Infection, Genetics and Evolution 106 (2022) 105380

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Phylogenetic group and virulence profile classification in *Escherichia coli* from distinct isolation sources in Mexico

José R. Aguirre-Sánchez [a], José B. Valdez-Torres [a], Nohemí Castro del Campo [b], Jaime Martínez-Urtaza [c], Nohelia Castro del Campo [a], Bertram G. Lee [d], Beatriz Quiñones [d], Cristóbal Chaidez-Quiroz [a,*]

[a] *Centro de Investigación en Alimentación y Desarrollo, Coordinación Regional Culiacán, Laboratorio Nacional para la Investigación en Inocuidad Alimentaria, 80110 Culiacán, Sinaloa, Mexico*
[b] *Facultad de Medicina Veterinaria y Zootecnia de la Universidad Autónoma de Sinaloa, 80260 Culiacán, Sinaloa, Mexico*
[c] *Department of Genetics and Microbiology, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain*
[d] *U.S. Department of Agriculture-Agricultural Research Service, Western Regional Research Center, Produce Safety and Microbiology Research Unit, Albany, CA 94710, United States*

ARTICLE INFO

Keywords:
E. coli
Phylogroups
Correspondence analysis
Virulence
Mexico
Genomics

ABSTRACT

Escherichia coli is a leading cause of human enteric diseases worldwide. The rapid and accurate causal agent identification to a particular source represents a crucial step in the establishment of safety and health measures in the affected human populations and would thus provide insights into the relationship of traits that may contribute for pathogen persistence in a particular reservoir. The objective of the present study was to characterize over two hundred *E. coli* strains from different isolation sources in Mexico by conducting a correspondence analysis to explore associations with the detected phylogenetic groups. The results indicated that *E. coli* strains, recovered from distinct sources in Mexico, were classified into phylogroups B1 (35.8%), A (27.8%), and D (12.3%) and were clustered to particular clades according to the predicted phylogroups. The results from correspondence analysis showed that *E. coli* populations from distinct sources in Mexico, belonging to different phylogroups, were not dispersed randomly and were associated with a particular isolation source. Phylogroup A was strongly associated with human sources, and the phylogroup B1 showed a significant relationship with food sources. Additionally, phylogroup D was also related to human sources. Phylogroup B2 was associated with herbivorous and omnivorous mammals. Moreover, common virulence genes in the examined *E. coli* strains, assigned to all phylogroups, were identified as essential markers for survival and invasion in the host. Although virulence profiles varied among the detected phylogroups, *E. coli* strains belonging to phylogroup D, associated with humans, were found to contain the largest virulence gene repertoire conferring for persistence and survival in the host. In summary, these findings provide fundamental information for a better characterization of pathogenic *E. coli*, recovered from distinct isolation sources in Mexico and would assist in the development of better tools for identifying potential transmission routes of contamination.

1. Introduction

Escherichia coli is a Gram-negative bacterium that resides and spreads among the gastrointestinal tract of warm-blooded animals (Gordon and Cowling, 2003). This bacterium sometimes contaminates surface water and may serve to trace fecal contamination and the presence of other waterborne pathogens, such as *Salmonella* and other enteric organisms (Odonkor and Ampofo, 2013). The identification of waterborne pathogens, using *E. coli* strains as indicator, derive from culture-based and molecular-based methods. However, several strains of *E. coli* are pathogenic to humans due to the presence of virulence factors, leading to a broad range of enteric human diseases such as diarrhea, colitis, dysentery, and hemolytic uremic syndrome. Other disease symptoms in humans also include extraintestinal diseases, such as sepsis and urinary tract infections (Kaper et al., 2004). The differentiation of intestinal pathogenic *E. coli* strains, based on virulence factors and the

* Corresponding author at: CIAD, Culiacán, Carretera El Dorado Km 5.5 Campo el Diez, 80110 Culiacán, Sinaloa, Mexico.
E-mail address: chaqui@ciad.mx (C. Chaidez-Quiroz).

https://doi.org/10.1016/j.meegid.2022.105380
Received 26 August 2022; Received in revised form 19 October 2022; Accepted 21 October 2022
Available online 22 October 2022

identification of disease mechanisms, has resulted in a classification into six pathotypes: enteropathogenic *E. coli* (EPEC), enterohaemorrhagic *E. coli* (EHEC), enterotoxigenic *E. coli* (ETEC), enteroinvasive *E. coli* (EIEC), enteroaggregative *E. coli* (EAEC) and diffusely-adherent *E. coli* (DAEC) (Kaper et al., 2004).

Techniques for differentiating *E. coli* have relied on the use of serological methods for the identification of surface antigens (Kauffmann, 1947) or on the use of sequence-based typing methods for identifying housekeeping genes by multilocus sequence typing (MLST) (Selander et al., 1986) or virulence gene profiles by PCR or whole genome sequencing (Gordon et al., 2008). To better address the epidemiological importance of pathogenic *E. coli,* a classification of *E. coli* strains into phylogroups has been previously developed based on the presence of specific gene targets (Clermont et al., 2013). As new sequence-based technologies emerge, phylogroup classification has been expanded by the additional identification of subdivisions within a phylogroup (Abram et al., 2021). A strategy for phylogroup classification is based on the quadruplex PCR assay, which is characterized by the addition of the *arpA* gene and two specific loci allowing the identification of seven phylogroups designated as A, B1, B2 C, D, E, and F (Clermont et al., 2013). The use of phylogroup classification has been employed in the study of ecological niches and lifestyles in bacterial pathogens and improves our understanding of population structure providing invaluable epidemiological information (Coura et al., 2015). Moreover, the implementation of sequence-based methods combined with the accumulation of *E. coli* sequenced genomes on public databases have allowed to study *E. coli* population structure and to investigate how virulence factors are maintained, acquired, or even shared with other bacterial species.

The classification of *E. coli* strains based on phylogroup designations has identified a relationship between different niches and lifestyles, indicating that *E. coli* strains are not randomly dispersed but rather exhibit both host taxonomic and environmental components (Souza et al., 1999). In an analysis of 152 *E. coli* strains, a total of 43.4% of those classified as phylogroup A were found to be human commensals (Escobar-Páramo et al., 2006). By contrast, strains isolated from animals fall mostly into phylogroup B1 (Higgins et al., 2007), suggesting an association between phylogenetic groups and host species (de Stoppe et al., 2017). In Mexico, a country with high enteric disease rate (Corzo-Ariyama et al., 2019), there is limited information on the population structure and characteristics of naturally-occurring *E. coli* strains. To develop better epidemiological tools using whole genome sequencing data, the present study analyzed the distribution and prevalence of a large number of *E. coli* strains with distinct phylogenetic groups and virulence profiles and from various isolation sources in Mexico, including clinical, animal and food. The findings from this study would thus provide fundamental information on the relationship of the genomic profiles of *E. coli* strains with their isolation source and disease outcome in humans.

2. Materials and methods

2.1. DNA extraction and genome sequencing

Four *E. coli* strains, previously isolated from milk-producing cows suffering from mastitis, were included in this study to expand the collection of strains for assessing the genotypic diversity of strains with importance to the food industry, and these strains from dairy cows were obtained from the collection of the National Food Safety Laboratory (LANIIA) at the Centro de Investigación en Alimentación y Desarrollo (CIAD), Culiacan, Mexico. DNA extraction was performed using the DNeasy Blood & Tissue Kit (QIAGEN, Mexico City, Mexico) according to the manufacturer's instructions. The concentration of the extracted DNA was determined using the Qubit dsDNA Broad Range Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA). For performing whole-genome sequencing, the genomic DNA from each *E. coli* strain was adjusted to a 0.2 ng/μL concentration, and aliquoted at a final amount of 1 ng for preparing genomic DNA libraries with the Nextera XT DNA Library Preparation Kit (Illumina Inc., San Diego, CA, USA). Subsequently, the prepared genomic libraries were sequenced using a MiSeq™ Reagent Kit v2 (300-cycle format) to obtain a 2 × 150 bp paired-end read output with a MiSeq™ System (Illumina, Inc.) at the Earlham Institute (Norwich Research Park, Norwich, United Kingdom). Additionally, whole genome sequences from a total of 208 *E. coli* strains, recovered from various clinical and environmental sources in Mexico were downloaded in FASTA format from publicly available assembly databases with the National Center for Biotechnology Information (NCBI) (Supplementary Table 1).

2.2. Quality control and assembly

Read quality control was performed using the command-line tool Cutadapt version 2.6 and the script wrapper Trim Galore version 0.5.0 (Krueger, 2015) with a minimum quality value of 30 for trimming the first 20 bp from the 5′ end of raw sequence reads. Duplicate reads were removed using Clumpify version 37.62 (Bushnell et al., 2017). The trimmed sequence reads were visualized and subjected to a quality check using FastQC version 0.11.8 (Wingett and Andrews, 2018) and were assembled *de novo* with the pipeline A5-miseq version 20,160,825 (Coil et al., 2015). The sequence assemblies with >180 contigs were oriented with ABACAS (Assefa et al., 2009) using the complete genome sequence of Shiga toxin-producing *E. coli* O111:H2 strain RM9322 (Accession number GCF_003112245.1) as the reference.

2.3. Phylogroup and ST determination

ClermonTyping based on the *in vitro* assay (Beghain et al., 2018) was used to determine *E. coli* phylogroups for all 212 genomes. Sequence type (ST) determination was performed by submitting genome sequences to the *E. coli* PubMLST database using the MLST scheme based on the sequence of the housekeeping genes *adk, fumC, gyrB, icd, mdh, purA*, and *recA* (Jolley et al., 2018).

2.4. Phylogenetic tree construction

A core genome alignment for 212 *E. coli* genomes was performed by the tool Parsnp v 1.2 using the closed genome of the *E. coli* strain 118UI (GCA_003627855.1) as reference. The resulted alignment was converted to multi-FASTA file using HarvestTools v 1.1.2 (Treangen et al., 2014) with the option of -M for generating a multi-fasta alignment option. A maximum likelihood approach with the GTRGAMMA model and 100 bootstraps were used to construct a phylogenic tree using the bioinformatics tool RAxML (Stamatakis, 2014). The resulted tree was displayed and edited with iTOL (Letunic and Bork, 2019).

2.5. Virulence genes identification

ABRicate v 1.0.1 (https://github.com/tseemann/abricate) was used in the present study for mass screening of contigs to identify virulence genes in the examined *E. coli* genomes with the Virulence Factor Database (VFDB) (Liu et al., 2019). Virulence genes were considered present based on a coverage of >90% and an identity >95%. The presence or absence of the virulence genes in the examined *E. coli* genomes was then reported as a matrix, and the iTOL Annotation Editor (Letunic and Bork, 2019) was used to construct the virulence gene matrix.

2.6. Statistical analysis

The metadata was employed to categorize a total of 212 *E. coli* strains, recovered from distinct sources and locations in Mexico. The diversity of the phylogroups from each isolation source was measured using the Simpson diversity index where a high diversity in the dataset is

indicated by index values close to 1 (Hunter and Gaston, 1988; Simpson, 1949). The Shannon diversity and evenness indexes were calculated using the pgirmess package version 2.0.0 (Giraudoux et al., 2018) with the default base of 2 with the R software version 4.2.0 (Team, 2022). To assess the phylogroup similarity between the isolation sources, the Pianka's index was calculated using the pgirmess package in the R software (Giraudoux et al., 2018). Furthermore, a correspondence analysis (Greenacre, 2007) was performed to determine the relative abundance of phylogroups distribution by isolation source and was constructed based on input from the contingency table, listing a total of three sources and seven phylogroups. Total inertia and mass were calculated to assess associations between categories and contributions to variation in the data. Relations among categories was presented in a symmetric biplot. The correspondence analysis was conducted using Minitab 18 (Arend, 1993).

3. Results

A total of 212 *E. coli* strains isolated from clinical, animal, and food sources in Mexico were classified according to the phylogenetic groups A, B1, B2, C, D, E, and F and MLST analysis (Supplementary Table S1). The results by MLST demonstrated a total 123 different STs, suggesting genetic diversity among the examined *E. coli* isolates. In particular, ST10 was the predominant type among 8% of the strains (Table S1), mostly recovered from clinical fecal samples. The MLST analysis revealed other strains also from clinical fecal samples were belonging to ST69 and ST131, types previously identified in pandemic lineages of extra-intestinal pathogenic *E. coli* (Riley, 2014). Moreover, the observed frequencies of the sources and phylogroup classification were further determined (Table 1), and the diversity indexes of the phylogroups for each source were calculated (Table 2). The results of the Simpson and Shannon indexes calculations showed that both human and animal sources were found to have significantly high diversities in the phylogenetic group distribution (Table 2). Subsequent calculation of the Shannon evenness index also demonstrated a more equal distribution of phylogroups for *E. coli* strains recovered from human and animal sources. By contrast, the distribution of *E. coli* strains recovered from food sources was found to have significantly lower diversity and evenness indexes (Table 2), and an explanation for these observations is based on the fact that strains from food sources were assigned to fewer phylogroups since no isolates were found to belong to phylogroups C, E and F, as shown in Table 1. To further analyze the relatedness in the distribution of the isolates in the phylogroups, the similarity index was calculated by comparing each pair of sources (Table 3), and the findings from these calculations indicated that the comparison in the phylogroup distribution between animal and food sources had the highest similarity index with a value of 0.91 (Table 3), based on the fact that both animal and food sources had the largest number of isolates assigned to phylogroup B1 (Table 1).

A correspondence analysis was conducted. Row profiles, row masses, and the average row profile, reported in Table 1, were employed for the subsequent analysis shown in Table 4. For row profiles, the phylogroup frequency by isolation source indicated a significant frequency of 0.645 for *E. coli* strains from food sources to be classified in phylogroup B1. Lower frequencies of 0.435 for phylogroup A and 0.406 for phylogroup B1 were observed for human and animal sources, respectively. In the analysis of row masses, the proportion of the isolation source in the entire dataset was examined. In particular, the animal source frequency was found to be 0.453 (96/212), while the frequencies of 0.401 (85/212) and 0.146 (31/212) corresponded to human and food sources, respectively. Moreover, examination of column masses described the proportion of phylogenetic groups in the entire dataset. In particular, the highest column mass was observed for phylogroup B1 with a frequency of 0.358 (76/212), followed by phylogroup A at a frequency of 0.278 (59/212). Lower frequencies in the column mass were observed for the remaining phylogroups B2, C, D, E, and F (Table 1).

As shown in Table 4, cell residuals indicated the difference between the observed metadata and the expected data, and the analysis of cell residuals measured either the overrepresentation (positive value) or the underrepresentation (negative value) of a phylogroup for a given source. In particular, the results in the present study indicated that phylogroup A was overrepresented in human sources, suggesting a strong relationship among the two categories, while a weak relationship (underrepresentation) was observed for the animal sources. By contrast, phylogroup B1 was underrepresented in human sources while overrepresented in both animal and food sources, suggesting a weak and strong relationship for these categories, correspondingly. Additionally, the variance, defined by cell inertia, was determined (Table 4), and the results showed that only phylogroups A and B1 had high inertia values of 0.33, indicating a significant variability in the association of these two phylogroups with a particular source.

The principal axes (components), inertias, their proportion, and cumulative values are shown in Supplementary Table 2. This analysis examined how many components were sufficient to influence the variation and relationship of *E. coli* strains from the various sources. The results indicated that component 1 corresponded to 63.96% of the original variation in the contingency table while component 2 corresponded to 36.04%. The findings indicated that both components account for all of the total variation in the table, resulting in a dimensional reduction for the observed representation of the variation, as shown in Table 1.

A symmetric biplot with two components was constructed to visualize the association among isolation sources and phylogroups (Fig. 1).

Table 2
Simpson and Shannon diversity indexes for each isolation source.

Source	Diversity		
	Simpson	Shannon diversity	Shannon evenness
Human	0.74	2.23	0.79
Animal	0.77	2.36	0.84
Food	0.25	1.33	0.47
Total	0.76	2.30	0.79

Table 3
Pianka's similarity index for each pair of isolation sources.

	Pianka's simlarity index		
	Human	Animal	Food
Human	–	0.74	0.70
Animal	–	–	0.91

Table 1
Contingency table with the distribution and frequencies of phylogenetic groups among *E. coli* strains from three different sources.

Isolation source	Phylogroups							Total (row mass)
	A	B1	B2	C	D	E	F	
Human	37 (0.435)	17 (80.200)	9 (0.106)	1 (0.012)	13 (0.153)	5 (0.059)	3 (0.035)	85 (0.401)
Animal	14 (0.146)	39 (0.406)	15 (0.156)	3 (0.031)	11 (0.115)	12 (0.125)	2 (0.021)	96 (0.453)
Food	8 (0.258)	20 (0.645)	1 (0.032)	0 (0.000)	2 (0.065)	0 (0.000)	0 (0.000)	31 (0.146)
Total (Column mass)	59 (0.278)	76 (0.358)	25 (0.118)	4 (0.019)	26 (0.123)	17 (0.080)	5 (0.024)	212

Table 4
Cell residuals and cell relative inertias among isolation sources and phylogroups.

Isolation source	Phylogroups							Row inertia
	A	B1	B2	C	D	E	F	
Human[1]	**13.344**	**−13.472**	−1.024	−0.604	2.575	−1.816	0.995	0.383
	0.187	**0.148**	0.003	0.006	0.016	0.012	0.012	
Animal	**−12.717**	4.585	3.679	1.189	−0.774	4.302	−0.264	0.276
	0.150	0.015	0.030	0.019	0.001	0.060	0.001	
Food	−0.627	**8.887**	−2.656	−0.585	−1.802	−2.486	−0.731	0.341
	0.001	**0.176**	0.048	0.015	0.021	0.062	0.018	
Column inertia	0.338	0.339	0.080	0.040	0.038	0.133	0.031	**0.1901**

[1] For each phylogroup per isolation source, the top cell corresponds to the residual and the bottom cell corresponds to the inertia.

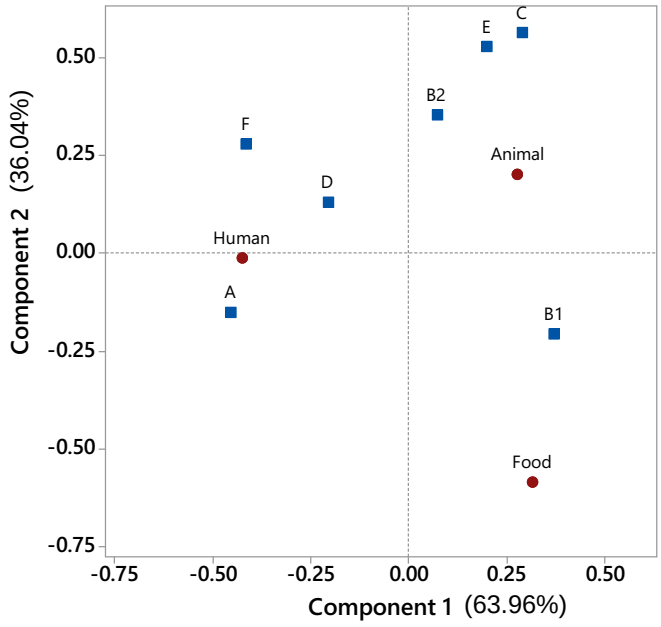

Fig. 1. Symmetric biplot from correspondence analysis based on rows (sources) and columns (phylogroups) from Table 1. The two-component describes a 100% of the total variation, resulting in 63.96% to be represented by the 1st component and 36.04% to be represented by the 2nd component.

Based on the data shown in Fig. 1 and Table 4, the data showed that the human source exhibited a significant positive association with phylogroup A and strong negative association with phylogroup B1. The animal source was negatively associated with phylogroup A while the food source was positively associated with phylogroup B1. These results were in agreement with the reported row masses of human and animal sources and the column mases of phylogroup A and B1, as displayed in Table 1.

To examine the genetic relatedness of the *E. coli* strains from Mexico and to identify specific virulence genes associated with the phylogroups, a maximum likelihood phylogenetic tree was constructed with a heatmap, showing the presence or absence of the examined virulence genes (Fig. 2A and B). The phylogenetic analysis revealed that specific clades for each of the phylogroups were detected (Fig. 2A and B). In more detail, the genomic content of the examined *E. coli* strains belonging to phylogroup B1 were found in the same clade, and the same finding was observed for the other phylogroups. The results also indicated that there was no association between a specific isolation source with a particular clade. High bootstraps values (>90) were observed in >90% of the phylogenetic tree branches (Fig. 2A and B).

As indicated by the presence/absence heatmap (Fig. 2A and B), unique virulence gene profiles were detected for certain phylogroups while other profiles were common among all phylogroups. Virulence operons coding for siderophore enterobactin (*entABCDEFS*), ferric transport (*fepABCDG*), type 1 fimbriae for adherence (*fimABCDEFGHI*), curli fiber (*csgBDFG*), *E. coli* common pilus (*ecpABCDE*), and the major outer protein A (*ompA*) in *E. coli* that can serve as a virulence factor for eukaryotic cell infection. The *chu* operon, encoding the hemin uptake system, was identified in the genomes for most strains belonging to phylogroups B2 and F. For phylogroup E, the *chu* operon genes, *chuTX,* were mostly absent while the *chuUVW* genes were present. By contrast, phylogroups A, B1, and C were found to lack the entire *chu* operon. As another virulence profile identified, the *shuATXSY* genes, encoding a membrane receptor, were present in most strains in phylogroups D and E. Interestingly, the analysis of virulence gene profiles indicated that the Shiga toxin (*stx*) genes were often present in *E. coli* strains belonging to phylogroup E (Fig. 2A and B). Moreover, the invasion gene *aslA* was common in all phylogroups except in the vast majority of strains in the phylogroup B1. Finally, all strains in phylogroup B2 lacked the *espY1*, *espY2*, *espY3*, and *espY4* genes, encoding effectors *via* the Type III Secretion System (T3SS), which were predominantly identified among strains belonging to phylogroups D and E, and also lacked the gene *espX2*, as observed in all phylogroups. In summary, the findings from these analyses have identified similarities and differences in the virulome content among *E. coli* strains belonging to the different phylogroups to enable an improved characterization of *E. coli* from food, animal and clinical isolation sources in Mexico.

4. Discussion

This study determined the occurrence of *E. coli* phylogroups and their association according to the isolation source. The results indicated that phylogroups B1 (35.8%), A (27.8%), D (12.3%), and B2 (11.8%) were the most prevalent among *E. coli* strains from different sources in Mexico. As indicated by the Shannon and Simpson diversity index analyses, the present study found that *E. coli* strains from human and animal sources had high diversities in the phylogroups distribution. In agreement with previous observations, a high diversity has also been previously reported for human and animal sources when examining the phylogenetic subgroup distribution data for *E. coli* strains from other geographical locations (Carlos et al., 2010). According to the associations of phylogroup frequency to isolation source, the results from this study indicated that phylogroup A was strongly linked with human sources, followed by phylogroup B1 with a lower frequency. As previously reported, both phylogroups A and B1 were highly prevalent phylogroups detected in the analysis of 100,000 publicly available *E. coli* genome sequences (Abram et al., 2021), and phylogroups A and B1, comprised mostly of the commensal *E. coli* strains (Singh et al., 2017), were detected as the most common from human sources (Duriez et al., 2001). Other reports examining the worldwide phylogroup distribution from human *E. coli* strains demonstrated a high prevalence of phylogroup A in human stool samples (Bailey et al., 2010; de Stoppe et al., 2017). The phylogroup A prevalence in the human host could potentially be influenced by the geographical distribution of the strains, the genetics of the human host, human microbiota, hygiene habits, and socioeconomic factors (Massot et al., 2016; Tenaillon et al., 2010).

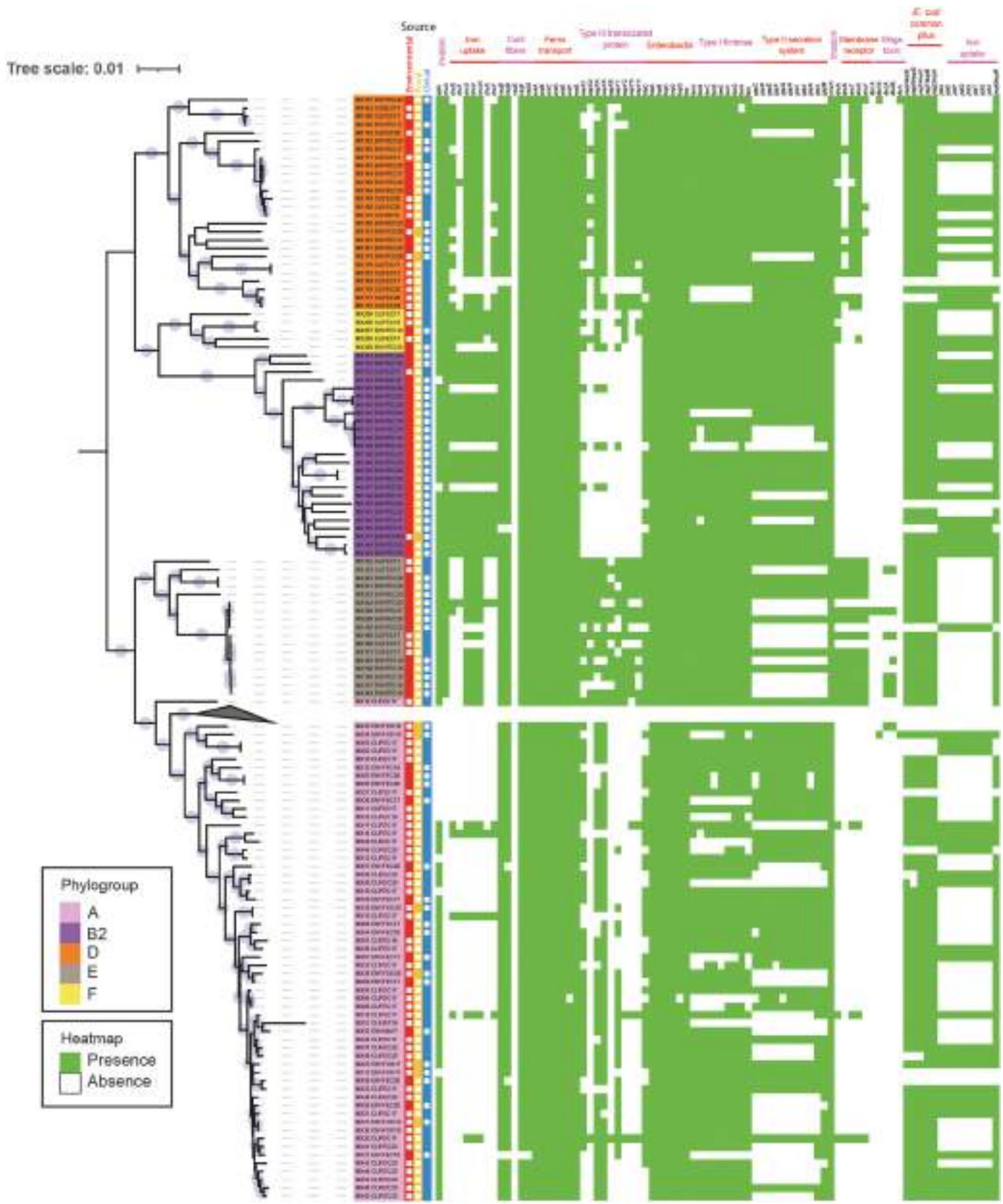

Fig. 2. Phylogenetic tree with the presence or absence of virulence factors in *E. coli* strains from various isolation sources in Mexico. Coloured labels represent the phylogroup designation, and bootstrapping >80 is displayed with blue dots. The presence or absence of the examined virulence genes are denoted in green and white, respectively. The virulence categories are shown at the top of the heatmap with each gene. Gray triangles represent collapsed clades to emphasize phylogroups visualization, and dashed lines connect tree branches to the strain label. Panel 2A shows phylogroups A, B2, D, E and F. Panel 2B shows phylogroups A, B1, and C. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

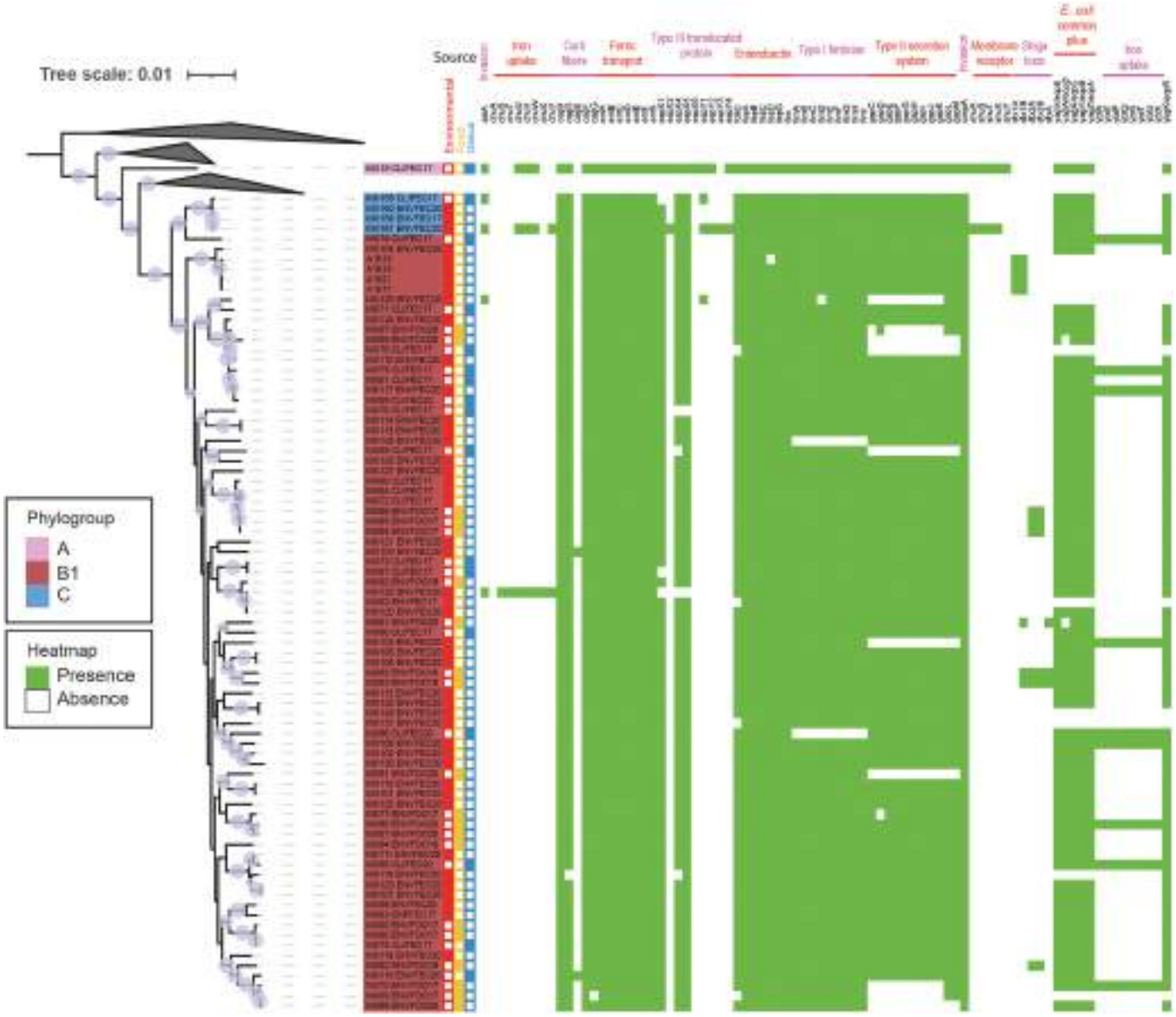

Fig. 2. (*continued*).

Interestingly, the scientific literature has described that phylogroups A and B1 comprise most of the commensal *E. coli* strains (Singh et al., 2017). These phylogroups have been identified in humans with different lifestyles and hygiene status (Massot et al., 2016), and this observation could explain the highest prevalence of phylogroup B1(76/212) and A (59/212) in Mexico. Although phylogroup D has not been reported as prevalent in the human host, the correlation analysis, conducted in the present study, indicated a frequency of 0.153 associated with clinical samples. This finding is in agreement with a previous report by Bailey et al. (2010) documenting a prevalence of 20.7% for phylogroup D in humans (Bailey et al., 2010). Moreover, phylogroup D had the largest virulence repertory genes, encoding for effector proteins translocated by the Type III Secretion System (T3SS), which is a complex system that enables *E. coli* strains to secrete and inject virulence determinants into host cells. These bacterial effectors confer persistence and survival of *E. coli* strains in the mammalian host cells by modulating various host cellular processes, including cytoskeleton rearrangements, apoptosis, phagocytosis as well as stimulation of the inflammatory response (Navarro-Garcia et al., 2016).

In the analysis of phylogroup association with the other examined sources, animal and food, phylogroup B2 showed the largest prevalence in wildlife and livestock samples, including bovines, pigs, and sea lions (Supplementary Table S1). Phylogroup B2 has been previously documented as prevalent among herbivorous and omnivorous mammals (Gordon and Cowling, 2003). Moreover, phylogroup E was found to be associated with cattle (Supplementary Table S1), and this finding was in agreement with published observations (Morcatti Coura et al., 2015). On identifying *E. coli* phylogroups linked to food samples, phylogroups A and B1 were detected as the most prevalent groups identified from different food sources, and, in particular, phylogroup B1 showed the higher correlation to the food samples. Limited published reports have examined the association of phylogroups with food samples, and an association of food samples derived from chicken, turkey, pork, and beef has been previously identified (Jakobsen et al., 2010; Kaesbohrer et al., 2019).

By constructing a heatmap to indicate either the presence or absence of virulence factors, the specific identification of markers in the examined *E. coli* strains belonging to certain phylogroups was determined in this study. For strains in phylogroups B2, D, E, and F, the virulence factor analysis confirmed the presence of the *chu* operon, required for iron

uptake using Fe^{2+} from hemoglobin in the host and for causing sepsis and infections of organs with very low iron conditions (Cherayil, 2011). For phylogroups D, E, and F, the *espY* gene family, coding for T3SS effectors, have been identified as markers in the identification of the *E. coli* pathotypes (Larzábal et al., 2018) and belonging to mainly phylogroup D (Finton et al., 2020). In the present study, the presence of the *espY* genes in phylogroup D strains could then be implemented for the identification of *E. coli* pathotypes in the strains identified in Mexico. Interestingly, *shuATSY* genes, iron-acquisition genes in *Shigella dysentery* and pathogenic *E. coli* (Kouse et al., 2013), were detected for phylogroups D and E in the present study. Furthermore, the presence of *shuA* in pathogenic *E. coli* has been positively correlated with virulence by *in vitro* and epidemiologic studies of human infections in uropathogenic *E. coli* (UPEC), EPEC, and EAEC (Okeke et al., 2004), enabling the potential identification of these pathotypes in phylogroup D strains.

The role of the *aslA* gene has been previously determined in the invasion by *E. coli* of brain microvascular endothelial cells (Hoffman et al., 2000). In the present study, *aslA* gene was predominantly absent in phylogroups B1 strains, and correlation analysis showed an association between phylogroup B1 and food sources. This finding could potentially be employed for a better characterization of the phylogroup B1 in strains collected from food sources in Mexico. Among the *E. coli* examined in this study in Mexico, the strains positive for the Shiga toxin (*stx*) genes, were mostly assigned to phylogroup E and to a lesser extent to phylogroup B1. In agreement with the findings of a recent publication, a survey of STEC/EHEC, recovered from cattle in Ireland, were also mostly assigned to phylogroup E (McCabe et al., 2019). By contrast, additional findings have reported phylogroup B1 as the predominant phylogroup classification for STEC/EHEC, and strains with phylogroup E were a small percentage of the total recovered strains (Badouei et al., 2015; Jajarmi et al., 2018; Mainda et al., 2016; Van Overbeek et al., 2020), indicating a difference in the prevalence of phylogroups among STEC, recovered from distinct sources and geographical locations.

In summary, this report found evidence that phylogroups identified in *E. coli* strains, recovered from clinical, animals, and food samples were found to harbor repertoires of virulence genes associated with iron uptake, fimbrial adhesion, curli fibers, membrane stability, and effector secretion. In addition, the use of comparative genomics and bioinformatics analyses provided evidence that the classification of virulence markers and phylogroups in *E. coli* could potentially be used for source tracking or for identifying transmission routes of contamination.

5. Conclusions

The present study employed the ClermontTyping method, a robust technique for *E. coli* phylogroup classification and grouping into phylogenetic clades (Beghain et al., 2018), to determine the associations between phylogroups, virulence profiles and isolation sources for a collection of hundreds of *E. coli* strains recovered in Mexico. The findings from this study identified phylogroup A as a prevalent and strongly associated phylogroup among *E. coli* strains from human sources. Additionally, phylogroup D was also related to human sources. Phylogroup B2 was associated with herbivorous and omnivorous mammals like cattle and pigs. Moreover, the B1 phylogenetic group was significantly related to food sources. Common virulence genes in the examined *E. coli* strains, assigned to all phylogroups, were identified as essential markers for survival and invasion in the host. In our study, phylogroup D, associated with humans, was found to have a high virulence gene content when compared to the other phylogroups. The correlation analysis showed that *E. coli* populations from distinct sources in Mexico, belonging to different phylogroups, were not dispersed randomly and were associated with a particular isolation source. In addition, virulence factor characterization showed a specific association among phylogroups, even exclusive presence of virulence profiles for certain phylogroups. These findings would contribute to a better characterization of the epidemiology of *E. coli* populations and the relationship with specific hosts and contamination sources.

Authors' contributions

Formal analysis and investigation: José Roberto Aguirre-Sánchez, José Benigno Valdez-Torres and Bertram G. Lee. Writing – original draft preparation: José Roberto Aguirre-Sánchez and Beatriz Quiñones. Writing – review and editing: José Roberto Aguirre-Sánchez, Cristóbal Chaidez-Quiroz, Beatriz Quiñones, Nohemí Castro del Campo, Nohelia Castro del Campo, and Bertram Lee. Resources: Nohelia Castro del Campo, Jaime Martinez-Urtaza and Beatriz Quiñones. Supervision: Cristóbal Chaidez-Quiroz and Jaime Martinez-Urtaza.

Ethical approval

Not required. This article does not contain any studies with human participants or animals performed by any of the authors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link to my data

Acknowledgments

This material was based in part upon work supported by the National Laboratory for Food Safety and Research (LANIIA) at Centro de Investigación y Desarrollo A. C. (CIAD) at Culiacán, Sinaloa and by the United States Department of Agriculture (USDA), Agricultural Research Service, CRIS Project 2030-42000-055-00D.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2022.105380.

References

Abram, K., Udaondo, Z., Bleker, C., Wanchai, V., Wassenaar, T.M., Robeson, M.S., Ussery, D.W., 2021. Mash-based analyses of *Escherichia coli* genomes reveal 14 distinct phylogroups. Commun. Biol. 4, 117. https://doi.org/10.1038/s42003-020-01626-5.

Arend, D.N., 1993. Choices (Version 4.0) [Computer software]. Champaign, US Army Corps Eng. Res. Lab. Rep. (No. CH7-22510).

Assefa, S., Keane, T.M., Otto, T.D., Newbold, C., Berriman, M., 2009. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics 25, 1968–1969.

Badouei, M.A., Jajarmi, M., Mirsalehian, A., 2015. Virulence profiling and genetic relatedness of Shiga toxin-producing *Escherichia coli* isolated from humans and ruminants. Comp. Immunol. Microbiol. Infect. Dis. 38, 15–20.

Bailey, J.K., Pinyon, J.L., Anantham, S., Hall, R.M., 2010. Distribution of human commensal Escherichia coli phylogenetic groups. J. Clin. Microbiol. 48, 3455–3456.

Beghain, J., Bridier-Nahmias, A., Le Nagard, H., Denamur, E., Clermont, O., 2018. ClermonTyping: an easy-to-use and accurate in silico method for *Escherichia* genus strain phylotyping. Microb. Genom. 4, e000192 https://doi.org/10.1099/mgen.0.000192.

Bushnell, B., Rood, J., Singer, E., 2017. BBMerge – accurate paired shotgun read merging via overlap. PLoS One 12, e0185056.

Carlos, C., Pires, M.M., Stoppe, N.C., Hachich, E.M., Sato, M.I.Z., Gomes, T.A.T., Amaral, L.A., Ottoboni, L.M.M., 2010. *Escherichia coli* phylogenetic group determination and its application in the identification of the major animal source of fecal contamination. BMC Microbiol. 10, 1–10.

Cherayil, B.J., 2011. The role of iron in the immune response to bacterial infection. Immunol. Res. 50, 1–9.

Clermont, O., Christenson, J.K., Denamur, E., Gordon, D.M., 2013. The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. Environ. Microbiol. Rep. 5, 58–65. https://doi.org/10.1111/1758-2229.12019.

Coil, D., Jospin, G., Darling, A.E., 2015. A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. Bioinformatics 31, 587–589.
Corzo-Ariyama, H.A., García-Heredia, A., Heredia, N., García, S., León, J., Jaykus, L., Solís-Soto, L., 2019. Phylogroups, pathotypes, biofilm formation and antimicrobial resistance of Escherichia coli isolates in farms and packing facilities of tomato, jalapeño pepper and cantaloupe from northern Mexico. Int. J. Food Microbiol. 290, 96–104.
Coura, F.M., de Araújo Diniz, S., Silva, M.X., Mussi, J.M.S., Barbosa, S.M., Lage, A.P., Heinemann, M.B., 2015. Phylogenetic group determination of *Escherichia coli* isolated from animals samples. Sci. World J. 2015.
de Stoppe, N.C., Silva, J.S., Carlos, C., Sato, M.I.Z., Saraiva, A.M., Ottoboni, L.M.M., Torres, T.T., 2017. Worldwide phylogenetic group patterns of *Escherichia coli* from commensal human and wastewater treatment plant isolates. Front. Microbiol. 8, 2512. https://doi.org/10.3389/fmicb.2017.02512.
Duriez, P., Clermont, O., Bonacorsi, S., Bingen, E., Chaventre, A., Elion, J., Picard, B., Denamur, E., 2001. Commensal *Escherichia coli* isolates are phylogenetically distributed among geographically distinct human populations. Microbiology 147, 1671–1676.
Escobar-Páramo, P., Le Menac'h, A., Le Gall, T., Amorin, C., Gouriou, S., Picard, B., Skurnik, D., Denamur, E., 2006. Identification of forces shaping the commensal *Escherichia coli* genetic structure by comparing animal and human isolates. Environ. Microbiol. 8, 1975–1984. https://doi.org/10.1111/j.1462-2920.2006.01077.x.
Finton, M.D., Meisal, R., Porcellato, D., Brandal, L.T., Lindstedt, B.-A., 2020. Whole genome sequencing and characterization of multidrug-resistant (MDR) bacterial strains isolated from a Norwegian university campus pond. Front. Microbiol. 11, 1273.
Giraudoux, P., Giraudoux, M.P., Mass, S., 2018. Package 'pgirmess.'. Spat. Anal. Data Min. F. Ecol.
Gordon, D.M., Cowling, A., 2003. The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects. Microbiology 149, 3575–3586. https://doi.org/10.1099/mic.0.26486-0.
Gordon, D.M., Clermont, O., Tolley, H., Denamur, E., 2008. Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. Environ. Microbiol. 10, 2484–2496. https://doi.org/10.1111/j.1462-2920.2008.01669.x.
Greenacre, M., 2007. Correspondence Analysis in Practice. Chapman and Hall/CRC.
Higgins, J., Hohn, C., Hornor, S., Frana, M., Denver, M., Joerger, R., 2007. Genotyping of *Escherichia coli* from environmental and animal samples. J. Microbiol. Methods 70, 227–235. https://doi.org/10.1016/J.MIMET.2007.04.009.
Hoffman, J.A., Badger, J.L., Zhang, Y., Huang, S.-H., Kim, K.S., 2000. *Escherichia coli* K1 aslA contributes to invasion of brain microvascular endothelial cells in vitro and in vivo. Infect. Immun. 68, 5062–5067.
Hunter, P.R., Gaston, M.A., 1988. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J. Clin. Microbiol. 26, 2465–2466.
Jajarmi, M., Askari Badouei, M., Imani Fooladi, A.A., Ghanbarpour, R., Ahmadi, A., 2018. Pathogenic potential of Shiga toxin-producing *Escherichia coli* strains of caprine origin: virulence genes, Shiga toxin subtypes, phylogenetic background and clonal relatedness. BMC Vet. Res. 14, 1–8.
Jakobsen, L., Kurbasic, A., Skjøt-Rasmussen, L., Ejrnæs, K., Porsbo, L.J., Pedersen, K., Jensen, L.B., Emborg, H.-D., Agersø, Y., Olsen, K.E.P., 2010. *Escherichia coli* isolates from broiler chicken meat, broiler chickens, pork, and pigs share phylogroups and antimicrobial resistance with community-dwelling humans and patients with urinary tract infection. Foodborne Pathog. Dis. 7, 537–547.
Jolley, K.A., Bray, J.E., Maiden, M.C.J., 2018. Open-access bacterial population genomics: BIGSdb software, the PubMLST. Org website and their applications. Wellcome Open Res. 3.
Kaesbohrer, A., Bakran-Lebl, K., Irrgang, A., Fischer, J., Kämpf, P., Schiffmann, A., Werckenthin, C., Busch, M., Kreienbrock, L., Hille, K., 2019. Diversity in prevalence and characteristics of ESBL/pAmpC producing E. coli in food in Germany. Vet. Microbiol. 233, 52–60.
Kaper, J.B., Nataro, J.P., Mobley, H.L.T., 2004. Pathogenic *Escherichia coli*. Nat. Rev. Microbiol. 2, 123–140. https://doi.org/10.1038/nrmicro818.
Kauffmann, F., 1947. Review: the serology of the coli group. J. Immunol. 57, 71–100.
Kouse, A.B., Righetti, F., Kortmann, J., Narberhaus, F., Murphy, E.R., 2013. RNA-mediated thermoregulation of iron-acquisition genes in *Shigella* dysenteriae and pathogenic Escherichia coli. PLoS One 8, e63781.
Krueger, F., 2015. Trim galore. A wrapper tool around Cutadapt FastQC to consistently apply Qual. Adapt. trimming to FastQ files 516, 517.
Larzábal, M., Marques Da Silva, W., Riviere, N.A., Cataldi, Á.A., 2018. Novel effector protein EspY3 of type III secretion system from Enterohemorrhagic *Escherichia coli* is localized in actin pedestals. Microorganisms 6, 112.
Letunic, I., Bork, P., 2019. Interactive tree of life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47, W256–W259. https://doi.org/10.1093/nar/gkz239.
Liu, B., Zheng, D., Jin, Q., Chen, L., Yang, J., 2019. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. Nucleic Acids Res. 47, D687–D692.
Mainda, G., Lupolova, N., Sikakwa, L., Bessell, P.R., Muma, J.B., Hoyle, D.V., McAteer, S. P., Gibbs, K., Williams, N.J., Sheppard, S.K., 2016. Phylogenomic approaches to determine the zoonotic potential of Shiga toxin-producing *Escherichia coli* (STEC) isolated from Zambian dairy cattle. Sci. Rep. 6, 1–8.
Massot, M., Daubié, A.-S., Clermont, O., Jaureguy, F., Couffignal, C., Dahbi, G., Mora, A., Blanco, J., Branger, C., Mentré, F., 2016. Phylogenetic, virulence and antibiotic resistance characteristics of commensal strain populations of *Escherichia coli* from community subjects in the Paris area in 2010 and evolution over 30 years. Microbiology 162, 642.
McCabe, E., Burgess, C.M., Lawal, D., Whyte, P., Duffy, G., 2019. An investigation of shedding and super-shedding of Shiga toxigenic *Escherichia coli* O157 and E. coli O26 in cattle presented for slaughter in the Republic of Ireland. Zoonoses Public Health 66, 83–91.
Navarro-Garcia, F., Ruiz-Perez, F., Larzábal, M., Cataldi, A., 2016. Secretion systems of pathogenic *Escherichia coli*. In: *Escherichia Coli* in the Americas. Springer, pp. 221–249.
Odonkor, S.T., Ampofo, J.K., 2013. *Escherichia coli* as an indicator of bacteriological quality of water: an overview. Microbiol. Res. (Pavia). 4, 2. https://doi.org/10.4081/mr.2013.e2.
Okeke, I.N., Scaletsky, I.C.A., Soars, E.H., Macfarlane, L.R., Torres, A.G., 2004. Molecular epidemiology of the iron utilization genes of enteroaggregative *Escherichia coli*. J. Clin. Microbiol. 42, 36–44.
Riley, L.W., 2014. Pandemic lineages of extraintestinal pathogenic *Escherichia coli*. Clin. Microbiol. Infect. 20, 380–390.
Selander, R.K., Caugant, D.A., Ochman, H., Musser, J.M., Gilmour, M.N., Whittam, T.S., 1986. Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. Appl. Environ. Microbiol. 51, 873–884.
Simpson, E.H., 1949. Measurement of diversity. Nature 163, 688.
Singh, T., Das, S., Ramachandran, V.G., Wani, S., Shah, D., Maroof, K.A., Sharma, A., 2017. Distribution of integrons and phylogenetic groups among enteropathogenic *Escherichia coli* isolates from children< 5 years of age in Delhi, India. Front. Microbiol. 8, 561.
Souza, V., Rocha, M., Valera, A., Eguiarte, L.E., 1999. Genetic structure of natural populations *Escherichia coli* in wild hosts on different continents. Appl. Environ. Microbiol. 65, 3373–3385.
Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.
Team, R.C., 2022. R: A Languaje and Environment for Statistical Computing.
Tenaillon, O., Skurnik, D., Picard, B., Denamur, E., 2010. The population genetics of commensal *Escherichia coli*. Nat. Rev. Microbiol. 8, 207–217.
Treangen, T.J., Ondov, B.D., Koren, S., Phillippy, A.M., 2014. The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol. 15, 524.
Van Overbeek, L.S., Wichers, J.H., Van Amerongen, A., Van Roermund, H.J.W., Van der Zouwen, P., Willemsen, P.T.J., 2020. Circulation of Shiga toxin-producing *Escherichia coli* phylogenetic group B1 strains between calve stable manure and pasture land with grazing heifers. Front. Microbiol. 11, 1355.
Wingett, S.W., Andrews, S., 2018. FastQ Screen: A Tool for Multi-genome Mapping and Quality Control. https://doi.org/10.12688/f1000research.15931.2.