *et al. BMC Genomics*   (2020) 21:377

Page 12 of 13

and unique genes by mapping predicted activities to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database and BRITE hierarchies using the KEGG Automated Assignment Server (KAAS) at https://www.genome.jp/kegg/kaas/ [53].

The percentage of shared genes among genomes of the 96 strains were visualized using the pairwise report and network chart features of centreseq [27]. Pairwise information on the core cluster counts were visualized by linking pairs of genomes that share genes at or below a given percentage threshold.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-06780-y.

Additional file 1: Table S1. Phenotypic characterization of *Citrobacter* and *Salmonella* strains.

Additional file 2: Table S2. List of reference genomes used in this study.

Additional file 3: Table S3. Average nucleotide identity (ANI) values between Citrobacter strains falsely identified as *Salmonella* and reference genomes.

Additional file 4: Table S4. Number of open reading frames (ORFS) and coding sequences (CDS) in the genomes of SalFoS strains.

Additional file 5: Table S5. Genes present in the pan-genome of 95 *Salmonella* and *Citrobacter* strains

Additional file 6: Table S6. List of *Citrobacter* genomes obtained from RefSeq database, June 2019.

Additional file 7: Table S7. KEGG functional classifications of *Citrobacter* specific unique genes.

Additional file 8. Supplementary information.

Abbreviations
ABC: Adenosine triphosphate binding cassette; ANI: Average nucleotide identity; API: Analytical profile index; ATCC: American type culture collection; HGT: Horizontal gene transfer; KEGG: Kyoto encyclopedia of genes and genomes; MALDI-(TOF): Matrix assisted laser desorption/ionization-(time of flight); (wg)MLST: (Whole genome) Multilocus sequence typing; QUAST: Quality assessment tool for genome assemblies; rRNA: Ribosomal ribonucleic acid; SalFoS: Salmonella foodborne Syst-OMICS database; SKESA: Strategic k-mer extension for scrupulous assemblies; SNP: Single nucleotide polymorphism

Authors' contributions
AVCP, NP, FD, AV, and ST designed the study, analyzed the data and wrote the manuscript. AVCP and NP performed the experiments. RCL, SB, and LG provided strains and sequence data. All authors reviewed and revised the manuscript. The author(s) read and approved the final manuscript.

Authors' information
Not applicable.

Availability of data and materials
All data supporting the findings of this study are available as tables or figures within the article and supplementary information. Genome sequences generated in this study are available from the *Salmonella* Foodborne Syst-OMICS database at https://salfos.ibis.ulaval.ca/. Some restrictions apply to the availability of these data, which were used under an agreement for the current study, and so are not publically available. Data are however available from the authors upon reasonable request and with permission from the *Salmonella* Syst-OMICS consortium.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors have no competing interests.

Author details
[1]Bureau of Microbial Hazards, Health Canada, Ottawa, Ontario, Canada. [2]Bureau of Food Surveillance and Science Integration, Health Canada, Ottawa, Ontario, Canada. [3]Laboratoire de santé publique du Québec, Ste-Anne-de-Bellevue, Quebec, Canada. [4]Institute for Integrative and Systems Biology (IBIS), Université Laval, Quebec, Quebec, Canada. [5]Department of Food Science and Agricultural Chemistry, Faculty of Agricultural and Environmental Sciences, McGill University, Ste-Anne-de-Bellevue, Quebec, Canada. [6]Food Science Department, University of Guelph, Guelph, Ontario, Canada.

References
1. Kirk MD, Pires SM, Black RE, Caipo M, Crump JA, Devleesschauwer B, et al. World Health Organization estimates of the global and regional disease burden of 22 foodborne bacterial, Protozoal, and viral diseases, 2010: a data synthesis. PLoS Med. 2015;12(12):e1001921.
2. Jackson BR, Griffin PM, Cole D, Walsh KA, Chai SJ. Outbreak-associated *Salmonella* serotypes and food commodities, United States, 1998-2008. Emerg Infect Dis. 2013;19(8):1239–44.
3. Reddy SP, Wang H, Adams JK, Feng PC. Prevalence and characteristics of *Salmonella* serotypes isolated from fresh produce marketed in the United States. J Food Prot. 2016;79(1):6–16.
4. Kozak GK, MacDonald D, Landry L, Farber JM. Foodborne outbreaks in Canada linked to produce: 2001 through 2009. J Food Prot. 2013;76(1):173–83.
5. Crosa JH, Brenner DJ, Ewing WH, Falkow S. Molecular relationships among the salmonelleae. J Bacteriol. 1973;115(1):307–15.
6. Kauffmann F, Edwards PR. A revised, simplified Kauffmann-white schema. Acta Pathol Microbiol Scand. 1957;41(3):242–6.
7. Gal-Mor O, Boyle EC, Grassl GA. Same species, different diseases: how and why typhoidal and non-typhoidal *Salmonella enterica* serovars differ. Front Microbiol. 2014;5:391.
8. Huehn S, La Ragione RM, Anjum M, Saunders M, Woodward MJ, Bunge C, et al. Virulotyping and antimicrobial resistance typing of *Salmonella enterica* serovars relevant to human health in Europe. Foodborne Pathog Dis. 2010; 7(5):523–35.
9. Shah DH, Paul NC, Sischo WC, Crespo R, Guard J. Population dynamics and antimicrobial resistance of the most prevalent poultry-associated *Salmonella* serotypes. Poult Sci. 2017;96(3):687–702.
10. Timme RE, Pettengill JB, Allard MW, Strain E, Barrangou R, Wehnes C, et al. Phylogenetic diversity of the enteric pathogen *Salmonella enterica* subsp. *enterica* inferred from genome-wide reference-free SNP characters. Genome Biol Evol. 2013;5(11):2109–23.
11. van der Walt ML, Steyn HC. The biochemical differentiation between *Salmonella* and *Citrobacter*. Onderstepoort J Vet Res. 1989;56(4):263–9.

Pilar *et al. BMC Genomics*   (2020) 21:377

Page 13 of 13

12. Gaillot O, di Camillo P, Berche P, Courcol R, Savage C. Comparison of CHROMagar *Salmonella* medium and hektoen enteric agar for isolation of salmonellae from stool samples. J Clin Microbiol. 1999;37(3):762–5.

13. Bell RL, Jarvis KG, Ottesen AR, McFarland MA, Brown EW. Recent and emerging innovations in *Salmonella* detection: a food and environmental perspective. Microb Biotechnol. 2016;9(3):279–92.

14. Ewing WH, Ball MM, Bartes SF, McWhorter AC. The biochemical reactions of certain species and bioserotypes of *Salmonella*. J Infect Dis. 1970;121(3):288–94.

15. Brenner DJ, Grimont PA, Steigerwalt AG, Fanning GR, Ageron E, Riddle CF. Classification of citrobacteria by DNA hybridization: designation of *Citrobacter farmeri* sp. nov., *Citrobacter youngae* sp. nov., *Citrobacter braakii* sp. nov., *Citrobacter werkmanii* sp. nov., *Citrobacter sedlakii* sp. nov., and three unnamed *Citrobacter* genomospecies. Int J Syst Bacteriol. 1993;43(4):645–58.

16. Warren JR, Farmer JJ 3rd, Dewhirst FE, Birkhead K, Zembower T, Peterson LR, et al. Outbreak of nosocomial infections due to extended-spectrum beta-lactamase-producing strains of enteric group 137, a new member of the family *Enterobacteriaceae* closely related to *Citrobacter farmeri* and *Citrobacter amalonaticus*. J Clin Microbiol. 2000;38(11):3946–52.

17. Borenshtein D, McBee ME, Schauer DB. Utility of the *Citrobacter rodentium* infection model in laboratory mice. Curr Opin Gastroenterol. 2008;24(1):32–7.

18. Werkman CH, Gillen GF. Bacteria producing Trimethylene glycol. J Bacteriol. 1932;23(2):167–82.

19. Edwards PR, Ewing WH, Edwards PR. Edwards and Ewing's identification of Enterobacteriaceae. 4th ed. New York: Elsevier; 1986. p. 536.

20. Delgado G, Souza V, Morales R, Cerritos R, Gonzalez-Gonzalez A, Mendez JL, et al. Genetic characterization of atypical *Citrobacter freundii*. PLoS One. 2013;8(9):e74120.

21. Retchless AC, Lawrence JG. Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. Proc Natl Acad Sci U S A. 2010; 107(25):11453–8.

22. Tang L, Liu SL. The 3Cs provide a novel concept of bacterial species: messages from the genome as illustrated by *Salmonella*. Antonie Van Leeuwenhoek. 2012;101(1):67–72.

23. Zahrt TC, Mora GC, Maloy S. Inactivation of mismatch repair overcomes the barrier to transduction between *Salmonella typhimurium* and *Salmonella typhi*. J Bacteriol. 1994;176(5):1527–9.

24. Matthews TD, Schmieder R, Silva GG, Busch J, Cassman N, Dutilh BE, et al. Genomic comparison of the closely-related *Salmonella enterica* Serovars Enteritidis, Dublin and Gallinarum. PLoS One. 2015;10(6):e0126883.

25. Porwollik S, Santiviago CA, Cheng P, Florea L, Jackson S, McClelland M. Differences in gene content between *Salmonella enterica* serovar Enteritidis isolates and comparison to closely related serovars Gallinarum and Dublin. J Bacteriol. 2005;187(18):6545–55.

26. Grimont PAD, Weill FX. Antigenic formulae of the *Salmonella* serovars. In: WHO Collaborating Centre for Reference and Research on Salmonella. 9th ed. Paris: Institut Pasteur; 2007.

27. Dussault FM. Centreseq: fast prokaryote core-genome analysis. 2019. https://github.com/BFSSI-Bioinformatics-Lab/centreseq.

28. Lan R, Reeves PR. Gene transfer is a major factor in bacterial evolution. Mol Biol Evol. 1996;13(1):47–55.

29. Riley MA, Lizotte-Waniewski M. Population genomics and the bacterial species concept. Methods Mol Biol. 2009;532:367–77.

30. Chung M, Munro JB, Tettelin H, Dunning Hotopp JC. Using core genome alignments to assign bacterial species. mSystems. 2018;3(6):e00236–18.

31. de Been M, Pinholt M, Top J, Bletz S, Mellmann A, van Schaik W, et al. Core genome multilocus sequence typing scheme for high- resolution typing of *Enterococcus faecium*. J Clin Microbiol. 2015;53(12):3788–97.

32. Kingry LC, Rowe LA, Respicio-Kingry LB, Beard CB, Schriefer ME, Petersen JM. Whole genome multilocus sequence typing as an epidemiologic tool for *Yersinia pestis*. Diagn Microbiol Infect Dis. 2016;84(4):275–80.

33. Kluytmans-van den Bergh MF, Rossen JW, Bruijning-Verhagen PC, Bonten MJ, Friedrich AW, Vandenbroucke-Grauls CM, et al. Whole-genome multilocus sequence typing of extended-spectrum-beta-lactamase-producing *Enterobacteriaceae*. J Clin Microbiol. 2016;54(12):2919–27.

34. Laing CR, Whiteside MD, Gannon VPJ. Pan-genome analyses of the species *Salmonella enterica*, and identification of genomic markers predictive for species, subspecies, and Serovar. Front Microbiol. 2017;8:1345.

35. Sela I, Wolf YI, Koonin EV. Theory of prokaryotic genome evolution. Proc Natl Acad Sci U S A. 2016;113(41):11399–407.

36. Dagan T. Phylogenomic networks. Trends Microbiol. 2011;19(10):483–91.

37. Boto L, Martinez JL. Ecological and temporal constraints in the evolution of bacterial genomes. Genes (Basel). 2011;2(4):804–28.

38. Hooper SD, Mavromatis K, Kyrpides NC. Microbial co-habitation and lateral gene transfer: what transposases can tell us. Genome Biol. 2009;10(4):R45.

39. Chaffron S, Rehrauer H, Pernthaler J, von Mering C. A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Res. 2010;20(7):947–59.

40. Wesche AM, Gurtler JB, Marks BP, Ryser ET. Stress, sublethal injury, resuscitation, and virulence of bacterial foodborne pathogens. J Food Prot. 2009;72(5):1121–38.

41. McMahon MA, Xu J, Moore JE, Blair IS, McDowell DA. Environmental stress and antibiotic resistance in food-related pathogens. Appl Environ Microbiol. 2007;73(1):211–7.

42. Wilson DA, Young S, Timm K, Novak-Weekley S, Marlowe EM, Madisen N, et al. Multicenter evaluation of the Bruker MALDI Biotyper CA system for the identification of clinically important Bacteria and yeasts. Am J Clin Pathol. 2017;147(6):623–31.

43. Naum M, Brown EW, Mason-Gamer RJ. Is 16S rDNA a reliable phylogenetic marker to characterize relationships below the family level in the *Enterobacteriaceae*? J Mol Evol. 2008;66(6):630–42.

44. Emond-Rheault JG, Jeukens J, Freschi L, Kukavica-Ibrulj I, Boyle B, Dupont MJ, et al. A Syst-OMICS approach to ensuring food safety and reducing the economic burden of salmonellosis. Front Microbiol. 2017;8:996.

45. Souvorov A, Agarwala R, Lipman DJ. SKESA: strategic k-mer extension for scrupulous assemblies. Genome Biol. 2018;19(1):153.

46. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–5.

47. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068–9.

48. Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35(11): 1026–8.

49. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.

50. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.

51. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–45.

52. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17(1):132.

53. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007;35(Web Server issue):W182–5.

Publisher's Note