frontiers in Microbiology

ORIGINAL RESEARCH
published: 21 December 2017
doi: 10.3389/fmicb.2017.02512

Worldwide Phylogenetic Group Patterns of *Escherichia coli* from Commensal Human and Wastewater Treatment Plant Isolates

*Nancy de Castro Stoppe[1, 2], Juliana S. Silva[2, 3, 4], Camila Carlos[1], Maria I. Z. Sato[5], Antonio M. Saraiva[2, 6], Laura M. M. Ottoboni[1] and Tatiana T. Torres[2, 4]**

[1] Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, Campinas, Brazil, [2] Núcleo de Pesquisa em Biodiversidade e Computação (BioComp-USP)-Universidade de São Paulo, São Paulo, Brazil, [3] Secretaria de Estado de Saúde de Mato Grosso, Cuiabá, Brazil, [4] Departamento de Genética e Biologia Evolutiva, Universidade de São Paulo, São Paulo, Brazil, [5] Departamento de Análises Ambientais, Companhia Ambiental do Estado de São Paulo-CETESB, São Paulo, Brazil, [6] Departamento de Engenharia de Computação e Sistemas Digitais, Escola Politécnica da USP, São Paulo, Brazil

OPEN ACCESS

Edited by:
*Hongyue Dang,
Xiamen University, China*

Reviewed by:
*Scott H. Harrison,
North Carolina Agricultural and Technical State University, United States
Qaisar Mahmood,
COMSATS Institute of Information Technology, Pakistan*

***Correspondence:**
*Tatiana T. Torres
tttorres@ib.usp.br*

Specialty section:
This article was submitted to Aquatic Microbiology, a section of the journal Frontiers in Microbiology

Received: *02 July 2017*
Accepted: *04 December 2017*
Published: *21 December 2017*

Citation:
Stoppe NdC, Silva JS, Carlos C, Sato MIZ, Saraiva AM, Ottoboni LMM and Torres TT (2017) Worldwide Phylogenetic Group Patterns of Escherichia coli from Commensal Human and Wastewater Treatment Plant Isolates. Front. Microbiol. 8:2512. doi: 10.3389/fmicb.2017.02512

Escherichia coli is an important microorganism in the gastrointestinal tract of warm-blooded animals. Commensal populations of *E. coli* consist of stable genetic isolates, which means that each individual has only one phylogenetic group (phylogroup). We evaluated the frequency of human commensal *E. coli* phylogroups from 116 people and observed that the majority of isolates belonged to group A. We also evaluated the frequency of phylogroups in wastewater samples and found a strong positive correlation between the phylogroup distribution in wastewater and human hosts. In order to find out if some factors, such as geographical location, and climate could influence the worldwide phylogroup distribution, we performed a meta-analysis of 39 different studies and 24 countries, including different climates, living areas, and feeding habits. Unexpectedly, our results showed no substructuring patterns of phylogroups; indicating there was no correlation between phylogroup distribution and geographic location, climate, living area, feeding habits, or date of collection.

Keywords: phylogenetic groups, *Escherichia coli*, multivariate analysis, wastewater, commensal strains

INTRODUCTION

Escherichia coli is a facultative anaerobic microorganism found in the gastrointestinal tract of warm-blooded animals, with which it maintains a mostly symbiotic relationship. *E. coli* has also been found in soil, water, and sediments not directly influenced by sewage discharges. Besides commensal strains, there exist also pathogenic variants of *E. coli*, capable of causing either intestinal or extraintestinal diseases. Pathogenic strains were probably derived from commensal strains following the horizontal acquisition of chromosomal and extrachromosomal genes and operons, as well as gene loss (Tallon et al., 2005; Whitman et al., 2006; Tenaillon et al., 2010). The pangenome structure of *E. coli*, which includes pathogenic and commensal isolates, comprises a core of 2,200 genes, a broad range of unique genes, and a reservoir with more than 13,000 genes, indicating that these bacteria have a high potential for diversity and pathogenesis (Rasko et al., 2008). *E. coli* populations within a host are shaped by multiple host and environmental factors. In spite of recombination events, these populations are commonly described as having a clonal structure (Tenaillon et al., 2010), as demonstrated by multilocus enzyme electrophoresis, serotyping, biotyping, random amplified polymorphic DNA, and restriction fragment length polymorphism (Selander and Levin, 1980; Caugant et al., 1985; Miller and Hartl, 1986; Desjardins et al., 1995). This clonal character derives from the fact that, at any given time, each person has a predominant strain that constitutes more than half of the isolated colonies. Intra-host diversity nevertheless varies among human populations. Analysis of human commensal *E. coli* from different climate regions showed that subjects in tropical areas exhibited more diverse *E. coli* microbiota than those in temperate areas (Escobar-Páramo et al., 2004). A number of host and environmental factors influence inter-host diversity. Many *E. coli* clones have a broad geographical distribution (Ochman and Selander, 1984) and are shared by hosts of different species. Although the abundance of different groups varies according to species, four phylogenetic groups are predominant in several human and animal populations (Clermont et al., 2000). These main groups (A, B1, B2, and D) were firstly identified by multilocus enzyme electrophoresis, and were also recovered by multilocus sequence typing using 2.6 million nucleotides of the *E. coli* genome (reviewed in Tenaillon et al., 2010), indicating that these groups are genetic entities. Despite not being necessarily monophyletic, the similarity of the results obtained by enzyme electrophoresis and sequence typing, illustrates that these groups are still useful to cluster strains in a relevant way (Tenaillon et al., 2010).

Among the tools for studying *E. coli* population genetics, phylogrouping triplex PCR has been widely applied owing to its simplicity and rapidity (Clermont et al., 2000; Tenaillon et al., 2010). This method can quickly assign *E. coli* isolates to one of the four major phylogenetic groups (phylogroup): A, B1, D, or B2, which makes it useful for population genetics, classification of extraintestinal pathogenic and commensal strains, and host-source relationships (Duriez et al., 2001; Escobar-Páramo et al., 2004; Orsi et al., 2007; Gordon et al., 2008; Unno et al., 2009; Carlos et al., 2010; Tenaillon et al., 2010).

Human commensal strains belong mostly to group A (43%; Escobar-Páramo et al., 2006; Li et al., 2010), however, in tropical areas both groups A and B1 are prevalent (Escobar-Páramo et al., 2004). Instead, strains isolated from animals fall mostly into group B1 (34–50%; Higgins et al., 2007; Ishii et al., 2007; Carlos et al., 2010), suggesting an association between phylogenetic groups and host species. Within the same host species, geography, climate, diet, gut morphology, body mass, sex, age, hygiene level, *inter alia* may be associated with the distribution of phylogroups (Gordon and Cowling, 2003; Escobar-Páramo et al., 2006).

Bailey et al. (2010a) studied the worldwide distribution of phylogroups, but without correlating them with geographic location. Escobar-Páramo et al. (2004) hypothesized that the geographic and climatic location of human populations could have an important role in defining the *E. coli* phylogroup distribution. These authors observed that the French population was distinct from the American and the African ones and the 10 different locations were separated in two clusters, tropical and temperate zones. However, no formal tests were performed to study the factors underlying such difference. Li et al. (2010) compared human commensal *E. coli* isolated from China with those studied by Escobar-Páramo et al. (2004) and observed that, even though China belonged to the temperate belt, it shared a phylogroup distribution similar to that found in the tropics. No further climate classification was used in these studies. Tenaillon et al. (2010) compiled the prevalence of *E. coli* groups in humans and reported a shift from A to B2 as the most frequent *E. coli* phylogroup group in France in 1980 and 2000, respectively. Additional data showed that group A was the most common phylogroup in Africa (Mali and Benin), Asia (Pakistan), Europe (Croatia), and South America (French Guiana, Colombia, and Bolivia), while B2 was the most common phylogroup in Europe (Sweden), North America (USA), developed Asia (Japan), and Oceania (Australia). Based on these results, the authors suggested that socioeconomic factors rather than geographic location or climate could determine the phylogroup distribution.

Another way for studying the distribution of *E. coli* strains is using wastewater isolates as surrogates for human commensal *E. coli*. Isolates from wastewater can be treated as a pool of clones derived from a local human population, thus reducing the sampling effort. Furthermore, there is no need for approval of the study by an ethics committee. For these reasons, *E. coli* isolates from wastewater have been used extensively to study the distribution of phylogroups (USEPA, 2005; Duran et al., 2006; Jiang et al., 2007; Kaneene et al., 2007; Ahmed et al., 2009; Fremaux et al., 2009; Silkie and Nelson, 2009; Kelty et al., 2012). Phylogroups D and B2 were predominant among *E. coli* isolated from wastewater samples, regardless of the climatic zone, accounting for 50% of isolates in a temperate area and more than 95% in a subtropical one (Anastasi et al., 2010; Mokracka et al., 2011). These phylogroups are frequently associated with pathogenic strains of *E. coli* (Clermont et al., 2000). Hence, Anastasi et al. (2010) suggest the use of phylogroup as a simple tool to determine if strains are pathogenic or not.

Although dispersal limitation and selection are the main processes influencing the distribution of free-living microbes, the host microenvironment also plays a major role in defining the population structure of commensal isolates. Here, we tested for the presence of discontinuity against the null hypothesis, in which the distribution of phylogenetic groups was limited by dispersal across space.

Differences in phylogroup distribution have been observed in many independent studies, and these differences were attributed to climate, feeding habits, geographic location, *inter alia*. However, no formal statistical framework was used to test the effect of these parameters in phylogroup distribution. To fill this gap, we compared the phylogenetic distribution of *E. coli* strains isolated from human and wastewater samples. Furthermore, we performed a meta-analysis of the worldwide phylogroup distribution from human *E. coli* isolates, and looked for patterns of association between the commensal isolates and host geographic location, climate, host feeding habits, or host living area, indicating local adaptation. We also tested if collection date could explain the distribution of phylogroups.

TABLE 1 | Distribution of commensal *E. coli* phylogenetic groups isolated from human feces.

#	City or region, country	Koppen climate classification*	Feeding habits	Living area	Approximated geographic coordinates		Phylogenetic group [*n* (%)]				References
							A	B1	B2	D	
1	São Paulo, Brazil	Aw	Western	Urban	23°32′S	46°38′W	56 (48.3)	6 (5.2)	19 (16.4)	35 (30.2)	This study
2	Sydney, Australia	Cfa	Western	Urban	33°52′S	151°12′E	39 (37.1)	10 (9.5)	31 (29.5)	25 (23.8)	Bailey et al., 2010a,b
3	Maroochydore, Australia	Cfa	Western	Urban	26°39′S	153°05′E	23 (15.2)	16 (10.6)	43 (28.5)	69 (45.7)	Vollmerhausen et al., 2011
4	Cotonou, Benin	Aw	Other	Urban	06°22′N	02°26′E	23 (50.0)	15 (32.6)	8 (17.4)	0 (0.0)	Escobar-Páramo et al., 2004
5	Alto de los Zarzos, Bolivia	Aw	Other	Rural	21°28′S	63°54′W	87 (77.0)	11 (10.0)	6 (5.0)	9 (8.0)	Pallechi et al., 2007
6	Villamontes, Bolivia	Aw	Other	Urban	21°15′S	63°32′W	23 (79.3)	0 (0.0)	0 (0.0)	6 (20.7)	Riccobono et al., 2012
7	São Paulo, Brazil	Aw	Western	Urban	23°32′S	46°38′W	38 (40.4)	8 (8.5)	12 (12.8)	36 (38.3)	Carlos et al., 2010
8	Calgary, Canada	Dfc	Western	NA	51°02′N	114°03′W	15 (13.9)	13 (12.0)	58 (53.7)	22 (20.4)	White et al., 2011
9	Fuzhou, China	Cfa	Other	Urban	26°04′N	119°17′E	142 (43.7)	76 (23.4)	52 (16.0)	55 (16.9)	Li et al., 2010
10	Beijing, China	Dwa	Other	Urban	39°54′N	116°24′E	11 (12.0)	0 (0.0)	44 (47.8)	37 (40.2)	Luo et al., 2011
11	Bogota, Colombia	Cfb	Western	Urban	04°35′N	74°04′W	16 (57.1)	1 (3.6)	7 (25.0)	4 (14.3)	Escobar-Páramo et al., 2004
12	Olib and Silba, Croatia	Csa	Western	NA	44°22′N	14°43′E	20 (35.1)	18 (31.6)	11 (19.3)	8 (14.0)	Duriez et al., 2001
13	Copenhagen, Denmark	Cfb	Western	NA	55°40′N	12°34′E	35 (20.5)	38 (22.2)	50 (29.2)	48 (28.1)	Damborg et al., 2009; Petersen et al., 2009; Jakobsen et al., 2010
14	Paris, France	Cfb	Western	Urban	48°51′N	02°15′E	113 (48.3)	21 (9.0)	43 (18.4)	57 (24.4)	Duriez et al., 2001; Escobar-Páramo et al., 2004; Leflon-Guibout et al., 2008
15	Brittany, France	Cfb	Western	Rural	48°12′N	02°55′W	14 (28.0)	13 (26.0)	12 (24.0)	11 (22.0)	Escobar-Páramo et al., 2004
16	Brest, France	Cfb	Western	Rural	48°23′N	04°29′W	3 (14.3)	5 (23.8)	7 (33.3)	6 (28.6)	Escobar-Páramo et al., 2004
17	Tours, France	Cfb	Western	Urban	47°23′N	00°37′E	6 (25.0)	5 (21.0)	7 (29.0)	6 (25.0)	Escobar-Páramo et al., 2004
18	Western France	Cfb	Western	Urban	48°06′N	01°41′W	12 (48.0)	3 (12.0)	5 (20.0)	5 (20.0)	Mereghetti et al., 2002
19	National Park, French Guiana	Af	Other	Rural	02°35′N	53°33′W	59 (63.4)	19 (20.4)	3 (3.2)	12 (12.9)	Escobar-Páramo et al., 2004

(Continued)

TABLE 1 | Continued

#	City or region, country	Koppen climate classification*	Feeding habits	Living area	Approximated geographic coordinates		Phylogenetic group [*n* (%)]				References
							A	B1	B2	D	
20	Munchengladbach, Germany	Cfb	Western	NA	48°08′N	11°34′E	16 (43.3)	10 (27.0)	6 (16.2)	5 (13.5)	Sorsa et al., 2007
21	Babylon, Iraq	BWh	Other	Urban	32°28′N	44°33′E	6 (60.0)	3 (30.0)	0 (0.0)	1 (10.0)	Abdul-Razzaq and Abdul-Lateef, 2011
22	Tokyo, Japan	Cfa	Other	Urban	35°41′N	139°41′E	30 (16.6)	17 (9.4)	91 (50.3)	43 (23.8)	Obata-Yasuoka et al., 2002; Harada et al., 2012
23	Kyoto, Japan	Cfa	Other	NA	35°00′N	135°46′E	21 (42.0)	16 (32.0)	10 (20.0)	2 (6.0)	Kanamaru et al., 2006
24	Seoul, Korea	Dwa	Other	Urban	37°34′N	126°58′E	78 (38.0)	37 (18.0)	47 (22.9)	43 (21.0)	Lee et al., 2010
25	Jeonnam Province, Korea	Dwa	Other	NA	34°52′N	126°59′E	42 (29.8)	48 (34.0)	0 (0.0)	31 (36.2)	Unno et al., 2009
26	Bay Reserve, Mali	As	Other	NA	13°00′N	05°00′E	13 (23.6)	32 (58.2)	1 (1.8)	9 (16.4)	Duriez et al., 2001
27	Oslo and Telemark, Norway	Dfb	Western	NA	59°54′N	10°45′E	6 (30.0)	5 (25.0)	3 (15.0)	6 (30.0)	Grude et al., 2007
28	Lahore, Pakistan	BSh	Other	Urban	31°32′N	74°20′E	74 (47.0)	28 (18.0)	19 (12.0)	36 (23.0)	Nowrouzian et al., 2009
29	Villa Real, Portugal	Csa	Western	NA	41°18′N	07°31′W	5 (8.6)	6 (10.3)	38 (65.5)	9 (15.5)	Silva et al., 2010
30	Angaiza, Peru	Af	Other	Rural	03°35′S	71°36′W	80 (72.0)	19 (17.0)	3 (3.0)	9 (8.0)	Bartoloni et al., 2009
31	Madrid, Spain	Csa	Western	Urban	40°25′N	03°42′W	28 (49.1)	11 (19.3)	4 (7.0)	14 (24.6)	Machado et al., 2005; Valverde et al., 2009
32	Barcelona, Spain	Csa	Western	Urban	41°23′N	02°10′E	40 (33.0)	23 (19.0)	20 (17.0)	37 (31.0)	Moreno et al., 2009
33	Gothenburg, Sweden	Cfb	Western	NA	57°42′N	11°58′E	147 (30.5)	61 (12.7)	205 (42.5)	69 (14.3)	Nowrouzian et al., 2005, 2006; Karami, 2007
34	Geneva and Ticino, Switzerland	Cfb	Western	Urban	46°05′N	07°30′E	4 (40.0)	1 (10.0)	2 (20.0)	3 (30.0)	Grasselli et al., 2009
35	Boise, United States	BSk	Western	Rural	43°33′N	116°22′W	32 (26.2)	17 (13.9)	35 (28.7)	38 (31.1)	Hannah et al., 2009
36	Minneapolis, United States	Dfb	Western	Urban	44°59′N	93°22′W	20 (13.6)	22 (15.0)	78 (53.1)	27 (18.4)	Zhang et al., 2002; Sannes et al., 2004; Johnson et al., 2005; Logue et al., 2012
37	Ann Arbor, United States	Dfa	Western	Urban	42°14′N	83°44′W	18 (20.5)	11 (12.5)	42 (47.7)	17 (19.3)	Zhang et al., 2002
38	Fargo, United States	Dfb	Western	Urban	46°52′N	96°47′W	33 (16.2)	32 (15.7)	110 (53.9)	29 (14.2)	Logue et al., 2012
	Total						1,428	677	1,132	879	

*Peel et al., 2007. *NA, not available.*

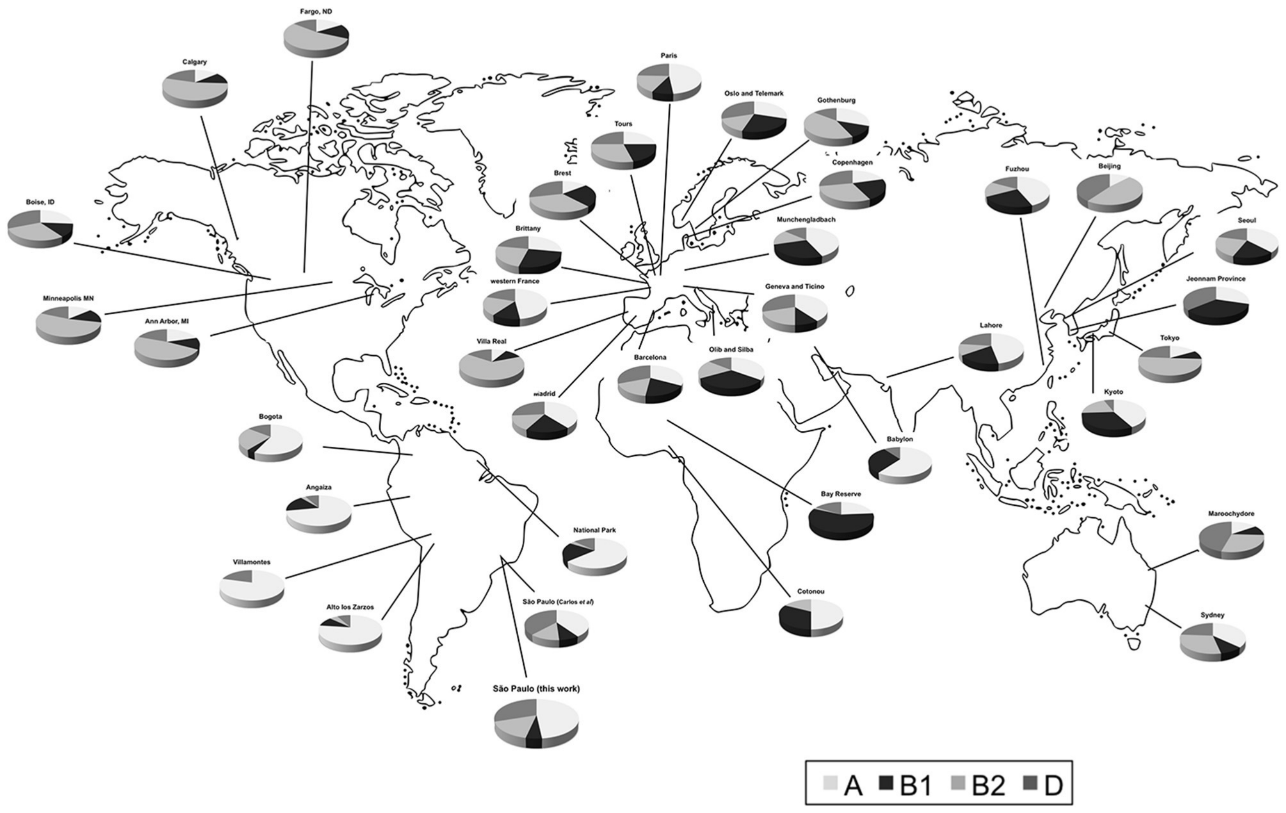

FIGURE 1 | Distribution of human commensal *E. coli* phylogenetic groups in selected studies from different sites.

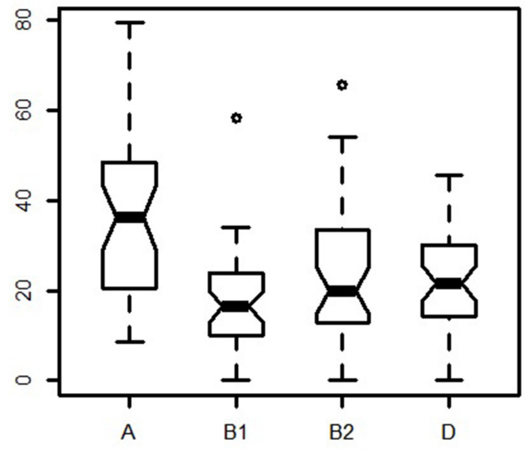

FIGURE 2 | Notched boxplot of meta-analysis phylogroup distribution.

MATERIALS AND METHODS

E. coli Isolation from Human Feces

Feces from 116 adult humans were collected using sterile swabs and Cary-Blair transport medium. They were streaked on Endo agar LES (Difco) and incubated for 24 h at 35°C. Three lactose-positive colonies (pink to dark-red with a metallic surface sheen) were picked from each sample and tested for citrate utilization, lactose fermentation, oxidase activity, L-lysine decarboxylase activity, motility, glucose and sucrose fermentation, tryptophan deamination, indole production, urea hydrolysis, and sulfide production. One typical *E. coli* profile strain from each individual host was re-isolated on nutrient agar, incubated for 24 h at 35°C, and kept at −70°C in tryptic soy broth (Difco) with 10% glycerol (v/v) for further analysis (ATCC, 2017).

The Research Ethics Committee of the State University of Campinas School of Medical Sciences has approved the present study (Permission 049/09) and all participants gave their informed written consent. Human samples were collected from healthy individuals, representing equal numbers of males and females living in the São Paulo Metropolitan area. The age interval of the subjects ranged from 19 to 79 years (average = 43) and BMI (body mass index) ranged from 17.5 to 40.9 (average = 26). The majority of subjects (>89%) had omnivore feeding habits.

E. coli Isolation from Raw Wastewater

Five wastewater treatment plants (WTPs) in the São Paulo Metropolitan area (Parque Novo Mundo, Jesus Netto, Barueri, São Lourenço da Serra, and Vinhedo) were sampled two to

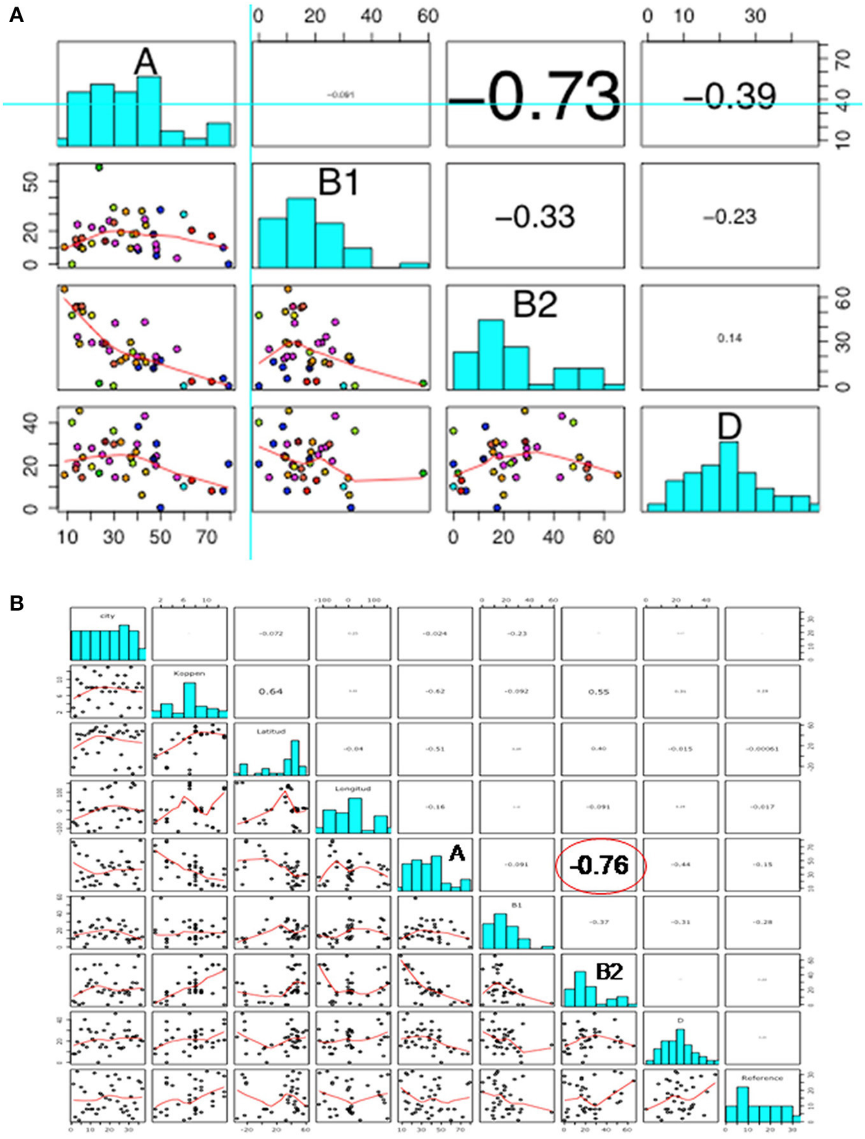

FIGURE 3 | Scatterplot showing **(A)** correlation among phylogroups; and **(B)** correlation among phylogroups, feeding habits, living area, and geographical location. In both scatterplots the figures are divided in three parts, representing: (1) upper panel, correlation analysis between variables; (2) diagonal panel, histogram of the phylogroup distribution data; and (3) lower panel, the smooth function of the data.

three times. All WTPs received domestic discharges, while Parque Novo Mundo and Barueri plants received also industrial discharges. Raw wastewater samples were collected in sterile bottles according to standard methods (APHA, 2010) and samples were analyzed using the membrane filter technique according to U.S. Environmental Protection Agency method 1603 (USEPA, 2002) as previously described by Stoppe et al. (2014).

DNA Isolation and Phylogenetic Grouping

Genomic DNA from *E. coli* strains was isolated with the Wizard Genomic DNA Purification Kit (Promega) according manufacturer's instructions. The phylogroup of *E. coli* strains was determined by triplex PCR as previously described by Clermont et al. (2000) and Stoppe et al. (2014).

Meta-Analysis of Human Isolates

Meta-analysis was performed using data from this and previous studies (**Table 1**). Commensal *E. coli* strains isolated from healthy humans were analyzed in terms of Koppen climate classification (Peel et al., 2007), feeding habits (western or other), living area (urban or rural), and geographic distance (km).

Data were selected according to the following criteria: (1) *E. coli* were isolated from feces of healthy individuals (both sexes); (2) one isolate per individual or representative strain was obtained; (3) the phylogroup classification was done according to

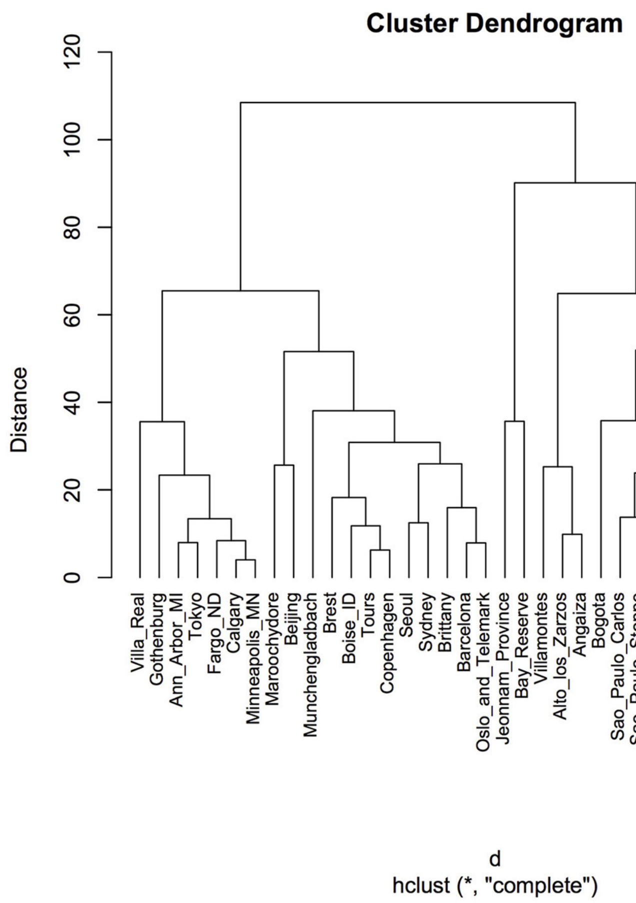

FIGURE 4 | Cluster dendrogram using Euclidean distance and hierarchical cluster analysis. Distances were calculated by using the function "dist" in the R package. A hierarchical cluster analysis was done using the set of dissimilarities and the function "hclust," with the "complete" linkage method. Using this method, when a cluster is formed, its distance to other objects is computed as the maximum distance between any object in the cluster and the other object. Iteratively, the algorithm joins the two most similar clusters until there is just a single cluster. At each iteration, distances between clusters are recomputed. "Distance" displayed in the y-axis is the value of the recomputed distance metric between clusters.

Clermont et al. (2000); (4) approximate geographic location (city or region) where the strains were obtained was recorded. In some cases, the same datasets were analyzed in two or more studies. In the present evaluation, we exerted care not to include the same datasets twice. We combined data of isolates from the same city or region (**Table 1** and **Figure 1**). We also used geographic coordinates and local climate to identify patterns in phylogroup distribution.

Visualization of Data Distribution and Correlation Analysis

Visualization of phylogroup distribution was achieved by means of a notched boxplot using the "*boxplot*" function from the "*graphics*" package in R (R Development Core Team, 2012a).

Next, we applied the "*pairs*" function from the same package with the following arguments: (1) "*upper.panel,*" to apply correlation analysis between variables; (2) "*diag.panel,*" to represent the frequency of distribution of the data using histograms; and (3) "*lower.panel,*" to smooth the data.

Analysis of Variance (ANOVA)

ANOVA was performed to test differences in phylogroup distribution with different response variables (city, country, Koppen climate classification, feeding habits, and living area). To this end, we used the "*aov*" function from the "*stats*" package in R (R Development Core Team, 2012b).

The *post-hoc* Tukey HSD test was used to identify significant differences between means using the function "*TukeyHSD*" from the "*stats*" package.

Clustering and Correspondence Analysis (CA)

The phylogroup distribution of the different sites was clustered hierarchically using R. We calculated the Euclidean distances by using the function "dist," and clustered the data points by using the function "htclust" testing two clustering methods (method = "complete," and method = "median"). Both functions, "dist" and "htclust," are part of the "*stats*" package. The dendrograms were plotted using the function "*plot*" from the "*graphics*" package.

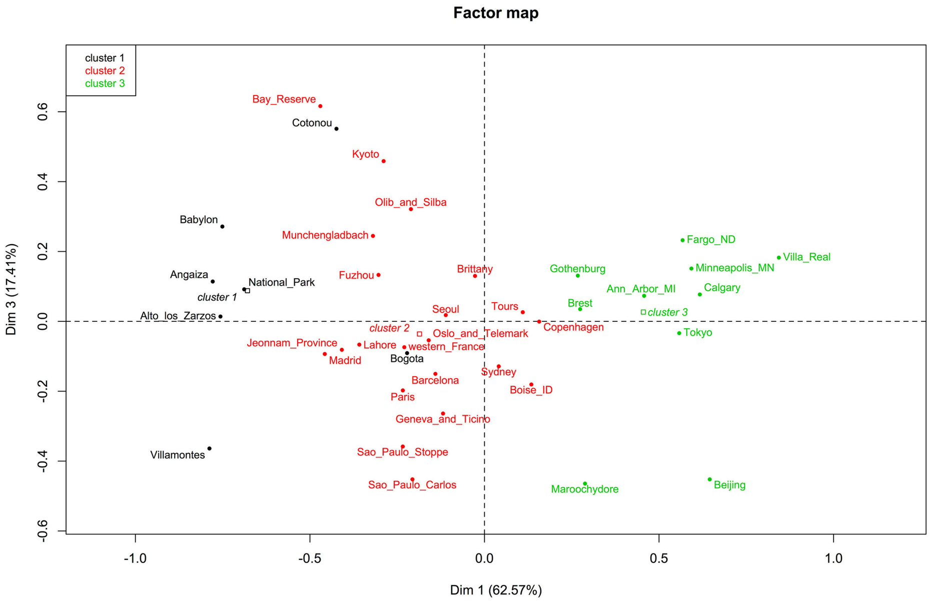

FIGURE 5 | Factor map showing three clusters.

To determine the existence of a correspondence between phylogroups and geographic location CA was performed using the packages "*ca*" (Nenadic and Greenacre, 2007), "*FactoMineR*" (Lê et al., 2008), and "*vcd*" (Meyer et al., 2014) in R. The agglomerative hierarchical clustering was done on the results of the factor analysis. To cluster the data and determine the optimal level of division, we used the method implemented in the R package "FactoMineR" (Lê et al., 2008). The function first built a hierarchical tree. Then the sum of the within-cluster inertia is calculated for each partition. The suggested partition is the one with the higher relative loss of inertia (Husson et al., 2010). The Ward criterion has to be used in the hierarchical clustering because it is based on the multidimensional variance (inertia). Locations (rows) were clustered using metric = "euclidean," and method = "ward," in the function HCPC. This allowed us to test whether phylogroups were could be used to group locations in meaningful clusters.

Probability Distribution Model: Poisson and Gamma Distributions

To calculate mean and variance for the Poisson test, we used the *ddply* function (Wickham, 2011) from the "*plyr*" package in R.

Given that the probability distribution model of phylogroup data was not known a priori, a more general probability model distribution, such as gamma, was used. To this end, we applied the parameters "*shape*" and "*rate,*" and the maximum-likelihood estimation (MLE) "*fitdistr*" function from the "*mass*" package in R (Ripley et al., 2013). Next, we generated a sample theoretical gamma distribution (cumulative distribution function, CDF) for the parameters "*shape*" and "*rate*" using the "*pgamma*" function from the "*stats*" package. Finally, to test the similarity between each phylogroup data distribution (empirical CDF) and a theoretical gamma distribution (gamma CDF), the Kolmogorov–Smirnov test (K–S test) was performed, using the "*ks.test*" function from the "*stats*" package.

Identification of Patterns Using Social Network Analysis *w-Clique* Metric

The *w-clique* metric was used to identify cohesive subgroups (clusters) in the network (Araújo et al., 2008). As described by Stoppe et al. (2014) and Silva et al. (2014), the *Dieta* program was first used to verify the nodes connected by strong interactions. These were encoded by a binary matrix (0/1), in which cells containing the number 1 represented interactions whose weights were higher than the average network weight (*w-cliques*; Araújo et al., 2008). Next, the *Pajek* program was used to transform the network from arcs to edges (Batagelj and Mrvar, 1998). The resulting matrix was analyzed using the *Ucinet* program (Borgatti et al., 2002) to identify the *w-cliques*.

Data Mining Classification

The "*decision tree*" algorithm (Witten et al., 2011) was used for data mining classification. To build the regression tree, we used the package "*rpart*" (Therneau et al., 2013) in R. To minimize the "*misclassification error*"—the percentage of data that the tree does not classify properly—we used different values for data training and testing.

Multivariate Analysis

Statistical analysis was based on two multivariate analysis methods: Mantel test (Oksanen, 2013) and clustering visualizations of multidimensional data (Hurley, 2004).

The former was used to analyze the correlation between two dissimilarity matrices, based on Pearson's product-moment correlation. The dissimilarity matrices were calculated with

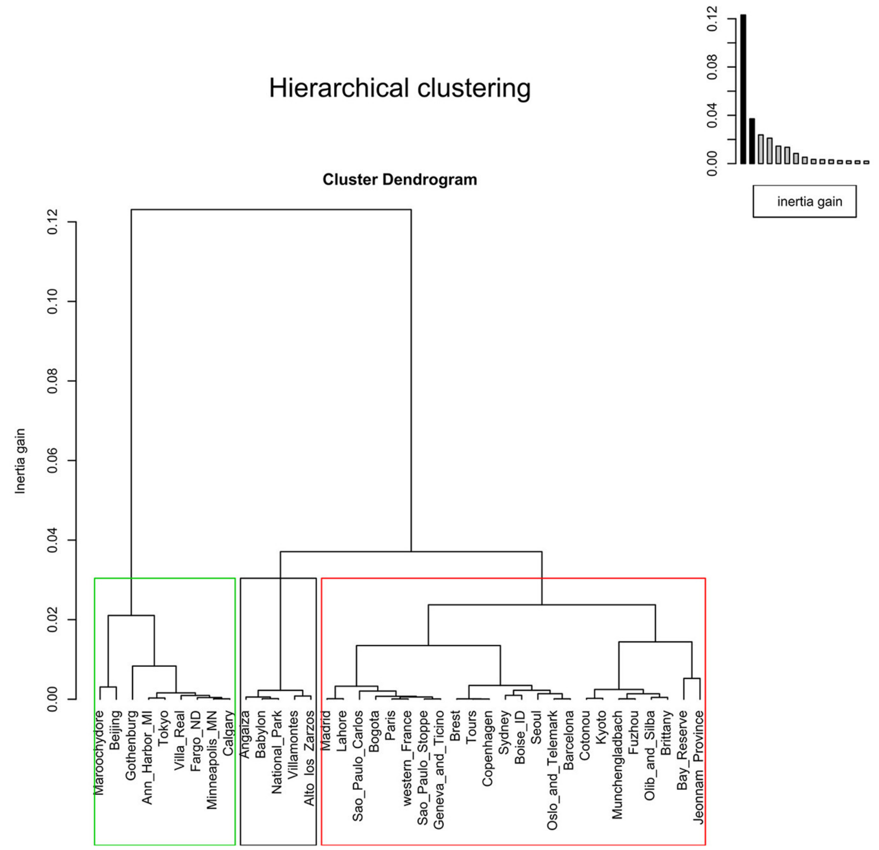

FIGURE 6 | Dendrogram showing three clusters with threshold of 0.03. The agglomerative hierarchical clustering was done on the results of the correspondence analysis performed using the function "ca" in R. To cluster the data and determine the optimal level of division, the method implemented in the R package FactoMineR was used. Each level at which the merging of clusters occurs corresponds to the associated reduction of the table's inertia. The height of the tree (y-axis) corresponds to the inertia gain. Locations were clustered using metric="euclidean," and method="ward," using the function "HCPC," and the optimal level of division was suggested as three clusters.

the "*vegdist*" function, using the community ecology package "*vegan*" in R (Oksanen, 2013), with the following coefficients as parameters: Bray–Curtis (data abundance) and Jaccard (data presence/absence; Legendre and Legendre, 2012).

For the latter, the dissimilarity matrices of the phylogroup distribution (raw data) were built using the Bray–Curtis coefficient (Legendre and Legendre, 2012). The similarity matrices were prepared as a complement to the dissimilarity matrices [1-vegdist (matrix, "Bray")].

The permutation of variables (hierarchical clustering order) was based on the classification of data, such as Koppen climate classification, feeding habits, living area, geographic distance, and collection date. This was presented graphically by the scatterplot matrix, which is commonly used for displaying multivariate data.

RESULTS

Phylogroup Distribution in Human and Wastewater Isolates

Briefly, 116 strains were isolated from different individuals (male and female adults with western feeding habits and average BMI of 26). Of these strains, 48.3% belonged to group A, 5.2% to group B1, 16.4% to group B2, and 30.2% to group D (**Table 1**). The phylogroup distribution did not differ significantly in terms of gender, age, or BMI.

Additional 150 strains were isolated from raw wastewater samples. Of these, 44% belonged to group A, 39% to group D, 9% to group B2 and 8% to group B1.

The phylogroup distribution was similar between wastewater and human samples; groups A and D were predominant and the phylogroup distribution was confirmed by a positive correlation (Mantel test, $r = 0.607$, $P = 0.046$).

Comparison of Worldwide Phylogroup Distribution in Humans and Wastewater

To obtain a more complete picture of the phylogroup distribution worldwide, we looked at published studies on WTP and human commensal isolates from Australia, Portugal, Spain, and the United States (Zhang et al., 2002; Sannes et al., 2004; Johnson et al., 2005; Machado et al., 2005; Boczek et al., 2006; Sabaté et al., 2008; Hannah et al., 2009; Moreno et al., 2009; Valverde et al., 2009; Anastasi et al., 2010; Bailey et al., 2010a,b; Figueira et al.,

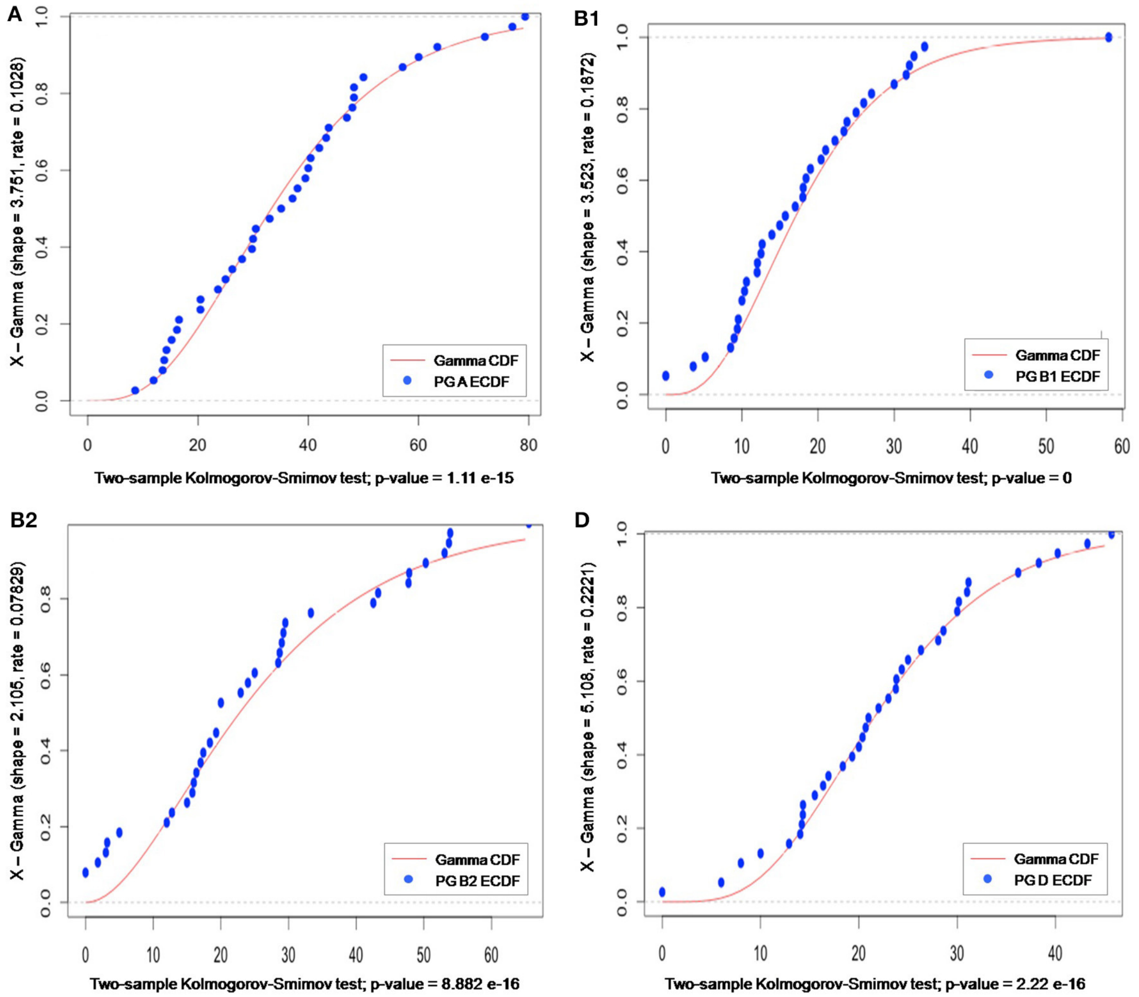

FIGURE 7 | Gamma distribution of meta-analysis phylogenetic data for groups A, B1, B2, and D. The gamma cumulative distribution function (CDF) is represented by a continuous red line and the phylogroup empirical CDF by blue dots. The "*shape*" and "*rate*" parameter values, derived from the gamma distribution for a specific sample space, are given on the y-axis label; and the *p-value* (generated by K–S test) is indicated on the x-axis label.

2011; Vollmerhausen et al., 2011; Logue et al., 2012; Silva et al., 2012). Contrary to our present findings, no significant correlation between phylogroup distribution in humans and wastewater was reported in Australia (Mantel test, $r = 0.003$, $P = 0.454$), Portugal (Mantel test, $r = -0.396$, $P = 0.600$), Spain (Mantel test, $r = -0.023$, $P = 0.575$), and the United States (Mantel test, $r = 0.070$, $P = 0.467$).

Meta-Analysis of Human Isolates

Next, we compared our distribution of *E. coli* phylogroups from commensal isolates with those from other worldwide studies.

Worldwide, group A was the most common type with a median of 36.1%. It was followed by D (median 21.5%), B2 (median 20%), and B1 (median 16.4%). The widest frequency range was observed for group A (from 8.7 to 79.3%), while B2 ranged from 0 to 53.9%, D from 0 to 45.7%, and B1 from 0 to 34%. Frequency differences across locations were highest for groups A and B1, and lowest for groups B2 and D (**Figure 2**).

A strong negative correlation was observed between groups A and B2 ($r = -0.73$), but not among other phylogroups (**Figure 3A**). When new variables (feeding habits, living area, or geographic location) were taken into account, the correlation among phylogroups did not change, suggesting that these other variables were independent (**Figure 3B**).

A comparison of phylogroup distributions between countries did not reveal any significant difference, except for groups A (ANOVA, $F = 2.83$, $P = 0.05$), B1 (ANOVA, $F = 2.39$, $P = 0.05$), and B2 (ANOVA, $F = 3.14$, $P = 0.05$). A Tukey HSD analysis of pairwise differences showed that Bolivia was significantly different from Portugal and the United States for both phylogroup A (respectively, $P = 0.04$ and $P = 0.01$) and phylogroup B2 (respectively, $P = 0.04$ and $P = 0.05$). For group B1, Mali was significantly different from Australia ($P = 0.03$), Bolivia ($P = 0.01$), Brazil ($P = 0.02$), China ($P = 0.03$), Colombia ($P = 0.03$), France ($P = 0.05$), and the

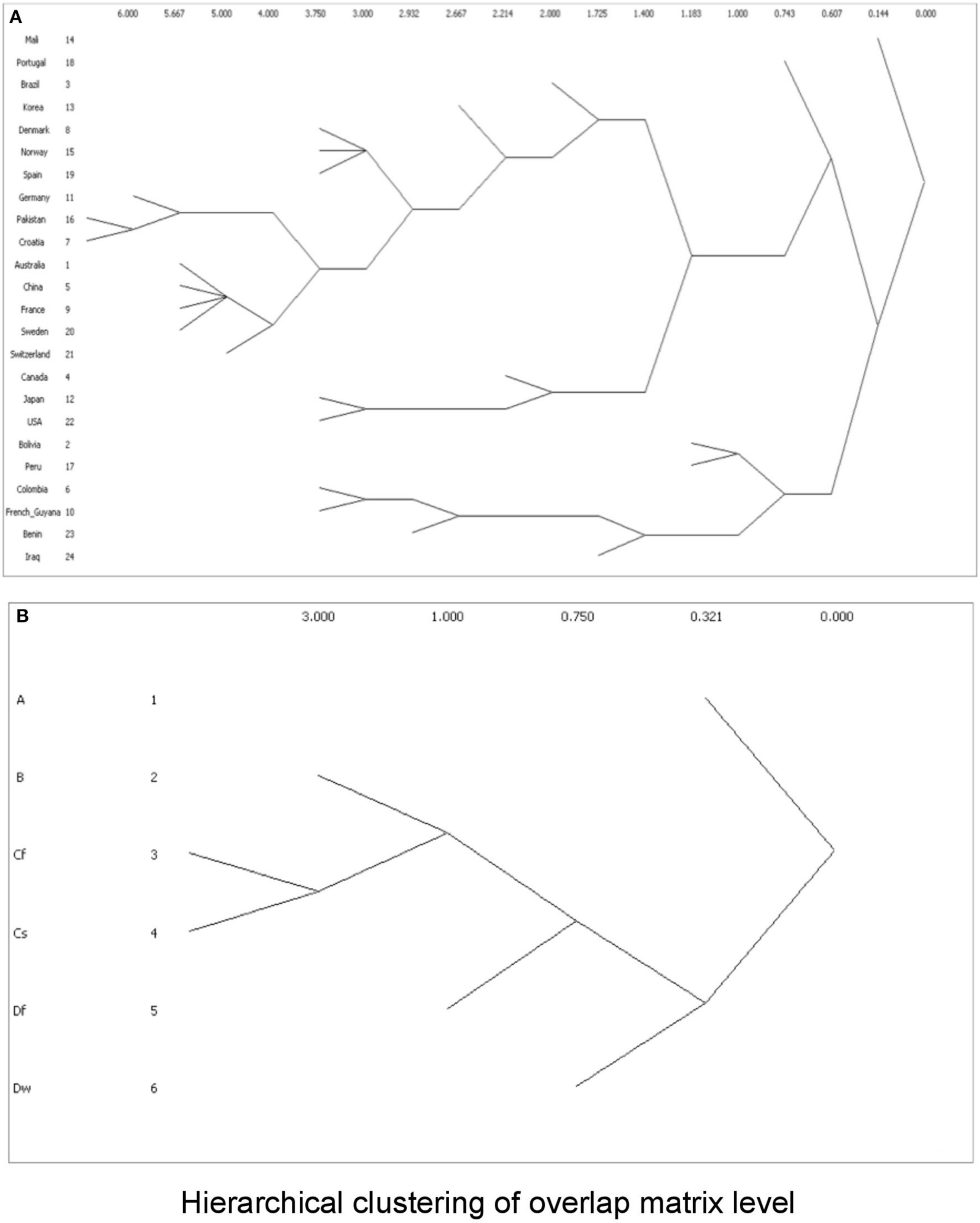

FIGURE 8 | Dendrograms showing clusters according to *w-clique* metric, based on **(A)** phylogroup distribution data and countries; or **(B)** Koppen climate classification.

United States ($P = 0.03$). Upon comparing the phylogroup distribution among the different climates (Koppen classification), only group A (ANOVA, $F = 0.009$, $P = 0.01$), tropical rain forest, and warm summer continental climates appeared significantly different ($P = 0.05$) from the other climates. However, following a less discriminative Koppen classification, with more countries for each climate type, the frequencies of both groups A and B2 differed significantly ($F = 0.002$, $P = 0.01$) in relation to climate. The phylogroup distribution in a tropical climate was significantly different from a temperate ($P = 0.02$) and a

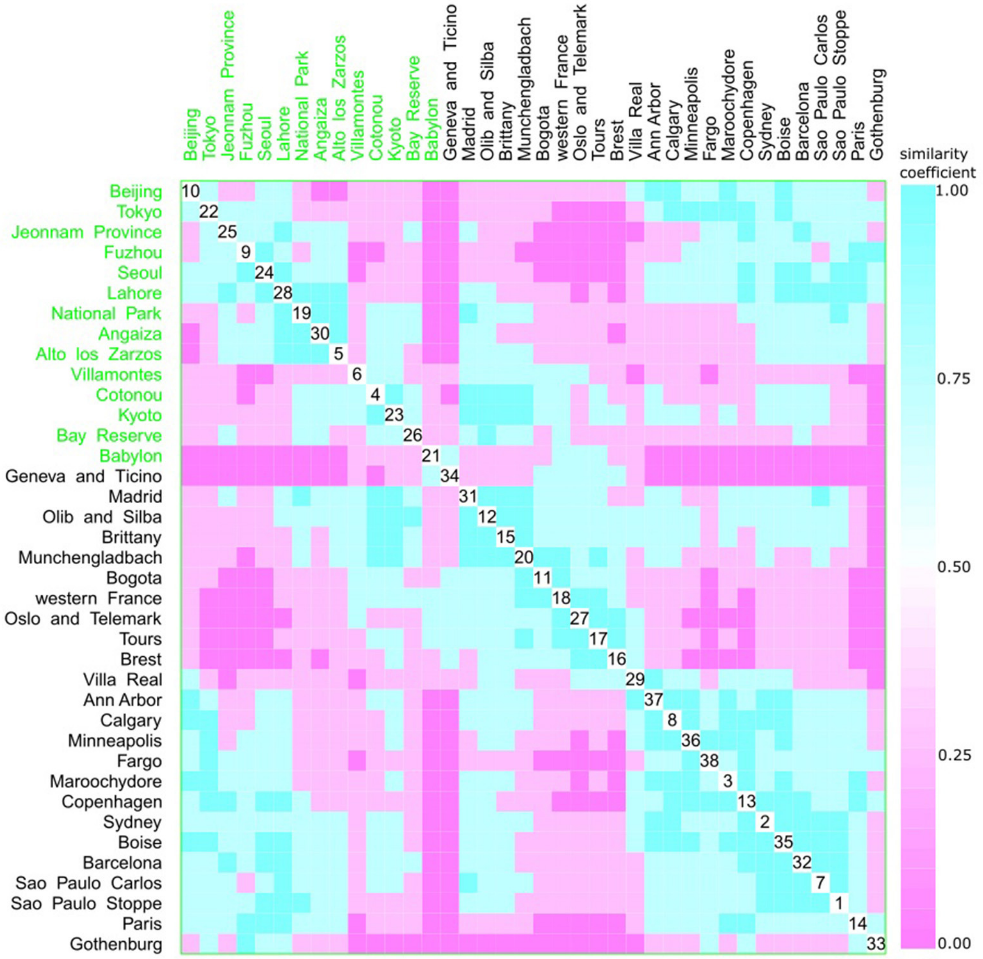

FIGURE 9 | Scatterplot matrix showing multivariate analysis between phylogenetic groups and feeding habits using clustering visualization of multidimensional data (Hurley, 2004). An identical color of the scatterplot clusters denotes a similar index value; thus, according to the Bray-Curtis coefficient, data have the same pattern ("*shape*" and "*rate*") of phylogroup distribution values. Components close to the scatterplot matrix diagonal consist of highly related variables. Three clusters can be seen, the first among cities, such as Beijing (10) and Alto de los Zarzos (5); the second is among Madrid (31) and Villa Real (29); and the last one among Villa Real (29) and Paris (33). Feeding habits: western with 24 locations, and 2529 isolates (black), and other diets with 14 locations, and 1587 isolates (green).

continental one ($P = 0.001$), irrespective of precipitation and season. Finally, the phylogroup distribution was comparable between the two different living areas, rural and urban, but differed significantly between groups A, B2, and D in terms of feeding habits, i.e., western or other diet.

Clustering according to Euclidean distance and the "complete method" of clustering showed two major groups that did not reflect geographic proximity, feeding habit, climate, or living area (**Figure 4**). The same pattern was observed when the method of clustering was "median" (data not shown).

A contingency table listing cities and phylogroups indicated a strong dependency, with a correlation coefficient of 0.526 (above the 0.2 threshold). Furthermore, the chi-square score was equal to 265.42 (degrees of freedom = 20), which was highly significant (well below alpha 0.01), suggesting geographic location might be a factor in phylogroup distribution.

The factor map showed three clusters. The first containing Alto de los Zarzos, Angaiza, Babylon, Bogota, Cotonou, National Park, and Villamontes. Here, no clear pattern could be observed, since these sites were geographically isolated, and differed in terms of both, climate and living area. The same could be said of the other two clusters (**Figure 5**). The dendrogram also showed three clusters but, as reported in the factor map, no pattern was observed (**Figure 6**).

As phylogroup data did not follow a discrete probability distribution, the use of a Poisson distribution was excluded. Instead, a gamma distribution was the probabilistic model that best fitted our data, as confirmed by calculating "*shape*" and "*rate*" parameters (using the MLE optimization algorithm) and running the K–S test (to identify the *p-value*). As gamma is a generalist continuous distribution model, it is typically used when the pattern of the data is not known (e.g., rainfall). In our case, the lower bound was zero, the upper one was not known, and no even

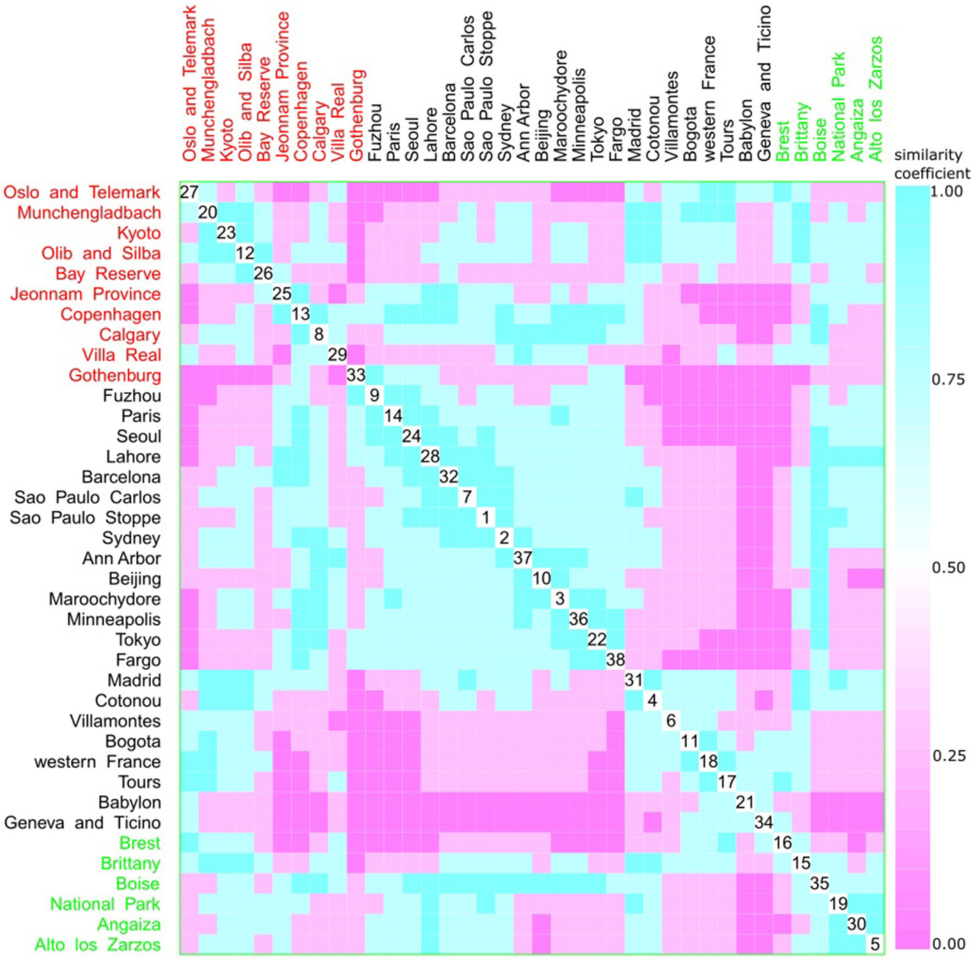

FIGURE 10 | Scatterplot matrix showing multivariate analysis between phylogenetic groups and living area using clustering visualization of multidimensional data. A cluster can be seen between Gothenburg (33) and Fargo (38). Living area: urban with 22 locations, and 2,448 isolates (black), rural with 6 locations, and 510 isolates (green), and not available with 10 locations, and 1,158 isolates (red).

distribution was observed around the mean, meaning that neither this model offered much information about the phylogroup data (**Figure 7**).

To cluster the phylogroup distribution according to geographic location and climate, we used *w-clique* metric. However, the clusters identified by the dendrogram did not reflect any geographic or climate classification (**Figure 8**). We also analyzed multiple factors simultaneously: geographic location and feeding habits; geographic location and living area; geographic location, feeding habits, and living area; climate and feeding habits; climate and living area; and climate, feeding habits, and living area. Even in this case, clusters failed to conform to any clear-cut subgroups (data not shown).

Data mining using a "*decision tree*" algorithm is a promising tool for simple classifications (Witten et al., 2011). Unfortunately, our data did not present a good level of confidence, voiding this option.

We also observed a weak correlation between geographic location and phylogroup distribution (Mantel test, $r = 0.2059$, $P = 0.007$), and moderate correlation between climate and phylogroup distribution (Mantel test, $r = 0.399$, $P = 0.001$).

A multivariate analysis using the multidimensional data method showed an unclear relationship between phylogroups and feeding habits, living area, Koppen climate groups, and geographic distance. Accordingly, three clusters could be observed between phylogroup and feeding habits, but each one consisted of both a western and other diet (**Figure 9**). The same was observed for living area, where one cluster contained cities from either urban or rural areas (**Figure 10**), for Koppen climate groups (**Figure 11**), geographic distance (**Figure 12**), and collection date (**Figure 13**).

Hurley (2004) proposed that the use of categorical variables with few levels could bias the results, resulting in inaccurate clusters. As feeding habits and living areas have only two

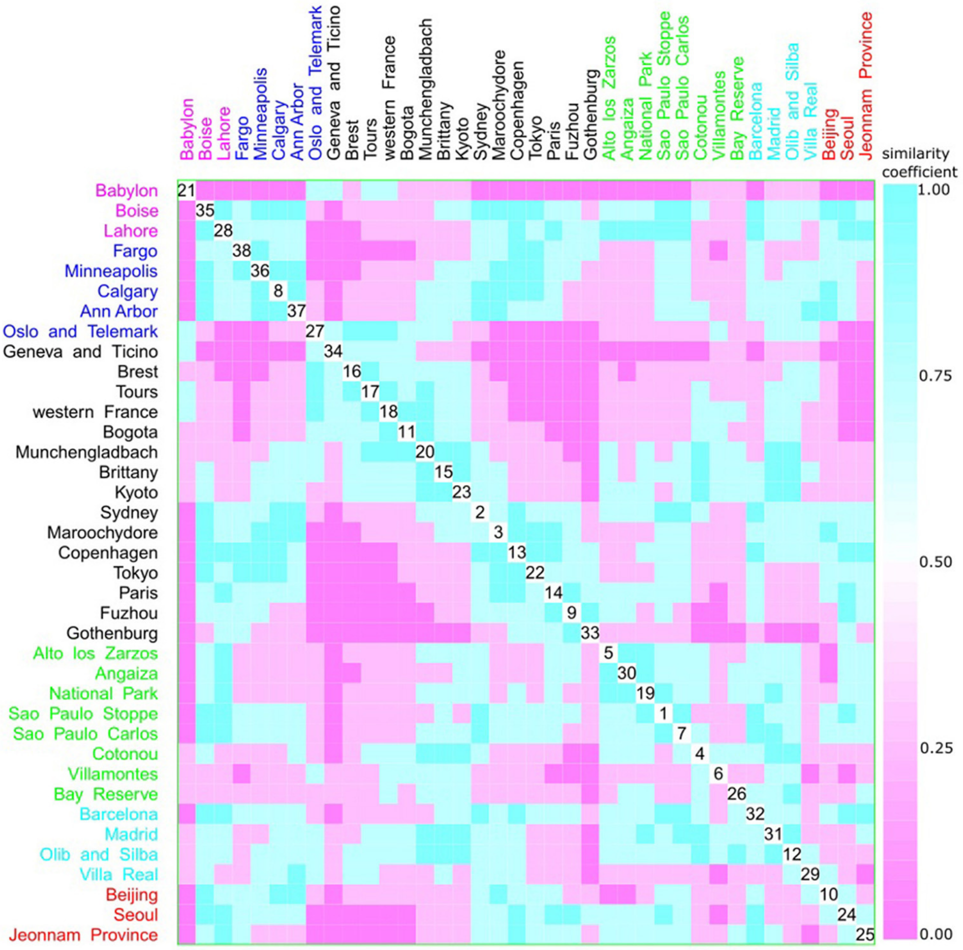

FIGURE 11 | Scatterplot matrix showing multivariate analysis between phylogenetic groups and Koppen climate classification using clustering visualization of multidimensional data. Two clusters can be seen, the first among Boise (35) and Ann Harbor (37); and the second among Oslo and Telemark (27) and Gothenburg (33). Koppen climate classification: warm temperate humid subtropical with 15 locations, and 1893 isolates (black), warm temperate Mediterranean with 4 locations, and 292 isolates (cyan), equatorial with 8 locations, and 657 isolates (green), arid desert and steppe with 3 locations, and 289 isolates (magenta), continental fully humid with 5 locations, and 567 isolates (blue), and continental hot summer with 3 locations, and 418 isolates (red).

classes each, western and other diets, and urban and rural areas respectively, this might explain the observed results (**Figures 9**, **10**). However, Koppen climate, and geographic distance comprise several categories and, despite that, only micro-clusters were observed (**Figures 11**, **12**), indicating that the human commensal *E. coli* phylogroup distribution did not correlate with either climate or geographic distance.

DISCUSSION

The methods described here were used independently for a number of other biological studies. For instance, triplex PCR (the Clermont method) was used to test whether phylogenetic group frequencies varied with the age and sex of hosts (Gordon et al., 2005). Johnson et al. (2005) used this technique to differentiate commensal and pathogenic *E. coli* strains. It was also used for the identification of the sources of fecal contamination (Carlos et al., 2010). For the statistical analysis, we used mainstream tools for exploratory data analysis. Some of the tests we applied are new or not widely used, but they also have examples of applications in the literature. For example, Tanner and Jackson (2012) and Stanley and Dunbar (2013) used Social Network Analysis (SNA) to study social organization in different species. We proposed a simple classification of polluted and unpolluted sites by using SNA metrics (Stoppe et al., 2014). Buttigieg and Ramette (2014) established a web-based resource to evaluate interaction between environmental factors and microorganisms in microbial ecology studies using multivariate analyses. To our knowledge, our manuscript is the first report of the use of the combined methods to test the worldwide distribution of *E. coli* subgroups. Previous research has shown that phylogroup distribution varied across locations. The observed differences were attributed to differences in climate, living area, and feeding habits (Escobar-Páramo et al.,

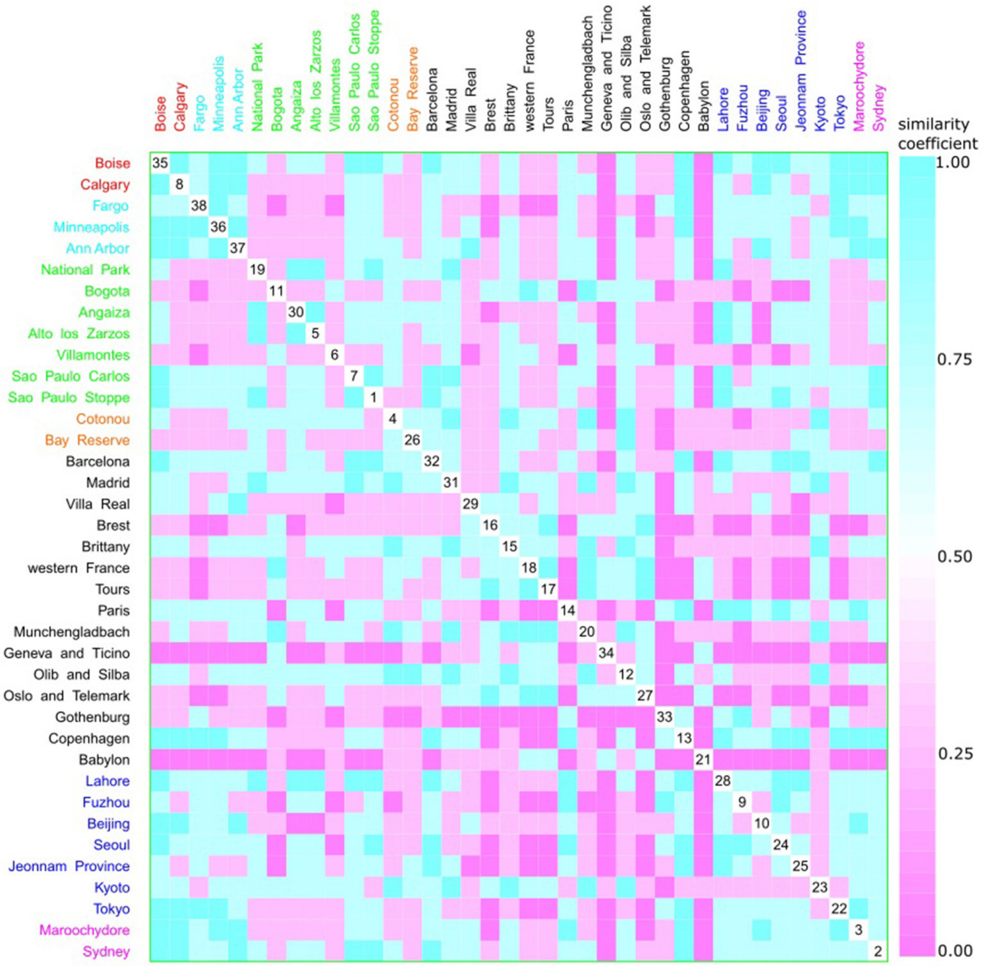

FIGURE 12 | Scatterplot matrix showing multivariate analysis between phylogenetic groups and geographic distance using clustering visualization of multidimensional data. Two clusters can be seen, the first among Boise (35) and Ann Harbor (37) and the second among Tours (17) and Villa Real (29). Geographic distance based on *k*-means clustering, in which each place belongs to the cluster with the nearest mean: cluster 1 with 2 locations, and 230 isolates (red), cluster 2 with 3 locations, and 439 isolates (cyan), cluster 3 with 7 locations, and 584 isolates (green), cluster 4 with 2 locations, and 101 isolates (orange), cluster 5 with 15 locations, and 1376 isolates (black), cluster 6 with 7 locations, and 1130 isolates (blue), and cluster 7 with 2 locations, and 256 isolates (magenta).

2004; Gordon et al., 2005; Nowrouzian et al., 2005, 2009; Karami, 2007; Pallechi et al., 2007; Vinue et al., 2008; Bartoloni et al., 2009; Hannah et al., 2009; Bailey et al., 2010b; Li et al., 2010; Nicolas-Chanoine et al., 2013). However, no formal statistical framework was used to test the effect of these parameters in phylogroup distribution. Here, we established a novel statistical workflow to test factors influencing worldwide distribution of *E. coli* subgroups, which may be of interest amongst other research groups with similar questions.

Commensal populations of *E. coli* comprise stable genetic isolates with very low rates of recombination, leading to a clonal population structure and allowing the characterization of major phylogenetic groups (Selander and Levin, 1980; Tenaillon et al., 2010). Hence, each host has only one phylogroup. The 116 *E. coli* strains isolated in this study provide a representative phylogroup distribution in this group of individuals.

The phylogroup frequency in human commensal *E. coli* strains was first described by Duriez et al. (2001), whereby the most common groups were A and B1, while B2 was the rarest. However, a later comparison of human commensal *E. coli* phylogenetic groups by Bailey et al. (2010a) coincided with the results from our study, indicating a predominance of group A and an underrepresentation of group B1.

Here, we observed a higher frequency of groups A and D in wastewater samples, whereas in previous studies either groups B2 or D (Anastasi et al., 2010; Mokracka et al., 2011), or groups A and B1 (Figueira et al., 2011) were predominant. This apparent discrepancy might imply that geographic location and climate play a role in determining phylogroup distribution, since environmental strains seem to be more susceptible to these factors. Some authors have suggested group A to be the best adapted to different environments (Skurnik et al., 2008;

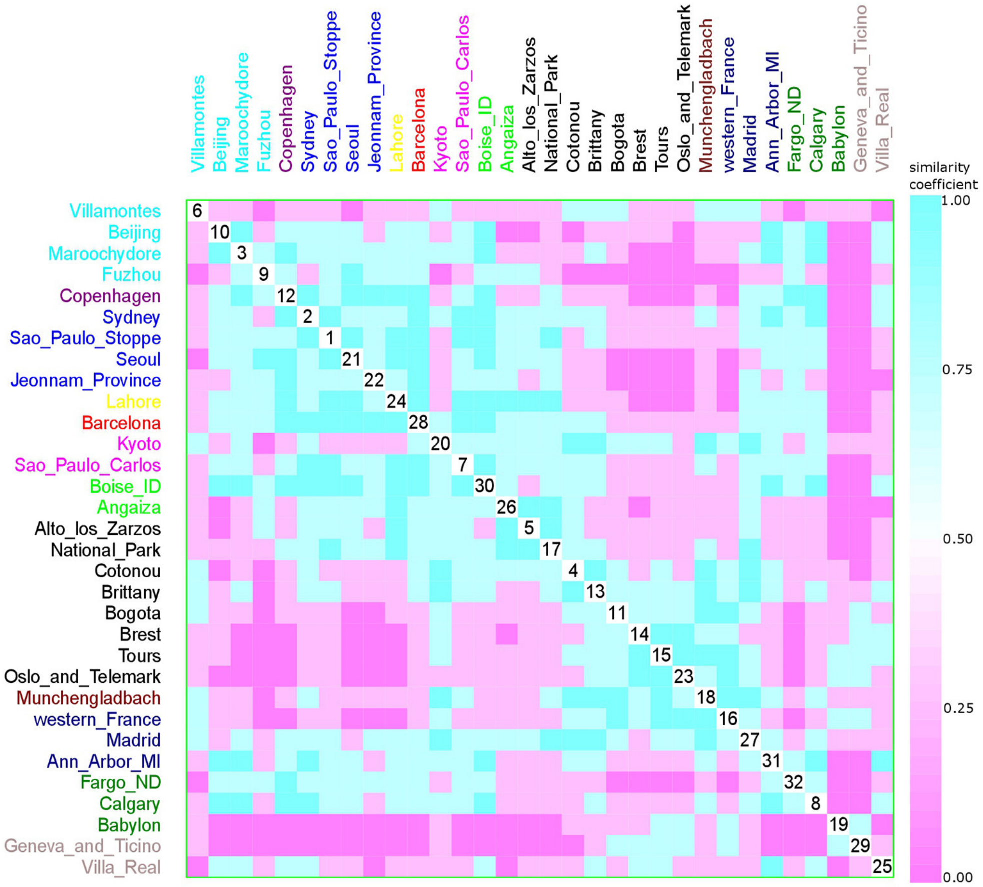

FIGURE 13 | Scatterplot matrix showing multivariate analysis between phylogenetic groups and collection date using clustering visualization of multidimensional data. Year of *E. coli* isolation: 1984 with 1 location, and 157 isolates (yellow), 1994 with 2 locations, and 144 isolates (magenta), 1999 with 8 locations, and 395 isolates (black), 2001 with 3 locations, and 170 isolates (dark blue), 2002 with 2 locations, and 233 isolates (green), 2003 with 1 location, and 120 isolates (red), 2004 with 1 location, and 171 isolates (violet), 2006 with 1 location, and 37 isolates (brown), 2007 with 2 locations, and 68 isolates (gray), 2008 with 4 locations, and 567 isolates (blue), 2009 with 4 locations, and 597 isolates (cyan), and 2010 with 3 locations, and 322 isolates (dark green). Samples that were collected across a large time frame were excluded from the analysis (Paris, Gothenburg, Olib e Silba, Bay Reserve, Tokyo, and Minneapolis).

Anastasi et al., 2012). The phylogroup distribution we observed in wastewater samples from Brazil follows those reported in Portugal (Figueira et al., 2011) and Spain (Sabaté et al., 2008), but not in Australia (Anastasi et al., 2010) or the United States (Boczek et al., 2006).

Bacterial isolates from sewage samples can be used in lieu of human feces samples (USEPA, 2005). Our results showed a positive correlation between phylogroup distributions in human and wastewater samples, reinforcing the use of isolates from urban wastewater treatment plants as surrogates for human fecal contamination. Conversely, *E. coli* commensal isolates collected in Australia, Portugal, Spain, and the United States did not show significant correlation with phylogroup distribution in isolates from WTPs samples. Contrary to these other studies, our feces and wastewater samples were collected from the same area (São Paulo Metropolitan region) and at the same time. Thus, it is possible that phylogroup frequencies fluctuate over time limiting the use of sewage samples as proxies for human ones.

Duriez et al. (2001) proposed that studies of *E. coli* commensal strains should include geographic, socioeconomic, and medical information since these enteric isolates were likely reservoirs of pathogenic strains. To this end, we compiled data from the literature that used the Clermont classification (Clermont et al., 2000), to find out if any phylogroup pattern might be related to climate, feeding habits, living area, and/or geographic location. Several studies compared Clermont phylogroups with climate (Escobar-Páramo et al., 2004; Gordon et al., 2005; Li et al., 2010), living area (Nowrouzian et al., 2005, 2009; Karami, 2007; Pallechi et al., 2007; Vinue et al., 2008; Bartoloni et al., 2009; Hannah et al., 2009; Bailey et al., 2010b; Nicolas-Chanoine et al., 2013), and feeding habits (Gordon et al., 2005), however only qualitative differences had been reported and none of the studies had formally tested the observed differences.

Here, we compared 38 samples from 24 countries and observed that groups A and B1 did not present overlapping frequencies, as opposed to groups B2 and D (**Figure 2**).

Moreover, we report a negative correlation between phylogroup A and B2 (**Figure 3**). Gordon et al. (2008) studied *E. coli* strains from different hosts (humans, other mammals, and birds) and environmental samples (water, soil, and sediments) from Europe, Africa, America, and Oceania and demonstrated the validity of Clermont's method for population studies due to its rapidity, low cost, and reliability for assigning *E. coli* strains to phylogenetic groups.

The differences between phylogroup frequencies on a worldwide scale did not correlate with geographical distance, not even in the case of Mali, which presented significant differences compared to several countries on other continents. Random location clustering was also observed in the dendrogram generated by Euclidean distance, corroborating the fact that phylogroup distribution did not correlate with geographic distance. Correspondence analysis resulted in three clusters, however, no clear pattern could be identified in terms of geographic location, climate, feeding habits, or living area. Based on the successful use of the *w-clique* metric by Stoppe et al. (2014) for differentiating between polluted and unpolluted sites in river samples, we used the same principle to cluster phylogroup distributions. Once again, the obtained clusters showed no grouping by geographic proximity or climate.

Data mining results showed weak tree classification. This might occur due to the algorithm used in constructing the tree, which automatically selected only part of the data (some phylogroups). Hence, a misclassification error higher than expected may be generated when a phylogroup that is not representative of the entire sample space is selected.

The Mantel test is widely used in ecological studies to correlate genetic markers and geographic distance (Bellay et al., 2011; Castillo-Rojas et al., 2013; Winter et al., 2013). Notwithstanding, our meta-analysis data confirmed only a weak correlation between geographic location and phylogroup distribution, suggesting that the distance between sites might not influence phylogroup distribution. Some authors proposed that the human hosts' climate influenced the *E. coli* phylogroup distribution (Escobar-Páramo et al., 2004; Gordon et al., 2005; Li et al., 2010). Our analysis showed a moderate correlation between climate and phylogroup distribution, suggesting this factor can have more influence on phylogroup distribution than geographic distance.

A meta-analysis comparison using clustering multivariate analysis failed to reveal any substructuring according to the parameters evaluated, even in the cases of geographic distance and climate, which have more levels of classification (**Figures 9–12**). In terms of feeding habits (**Figure 9**), it was possible to observe three different clusters; however, each cluster comprised either western habits or other diets. In the case of living area, only one cluster was observed that contained only urban areas, although not all of them (**Figure 10**).

The similarity matrix according to Koppen climate classification revealed two clusters. The first one (identifier 35–37) encompassed only arid (B) and continental (D) climates, yet, some sites with these climate types were not included in the cluster. The other cluster included cities with different climates and no obvious pattern (**Table 1** and **Figure 11**).

When looking for patterns of geographic proximity, we observed two clusters (**Figure 12**). The first one included cities located in North America, while the second included only cities in Europe. These results suggest that geographic location could influence phylogroup distribution. The lack of additional clusters indicates that other factors might also be at play.

Only a few studies on microbial biogeography of commensal microorganisms and their hosts have been published to date. According to Tenaillon et al. (2010) socioeconomic factors and hygiene are important determinants of phylogroup distribution, as may be diet. In developed countries, the consumption of industrialized food has increased over the past decades, whereas in developing countries the consumption of carbohydrates and fresh food are more common. This may account for the shift from A as the predominant group in developed countries and in France in 1980, to B2 being the main group in France in 2000. Nevertheless, other factors should be taken into account when determining the phylogroup distribution. In Asia, where high levels of carbohydrates are consumed, phylogroup prevalence has shifted from A (Kyoto, Japan and Seoul, Korea) (Kanamaru et al., 2006; Lee et al., 2010) to B2 (Beijing, China and Tokyo, Japan) or D (Jeonnam Province, Korea) (Unno et al., 2009). These observations suggest that neither socioeconomic factors nor diet alone is sufficient to determine phylogroup distribution patterns. Among the other variables taken in consideration in this study, none significantly affected phylogroup distribution.

Our spatial autocorrelation analysis did not reveal any specific phylogroup patterns. There is evidence of phylogroup distribution substructuring, but geographic distance does not limit *E. coli* dispersion. For free-living microorganisms, dispersion over large distances is extremely rare, but for commensal microbes it can be much easier because they can travel with their hosts. The relationship between phylogroup and their hosts is not fully understood and it must involve many other factors, besides host climate living area or feeding habits. Further research is warranted as the existence of patterns relating phylogroup distribution with host-associated factors could facilitate prediction of which *E. coli* strain is prevalent in different climate or geographic locations. Accordingly, it might act as a proxy for laboratory analysis and simplify the identification of potential pathogenic strains.

AUTHOR CONTRIBUTIONS

NS, LO, MS, and TT: Conceived and designed the work; NS and CC: Performed the experiments; NS, JS, and TT: Analyzed the data, prepare the figures, and wrote the manuscript; MS, AS, LO, and TT: Ensured the financial and material resources; NS, JS, CC, MS, AS, LO, and TT: Reviewed the final draft.

ACKNOWLEDGMENTS

We thank the São Paulo State Research Foundation (FAPESP), and the Coordination for the Improvement of Higher Education Personnel (CAPES) for supporting this work.

NS received a fellowship from FAPESP (project no. 2010/15222-0). JS had financial support from IFMT, CAPES, and FAPEMAT. TT received a research fellowship from the National Council for Technological and Scientific Development (CNPq). The Research Center on Biodiversity and Computing (BioComp) was supported by the USP Provost's Office for Research. We thank Raul Jimenez Rosenberg from CONABIO, Mexico for helping us with our data analysis, and Elayse Maria Hachich from CETESB for allowing the use of Microbiological Laboratory Facilities. The authors wish to thank the reviewers for their constructive comments that helped us to improve our analyses and the manuscript.

REFERENCES

Abdul-Razzaq, M. S., and Abdul-Lateef, L. A. (2011). Molecular phylogeny of *Escherichia coli* isolated from clinical samples in Hilla, Iraq. *Afr. J. Biotechnol.* 10, 15783–15787. doi: 10.5897/AJB11.1273

Ahmed, W., Goonetileke, A., Powell, D., Chauhan, K., and Gardner, T. (2009). Comparison of molecular markers to detect fresh sewage in environmental waters. *Water Res.* 43, 4908–4917. doi: 10.1016/j.watres.2009.09.047

Anastasi, E. M., Matthews, B., Gundogdu, A., Vollmerhausen, T. L., Ramos, N. L., Stratton, H., et al. (2010). Prevalence and persistence of *Escherichia coli* strains with uropathogenic virulence characteristics in sewage tratment plants. *Appl. Environ. Microbiol.* 76, 5788–5786. doi: 10.1128/AEM.00141-10

Anastasi, E. M., Matthews, B., Stratton, H. M., and Katouli, M. (2012). Pathogenic *Escherichia coli* found in sewage treatment plants and environmental waters. *Appl. Environ. Microbiol.* 78, 5536–5541. doi: 10.1128/AEM.00657-12

APHA (2010). "Samples," in *Standard Methods for Examination of Water and Wastewater*, Section 9060. American Public Health Association. Available online at: http://standardmethods.org (Accessed February 18, 2010).

Araújo, M. S., Guimarães, P. R., Svanbäck, R., Pinheiro, A., Guimarães, P., Reis, S. F., et al. (2008). Network analysis reveals contrasting effects of intraspecific competition on individual vs. population diets. *Ecology* 89, 1981–1993. doi: 10.1890/07-0630.1

ATCC (American Type Culture Collection). (2017). *Bacterial Culture Guide - Tips and Techniques for Culturing Bacteria and Bacteriohages.* Available online at: https://www.atcc.org/~/media/PDFs/Culture%20Guides/ATCC_Bacterial_Culture_Guide.ashx (Accessed December 8, 2017).

Bailey, J. K., Pinyon, J. L., Annantham, S., and Hall, R. M. (2010a). Distribution of human commensal *Escherichia coli* phylogenetic groups. *J. Clin. Microbiol.* 48, 3455–3456. doi: 10.1128/JCM.00760-10

Bailey, J. K., Pinyon, J. L., Annantham, S., and Hall, R. M. (2010b). Commensal *Escherichia coli* of healthy humans: a reservoir for antibiotic-resistance determinants. *J. Med. Microbiol.* 59, 1331–1339. doi: 10.1099/jmm.0.022475-0

Bartoloni, A., Palecchi, L., Rodríguez, L., Fernandez, C., Mantella, A., Bartalesi, F., et al. (2009). Antibiotic resistance in a remote Amazonas community. *Int. J. Antimicrob. Agents* 33, 125–129. doi: 10.1016/j.ijantimicag.2008.07.029

Batagelj, V., and Mrvar, A. (1998). Pajek – program for large network analysis. *Connections* 21, 47–57.

Bellay, S., Lima Junior, D. P., Takemoto, R. M., and Luque, J. L. (2011). A host-endoparasite network of Neotropical marine fish: are there organizational patterns? *Parasitology* 138, 1945–1952. doi: 10.1017/S0031182011001314

Boczek, L. A., Johnson, C. H., Rice, E. W., and Kinkle, B. K. (2006). The widespread occurrence of the enterohemolysin gene ehlyA among environmental strains of *Escherichia coli. FEMS Microbiol. Lett.* 254, 281–284. doi: 10.1111/j.1574-6968.2005.00035.x

Borgatti, S. P., Everett, M. G., and Freeman, L. C. (2002). *Ucinet for Windows: Software for Social Network Analysis.* Harvard, MA: Analytic Technologies.

Buttigieg, P. L., and Ramette, A. (2014). A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses. *FEMS Microbiol. Ecol.* 90, 543–550. doi: 10.1111/1574-6941.12437

Carlos, C., Aires, M. M., Stoppe, N. C., Hachich, E. M., Sato, M. I. Z., Gomes, T. A. T., et al. (2010). *Escherichia coli* phylogenetic group determination and its application in the identification of the major animal source of fecal contamination. *BMC Microbiol.* 10:161. doi: 10.1186/1471-2180-10-161

Castillo-Rojas, G., Mazari-Hiriart, M., Leon, S. P., Amieva-Fernandez, R. I., Agis-Juarez, R. A., Huebner, J., et al. (2013). Comparison of *Enterococcus faecium* and *Enterococcus faecalis* strains isolated from water and clinical samples: antimicrobial susceptibility and genetic relationships. *PLosONE* 8:e59491. doi: 10.1371/journal.pone.0059491

Caugant, D. A., Levin, B. R., Orskov, F., Svanborg Eden, C., and Selander, R. K. (1985). Genetic diversity in relation to serotype in *Escherichia coli*. *Infect. Immun.* 49, 407–413.

Clermont, O., Bonacorsi, S., and Bingen, E. (2000). Rapid and simple determination of the *Escherichia coli* phylogenetic group. *Appl. Environ. Microbiol.* 66, 4555–4558. doi: 10.1128/AEM.66.10.4555-4558.2000

Damborg, P., Nielsen, S. S., and Guardabassi, L. (2009). *Escherichia coli* shedding patterns in humans and dogs: insights into within-household transmission of phylotypes associated with urinary tract infections. *Epidemiol. Infect.* 137, 1457–1464. doi: 10.1017/S095026880900226X

de Castro Stoppe, N., Silva, J. S., Torres, T. T., Carlos, C., Hachich, E. M., Sato, M. I. Z., et al. (2014). Clustering of water bodies in unpolluted and polluted environments based on *Escherichia coli* phylogroup abundance using a simple interaction database. *Genet. Mol. Biol.* 37, 694–714. doi: 10.1590/S1415-47572014005000016

Desjardins, P., Picard, B., Kaltenbock, B., Ellion, J., and Denamur, E. (1995). Sex in *Escherichia coli* does not disrupt the clonal structure of the population: evidence from random amplified polymorphic DNA and restriction-fragment-length polymorphism. *J. Mol. Evol.* 41, 440–448. doi: 10.1007/BF00160315

Duran, M., Haznedaroglu, B. Z., and Zitomer, D. H. (2006). Microbial source tracking using host specific FAME profiles of fecal coliforms. *Water Res.* 40, 67–74. doi: 10.1016/j.watres.2005.10.019

Duriez, P., Clermont, O., Bonacorsi, S., Bingen, E., Chaventré, A., Elion, J., et al. (2001). Commensal *Escherichia coli* isolates are phylogenetically distributed among geographically distinct human populations. *Microbiology* 147, 1671–1676. doi: 10.1099/00221287-147-6-1671

Escobar-Páramo, P., Grenet, K., Menac'h, A. L., Rode, L., Salgado, E., Amorin, C., et al. (2004). Large-scale population structure of human commensal *Escherichia coli* isolates. *Appl. Environ. Microbiol.* 70, 5698–5700. doi: 10.1128/AEM.70.9.5698-5700.2004

Escobar-Páramo, P., Le Menac'h, A., LeGall, T., Amorin, C., Gouriou, S., Picard, B., et al. (2006). Identification of forces shaping the commensal *Escherichia coli* genetic structure by comparing animal and human isolates. *Environ. Microbiol.* 8, 1975–1984. doi: 10.1111/j.1462-2920.2006.01077.x

Figueira, V., Serra, E., and Manaia, C. M. (2011). Differential patterns of antimicrobial restance in population subsets of *Escherichia coli* isolated from waste- and surface waters. *Sci. Total Environ.* 409, 1017–1023. doi: 10.1016/j.scitotenv.2010.12.011

Fremaux, B., Gritzfeld, J., and Yost, C. K. (2009). Evaluation of host-specific Bacteroidales 16S rRNA gene markers as a complementary tool for detecting fecal pollution in a prairie watershed. *Water Res.* 43, 4838–4849. doi: 10.1016/j.watres.2009.06.045

Gordon, D. M., and Cowling A. (2003). The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects. *Microbiology* 149, 3575–3586. doi: 10.1099/mic.0.26486-0

Gordon, D. M., Clermont, O., Tolley, H., and Denamur, E. (2008). Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. *Environ. Microbiol.* 10, 2484–2496. doi: 10.1111/j.1462-2920.2008.01669.x

Gordon, D. M., Stern, S. E., and Collignon, P. J. (2005). Influence on the age and sex of human hosts on the distribution of *Escherichia coli* ECOR groups and virulence traits. *Microbiology* 151, 15–23. doi: 10.1099/mic.0.27425-0

Grasselli, E., François, P., Gutacker, M., Gettler, E., Benagli, C., Convert, M., et al. (2009). Evidence of horizontal gene transfer between human and animal isolates *Escherichia coli* strains identified by microarray. *FEMS Immunol. Med. Microbiol.* 53, 351–358. doi: 10.1111/j.1574-695X.2008.00434.x

Grude, N., Potaturkina-Nesterova, N. I., Jenkins, A., Strand, L., Nowrouzian, F. L., Nyhus, J., et al. (2007). A comparison of phylogenetic group, virulence factors and antibiotic resistance in Russian and Norwegian isolates of

Escherichia coli from urinary tract infection. *Clin. Microbiol. Infect.* 13, 208–211. doi: 10.1111/j.1469-0691.2006.01584.x

Hannah, E. L., Johnson, J. R., Angulo, F., Haddadin, B., Williamson, J., and Samore, M. H. (2009). Molecular analysis of antimicrobial-susceptible and -resistant *Escherichia coli* from retail meats and human stool and clinical specimens in a rural community setting. *Foodborne Pathog. Dis.* 6, 285–295. doi: 10.1089/fpd.2008.0176

Harada, K., Okada, E., Shimizu, T., Kataoka, Y., Sawada, T., and Takahashi, T. (2012). Antimicrobial resistance, virulence profiles, phylogenetic groups of Escherichia coli isolates: a comparative analysis between dogs and their owners in Japan. *Comp. Immunol. Microbiol. Infect. Dis.* 35, 139–144. doi: 10.1016/j.cimid.2011.12.005

Higgins, J., Hohn, C., Hornor, S., Frana, M., Denver, M., and Joerger, R. (2007). Genotyping of *Escherichia coli* from environmental and animal samples. *J. Microbiol. Methods* 70, 227–235. doi: 10.1016/j.mimet.2007.04.009

Hurley, C. B. (2004). Clustering visualizations of multidimensional data. *J. Comp. Graph. Stat.* 13, 788–806. doi: 10.1198/106186004X12425

Husson, F., Josse, J., and Pagès, J. (2010). *Principal Component Methods-Hierarchical Clustering-Partitional Clustering: Why Would we Need to Choose for Visualizing Data*? Technical Report, Agrocampus.

Ishii, S., Meyer, K. P., and Sadowsky, M. J. (2007). Relationship between phylogenetic groups, genotypic clusters, and virulence factors of *Escherichia coli* strains form diverse human and animal sources. *Appl. Environ. Microbiol.* 73, 5703–5710. doi: 10.1128/AEM.00275-07

Jakobsen, L., Kurbasic, A., Skjøt-Rasmussen, L., Ejrnaes, K., Porsbo, L. J., Pedersen, K., Jensen, L. B., et al. (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. doi: 10.1089/fpd.2009.0409

Jiang, S. C., Chu, W., Olson, B. H., He, J. W., Choi, S., Zhang, J., et al. (2007). Microbial source tracking in a small southern California urban watershed indicates wild animals and growth as the source of fecal bacteria. *Appl. Microbiol. Biotechnol.* 76, 927–934. doi: 10.1007/s00253-007-1047-0

Johnson, J. R., Owens, K., Gajewski, A., and Kuskowski, M. A. (2005). Bacterial characteristics in relation to clinical sources of *Escherichia coli* isolates from women with acute cystitis or pyelonefritis and uninfected women. *J. Clin. Microbiol.* 43, 6064–6072. doi: 10.1128/JCM.43.12.6064-6072.2005

Kanamaru, S., Kurazono, H., Nakano, M., Terai, A., Ogawa, O., and Yamamoto, S. (2006). Subtyping of uropathogenic *Escherichia coli* according to the pathogenicity island encoding uropathogenic-specific protein: comparison with phylogenetic groups. *Int. J. Urol.* 13, 754–760. doi: 10.1111/j.1442-2042.2006.01398.x

Kaneene, J., Miller, R., Sayah, R., Johnson, Y. J., Gilliland, D., and Gardiner, J. C. (2007). Considerations when using discriminant function analysis of antimicrobial resistance profiles to identify sources of fecal contamination of surface water in Michigan. *Appl. Environ. Microbiol.* 73, 2878–2890. doi: 10.1128/AEM.02376-06

Karami, N. (2007). *Antibiotic resistance and fitness of Escherichia coli in the Infantile Commensal Microbiota*. Ph.D. thesis, University of Gothenburg. Available online at: https://gupea.ub.gu.se/handle/2077/4418

Kelty, C. A., Varma, M., Sivaganesan, M., Haugland, R. A., and Shanks, O. C. (2012). Distribution of genetic marker concentrations for fecal indicator bacteria in sewage and animal feces. *Appl. Environ. Microbiol.* 78, 4225–4232. doi: 10.1128/A. E. M.07819-11

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR, An R package for multivariate analysis. *J. Stat. Softw.* 25, 1–18. doi: 10.18637/jss.v025.i01

Lee, S., Yu, J. K., Park, K., Oh, E. J., Kim, S. Y., and Park, Y. J. (2010). Phylogenetic groups and virulence factors in pathogenic and commensal strains of *Escherichia coli* and their association with blaCTX-M. *Ann. Clin. Lab. Sci.* 40, 361–367.

Leflon-Guibout, V., Blanco, J., Amaqdouf, K., Mora, A., Guize, L., and Nicolas-Chanoine, M.-H. (2008). Absence of CTX-M enzymes but high prevalence of clones, including clone ST131, among fecal *Escherichia coli* isolates from healthy subjects living in the area of Paris, France. *J. Clin. Microbiol.* 46, 3900–3905. doi: 10.1128/JCM.00734-08

Legendre, P., and Legendre, L. (2012). *Numerical Ecology, 3rd Edn.* Oxford: Elsevier.

Li, B., Sun, J. Y., Han, L. Z., Huang, X. H., Fu, Q., and Ni, Y. X. (2010). Phylogenetic groups and pathogenicity islands markers in fecal *Escherichia coli* isolates from asymptomatic humans in China. *Appl. Environ. Microbiol.* 76, 6698–6700. doi: 10.1128/AEM.00707-10

Logue, C. M., Doetkott, C., Mangiamele, P., Wannemuehler, Y. M., Johnson, T. J., Tivendale, K. A., et al. (2012). Genotypic and phenotypic traits that distinguish neonatal-meningitis-associated *Escherichia coli* from fecal *E.coli* isolates from healthy human hosts. *Appl. Environ. Microbiol.* 78, 5824–5830. doi: 10.1128/AEM.07869-11

Luo, Y., Cui, S., Li, J., Yang, J., Lin,. L., Hu, C., et al. (2011). Characterization of *Escherichia coli* isolates from healthy food handlers in hospital. *Microb. Drug Resist.* 17, 443–448. doi: 10.1089/mdr.2011.0032

Machado, E., Cantón, R., Baquero, F., Galán, J. C., Rollán, A., Peixe, L., et al. (2005). Integron content of extended-spectrum-beta-lactamase-producing *Escherichia coli* strains over 12 years in a single hospital in Madrid, Spain. *Antimicrob. Agents Chemother.* 49, 1823–1829. doi: 10.1128/AAC.49.5.1823-1829.2005

Mereghetti, L., Tayoro, J., Watt, S., Lanotte, P., Loulergue, J., Perrotin, D., et al. (2002). Genetic relationship between *Escherichia coli* strains isolated from the intestinal flora and those responsible for infectious diseases among patients hospitalized in intensive care units. *J. Hosp. Infect.* 52, 43–51. doi: 10.1053/jhin.2002.1259

Meyer, D., Zeileis, A., and Hornik, K. (2014). *vcd: Visualizing Categorical Data, R Package Version 1.3-2.* 03 Aug. 2014. Available online at: http://cran.r-project.org/ (Accessed August 13, 2014).

Miller, R. D., and Hartl, D. L. (1986). Biotyping confirms a nearly clonal population structure in *Escherichia coli*. *Evolution* 40, 1–12. doi: 10.1111/j.1558-5646.1986.tb05712.x

Mokracka, J., Kokzura, R., Jabłonska, L., and Kaznowski, A. (2011). Phylogenetic groups, virulence genes and quinolone resistance of integron-bearing *Escherichia coli* strains isolated from a wastewater treatment plant. *Antonie Van Leeuwenhoek* 99, 817–824. doi: 10.1007/s10482-011-9555-4

Moreno, E., Johnson, J. R., Pérez, T., Pratts, G., Kuskowski, M. A., and Andreu, A. (2009). Structure and urovirulence of the fecal *Escherichia coli* population among healthy women. *Microbes Infect.* 11, 274–280. doi: 10.1016/j.micinf.2008.12.002

Nenadic, O., and Greenacre, M. (2007). Correspondence Analysis in R, with two- and three-dimensional graphics: the ca package. *J. Stat. Softw.* 20, 1–13. doi: 10.18637/jss.v020.i03

Nicolas-Chanoine, M.-H., Gruson, C., Bialek-Davenet, S., Bertrand, X., Thomas-Jean, F. T., Bert, F., et al. (2013). 10-fold increase (2006-11) in the rate of healthy subjects with extended-spectrum β-lactamase-producing *Escherichia coli* faecal carriage in a Parisian check-up centre. *J. Antimicrob. Chemother.* 68, 562–568. doi: 10.1093/jac/dks429

Nowrouzian, F. L., Adlerberth, I., and Wold, W. E. (2006). Enhanced persistence in the colonic microbiota of *Escherichia coli* strains belonging to phylogenetic group B2: role of virulence factors and adherence to colonic cells. *Microbes Infect.* 8, 834–840. doi: 10.1016/j.micinf.2005.10.011

Nowrouzian, F. L., Ostblom, A. E., Wold, W. E., and Adlerberth, I. (2009). Phylogenetic group B2 *Escherichia coli* strains from the bowel microbiota of Pakistani infants carry few virulence genes and lack of capacity for long-term persistence. *Clin. Microbiol. Infect.* 15, 466–472. doi: 10.1111/j.1469-0691.2009.02706.x

Nowrouzian, F. L., Wold, W. E., and Adlerberth, I. (2005). *Escherichia coli* strains belonging to phylogenetic group B2 have superior capacity to persist in the intestinal microflora to infants. *J. Infect. Dis.* 191, 1078–1083. doi: 10.1086/427996

Obata-Yasuoka, M., Ba-Thein, W., Tsukamoto, T., Yoshikawa, H., and Hayashi, H. (2002). Vaginal *Escherichia coli* share common virulence factor profiles, serotypes and phylogeny with other extraintestinal *E.coli*. *Microbiology* 148, 2745–2752. doi: 10.1099/00221287-148-9-2745

Ochman, H., and Selander, R. K. (1984). Standard reference strains of *Escherichia coli* from natural populations. *J. Bacteriol.* 157, 690–693.

Oksanen, J. (2013). *Multivariate Analysis of Ecological Communities in R: vegan tutorial.* Version February 08, 2013. Available online at: http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf (Accessed July 29, 2013).

Orsi, R. H., Stoppe, N. C., Sato, M. I., Gomes, T. A., Prado, P. I., Manfio, G. P., et al. (2007). Genetic variability and pathogenicity potential of *Escherichia* *coli* isolated from recreational water reservoirs. *Res. Microbiol.* 158, 420–427. doi: 10.1016/j.resmic.2007.02.009

Pallechi, L., Lucchetti, C., Bartolini, A., Bartalesi, F., Mantella, A., Gamboa, H., et al. (2007). Population structure and resistance genes in antibiotic-resistant bacteria from a remote community with a minimal antibiotic exposure. *Antimicrob. Agents Chemother* 51, 1179–1184. doi: 10.1128/AAC.01101-06

Peel, M. C., Finlayson, B. L., and McMahon, T. A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrol. Earth Syst. Sci.* 11, 1633–1644. doi: 10.5194/hess-11-1633-2007

Petersen, A. M., Nielsen, E. M., Litrup, E., Brynskov, L., Mirsepasi, H., and Krogfelt, K. A. (2009). A phylogenetic group of *Escherichia coli* associated with active left-sided inflammatory bowel disease. *BMC Microbiol.* 9:171. doi: 10.1186/1471-2180-9-171

R Development Core Team (2012a). *The R Stats Package: R Statistical Functions, Version 2.15.0.* 30 March 2012. Available online at: http://stat.ethz.ch/R-manual/R-patched/library/stats/html/00Index.html (Accessed May 20, 2014).

R Development Core Team (2012b). *The R Graphics Package: R Functions for Base Graphics.* Version 2.15.0, 30 Mar. 2012. Available online at: http://stat.ethz.ch/R-manual/R-patched/library/graphics/html/00Index.html (Accessed May 20, 2014).

Rasko, D. A., Rosovitz, M., Myers, G. S., Mongodin, E. F., Fricke, W. F., Gajer, P., et al. (2008). The pangenome structure of *Escherichia coli* commensal and pathogenic isolates. *J. Bacteriol.* 190, 6881–6893. doi: 10.1128/J. B.00619-08

Riccobono, E., Pallechi, L., Mantella, A., Bartalesi, F., Zeballos, I. C., Trigoso, C., et al. (2012). Carriage of antibiotic-resistant *Escherichia coli* among healthy children and home-raised chickens: a household study in a resource-limited setting. *Microb. Drug Resist.* 18, 83–87. doi: 10.1089/mdr.2011.0003

Ripley, B., Venables, B., Hornik, K., Gebhardt, A., and Firth, D. (2013). *Support Functions and Datasets for Venables and Ripley's MASS.* Version 7.3-23. Available online at: http://www.stats.ox.ac.uk/pub/MASS4/ (Accessed April 05, 2014).

Sabaté, M., Prats, G., Moreno, E., Ballesté, E., Blanch,. A. R., and Andreu, A. (2008). Virulence and anitmicrobial resistance profiles among *Escherichia coli* strains isolated from human and animal wastewater. *Res. Microbiol.* 159, 288–293. doi: 10.1016/j.resmic.2008.02.001

Sannes, M. R., Kuskowski, M. A., Owens, K., Gajewski, A., and Johson, J. (2004). Virulence factors profile and phylogenetic background of Escherichia coli isolates from veterans with bacteremia and uninfected control subjects. *J. Infect. Dis.* 190, 2121–2128. doi: 10.1086/425984

Selander, R. K., and Levin, B. R. (1980). Genetic diversity and structure in *Escherichia coli* populations. *Science* 210, 545–547. doi: 10.1126/science.6999623

Silkie, S., and Nelson, K. L. (2009). Concentrations of host-specific and generic fecal markers measured by quantitative PCR in raw sewage and fresh animal feces. *Water Res.* 43, 4860–4871. doi: 10.1016/j.watres.2009.08.017

Silva, J. S., Stoppe, N. C., Torres, T. T. Ottoboni, L. M. M., and Saraiva, A. M. (2014). "Social network analysis metrics and their application in microbiological network studies," in *5th Workshop on Complex Networks CompleNet 2014*, Bologna. 549, 251–260.

Silva, N., Igrejas, G., Figueiredo, N., Goncalves, A., Radhouani, H., Rodrigues, J., et al. (2010). Molecular characterization of antimicrobial resistance in enterococci and *Escherichia coli* isolates from European wild rabbit (*Oryctolagus cuniculus*). *Sci. Total Environ.* 408, 4871–4876. doi: 10.1016/j.scitotenv.2010.06.046

Silva, N., Igrejas, G., Gonçalves, A., and Poeta, P. (2012). Commensal gut bacteria, distribution of Enterococcus species and prevalence of *Escherichia coli* phylogenetic groups in animals and humans in Portugal. *Ann. Microbiol.* 62, 449–459. doi: 10.1007/s13213-011-0308-4

Skurnik, D., Bonet, D., Bernède-Baudin, C., Michel, R., Baleire, C., Chau, F., et al. (2008). Characteristics of human intestinal *Escherichia coli* with changing environments. *Environ. Microbiol.* 10, 2132–2137. doi: 10.1111/j.1462-2920.2008.01636.x

Sorsa, L. J., Feldmann, F., Hildinger, K., Dufke, S., and Schubert, S. (2007). Characterization of four novel genomic regions of uropathogenic *Escherichia coli* highly associated with the extraintestinal virulent phenotype: a jigsaw puzzle of genetic modules. *J. Med. Microbiol.* 297, 83–95. doi: 10.1016/j.ijmm.2006.11.007

Stanley, C. R., and Dunbar, R. I. M. (2013). Consistent social structure and optimal clique size revealed by social network analysis of feral goats, Capra hircus. *Anim. Behav*. 85, 771–779. doi: 10.1016/j.anbehav.2013.01.020

Tallon, P., Magajna, B., Lofranco, C., and Leung, K. T. (2005). Microbial indicators of faecal contamination in water: a current perspective. *Water Air Soil Poll.* 166, 139–166. doi: 10.1007/s11270-005-7905-4

Tanner, C. J., and Jackson, A. L. (2012). Social structure emerges via the interaction between local ecology and individual behaviour. *J. Anim. Ecol.* 81, 260–267. doi: 10.1111/j.1365-2656.2011.01879.x

Tenaillon, O., Skurnik, D., Picard, B., and Denamur, E. (2010). The population genetics of commensal *Escherichia coli*. *Nat. Rev. Microbiol.* 8, 207–217. doi: 10.1038/nrmicro2298

Therneau, T., Atkinson, B., and Ripley, Y. B. (2013). *Recursive Partitioning and Regression Trees.* Version 4.1-1. Available online at: http://cran.r-project.org/web/packages/rpart/ (Accessed April 20, 2014).

Unno, T., Han, D., Jang, J., Lee, S. N., Ko, G. P., Choi, H. Y., et al. (2009). Absence of *Escherichia coli* phylogenetic group B2 strains in human and domesticated animals from Jeonnam Province, Republic of Korea. *Appl. Environ. Microbiol.* 75, 5659–5666. doi: 10.1128/AEM.00443-09

USEPA (2005). *Microbial Source Tracking Guide Document.* Washington, DC: Office of Research and Development.

USEPA (2002). *Method 1603 Escherichia coli (E. coli) in Water by Membrane Filtration using Modified Membrane-Thermotolerant Escherichia coli Agar (Modified m-TEC), EPA 821-R-02-023.* Cincinatti, OH: Environmental Protection Agency, Office of Research and Development.

Valverde, A., Canton, R., Garcillan-Barcia, M., Novais, A., Gallan, J. C., Alvarado, A., et al. (2009). Spread of blaCTX-M-14 is driven mainly by lncK plasmids disseminated among *Escherichia coli* phylogroups A, B1 and D in Spain. *Antimicrob. Agents Chemother.* 53, 5204–5212. doi: 10.1128/AAC.01706-08

Vinué, L., Sáenz, Y., Somalo, S., Escudero, E., Moreno, M. A., Ruiz-Larrea, F., et al. (2008). Prevalence and diversity of integrons and associated resistance genes in faecal *Escherichia coli* isolates of healthy humans in Spain. *J. Antimicrob. Chemother* 62, 934–937. doi: 10.1093/jac/dkn331

Vollmerhausen, T. L., Ramos, N. L., Gündogdu, A., Robinson, W., Brauner, A., and Katouli, M. (2011). Population structure and uropathogenic virulence-associated genes of faecal *Escherichia coli* from healthy young and elderly adults. *J. Med. Microbiol.* 60, 574–581. doi: 10.1099/jmm.0.027037-0

White, A. P., Sibley, K. A., Sibley, C. D., Wasmuth, J. D., Schaefer, R., Surette, M. G., et al. (2011). Intergenic sequence comparison of *Escherichia coli* isolates reveals lifestyle adaptations but not host specificity. *Appl. Environ. Microbiol.* 77, 7620–7632. doi: 10.1128/AEM.05909-11

Whitman, R. L., Nevers, M. B., and Byappanahalli, N. (2006). Examination of watershed-wide distribution of *Escherichia coli* along Southern Lake Michigan: and integrated approach. *Appl. Environ. Microbiol.* 72, 7301–7310. doi: 10.1128/AEM.00454-06

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *J. Stat. Softw.* 40, 1–29. doi: 10.18637/jss.v040.i01

Winter, C., Matthews, B., and Suttle, C. A. (2013). Effects of environmental variation and spatial distance on Bacteria, Archaea and viruses in sub-polar and arctic waters. *ISME J.* 7, 1507–1518. doi: 10.1038/ismej.2013.56

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edn.* Burlington, MA: Morgan Kaufmann.

Zhang, L., Foxman, B., and Marrs, C. (2002). Both urinary and rectal *Escherichia coli* isolates are dominated by strains of phylogenetic group B2. *J. Clin. Microbiol.* 40, 3951–3955. doi: 10.1128/JCM.40.11.3951-3955.2002

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.