AMERICAN SOCIETY FOR MICROBIOLOGY mSphere®

RESEARCH ARTICLE
Host-Microbe Biology

Genomic Epidemiology and Evolution of *Escherichia coli* in Wild Animals in Mexico

Robert Murphy,[a,b] Martin Palm,[c,d] Ville Mustonen,[e,f] Jonas Warringer,[c] Anne Farewell,[c] Leopold Parts,[g,h] Danesh Moradigaravand[b]

[a]University of Copenhagen, Department of Biology, Section for Ecology and Evolution, Copenhagen, Denmark
[b]Center for Computational Biology, Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, United Kingdom
[c]Department for Chemistry and Molecular Biology, University of Gothenburg, Gothenburg, Sweden
[d]Centre for Antibiotic Resistance Research at the University of Gothenburg, Gothenburg, Sweden
[e]Organismal and Evolutionary Biology Research Programme, Department of Computer Science, Institute of Biotechnology, University of Helsinki, Helsinki, Finland
[f]Helsinki Institute for Information Technology, Helsinki, Finland
[g]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, United Kingdom
[h]Department of Computer Science, University of Tartu, Tartu, Estonia

ABSTRACT *Escherichia coli* is a common bacterial species in the gastrointestinal tracts of warm-blooded animals and humans. Pathogenicity and antimicrobial resistance in *E. coli* may emerge via host switching from animal reservoirs. Despite its potential clinical importance, knowledge of the population structure of commensal *E. coli* within wild hosts and the epidemiological links between *E. coli* in nonhuman hosts and *E. coli* in humans is still scarce. In this study, we analyzed the whole-genome sequencing data of a collection of 119 commensal *E. coli* strains recovered from the guts of 55 mammal and bird species in Mexico and Venezuela in the 1990s. We observed low concordance between the population structures of *E. coli* isolates colonizing wild animals and the phylogeny, taxonomy, and ecological and physiological attributes of the host species, with distantly related *E. coli* strains often colonizing the same or similar host species and distantly related host species often hosting closely related *E. coli* strains. We found no evidence for recent transmission of *E. coli* genomes from wild animals to either domesticated animals or humans. However, multiple livestock- and human-related virulence factor genes were present in *E. coli* of wild animals, including virulence factors characteristic of Shiga toxin-producing *E. coli* (STEC) and atypical enteropathogenic *E. coli* (aEPEC), where several isolates from wild hosts harbored the locus of enterocyte effacement (LEE) pathogenicity island. Moreover, *E. coli* isolates from wild animal hosts often harbored known antibiotic resistance determinants, including those against ciprofloxacin, aminoglycosides, tetracyclines, and beta-lactams, with some determinants present in multiple, distantly related *E. coli* lineages colonizing very different host animals. We conclude that genome pools of *E. coli* colonizing the guts of wild animals and humans share virulence and antibiotic resistance genes, underscoring the idea that wild animals could serve as reservoirs for *E. coli* pathogenicity in human and livestock infections.

IMPORTANCE *Escherichia coli* is a clinically important bacterial species implicated in human- and livestock-associated infections worldwide. The bacterium is known to reside in the guts of humans, livestock, and wild animals. Although wild animals are recognized as potential reservoirs for pathogenic *E. coli* strains, the knowledge of the population structure of *E. coli* in wild hosts is still scarce. In this study, we used fine resolution of whole-genome sequencing to provide novel insights into the evolution of *E. coli* genomes from a small yet diverse collection of strains recovered within a broad range of wild animal species (including mammals and birds), the coevolution of *E. coli* strains with their hosts, and the genetics of pathogenicity

Citation Murphy R, Palm M, Mustonen V, Warringer J, Farewell A, Parts L, Moradigaravand D. 2021. Genomic epidemiology and evolution of *Escherichia coli* in wild animals in Mexico. mSphere 6:e00738-20. https://doi.org/10.1128/mSphere.00738-20.

Editor Maria L. Marco, University of California, Davis

Copyright © 2021 Murphy et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Leopold Parts, leopold.parts@sanger.ac.uk, or Danesh Moradigaravand, d.moradigaravand@bham.ac.uk.

Received 22 July 2020
Accepted 11 December 2020
Published 6 January 2021

of *E. coli* strains in wild hosts in Mexico. Our results provide evidence for the clinical importance of wild animals as reservoirs for pathogenic strains and highlight the need to include nonhuman hosts in the surveillance programs for *E. coli* infections.

KEYWORDS *Escherichia coli*, genomic epidemiology, host-pathogen interaction, infectious diseases, whole-genome sequencing, wild animals

Escherichia coli is the most prevalent aerobic bacterial species that resides in the intestines and feces of warm-blooded animals and dominates the corresponding microbiomes (1). Hosts provide the bacterium with a constant supply of nutrients and protection against environmental stresses, and the commensal nature of *E. coli* may facilitate its dissemination across hosts (2, 3). Pathogenic and antimicrobial-resistant (AMR) clones of *E. coli* have spread rapidly over recent years, and understanding the ecological origins of these strains has therefore become increasingly important. A significant number of acute *E. coli* infections are known to have zoonotic origins (2). Because half of the total natural *E. coli* population is estimated to inhabit environmental sites, nonhuman hosts and settings are large potential reservoirs for pathogenic and AMR strains and genes (4).

Despite its likely importance for human health, the genetic diversity of commensal *E. coli* within wild hosts is not well understood, primarily due to the difficulty of recovering samples. Some studies have suggested that *E. coli* and colonized hosts coevolve, such that the genomic characteristics of *E. coli* depend on the host species (5, 6). While neutral evolutionary forces such as genetic drift likely dictate most of the *E. coli* genetic diversity, microenvironments of gastrointestinal tracts of the hosts may exert powerful selection pressures that influence *E. coli* feeding habits and diet and contribute to the phenotypic differentiation of commensal strains. These factors have led to *E. coli* strains from wild animals often falling into other genetic and phenetic clades, and thus into other phylogroups, than isolates retrieved from humans (5–8).

Reports have described high multidrug resistance in individual environmentally sourced *E. coli* isolates, but strains found in wild animals generally display lower AMR than those found in livestock and nonanimal environmental samples. Further, proximity of wildlife to human settlements seems to influence the AMR of gut microbiomes in wild hosts, likely due to the closely associated antibiotic pollution of land and water environments (9, 10). Interactions between humans and livestock have also been reported to catalyze the colonization of wild-life by AMR *E. coli* in Nairobi, Kenya. But whether wild animals predominantly act as sources or sinks in AMR evolution is still unclear (11, 12). The distribution of genes encoding virulence factors in environmental *E. coli* isolates is still an understudied area, although accumulating evidence shows high genetic relatedness between pathogenic strains infecting livestock and those infecting humans. This suggests that jumps between animal and human hosts do occur at epidemiological time scales (13, 14). Furthermore, a recent study on human and environmental strains in Australia found a dominant role for horizontal gene transfer in spreading virulence factor and AMR genes across hosts and showed that host phylogeny and habitat can shape the *E. coli* genetic diversification (8).

Exploration of the potential environmental, and more specifically zoonotic, origins of AMR and pathogenic *E. coli* strains and genes requires studies of genetically different bacterial isolates from a wide diversity of sources and geographical regions. Here, we examined the whole-genome sequences of 119 commensal *E. coli* isolates recovered from the fecal samples of 55 wild mammal and bird species from North America, predominantly from Mexico (5). With an estimated 2,000 different resident mammal and bird species, Mexico hosts 10 to 12% of the corresponding worldwide diversity (15). This allowed us to scrutinize the host-pathogen evolution across a wide range of wild host populations, at a regional level.

Our results indicate that *E. coli* populations in wild hosts are only weakly associated with the taxonomy and ecological and physiological attributes of the host species. Furthermore, we detected a few incidents of epidemiological links between animal and human hosts. We also found that *E. coli* isolates from these host types were clearly mixed into local populations and that antibiotic resistance and virulence genes had been shared between strains from wild and domesticated/livestock animal hosts. These results suggest that wild hosts indeed can serve as reservoirs for *E. coli* pathogens and underscore the importance of large-scale population genomics studies of *E. coli* across multiple host species.

RESULTS

We sequenced 119 strains isolated from 55 wild animal host species predominantly in Mexico and found them to capture much of the known global *E. coli* genetic diversity. Indeed, our wild-host collection contained representatives of all of the major known phylogroups of *E. coli*, with group B1 (55 strains; 47% of the total) being most prevalent, followed by B2 (21 strains; 18%), A (17 strains; 14%), D (15 strains; 13%), and E (7 strains; 6%) (Fig. 1A). The high frequency of B1 strains is consistent with previous epidemiological reports on *E. coli* isolated from domesticated animals but stands in contrast to the high prevalence of phylogroups B2 and A among *E. coli* isolates from human hosts, indicating an association between the *E. coli* phylogenetic structure and the type of natural host colonized by *E. coli* (16). Analyzing the genomes of our *E. coli* isolates from wild hosts together with those of many previously sequenced strains with diverse origins, we found that *E. coli* from domesticated/livestock animals and North America were disproportionately likely to share phylogenetic origins with our wild *E. coli* strains (see Fig. S1A and B in the supplemental material). This suggests a regional dissemination of specific *E. coli* clones (sequence types) across both domesticated/livestock and wild animals in North America, which motivated us to examine the prevalence of transmissions and recent divergences at a fine resolution.

We compared the phylogenetic trees for our collection of *E. coli* strains from wild animals and their hosts to understand the long-term concordance between the evolutionary histories of *E. coli* and their host species and the extent to which the genetic distance between the host species agrees with the regional population structure of *E. coli*. Both comparisons of host and *E. coli* distance matrices ($P = 0.0001$; Mantel test) (Fig. 2A) and comparisons of distances between phylogenetic trees for *E. coli* strains and hosts to distances in randomized trees ($P = 0.003$; 1,000 tests) (Fig. 2B) rejected completely random observations. Despite this, we found only a moderate correlation of 0.47 between the genetic distance matrices for *E. coli* strains and hosts (Fig. 2A), with closely related *E. coli* strains sometimes colonizing divergent wild hosts and closely related wild animal species sometimes hosting distantly related *E. coli*. The limited genetic association between *E. coli* isolates and their wild hosts was also evident at higher taxonomic levels, with only weak genetic clustering of *E. coli* according to the host class, order, and family (Fig. 1B). This was further confirmed by the extensive overlap in the distributions of single-nucleotide-polymorphism (SNP) distances for *E. coli* pairs colonizing host species from the same taxonomic groups and those of pairs colonizing different taxonomic groups (Fig. 1C), as 0.95, 0.95, and 0.96 of ranges of distributions overlapped for taxonomic ranks of class, order, and family, respectively. Moreover, we found the accessory genome of our *E. coli* colonizing wild hosts to have evolved in concert with their core genome ($P = 0.0001$; Pearson's $R = 0.85$; Mantel test on distance matrices for core genome and accessory genes) (Fig. S2). Together, these results provide evidence of regional gene flow across *E. coli* lineages colonizing wild-animal hosts in Mexico (see Discussion).

Host adaptation is a consequence of diversifying selection across lineages, but it may be influenced by random effects due to, e.g., population structure (17). To examine the extent to which selection has shaped genetic variation in our *E. coli* collection from wild-animal hosts, we compared the rates of nonsynonymous and synonymous single-nucleotide evolution (K_a/K_s) since their last shared common ancestor. Of 3,529

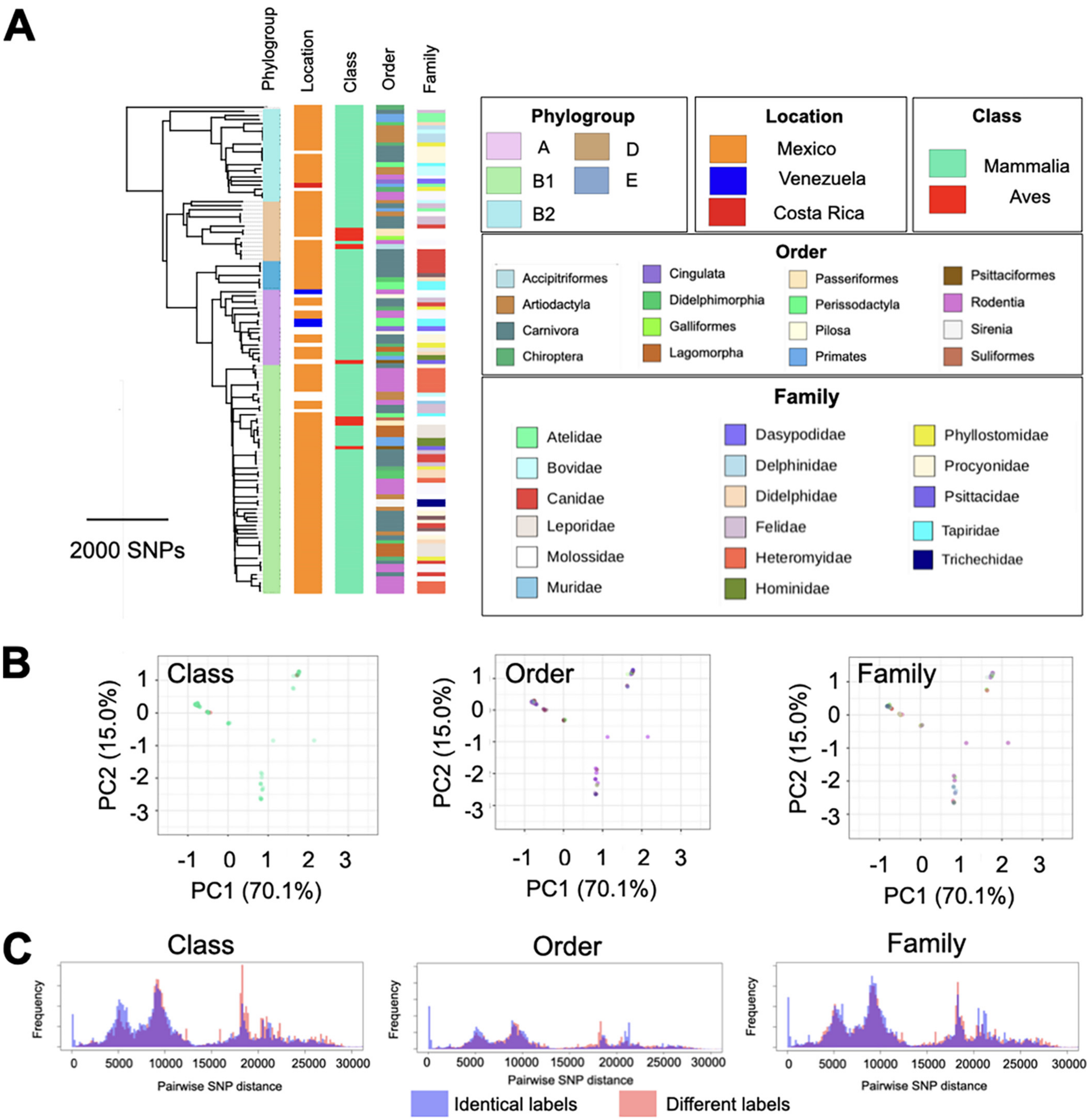

FIG 1 Phylogenetic distribution of host specificity and cluster analysis. (A) Phylogenetic tree of our *E. coli* strains from wild-animal hosts and its association with host taxonomy, at different taxonomic levels. Families of host species colonized by only one *E. coli* strain in our collection are not shown. (B) Principal-component analysis of our *E. coli* strains, with labels representing the phylogroup of the *E. coli* isolate and the taxonomic rank of the host species. Each color corresponds to one taxonomic rank, shown in panel A. (C) Distribution of pairwise SNP distances for *E. coli* strains from hosts belonging to the same (red) and different (blue) taxonomic ranks.

genes in the core genome, 242 had a K_a/K_s value above 1 in at least one strain, with an average of 11.7 genes, i.e., 0.3% of total genes, per strain falling in this category (Fig. S3A and B; Table S3). The number of genes under strong positive selection did not show any link with the type of wild-animal host from which the strains had been isolated (Fig. S3B). The strongly selected genes encoded proteins involved in a broad range of functions, with genes encoding energy production, carbohydrate and ion metabolism and transport, and signal transduction proteins being slightly overrepresented

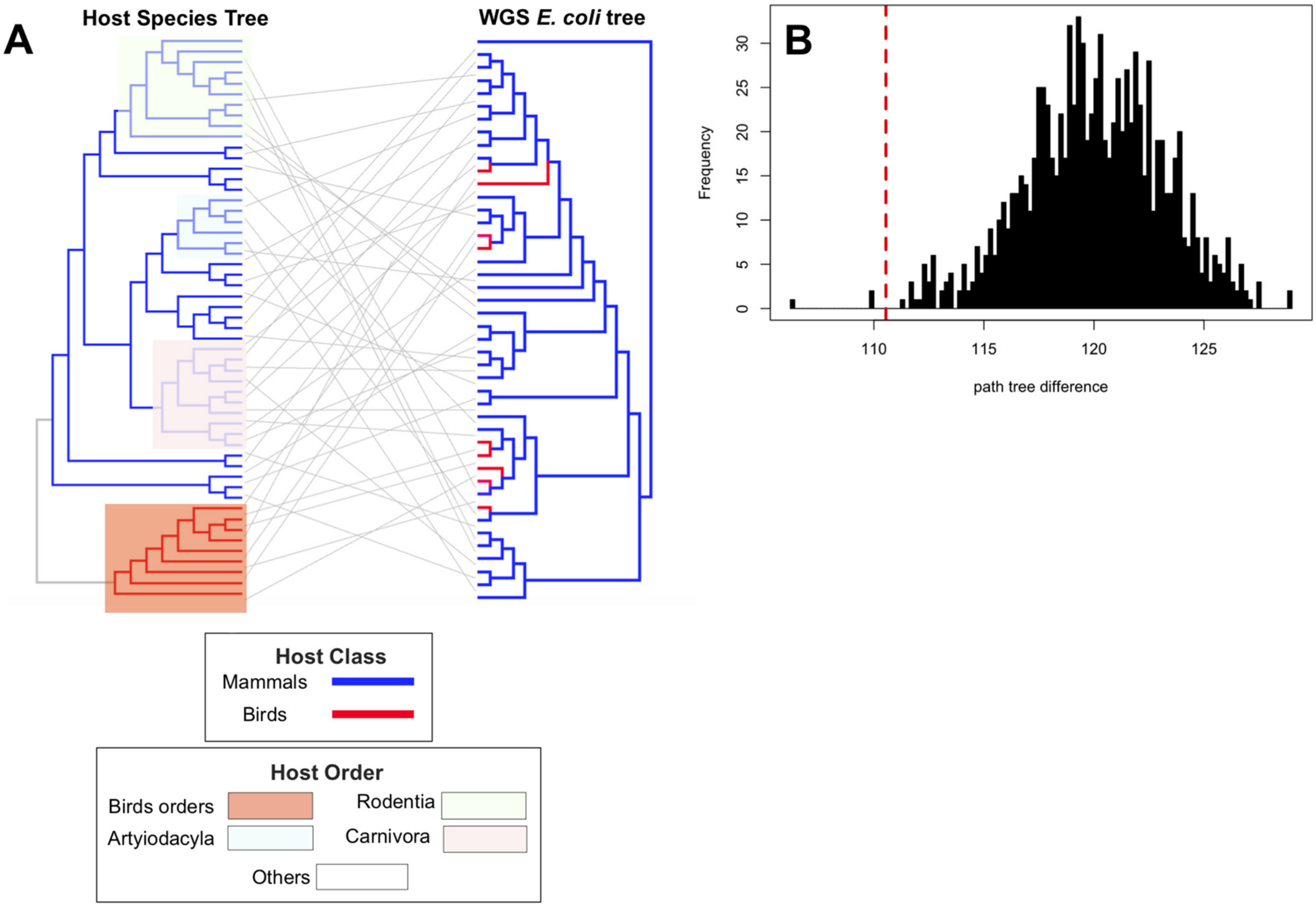

FIG 2 Concordance between host and *E. coli* phylogenetic trees. (A) Phylogenetic tree of the whole-genome sequencing of *E. coli* strains and the tree of life (TOL) for host strains. For host species for which more than one isolate were available in the data set, one strain was randomly drawn. Clades for bird and major mammalian orders are highlighted. (B) The frequency (*y* axis) of path tree differences relative to the *E. coli* tree (*x* axis) for 1,000 random shuffling of tree tips of the host tree in panel A (black bars), contrasted to the observed value from unpermuted data (red dashed line).

(Fig. S3B). This pattern agrees with a complex nature of the *E. coli* adaption to colonizing the guts of different wild-animal hosts and the degree of genome-wide selection having been little influenced by the type of host species colonized.

We next probed the genome evolution and epidemiology of our *E. coli* isolates colonizing wild animals in relation to those of the external collection of *E. coli* isolates coming from other hosts. Our analysis did not identify *E. coli* isolates with zero, or close to zero, SNP distance within wild hosts or food animal sources, which would suggest direct, recent transmissions. However, the analysis of SNP clusters revealed two links between wild hosts and clinical strains. One link weakly (38 SNPs) connected a sample (SAMEA4607586) from the avian species *Aratinga canicularis* to two clinical strains from Mexico (SNP cluster ID PDS000073768.1), and one moderately strongly (17 SNPs) connected an *E. coli* isolate from the avian species of *Gallus gallus* and 7 clinical strains (SNP cluster PDS000066827.4). One and nine of the clinical strains were from France and the United States, respectively. Interestingly, of 22 external strains isolated from wild hosts in Mexico, 17 formed parts of SNP clusters with our strains, showing extensive circulation of *E. coli* across wild hosts in Mexico (Table S1). The phylogenetic analysis revealed genetic similarity between our *E. coli* isolates colonizing wild animals and *E. coli* isolates colonizing domesticated animals in the B1 phylogroup, where one-third of our *E. coli* isolates from wild animals clustered with external lineages isolated from domesticated/livestock animals (*n* = 96), food (*n* = 12), and environmental sources (*n* = 13) (see Materials and Methods). We reconstructed the Bayesian tree of these 158 strains and found their last common ancestor to have lived about 1,000 years ago, with

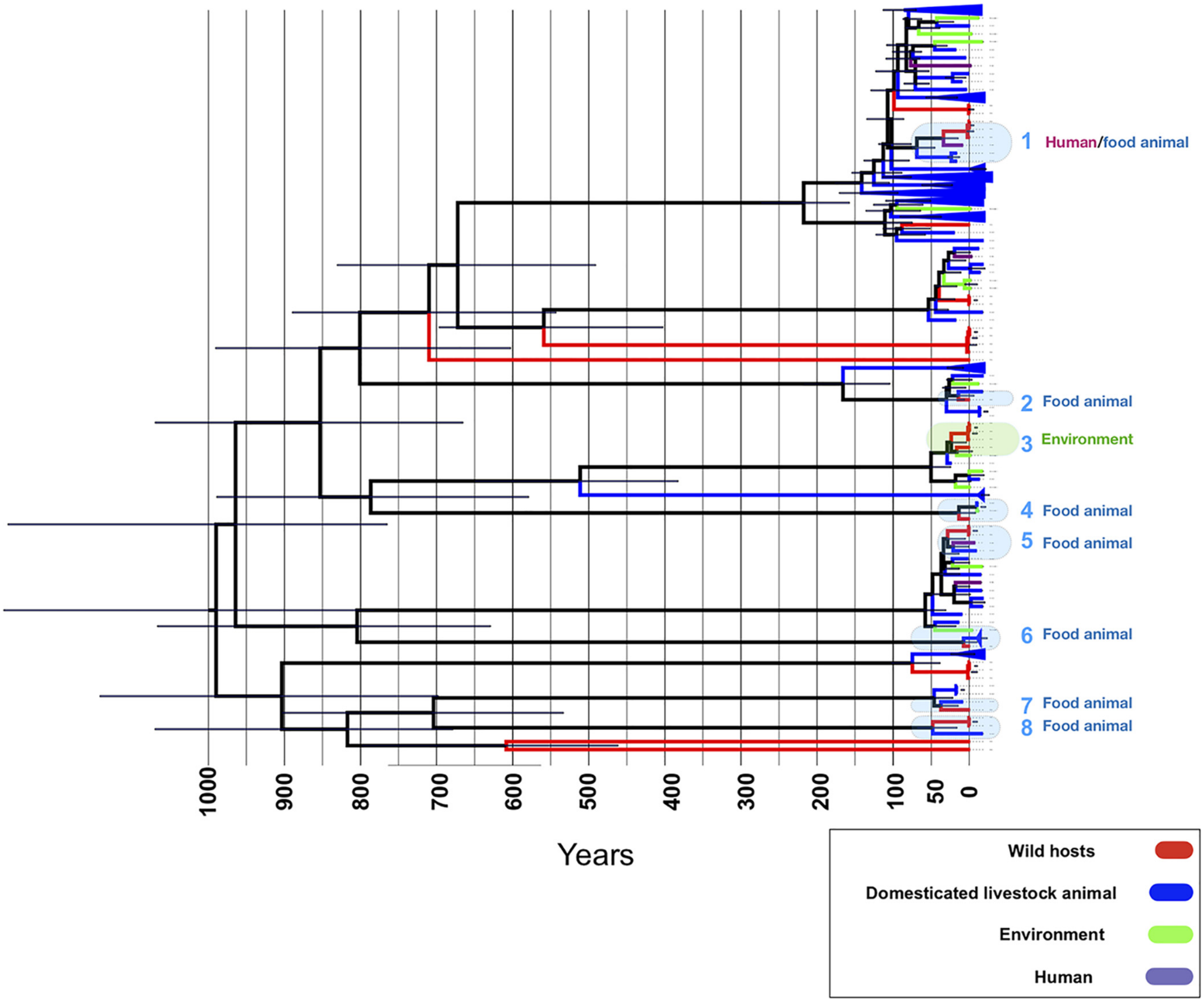

FIG 3 Recent mixing of wild and nonwild host lineages. Bayesian tree for strains in a clade belonging to phylogroup B1. The shaded boxes show putative host jump events between wild hosts and other sites, i.e., domesticated animals, environment, and humans, over the past 100 years. The error bar shows the 95% confidence interval from the Bayesian tree.

a substantial expansion of the clade over the past 100 years (Fig. 3). We identified eight incidents of strains jumping between wild animals and other sources in this clade, all during the last 100 years and all but one during the last 50 years (Fig. 3). One incident involved *E. coli* jumping between wild hosts residing in city regions and domesticated/livestock animals. These *E. coli* host switches between wild animals and other sources may reflect anthropogenic intervention in the habitats of their wild hosts, and the rapid urban and agricultural growth and environmental degradation in Mexico over the past decades (18).

The incidents of *E. coli* jumping between wild and domesticated animals led us to examine whether *E. coli* isolates colonizing the former harbor any known human- or food-animal-linked virulence factors. We identified a range of virulence factor genes, including four types of toxin genes, two adhesin genes, two iron chelators, and three transporters. These were present in *E. coli* isolates colonizing different wild animals (Fig. 4A). The frequency of virulence factors was on average higher for strains recovered from Primate (11.5 genes per isolate), Rodentia (9.5 genes per isolate) and Carnivora (12.5 genes per isolate) host species (Fig. 4A and B). Some host species not

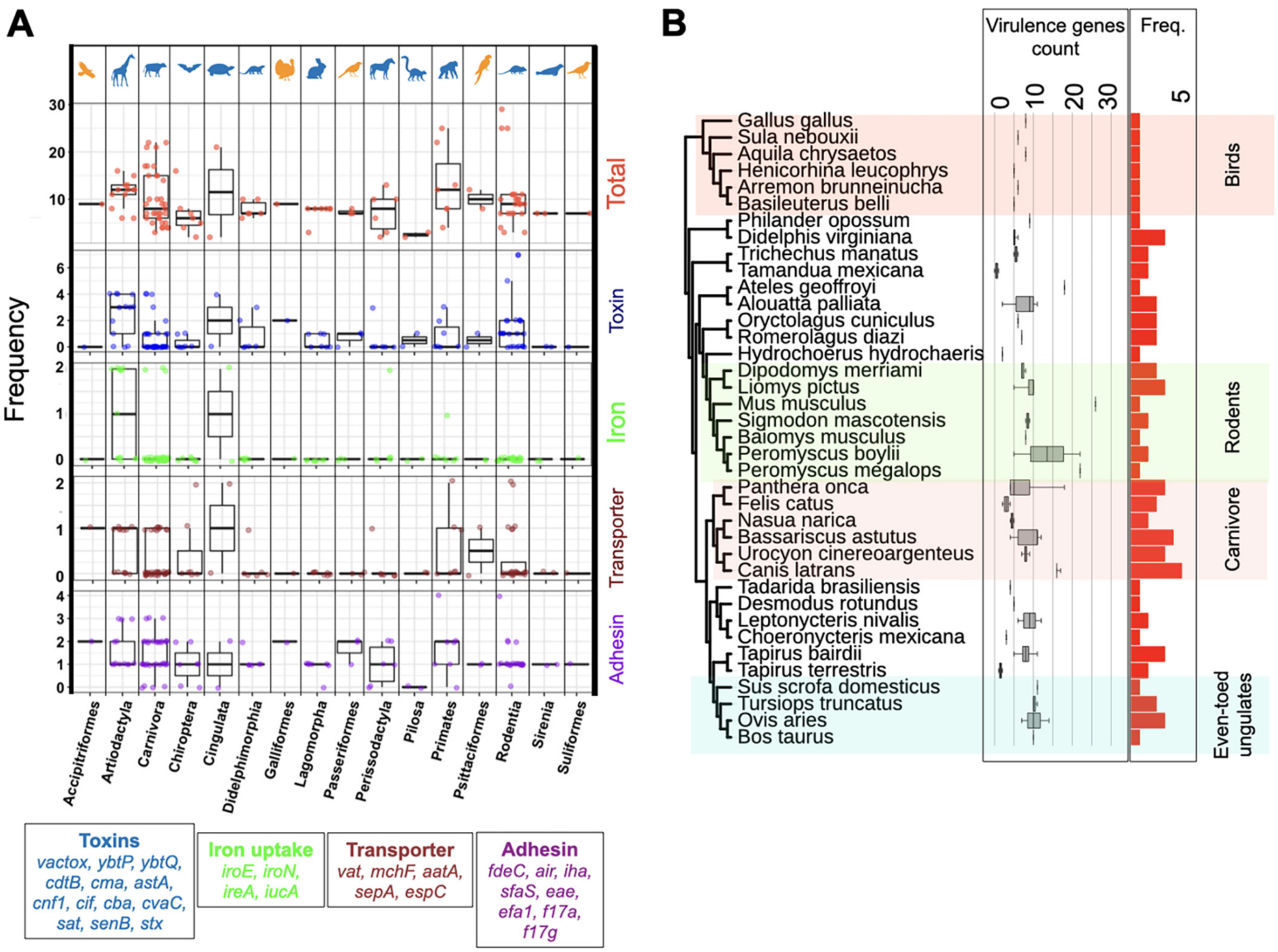

FIG 4 Distribution of virulence factor genes. (A) Frequency of virulence factors genes across functional groups and taxonomic orders. (B) Phylogenetic distribution of *E. coli* virulence genes across wild animal host species. The tree shows the tree of life for hosts, where major orders are shown in shaded boxes. Bar plots show the frequency of genes. Horizontal box plots represent the distribution of virulence genes for strains recovered from each host across host orders.

closely related to humans, such as avian species, were colonized by strains carrying a high number of virulence factors (Fig. 4A and B), suggesting that the pattern is not simply a reflection of the higher frequency of human- and livestock-associated genes in the database.

Because both the physiology and ecology of the host species can affect the virulence factors encoded in the genomes of infectious bacteria, we examined the relationship between the number of virulence genes in *E. coli* isolates colonizing wild animals and the 45 such features in the panTHERIA database. A previous study on four virulence genes revealed that the body mass of the host species can be positively linked with the number of virulence factors present in the gut microbiome, and this was attributed to the gut complexity (19). However, our analysis of many more virulence genes showed no such correlation, considering either adult, neonate, or weaning body mass (Fig. S4A). Only the terrestriality, i.e., adaptation to living on land ($P = 0.03$; Spearman's $\rho = -0.18$), habitat breadth ($P = 0.05$; Spearman's $\rho = -0.24$), diet breadth ($P = 0.02$; Spearman's $\rho = -0.26$), and social group size ($P = 0.02$; Spearman's $\rho = 0.27$) of hosts correlated significantly with *E. coli* virulence gene counts. More diverse habitats and diets of the hosts were associated with fewer virulence genes, and the formation of larger social groups was associated with more virulence genes in *E. coli* isolates colonizing these hosts (Fig. S4A and B). Larger social groups, as observed in Carnivora,

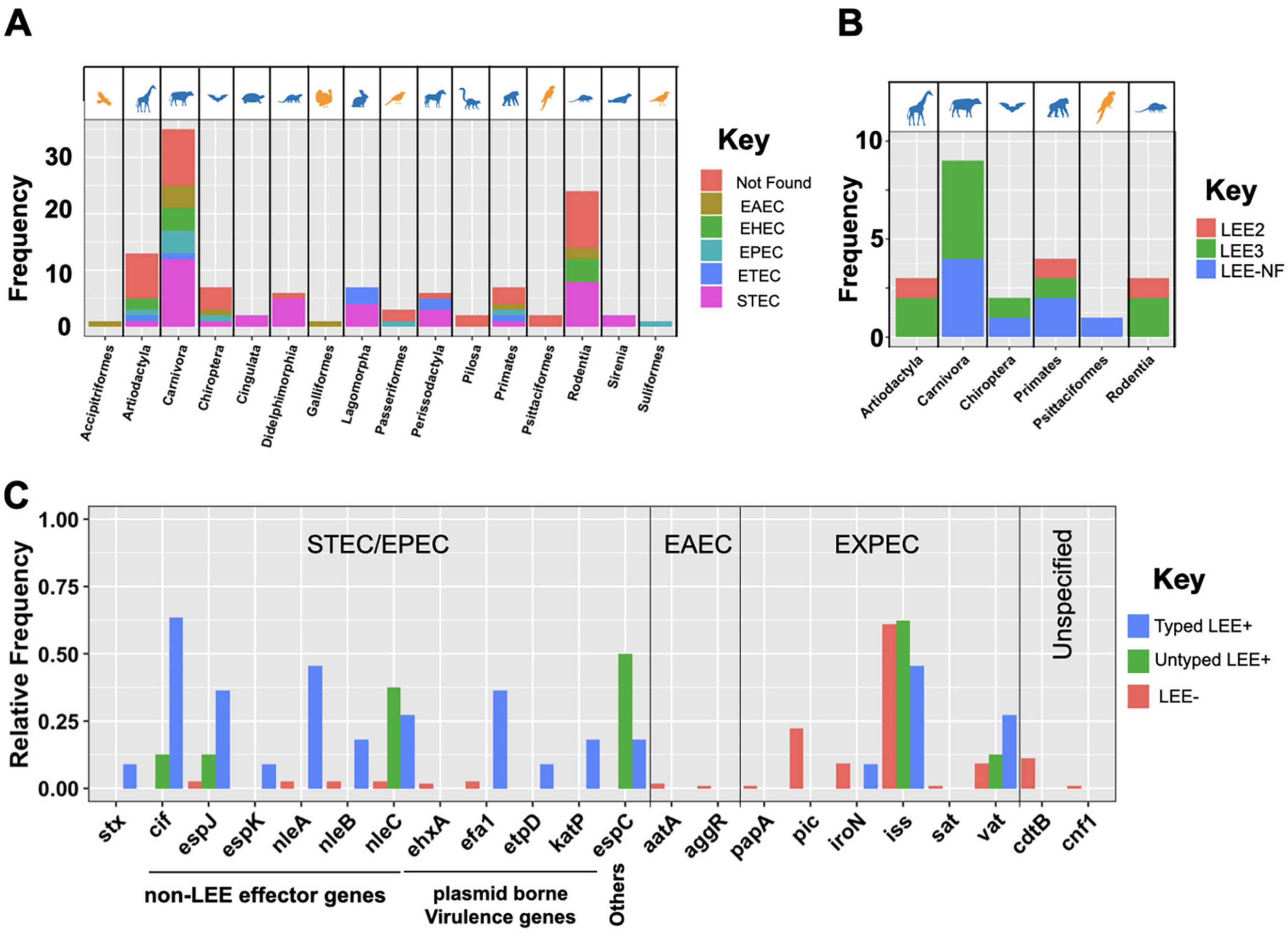

FIG 5 Sharing of serotypes and distribution of LEE genes and effectors genes across hosts. (A) Distribution of serotypes shared between *E. coli* isolates colonizing wild hosts and known pathovars across taxonomic orders of hosts. (B) Distribution of typed and nontyped LEE families across taxonomic orders of hosts. (C) Distribution of virulence genes and LEE effector genes in typed LEE-positive, untyped LEE-positive, and LEE-negative strains.

Artiodactyla, and Primates (Fig. S4C), are known to increase the social transmissions of infectious agents, such as *E. coli*, in animal societies, which may facilitate the dispersion of virulence genes among these infectious agents (20). Although a larger sample set is needed to examine the impact of potential confounding factors, the findings further support the idea that a complex network of host- and environment-related factors shapes the genomic characteristics of commensal *E. coli* strains.

Certain *E. coli* serotypes, which reflect O, H, and K antigen variation, are recognized to cause virulence in human- and livestock-associated infection. We found 53 and 14 serotypes to be shared between our *E. coli* strains colonizing wild animal hosts and those in domestic animal and human infections, respectively (Table S1). In total, we identified 71 distinct serogroups and 14 strains among *E. coli* isolates colonizing wild animals that were not typeable according to known serotype patterns, further underscoring their broad diversity. The serogroups of 74 strains overlapped with those of known pathovars, including non-O157 Shiga toxin-producing *E. coli* (STEC) ($n = 40$), enterotoxigenic *E. coli* (ETEC) strains ($n = 12$), enteropathogenic (EPEC) strains ($n = 11$), and enteroaggregative *E. coli* (EAEC) strains ($n = 11$), across hosts (Fig. 5A; Table S1). The pathovars are recognized to have nonhuman sources and are known to be acquired by humans via direct contact with either animals or their feces, in petting zoos and on farms (for STEC), or through the consumption of contaminated water and food (for EAEC and ETEC), as previously reported in Mexico (21, 22). ETEC is also an important cause of diarrhea in domestic animals, notably in calves and piglets (23). Two strains from our wild host collection shared serotypes with human-pathogenic strains and contained the genetic virulence hallmarks of their associated pathovars. One strain belonged to O111:H8, a clinically relevant enterohemorrhagic *E. coli* (EHEC) serotype, and contained both the enterocyte effacement (LEE) pathogenicity island (PAI) and the toxin stx_2 gene. This strain was recovered from a wild sheep close to a city. The other strain belonged to O78:H34 and was isolated from a parakeet carrying EAEC virulence genes, including the gene for the plasmid-encoded, heat-stable enterotoxin toxin (EAST-1) and *aatA* and *aggR*, encoding a transporter of a virulence protein and a virulence regulator, respectively (Fig. 5). The serogroup was recently isolated from free pigeons in Brazil, showing the circulation of the pathovar among birds (24). None of the 74 strains whose serogroups were associated with STEC and ETEC pathovars were found to carry a toxin gene. Since our strains were recovered from feces, the virulence ability of their shared serotypes outside the gut in wild hosts is unclear. However, our findings are in line with the idea that virulent pathovars in food animals have emerged as a result of the acquisition of virulence factor genes by isolates belonging to serotypes of wild-host origin.

We found the pathogenicity island LEE, which is a hallmark of STEC and EPEC pathovars, in 21 of the *E. coli* lineages from wild-animal hosts, and these hosts belonged to six different taxonomic orders (Fig. 5B; Table S1). The LEE encodes factors required for the colonization of the human intestine (25). However, the absence of the plasmid carrying *E. coli* adherence factors (pEAF) led us to classify these isolates as atypical EPEC (aEPEC), an *E. coli* class that is widely spread across food animals and humans (26). Our LEE-positive *E. coli* strains also harbored other virulence factors that are typical of EAEC and extraintestinal pathogenic *E. coli* (ExPEC) pathovars and that affect pathogenicity (Fig. 5C). This included genes normally located on STEC virulence plasmids, such as pO157, pO26, *espP*, and *nle*, all of which were significantly more frequent among our LEE-positive strains than among our LEE-negative strains ($P < 0.001$ for two-sided Bonferroni-corrected Fisher's exact test) (Fig. 5C). We found that 2 and 11 strains, all in the B1, E, and D phylogroups, carried the LEE2 and LEE3 variants, respectively, while 8 strains, mainly in the B2 phylogroup, carried a nontypeable LEE. All three locus types were broadly distributed among host taxonomic families, in agreement with their benefitting *E. coli* colonization of animal guts in a general sense, as previously proposed for bovine hosts (13). Our findings also agree with the virulence ability of aEPEC strains spanning a broad host range and with the idea that virulence in STEC and EPEC strains has evolved by commensal strains acquiring virulence factors sequentially (26). We also identified a set of ExPEC-associated genes, encoding toxins (*pic*, *sat*, and *vat*), iron uptake (*iroN* and *iha*), and serum resistance (*iss*) proteins (Fig. 5C). Whether the presence of these genes is sufficient for a strain to cause sepsis or bacteremia is unclear, since the transition between asymptomatic colonization of the guts to the spreading of the bacteria into the bloodstream strains is poorly understood for ExPEC strains (27).

We tested the sensitivity of our collection of *E. coli* isolates from wild animal hosts to antibiotics in common use against human *E. coli* infections and found them to be mostly sensitive, except that 65% of the strains were resistant to ampicillin (Table S1). Their general susceptibility to antibiotics agrees with the lack of historical exposure of *E. coli* colonizing wild animals to therapeutic levels of antimicrobials. However, despite their general sensitivity to antibiotics, we found a range of AMR genes against beta-lactamase, aminoglycosides, sulfonates, and ciprofloxacin in the genomes of different *E. coli* lineages colonizing different host species (Fig. S5). This discordance between AMR phenotypes and genotypes points to regulation mechanisms or other epistatic effects that reduce the phenotypic penetrance of these resistance genes. The genomic context of these AMR genes turned out to be diverse, with genetic linkage to a range of phage genes and insertion sequence (IS) elements, including to IS91 and IS10. For AMR genes located on sufficiently long contigs, we explored the genomic context and found similarity with broad-host-range Col plasmid ($n = 21$) and chromosomal ($n = 3$) regions. The genomic contexts varied across host species; for example, while one strain from a member of the Pilosa, a placental mammal, harbored a distinct AMR gene cassette consisting of *tet*, *str*, and *sul* genes, we found the AMR genes of four other strains, from different mammalian species, to be sporadically distributed across the genomes. Besides plasmid-borne resistance determinants, we identified a set of ciprofloxacin resistance mutations in the *parE*, *parC*, and *gyrA* genes which emerged independently across lineages (Fig. S5). The strains had been recovered from Carnivora, Rodentia, and Passeriformes species. Four of the isolates belonged to the clinically relevant O17/77:H18 serotype, which forms a highly relevant pathogenic group in phylogroup D that was a clinical threat in the 1990s, predominantly in North America (28). Ciprofloxacin was introduced into clinical settings in the 1980s (29), prior to the sampling time period of our collection. The presence of ciprofloxacin resistance determinants in wild hosts, therefore, suggests that either rapidly emerging resistance was transmitted from wild hosts into human settings prior to the sampling time period or resistance preexisted in wild-host reservoirs.

DISCUSSION

We examined a collection of *E. coli* strains in wild hosts in Mexico to understand the regional genomic epidemiology of these strains. We integrated available data on the host species with the *E. coli* whole-genome sequencing data to understand the host-associated population structure in the collection. Despite the limited size of the collection and its regional nature, we found it to be genetically diverse, containing representatives from all major phylogroups of *E. coli*. We also found some of our strains to belong to local *E. coli* populations also colonizing lineages of domesticated/companion animals in the region. Moreover, some of our wild-animal strains harbored virulence and AMR genes that they shared with lineages identified as pathogenic to human and livestock animals.

The absence of strong evidence for transmission of *E. coli* from wild animals, predominantly in Mexico, to human hosts suggests that wild hosts are not immediate infection sources in human outbreak networks in the region. However, as our sampling of *E. coli* isolates colonizing wild animals in the region is far from exhaustive, we cannot exclude the possibility that such transmissions have occurred. The problem of nonexhaustive sampling is prevalent in genomic epidemiological studies, most of which have reported a clear genetic distinction between *E. coli* isolates found in humans, food animals, and other sources and few incidents of *E. coli* transmission between them (30–32). The problem is an issue not only of sample size but also of the breadth of sources from which they are obtained, and expanding both such that strong conclusions can be drawn from these negative results will remain a challenge.

We note that since our strains were recovered from feces, we are unable to ascertain whether the presence of pathovar-associated genes in human and livestock strains is sufficient to cause virulence, when introduced into the bloodstream in wild hosts. However, our findings are consistent with *E. coli* isolates that colonize wild-animal hosts serving as a source/sink for known pathogenic strains, serotypes, and genes. Nonhuman origins for human EPEC strains were reported in a recent large-scale genomic study, but the study examined only livestock sources (13). Our results complement these findings and suggest that evolution of virulent strains in some cases can be traced back to wild-animal sources, highlighting the role of *E. coli* host diversity in facilitating human infections. Besides virulent strains, our results demonstrate that wild animals serve as reservoirs for antibiotic resistance and virulence genes and that these can be transmitted to strains acting as human pathogens. This is in line with a recent study that also demonstrated a major role of horizontal gene transfer (HGT), through mobile genetic elements, in the recent spread of beneficial *E. coli* genes across niches in Australia (8). A more dominant role of HGT, compared with mutations arising *de novo* in the core *E. coli* genomes, was also shown in an *in vivo* study on the colonization of the mouse gut by commensal *E. coli* strains (33).

Besides the limitations in scope imposed by sample size, we did not examine the intrahost diversity of *E. coli* strains. Genetically distinct strains reside within the gut, and the genetic composition of *E. coli* genomes varies systematically across the different regions of the gut. One or two resident *E. coli* clones most often dominate the microbial community of the gut of any one individual (34), and the strains recovered from each species in our study are likely to correspond to one such clone. However, different clones may dominate in different individuals of the same species, and a broader sampling of each host species is therefore required to better understand the *E. coli*-host species interaction. Multisite sampling of individuals would also allow us to examine whether virulence genes present in the dominant clone confer any fitness advantage over other clones and whether these advantages persist across the species or even across larger taxonomic distances. Our study also neglected the degree of expression of antibiotic and virulence genes (35), which helps determine the extent to which they exert their function and which could shed light on the discordance between the presence of such genes in the genome and the absence of an evident functional activity, in terms of antibiotic resistance.

Studies on *E. coli* genomics have largely focused on pathogenic clinical strains under therapeutic conditions. However, to understand the evolutionary trajectories from commensalism to pathogenicity, we must also decipher the genetics of commensalism. Ours is one of a growing number of studies that focus on *E. coli*'s nonhuman natural habitats and that seek to describe the distribution of genes and properties in the global population of *E. coli* strains. The insights from these studies not only facilitate the diagnosis and tracking of infectious *E. coli* strains at an epidemiological level but also may help to pinpoint genetic biomarkers for pathogenicity, which are potential targets for the development of therapeutic agents.

MATERIALS AND METHODS

Strain acquisition, sequencing, and genome assembly. We acquired a systematic collection of commensal *E. coli* isolates from a previous study (5). The collection comprised 119 fecal strains from hosts belonging to 55 animal species, 31 families, and 16 orders. Of these, 110 and 9 strains were from mammals and birds, respectively. Also, 110 strains were recovered from Mexico and the rest were isolated in Venezuela and Costa Rica during the 1990s. The antimicrobial susceptibility testing was conducted on the whole collection for 8 antimicrobials clinically approved for treating *E. coli* infections, including beta-lactams (ampicillin, cefotaxime, ceftazidime, cefuroxime, and cephalothin), aminoglycosides (gentamicin and tobramycin), ciprofloxacin, and trimethoprim, as described in reference 36. The full description of the strains with metadata is available in Table S1.

DNA was extracted with the QIAxtractor (Qiagen) kit according to the manufacturer's instructions. We prepared Illumina sequencing libraries with a 450-bp insert size and performed sequencing on an Illumina HiSeq2000 sequencing machine with paired-end read lengths of 100 bp. Ninety-six samples were multiplexed to yield an average depth of coverage of ~85-fold. Reads were then assembled and improved with an automated pipeline, based on Velvet with default parameters. Assemblies were annotated with an improvement assembly and Prokka-based annotation pipeline, respectively (37–39). Details on assembly statistics, access codes for annotated assemblies, and gene annotations are available in Table S1. Roary, with the sequence identity value of 95% for orthologous groups, was used to create a pan-genome from annotated contigs (40). The Roary output file is available on the GitHub directory for the project (www.github.com/dmoradigaravand/WildHostEcoliMexico). Roary identified 24,060 genes, composed of 2,855 genes in the core genome (present in at least 99% of strains), 357 genes in the soft core genome (present in 95% to 99% of strains), 2,141 genes in the shell genome (present in 15% to 95% of strains), and 18,707 genes in the cloud genome (present in up to 15% of strains). Multilocus sequence typing was performed on assemblies using a publicly accessible typing tool and database (www.github.com/sanger-pathogens/mlst_check) with default parameter values to identify sequence type (ST) clones. Multilocus sequence typing (MLST) results are provided in Table S1. We identified phylogroups using ClermonTyping (41).

We contextualized our collection with *E. coli* strains from the environment, livestock/domesticated animals, and humans in the publicly available Enterobase data set (https://enterobase.warwick.ac.uk/). Since we were primarily interested in recent evolution and transmissions between *E. coli* in wild hosts and other hosts, we retrieved genomic data and metadata for all strains with an identical ST with at least one strain in our collection on 26 April 2020. We included only strains for which prior consent was obtained from the strain's owners. In total, genomic data for 1,868 strains were retrieved. The accession numbers and associated metadata are provided in Table S2. We then classified strains based on their source of isolation, i.e., environment, livestock/domesticated animals, and human associated. We used the above-mentioned pipeline to assemble the paired-end short reads and annotate the assemblies for external samples.

We also examined the genetic relatedness between our strains and 138,507 *E. coli* strains, available on the NCBI Pathogen Detection database (www.ncbi.nlm.nih.gov/pathogens) on 4 December 2020. We determined whether our strains fell in any SNP cluster, i.e., isolates that differ by $<$50 SNPs, with external strains from wild-animal, food animal, and human hosts. Assuming a substitution rate of two SNPs/year (42) for *E. coli* genomes, the SNP distance of 50 corresponds to 25 years. The accession numbers for these clusters, which are indicative of recent divergences, are provided in Table S1. In addition to our strains, we also examined a total of 22 external strains from wild hosts of Mexican origin in the NCBI database to detect whether they are closely related to any other strain from human or food animal hosts. The accession numbers for these strains and their origins are provided in Table S1.

Mapping, variant calling, and phylogenetic analysis. We mapped short-read sequences to the *E. coli* K-12 sequence (Biosample ID SAMN02604091), with SMALT v 0.7.4 (https://www.sanger.ac.uk/tool/smalt-0/), with a minimum score of 30 for mapping. SAMtools and BCFtools were then employed to annotate SNPs (43). SNPs at sites in which SNPs were present in less than 75% of reads were excluded. We extracted SNPs from the core-genome alignment produced by Roary and mapped them to the *E. coli* K-12 reference genome using the script available at https://github.com/sanger-pathogens/snp-sites. The SNP alignment file for the core genome is provided in the GitHub directory (www.github.com/dmoradigaravand/WildHostEcoliMexico).

Because *E. coli* genomes are too divergent to map to a single genome, we adopted an alignment-free approach to analyze the phylogenetic tree for the collection composed of wild-type and external strains. To this end, we first enumerated *k*-mers with a size of 50 from assemblies with the frequency-based substring mining (fsm-lite) package (www.github.com/nvalimak/fsm-lite). We subsequently counted the number of identical *k*-mers for pairs of isolates to produce a similarity matrix, which was then converted into a distance matrix. The distance matrix was used as input for the ape (44) and phangorn (45) packages to produce a neighbor-joining phylogenetic tree. The tree was visualized with iTOL (46) and Figtree (http://tree.bio.ed.ac.uk/software/figtree/).

Virulence factors, antimicrobial resistance gene identification, and *in silico* serotyping and LEE typing. Virulence factors and antimicrobial resistance genes were identified with the VirulenceFinder and ResFinder (both chromosomal and plasmid-borne genes) online servers (https://cge.cbs.dtu.dk/services/) using the Virulence Factor Data Base (VFDB) (47) and ResFinder database (48), respectively. We employed a loose similarity and a minimum length cutoff of 60% to ensure that divergent genes were detected. We merged the set of identified genes with those reported on the NCBI Pathogen Detection database, i.e., AMRFinderPlus, for our strains (49). We used the virulence gene panels defined in references 13 and 27 for *E. coli* pathovars to determine the pathogenicity of strains from human and food animals. We provide the list of identified resistance and virulence factors in Table S1.

The genomic context of the AMR genes was explored in two ways. First, we searched the nucleotide database to find similar annotated genomic regions with the contig that contains the resistance gene with BLASTN. Second, to further examine whether genes are located on the plasmid or the chromosome, we also utilized PlasmidSPAdes (version 3.9.0) (50) to first reconstruct plasmid assemblies and then screened the contigs for the AMR gene with BLAST, as part of the assembly graph viewer Bandage (51). We identified LEE and serotypes with the typing method in the srst2 package, using a similarity threshold of 60%. We then confirmed the presence of virulence factor genes by running BLASTN against assemblies. For the O antigens produced by the Wzy-dependent pathway, variations in the unique genes *wzx* (encoding an O-antigen flippase) and *wzy* (encoding an O-antigen polymerase) were examined (21). For the ABC transporter-dependent pathway, variations in *wzm* (encoding an O-antigen ABC transporter permease gene) and *wzt* (encoding an ABC transporter ATP-binding gene), involved in O-antigen synthesis, were studied. All the databases for typing are available in the srst2 package, and results of typing are provided in Table S1. To link serotypes with pathovars, we conducted a search in available online literature sources on 26 April 2020 to find whether the serotype has been reported in association with a pathovar. The links to sources are provided in Table S1.

Association with ecological and taxonomical attributes of host species. We obtained the tree of life for the wild host species with the R package rotl (22) and visualized the concordance between the host tree and the core genome tree of colonizing *E. coli* strains with Dendroscope (52). We used the treedist function in the ape package to compute the distance matrix from the phylogenetic tree. For *E. coli* strains, the distance matrix was obtained from pairwise Hamming distances between core genome sequences. We then used a Mantel test with 1,000 permutations as part of ade4 package (53) to assess the correlation between the distance matrices for *E. coli* genomes and that for host species. To compute the difference between the phylogenetic trees of *E. coli* strains and hosts, we used the treedist function, as part of the phangorn package in R. By doing so, we computed the square root of the sum of squares of differences in path length between each pair of tips in two trees (54). The path is defined as the number of edges within the tree that must be traversed to navigate from one tip to the other.

We dissected the relationship between virulence ability, measured as the total number of virulence genes, and ecological and physiological attributes of each host species in the panTHERIA database (55). The database includes a comprehensive species-level data set of life history and ecological and geographical traits of all known extant mammals. Spearman's rank correlation coefficient values were computed to assess the significance of the correlation between virulence gene count and attributes.

Positive selection analysis. We analyzed positive selection by reconstructing the ancestral sequence for each gene in the core genome, identified by Roary, with FastML using the general time-reversible (GTR) model as the evolutionary model for nucleotide substitution (56). Subsequently, the seqinR 1.0–2 package (57) was employed to compute the K_a and K_s values for each strain, in comparison to the ancestral sequence. We left out the strains with no synonymous changes, i.e., a K_s of 0. For functional enrichment analysis, COG (clusters of orthologous groups) categories of genes were extracted from the annotation by Prokka and assigned to functional classes. We used COG categories (functional groups) for *Escherichia coli* K-12 substrain MG1655 on https://www.ncbi.nlm.nih.gov/research/cog/. We repeated the analysis on genes that were present in 3,529 strains, which corresponds to >70% of strains. The results of positive selection analysis for these genes are detailed in Table S3.

Bayesian analysis. We constructed a Bayesian tree using BEAST (58) to date the recent mixing between *E. coli* from wild hosts and other strains in a clone in the B1 phylogroup. The clone was identified with the clustering tool in the adegenet package (59). To this end, we used the pairwise SNP distance matrix, reconstructed from multiple alignment of genomes mapped to the reference genome. We then used the gengraph function as part of the adegenet package to compute the clusters in the population. The function is based on hierarchical clustering of the pairwise SNP distance measures and involves a hyperparameter for the number of clusters. To tune the hyperparameter, we screened the SNP cutoff value for identifying clusters in the wild host and global collection and used the clustering that remained unchanged for the highest number of SNP cutoff values. As a result, the most robust clones were identified. In total, we found 90 clusters, including 14 clusters consisting of a total of 158 strains. These 14 clusters belonged to the phylogroup B1 and contained a high number, i.e., 30/119, of *E. coli* strains from wild hosts. We then extracted the genomes of these strains from the multiple alignment.

The multiple alignment encompassed 2,176 variant sites and included 128 strains from the global collection. We ran Gubbins (60) with 5 iterations to remove hypervariable sites from the genome alignment and produced a neighbor-joining phylogenetic tree. To assess the strength of the temporal signal, we plotted the root-to-tip distance versus year of isolation and performed 10,000 bootstraps with randomized years to attain a distribution for R^2 values. Subsequently, we compared the R^2 value for the data distribution with the simulated distribution. The temporal signal for the data set was stronger than 95% of signals for bootstrapped samples. We provided the SNP alignment file for the strains the clade in the GitHub directory (www.github.com/dmoradigaravand/WildHostEcoliMexico).

The multiple alignment was then used as input for BEAST. We examined a range of prior models, including a strict molecular clock and a log-normal model of a relaxed molecular clock with constant population size. Markov chain Monte Carlo (MCMC) simulations were performed three times for 50 million generations with sampling every 10 generations. A cutoff 100 was chosen for the effective sample size (ESS) of key parameters, i.e., the substitution rate, the tree root height, and the population size, for the convergence of simulations. We used TreeAnnotator v1.10.4 to aggregate trees after removing 0.2 of the tree as the burn-in phase. The 95% highest posterior interval (HPI) was used to report the certainty on ages of ancestral nodes.

Data availability. Short-read data were submitted to the European Nucleotide Archive under the BioProject accession number PRJEB23294.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, TIF file, 0.9 MB.

FIG S2, TIF file, 0.7 MB.

FIG S3, TIF file, 0.5 MB.

FIG S4, TIF file, 0.6 MB.

FIG S5, TIF file, 0.5 MB.

TABLE S1, CSV file, 0.1 MB.

TABLE S2, CSV file, 0.1 MB.

TABLE S3, CSV file, 0.2 MB.

ACKNOWLEDGMENTS

We thank Jan Michels for providing the strains.

This work was partially funded by a grant from the Centre for Antibiotic Resistance Research (CARe) at the University of Gothenburg to A.F. and grant number 2016-06503 from the Joint Programming Initiative on Antimicrobial Resistance (JPIAMR) to J.W. and A.F. This work was in part supported by the Academy of Finland (grant 313270 to V.M.). L.P. was supported by Wellcome and the Estonian Research Council (IUT34-4). D.M. was supported by the Joint Programming Initiative on Antimicrobial Resistance (JPIAMR) via MRC grant MR/R004501/1. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

1. Tenaillon O, Skurnik D, Picard B, Denamur E. 2010. The population genetics of commensal Escherichia coli. Nat Rev Microbiol 8:207–217. https://doi.org/10.1038/nrmicro2298.
2. García A, Fox JG, Besser TE. 2010. Zoonotic enterohemorrhagic escherichia coli: a One Health perspective. ILAR J 51:221–232. https://doi.org/10.1093/ilar.51.3.221.
3. Radhouani H, Silva N, Poeta P, Torres C, Correia S, Igrejas G. 2014. Potential impact of antimicrobial resistance in wildlife, environment, and human health. Front Microbiol 5:23. https://doi.org/10.3389/fmicb.2014.00023.
4. Wright GD. 2007. The antibiotic resistome: the nexus of chemical and genetic diversity. Nat Rev Microbiol 5:175–186. https://doi.org/10.1038/nrmicro1614.

5. Souza V, Rocha M, Valera A, Eguiarte LE. 1999. Genetic structure of natural populations of Escherichia coli in wild hosts on different continents. Appl Environ Microbiol 65:3373–3385. https://doi.org/10.1128/AEM.65.8.3373-3385.1999.
6. Mercat M, Clermont O, Massot M, Ruppe E, De Garine-Wichatitsky M, Miguel E, Fox HV, Cornelis D, Andremont A, Denamur E, Caron A. 2016. Escherichia coli population structure and antibiotic resistance at a buffalo/cattle interface in southern Africa. Appl Environ Microbiol 82:1459–1467. https://doi.org/10.1128/AEM.03771-15.
7. Gordon DM, Cowling A. 2003. The distribution and genetic structure of Escherichia coli in Australian vertebrates: host and geographic effects. Microbiology (Reading) 149:3575–3586. https://doi.org/10.1099/mic.0.26486-0.
8. Touchon M, Perrin A, De Sousa JAM, Vangchhia B, Burn S, O'Brien CL, Denamur E, Gordon D, Rocha EPC. 2020. Phylogenetic background and habitat drive the genetic diversification of Escherichia coli. PLoS Genet 16:e1008866. https://doi.org/10.1371/journal.pgen.1008866.
9. Martinez JL. 2009. Environmental pollution by antibiotics and by antibiotic resistance determinants. Environ Pollut 157:2893–2902. https://doi.org/10.1016/j.envpol.2009.05.051.
10. Gothwal R, Shashidhar T. 2015. Antibiotic pollution in the environment: a review. Clean Soil Air Water 43:479–489. https://doi.org/10.1002/clen.201300989.
11. Hassell JM, Ward MJ, Muloi D, Bettridge JM, Robinson TP, Kariuki S, Ogendo A, Kiiru J, Imboma T, Kang'ethe EK, Öghren EM, Williams NJ, Begon M, Woolhouse MEJ, Fèvre EM. 2019. Clinically relevant antimicrobial resistance at the wildlife–livestock–human interface in Nairobi: an epidemiological study. Lancet Planet Heal 3:e259–e269. https://doi.org/10.1016/S2542-5196(19)30083-X.
12. Sato G, Oka C, Asagi M, Ishiguro N. 1978. Detection of conjugative R plasmids conferring chloramphenicol resistance in Escherichia coli isolated from domestic and feral pigeons and crows. Zentralblatt Bakteriol Orig A 241:407–417.
13. Arimizu Y, Kirino Y, Sato MP, Uno K, Sato T, Gotoh Y, Auvray F, Brugere H, Oswald E, Mainil JG, Anklam KS, Döpfer D, Yoshino S, Ooka T, Tanizawa Y, Nakamura Y, Iguchi A, Morita-Ishihara T, Ohnishi M, Akashi K, Hayashi T, Ogura Y. 2019. Large-scale genome analysis of bovine commensal Escherichia coli reveals that bovine-adapted E. coli lineages are serving as evolutionary sources of the emergence of human intestinal pathogenic strains. Genome Res 29:1495–1505. https://doi.org/10.1101/gr.249268.119.
14. de Been M, Lanza VF, de Toro M, Scharringa J, Dohmen W, Du Y, Hu J, Lei Y, Li N, Tooming-Klunderud A, Heederik DJJ, Fluit AC, Bonten MJM, Willems RJL, de la Cruz F, van Schaik W. 2014. Dissemination of cephalosporin resistance genes between Escherichia coli strains from farm animals and humans by specific plasmid lineages. PLoS Genet 10:e1004776. https://doi.org/10.1371/journal.pgen.1004776.
15. Sarukhan J, Soberón Mainero J. 2016. Capital natural de México. Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO), Mexico City, Mexico.
16. Pallecchi L, Lucchetti C, Bartoloni A, Bartalesi F, Mantella A, Gamboa H, Carattoli A, Paradisi F, Rossolini GM. 2007. Population structure and resistance genes in antibiotic-resistant bacteria from a remote community with minimal antibiotic exposure. Antimicrob Agents Chemother 51:1179–1184. https://doi.org/10.1128/AAC.01101-06.
17. Sheppard SK, Guttman DS, Fitzgerald JR. 2018. Population genomics of bacterial host adaptation. Nat Rev Genet 19:549–565. https://doi.org/10.1038/s41576-018-0032-z.
18. Bray DB, Klepeis P. 2005. Deforestation, forest transitions, and institutions for sustainability in southeastern Mexico, 1900–2000. Environ Hist 11:195–223. https://doi.org/10.3197/096734005774434584.
19. 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.
20. Nunn CL, Jordan F, Mc-Cabe CM, Verdolin JL, Fewell JH. 2015. Infectious disease and group size: more than just a numbers game. Philos Trans R Soc Lond B Biol Sci 370:20140111. https://doi.org/10.1098/rstb.2014.0111.
21. Samuel G, Reeves P. 2003. Biosynthesis of O-antigens: Genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly. Carbohydr Res 338:2503–2519. https://doi.org/10.1016/j.carres.2003.07.009.
22. Michonneau F, Brown JW, Winter DJ. 2016. rotl: an R package to interact with the Open Tree of Life data. Methods Ecol Evol 7:1476–1481. https://doi.org/10.1111/2041-210X.12593.
23. Nagy B, Fekete PZ. 1999. Enterotoxigenic Escherichia coli (ETEC) in farm animals. Vet Res 30:259–284.
24. Borges CA, Maluta RP, Beraldo LG, Cardozo MV, Guastalli EAL, Kariyawasam S, DebRoy C, Ávila FA. 2017. Captive and free-living urban pigeons (Columba livia) from Brazil as carriers of multidrug-resistant pathogenic Escherichia coli. Vet J 219:65–67. https://doi.org/10.1016/j.tvjl.2016.12.015.
25. Schmidt MA. 2010. LEEways: tales of EPEC, ATEC and EHEC. Cell Microbiol 12:1544–1552. https://doi.org/10.1111/j.1462-5822.2010.01518.x.
26. Ingle DJ, Tauschek M, Edwards DJ, Hocking DM, Pickard DJ, Azzopardi KI, Amarasena T, Bennett-Wood V, Pearson JS, Tamboura B, Antonio M, Ochieng JB, Oundo J, Mandomando I, Qureshi S, Ramamurthy T, Hossain A, Kotloff KL, Nataro JP, Dougan G, Levine MM, Robins-Browne RM, Holt KE. 2016. Evolution of atypical enteropathogenic E. Coli by repeated acquisition of LEE pathogenicity island variants. Nat Microbiol 1:15010. https://doi.org/10.1038/nmicrobiol.2015.10.
27. Sarowska J, Futoma-Koloch B, Jama-Kmiecik A, Frej-Madrzak M, Ksiazczyk M, Bugla-Ploskonska G, Choroszy-Krol I. 2019. Virulence factors, prevalence and potential transmission of extraintestinal pathogenic Escherichia coli isolated from different sources: recent reports. Gut Pathog 11:10. https://doi.org/10.1186/s13099-019-0290-0.
28. Griffin PM, Manges AR, Johnson JR. 2012. Food-borne origins of Escherichia coli causing extraintestinal infections. Clin Infect Dis 55:712–719. https://doi.org/10.1093/cid/cis502.
29. Andersson MI. 2003. Development of the quinolones. J Antimicrob Chemother 51:1–11. https://doi.org/10.1093/jac/dkg212.
30. Ludden C, Raven KE, Jamrozy D, Gouliouris T, Blane B, Coll F, de Goffau M, Naydenova P, Horner C, Hernandez-Garcia J, Wood P, Hadjirin N, Radakovic M, Brown NM, Holmes M, Parkhill J, Peacock SJ. 2019. One Health genomic surveillance of Escherichia coli demonstrates distinct lineages and mobile genetic elements in isolates from humans versus livestock. mBio 10:e02693-18. https://doi.org/10.1128/mBio.02693-18.
31. McEwen SA, Collignon PJ. 2018. Antimicrobial resistance: a One Health perspective. Microbiol Spectr 6:ARBA-0009-2017. https://doi.org/10.1128/microbiolspec.ARBA-0009-2017.
32. Johnson TJ, Logue CM, Johnson JR, Kuskowski MA, Sherwood JS, Barnes HJ, Debroy C, Wannemuehler YM, Obata-Yasuoka M, Spanjaard L, Nolan LK. 2012. Associations between multidrug resistance, plasmid content, and virulence potential among extraintestinal pathogenic and commensal Escherichia coli from humans and poultry. Foodborne Pathog Dis 9:37–46. https://doi.org/10.1089/fpd.2011.0961.
33. Frazão N, Sousa A, Lässig M, Gordo I. 2019. Horizontal gene transfer overrides mutation in Escherichia coli colonizing the mammalian gut. Proc Natl Acad Sci U S A 116:17906–17915. https://doi.org/10.1073/pnas.1906958116.
34. Winfield MD, Groisman EA. 2003. Role of nonhost environments in the lifestyles of Salmonella and Escherichia coli. Appl Environ Microbiol 69:3687–3694. https://doi.org/10.1128/aem.69.7.3687-3694.2003.
35. Klemm P, Hancock V, Schembri MA. 2007. Mellowing out: adaptation to commensalism by Escherichia coli asymptomatic bacteriuria strain 83972. Infect Immun 75:3688–3695. https://doi.org/10.1128/IAI.01730-06.
36. Moradigaravand D, Palm M, Farewell A, Mustonen V, Warringer J, Parts L. 2018. Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data. PLoS Comput Biol 14:e1006258. https://doi.org/10.1371/journal.pcbi.1006258.
37. Page AJ, De Silva N, Hunt M, Quail MA, Parkhill J, Harris SR, Otto TD, Keane JA. 2016. Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. Microb Genom 25:e000083. https://doi.org/10.1099/mgen.0.000083.
38. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.
39. Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829. https://doi.org/10.1101/gr.074492.107.
40. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693. https://doi.org/10.1093/bioinformatics/btv421.
41. 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.
42. Lee H, Popodi E, Tang H, Foster PL. 2012. Rate and molecular spectrum

of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. Proc Natl Acad Sci U S A 109: E2774–E2783. https://doi.org/10.1073/pnas.1210309109.

43. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
44. Paradis E, Schliep K. 2019. Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35:526–528. https://doi.org/10.1093/bioinformatics/bty633.
45. Schliep KP. 2011. phangorn: phylogenetic analysis in R. Bioinformatics 27:592–593. https://doi.org/10.1093/bioinformatics/btq706.
46. 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.
47. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q. 2005. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res 33:D325–D328. https://doi.org/10.1093/nar/gki008.
48. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 67:2640–2644. https://doi.org/10.1093/jac/dks261.
49. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, Tyson GH, Zhao S, Hsu CH, McDermott PF, Tadesse DA, Morales C, Simmons M, Tillman G, Wasilenko J, Folster JP, Klimke W. 2019. Validating the AMRFINder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. Antimicrob Agents Chemother 63:e00483-19. https://doi.org/10.1128/AAC.00483-19.
50. Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A, Pevzner PA. 2016. plasmidSPAdes: assembling plasmids from whole genome sequencing data. Bioinformatics 32:3380–3387. https://doi.org/10.1093/bioinformatics/btw493.
51. Wick RR, Schultz MB, Zobel J, Holt KE. 2015. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics 31:3350–3352. https://doi.org/10.1093/bioinformatics/btv383.
52. Huson DH, Scornavacca C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol 61:1061–1067. https://doi.org/10.1093/sysbio/sys062.
53. Dray S, Dufour AB. 2007. The ade4 package: implementing the duality diagram for ecologists. J Stat Softw 22:1–20. https://doi.org/10.18637/jss.v022.i04.
54. Penny D, Hendy MD. 1985. The use of tree comparison metrics. Syst Zool 34:75–82. https://doi.org/10.2307/2413347.
55. Jones KE, Bielby J, Cardillo M, Fritz SA, O'Dell J, Orme CDL, Safi K, Sechrest W, Boakes EH, Carbone C, Connolly C, Cutts MJ, Foster JK, Grenyer R, Habib M, Plaster CA, Price SA, Rigby EA, Rist J, Teacher A, Bininda-Emonds ORP, Gittleman JL, Mace GM, Purvis A. 2009. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. Ecology 90:2648–2648. https://doi.org/10.1890/08-1494.1.
56. Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, Pupko T. 2012. FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic Acids Res 40:W580–W584. https://doi.org/10.1093/nar/gks498.
57. Charif D, Lobry JR. 2007. SeqinR 1.0–2: a contributed package to the R Project for Statistical Computing devoted to biological sequences retrieval and analysis., p 207–232. *In* Bastolla U, Porto M, Roman HE, Vendruscolo M (ed), Structural approaches to sequence evolution. Springer, New York, NY.
58. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol 10:e1003537. https://doi.org/10.1371/journal.pcbi.1003537.
59. Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405. https://doi.org/10.1093/bioinformatics/btn129.
60. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res 43:e15. https://doi.org/10.1093/nar/gku1196.
61. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41. https://doi.org/10.1186/1471-2105-4-41.