Current Research in Microbial Sciences 3 (2022) 100093

Contents lists available at ScienceDirect

Current Research in Microbial Sciences

journal homepage: www.sciencedirect.com/journal/current-research-in-microbial-sciences

Phylogeny and potential virulence of cryptic clade *Escherichia coli* species complex isolates derived from an arable field trial

Katrin MacKenzie [a], Jacqueline Marshall [b], Frank Wright [a], George Gunn [c], Nicola Holden [b,d,*]

[a] Biomathematics & Statistics Scotland, Invergowrie, Dundee DD2 5DA, Scotland, UK
[b] The James Hutton Institute, Invergowrie, Dundee DD2 5DA, Scotland, UK
[c] SRUC, Veterinary Epidemiology, An Lòchran, Inverness IV2 5NA, Scotland, UK
[d] SRUC, Department of Rural Land Use, Craibstone Estate, Aberdeen AB21 9YA, Scotland, UK

ARTICLE INFO

Keywords:
Phylogenomics
Cryptic clades
E. marmotae
Naturalised
Pan-genome

ABSTRACT

Analysis of *Escherichia coli* taxonomy has expanded into a species-complex with the identification of divergent cryptic clades. A key question is the evolutionary trajectory of these clades and their relationship to isolates of clinical or veterinary importance. Since they have some environmental association, we screened a collection of *E. coli* isolated from a long-term spring barley field trial for their presence. While most isolates clustered into the enteric-clade, four of them clustered into Clade-V, and one in Clade-IV. The Clade -V isolates shared >96% intra-clade average nucleotide sequence identity but <91% with other clades. Although pan-genomics analysis confirmed their taxonomy as Clade -V (E. marmotae), retrospective phylogroup PCR did not discriminate them correctly. Differences in metabolic and adherence gene alleles occurred in the Clade -V isolates compared to *E. coli* sensu scricto. They also encoded the bacteriophage phage-associated cyto-lethal distending toxin (CDT) and antimicrobial resistance (AMR) genes, including an ESBL, blaOXA-453. Thus, the isolate collection encompassed a genetic diversity, and included cryptic clade isolates that encode potential virulence factors. The analysis has determined the phylogenetic relationship of cryptic clade isolates with *E. coli* sensu scricto and indicates a potential for horizontal transfer of virulence factors.

Introduction

E. coli is the best-studied biological species: as a model organism in bacteriology and more widely in metabolism, as biotechnological tool and workhorse, and as a pathogen of clinical and veterinary concern. The archetypal isolate, MG1655 was the first *E. coli* to be whole genome sequenced and among the first 30 bacterial genome sequences (Blattner et al., 1997). Since then the number of bacterial genome sequences has increased to tens of thousands (Land et al., 2015), and at the time of writing there are >24,000 submitted genomes for *E. coli* (taxid:562) within the NCBI database and >124,000 assembly records within the EBI database.

Since some members of the *E. coli* species are harmful, various differentiation approaches to distinguish pathogenic from non-threatening isolates work at different levels of resolution. These range from classification for a limited set of phylogroups based the presence/absence of four genomic sequences that are not uniformly shared between isolates or pathotypes (Clermont et al., 2000), to sub-species typing based on allelic variation in a limited number of conserved sequences (< 10) (Maiden, 2006), or high resolution typing based on single nucleotide variants across the whole genome (Dallman et al., 2015). Whole genome sequencing has enhanced detection and epidemiological investigations of pathogenic isolates, although collections are dominated by clinical and veterinary pathogenic isolates.

E. coli is a diverse species and genome sequences exhibit recombination and horizontal gene transfer (Welch et al., 2002). Phylogenomics analysis has been used to determine the extent of diversity within the species, albeit constrained to the available genome sequences at the time of analysis (Touchon et al., 2009; Lukjancenko et al., 2010). A phylogroup classification revealed a far wider diversity by including larger number of isolates from different sources (Clermont et al., 2011) and identified ‘cryptic’ clades that are taxonomically divergent (Walk et al., 2009). They were validated from genomic comparisons of a small number of these novel isolates (Luo et al., 2011) and classified a

Subject category: population genomics and phylogeography

* Corresponding author at: SRUC, Department of Rural Land Use, Craibstone Estate, Aberdeen AB21 9YA, Scotland, UK
E-mail address: nicola.holden@sruc.ac.uk (N. Holden).

https://doi.org/10.1016/j.crmicr.2021.100093
Received 22 October 2021; Received in revised form 8 December 2021; Accepted 9 December 2021
Available online 20 December 2021

potential new species (Liu et al., 2015). The clades are genetically distinct from the 'enteric' clade, *E. coli* sensu stricto, which contains the vast majority of sequenced isolates. The cryptic clade isolates also appear to be ecologically distinct and have been sourced from environmental habitats. Although there appears to be an environmental association for cryptic clade, there is a paucity of sequence data for isolates within the *E. coli* species complex from environmental sources, increasing uncertainty for any correlations.

We previously isolated *E. coli* from an arable field trial site that was amended with bulky fertilisers: bovine slurry or green compost, at a sampling prevalence of ~ 10%. Surprisingly, there was a lack of correlation between *E. coli* isolated from soil or barley plants with isolates sourced from the fertiliser inputs, showing that these isolates were genetically distinct from those associated with farm animals, and potentially environmentally-persistent (Holden et al., 2013). This is supported by the occurrence of environmentally-adapted isolates elsewhere, e.g. in freshwater beaches (Walk et al., 2007) or in soil (Brennan et al., 2010), supporting the hypothesis that some *E. coli* isolates are naturalised to the wider environment and not necessarily mammalian-derived. A key question is the evolutionary history and trajectory of these isolates and their relationship to isolates of clinical/veterinary importance {Walk et al., 2009, Cryptic lineages of the genus Escherichia}. Therefore, the aim was to determine whether there were obvious genetic and functional components that distinguished divergent isolates from *E. coli* sensu sricto with respect to wider collections, including the Enterobase catalogue (Zhou et al., 2020; Zhou et al., 2020). We took a bioinformatics approach for genomic analysis, using tools to determination of phylogeny from unfinished genome sequences. Iterative analyzes enabled refinement of the pipeline and selection of the appropriate reference scaffolds, allowing functional predictions to be made for gene content.

Materials and methods

Isolate sequencing and bioinformatics analysis

E. coli isolates were whole genome sequenced by Illumina short-read technology, from paired-end libraries, as described previously (Lupolova, 2017) (Table S1). Genome data for the isolates has been deposited in the European Nucleotide Archive (ENA - http://www.ebi.ac.uk/ena) under the Study number PRJEB22793 for four of the isolates (Isolate name, accession numbers: 5002, ERS1942968; 5034, ERS1942966; 5035, ERS1942967; 5088, ERS1942965), and the remainder of the isolates that were first published elsewhere (Lupolova et al., 2017), deposited in ENA (project number PRJEB22630) and in Edinburgh Datashare (https://doi.org/10.7488/ds/2102). All accession numbers and metadata for environmental isolates in this study are listed in Table S1a and reference isolates (Table S1b).

Sequence read files were quality checked and assembled against *E. coli* isolate Sakai (BA000007) in the first instance and following an initial round of alignment using Phyla-AMPHORA to determine clade/sub-clade positioning, were re-assembled against a more appropriate reference (Table S1b). The genomes were processed by Phyla-AMPHORA (Wang and Wu, 2013), using the contigs to obtain translated ORFs, which generated 295 suitable loci for comparison using Hidden Markov Models (HMM). The references genomes, obtained from Genbank, were similarly searched. The loci were aligned using ClustalOmega for the combined total of 59 genomes (Sievers et al., 2011). Maximum-likelihood trees (PhyML (Guindon et al., 2010)) were estimated for each locus and, using the sum of the branch lengths for the trees, the alignments were checked for errors by visual inspection of those with large values for this sum, excluding one alignment. The alignments for the remaining 293 loci were concatenated and a tree estimated from this using IQ-Tree (Nguyen et al., 2015). Alignment generation, tree estimation for individual loci and a concatenated alignment were automated. The coordinates for the loci were used in the tree estimation to allow a separate model of evolution to be used for each, with ultrafast bootstrap approximation was used (Minh et al., 2013). Although the evolutionary model is allowed to vary across the loci, it was assumed that the tree topology is the same for each of the 293 loci. The 293 loci were treated as separate partitions and independent evolutionary rates were estimated for each (Chernomor et al., 2016). The proportion of variable sites across each of the 293 loci was calculated as the sum of sites with at least two copies of at least two nucleotides divided by the sequence length. Average nucleotide identity (ANI) was calculated using the calculator tool on the www.enve-omics.ce.gatech.edu website (Rodriguez-R and Konstantinidis, 2014). All the genomes were compared to reference sequences from isolates Sakai, SMS-3–5, TW15838, TW09307 and TW09276 (Table S1). Pan-genome analysis was carried out for all of the barley trial isolates together and separately for Clade-V (C-V) isolates selected from Enterobase (Table S2) based on a Hierarchical Clustering threshold of HC2350 (Zhou et al., 2020). Sequences were first annotated by Prokka (Seemann, 2014) and the CDS output used with the Roary tool (Page et al., 2015) to generate presence/absence matrices. The files were exported into MS Excel for filtering and searching via gene accession numbers. Maximum likelihood phylogenetic trees were visualised in FigTree (http://tree.bio.ed.ac.uk/software/figtree/) or the Enterobase tool GrapeTree (Zhou et al., 2018). PCR detection for enteric clade phylogroups or cryptic clades was done according to the published methods (Clermont et al., 2011; Clermont et al., 2013), using the recommended primer sequences and PCR reaction conditions.

Functional gene detection

Functional ontology analysis was applied to the Roary gene difference output for C-V 'vs' the enteric-clade isolates. C-V unique genes were filtered and the annotated genes searched for functional groups using the Panther classification system (Mi et al., 2013) against the *E. coli* reference gene set ($n = 4306$). Enriched functional groups were identified from a Fisher statistical over-representation test and orthologolous gene lists generated. Orthologues of the C-V and C-IV isolate fimbrial gene clusters were identified by BLASTn analysis using the *E. coli* Sakai loc2, loc5 and loc10 gene cluster reference sequences (Low et al., 2006).

The presence mobile elements was screened from whole genome fasta files using on-line tools for plasmids with PlasmidFinder (Carattoli et al., 2014) and phage with PHASTER (Arndt et al., 2016). Plasmid replication origins were confirmed by end-point PCR detection with published primers (Carattoli et al., 2005). The cdt gene cluster was detected with primers that encompassed cdtA (5′-GGGATCGGTGATTCACCTT) to cdtC (5′-GGAGACATTATTGCCGGAGATA), generating a 2031 bp product. Antimicrobial resistance gene detection used the ARBicate tool in Galaxy, which combines databases (card; argannot; resfinder; ncbi; ecoli_vf; plasmidfinder; vfdb; ecoh) (https://github.com/tseemann/abricate) and point mutations that confer resistance with ResFinder (Bortolaia et al., 2020). CRISPR sequences were detected using the CRISPRCasFinder tool (Couvin et al., 2018).

Functional assays in vitro

For all microbiology assays, bacterial cultures were grown from saturated cultures in LB medium at 37 °C, diluted in the relevant medium and incubated at the relevant temperature. At least two independent experimental repeats were made for each. Growth rates were determined from changes in optical density of cultures over time, from two replicated cultures. Bacteria were sub-cultured in fresh LB medium at a dilution of 1:100 in 10 ml and incubated at 18 °C. Samples were taken at set time periods and the cell density measured at 600 nm. Growth rates (cell density changes per minute) were estimated from the DMFit model (Baranyi and Roberts, 1994) as an Excel add-in (via the ComBase website: https://www.combase.cc/index.php/en/). Motility was assessed by the distance a colony front travelled over semi-solid agar. Bacteria cell density was adjusted to OD600 of 1.0 and ~ 1 × 105 cfu stabbed into LB medium containing 0.25% agar, and incubated at 37 °C for 16 h. The diameter of the colony was measured and the average taken from four independent colonies.

Biofilm formation was assessed from static cultures grown in polystyrene multi-well plates. Bacteria cell density was adjusted to OD600 of 0.02 in fresh LB medium and 200 μl (~4 × 107 cfu) were transferred to wells in a 96-well microtitre plate (4 replicate samples of each) and incubated at 37 °C for ~18 h. Non-adherent cells were removed and adherent cells remaining in the wells washed with PBS, stained with 1% crystal violet for 15 min at room temperature, which was solubilised in 80:20% ethanol: acetone. The amount of retained dye was measured in a spectrophotometer at 560 nm. Identification of the isolates based on utilisation of 12 substrates used the Gram Negative Bacilli 12E multiwell pate format (Microbact™, Oxoid, Thermo Scientific), according to the manufactures instructions. The Microbact™ computer aided identification package was used to assess the results, with the closest match of organism name provided from the percentage share of the probability for that organism as a part of the total probabilities for all choices.

Results

Phylogenomics analysis

Whole genome sequences were determined previously for a set of *E. coli* isolates (42) with diverse multi-locus sequence types (MLST), sourced from an arable field trial (Holden et al., 2013) from soil ($n = 13$), barley roots ($n = 15$), clover roots ($n = 3$), barley grain ($n = 6$), compost ($n = 2$), bovine slurry ($n = 2$) and lettuce root ($n = 1$) (Table S1a). Their phylogeny was determined from a gene alignment approach using iterative rounds of selection with Phyla-AMPHORA (Wu and Eisen, 2008; Wu and Scott, 2012). A total of 59 isolate genomes comprising 37 of the environmental isolates plus 22 reference genomes selected for genetic diversity (Table S1b) generated 293 unique locus markers for the final phylogeny. An iterative approach increased the number of common genes for the alignment by >3.5-fold and aided in the identification of (presumptive) false positive isolates. Of the 293 loci included, 29 had no variation between the isolates and for another 170, the level of phylogenetic information was less than 5%. An alignment was generated for each translated locus using ClustalOmega (Sievers et al., 2011) and a phylogenetic tree was estimated by concatenating the 293 individual alignments (Fig. 1). This clustered the majority of the environmental isolates within the 'enteric' clade. Four divergent isolates clustered with the sequenced environmental C-V isolate TW09308 (Luo et al., 2011), JHI_5040, JHI_5085, JHI_5133 and JHI_5041. Bootstrap support for these clades was 100%. Those in the enteric clade clustered to a range of reference isolates, including the pathogenic isolates 2011C_3493 belonging to serogroup O104:H4 (JHI_5034, JHI_5067, JHI_5077) and CFT073 (JHI_5075, JHI_5079), as well as the archetypal K-12 isolate MG1655 (5025). There was indication of five pairs/trios based on low or no genetic distances: JHI_5040 & JHI_5085; JHI_5042 & JHI_5084; JHI_5137, JHI_5138 & 5139; JHI_5009 & JHI_5036; JHI_5067 & JHI_5077.

Average nucleotide identity (ANI) showed that the C-V isolates shared degree of similarity of less than 91% with those in the other clades, but >96% within the same clade (Supplementary Fig. 1). Since genomes from members of the same population share high sequence identity, greater than 94%, it confirms placement of the cryptic C-V isolates (Konstantinidis and Tiedje, 2007). In comparison, E. albertii resulted in ANI values that were <90% for the other non-E. albertii isolates.

PCR validation of clade status

A PCR-based method for cryptic clade detection has been reported (Clermont et al., 2011), based on variations within aes or chuA loci. The phylogrouping PCR (Clermont et al., 2013) with the quadraplex primer set placed three of the isolates (JHI_5040, JHI_5041, JHI_5133) as potential cryptic clade isolates with a positive signal for chuA (476 bp) (Suppl. Fig. 2), supporting the phylogenomics data. Isolate JHI_5085, which has similar genotype to isolate JHI_5040 (Fig. 1) was not assigned to the same category. Furthermore, the singleplex PCR for C-V yielded products of the same size as those in the enteric clade isolate Sakai, i.e. arp and chuA 288 bp, placing isolate JHI_5085 as phylogroup D or E. The presence of other, potentially cross-reactive or non-specific bands indicated that the phylogroup primers were insufficient to accurately discriminate these isolates phylogenetically.

Genomic complement of clade-IV and clade-V isolates

The genomic complement for the environmental isolates identified in C-V was performed using a presence/absence matrix approach from the pan genome (Page et al., 2015), supplemented with additional environmental isolates and selected C-IV and C-V and enteric-clade reference isolates from Enterobase (Table S1b). The clustering of the 293-locus alignment was retained, with three distinct groups comprising the enteric-clade reference isolates, the C-III (TW09231) and C-IV (TW14182, JHI_5145) isolates, and the C-V isolates (Fig. 2). Sub-divisions in the core genome were evident for the C-III & C-IV cluster and the enteric cluster (Fig. 2A). Alignment of the accessory genomes based on the presence/absence complement formed two main branches, placing the C-III and C-IV isolates together with the enteric isolates, while the other branch comprising the C-V isolates (Fig. 2B). The C-V environmental isolates (JHI_5041, JHI_5133, JHI_5040, JHI_5085) clustered with isolates designated as E. marmotae (HT073016, B116), and JHI_5145 isolate clustered with C-IV isolates.

The pan-genome was expanded to place the four environmental C-V isolates into a wider context alongside isolates designated as C-V from those in the Enterobase database (an additional 108) based on their hierarchical clustering group (Zhou et al., 2020) (Table S1c). This showed the degree of core genome similarity between isolates JHI_5038, JHI_5040 and JHI_5085 with TW14263 (Fig. 3A), which was retained, to a lesser extent in the accessory binary tree (Fig. 3B). The analysis showed that isolate JHI_5038 appeared to be clonal with JHI_5085 as they clustered together. The source-of-isolation metadata had no apparent impact on phylogenetic clustering for the wider group of C-V isolates ($n = 112$) (Fig. 3C). The analysis also aided downstream functional assessment.

Functional assessment of environmental E. coli isolates

Assessment for potential functions based on the most variable common 293 loci were almost entirely associated with metabolic functions and included three loci involved in folate biosynthesis and binding: FolA (P0ABQ4), FolK (P26281), YgfZ (P0ADE8). Other metabolic-associated loci included GtnX (P46846), TilS (P52097), MutT (P08337), LpxK (P27300) and the DNA processing protein Smf (P30852). Since this indicated divergence in metabolic potential, potential functions in the accessory genome of the environmental C-V isolates (JHI_5040, JHI_5041, JHI_5085, JHI_5133) were assessed. Annotated genes in the C-V environmental isolates with different sequence alleles in comparison with the enteric-clade isolates were used to identify enriched functional gene ontology groups, against the *E. coli* reference gene list. The two functional groups that were significantly enriched were in 'Cell Projection in Cellular Components' (GO:0,042,995) increased by 3.87-fold ($n = 16$ genes; $p = 0.0000267$) and 'Catalytic Activity in Molecular Functions' (GO:0,003,824) increased by 1.28-fold ($n = 199$ genes; $p = 0.0000245$).

The catalytic activity genes included a range of metabolism-associated gene functions (Table S1d) and included some contiguous/operon gene clusters. 21 sequence-type divergent environmental

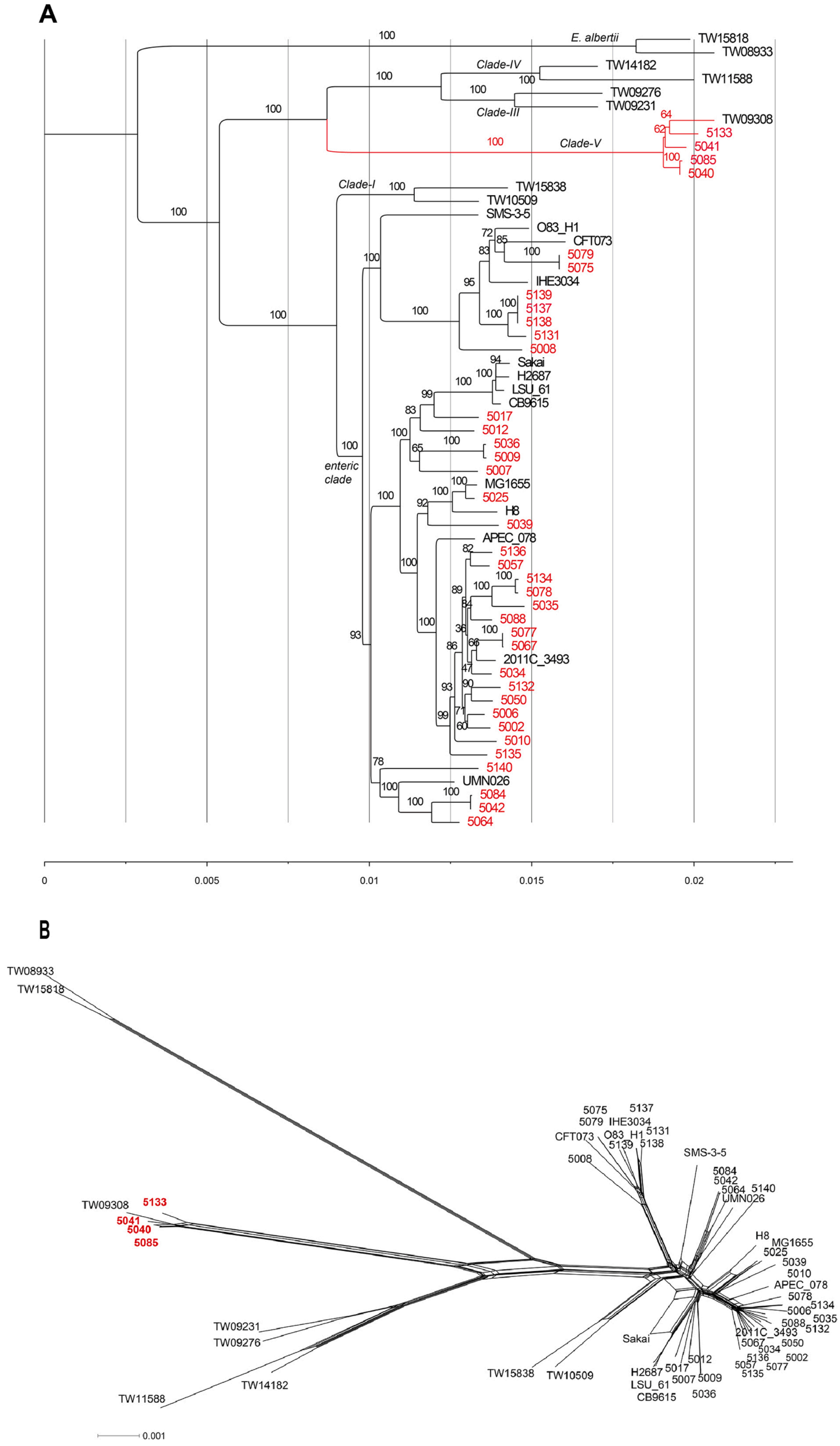


Fig. 1. Phylogenetic tree based on alignment of 293 core genes from 59 Escherichia spp. genomes. (A) 293-locus MLST tree (partition for each locus) with *E. coli* clades indicated and an outgroup of two E. albertii isolates. Isolate names for the arable field trial collection are labelled in red font, with reference isolates in black, and the C-V branch labelled red. Bootstrap values are adjacent to branches and the branch length scale axis indicated. The tree was mid-point rooted and annotated in FigTree (Andrew Rambaut). (B) Neighbour-net tree of the same group, annotated in the same way (Splits-tree).

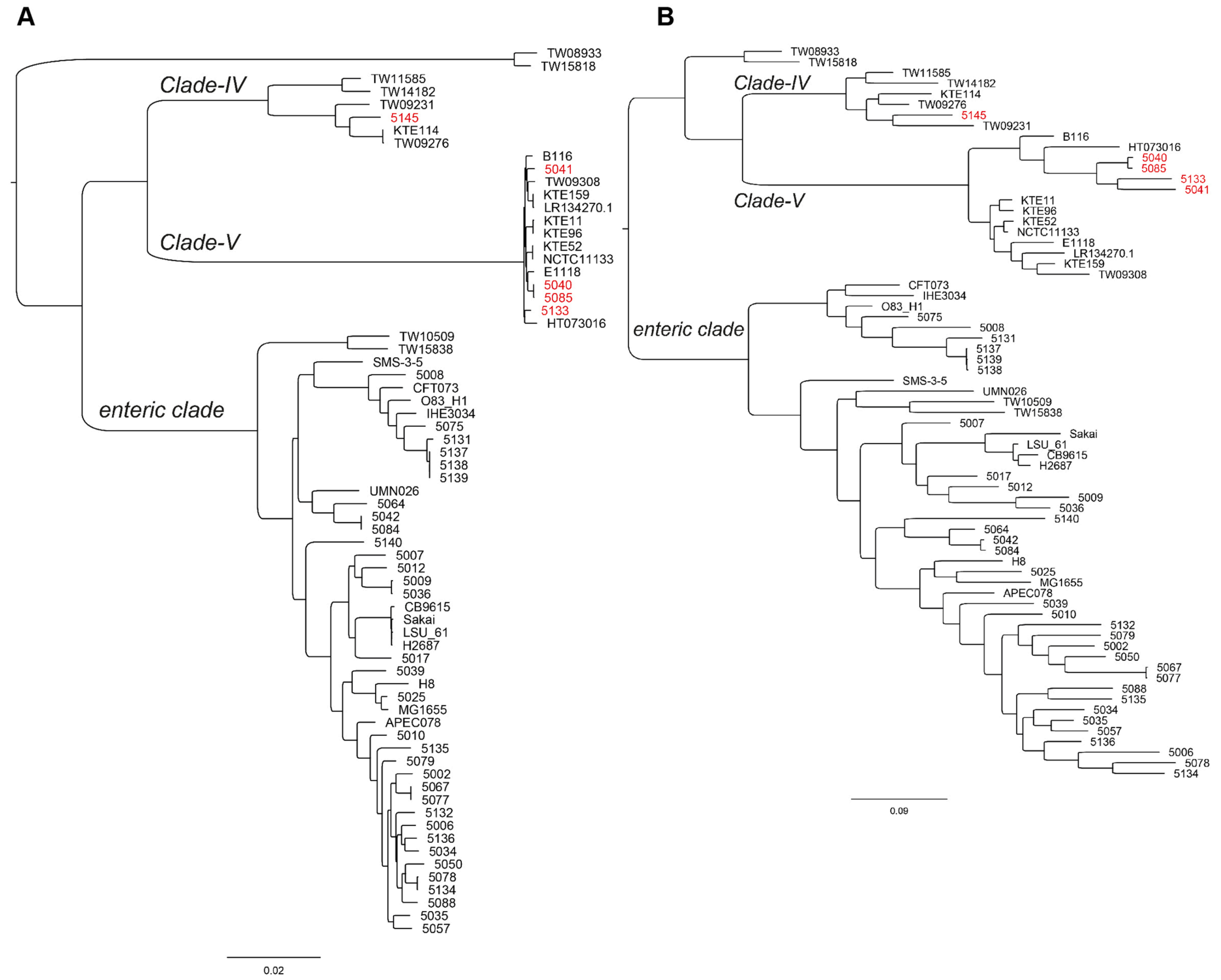


Fig. 2. Low Pilmore isolates pan-genore core & accessory phylogenies. The core (A) and accessory (presence/absence binary) (B) genomes phylogenetic trees of the environmental *E. coli* isolates with representative 'reference' isolates selected from Enterobase (FigTree). Shaded parts identify the cryptic clade clusters and red font used to identify isolates in cryptic clade V (JHI_5040, JHI_5041, JHI_5085, JHI_5133) and IV (JHI_5145). All other isolates and reference strains are within the enteric-clade, cryptic C-I or an outgroup (E. fergusonii).

isolates including three C-V isolates were tested on their ability to utilise 12 different substrates that are often employed for identification of Gram-negative bacilli. The C-V isolates showed a typical *E. coli* metabolic profile for these substrates and generated the same percentage probability of *E. coli* speciation as the reference isolate Sakai (Table S1e). C-V genes for lysine and glucose metabolism that were enriched in the Catalytic Activities functional group, correlated with lysine decarboxylase and acid from glucose substrate utilisation. Growth kinetics were determined at an environmental-relevant temperature and showed that three of the C-V isolates (JHI_5041, JHI_5085, JHI_5133) had faster maximal growth rates than *E. coli* isolate Sakai, in rich, undefined (LB) medium at 18 °C (Fig. 4a,b).

The cell projection genes were mostly composed of multiple genes within gene clusters, included six fimbrial loci (Table S1e). Although these were annotated as orthologues to fimbrial clusters in the reference sequence (isolate Sakai), alignment at the nucleotide level showed substantial divergence. Orthologues of fimbrial loci loc2, loc5 and loc10 of the C-V isolate showed only partial homology that was limited to the chaperone and usher subunits, with only near-complete coverage to the loc10 usher CDS (average 66.1% nt identity for 2540/2640 nt). In contrast, the C-IV isolate (5145) encoded a complete and highly-conserved loc5 gene cluster with 99.8% overlap over 6751 nt at 86.7% nt identity across the cluster, compared to the reference (*E. coli* Sakai). Motility and biofilm formation were tested for the C-V isolates and compared to two enteric-clade isolates (Sakai, JHI_5025). Isolate JHI_5085 produced the greatest extent of biofilm formation and was the most motile, while enteric-clade isolates (Sakai and JHI_5025) exhibited high and low extremes for both phenotypes (Fig. 4). Isolates of the same C-V genotype (JHI_5040 & JHI_5085) exhibited differences in motility.

Mobile elements in the environmental clade-V isolates

The presence of bacteriophage and plasmids was determined in the C-V and C-IV isolates, together with genes for antimicrobial resistance and CRISPR-Cas systems (Table 1). The antimicrobial resistance gene mdfA, a multidrug efflux pump commonly associated with *E. coli* (Edgar and Bibi, 1997), was encoded by all C-V and the C-IV isolate with an intact CDS despite some SNPs and a 9 bp deletion in isolate JHI_5133 (Supplementary Fig. 3). Partial sequence for a β-lactam (ESBL) gene, blaOXA-453 (accession number KR061507) was detected in one of the

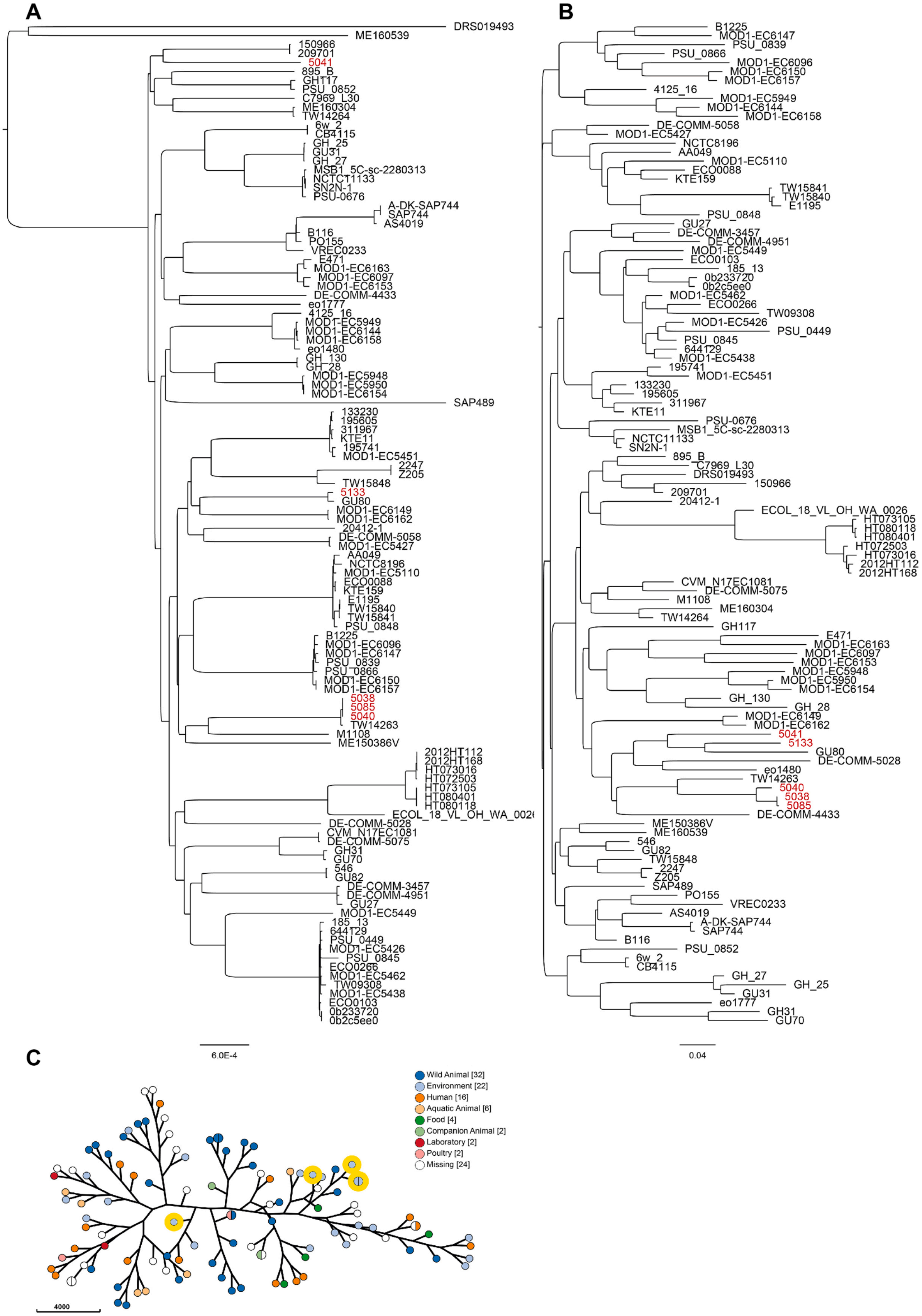
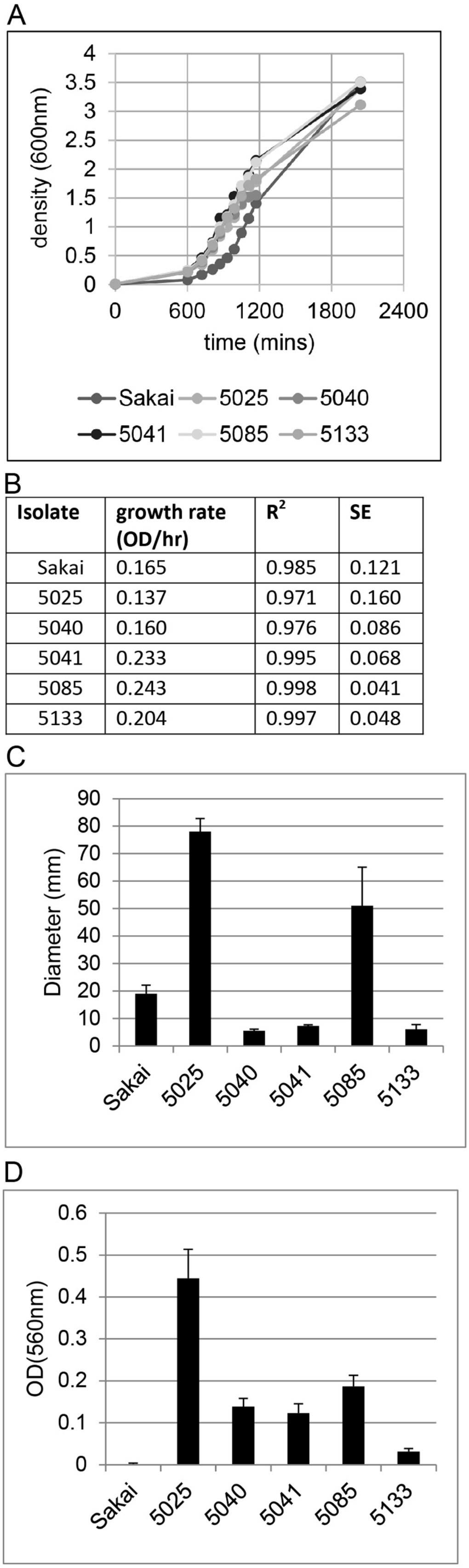


Fig. 3. Core and accessory genomes of selected cryptic Clade-V isolates. Phylogenetic trees (maximum-likelihood) of the core (A) and accessory (presence/absence binary) (B) genomes of the environmental C-V isolates with a selection of published C-V genomes from Enterobase (FigTree). C-V isolates (JHI_5038, JHI_5040, JHI_5041, JHI_5085, JHI_5133) are indicated in red font. The phylogeny was overlaid with source-of-isolation metadata (C) in a GrapeTree representation, with the environmental Low Pilmore isolates highlighted in yellow.



Isolate	growth rate (OD/hr)	R^2	SE
Sakai	0.165	0.985	0.121
5025	0.137	0.971	0.160
5040	0.160	0.976	0.086
5041	0.233	0.995	0.068
5085	0.243	0.998	0.041
5133	0.204	0.997	0.048

(caption on next column)

Fig. 4. Growth, motility and biofilm characteristics of Clade-V isolates. Growth (A) of the isolates in LB medium at 18 °C with aeration was measured from cell density (optical density at 600 nm) over time and (B) rates were estimated from a DMFit model and expressed as changes in density (OD600) per hour. The R2 and standard error (SE) of the model fit is provided. Isolates Sakai and 5025 were included as enteric-clade comparators. Motility (C) measured as colony diameter, and biofilm formation (D) measured from amount of adherent cells that retained crystal violet dye, were assessed from cultures grown in LB medium at 37 °C. The average and standard deviation are charted. In all cases, the reference isolate *E. coli* Sakai and enteric-clade isolate 5025 were used as comparators.

isolates (JHI_5085). The sequence was identical to >100 Campylobacter sp. sequences along the entire contig node (433 nt), annotated as the OXA-61 family. Isolate JHI_5085 contains the same ‘G’ SNP variant within the TATA -10 promoter box that represses expression in C. jejuni type strain NCTC 11,168 (Zeng et al., 2014). Orthologues of blaOXA-453 allele from isolate JHI_5085 were restricted to Campylobacter isolates and were not present in other genomes in the Escherichia genus. In comparison, seven of the additional 108 *E. coli* C-V isolates catalogued in Enterobase encoded at least one ESBL, and one encoded multiple non-ESBL antibiotic resistance genes (ARG) (Table S1c).

Plasmid DNA sequence was detected in two of the four C-V isolates: IncFIB in both JHI_5041 and JHI_5133, and IncFII and IncFIC present in each, respectively (Table 1). PCR screening for plasmid ori DNA supported presence of the IncFIB replicon in the JHI_5041 genome, and detected replicon A in isolate JHI_5085. Phage DNA sequence was detected in all C-V and C-IV isolates, each with six to nine prophage regions, detected by PHASTER analysis (Arndt et al., 2016) (Table 1). All of the C-V isolates encoded a CRISPR-Cas system, with the Class I Cas cluster (Makarova et al., 2011), although the C-IV isolate had no detectable CRISPR system beyond expected evidence level thresholds (Couvin et al., 2018).

All four of the C-V isolates encoded a phage-associated cyto-lethal distending toxin (CDT/CLDT) gene cluster. Partial prophage alignment to bacteriophage CdtI (accession AB285204.1) occurred for the 5′-end and to bacteriophage Lambda_2G7b (accession NC_049954) at the 3′-end. However, the sequence did align completely to some of C-V (E. marmotae) isolates (Fig. 5A) including the cdt gene cluster, although one C-V isolate contained an addition CDS (isolate NCTC8196, accession LR134270) (Table S1f). The sequence also aligned with sequence from *E. coli* strain 2009C-3133, a non-O157 STEC associated with clinical disease (Lindsey et al., 2015), although many STEC encode phage-associated cdt gene clusters (Janka et al., 2003). 33 of additional 108 C-V sequences from Enterobase also encoded the cdt cluster, but with no obvious association of the phylogenetic cluster (Fig. 5C,D).

Discussion

Phylogenetic analysis has identified environmental Escherichia species isolates that are genetically placed within the cryptic clades of the *E. coli* species complex. Their prevalence for this set within the species complex is in-line with that reported previously of ~ 10% (Clermont et al., 2011). Although presumptive identification and traditional phenotypic and biochemical approaches placed the isolates as *E. coli*, phylogenomics analysis identifies them as cryptic clade C-V (E. marmotae) and C-IV {Kim et al., 2014 #3279;Richter and Rosselló-Móra, 2009 #3280;Walk et al., 2009, Cryptic lineages of the genus Escherichia}. The incomplete classification based on PCR-detection was likely a reflection on the paucity of sequence data for isolates from environmental sources, compared to pathogenic (clinical & veterinary) or animal sources, from which to base specific amplicons (Clermont et al., 2011; Luo et al., 2011).

Multiple approaches are available for determination of phylogeny from unfinished genome sequences. The Phyla-AMPHORA approach was used as a rapid and semi-automated determination for diverse

Table 1
Plasmid and Phage complement of C-V and C-IV isolates.

Isolate	Clade	Plasmids: origin PCR*/ genome hits	ARG	Phage	CRISPR-Cas
JHI_5040	V	IB/No hits	mdfA	6 regions: 2 intact, 2 incomplete, 2 questionable	Class 1 cluster (EL = 4)
JHI_5041	V	IB/IncFIB; IncFII	mdfA	7 regions: 5 intact, 2 incomplete	Class 1 cluster (EL = 4)
JHI_5085	V	A0/No hits	blaOXA-453; mdfA	9 regions: 3 intact, 4 incomplete, 2 questionable	Class 1 cluster (EL = 4)
JHI_5133	V	NT/IncFIB; IncFIC	mdfA	8 regions: 3 intact, 3 incomplete, 2 questionable	Class 1 cluster (EL = 4,3)
JHI_5145	IV	NT/Col156; IncFII; IncI1-I	mdfA	7 regions: 5 intact, 2 incomplete	ND

* Presence of plasmid- and phage-associated genes detected from computational screening of bacteria genomes, except for plasmid origins that were detected from PCR screening.

NT not tested

ND not detected

EL 'Evidence level' relating to CRISPR spacing (Couvin et al., 2018)

ARG AMR genes.

taxonomies (Wu and Eisen, 2008; Wu and Scott, 2012; Wang and Wu, 2013), with 293 loci sufficient to discriminate this group to the sub-species level, in keeping with other whole genome or core MLST (wgMLST, cMLST) (Maiden et al., 2013). Although the loci are defined as 'core', they may be present on shared accessory elements and as such the term is limited to only this set of genomes. Further, isolates that were excluded from the selection on the basis of identical loci may not be truly clonal and encode differences elsewhere in the genome. Subsequent pan-genome analysis for the C-V isolates generated a different set of core and accessory genomes when combined with the wider set of isolates from Enterobase (Zhou et al., 2020). Use of the pipeline in an iterative manner enabled refinement based on more appropriate reference scaffolds.

The *E. coli* species complex is characterised by its metabolic flexibility and as a mesophile, by its ability to grow and persist under an array of physio-chemcio environments. The catabolite profiles of the cryptic clade isolates were indistinguishable from others classed as *E. coli*, in-line with previous reports (Walk et al., 2009). This raises questions about current systematic classifications within the *E. coli* species complex, and highlights the need to link ecological and phenotypic data to genome sequences to help understand the functional role of any organism within its environment (Garrity and Craft, 2016). Functional ontology searches highlighted orthologous genes for fimbrial gene clusters in the C-V isolates, indicating host adherence as a potential functional difference. The high level of divergence of the loc2, loc5 and loc10 orthologues from the annotated genes in the enteric-clade reference isolate (Low et al., 2006) suggests differences in target binding substrates and could indicate novel adherence gene clusters. Furthermore, differences in motility and biofilm formation overlaid the genetic diversity within the set of four C-V isolates, indicative of different regulatory networks (Martinez-Antonio et al., 2008), consistent with enhanced biofilm formation observed with other C-V isolates (Ingle et al., 2011). Such differences reflected temperature-dependant expression and functional binding of a near-ubiquitous adherence factor, type 1 fimbriae, within the field trial isolates dataset (Marshall et al., 2016).

Partial sequence of the ESBL gene blaOXA-453 was detected in the C-V isolate JHI_5085, which was derived from the rhizosphere of a barley plant with no apparent link to (large) animals, as it occurred a control, inorganic N-fertiliser treated plot in a long-term experimental barley trial (Holden et al., 2013). blaOXA-453 is associated with Camplyobacter sp. (accession number KR061507), suggesting transmission via a broad-host range mechanism (Partridge et al., 2018). The C-V isolate JHI_5085 shared the 'G' substitution in the promoter region with C. jejuni type strain NCTC 11,168 that supresses expression (Zeng et al., 2014), in-line with C-V isolate JHI_5085 sensitivity to ampicillin (Holden et al., 2013). AMR genes were detected in eight additional Enterobase C-V isolates, with at least one ESBL in seven of the isolate genomes, including blaKPC-2 in ESC_KA9267AA that is normally associated with Klebsiella. Their presence is indicative of horizontal gene transfer of β-lactamases within the Proteobacteria (Ellabaan et al., 2021). Plasmid sequences were not detected in the short-read sequence of C-V isolate JHI_5085, and PCR screening did not detect replicons for 18 Inc-type plasmids (Holden et al., 2013), which may indicate chromosomal integration or lack of detection. The closely related strain (JHI_5040) did not contain the blaOXA-453 sequence, but an Inc FIB plasmid replicon was detected by PCR.

Multiple prophage sequences were detected in the C-V isolates along with anti-phage systems, including CRISPR-Cas (Class I) systems. Prophage sequences sometimes carry toxin genes and all four of the C-V isolates and 34 of the additional 108 C-V Enterobase isolates encoded CDT apparently embedded in prophage sequences. However, a complete phage sequence match did not occur with current repositories and may indicate an un-recorded phage. CDT is an intracellular genotoxin that initiates a cell cycle arrest at the G(2)/M stage prior to mitosis (Smith and Bayles, 2006). It is normally associated with diarrheagenic *E. coli* and E. albertii, but has also been detected in non-clinical isolates (Allué-Guardia et al., 2011), including C-V *E. coli* isolates (Ingle et al., 2011) and avian pathogenic E. albertii (reclassified from *E. coli*) (Foster et al., 1998). CDT is present in other genera and it is possible that the C-V isolates acquired the cdt-phage through horizontal transfer from non-Escherichia sources. The presence of partial sequence matches with sequence surrounding the cdt cluster in C-V isolates (E. marmotae) and a STEC *E. coli* (Lindsey et al., 2015) is notable given the different geographical regions of isolation, and indicates common specificity in this group of isolates for phage recognition.

The environmental transfer of AMR and toxins on mobile genetic elements is a key risk question relevant to the One Health agenda. ESBL transfer between isolates clearly occurs but appears to be restricted, as found from comparison of genetically distinct livestock-associated and bloodstream *E. coli* isolates that encoded identical ESBL genes but had different surrounding mobile elements (Ludden et al., 2019). In a separate study for plasmid and ESBL carriage for *E. coli*, human-human transmission was found to be the principle source, followed by food (Mughini-Gras et al., 2019). Transfer of CDT on bacteriophage has been less well studied, and although CDT occurs in a range of Gram negative genera, it is not ubiquitous within a species, with apparent genotype association, e.g. for the STEC group (Janka et al., 2003). The lack of any clear phylogenetic cluster of niche association with the C-V isolates may reflect genetic diversity within this group. Yet the functional role and transferability of this toxin warrants further study.

Isolation of *E. coli* from non-animal sources in the wider environment has led to the concept of naturalisation of *E. coli* to environmental habitats. The current set of available C-V isolate genomes is relatively limited and hampered by lack of metadata, so that it is not yet possible to make definitive statements about the original source of these isolates, and whether they pose any threat to human health, as for other naturalised *E. coli* (Ishii and Sadowsky, 2008; Ishii et al., 2009; Jang et al., 2017). Large-scale genomic epidemiology will help to define the taxonomy and evolutionary trajectory of the *E. coli* species complex and is required to identify reservoir jumps for Escherichia spp. (Mills and Lee, 2019) coupled with longitudinal studies to witness such jumps.

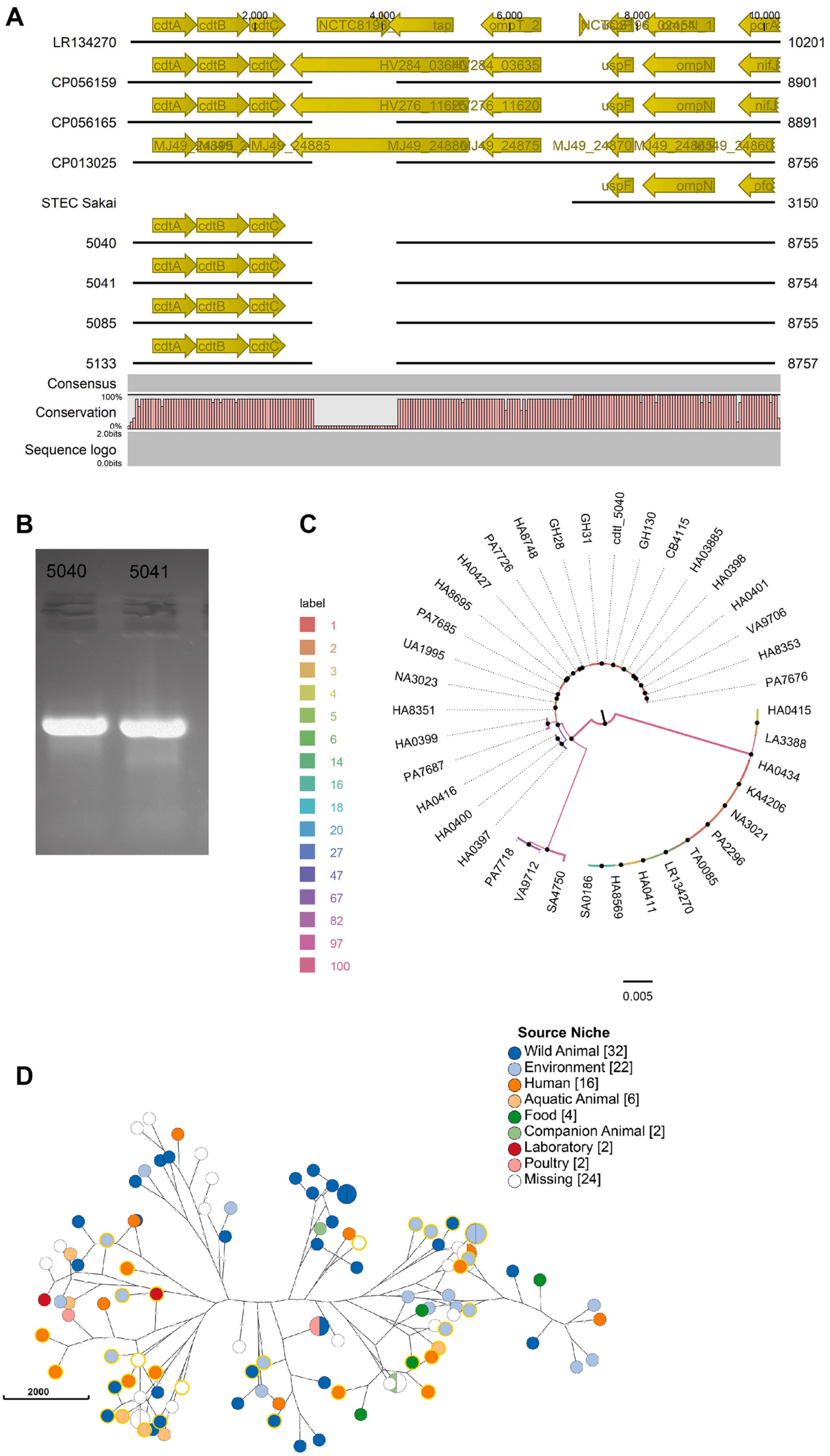


Fig. 5. The cdt locus of the Low Pilmore C-V isolates. Alignment of cdt loci and extended genomic location against selected other C-V isolates and STEC isolate Sakai (cdt-) (A) and PCR detection of cdt locus (2031 bp) in isolates JHI_5040 and JHI_5041 (B). Phylogeny of 112 C-V isolates include isolate JHI_5040 (C), overlaid with source-of-isolation metadata (GrapeTree representation), with cdt locus positive genomes highlighted in yellow.

Funding information

The work done by KMcK, JM, FW, GG, and NH was supported by funding from Scottish Government as part of it's 2016–2021 Environment, Agriculture and Food Strategic Research Programme (project code RD2.2.6).

Ethical statement

N/A.

Declaration of Competing Interest

the authors declare that there is no conflict of interest.

Acknowledgments

We thank Leighton Pritchard (Strathclyde university) for useful discussion, and Claus Dieter-Meyer and Glenn Marion (BioSS) for manuscript critique and Mart Achtman for Enterobase discussions. We dedicate this manuscript to the memory of Frank Wright (BioSS) and his contribution to phylogenetics research and teaching.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.crmicr.2021.100093.

References

Allué-Guardia, A., García-Aljaro, C., Muniesa, M., 2011. Bacteriophage-encoding cytolethal distending toxin type V gene induced from nonclinical *Escherichia coli* isolates. Infect. Immun. 79, 3262–3272. https://doi.org/10.1128/iai.05071-11.

Arndt, D., Grant, J.R., Marcu, A., Sajed, T., Pon, A., Liang, Y., Wishart, D.S., 2016. PHASTER: a better, faster version of the PHAST phage search tool. Nucl. Acids Res. 44, W16–W21. https://doi.org/10.1093/nar/gkw387.

Baranyi, J., Roberts, T.A., 1994. A dynamic approach to predicting bacterial growth in food. Int. J. Food Microbiol. 23, 277–294.

Blattner, F.R., Plunkett, G., Bloch, C.A., et al., 1997. The complete genome sequence of *Escherichia coli* K-12. In: Science, 277, pp. 1453–1462 (New York, NY.

Bortolaia, V., Kaas, R.S., Ruppe, E., et al., 2020. ResFinder 4.0 for predictions of phenotypes from genotypes. J. Antimicrob. Chemother. 75, 3491–3500. https://doi.org/10.1093/jac/dkaa345.

Brennan, F.P., Abram, F., Chinalia, F.A., Richards, K.G., O'Flaherty, V., 2010. Characterization of environmentally persistent *Escherichia coli* isolates leached from an Irish soil. Appl. Environ. Microbiol. 76, 2175–2180. https://doi.org/10.1128/aem.01944-09.

Carattoli, A., Bertini, A., Villa, L., Falbo, V., Hopkins, K.L., Threlfall, E.J., 2005. Identification of plasmids by PCR-based replicon typing. J. Microbiol. Methods 63, 219–228. https://doi.org/10.1016/j.mimet.2005.03.018.

Carattoli, A., Zankari, E., García-Fernández, A., Voldby Larsen, M., Lund, O., Villa, L., Møller Aarestrup, F., Hasman, H., 2014. *In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob. Agents Chemother. 58, 3895–3903. https://doi.org/10.1128/aac.02412-14.

Chernomor, O., von Haeseler, A., Minh, B.Q., 2016. Terrace aware data structure for phylogenomic inference from supermatrices. Syst. Biol. 65, 997–1008. https://doi.org/10.1093/sysbio/syw037.

Clermont, O., Bonacorsi, S., Bingen, E., 2000. Rapid and simple determination of the *Escherichia coli* phylogenetic group. Appl. Environ. Microbiol. 66, 4555–4558.

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

Clermont, O., Gordon, D.M., Brisse, S., Walk, S.T., Denamur, E., 2011. Characterization of the cryptic escherichia lineages: rapid identification and prevalence. Environ. Microbiol. 13, 2468–2477. https://doi.org/10.1111/j.1462-2920.2011.02519.x.

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E.P.C., Vergnaud, G., Gautheret, D., Pourcel, C., 2018. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucl. Acids Res. 46, W246–W251. https://doi.org/10.1093/nar/gky425.

Dallman, T.J., Ashton, P.M., Byrne, L., et al., 2015. Applying phylogenomics to understand the emergence of Shiga-toxin-producing *Escherichia coli* O157:H7 strains causing severe human disease in the UK. Microb. Genom. 1 https://doi.org/10.1099/mgen.0.000029.

Edgar, R., Bibi, E., 1997. MdfA, an *Escherichia coli* multidrug resistance protein with an extraordinarily broad spectrum of drug recognition. J. Bacteriol. 179, 2274–2280. https://doi.org/10.1128/jb.179.7.2274-2280.1997.

Ellabaan, M.M.H., Munck, C., Porse, A., Imamovic, L., Sommer, M.O.A., 2021. Forecasting the dissemination of antibiotic resistance genes across bacterial genomes. Nat. Commun. 12, 2435. https://doi.org/10.1038/s41467-021-22757-1.

Foster, Ross, Pennycott, Hopkins, McLaren, 1998. Isolation of *Escherichia coli* O86:K61 producing cyto-lethal distending toxin from wild birds of the finch family. Lett. Appl. Microbiol. 26, 395–398. https://doi.org/10.1046/j.1472-765X.1998.00359.x.

Garrity, G.M., Kraft, C.S., 2016. A new genomics-driven taxonomy of bacteria and archaea: are we there yet? J. Clin. Microbiol. 54 (8), 1956–1963. https://doi.org/10.1128/JCM.00200-16.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321. https://doi.org/10.1093/sysbio/syq010 syq010 [pii].

Holden, N.J., Wright, F., MacKenzie, K., Marshall, J., Mitchell, S., Mahajan, A., Wheatley, R., Daniell, T.J., 2013. Prevalence and diversity of *Escherichia coli* isolated from a barley trial supplemented with bulky organic soil amendments: green compost and bovine slurry. Lett. Appl. Microbiol. 58, 205–212. https://doi.org/10.1111/lam.12180.

Ingle, D.J., Clermont, O., Skurnik, D., Denamur, E., Walk, S.T., Gordon, D.M., 2011. Biofilm formation by and thermal niche and virulence characteristics of Escherichia spp. Appl. Environ. Microbiol. 77 (8), 2695–2700. https://doi.org/10.1128/AEM.02401-10.

Ishii, S., Sadowsky, M.J., 2008. *Escherichia coli* in the environment: implications for water quality and human health. Microbes Environ. 23, 101–108. https://doi.org/10.1264/jsme2.23.101.

Ishii, S., Yan, T., Vu, H., Hansen, D.L., Hicks, R.E., Sadowsky, M.J., 2009. Factors controlling long-term survival and growth of naturalized *Escherichia coli* populations in temperate field soils. Microbes Environ. 25, 8–14. https://doi.org/10.1264/jsme2.ME09172.

Jang, J., Hur, H.-G., Sadowsky, M.J., Byappanahalli, M.N., Yan, T., Ishii, S., 2017. Environmental *Escherichia coli*: ecology and public health implications - a review. J. Appl. Microbiol. https://doi.org/10.1111/jam.13468 n/a-n/a.

Janka, A., Bielaszewska, M., Dobrindt, U., Greune, L., Schmidt, M.A., Karch, H., 2003. Cytolethal distending toxin gene cluster in enterohemorrhagic *Escherichia coli* O157: H− and O157:H7: characterization and evolutionary considerations. Infect. Immun. 71, 3634–3638. https://doi.org/10.1128/iai.71.6.3634-3638.2003.

Kim, M., Oh, H.-S., Park, S.-C., Chun, J., 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int. J. Syst. Evol. Microbiol. 64, 346–351. https://doi.org/10.1099/ijs.0.059774-0.

Konstantinidis, K.T., Tiedje, J.M., 2007. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. Current Opinion in Microbiology. https://doi.org/10.1016/j.mib.2007.08.006.

Land, M., Hauser, L., Jun, S.-R., et al., 2015. Insights from 20 years of bacterial genome sequencing. Funct. Integr. Genom. 15, 141–161. https://doi.org/10.1007/s10142-015-0433-4.

Lindsey, R.L., Knipe, K., Rowe, L., Garcia-Toledo, L., Loparev, V., Juieng, P., Trees, E., Strockbine, N., Stripling, D., Gerner-Smidt, P., 2015. Complete genome sequences of two shiga toxin-producing *escherichia coli* strains from serotypes O119:H4 and O165: H25. Genome Announc. 3, e01496–e01515. https://doi.org/10.1128/genomeA.01496-15.

Liu, S., Jin, D., Lan, R., Wang, Y., Meng, Q., Dai, H., Lu, S., Hu, S., Xu, J., 2015. Escherichia marmotae sp. nov., isolated from faeces of Marmota Himalayana. Int. J. Syst. Evol. Microbiol. 65, 2130–2134. https://doi.org/10.1099/ijs.0.000228.

Low, A.S., Holden, N., Rosser, T., Roe, A.J., Constantinidou, C., Hobman, J.L., Smith, D. G., Low, J.C., Gally, D.L., 2006. Analysis of fimbrial gene clusters and their expression in enterohaemorrhagic *Escherichia coli* O157:H7. Environ. Microbiol. 8, 1033–1047.

Ludden, C., Raven, K.E., Jamrozy, D., et al., 2019. One Health genomic surveillance of *escherichia coli* demonstrates distinct lineages and mobile genetic elements in isolates from humans versus livestock. MBio 10, e02693–e02718. https://doi.org/10.1128/mBio.02693-18.

Lukjancenko, O., Wassenaar, T.M., Ussery, D.W., 2010. Comparison of 61 sequenced *Escherichia coli* genomes. Microb. Ecol. 60, 708–720. https://doi.org/10.1007/s00248-010-9717-3.

Luo, C., Walk, S.T., Gordon, D.M., Feldgarden, M., Tiedje, J.M., Konstantinidis, K.T., 2011. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. Proc. Natl Acad. Sci. U. S. A. 108, 7200–7205. https://doi.org/10.1073/pnas.1015622108, 1015622108 [pii].

Lupolova, N., 2017. Patchy promiscuity: machine learning applied to predict the host specificity of *Salmonella enterica* and *Escherichia coli*. University of Edinburgh. https://doi.org/10.7488/ds/2102.

Lupolova, N., Dallman, T.J., Holden, N.J., Gally, D.L., 2017. Patchy promiscuity: machine learning applied to predict the host specificity of *Salmonella enterica* and *Escherichia coli*. Microbial Genomics 3. https://doi.org/10.1099/mgen.0.000135.

Maiden, M.C.J., 2006. Multilocus sequence typing of bacteria. Annu. Rev. Microbiol. 60, 561–588. https://doi.org/10.1146/annurev.micro.59.030804.121325.

Maiden, M.C.J., van Rensburg, M.J.J., Bray, J.E., Earle, S.G., Ford, S.A., Jolley, K.A., McCarthy, N.D., 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat. Rev. Microbiol. 11, 728–736. https://doi.org/10.1038/nrmicro3093. http://www.nature.com/nrmicro/journal/v11/n10/abs/nrmicro3093.html#supplementary-information.

Makarova, K.S., Haft, D.H., Barrangou, R., et al., 2011. Evolution and classification of the CRISPR–cas systems. Nat. Rev Microbiol. 9, 467–477. https://doi.org/10.1038/nrmicro2577.

Marshall, J., Rossez, Y., Mainda, G., Gally, D.L., Daniell, T., Holden, N., 2016. Alternate thermoregulation and functional binding of *Escherichia coli* type 1 fimbriae in environmental and animal isolates. FEMS Microbiol. Lett. 363, fnw251. https://doi.org/10.1093/femsle/fnw251.

Martinez-Antonio, A., Janga, S.C., Thieffry, D., 2008. Functional organisation of *Escherichia coli* transcriptional regulatory network. J. Mol. Biol. 381, 238–247. https://doi.org/10.1016/j.jmb.2008.05.054.

Mi, H., Muruganujan, A., Casagrande, J.T., Thomas, P.D., 2013. Large-scale gene function analysis with the PANTHER classification system. Nat. Protoc. 8, 1551. https://doi.org/10.1038/nprot.2013.092. https://www.nature.com/articles/nprot.2013.092#supplementary-information.

Mills, M.C., Lee, J., 2019. The threat of carbapenem-resistant bacteria in the environment: evidence of widespread contamination of reservoirs at a global scale. Environ. Pollut. 255, 113143 https://doi.org/10.1016/j.envpol.2019.113143.

Minh, B.Q., Nguyen, M.A.T., von Haeseler, A., 2013. Ultrafast approximation for phylogenetic bootstrap. Mol. Biol. Evol. 30, 1188–1195. https://doi.org/10.1093/molbev/mst024.

Mughini-Gras, L., Dorado-García, A., van Duijkeren, E., et al., 2019. Attributable sources of community-acquired carriage of *Escherichia coli* containing beta-lactam antibiotic resistance genes: a population-based modeling study. Lancet Planet. Health 3, e357–e369. https://doi.org/10.1016/S2542-5196(19)30130-5.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274. https://doi.org/10.1093/molbev/msu300.

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., Fookes, M., Falush, D., Keane, J.A., Parkhill, J., 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31, 3691–3693. https://doi.org/10.1093/bioinformatics/btv421.

Partridge, S.R., Kwong, S.M., Firth, N., Jensen, S.O., 2018. Mobile genetic elements associated with antimicrobial resistance. Clin. Microbiol. Rev. 31, e00088–e00117. https://doi.org/10.1128/cmr.00088-17.

Richter, M., Rosselló-Móra, R., 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl Acad. Sci. U. S. A. 106, 19126–19131. https://doi.org/10.1073/pnas.0906412106.

Rodriguez-R, L., Konstantinidis, K., 2014. Bypassing Cultivation to Identify Bacterial Species, 9. ASM, pp. 111–118 p.^pp.

Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069. https://doi.org/10.1093/bioinformatics/btu153.

Sievers, F., Wilm, A., Dineen, D., et al., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Mol. Syst. Biol. 7 https://doi.org/10.1038/msb.2011.75.

Smith, J.L., Bayles, D.O., 2006. The contribution of cytolethal distending toxin to bacterial pathogenesis. Crit. Rev. Microbiol. 32, 227–248. https://doi.org/10.1080/10408410601023557.

Touchon, M., Hoede, C., Tenaillon, O., et al., 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLos Genet. 5, e1000344 https://doi.org/10.1371/journal.pgen.1000344.

Walk, S.T., Alm, E.W., Calhoun, L.M., Mladonicky, J.M., Whittam, T.S., 2007. Genetic diversity and population structure of *Escherichia coli* isolated from freshwater beaches. Environ. Microbiol. 9, 2274–2288.

Walk, S.T., Alm, E.W., Gordon, D.M., Ram, J.L., Toranzos, G.A., Tiedje, J.M., Whittam, T. S., 2009. Cryptic lineages of the genus Escherichia. Appl. Environ. Microbiol. 75 (20), 6534–6544. https://doi.org/10.1128/aem.01262-09.

Wang, Z., Wu, M., 2013. A phylum-level bacterial phylogenetic marker database. Mol. Biol. Evol. 30, 1258–1262. https://doi.org/10.1093/molbev/mst059.

Welch, R.A., Burland, V., Plunkett, G., et al., 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc. Natl Acad. Sci. 99, 17020–17024. https://doi.org/10.1073/pnas.252529799.

Wu, M., Eisen, J.A., 2008. A simple, fast, and accurate method of phylogenomic inference. Genome Biol. 9, R151. https://doi.org/10.1186/gb-2008-9-10-r151.

Wu, M., Scott, A.J., 2012. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics 28, 1033–1034. https://doi.org/10.1093/bioinformatics/bts079.

Zeng, X., Brown, S., Gillespie, B., Lin, J., 2014. A single nucleotide in the promoter region modulates the expression of the β-lactamase OXA-61 in Campylobacter jejuni. J. Antimicrob. Chemother. 69, 1215–1223. https://doi.org/10.1093/jac/dkt515.

Zhou, Z., Alikhan, N.-F., Mohamed, K., Fan, Y., Group tAS & Achtman M, 2020. The Enterobase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. Genome Res. 30, 138–152. https://doi.org/10.1101/gr.251678.119.

Zhou, Z., Alikhan, N.-F., Mohamed, K., Fan, Y., the Agama Study Group, Achtman, M., 2020. The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. Genome Research. https://doi.org/10.1101/gr.251678.119.

Zhou, Z., Alikhan, N.-.F., Sergeant, M.J., Luhmann, N., Vaz, C., Francisco, A.P., Carriço, J.A., Achtman, M., 2018. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. Genome Res. 28, 1395–1404. https://doi.org/10.1101/gr.232397.117.