








A role for ColV plasmids in the evolution of pathogenic *Escherichia coli* ST58

Cameron J. Reid ^{1✉}, Max L. Cummins ¹, Stefan Börjesson ^{2,3}, Michael S. M. Brouwer⁴, Henrik Hasman⁵, Anette M. Hammerum⁵, Louise Roer⁵, Stefanie Hess⁶, Thomas Berendonk⁷, Kristina Nešporová ^{8,9}, Marisa Haenni¹⁰, Jean-Yves Madec¹⁰, Astrid Bethe ^{11,12}, Geovana B. Michael^{11,12}, Anne-Kathrin Schink^{11,12}, Stefan Schwarz ^{11,12}, Monika Dolejska^{8,9,13} & Steven P. Djordjevic ^{1✉}

Escherichia coli ST58 has recently emerged as a globally disseminated uropathogen that often progresses to sepsis. Unlike most pandemic extra-intestinal pathogenic *E. coli* (ExPEC), which belong to pathogenic phylogroup B2, ST58 belongs to the environmental/commensal phylogroup B1. Here, we present a pan-genomic analysis of a global collection of 752 ST58 isolates from diverse sources. We identify a large ST58 sub-lineage characterized by near ubiquitous carriage of ColV plasmids, which carry genes encoding virulence factors, and by a distinct accessory genome including genes typical of the Yersiniabactin High Pathogenicity Island. This sub-lineage includes three-quarters of all ExPEC sequences in our study and has a broad host range, although poultry and porcine sources predominate. By contrast, strains isolated from cattle often lack ColV plasmids. Our data indicate that ColV plasmid acquisition contributed to the divergence of the major ST58 sub-lineage, and different sub-lineages inhabit poultry, swine and cattle.

¹iThree Institute, University of Technology Sydney, Ultimo, NSW 2007, Australia. ²Department of Animal Health and Antimicrobial Strategies, National Veterinary Institute (SVA), 75189 Uppsala, Sweden. ³Department of Microbiology, Public Health Agency of Sweden, 17182 Solna, Sweden. ⁴Wageningen Bioveterinary Research, Lelystad, Netherlands. ⁵Department of Bacteria, Parasites and Fungi, Statens Serum Institut, Copenhagen S, Denmark. ⁶Institute of Microbiology, Technische Universität Dresden, Dresden, Germany. ⁷Institute of Hydrobiology, Technische Universität Dresden, Dresden, Germany. ⁸CEITEC VETUNI, University of Veterinary Sciences Brno, Brno, Czech Republic. ⁹Department of Biology and Wildlife Disease, Faculty of Veterinary Hygiene and Ecology, University of Veterinary Sciences Brno, Brno, Czech Republic. ¹⁰Université de Lyon-ANSES, Unité Antibiorésistance et Virulence Bactériennes, Lyon, France. ¹¹Institute of Microbiology and Epizootics, Centre for Infection Medicine, Department of Veterinary Medicine, Freie Universität Berlin, Berlin, Germany. ¹²Veterinary Centre for Resistance Research (TZR), Freie Universität Berlin, 14163 Berlin, Germany. ¹³Biomedical Center, Charles University, Charles, Czech Republic. ✉email: Cameron.Reid@uts.edu.au; Steven.Djordjevic@uts.edu.au

E*scherichia coli* predominantly live as harmless commensals in the gastrointestinal tract of mammals and birds. They also reside, independent of a host, in environmental habitats including water, soil and sediments. If they possess or acquire factors that allow them to adapt to niches in humans that are outside the gut, they can cause disease. The *E. coli* pathotype that accounts for the vast majority of human extra-intestinal pathologies—urinary tract infections, pyelonephritis, sepsis and meningitis—is known as extra-intestinal pathogenic *E. coli* (ExPEC). The pathologies it causes place a significant burden on health systems globally¹.

ST58 (clonal complex (CC) 155) is a persistent ExPEC clonal group that is unusual among ExPEC strains by belonging to phylogroup B1, one of eight stable phylogroups used to classify *E. coli*. Most ExPEC strains, including those in the globally dominant ST131 clonal group, are members of phylogroup B2². The emergence of ST58 is perhaps best highlighted by a recent study that found the proportion of B1-ST58 isolated from bloodstream infections in the Paris region has more than doubled over a 12 year period in contrast to stable proportions of infections caused by phylogroup B2 isolates³. Beyond this study, *E. coli* ST58 is increasingly responsible for both sporadic and persistent cases of bloodstream infections in humans across the globe^{4–12}. As well as colonising humans, ST58 has been identified in healthy and diseased food-producing animals (cattle, poultry, swine) and in poultry farm-associated flies, manure and water^{8,13–21}. Early reports of ST58 postulated a wild animal source and it has since been reported in commensal and pathogenic cases from wild, captive and companion animals, with wild birds from both urban and pristine environments emerging globally as an important source^{22–33}. Food sources of ST58 include chicken and turkey meat, barley and oats, raw meat-based pet food and store-bought produce^{8,34–37}. Further concern arises from widespread observations of ST58 carrying antimicrobial resistance genes (ARGs). Human faecal isolates of ST58 from healthy individuals in Tunisia, Sweden and the Netherlands^{38–41} have been found harbouring genes conferring resistance to third-generation antimicrobials, as have ST58 isolates from food-producing animals^{8,13–21} (*bla*_{CTX-M} genes encoding for extended-spectrum beta-lactamase (ESBL) production). Multidrug-resistant (MDR) and ESBL-producing ST58 also contaminate soil, rivers, man-groves and wastewater^{35,42–44}.

We have limited knowledge of how ST58 emerged as a human pathogen. How did it come to evolve within phylogroup B1, which is rarely pathogenic?² Though host factors are the greatest predictor of extra-intestinal virulence, intrinsic factors include adhesins, toxins, protectins and iron acquisition systems⁴⁵. Phage and plasmid mobilised iron acquisition systems in particular, including yersiniabactin (HPI) and aerobactin among others, have been shown to play a major role in intrinsic virulence across the genus *Escherichia*⁴⁶. The majority of existing studies reporting ST58, and indeed most studies on ExPEC, are limited by their employment of ESBL-selection criteria. Although clinically coherent, this selection criteria ignores the ecological complexity of AMR and pathogen emergence. The consequences are significant—ESBL-selection criteria drive misleading generalisations about the AMR status of clonal groups and actively prevent a complete understanding of their evolutionary history, particularly with regard to the identification of non-AMR traits that play significant roles in fitness and pathogenesis⁴⁷.

We recently reported an ST58 strain carrying a ColV plasmid that caused urosepsis in a Sydney Hospital⁴. ColV plasmids are a subset of typically conjugative F plasmids abundant in poultry and have also been reported in healthy human faecal commensal *E. coli* and ExPEC⁴⁸. They have known pathogenic properties and traits involved in intestinal fitness^{4,49–53}, and carry a variety of

ARGs, class 1 integrons and mercury resistance transposons⁴⁸. Therefore, there are numerous selective pressures that may contribute to their persistence in animal and human hosts. ColV plasmids are carried at high levels in important human pathogens with poultry associations including ExPEC clonal groups ST95 and ST117, though their relative abundance in other STs is unknown⁵⁴.

In this work, we perform a pan-genomic epidemiological analysis of a global collection of 752 ST58 isolated from humans, animals and environmental sources. Hypothesising that ColV plasmids had a role in the emergence of ST58 as a pathogen, we query their presence and source distribution in a collection of 34,364 draft *E. coli* genome assemblies from Enterobase. Our findings support the role of ColV plasmids and non-human sources in the evolution of pathogenic ST58 as well as the idea that pathogen emergence should be understood within a One Health framework, resulting from a complex of mechanistic drivers and networked pathways between environments and hosts rather than discretely acting selective pressures.

Results

A diverse collection of ST58 genomes. We first built a collection of *E. coli* ST58 genome sequences from isolates that were diverse in origin—temporally, by source and geographically (Fig. 1(a–c)). In total, our genome collection comprised 752 whole-genome sequences of *E. coli* ST58 isolates collected between 1970 and 2019 from 6 niches sub-divided into 15 sources, Fig. 1(b) across 33 countries on 6 continents (Supplementary Data 1). Of the 752 sequences, 178 were sourced from in-house and collaborator collections and the remainder from Enterobase. Within the collection, 92% of sequences (692/752) were from isolates collected after the year 2000 (See Fig. S1 for full range). The most common niche represented in the collection was livestock, accounting for 68% of the collection (514/752). Livestock was followed distantly by wild animals (12%, 93/752) and humans (11%, 79/752). Livestock isolates were predominantly from bovine (35%, 262/752), poultry (17%, 125/752) and porcine sources (14%, 106/752). Among wild animals isolates, 11% were from avian sources (84/752). More than half of the human isolates were from people suffering from extra-intestinal *Escherichia coli* pathologies (ExPEC; 6% of the total collection, 43/752). The most represented continents in the collection were North America (60%, 454/752) and Europe (25%, 189/752).

***E. coli* ST58 contains a major sub-lineage rich in ColV virulence plasmids.** To investigate evolutionary relationships within our collection of ST58 genomic sequences, we first generated a global ST58 phylogeny inferred from the core gene multi-alignment.

The core gene phylogeny was grouped into six major clusters by fastbaps (designated BAP1–6; See Methods), encompassing 750 sequences. The outgroup strain, belonging to ST155, and two ST58 strains on divergent branches near the root comprised BAP clusters 7, 8 and 9, which are hereafter excluded from cluster-based analyses. Each major cluster was characterised by source, serotype, *fimH* allele, F plasmid replicon sequence type (F RSTs) and inference of ColV plasmid presence. Two of the clusters included the vast majority of collection sequences: BAP6 ($n = 205$) and BAP2 ($n = 363$), (Fig. 2). BAP6 sequences were more commonly from bovine than other sources (bovine: 142/205, 69%) but they displayed a diversity of F plasmid RSTs, serotypes and *fimH* alleles (Figs. 3, S2, S3). In contrast, BAP2 exhibited a greater distribution of sources and dominance of specific serotypes and *fimH* alleles (Fig. 3). The most striking feature of the BAP2 cluster was the high proportion of sequences

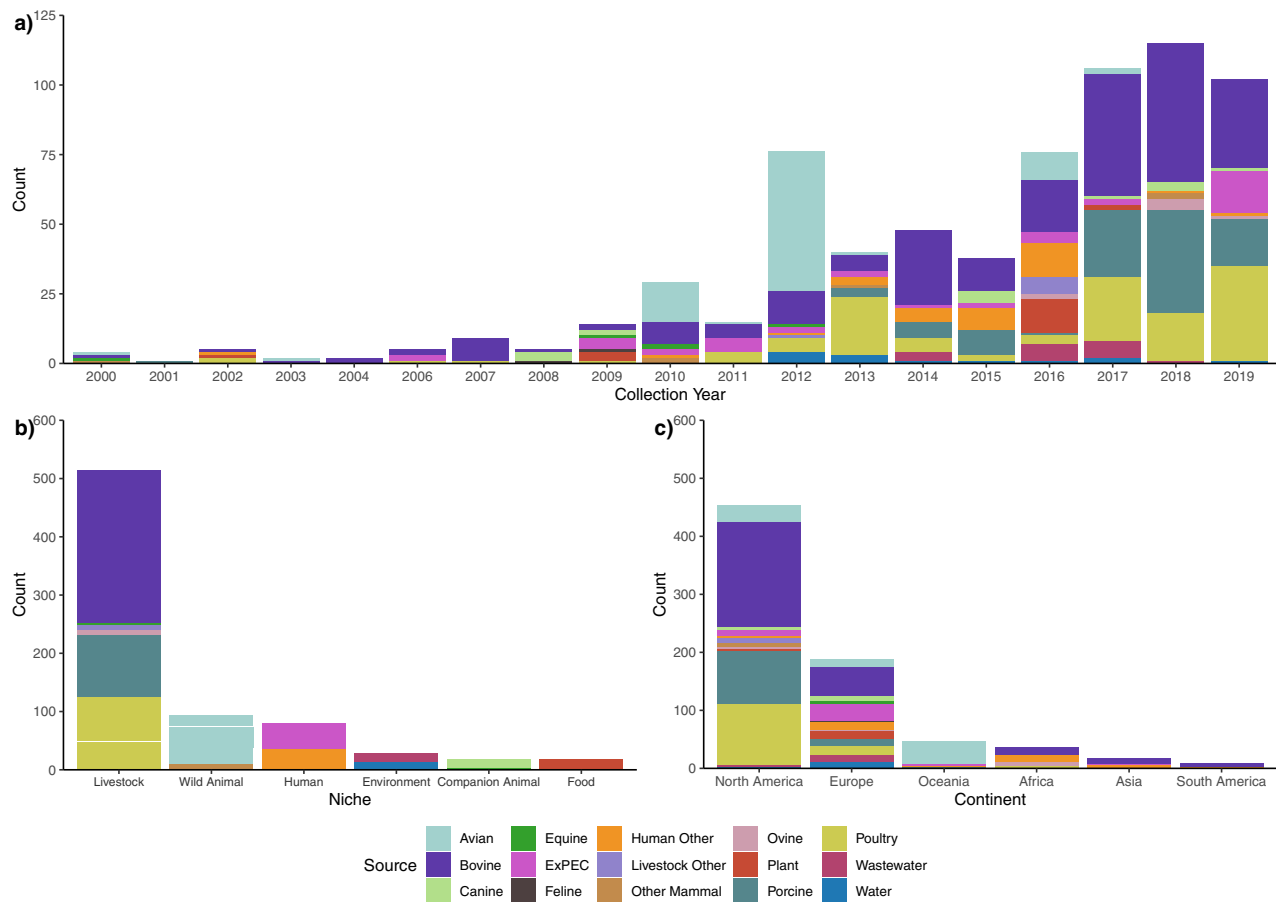


Fig. 1 ST58 genome collection metadata. Summary of metadata of 752 *E. coli* ST58 strains stratified by source. **a** Collection year (2000–2019; for full year distribution see Fig. S1); **b** Niche; **c** Continent.

that featured ColV F plasmids (308/363, 85%). These plasmids were represented predominantly by three RSTs: F2:B1, F18:A6:B1 and F18:B1 (Figs. 3b, S4, Supplementary Data 2).

The initial screening of the cluster sequences for the presence of a ColV plasmid was based on the Liu criteria (see Methods, Fig. S5). However, as the Liu criteria are limited by screening for specific genes without consideration of plasmid backbone structure, we sought to corroborate the identification of ColV plasmids via an additional methodology. For this purpose, we aligned de novo assemblies to the backbone of archetypal F2:B1 ColV plasmid pCERC4 and visualised the alignment as a heatmap of binned nucleotide identities (Fig. 4). Overall, results obtained using each methodology corresponded well. Coverage patterns of ColV-positive (ColV+) sequences outside the BAP2 cluster were similar to those within it.

Most of the ColV+ genomes that were identified in our collection originated from sources associated with food production (poultry, porcine, bovine) and extra-intestinal disease in humans (ExPEC) (Fig. 5, Fig. S6). For three of these sources, the vast majority of genome sequences were ColV+: poultry (109/125, 87%), porcine (77/106, 73%) and ExPEC (33/43, 77%). In contrast, the proportion of bovine sequences that were ColV+ was notably less (48/262, 18%).

The BAP2 cluster displays a distinctive accessory genome.

Having identified at the core genome level a distinct ColV+ sub-lineage within ST58—the BAP2 cluster—we hypothesised that it would also exhibit fundamental differences in its accessory

genome relative to the remainder of the phylogeny. To test this we performed a pangenome-wide association study (pan-GWAS), using BAP2 cluster membership as the test variable. Within the pangenome, 78 genes coding for non-hypothetical proteins were over-represented in the BAP2 cluster and 55 were under-represented (these genes are therefore associated with non-BAP2 ST58). We also performed this analysis for the other BAP groups; however, only BAP6 contained genes that met our significance threshold ($1E-50$) for over- or under-representation, and these *p*-values were generally far lower than any of the genes associated with BAP2 (Fig. S7, Supplementary Data 3). Mapping genes identified in the BAP2 analysis back to the core gene phylogeny revealed that a sub-group of approximately 294 genomes within BAP2 contained a highly conserved accessory gene profile (Fig. 6). In addition to the expected identification of genes present on ColV plasmids, several other genes stood out in the BAP2-associated accessory gene profile. The two most highly associated genes in the cluster, *ugd* and *galF*, both encode enzymes involved in outer membrane lipopolysaccharide biosynthesis. Two prophage integrase *intA* genes were associated with the clade, one of which was exclusively associated with the aforementioned 294 sequence sub-group in the BAP2 cluster. *mlrA*, a regulator of curli biosynthesis and biofilm formation was also exclusively associated with this sub-group. *fyuA*, a marker gene for Yersiniabactin High Pathogenicity Island (HPI), which encodes an iron uptake system that is a major contributor to intrinsic extra-intestinal virulence across the genus *Escherichia*, was also associated with BAP2 indicating that most of these ST58 have acquired an HPI-like pathogenicity island. This was

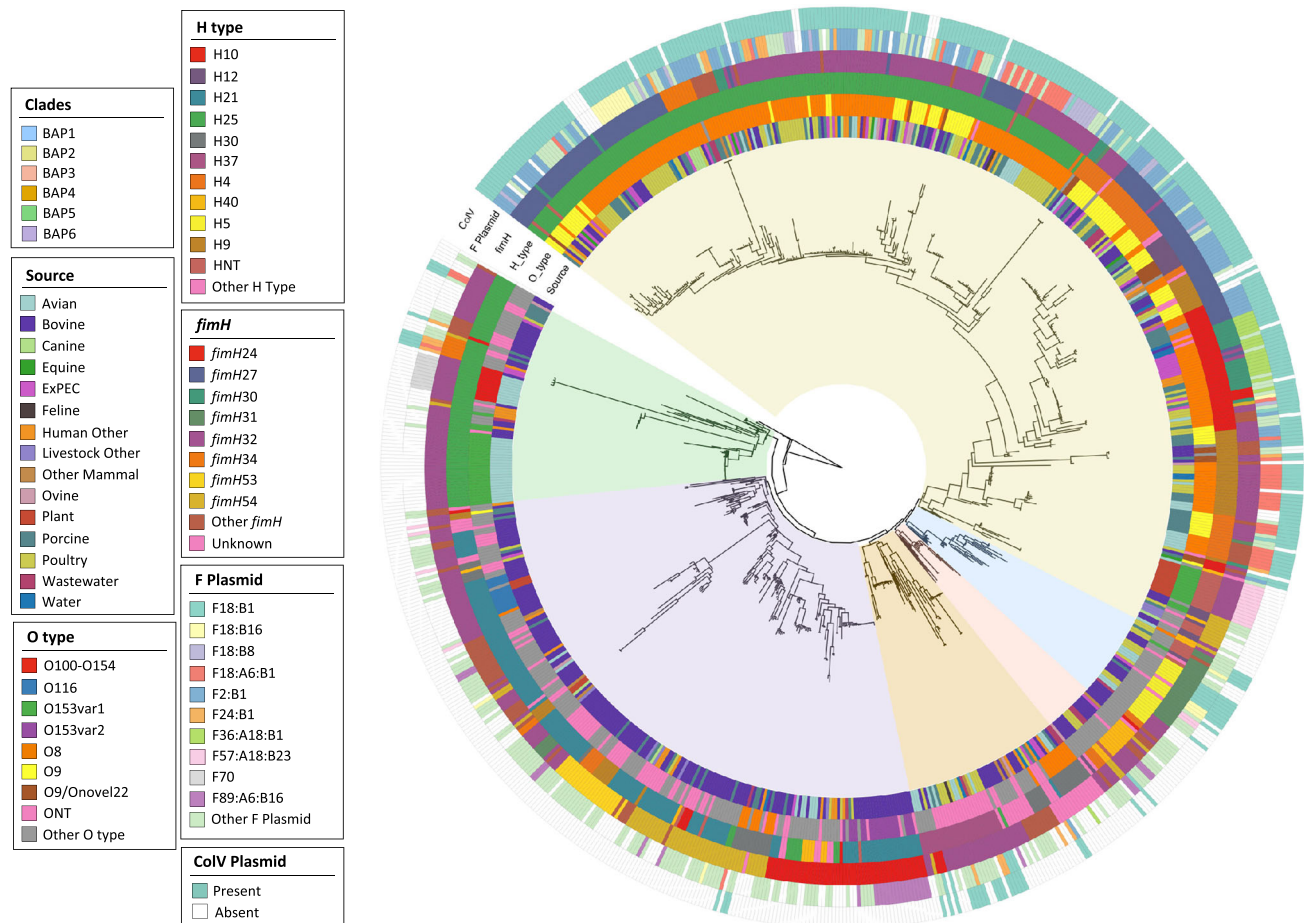


Fig. 2 Phylogenetic tree with metadata. Maximum-likelihood core gene phylogeny of 752 *E. coli* ST58 rooted on ST155 strain MOD-EC5019 with genotypic information and metadata.

supported by virulence gene screening which identified both *fyuA* (306/363; 84%) and cognate HPI marker, *irp2* (303/363; 83%), in a high proportion of BAP2 sequences (Fig. S9).

Twenty-seven genes were present in both BAP2 over- and under-represented groups, indicative of gene families that are likely core to ST58 but have alternative sequences in BAP2 compared to the remainder of the phylogeny (Supplementary Data 3). Most of these paralogous genes were involved in metabolism, membrane transport processes and DNA transcription, notably including genes of the *bcs* operon involved in cellulose metabolism, an abundant polysaccharide in the bovine rumen. These results suggest that the BAP2 cluster likely displays a number of functional differences to the remainder of ST58 in relation to multiple biological processes.

Overall the pan-GWAS analysis implies that both sequence divergence in core ST58 genes and multiple instances of horizontal gene transfer events involving plasmids, phages and genomic islands, have contributed to the evolution of the BAP2 cluster.

ColV+ and BAP2 genomes carry more antimicrobial resistance genes and virulence genes. We screened our collection of ST58 *E. coli* genomes for antimicrobial resistance genes (ARGs) and virulence-associated genes (VAGs). Strain-wise totals for each were compared with respect to BAP cluster membership and ColV status. This analysis revealed that, BAP2 sequences carried significantly more ARGs on average than the other clusters except BAP1 (Pairwise Wilcoxon test with Benjamini–Hochberg

adjusted *p*-values: vs BAP3 $p = 8.63e-8$; BAP4 $p = 0.036$; BAP5 $p = 6.26e-6$; BAP6 $p = 1.38e-28$), and more VAGs than all other clusters (Pairwise Wilcoxon test with Benjamini–Hochberg adjusted *p*-values: vs BAP1 $p = 5.23e-5$; BAP3 $p = 3.1e-9$; BAP4 $p = 2.72e-16$; BAP5 $p = 1.67e-21$; BAP6 $p = 5.96e-70$) (Fig. 7a, b). Similarly, ColV+ strains carry more ARGs (Two-sided Wilcoxon Rank-sum test: $p = 9.2e-34$) and VAGs (Two-sided Wilcoxon Rank-sum test: $p = 1.28e-112$) than ColV- strains (Fig. 7c, d).

A wide variety of ARGs were identified in the total collection including genes conferring resistance to older antimicrobial compounds such as ampicillin (*bla*_{TEM-1B}; 248; 33%), streptomycin (*strAB*; 296; 39% and 294; 39%), sulphonamides (*sul2*; 279; 37%), tetracyclines (*tet(A)*; 272; 36% and *tet(B)*; 165; 22%) and trimethoprim (*dfrA5*; 127; 17%). Genes mediating resistance to third-generation cephalosporins and carbapenems were comparatively less common. ESBL-encoding genes included *bla*_{CTX-M-1} (49; 7%), *bla*_{CTX-M-55} (23; 3%) and *bla*_{CTX-M-15} (21; 3%). AmpC beta-lactamase *bla*_{CMY-2} (61; 8%) was also identified. The class 1 integron-integrase gene, *intI* was present in 245 sequences (33%). Integrons are a major driver of bacterial evolution and *intI* is commonly linked with ARGs. Mercury resistance gene *merA* was present in 22% of sequences (164/752). Point mutations associated with resistance to fluoroquinolones were only identified in 9% of all sequences and 12% of BAP2 sequences (Fig. S8).

Nearly all the most common VAGs in the full collection (those present in more than 200 sequences) were concentrated within the BAP2 cluster and known to be carried on ColV plasmids. Abundant non-ColV VAGs included those encoding increased

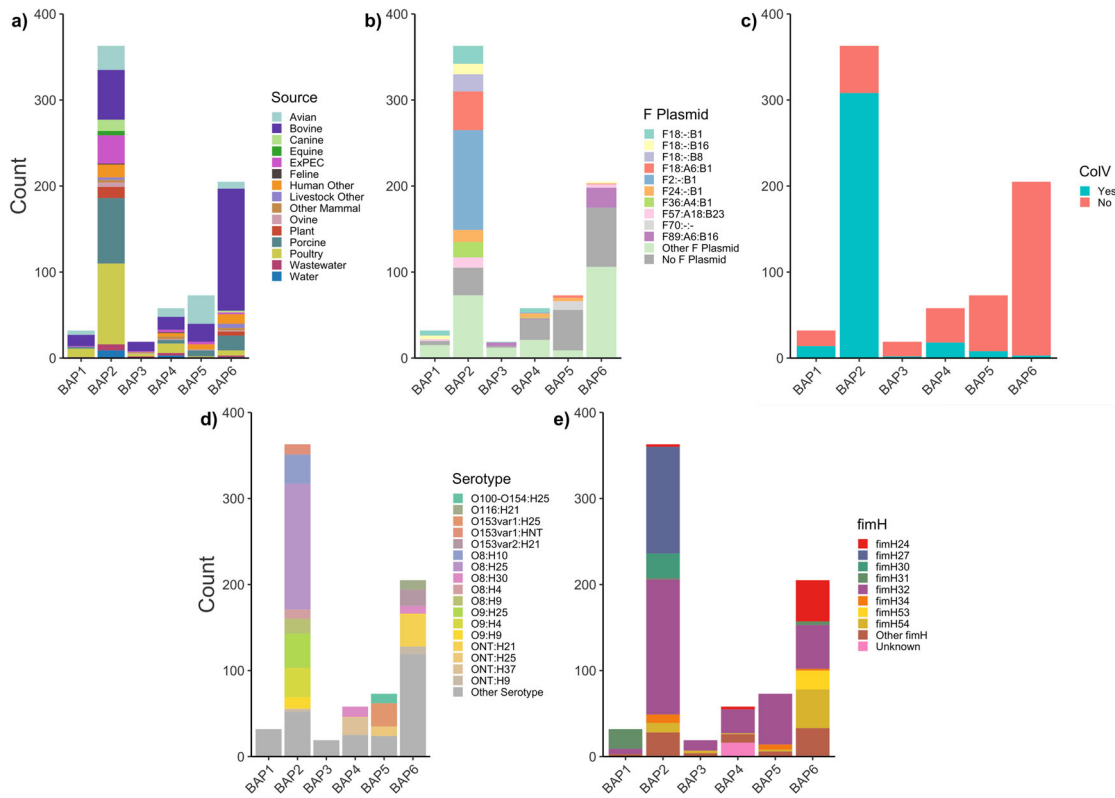


Fig. 3 Metadata distribution across BAP clusters. Distribution of **a** source; **b** F plasmid replicon sequence type (RST); **c** ColV carriage; **d** serotype and **e** *fimH* allele by BAP cluster for 750 ST58 genome sequences.

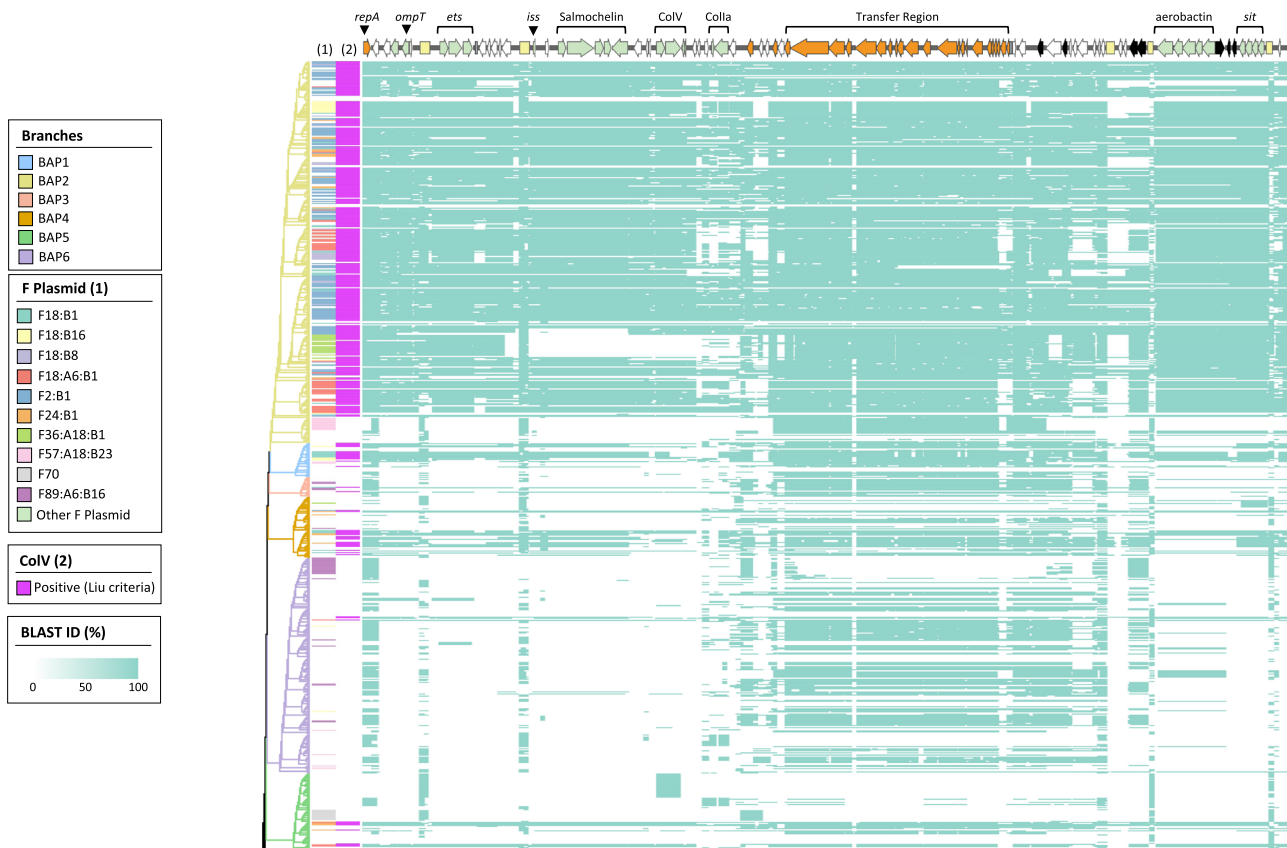


Fig. 4 Alignment of ST58 sequences to ColV plasmid pCERC4. Heatmap of nucleotide identity across 100 bp segments of the pCERC4 backbone. Tree branches are coloured by BAP cluster. F plasmid replicon sequence type (RST) and ColV presence are indicated in panels (1) and (2).

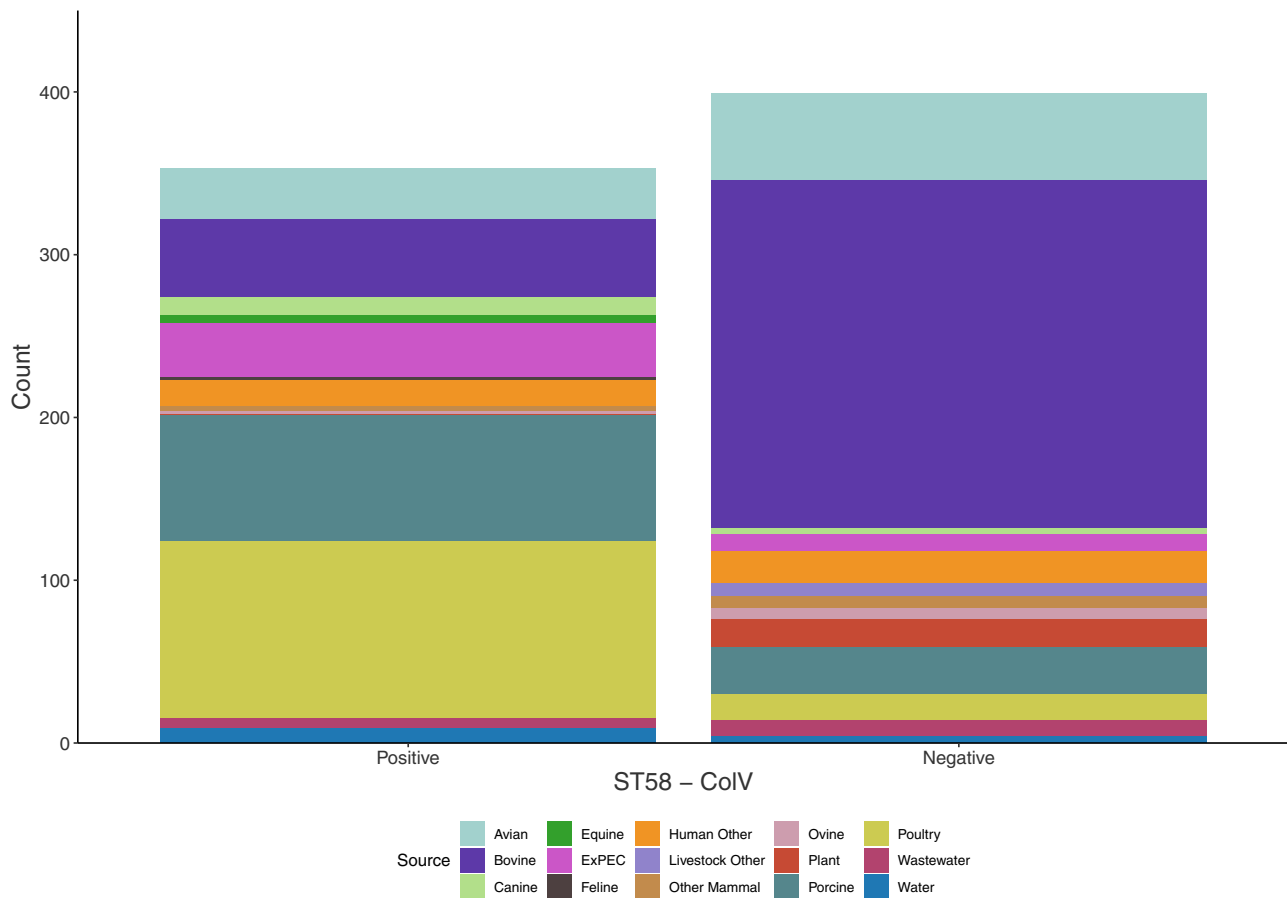


Fig. 5 ColV carriage by source. ColV carriage stratified by source in the ST58 genome collection.

serum survival protein *iss* (583; 78%; note: *iss* is frequently observed on ColV plasmids but is not included in the marker gene criteria); colonisation factor and fimbrial adhesin *fimH* (730; 97%); ferric dicitrate receptor *fecA* (412; 55%); high pathogenicity island marker and yersiniabactin iron receptor *fyuA* (347; 46%), and bacterial defence factor microcin transport protein *mchF* (324; 43%). Toxin carriage was rare though enterohaemolytic *E. coli* (EHEC) marker genes, including *stx* variants and *subA*, respectively, were identified in some strains (Fig. S9).

Aside from F plasmids, plasmids from incompatibility group IncI1 were the most commonly identified (228/752, 30%). IncI1 plasmids were present in all major clusters, but were diverse by IncI1 pMLST. (Fig. S10). Sixty-seven sequences that contained an IncI1 replicon did not have an identifiable pMLST. Isolates sourced from poultry had the highest intra-source proportion of IncI1-positive sequences (67/125; 53.6%).

The high ColV carriage rate in ST58 is comparable to other emerging ExPEC with major reservoirs in food animals. To date, most data on the presence of ColV plasmids in *E. coli* has been obtained from avian or human ExPEC isolates. In addition to these sources, we have observed, in high proportions, ColV plasmids in ST58 BAP2 genomes originating from other food production animals, particularly pigs. Thus, we wanted to investigate how common it was for ColV plasmids to be present in the genomes of *E. coli* isolated from other sources, as well as the range of *E. coli* STs hosting ColV plasmids. To do this, we curated and analysed a collection of 34,364 draft *E. coli* genome assemblies from Enterobase (Supplementary Data 4). We found that within this collection, poultry (2328/4260, 55%) was by far

the most dominant source for ColV+ *E. coli*, followed by porcine (526/2683, 20%) and human ExPEC (735/4465, 16%; Fig. 8a). Low carriage was observed in human other (523/15088; 3%) and bovine sources (176/6926; 3%). The rate of ColV carriage in all genomes was 13% (4370/34,364; Fig. 8b). ColV carriage rates summarised by ST and Source were compared for fifteen STs containing more than 100 assemblies and ColV carriage rate greater than 10% (Fig. 8c, d). Whilst the methodology for this collection of sequences captured only 588 ST58 draft genome assemblies (most of which are present in the primary collection under analysis in this manuscript), the ColV carriage rate of 48% (281/588) was only slightly higher than the estimate for the primary collection at 46% (353/752). Only five STs (G-ST117, C-ST88, B1-ST162, A-ST93 and C-ST23) had higher ColV carriage rates than ST58. Four of these STs belong to the 'environmental' phylogroups A, B1 and C whilst globally dominant APEC ST117 now belongs to phylogroup G, having previously been considered D or F⁵⁵. Poultry-sourced sequences dominated these top six STs though distributions varied, with ST88 displaying more porcine sequences. ST95, ST73 and ST12, all of which belong to 'pathogenic' phylogroup B2 contrastingly displayed mostly human sources, either ExPEC or other. Almost all of the STs present feature in numerous reports of human or animal infection, or in conjunction with concerning AMR genotypes.

Genomic linkages between human and non-human source ST58. We hypothesised that a proportion of ST58 that cause human extra-intestinal infections are closely related to ST58 found in non-human sources. To test this, we determined pairwise SNP counts between all sequences (Fig. S11) and then

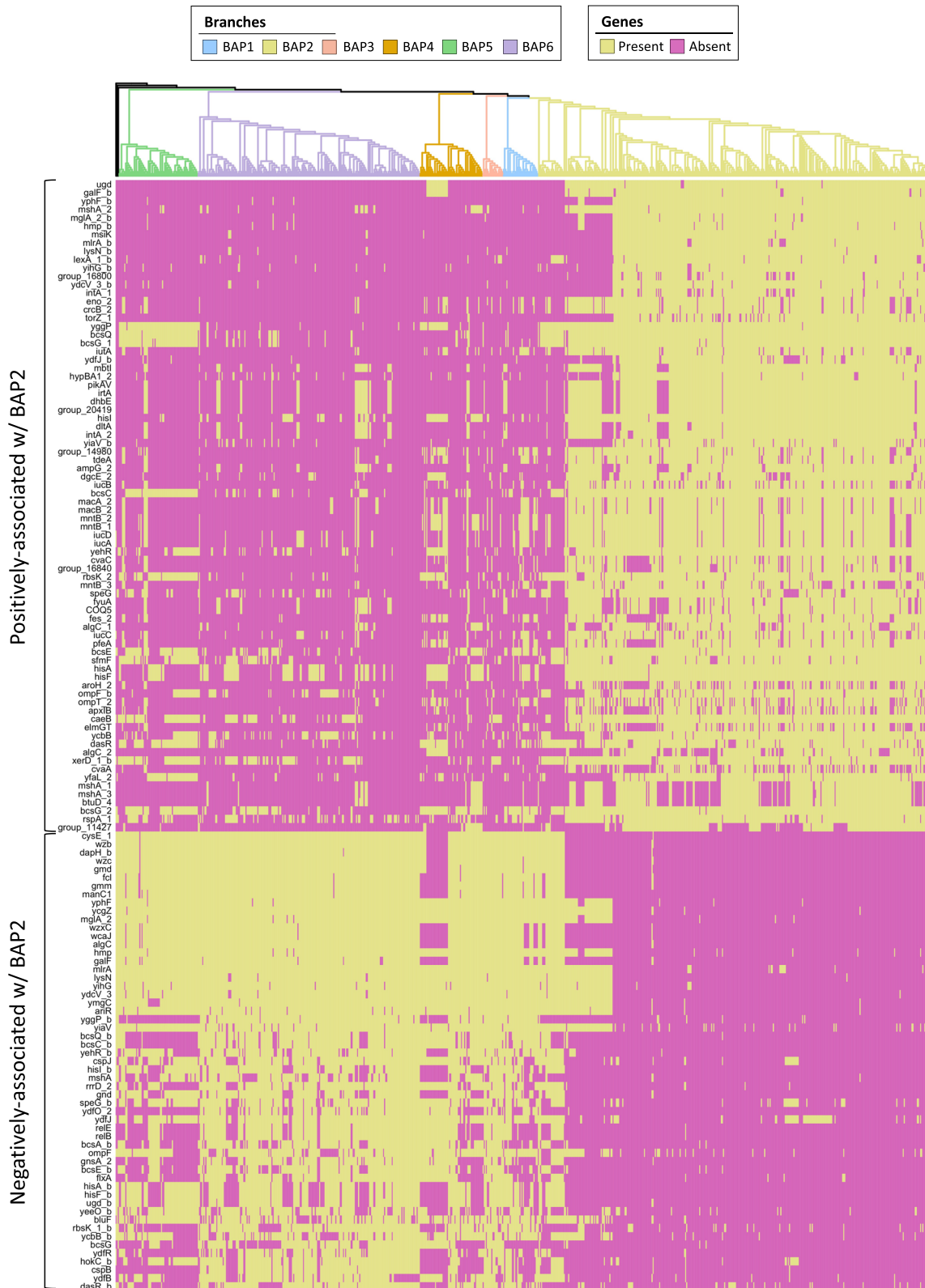


Fig. 6 Genes associated with the BAP2 cluster. Presence/absence (yellow/magenta) of genes positively and negatively associated with the BAP2 cluster mapped to the phylogeny. Tree branches indicate BAP clusters.

compared human source sequences (ExPEC, other) with pairwise SNP counts of 20 or less to sequences from the remaining 13 source categories. For perspective, 20 SNPs across the 2.8Mbp core gene alignment constitutes a core nucleotide divergence of just 0.0007%. We identified 135 pairwise cases of ≤ 20 SNPs that

occurred between 26 humans (15 ExPEC, 11 other) sequences and 34 non-human sequences representing 11 of the 13 non-human sources (Fig. 9). This analysis revealed a close linkage between geographically and temporally distinct strains. One cluster comprised 8 ExPEC from Denmark and 3 human others

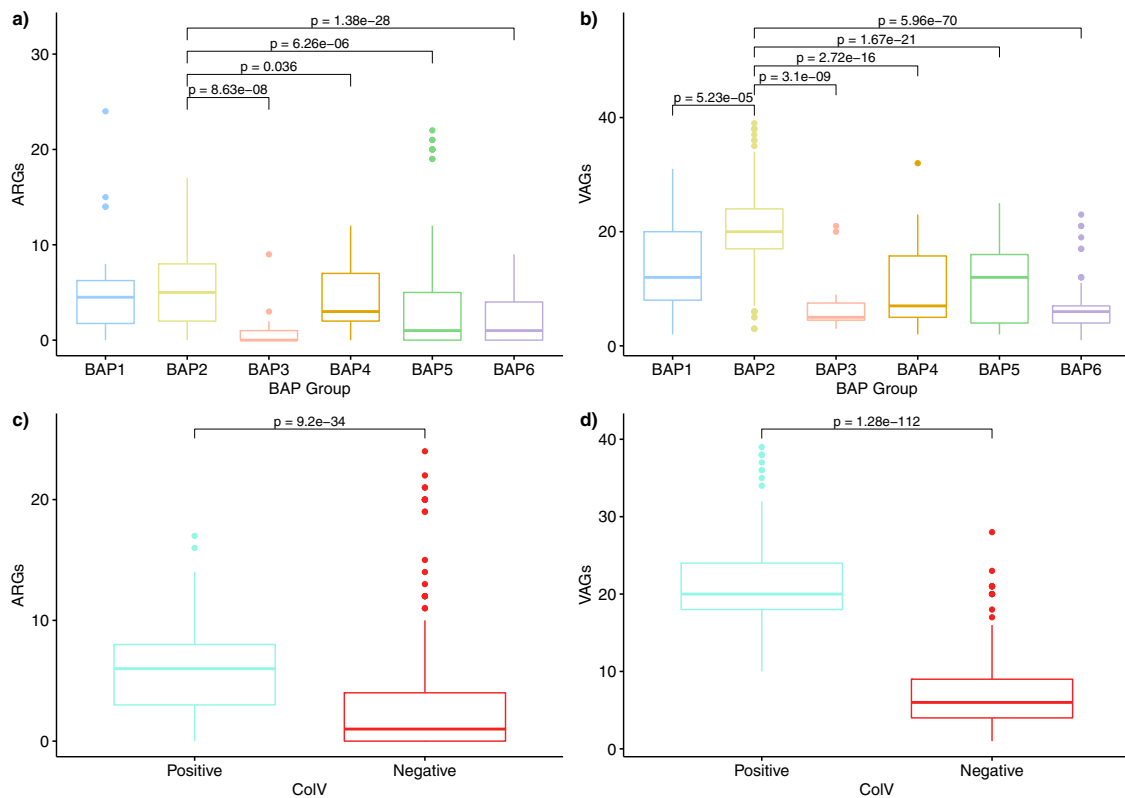


Fig. 7 Relationships of ARG and VAG carriage with BAP cluster and ColV carriage. Boxplots comparing total ARG and VAG counts by BAP clusters (**a, b**; $n = 750$ biologically independent ST58 genome sequences) and ColV status for (**c, d**; $n = 752$ biologically independent ST58 genome sequences). **a** ARG carriage by BAP cluster (Pairwise Wilcoxon test with Benjamini-Hochberg adjusted p -values); **b** VAG carriage by BAP cluster; (Pairwise Wilcoxon test with Benjamini-Hochberg adjusted p -values); **c** ARG carriage by ColV carriage (Two-sided Wilcoxon Rank-sum test); **d** VAG carriage by ColV carriage (Two-sided Wilcoxon Rank-sum test). Upper and lower limits of the boxes represent 75th and 25th quartile, centre line represents the median, whiskers extend to $1.5 \times$ IQR and values outside these ranges are represented by dots. ARG maximum value is 24 and minimum value is 0. VAG maximum value is 39 and minimum value is 1. P -values for significant differences are shown. (n.b. Only significant differences between BAP2 and other groups are annotated with p -values in **a** and **b**).

(2 faecal swabs, 1 sputum; patient health status unknown) from Thailand with close linkages to poultry and porcine strains from the United States as well as with water and bovine origin strains from Sweden. Among these, ExPEC strain ESBL20140051 from Denmark was separated by only a single SNP from three water strains and one bovine strain, all of which were from Sweden. This is particularly striking given the low numbers of human-origin sequences in the collection and implies that far more genomic links exist between ST58 that cause infections and those present in animal and environmental reservoirs.

Discussion

Here, in a genomic epidemiological study, we examined 752 whole genomes of *E. coli* ST58 from a variety of sources in order to provide some explanation for its emergence as a human pathogen. We identified the BAP2 cluster; a large, divergent lineage of strains with a broad host range. This cluster had near-ubiquitous carriage of ColV-like plasmids, reduced diversity of adaptive colonisation-related traits such as *fimH* and serotypes, and distinctive accessory gene content, including carriage of yersiniabactin genomic island marker genes and an additional prophage. Among the collection's strains, the BAP2 cluster contained the vast majority of strains isolated from extra-intestinal infections, poultry and swine. Our data indicate that complex interactions between mobile genetic elements, phylogenomic background, and various hosts comprise the mechanisms and

networks through which *E. coli* ST58 has emerged as a human pathogen⁵⁶.

Limitations. Epidemiological studies that leverage publicly available genomic data are typically beset by uncertainty as to whether the total dataset represents a reliable proxy for the genomic and source diversity of the entire population. Whilst our source distribution was dominated by cattle, poultry and swine, we believe we have observed most of the genomic diversity present in ST58 because of the large pangenome, variety of ARGs, VAGs and plasmid replicons, and clear population structure. The major question remaining is whether most ST58 from human faeces and ExPEC belong to the BAP2 cluster. Though not included in our study collection, ST58 sepsis isolates from multiple hospitals in Paris predominantly display O8/O9:H25 serotypes, as well as ColV and HPI marker genes, all of which were typical of BAP2 ST58³. In conjunction with our data, this tends to suggest that most ST58 ExPEC will belong to this cluster, however, an expanded dataset with more commensal and pathogenic human sequences would increase our confidence. Some form of virulence characterisation of non-clinical isolates in the BAP2 cluster would have been desirable in order to solidify the linkage between their genomic traits and phenotypes; however, this was outside the scope of the study. Nonetheless, the sharing of genomic elements with well-documented virulence traits between clinical and non-clinical isolates in the cluster strongly supports a degree of innate virulence. Similarly, though

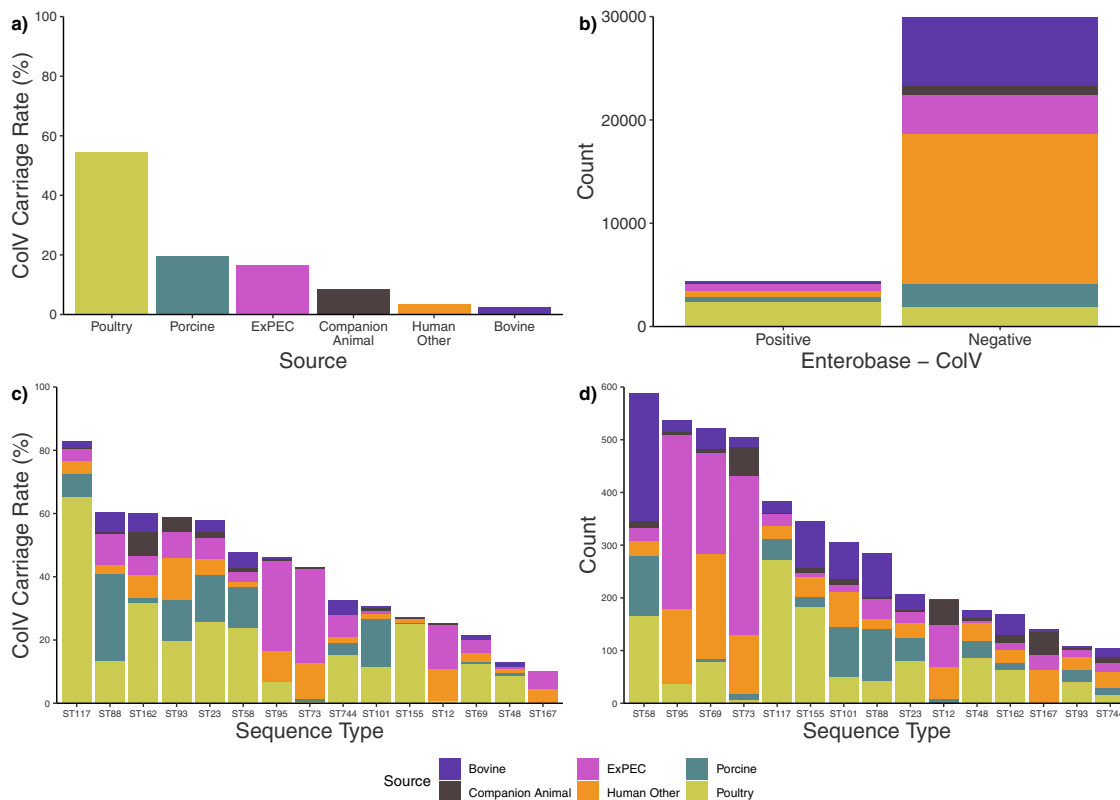


Fig. 8 ColV carriage in 34,364 *E. coli* genomes. Summary of sources, STs and ColV carriage from 34,364 Enterobase *E. coli* assemblies. **a** Proportional ColV carriage in each source; **b** Absolute ColV carriage stratified by source; **c** Proportional ColV carriage in the most abundant ColV carrying STs, stratified by source; **d** Absolute counts of STs from **c**, stratified by source.

antimicrobial resistance phenotypes were not available for all strains, concordance between acquired ARG presence and resistance phenotype is typically very high^{57,58}.

ColV plasmids in ST58. Our phylogenetic analysis revealed that the BAP2 cluster of ST58 strains has acquired a diversity of ColV plasmids. The majority of the ColV plasmids observed carried the full repertoire of archetypal ColV genes and operons. The importance of ColV and plasmids generally in the evolution of multiple pandemic ExPEC lineages is increasingly acknowledged^{49,59,60}. Carriage of ColV plasmids may prime strains to cause extra-intestinal infections in humans. ColV plasmids typically encode multiple genetic loci including siderophores involved in iron acquisition and transport (aerobactin and salmochelin), *iss* conferring copy number-associated increased survival in human serum, outer membrane vesicle and protease production (*hlyF* and *ompT*) and putative ABC transporter *ets*⁴⁸. This gene repertoire allows strains that acquire them to develop increased virulence in animal models of neonatal meningitis, urinary tract infection and sepsis^{53,61}, increased killing of chicken embryos, increased growth in urine and colonise the murine kidney⁵¹. Aerobactin specifically is associated with cystitis, pyelonephritis and bacteraemia⁶². Carriage of yersinia-bactin (*fyuA*, *irp2*, HPI) by the vast majority of strains in the BAP2 cluster also strongly supports their innate virulence⁴⁶. Furthermore, the ColV-based *sit* operon is associated with innate virulence⁴⁶. In addition to pathogenic properties, ColV+ *E. coli* have been shown to outcompete ColV- strains in the human gut⁶³. Aerobactin and salmochelin carriage are likely to play a role here, being implicated in intestinal persistence and gut colonisation^{64–66}. In this regard, ColV carriage may not only

contribute to BAP2 ST58 pathogenicity but also to lineage expansion via the presence of factors advantageous for intestinal colonisation and persistence.

Implications for antimicrobial resistance. Our findings, together with the literature, suggest the acquisition of ColV plasmids heightens the likelihood that a strain will also acquire AMR determinants. In our ST58 genome collection, we observed that, on average, ColV+ strains carried more ARGs than ColV- strains. In line with this, ARG loci primed for further resistance gene acquisition (e.g. co-carriage of *intI1* (class 1 integron), *dfrA5* (a trimethoprim resistance gene) and a globally disseminated IS26 deletion signature⁶⁷) could be localised to the same assembly contig as ColV genes and were observed in nearly a third of our ColV+ sequences. ARG loci such as these on plasmids tend to act as hotspots for stepwise gene acquisition via multiple mechanisms^{68–70}. In healthy humans, ARG loci evolving in conjunction with diverse mobile genetic elements (MGEs), such as IS26, on different lineages of ColV plasmids have been described in commensal strains of *E. coli* isolated from the faeces of healthy humans^{48,71,72}. ColV plasmids also circulate in the environment with determinants conferring resistance to critically important antimicrobials—observed in an MDR ST58 strain isolated from a pig⁷³ (ColV-like plasmid carrying *bla*_{CTX-M-15}) and an MDR *E. coli* ST131-H22 strain isolated from agricultural soil⁷⁴ (a transposable *ISEcp1-bla*_{CTX-M-15} unit inserted upstream of an IS26-truncated copy of *Tn2* on an F2:B1 ColV plasmid backbone). Overall this highlights the risk of ColV plasmids acting as backbones for the acquisition and dissemination of ARGs, though they are most likely selected for traits other than AMR in the first instance.

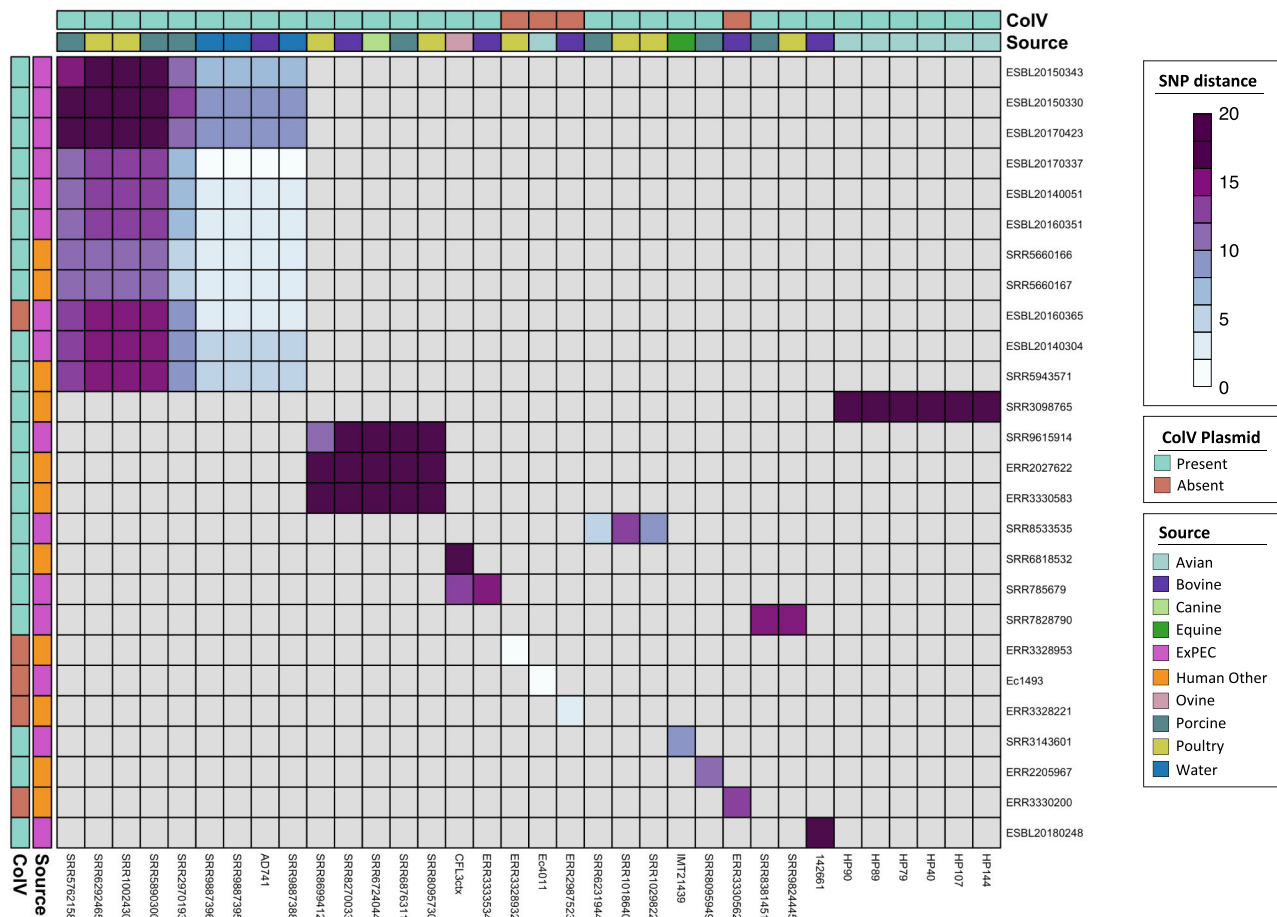


Fig. 9 Close relationships between human and non-human ST58. Heatmap of human vs. non-human strain pairs differing by ≤ 20 SNPs. Scale from white to dark purple represents SNPs from lowest to highest. Grey squares represent more distantly related sequence pairs. Heatmap is clustered to show groups of closely related sequences.

Our analysis of ST58 highlights the importance of screening for AMR in ExPEC without selective bias. Many studies select isolates for study on the basis of their carrying ESBL-encoding genes. However, such methodology can overlook precursor conditions that allow for resistance to critically important antimicrobials to be acquired. This is demonstrated above by ColV carriage as well as Inc1 plasmids carried by ST58. Inc1 plasmids are known primarily for their carriage of ESBL genes (*bla*_{CTX-M-1}, *bla*_{CTX-M-15} and *bla*_{CMY-2}). The extent of reporting on ESBL-carrying Inc1 plasmids in ST58 would suggest these plasmids are common in ST58^{8,13,35,39}. Interestingly, we found that whilst Inc1 plasmids are found in 30% of ST58, no ESBL gene was present in greater than 8% of sequences. This suggests that Inc1 plasmids were selected in ST58 in sources where ESBL-selection pressure was not a factor, yet these plasmids may provide a convenient scaffold for ESBL gene acquisition when the relevant selection pressure is encountered. This is conceptually reflective of the evolution of ARG loci on some ColV plasmids, which evidently occurred subsequent to their primary evolution and widespread dissemination^{48,66}. Thus, understanding the ecology and lineage associations of plasmids without antimicrobial selection bias is critical to a holistic understanding of factors that influence the emergence and spread of ARGs.

Ecology of ColV plasmids. Our findings illustrate the relationship between the source, MGEs (mediators of horizontal gene

transfer) and genomic background in pathogen emergence. The Enterobase analysis supports the role of ColV in pathogenicity in a proportion of human extra-intestinal infections. Of 34,364 draft *E. coli* genome assemblies from Enterobase, ColV plasmids were present in 16% of ExPEC strains. In ST58, we observed the presence of ColV plasmids in more than 75% of the ExPEC strains. However, ExPEC strains are not the dominant reservoirs of ColV plasmids. Rather, our data supports the widely held opinion that poultry and associated meat are the major reservoirs for ColV plasmids with a carriage rate of 55%^{54,75}. Unexpectedly, we found that porcine sources were also a non-trivial secondary reservoir of ColV plasmids, with 20% carriage. ColV plasmids have previously been reported in pigs; however, it is unclear whether ColV plasmids were historically common in production systems or whether they represent a relatively recent, direct or indirect incursion from poultry^{76,77}. Further analysis of plasmid diversity in the species collection could be useful in this regard. Consistent with the Enterobase analysis, our collection of 752 genomes of ST58 revealed that poultry and porcine-sourced ST58 were more often ColV+ than ColV-. Interaction between ColV plasmids, phylogenomic background and source was also evident in the variable proportions of isolation sources displayed by the major ColV+ STs.

To what extent is the human gut a reservoir of ColV+ ST58 and ColV plasmids in general? It remains difficult to quantify. Human faecal commensals are the origin of the vast majority of ExPEC infections^{78,79} and yet very few sequences are available or

identifiable online due to the strong bias towards human diarrhoeagenic isolates and lack of metadata, respectively. Nonetheless, evidence pointing to the human gut as a reservoir comes from our study and others^{48,63}. As previously mentioned, the carriage of multiple siderophore systems such as aerobactin and salmochelin is considered to be advantageous for gut colonisation^{64–66}. Detailed information on ColV carriage in large collections of geographically diverse human faecal commensals is urgently required as it is the critical missing link between primary sources of ColV plasmids and ColV+ ExPEC infections.

Plasmids, sources and phylogeny in pathogen emergence. The importance of interrelation between plasmids and sources for understanding pathogen emergence is highlighted by recent literature. Evolutionary models demonstrate the role of horizontal gene transfer and microbial migration in decoupling ecologically important traits from genomic backgrounds in a specific niche and allowing horizontal sweeps of these traits under selective pressure⁸⁰. Experimental work similarly demonstrated that different microbial habitats carry distinct sets of MGEs, which could drive the evolution of immigrant *E. coli* by providing a pool of MGE-associated functional elements advantageous to survival in that niche⁸¹. Shaw and Matlock have also reported the effect of niche on the distribution of Enterobacterial plasmids and F plasmids, respectively^{82,83}.

The dominance of poultry and porcine sources in the ST58 BAP2 cluster is therefore likely to be because ColV plasmids and other BAP2-associated genes are abundant and provide a fitness advantage within niches associated with these hosts. The lower diversity of key adaptive traits (e.g. *fimH* alleles and serotypes) in the BAP2 cluster compared to other clusters is suggestive of evolutionary convergence commensurate with selective pressure exerted by a specific host or hosts^{84,85}. Overall this suggests that poultry and swine have influenced the evolution and emergence of BAP2 ST58.

As well as source, the phylogenomic background is also known to affect the propensity for certain lineages to horizontally acquire genetic material and conversely horizontal acquisition influences core genomic changes⁸⁶. Phylogroup B1, to which ST58 belongs, is considered to be particularly receptive to horizontally acquired genetic material⁸¹. The abundance of ColV plasmids in one particular lineage however suggests that BAP2 ST58 may be more susceptible to plasmid acquisition, more suited to niches where ColV plasmids are abundant or a combination of both. Similarly, the phylogenetic divergence of the cluster may have been influenced by compensatory SNPs accumulated as a result of the acquisition of ColV plasmids and other MGEs⁸⁶.

With all of the above in mind, it appears that the ST58 BAP2 cluster has emerged via multiple horizontal acquisitions of ColV plasmids, phage and additional accessory genes as well as core genomic adaptations both favourable to and resulting from these evolutionary events. We strongly suspect that lineage expansion has been significantly influenced by ColV plasmids and their association with hosts such as poultry and swine. Siderophore iron acquisition systems found on ColV plasmids and the HPI are also likely to have been influential via their conferral of both fitness-related and pathogenic traits. The order in which specific genomic events actually occur in conjunction with transfer between different sources is unknown, but can be conceptualised as a constant push-pull relationship between core genomic adaptations and HGT occurring within a complex network of interacting sources and MGEs under selection within and between them.

Conclusion

Our epidemiological data indicate that *E. coli* ST58 has emerged as a prominent sequence type and human pathogen through the

interplay of mobile genetic elements, ecology, and genomic background. We identified a sub-lineage of ST58 that has acquired extrachromosomal and chromosomally-located mobile genetic elements, most notably a diversity of ColV plasmids and the Yersiniabactin High Pathogenicity Island. Poultry and swine are implicated in the emergence of this distinctive sub-lineage, with the contribution of the collective human gut yet to be fully appreciated. Carriage of ColV plasmids in ST58 exemplify selection offering *E. coli* fitness advantages where it is commensal or free-living and virulence capabilities once in the urinary tract or blood. Although ColV plasmids may not be primarily selected for their capacity to confer resistance to antimicrobials, their frequent carriage of ARGs is an additional threat to those who develop infections caused by ColV+ ST58. Our study highlights the importance of genomic epidemiological investigations that avoid antimicrobial selection bias and take plasmid ecology and non-human sources into consideration in order to develop a holistic understanding of factors influencing *E. coli* pathogenesis.

Methods

Collaborator sequences. A global collection of 190 ST58 *E. coli* sequences was compiled from in-house datasets and collaborators in Australia, Canada, Czech Republic, Denmark, France, Germany, Netherlands, Poland, Romania, Sweden, United Kingdom and United States. These sequences represented a wide range of sources including human urinary tract and blood infections, wild birds, cattle, companion animals, produce, swine, poultry, wastewater and surface water. Detailed methods regarding culture and DNA isolation for the generation of in-house sequences are available in their respective publications listed as PMID accessions in Supplementary Data 1. All data from collaborators were sequenced on Illumina platforms (see SRA entries for specific instruments) and was received as raw *fastq* paired reads. Individual strain and DNA isolation methods were not provided for collaborator sequences. In-house and collaborator reads were directly uploaded to an Enterobase workspace for quality control (QC) and preliminary analyses in the context of other ST58 sequences. Twelve sequences were excluded as they were determined not to belong to ST58, failed assembly or QC performed by Enterobase. Eight sequences were excluded in preliminary phylogenetic analyses. The final number of 'in-house' sequences in the collection was 178. Of these, 158 were previously unpublished and have been uploaded to SRA under BioProject PRJNA727368. Full metadata and accession numbers for all sequences are available in Supplementary Data 1.

Publicly available sequences. Enterobase was queried on 04/12/2019 for released ST58 whole-genome sequences with available metadata for source, collection year, continent and country. SRA and ENA accession numbers were extracted and parallel-fastq-dump 0.6.6 was used to download 614 read sets with the following flags: `-skip-technical,-read-filter pass,-dumpbase,-split-files,-clip`. Enterobase sequences were named for analysis with their NCBI SRA or EBI ENA accession number beginning with SRR or ERR, respectively. Full metadata and accession numbers for these sequences are available in Supplementary Data 1.

Curation and metadata processing. To create consistency and ensure cogency within source information from Enterobase we curated the metadata in R 3.6.3 with a range of regular expression substitutions and some unavoidable manual curation utilising the three source information columns from Enterobase to classify sequences by niche and source. Defined niches and their respective sources in parentheses were: livestock (bovine, equine, porcine, poultry, livestock other, ovine); human (ExPEC, human other); wild animal (avian, other mammal); companion animal (canine, feline, avian); food (plant); environment (water, wastewater). Forty sequences were excluded during this process due to irreconcilable source information. The scripts used to process the raw metadata, as well as the pre- and post-manual curation datasets have been made available as a GitHub repository for reproducibility and transparency (See 'Data availability' and 'Code availability' below). The final collection numbered 752 sequences comprising 178 in-house/collaborator sequences and 574 publicly available sequences.

De novo assembly and annotation. Raw sequence reads were de novo assembled with Shovill 1.0.4 with default settings and a minimum contig length of 200 bp. The resulting assemblies were annotated by Prokka 1.14.5, run with default settings and the `-kingdom` and `-genus` flags set to 'Bacteria' and 'Escherichia', respectively.

Core and pangenome identification. The collection of 752 prokka-annotated sequences and outgroup strain MOD-EC5019 (ST155) were analysed with Roary 3.13.0 to infer core and pangenomes of *E. coli* ST58⁸⁷. `-v` and `-e` flags were used to produce an alignment of core genes with MAFFT 7.455, which formed the basis of

phylogenomic and SNP analyses described below. Aside from these flags, default settings were used with paralog splitting on, 95% minimum identity for BLASTp and core genes defined as those present in 99% of isolates (this prevents the outgroup strain from affecting core genome estimation). The resulting core gene alignment comprised 2,807,790 bp and 3023 core genes.

Phylogenomic and SNP analyses. A maximum-likelihood phylogenetic tree was inferred from core gene alignment with IQTree 2.0.3 using the GTR + F + R4 model of nucleotide substitution and 1000 bootstrap replicates and rooted on the outgroup strain. fastbaps 1.0.6 was utilised to define clusters from the core gene alignment conditioned on the core gene tree and these designations were subsequently used in downstream analyses⁸⁸. SNPs were identified by running snp-sites 2.5.1 on the core gene alignment, generating a core SNP alignment of 30,771 SNP sites. Pairwise SNPs between strains were counted from the core SNP alignment with snp-dists 0.6.3.

Pan-GWAS. Scoary calculates associations between gene presence/absence (as determined by Roary) and phenotypic or metadata traits of the strains. We defined the fastbaps clusters as traits and used Scoary 1.6.16 with the `-no_pairwise` flag (to identify over-representation as opposed to causation) and a Benjamini–Hochberg corrected *p*-value threshold of 1E-50 to identify genes that were over- or under-represented in those clusters.

Gene and mutation screening. Antimicrobial resistance genes, virulence-associated genes, plasmid replicons, F plasmid types and serotypes were identified via a read-mapping approach with ARIBA 2.13.3⁸⁹, using publicly available ResFinder, VirulenceFinder, PlasmidFinder and SerotypeFinder databases from the Centre for Genomic Epidemiology and custom databases available at https://github.com/maxcummins/custom_DBs/blob/master/EC_custom.fa^{90–94}. SNP-mediated resistance genotypes were predicted with PointFinder 3.1.0⁹⁵, downloaded as a Python script from <https://bitbucket.org/genomicEpidemiology/pointfinder/src/master/>. *fimH* alleles were identified by Enterobase. ABRicate 0.9.8, a BLAST-based screening tool, was additionally used with default settings to identify Inc11 pMLST and the *dfrA5-IS26* deletion signature.

Inference of ColV plasmid carriage. We used ABRicate 0.9.8 with default settings to implement the ColV screening methodology described by Liu *et al* to identify de novo assemblies that putatively carried a ColV plasmid⁷⁵. Briefly, the Liu criteria considers a strain ColV-positive if it carries at least one or more genes from four or more of the following six gene sets (i) *cvaABC* and *cvi* (the ColV operon), (ii) *iroBCDEN* (the salmochelin operon), (iii) *iucABCD* and *iutA* (the aerobactin operon), (iv) *etsABC*, (v) *ompT* and *hlyF*, and (vi) *sitABCD*. Thresholds of $\geq 90\%$ nucleotide identity and $\geq 95\%$ length coverage were applied post-ABRicate in RStudio 1.4.1106 with R 4.0.5 to determine positive hits for a gene and sum the group counts to infer presence or absence.

To strengthen the evidence for ColV plasmids in our collection we aligned assemblies to the backbone of pCERC4, an archetypal F2:B1 ColV plasmid isolated from an ST95 *E. coli* strain from healthy human faecal flora. Briefly, the 126,270 bp pCERC4 backbone was defined by removing resistance regions from the complete pCERC4 sequence (RefSeq: NZ_KU578032.1) as described by Moran and Hall⁴⁸. Assemblies were aligned against the backbone with ABRicate and outputs concatenated for import into R. We used a custom R script to extract hit ranges for each sequence and represented them as a series of percentage nucleotide identities across 100 bp segments of the plasmid backbone. These percentages were then visualised against a schematic backbone of pCERC4, generated with SnapGene 5.2 (GSL Biotech), as a heatmap in gtree.

Enterobase ColV analysis. To compare ColV carriage across *E. coli* STs we filtered the Enterobase *Escherichia/Shigella* database for strains deposited prior to January 12, 2020. To be included in the study, strains were required to have metadata relating to their source, geographical and temporal origins. Serotypes, Achtman MLST and *fimH* alleles were also downloaded from Enterobase. Sources were curated in R as well as manually in Excel in order to reduce the heterogeneous Enterobase metadata into six categories: Bovine, Porcine, Poultry, ExPEC, Human Other and Companion Animal. Following removal of *Shigella sp.*, 35,254 genomic assemblies were then downloaded from Enterobase using a third-party custom script available at <https://github.com/C-Connor/EnterobaseGenomeAssemblyDownload>. Genomes less than 4.5 Mbp or greater than 6.5 Mbp were excluded. ABRicate 0.9.8 was used to screen for ColV related genes and the Liu criteria was applied to infer ColV carriage as described above. Metadata, accession numbers and additional information on the cohort of strains under analysis are available in Supplementary Data 4.

Statistics. All statistical analysis were performed in R 4.0.5 and scripts are available as described in ‘Code availability’ below. The Kruskal–Wallis test was used to determine whether the mean of total ARGs or VAGs per sequence was different based on BAP group membership and between ColV+/- status. Pairwise

Wilcoxon test with Benjamini–Hochberg *p*-value correction for multiple testing was then used to calculate the significance of the pairwise differences between BAP groups. Two-sided Wilcoxon rank-sum test was used to calculate the significance of the difference between ColV+ and ColV- sequences.

Data processing and visualisation. All processing, analyses and data visualisation were performed in RStudio 1.4.1106 with R 4.0.5 and can be reproduced with the scripts linked in ‘Code availability’. Some manual editing of figures for consistency of presentation was performed in Microsoft PowerPoint.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All ST58 raw sequence read data generated for the first time in this study have been deposited in NCBI BioProject [PRJNA4727368](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA4727368) with individual accession numbers available in Supplementary Data 1. All other ST58 raw sequence read data used in this study are available via either the NCBI Sequence Read Archive [<https://www.ncbi.nlm.nih.gov/sra>] or the EMBL-EBI European Nucleotide Archive [<https://www.ebi.ac.uk/ena/browser/home>] via accession numbers listed in Supplementary Data 1. Genome assemblies from the Enterobase collection are available at <https://enterobase.warwick.ac.uk/species/index/ecoli> via assembly barcodes listed in Supplementary Data 4. Source data are provided with this paper.

Code availability

To support the reproducibility of this work we have made all the raw data generated from the original sequence reads, and the R scripts used to process, analyse and visualise this data available at https://github.com/CJREID/ST58_project.

Received: 2 June 2021; Accepted: 11 January 2022;

Published online: 03 February 2022

References

1. Foxman, B. The epidemiology of urinary tract infection. *Nat. Rev. Urol.* **7**, 653–660 (2010).
2. Manges, A. R. *et al.* Global extraintestinal pathogenic *Escherichia coli* (ExPEC) lineages. *Clin. Microbiol. Rev.* **32**, 1–25 (2019).
3. Royer, G. *et al.* Phylogroup stability contrasts with high within sequence type complex dynamics of *Escherichia coli* bloodstream infection isolates over a 12-year period. *Genome Med.* **13**, 77 (2021).
4. McKinnon, J., Roy Chowdhury, P. & Djordjevic, S. P. Genomic analysis of multidrug-resistant *Escherichia coli* ST58 causing urosepsis. *Int. J. Antimicrob. Agents* **52**, 430–435 (2018).
5. Roer, L. *et al.* WGS-based surveillance of third-generation cephalosporin-resistant *Escherichia coli* from bloodstream infections in Denmark. *J. Antimicrob. Chemother.* **72**, 1922–1929 (2017).
6. Flament-Simon, S.-C. *et al.* High prevalence of ST131 subclades C2-H30Rx and C1-M27 among extended-spectrum β -lactamase-producing *Escherichia coli* causing human extraintestinal infections in patients from two hospitals of Spain and France during 2015. *Front. Cell Infect. Microbiol.* **10**, 125 (2020).
7. Mamani, R. *et al.* Sequence types, clonotypes, serotypes, and virotypes of extended-spectrum β -lactamase-producing *Escherichia coli* causing bacteraemia in a Spanish hospital over a 12-year period (2000 to 2011). *Front. Microbiol.* **10**, 1530 (2019).
8. Pietsch, M. *et al.* Whole genome analyses of CMY-2-producing *Escherichia coli* isolates from humans, animals and food in Germany. *BMC Genomics* **19**, 601 (2018).
9. Day, M. J. *et al.* Population structure of *Escherichia coli* causing bacteraemia in the UK and Ireland between 2001 and 2010. *J. Antimicrob. Chemother.* **71**, 2139–2142 (2016).
10. Wang, S. *et al.* Antimicrobial resistance and molecular epidemiology of *Escherichia coli* causing bloodstream infections in three hospitals in Shanghai, China. *PLoS ONE* **11**, e0147740 (2016).
11. Ireng, L. M. *et al.* Whole-genome sequences of multidrug-resistant *Escherichia coli* in South-Kivu Province, Democratic Republic of Congo: characterization of phylogenomic changes, virulence and resistance genes. *BMC Infect. Dis.* **19**, 137 (2019).
12. Paramita, R. I. *et al.* Genome-based characterization of *Escherichia coli* causing bloodstream infection through next-generation sequencing. *PLoS ONE* **15**, e0244358 (2020).
13. Dahmen, S., Métayer, V., Gay, E., Madec, J.-Y. & Haenni, M. Characterization of extended-spectrum beta-lactamase (ESBL)-carrying plasmids and clones of

- Enterobacteriaceae causing cattle mastitis in France. *Vet. Microbiol.* **162**, 793–799 (2013).
14. Nüesch-Inderbinen, M. et al. Molecular types, virulence profiles and antimicrobial resistance of *Escherichia coli* causing bovine mastitis. *Vet. Rec. Open* **6**, e000369 (2019).
 15. Ali, A., Ali, Q., Ali, R. & Mohsin, M. Draft genome sequence of an extended-spectrum β -lactamase-producing *Escherichia coli* ST58 isolate from cattle in Pakistan. *J. Glob. Antimicrob. Resist.* **21**, 303–305 (2020).
 16. Borges, C. A., Tarlton, N. J. & Riley, L. W. *Escherichia coli* from commercial broiler and backyard chickens share sequence types, antimicrobial resistance profiles, and resistance genes with human extraintestinal pathogenic *Escherichia coli*. *Foodborne Pathog. Dis.* **16**, 813–822 (2019).
 17. Maamar, E. et al. High prevalence of extended-spectrum and plasmidic AmpC β -lactamase-producing *Escherichia coli* from poultry in Tunisia. *Int. J. Food Microbiol.* **231**, 69–75 (2016).
 18. Chah, K. F. et al. Detection and molecular characterisation of extended-spectrum β -lactamase-producing enteric bacteria from pigs and chickens in Nsukka, Nigeria. *J. Glob. Antimicrob. Resist.* **15**, 36–40 (2018).
 19. Song, J., Oh, S.-S., Kim, J., Park, S. & Shin, J. Clinically relevant extended-spectrum β -lactamase-producing *Escherichia coli* isolates from food animals in South Korea. *Front. Microbiol.* **11**, 604 (2020).
 20. Ahmed, S., Olsen, J. E. & Herrero-Fresno, A. The genetic diversity of commensal *Escherichia coli* strains isolated from non-antimicrobial treated pigs varies according to age group. *PLoS ONE* **12**, e0178623 (2017).
 21. Blaak, H. et al. Detection of extended-spectrum beta-lactamase (ESBL)-producing *Escherichia coli* on flies at poultry farms. *Appl. Environ. Microbiol.* **80**, 239–246 (2014).
 22. Vignaroli, C. et al. Multidrug-resistant and epidemic clones of *Escherichia coli* from natural beds of Venus clam. *Food Microbiol.* **59**, 1–6 (2016).
 23. Cortés-Cortés, G. et al. Detection and molecular characterization of *Escherichia coli* strains producers of extended-spectrum and CMY-2 type beta-lactamases, isolated from Turtles in Mexico. *Vector Borne Zoonotic Dis.* **16**, 595–603 (2016).
 24. Furlan, J. P. R., Lopes, R., Gonzalez, I. H. L., Ramos, P. L. & Stehling, E. G. Comparative analysis of multidrug resistance plasmids and genetic background of CTX-M-producing *Escherichia coli* recovered from captive wild animals. *Appl. Microbiol. Biotechnol.* <https://doi.org/10.1007/s00253-020-10670-4> (2020).
 25. Fuentes-Castillo, D. et al. Genomic characterization of multidrug-resistant ESBL-producing *Escherichia coli* ST58 causing fatal colibacillosis in critically endangered Brazilian merganser (*Mergus octosetaceus*). *Transbound. Emerg. Dis.* **68**, 258–266 (2021).
 26. Mattioni Marchetti, V. et al. Deadly puppy infection caused by an MDR *Escherichia coli* O39 blaCTX-M-15, blaCMY-2, blaDHA-1, and aac(6)-Ib-cr - positive in a breeding Kennel in Central Italy. *Front. Microbiol.* **11**, 584 (2020).
 27. de Carvalho, M. P. N. et al. International clones of extended-spectrum β -lactamase (CTX-M)-producing *Escherichia coli* in peri-urban wild animals, Brazil. *Transbound Emerg. Dis.* <https://doi.org/10.1111/tbed.13558> (2020).
 28. Batalha de Jesus, A. A. et al. High-level multidrug-resistant *Escherichia coli* isolates from wild birds in a large urban environment. *Microb. Drug Resist.* **25**, 167–172 (2019).
 29. Wyrsh, E. R. et al. Whole-genome sequence analysis of environmental *Escherichia coli* from the faeces of straw-necked ibis (*Threskiornis spinicollis*) nesting on inland wetlands. *Microb. Genom.* **6**, 1–16 (2020).
 30. Wyrsh, E. R., Chowdhury, P. R., Jarocki, V. M., Brandis, K. J. & Djordjevic, S. P. Duplication and diversification of a unique chromosomal virulence island hosting the subtilase cytotoxin in *Escherichia coli* ST58. *Microb. Genom.* **6**, (2020).
 31. Zurfluh, K. et al. Antimicrobial resistant and extended-spectrum β -lactamase producing *Escherichia coli* in common wild bird species in Switzerland. *Microbiologyopen* **8**, e845 (2019).
 32. Jamborova, I. et al. Plasmid-mediated resistance to cephalosporins and fluoroquinolones in various *Escherichia coli* sequence types isolated from rooks wintering in Europe. *Appl. Environ. Microbiol.* **81**, 648–657 (2015).
 33. Skurnik, D. et al. Emergence of antimicrobial-resistant *Escherichia coli* of animal origin spreading in humans. *Mol. Biol. Evol.* **33**, 898–914 (2016).
 34. Zurfluh, K. et al. Key features of mcr-1-bearing plasmids from *Escherichia coli* isolated from humans and food. *Antimicrob. Resist. Infect. Control* **6**, 91 (2017).
 35. Ben Said, L. et al. Detection of extended-spectrum beta-lactamase (ESBL)-producing Enterobacteriaceae in vegetables, soil and water of the farm environment in Tunisia. *Int. J. Food Microbiol.* **203**, 86–92 (2015).
 36. Nüesch-Inderbinen, M., Treier, A., Zurfluh, K. & Stephan, R. Raw meat-based diets for companion animals: a potential source of transmission of pathogenic and antimicrobial-resistant Enterobacteriaceae. *R. Soc. Open Sci.* **6**, 191170 (2019).
 37. Reid, C. J., Blau, K., Jechalke, S., Smalla, K. & Djordjevic, S. P. Whole genome sequencing of *Escherichia coli* from store-bought produce. *Front. Microbiol.* **10**, 3050 (2019).
 38. Ferjani, S. et al. Community fecal carriage of broad-spectrum cephalosporin-resistant *Escherichia coli* in Tunisian children. *Diagn. Microbiol. Infect. Dis.* **87**, 188–192 (2017).
 39. Ben Sallem, R. et al. IncI1 plasmids carrying bla(CTX-M-1) or bla(CMY-2) genes in *Escherichia coli* from healthy humans and animals in Tunisia. *Microb. Drug Resist.* **20**, 495–500 (2014).
 40. Sütterlin, S. et al. Silver resistance genes are overrepresented among *Escherichia coli* isolates with CTX-M production. *Appl. Environ. Microbiol.* **80**, 6863–6869 (2014).
 41. Teunis, P. F. M. et al. Time to acquire and lose carriage of ESBL/pAmpC producing *E. coli* in humans in the Netherlands. *PLoS ONE* **13**, e0193834 (2018).
 42. Chen, P.-A. et al. Characteristics of CTX-M extended-spectrum β -lactamase-producing *Escherichia coli* strains isolated from multiple rivers in Southern Taiwan. *Appl. Environ. Microbiol.* **82**, 1889–1897 (2016).
 43. Sacramento, A. G. et al. Genomic analysis of MCR-1 and CTX-M-8 co-producing *Escherichia coli* ST58 isolated from a polluted mangrove ecosystem in Brazil. *J. Glob. Antimicrob. Resist.* **15**, 288–289 (2018).
 44. Ben Said, L. et al. Characteristics of extended-spectrum β -lactamase (ESBL)- and pAmpC beta-lactamase-producing Enterobacteriaceae of water samples in Tunisia. *Sci. Total Environ.* **550**, 1103–1109 (2016).
 45. Denamur, E., Clermont, O., Bonacorsi, S. & Gordon, D. The population genetics of pathogenic *Escherichia coli*. *Nat. Rev. Microbiol.* <https://doi.org/10.1038/s41579-020-0416-x> (2020).
 46. Galardini, M. et al. Major role of iron uptake systems in the intrinsic extraintestinal virulence of the genus *Escherichia* revealed by a genome-wide association study. *PLoS Genet.* **16**, e1009065 (2020).
 47. Manges, A. R. *Escherichia coli* causing bloodstream and other extraintestinal infections: tracking the next pandemic. *Lancet Infect. Dis.* **19**, 1269–1270 (2019).
 48. Moran, R. A. & Hall, R. M. Evolution of regions containing antibiotic resistance genes in FII-2-FIB-1 ColV-Colla virulence plasmids. *Microb. Drug Resist.* **24**, 411–421 (2018).
 49. Johnson, T. J. Role of plasmids in the ecology and evolution of “High-Risk” extraintestinal pathogenic *Escherichia coli* clones. *Ecosal Plus* **9**, 1–17 (2021).
 50. Johnson, T. J. et al. Horizontal gene transfer of a ColV plasmid has resulted in a dominant avian clonal type of *Salmonella enterica* serovar Kentucky. *PLoS ONE* **5**, e15524 (2010).
 51. Skyberg, J. A. et al. Acquisition of avian pathogenic *Escherichia coli* plasmids by a commensal *E. coli* isolate enhances its abilities to kill chicken embryos, grow in human urine, and colonize the murine kidney. *Infect. Immun.* **74**, 6287–6292 (2006).
 52. Xu, W.-Y., Li, Y.-J. & Fan, C. Different loci and mRNA copy number of the increased serum survival gene of *Escherichia coli*. *Can. J. Microbiol.* **64**, 147–154 (2018).
 53. Lemaitre, C. et al. A conserved virulence plasmidic region contributes to the virulence of the multiresistant *Escherichia coli* meningitis strain S286 belonging to phylogenetic group C. *PLoS ONE* **8**, e74423 (2013).
 54. Cummins, M. L. et al. Whole genome sequence analysis of Australian avian pathogenic *Escherichia coli* that carry the class 1 integrase gene. *Microb. Genom.* **5**, 1–13 (2019).
 55. Clermont, O. et al. Characterization and rapid identification of phylogroup G in *Escherichia coli*, a lineage with high virulence and antibiotic resistance potential. *Environ. Microbiol.* **21**, 3107–3117 (2019).
 56. Dunn, S. J., Connor, C. & McNally, A. The evolution and transmission of multi-drug resistant *Escherichia coli* and *Klebsiella pneumoniae*: the complexity of clones and plasmids. *Curr. Opin. Microbiol.* **51**, 51–56 (2019).
 57. Tyson, G. H. et al. WGS accurately predicts antimicrobial resistance in *Escherichia coli*. *J. Antimicrob. Chemother.* **70**, 2763–2769 (2015).
 58. Feldgarden, M. et al. Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrob. Agents Chemother.* **63**, (2019).
 59. Cummins, M. L., Reid, C. J. & Djordjevic, S. P. F Plasmid Lineages in *Escherichia coli* ST95: Implications for Host Range, Antibiotic Resistance, and Zoonoses. *mSystems.* <https://doi.org/10.1128/msystems.01212-21>, e0121221 (2022) Epub ahead of print.
 60. Koraimann, G. Spread and persistence of virulence and antibiotic resistance genes: a ride on the F plasmid conjugation module. *Ecosal Plus* **8**, 1–23 (2018).
 61. Agüero, M. E., de la Fuente, G., Vivaldi, E. & Cabello, F. ColV increases the virulence of *Escherichia coli* K1 strains in animal models of neonatal meningitis and urinary infection. *Med. Microbiol. Immunol.* **178**, 211–216 (1989).
 62. Johnson, J. R. Virulence factors in *Escherichia coli* urinary tract infection. *Clin. Microbiol. Rev.* **4**, 80–128 (1991).
 63. Smith, H. W. & Huggins, M. B. Further observations on the association of the colicine V plasmid of *Escherichia coli* with pathogenicity and with survival in the alimentary tract. *J. Gen. Microbiol.* **92**, 335–350 (1976).

64. Nowrouzian, F., Adlerberth, I. & Wold, A. E. P fimbriae, capsule and aerobactin characterize colonic resident *Escherichia coli*. *Epidemiol. Infect.* **126**, 11–18 (2001).
65. Nowrouzian, F., Wold, A. E. & Adlerberth, I. P fimbriae and aerobactin as intestinal colonization factors for *Escherichia coli* in Pakistani infants. *Epidemiol. Infect.* **126**, 19–23 (2001).
66. Searle, L. J., Méric, G., Porcelli, I., Sheppard, S. K. & Lucchini, S. Variation in siderophore biosynthetic gene distribution and production across environmental and faecal populations of *Escherichia coli*. *PLoS ONE* **10**, e0117906 (2015).
67. Dawes, F. E. et al. Distribution of class 1 integrons with IS26-mediated deletions in their 3'-conserved segments in *Escherichia coli* of human and animal origin. *PLoS ONE* **5**, e12754 (2010).
68. Harmer, C. J., Moran, R. A. & Hall, R. M. Movement of IS26-associated antibiotic resistance genes occurs via a translocatable unit that includes a single IS26 and preferentially inserts adjacent to another IS26. *MBio* **5**, e01801–e01814 (2014).
69. Harmer, C. J. & Hall, R. M. IS26-mediated formation of transposons carrying antibiotic resistance genes. *mSphere* **1**, 1–8 (2016).
70. Partridge, S. R., Zong, Z. & Iredell, J. R. Recombination in IS26 and Tn2 in the evolution of multiresistance regions carrying blaCTX-M-15 on conjugative IncF plasmids from *Escherichia coli*. *Antimicrob. Agents Chemother.* **55**, 4971–4978 (2011).
71. Moran, R. A., Anantham, S., Pinyon, J. L. & Hall, R. M. Plasmids in antibiotic susceptible and antibiotic resistant commensal *Escherichia coli* from healthy Australian adults. *Plasmid* **80**, 24–31 (2015).
72. Moran, R. A., Holt, K. E. & Hall, R. M. pCERC3 from a commensal ST95 *Escherichia coli*: a ColV virulence-multiresistance plasmid carrying a sul3-associated class 1 integron. *Plasmid* **84–85**, 11–19 (2016).
73. Hayer, S. S. et al. Genetic determinants of resistance to extended-spectrum cephalosporin and fluoroquinolone in *Escherichia coli* isolated from diseased pigs in the United States. *mSphere* **5**, 1–24 (2020).
74. Lopes, R., Furlan, J. P. R., Dos Santos, L. D. R., Gallo, I. F. L. & Stehling, E. G. Colistin-resistant mcr-1-positive *Escherichia coli* ST131-H22 carrying blaCTX-M-15 and qnrB19 in agricultural soil. *Front. Microbiol.* **12**, 659900 (2021).
75. Liu, C. M. et al. *Escherichia coli* ST131-H22 as a Foodborne Uropathogen. *MBio* **9**, 1–11 (2018).
76. Reid, C. J., McKinnon, J. & Djordjevic, S. P. Clonal ST131-H22 *Escherichia coli* strains from a healthy pig and a human urinary tract infection carry highly similar resistance and virulence plasmids. *Microb. Genom.* **5**, 1–12 (2019).
77. Flament-Simon, S.-C. et al. Whole genome sequencing and characteristics of mcr-1-harboring plasmids of porcine *Escherichia coli* isolates belonging to the high-risk clone O25b:H4-ST131 clade B. *Front. Microbiol.* **11**, 387 (2020).
78. Matsui, Y. et al. Multilocus sequence typing of *Escherichia coli* isolates from urinary tract infection patients and from fecal samples of healthy subjects in a college community. *Microbiologyopen* **9**, 1225–1233 (2020).
79. Thänert, R. et al. Comparative genomics of antibiotic-resistant uropathogens implicates three routes for recurrence of urinary tract infections. *MBio* **10**, 1–16 (2019).
80. Niehus, R., Mitri, S., Fletcher, A. G. & Foster, K. R. Migration and horizontal gene transfer divide microbial genomes into multiple niches. *Nat. Commun.* **6**, 8924 (2015).
81. Touchon, M. et al. Phylogenetic background and habitat drive the genetic diversification of *Escherichia coli*. *PLoS Genet.* **16**, e1008866 (2020).
82. Shaw, L. P. et al. Niche and local geography shape the pangenome of wastewater- and livestock-associated Enterobacteriaceae. *BioRxiv* <https://doi.org/10.1101/2020.07.23.215756> (2020).
83. Matlock, W. et al. Genomic network analysis of environmental and livestock F-type plasmid populations. *ISME J.* <https://doi.org/10.1038/s41396-021-00926-w> (2021).
84. Milkman, R., Jaeger, E. & McBride, R. D. Molecular evolution of the *Escherichia coli* chromosome. VI. Two regions of high effective recombination. *Genetics* **163**, 475–483 (2003).
85. Sokurenko, E. V. et al. Selection footprint in the FimH adhesin shows pathoadaptive niche differentiation in *Escherichia coli*. *Mol. Biol. Evol.* **21**, 1373–1383 (2004).
86. McNally, A. et al. Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. *PLoS Genet.* **12**, e1006280 (2016).
87. Page, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
88. Tonkin-Hill, G., Lees, J. A., Bentley, S. D., Frost, S. D. W. & Corander, J. Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Res.* **47**, 5539–5549 (2019).
89. Hunt, M. et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb. Genom.* **3**, e000131 (2017).
90. Zankari, E. et al. Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* **67**, 2640–2644 (2012).
91. Joensen, K. G. et al. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *J. Clin. Microbiol.* **52**, 1501–1510 (2014).
92. Carattoli, A. et al. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* **58**, 3895–3903 (2014).
93. Joensen, K. G., Tetzschner, A. M. M., Iguchi, A., Aarestrup, F. M. & Scheutz, F. Rapid and easy in silico serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. *J. Clin. Microbiol.* **53**, 2410–2426 (2015).
94. Reid, C. J., DeMaere, M. Z. & Djordjevic, S. P. Australian porcine clonal complex 10 (CC10) *Escherichia coli* belong to multiple sublineages of a highly diverse global CC10 phylogeny. *Microb. Genom.* **5**, 1–10 (2018).
95. Zankari, E. et al. PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. *J. Antimicrob. Chemother.* **72**, 2764–2768 (2017).

Acknowledgements

We would like to thank: Frank Hansen, Karin Sixhøj Pedersen and Hülya Kaya for their excellent technical assistance; Jonas Bonnedahl, Linköping University; Sara Byfors, Public Health Agency of Sweden; Catarina Flink and Mia Egervärn, Swedish Food Agency; Ivan Literak, Jaroslav Hrabak, Iva Kutilova and Ivana Jamborova for providing strains and help in the laboratory. Kay Anantanawat for genome sequencing support at iThree Institute. Ian Charles and Garry Myers for their helpful comments on the manuscript. Fiona MacIver for editing. This study was partially funded by Czech Science Foundation Grant no. 18-23532 S awarded to M.D. It was also supported by an Australian Government Medical Research Future Fund project (MRFF75873), the Australian Centre for Genomic Epidemiological Microbiology (AusGEM), a strategic research initiative between the New South Wales Department of Primary Industries and the University of Technology Sydney and Australian Research Council Linkage Project LP150100912.

Author contributions

C.J.R. contributed conceptualisation, data curation, formal analysis, investigation methodology, project administration, software, validation, visualisation, writing—original draft and writing review and editing; M.L.C. contributed data curation, investigation, methodology, software, validation, writing—review and editing; S.B. contributed data curation, investigation, resources, writing—review and editing; M.B. contributed data curation, investigation, resources, writing—review and editing; H.H. contributed data curation, investigation, resources, writing—review and editing; A.M.H. contributed data curation, investigation, resources; L.R. contributed data curation, investigation, resources; S.H. contributed data curation, investigation, resources, writing—review and editing; T.B. contributed data curation, investigation, resources, writing—review and editing; K.N. contributed data curation, investigation; M.H. contributed data curation, investigation, resources, writing—review and editing; J.-Y.M. contributed data curation, investigation, resources; A.B. contributed data curation, investigation; G.B.M. contributed data curation, investigation; A.-K.S. contributed data curation, investigation; S.S. contributed conceptualisation, data curation, project administration, resources, supervision, writing—review and editing; M.D. contributed conceptualisation, data curation, funding acquisition, project administration, resources, supervision, writing—review and editing; S.P.D. contributed conceptualisation, funding acquisition, project administration, resources, supervision, writing—review and editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-28342-4>.

Correspondence and requests for materials should be addressed to Cameron J. Reid or Steven P. Djordjevic.

Peer review information *Nature Communications* thanks Erick Denamur and the other, anonymous, reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022