Remarkable genomic diversity among *Escherichia* isolates recovered from healthy chickens

Nicholas M. Thomson[1], Rachel Gilroy[1], Maria Getino[2,3], Ebenezer Foster-Nyarko[1,4], Arnoud H.M. van Vliet[3], Roberto M. La Ragione[3,5] and Mark J. Pallen[1,3,6]

[1] Quadram Institute Bioscience, Norwich, Norfolk, United Kingdom
[2] NIHR Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, Department of Infectious Disease, Imperial College London, London, United Kingdom
[3] Department of Pathology and Infectious Diseases, School of Veterinary Medicine, University of Surrey, Guildford, Surrey, United Kingdom
[4] Department of Infection Biology, London School of Hygiene & Tropical Medicine, University of London, London, United Kingdom
[5] Department of Microbial Sciences, School of Biosciences and Medicine, University of Surrey, Guildford, Surrey, United Kingdom
[6] School of Biological Sciences, University of East Anglia, Norwich, Norfolk, United Kingdom

ABSTRACT

The genus *Escherichia* has been extensively studied and it is known to encompass a range of commensal and pathogenic bacteria that primarily inhabit the gastrointestinal tracts of warm-blooded vertebrates. However, the presence of *E. coli* as a model organism and potential pathogen has diverted attention away from commensal strains and other species in the genus. To investigate the diversity of *Escherichia* in healthy chickens, we collected fecal samples from antibiotic-free Lohmann Brown layer hens and determined the genome sequences of 100 isolates, 81 of which were indistinguishable at the HC0 level of the Hierarchical Clustering of Core Genome Multi-Locus Sequence Typing scheme. Despite initial selection on CHROMagar Orientation medium, which is considered selective for *E. coli*, *in silico* phylotyping and core genome single nucleotide polymorphism analysis revealed the presence of at least one representative of all major clades of *Escherichia*, except for *E. albertii, Shigella*, and *E. coli* phylogroup B2 and cryptic clade I. The most frequent phylogenomic groups were *E. coli* phylogroups A and B1 and *E. ruysiae* (clades III and IV). We compiled a collection of reference strains isolated from avian sources (predominantly chicken), representing every *Escherichia* phylogroup and species, and used it to confirm the phylogeny and diversity of our isolates. Overall, the isolates carried low numbers of the virulence and antibiotic resistance genes typically seen in avian pathogenic *E. coli*. Notably, the clades not recovered are ones that have been most strongly associated with virulence by other studies.

Subjects Bioinformatics, Genomics, Microbiology, Molecular Biology, Veterinary Medicine
Keywords Escherichia, Genomic diversity, Chickens, Phylogenomics, Commensal, Cryptic clades

Submitted 12 November 2021
Accepted 23 January 2022
Published 1 March 2022

Corresponding author
Mark J. Pallen,
mark.pallen@quadram.ac.uk

Academic editor
Jonathan Thomas

Additional Information and Declarations can be found on page 17

DOI **10.7717/peerj.12935**

How to cite this article Thomson NM, Gilroy R, Getino M, Foster-Nyarko E, van Vliet AH, La Ragione RM, Pallen MJ. 2022. Remarkable genomic diversity among *Escherichia* isolates recovered from healthy chickens. **PeerJ 10:e12935 DOI 10.7717/peerj.12935**

INTRODUCTION

Members of the genus *Escherichia* are common inhabitants of the gastrointestinal tracts of warm-blooded animals (*Tenaillon et al., 2010*; *Jang et al., 2017*). Although many isolates are believed to be beneficial or harmless components of a healthy microbiome, the genus has received a large share of research attention as some members—notably certain strains of *Escherichia coli*—cause disease and/or carry transferable antibiotic resistance genes (*Kaper, Nataro & Mobley, 2004*; *Cummins et al., 2019*).

Much effort has been expended to develop techniques for identifying and categorizing members of *Escherichia* and understanding the population structure of the genus. Classical phenotypic techniques identified species by morphology and biochemical tests and divided strains into pathotypes, based on site of infection, *e.g*., uropathogenic *E. coli* (UPEC) causing urinary tract infections; and serotypes, based on antibody recognition of variation in lipopolysaccharides, flagella and fimbriae (*Fratamico et al., 2016*). These methods have gradually given way to genome-based approaches classifying isolates with multi-locus enzyme electrophoresis studies (*Selander et al., 1987*) and multi-locus sequence typing (*Maiden et al., 1998*). Whole genome sequencing technologies have recently facilitated the comparison of strains in ever greater detail (*Uelze et al., 2020*), leading to new classification schemes based on average nucleotide identity (*Konstantinidis & Tiedje, 2005*) and whole-genome nucleotide polymorphism (*Schürch et al., 2018*).

The avian pathogenic *E. coli* (APEC) pathotype causes colibacillosis in chickens, turkeys and other avian species, and is responsible for significant morbidity and mortality in the worldwide poultry industry (*Dziva & Stevens, 2008*). Certain serotypes and virulence factors are frequently associated with APEC but it remains challenging to distinguish categorically between commensal and virulent strains (*Mehat, van Vliet & La Ragione, 2021*).

The large number of high-quality *Escherichia* genomes from diverse geographical and biological sources has forced a re-thinking of classification. Thus, although *Shimwellia blattae* was initially placed within *Escherichia*, it is now assigned to a separate genus (*Priest & Barker, 2010*). By contrast, although named as if a genus, *Shigella* is now thought merely to represent a series of pathovars of *E. coli* (*Yang et al., 2005*; *Devanga Ragupathi et al., 2018*). Nevertheless, the assignment of strains of *E. coli* to seven phylogroups (A, B1, B2, C, D, E and F) has proven a robust finding that holds for phylogenetic trees built from core genome alignments of large numbers of isolates. Phylogroups G and H have recently been proposed (*Lu et al., 2016*; *Clermont et al., 2019*), while five of what were initially called 'cryptic clades' (*Walk et al., 2009*) have been assigned to species: Clade I to *E. coli* (*Clermont et al., 2013*), Clade II to *E. whittamii* (*Gilroy et al., 2021*), Clades III and IV together to *E. ruysiae* (*van der Putten et al., 2021*), and Clade V to *E. marmotae* (*Liu et al., 2015*).

The accumulation of genomic data for *Escherichia* has been strongly biased towards clinical isolates from humans and economically important animals (*Touchon et al., 2020*). This hampers efforts to achieve a complete understanding of *Escherichia* diversity, ecology, and population genetics. Here, we report phylogenomic analyses on 100 *Escherichia* isolates cultured from fecal samples from a small flock of healthy adult layer chickens, 81 of which we determined to be unique (Fig. 1).

MATERIALS AND METHODS

Sample collection and storage

The project was approved by the University of Surrey's NASPA Ethical Review Assessment Committee with project number NERA-2018-011. Fresh fecal samples from healthy adult Lohmann Brown layer hens on a farm in the south-east of England were collected to isolate presumptive *E. coli*. The birds (flock size = 24) were purchased from a single flock in April 2017 at 20 weeks old and kept in a large outdoor run with a substrate of stone chippings and small turf enrichment beds during the day and in a coop overnight. They were fed Farmgate Layer pellets and Mash (ForFarmers Ltd, Bury St Edmunds, UK), according to the manufacturer's instructions, and no antibiotics were used. Sampling was carried out between July and November 2018. Once a month, five freshly voided fecal samples were collected from different hens. 1:10 serial dilutions of each sample were plated (100 μL) onto MacConkey agar to obtain single colonies. Several colonies with different morphologies were selected and streaked to fresh MacConkey agar plates to isolate single colonies. Identification as presumptive *E. coli* was confirmed by a purple colony phenotype on CHROMagar Orientation Medium (CHROMagar, Paris, France) and a positive indole test. Multiple *E. coli*-positive isolates from each sample were tested by GTG_5 PCR (*Mohapatra, Broersma & Mazumder, 2007*) to minimize the collection of identical genotypes. Isolates were stored using the Microbank Bacterial and Fungal Preservation system (Pro-Lab Diagnostics, Richmond Hill, Canada) at −80 °C.

Antimicrobial resistance screening

Isolates were screened for resistance against a panel of eight antibiotics (gentamicin, ampicillin, trimethoprim, chloramphenicol, nitrofurantoin, cefpodoxime, meropenem and ciprofloxacin) by the disk diffusion method, according to EUCAST guidelines (https://www.eucast.org/ast_of_bacteria). They were additionally subjected to a minimum inhibitory concentration assay for colistin by broth dilution, according to EUCAST guidelines.

DNA extraction and genome sequencing

A total of 100 isolates were submitted for whole genome sequencing, comprised of multiple isolates per fecal sample from the first 3 months and one isolate per sample for the remaining 2 months (Table S1). Genomic DNA was extracted from overnight 1 mL cultures in Lysogeny Broth (LB; Miller's formulation), maintained at 37 °C statically in sterile deep-well plates prepared from single colonies taken from overnight Lysogeny Agar plates, according to a previously described 96-well plate lysate method (*Foster-Nyarko et al., 2020*). Briefly, cultures were pelleted at 3,500 × g then resuspended and pre-treated with lysozyme, proteinase K and RNase A before lysing with 10% (w/v) sodium dodecyl sulphate in Tris-EDTA (pH 8.0). The DNA was retained on AMPure XP solid-phase reversible immobilization beads (Beckman Coulter, High Wycombe, UK), eluted in

Tris/HCl, pH 8.0 and quantified using the Qubit high-sensitivity double-stranded DNA assay (Invitrogen, MA, USA).

For sequencing, samples were normalized to 0.5 ng.µL^{-1} with 10 mM Tris-HCl prior to library preparation, which was carried out with a modified Nextera XT DNA protocol (*Foster-Nyarko et al., 2020*). The pooled libraries were run at a final concentration of 1.8 pM using the Illumina NextSeq 500 platform (Illumina, CA, USA) following the manufacturer's recommended denaturation and loading procedures, which included a 1% PhiX spike, with 150 bp paired-end reads.

Genome assembly

Raw sequence data were demultiplexed and converted to fastq format by bcl2fastq v.2.20 (Illumina, CA, USA). Reads from multiple lanes were concatenated into single forward and reverse read files then uploaded to Enterobase (www.enterobase.warwick.ac.uk) for automated quality control and assembly *via* the QAssembly pipeline (*Zhou et al., 2020*). All assemblies exceeded the minimum quality requirements for Enterobase (genome length between 3.7–6.4 Mbp; $N_{50} > 20$ kb; number of contigs ≤ 800; proportion of Ns < 3%; >70% of contigs assigned correctly by Kraken 2 (*Wood, Lu & Langmead, 2019*)). The final assemblies were downloaded for further analyses, which were all run on the Cloud Infrastructure for Microbial Bioinformatics (*Connor et al., 2016*).

Phylogenomic analysis

Initial phylogenomic analysis was carried out in Enterobase for *in silico* serotyping, *fimH* allele typing (*Roer et al., 2017*), multi-locus sequence typing according to the Warwick seven-gene scheme and phylotyping (*Beghain et al., 2018*), and to construct neighbor-joining trees with the NINJA algorithm in GrapeTree (*Zhou et al., 2018*), based on the Hierarchical Clustering of Core Genome MLST (HierCC) scheme (*Zhou et al., 2020*). HierCC was also used to de-replicate the strain collection by selecting a single representative isolate from every identical cluster of genomes at the HC0 level. Representative isolates were determined according to the quality score-based system employed as part of dRep v2.5.4, reliant on genome completion, contamination and N50 metrics (*Olm et al., 2017*). The assemblies for the 33 strains involved in clustering were used to construct a core genome alignment using Snippy v.4.3.2 (https://github.com/tseemann/snippy). A matrix of pairwise single nucleotide polymorphisms (SNPs) was then compiled with snp-dists v.0.7.0 (https://github.com/tseemann/snp-dists). *E. coli* MG1655 (NC_000913.3) was used as the reference.

To confirm the identities and phylogenetic relationships of the final isolate collection, a second core genome alignment was reconstructed, containing all strains. In addition, we collated a panel of reference genomes (Table S2) for every phylogroup of *E. coli* and all other *Escherichia* species, including *Shigella flexneri*, all of which had been isolated from avian hosts (primarily chickens). These reference genomes were downloaded from the NCBI assembly archive and included in the core genome alignment, which was then used to reconstruct a core SNP phylogenetic tree with *Salmonella bongori* and *Salmonella enterica* as outgroups. IQ-TREE v.2.0.3 (*Nguyen et al., 2015*) was used with 1,000 bootstrap replications for maximum-likelihood inference of phylogenetic relationships with the best fitting model (TVM+F+ASC+R4) automatically selected by ModelFinder (*Kalyaanamoorthy et al., 2017*). The resulting tree was visualized and combined with other data using iTOL v.5 (*Letunic & Bork, 2021*).

Analysis of gene and plasmid content

To process shared gene content across our selected genome catalogue, we used the pangenomics pipeline (*Delmont & Eren, 2018*) as implemented in anvi'o v.7.0 (*Eren et al., 2015*) with open reading frames predicted using prodigal v.2.6 (*Hyatt et al., 2010*) and annotation using the NCBI's Clusters of Orthologous Groups database (*Tatusov et al., 2003*). Using NCBI BLAST, the similarity between gene pairs was quantified and subsequently the Markov Cluster algorithm determined clusters of homologous genes, with a *minbit heuristics* threshold of 0.5 to eliminate weak matches and an MCL inflation of 10 for closely related genomes. FastANI (*Jain et al., 2018*) was applied for the calculation of average nucleotide identity between genomes. The program anvi-display-pan provided an interactive visualization of the pangenome, with imported average nucleotide identity values being depicted as part of this visualization.

ABRicate v.1.0.1 (https://github.com/tseemann/abricate) was used to search assemblies for genes related to antibiotic resistance and virulence and for plasmid replicons by comparison to the NCBI AMRFinderPlus (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA313047), ecoli_VF (https://github.com/phac-nml/ecoli_vf) and PlasmidFinder (*Carattoli et al., 2014*) databases, respectively. A custom database was prepared for detection of 24 APEC-related virulence genes based on a literature search (File S1). In each case, identification was defined by minimum coverage of 90% and minimum identity of 80% of the respective nucleotide sequences. To detect plasmid-encoded virulence-associated genes we first identified contigs derived from plasmids using platon v.1.6 in accuracy mode (*Schwengers et al., 2020*), then searched those contigs against our custom database using ABRicate as described above.

RESULTS

High phylogenomic diversity of isolates recovered

One hundred isolates provisionally identified as *E. coli* were collected from healthy layer hens over a 5-month period. An indication of the diversity of *Escherichia* recovered was provided by the isolation of 1–7 isolates from up to six phylogroups per fecal sample (Table 1). When cultured on MacConkey agar, 15 isolates were non-lactose-fermenters but positive for other *E. coli*-specific attributes (*Nicoletti et al., 1988*).

Following short read whole genome shotgun sequencing and assembly, the HierCC feature within Enterobase identified 33 out of the 100 isolates as belonging to one of 14 clusters at the HC0 level, meaning that the members of each cluster are indistinguishable from each other at every core genome locus interrogated (Table 2). The 14 HC0 clusters all contained isolates taken from different fecal samples, spanning the duration of the study, except cluster HC0:148574, which contained three isolates that all came from a single sample taken in September 2018. A core genome alignment of the 33 strains

Table 1 Collection dates, age of birds and number of isolates recovered for each fecal sample.

Faecal sample	Date collected	Age of birds (weeks)	GTG_5 unique isolates	HC0 unique isolates	Phylogroups
1	09/07/2018	84	5	5	A, B1
2	09/07/2018	84	6	5	A, B1, (II)
3	09/07/2018	84	6	6	A, B1, C
4	09/07/2018	84	6	3	A, Ef
5	09/07/2018	84	6	4	A, D, (V)
6	13/08/2018	89	7	7	A, B1, F, Ef, (III), (V)
7	13/08/2018	89	5	4	A, B1
8	13/08/2018	89	6	5	A, Ef, (III), (V)
9	13/08/2018	89	6	6	A, B1, (V)
10	13/08/2018	89	5	3	B1, (III), (IV)
11	10/09/2018	93	6	6	A, B1, E, Ef, (III)
12	10/09/2018	93	7	5	A, B1, E, Ef
13	10/09/2018	93	7	5	A, B1, (III)
14	10/09/2018	93	7	4	A, (IV), (V)
15	10/09/2018	93	6	4	A, (III), (V)
16	08/10/2018	97	1	1	B1
17	08/10/2018	97	1	1	(IV)
18	08/10/2018	97	1	1	(III)
19	08/10/2018	97	1	1	A
20	08/10/2018	97	1	1	D
21	12/11/2018	102	1	1	(V)
22	12/11/2018	102	1	1	D
23	12/11/2018	102	1	1	B1
24	12/11/2018	102	1	1	A
			100	81	

Note:

Following culturing on MacConkey agar and CHROMagar Orientation medium, 100 presumptive *E. coli* were isolated. Multiple colonies per fecal sample were screened by GTG_5 PCR to reduce the collection of identical strains. Whole-genome sequence analysis revealed the large diversity of *Escherichia* present despite apparently selecting for *E. coli*, although only 81 of the initial strains were distinguishable at the HC0 level using HierCC clustering (see Materials and Methods). Phylogroups of *E. coli* are referred to by individual letters; cryptic clades are referred to by Roman numerals in parentheses; Ef, *E. fergusonii*.

involved in clustering was used to construct a matrix of pairwise SNP distances, confirming the close relationships between clustered strains, with 2–257 SNPs detected (Table 2, Table 3). Since we were primarily interested in assessing species diversity within our samples, we therefore removed 19 isolates from the analysis by keeping only the best quality assembly from each HC0 cluster as a representative (Table S3).

In silico evaluation of Warwick seven-gene MLSTs for the remaining 81 isolates revealed the presence of 45 different sequence types, of which 30 were represented by a single isolate each. ST48 (nine isolates), ST10 (seven isolates) and ST2456 (five isolates) were the most numerous. HC0 isolate clusters correlated with MLSTs, supporting the close relationship between strains in each cluster (Table S1). Predicted phylogroups also suggested considerable genetic diversity among the isolates, with at least one representative from all seven *E. coli* phylogroups except B2, every 'cryptic clade' except clade I, and five

Table 2 **Clustering of isolates by HierCC level HC0 and selection of representative isolate for each cluster.**

HC0 cluster	MLST	Isolates	Core SNP distance	Representative isolate
148525	5643	021, 043	124	021
148529	155	007, 037	21	037
148537	48	024, 032	11	032
148541	1112	026, 067	6	067
148542	1638	006, 055	11	006
148547	5573	025, 041, 057, 073, 087	See Table 3	025
148548	1276	034, 089	18	034
148555	48	018, 042	20	042
148556	7059	045, 085	55	045
148564	1844	019, 050, 078	See Table 3	050
148571	6540	063, 080	257	063
148574	2456	068, 069, 071	See Table 3	068
148603	48	004, 081	2	004
148607	11513	020, 058	143	058

Table 3 **SNP distance matrices for HC0 clusters of >2 isolates.**

SNP distance matrix for Cluster HC0_148547

Strains	041	057	073	087
025	9	9	9	4
041	0	5	11	11
057		0	13	8
073			0	10

SNP distance matrix for Cluster HC0_148564

	050	078
019	16	15
050	0	14

SNP distance matrix for Cluster HC0_148574

	069	071
068	8	14
069	0	10

Note:
Table 2 provides SNP distances for 11 HC0 clusters comprising pairs of isolates. The remaining three clusters contained >2 isolates, requiring a matrix for full pairwise comparison.

E. fergusonii (Table 1, Fig. S1). *In silico* prediction of O:H serotype and *fimH* alleles also pointed towards considerable diversity, with 30 different O antigens, 22 different H antigens and 24 different *fimH* alleles detected (Table S1). Isolates outside of *E. coli sensu stricto* had increased incidence of ‘undetermined’ results for O antigen and *fimH*, suggesting that current databases do not cover the whole diversity of these features for the genus. The diversity of strains did not have a significant temporal component as isolates

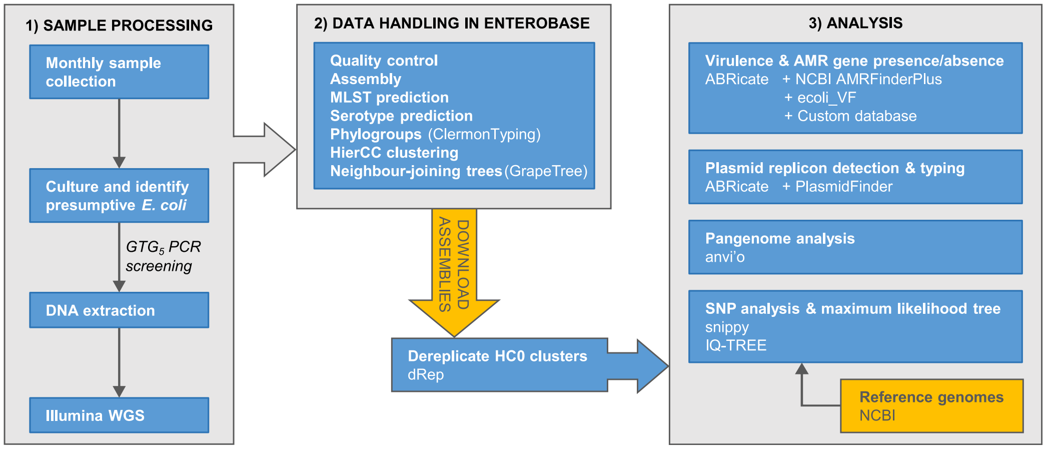

Figure 1 **Summary of the major stages of sample collection, processing and data analysis and the primary programs and databases used.** Full-size DOI: 10.7717/peerj.12935/fig-1

collected in different months were evenly distributed throughout a neighbor-joining phylogenetic tree based on the HierCC results (Fig. S1).

A core SNP maximum likelihood tree containing the 81 isolates and reference genomes for every clade of *Escherichia* was in good agreement with the neighbor-joining tree, and the predicted phylogroups clustered with their respective reference genomes (Fig. 2). This validates the reference genome collection, indicating its value for providing a scaffold phylogeny for future studies. Three isolates that were identified as 'group E or clade I' by *in silico* phylotyping were assigned to phylogroup D based on the core SNP alignment. Isolate Surrey070 was also reassigned, as it clustered more closely with phylogroup E than its predicted phylogroup of D.

The pangenome of isolates reflects phylogenomic diversity

More than 500,000 genes were identified within the 81 study isolates and 33 reference genomes before organization into 14,421 gene clusters, formed by grouping homologous genes according to amino acid similarity, with a *minbit heuristics* threshold of 0.5 and an MCL inflation of 10. The core genome (genes present in all genomes) contained 2,449 gene clusters and the accessory genome (genes present in >1 but not all genomes) contained 8,331 gene clusters, while 3,641 singleton clusters (present in only one genome) were identified. All accessory and core genes were hierarchically clustered according to distribution pattern (Fig. 3). Functional hits were assigned to gene clusters according to NCBIs Clusters of Orthologous Groups database, with 2,766 unique groups annotated (Table S4). A total of 94% of core gene clusters were annotated compared to the considerably lower proportion annotated for both accessory (50%) and singleton (30%) gene clusters. Functional pathways assigned to annotated gene clusters were primarily associated with translation, ribosomal structure, and biogenesis (9.5% of annotated core gene clusters) with accessory gene clusters associated with prophages and transposons

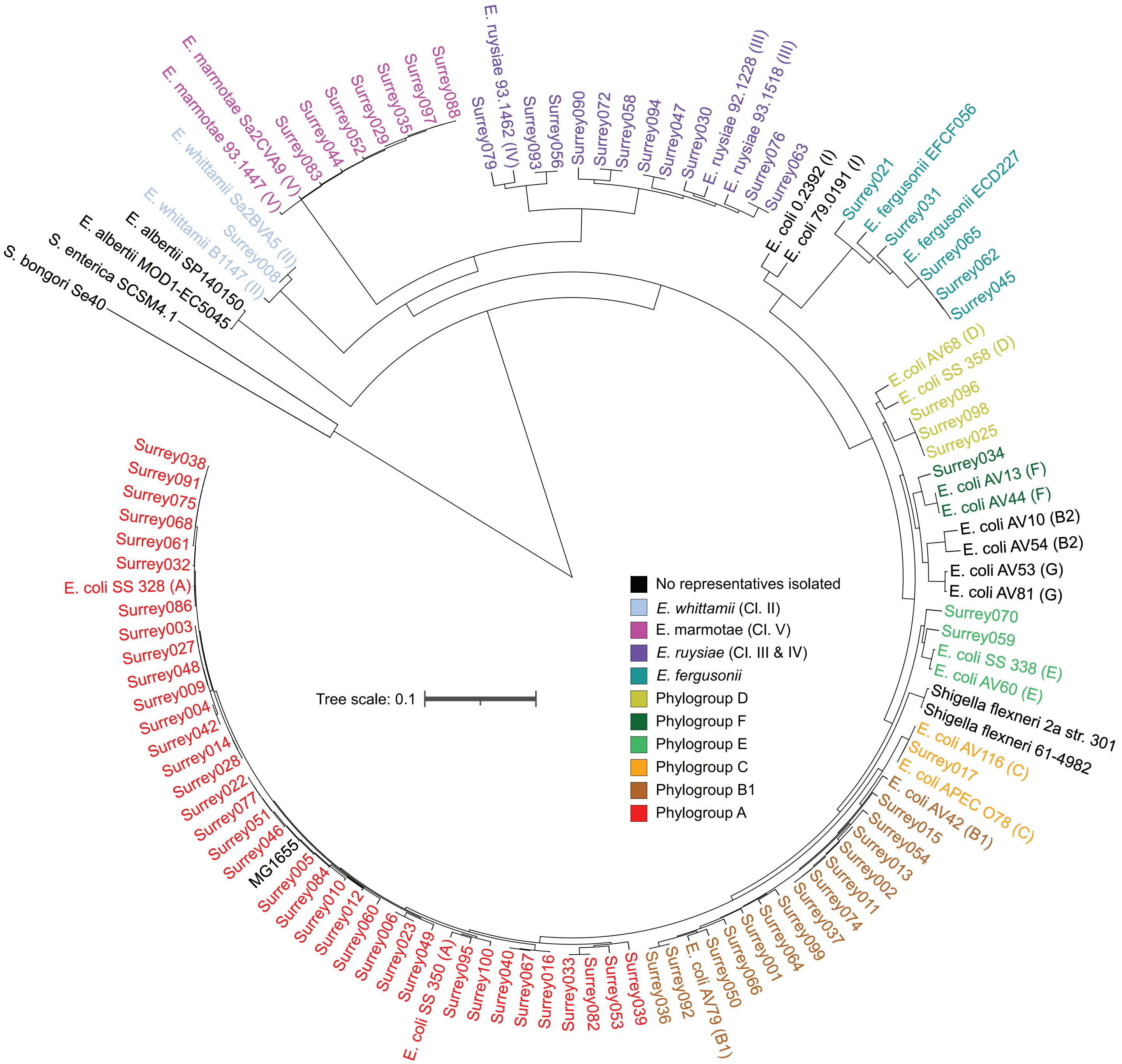

Figure 2 Core SNP maximum likelihood tree of the final 81 *Escherichia* isolates, including reference sequences for every *Escherichia* species and all phylogroups of *E. coli*. Names of our isolates and their reference strains are coloured according to their species/phylogroup. Names of reference strains for clades not found in our samples are coloured black. Reference strains are named with their species and strain name, with phylogroups given in parentheses. *Salmonella bongori* and *Salmonella. enterica ser. Typhi* are included as outgroups.

Full-size DOI: 10.7717/peerj.12935/fig-2

(9.8% of annotated accessory gene clusters). Both core and accessory splits had prominent proportions of gene clusters annotated with carbohydrate transport and metabolism functions (7.2% and 10.1% of annotated gene clusters, respectively).

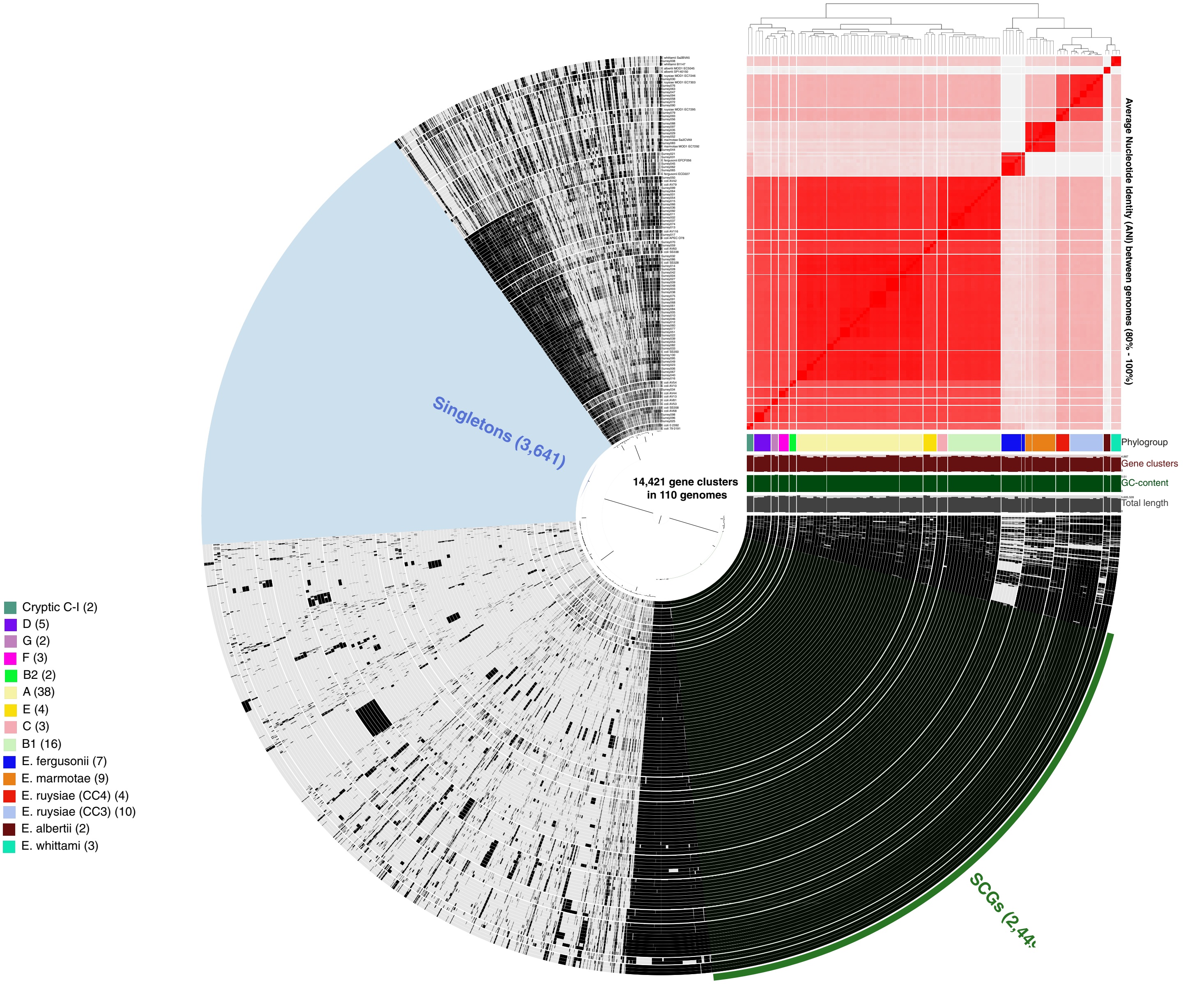

Figure 3 Anvi'o representation of the *Escherichia* pangenome showing 81 *Escherichia* isolates recovered from chicken faeces and 33 publicly available reference *Escherichia* isolates from poultry species. Each layer represents a single genome, with black colouring signifying the presence of a gene cluster. Gene clusters are organised according to their distribution across the genome, with co-occurring genes shown closer together. The heatmap at the top-right of the image represents average nucleotide identity across all included genomes, with darker red colours indicating a higher percentage of average nucleotide identity. Assigned phylogroup is shown by the colour bar on the right of the image. Singleton gene clusters (present only in one genome; *n* = 3,641) are highlighted in blue while core gene clusters (present in all genomes; *n* = 2,449) are shown in green.

Full-size DOI: 10.7717/peerj.12935/fig-3

Plasmid incompatibility groups are shared throughout *Escherichia*

A total of 25 different plasmid replicons were identified by comparison with the PlasmidFinder database (Table S5). Eight strains did not contain a plasmid replicon, while two strains (Surrey037 and Surre074) contained seven different replicons (Fig. 4).

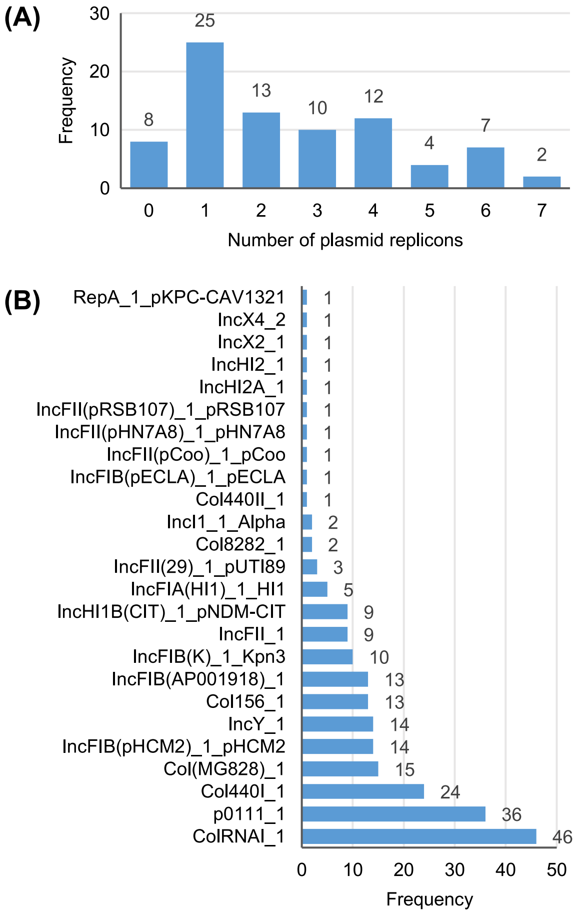

Figure 4 Summary of the plasmid replicon contents of 81 *Escherichia* isolates. (A) Number of isolates containing 0–7 distinct plasmid replicons from the PlasmidFinder database. (B) Frequency of each plasmid replicon identified from the PlasmidFinder database. Some isolates contained >1 of the same replicon type (see main text). Full-size DOI: 10.7717/peerj.12935/fig-4

The overall rate of plasmid carriage was low, with a mode of 1 replicon and median of 2. There were 14 instances of an isolate containing multiple replicons of the same identity, although these were confined to the Col(MG828)_1, Col440I_1 and ColRNAI_1 groups. The most common replicons were ColRNAI_1 (46 replicons in 35 isolates) and p0111_1 (36 replicons in 36 isolates). Of the broad plasmid incompatibility groups, Col plasmids were most frequently found (101 replicons), partly because of their propensity for multiple replicons per isolate, with IncF (58 replicons) and p0111_1 (36 replicons) also highly represented (Fig. 5). There were no clear associations between plasmid incompatibility groups and phylogenomic groups, which is consistent with frequent transfer of genetic material within the genus (*Shaw et al., 2021*).

Overall low virulence potential amongst isolates

To assess the potential for our isolates to cause disease in chickens we looked for genomes containing ≥4 of the virulence-associated genes *iutA, hlyF, iss, iroN* and *ompT*. This panel has been proposed to identify APEC more accurately than classical serotyping methods

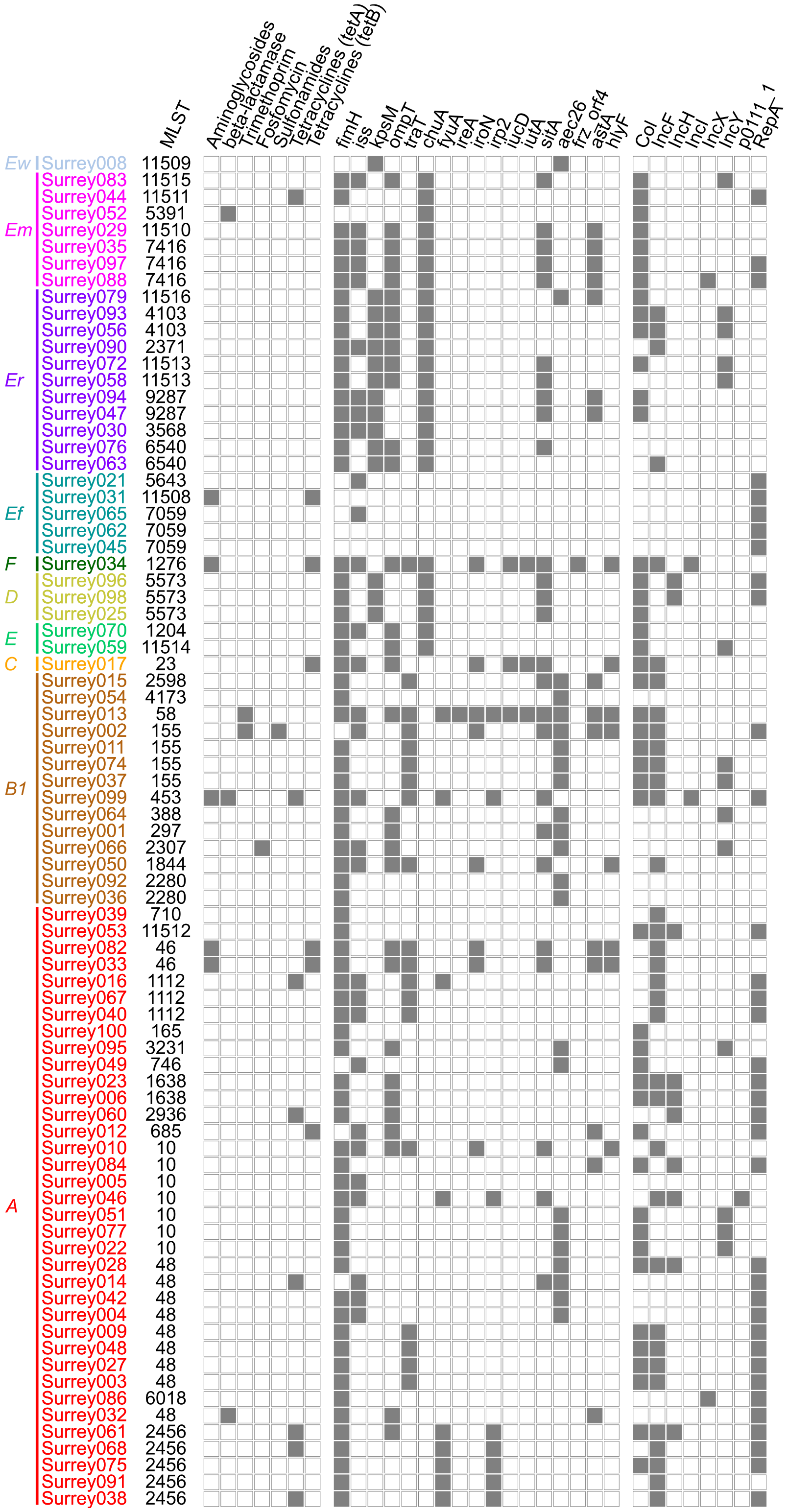

Figure 5 Presence (solid squares) and absence (open squares) of predicted antibiotic resistance genes, virulence-associated genes and plasmid replicons in the final 81 isolates. Isolates are arranged according to the phylogenetic tree shown in Fig. 2 with their names coloured according to phylogenetic group. Ew, *E. whittamii*; Em, *E. marmotae*; Er, *E. ruysiae*; Ef, E. *fergusonii; A–F, E. coli*

Figure 5 (continued)
phylogroups. Virulence-associated genes shown here are from the custom panel of APEC-associated genes. Members of the panel not represented here were not present in any isolates. Plasmid replicon and predicted antibiotic resistance gene data have been condensed to show major classes. Therefore, presence indicates ≥1 members of that class were detected. Full gene and plasmid detection data are available in Tables S5–S7. Full-size DOI: 10.7717/peerj.12935/fig-5

(*Johnson et al., 2008*). Within our isolates, Surrey013, Surrey017 and Surrey034 carried all five genes, while Surrey010 and Surrey050 carried all except *iutA*. However, genes are more strongly correlated with pathogenicity if they are carried on a plasmid rather than the chromosome (*Johnson et al., 2008*). Both Surrey013 and Surrey034 had plasmid-encoded copies of four of the genes, qualifying them as APEC. Surrey017 and Surrey050 carried four plasmid-encoded virulence-associated genes each, but not all were from the panel of five genes so are not counted as APEC under these criteria (Table S6). A further 16 strains carried one or two virulence-associated genes on plasmids.

An alternative diagnostic strategy defines APEC by categorization into any of four associations of virulence (*Schouler et al., 2012*). By these criteria 5/81 strains are categorized as APEC: Surrey034 (*iutA*+, P(F11)−, frz_{orf4}+), Surrey001, Surrey002, Surrey014 and Surrey015 (*iutA*−, *sitA*+, *aec26*+). Surrey013 and Surrey017 might additionally be considered marginal as they were only discounted due to their predicted O8 serotype, rather than O78. Therefore, the two diagnostic measures combined indicate that 6/81 strains are APEC and 3/81 have increased virulence potential but fall short of the cut-off. Overall, only a minority of our isolates are likely to be APEC according to their complement of genetic virulence determinants.

APEC fall into the wider pathotype of Extraintestinal Pathogenic *E. coli* (ExPEC), which includes named groups causing disease in humans: Neonatal Meningitis *E. coli* (NMEC), Sepsis-Associated *E. coli* (SEPEC) and Uropathogenic *E. coli* (UPEC) (*Sarowska et al., 2019*). APEC may be closely related to human ExPEC and may act as a reservoir for ExPEC virulence-associated genes (*Cummins et al., 2019*). Therefore, to assess the wider pathogenicity potential of our strains, we compiled a panel of 24 virulence-associated genes. These included the APEC-associated genes above and other typical ExPEC virulence-associated genes (Table 4). Sixteen of these genes were present in at least one isolate (Fig. 5). Three isolates (Surrey031, Surrey045 and Surrey062) contained none of the genes and Surrey013 contained the most with 14/24 detected. The modal isolate contained two genes from the panel, and the median contained three genes.

Other commonly identified virulence-associated genes from comparison with the ecoli_VF database (a much larger database covering *E. coli* in general) included *fimA–I, yagV–Z, ompA, entA–F/S, fepA–D/G* and *fes.* Genes associated with the general secretory pathway (*gspC–M*) (*Francetic & Pugsley, 1996*) and type III secretion systems 1 and 2 (*espL, R, X, Y; eivA, C, E–G, I, J*) were also frequently identified (*Franzin & Sircili, 2015*; *Fox et al., 2020*). Many of these genes are associated with intestinal *E. coli* infections. Full predicted virulence-associated gene content of all strains, from comparison with the ecoli_VF database are available in Table S7.

Table 4 **Members of the panel of virulence-associated genes used to predict virulence potential of isolates in this study.**

Gene	Function	Protein	N° isolates	Ref.
afa/dra	Adhesion	Afimbrial adhesin	0	(*Johnson et al., 2003*)
fimH		Type 1 D-mannose specific adhesin	70	(*Cummins et al., 2019*)
papA/felA		P fimbrial adhesin (type F11)	0	(*Kariyawasam & Nolan, 2011*; *Schouler et al., 2012*)
papC		Outer membrane usher	0	(*Dodson et al., 1993*; *Schouler et al., 2012*)
papG		P fimbrial tip adhesin	0	(*Johnson & Brown, 1996*; *Schouler et al., 2012*)
sfa/foc		S fimbrial adhesin	0	(*Johnson et al., 2003*)
iss	Host evasion	Increased serum survival lipoprotein	30	(*Johnson et al., 2008*)
kpsM		Polysialic acid transport protein	15	(*Stromberg et al., 2017*)
ompT		Outer membrane protease	32	(*Johnson et al., 2008*)
traT		Complement resistance protein	19	(*Cummins et al., 2019*)
chuA	Iron acquisition	TonB-dependent heme receptor	24	(*Spurbeck et al., 2012*)
fyuA		Siderophore yersiniabactin receptor	9	(*Stromberg et al., 2017*)
ireA		Iron-regulated outer membrane virulence protein	1	(*Cummins et al., 2019*)
iroN		Siderophore salmochelin receptor	8	(*Johnson et al., 2008*)
irp2		Yersiniabactin biosynthetic protein	8	(*Cummins et al., 2019*)
iucD		L-lysine N6-monooxygenase (aerobactin synthesis)	3	(*Cummins et al., 2019*)
iutA		Siderophore aerobactin receptor	3	(*Johnson et al., 2008*; *Schouler et al., 2012*)
sitA		Fe/Mn ABC transporter substrate binding	26	(*Schouler et al., 2012*)
icmH/ aec26	Secretion	type IVB secretion system protein	23	(*De Pace et al., 2010*; *Schouler et al., 2012*)
frz_{orf4}	Sugar transport	PTS fructose transporter subunit IIC	1	(*Rouquet et al., 2009*; *Schouler et al., 2012*)
astA	Toxin	Heat-stable enterotoxin EAST1	15	(*Kaper, Nataro & Mobley, 2004*)
hlyF		SDR Family oxidoreductase	8	(*Johnson et al., 2008*)
tsh		Temperature-sensitive haemagglutinin	0	(*Cummins et al., 2019*)
vat		Vacuolated autotransporter toxin	0	(*Spurbeck et al., 2012*)

Low carriage of predicted antimicrobial resistance genes

From our comparison with the NCBI AMRFinderPlus database, we predicted low levels of antimicrobial resistance. Among the isolates, 62/81 had no predicted resistance genes (Table S8). However, five isolates carried three predicted resistance genes, one carried two genes, and 13 carried one gene. Fourteen isolates carried a gene for predicted tetracycline resistance, nine carried a predicted aminoglycoside resistance gene, three carried a predicted β-lactamase class *bla-TEM* gene, two carried a predicted trimethoprim resistance gene, and predicted genes for fosfomycin and sulfonamide resistance were found in a single isolate each (Fig. 5). Every isolate except three carried a single copy of a β-lactamase class *blaEC* gene (one of *blaEC*, $blaEC_{8}$, $blaEC_{13,}$ $blaEC_{15}$, $blaEC_{18}$ or $blaEC_{19}$). These genes are frequently detected in genomes of β-lactam-susceptible isolates and therefore were discounted from the analysis. We did not experimentally verify the predicted resistance phenotypes; however, a preliminary screening following colony isolation also revealed low levels of phenotypic resistance to nine antibiotics (Table S1).

DISCUSSION

Our data reveal remarkable diversity within the genus *Escherichia* from such a commonplace setting and small sample size. After minimal selection during sample collection, we were able to recapitulate almost the entire phylogeny of *Escherichia*. However, our initial selection for *E. coli* identity may still have excluded some isolates with phenotypes that differ from typical *E. coli*. On the other hand, the isolation of so many other *Escherichia* species raises doubts over the specificity of CHROMagar Orientation Medium for *Escherichia coli sensu stricto*. This finding may have implications for the use of CHROMagar Orientation Medium in clinical research and diagnostics, and warrants further investigation into the growth and morphology of different members of the *Escherichia* genus on this medium. We confirmed the diversity and identity of the isolates by several different methods, including PCR, *in silico* phylotyping and MLST, and core genome analysis. Our final core genome-derived phylogenetic tree was also in agreement with previously published phylogenies for the genus (*Zhou et al., 2020*; *Abram et al., 2021*; *Gilroy et al., 2021*).

The generally low carriage of virulence-associated genes within the isolates suggests that the majority are non-pathogenic and more likely represent commensal members of the gut microbiota (*Wigley, 2015*). By extension, this also supports the idea that although many strains are capable of opportunistic infection, most frank pathogens in the *Escherichia* genus belong to relatively well-defined pathotypes (*Johnson et al., 2008*). In support of this conclusion, three key phylogroups, B2, G and clade I, which have been associated with virulence in poultry (*Papouskova et al., 2020*; *Mehat, van Vliet & La Ragione, 2021*), were missing from our collection. This could be due to the relatively small sample size and small flock or might reflect that we only sampled healthy birds. Equally, the most common phylogroup in this study was phylogroup A, which has been linked to commensalism in chickens and other omnivores (*Smati et al., 2015*). *E. albertii* and *Shigella* were also absent from our collection. They are less well studied in birds, but have also both been linked to virulence in poultry (*Shi et al., 2014*; *Gomes et al., 2020*). We cannot rule out the possibility that APEC strains were simply lower in abundance and so were overlooked during culturing. As we only sampled healthy birds, it seems likely that this is a true reflection of the prevalence of these phylogroups among our study population. The limitations of our study prevent us from inferring any general conclusions about the distribution of *Escherichia* strains within chickens in other settings. Nevertheless, if such striking diversity is present in a few birds in a single location, it surely suggests that similar diversity (although perhaps with different population structures) has so far been overlooked elsewhere.

Although we did not track individual birds temporally and so cannot verify stable colonization of isolates within birds, most fecal samples yielded multiple distinct isolates. There were no clear indications of niche adaptation based on the pangenome and plasmid content of the strains, as gene clusters and plasmid replicons were spread fairly evenly among phylogenomic groups, and most phylogroups contained few isolates. The use of only short-read data also limited the completeness of our assemblies and our ability to identify plasmids and plasmid-encoded genes (*Arredondo-Alonso et al., 2017*). Further work involving more isolates and using both short- and long-read data will be necessary to identify unique, characteristic genetic components for each species and phylogroup, and to establish their spatial and functional niches within the chicken gut.

Recent studies have found similarly diverse populations of *Escherichia* after sampling of healthy pigs (*Ahmed, Olsen & Herrero-Fresno, 2017*; *Bok et al., 2020*), cattle and sheep (*Shaw et al., 2021*), and non-human primates (*Foster-Nyarko et al., 2020*), although it is notable that representatives of the 'cryptic clades' are scarce within these data. The reasons for this scarcity are unclear but are likely to include the use of different selection methods and the focus on virulent/AMR strains and *E. coli sensu stricto*. *Escherichia* diversity varies among samples from healthy humans (*Bok et al., 2018*; *Foster-Nyarko et al., 2021a*). High *Escherichia* species diversity has been detected within chickens previously (*Vounba et al., 2019a*; *Foster-Nyarko et al., 2021b*), although not from such a narrow pool of host birds as in the current study. Members of the cryptic clades have also previously been found in relatively high abundance in wild birds of various species (*Clermont et al., 2011*) but a definitive link to chickens has not been established. A recent metagenomics study included 60 different fecal samples from the same birds used here. Culturing of 3 of those samples yielded isolates of *E. marmotae, E. whittamii* and phylogroups A, B1, C, D and E, reflecting the diversity found in this study (*Gilroy et al., 2021*). However, most studies in chickens have taken samples from diseased birds and/or have focused on antimicrobial resistance and pathogenicity of the isolates recovered (*Braga et al., 2016*; *Cummins et al., 2019*; *Vounba et al., 2019b*; *Papouskova et al., 2020*; *Kubelová et al., 2021*).

CONCLUSIONS

The dearth of whole-genome sequencing studies that include all members of *Escherichia* (and sufficient metadata) skews our understanding of the genus and makes it impossible to discern patterns within or between sampling environments, geographical regions, diets, health status and so on. Consequently, there is much still to learn about the population structure and ecology of *Escherichia* (*Lagerstrom & Hadly, 2021*). We were surprised to find that isolates representing almost the entire known phylogeny of *Escherichia* were recovered from fecal samples from a small flock of healthy layer chickens, including species that have only recently been recognized (previously known as cryptic clades) and for which chicken-associated isolates have not been widely reported. These isolates had low carriage rates of antimicrobial resistance genes and virulence factors, suggesting that similar isolates might often be overlooked in studies that focus on these traits. Our findings highlight the surprising diversity of *Escherichia* harbored by even an individual chicken and emphasize the need to broaden the focus of research to encapsulate the full variety of species.

ACKNOWLEDGEMENTS

We thank Mr Steven Rudder and Mr David Baker for their assistance with sequencing, and the farm for assisting with the study.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Nicholas M. Thomson, Rachel Gilroy and Mark J. Pallen were supported by the BBSRC Institute Strategic Programme Microbes in the Food Chain BB/R012504/1 and its constituent project BBS/E/F/000PR10351. Maria Getino, Arnoud H.M. van Vliet and Roberto M. La Ragione were supported by the European Joint Programme One Health EJP. This project received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 773830. There was no additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
BBSRC Institute Strategic Programme Microbes in the Food Chain BB/R012504/1 and constituent project BBS/E/F/000PR10351.
European Joint Programme One Health EJP.
EU Horizon 2020 Research and Innovation Programme: 773830.

Competing Interests

The authors declare that there are no competing interests.

Author Contributions

- Nicholas M. Thomson conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Rachel Gilroy conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Maria Getino conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Ebenezer Foster-Nyarko analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Arnoud H.M. van Vliet analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Roberto M. La Ragione conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Mark J. Pallen conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.

Animal Ethics

The following information was supplied relating to ethical approvals (*i.e.*, approving body and any reference numbers):

The study was approved by the University of Surrey's NASPA Ethical Review Assessment Committee with project number NERA-2018-011.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

The raw sequencing reads and assemblies used in this work are available at NCBI SRA: PRJNA757375.

The individual accession numbers of the sequencing reads and assemblies for the isolates are available in Table S1.

Data Availability

The following information was supplied regarding data availability:

The raw data is available in the Supplemental Files.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.12935#supplemental-information.

REFERENCES

Abram K, Udaondo Z, Bleker C, Wanchai V, Wassenaar TM, Robeson MS, Ussery DW. 2021. Mash-based analyses of Escherichia coli genomes reveal 14 distinct phylogroups. *Communications Biology* **4(1)**:570 DOI 10.1038/s42003-020-01626-5.

Ahmed S, Olsen JE, Herrero-Fresno A. 2017. The genetic diversity of commensal Escherichia coli strains isolated from nonantimicrobial treated pigs varies according to age group. *PLOS ONE* **12(5)**:e0178623 DOI 10.1371/journal.pone.0178623.

Arredondo-Alonso S, Willems RJ, van Schaik W, Schürch AC. 2017. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microbial Genomics* **3(10)**:e000128 DOI 10.1099/mgen.0.000128.

Beghain J, Bridier-Nahmias A, Le Nagard H, Denamur E, Clermont O. 2018. ClermonTyping: an easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. *Microbial Genomics* **4(7)**:1–8 DOI 10.1099/mgen.0.000192.

Bok E, Kożańska A, Mazurek-Popczyk J, Wojciech M, Baldy-Chudzik K. 2020. Extended phylogeny and extraintestinal virulence potential of commensal Escherichia coli from piglets and sows. *International Journal of Environmental Research and Public Health* **17(1)**:366 DOI 10.3390/ijerph17010366.

Bok E, Mazurek J, Myc A, Stosik M, Wojciech M, Baldy-Chudzik K. 2018. Comparison of commensal escherichia coli isolates from adults and young children in Lubuskie province, Poland: virulence potential, phylogeny and antimicrobial resistance. *International Journal of Environmental Research and Public Health* **15(4)**:1–19 DOI 10.3390/ijerph15040617.

Braga JFV, Chanteloup NK, Trotereau A, Baucheron S, Guabiraba R, Ecco R, Schouler C. 2016. Diversity of Escherichia coli strains involved in vertebral osteomyelitis and arthritis in broilers in Brazil. *BMC Veterinary Research* **12(1)**:1–12 DOI 10.1186/s12917-016-0762-0.

Carattoli A, Zankari E, Garciá-Fernández A, Larsen MV, Lund O, Villa L, Aarestrup FM, Hasman H. 2014. In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrobial Agents and Chemotherapy* **58(7)**:3895–3903 DOI 10.1128/AAC.02412-14.

Clermont O, Christenson JK, Denamur E, Gordon DM. 2013. The Clermont Escherichia coli phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environmental Microbiology Reports* **5(1)**:58–65 DOI 10.1111/1758-2229.12019.

Clermont O, Dixit OVA, Vangchhia B, Condamine B, Dion S, Bridier-Nahmias A, Denamur E, Gordon D. 2019. Characterization and rapid identification of phylogroup G in Escherichia coli, a lineage with high virulence and antibiotic resistance potential. *Environmental Microbiology* **21(8)**:3107–3117 DOI 10.1111/1462-2920.14713.

Clermont O, Gordon DM, Brisse S, Walk ST, Denamur E. 2011. Characterization of the cryptic Escherichia lineages: rapid identification and prevalence. *Environmental Microbiology* **13(9)**:2468–2477 DOI 10.1111/j.1462-2920.2011.02519.x.

Connor TR, Loman NJ, Thompson S, Smith A, Southgate J, Poplawski R, Bull MJ, Richardson E, Ismail M, Thompson SE, Kitchen C, Guest M, Bakke M, Sheppard SK, Pallen MJ. 2016. CLIMB (the cloud infrastructure for microbial bioinformatics): an online resource for the medical microbiology community. *Microbial Genomics* **2(9)**:e000086 DOI 10.1099/mgen.0.000086.

Cummins ML, Reid CJ, Chowdhury PR, Bushell RN, Esbert N, Tivendale KA, Noormohammadi AH, Islam S, Marenda MS, Browning GF, Markham PF, Djordjevic SP. 2019. Whole genome sequence analysis of Australian avian pathogenic Escherichia coli that carry the class 1 integrase gene. *Microbial Genomics* **5(2)**:299 DOI 10.1099/mgen.0.000250.

De Pace F, Nakazato G, Pacheco A, De Paiva JB, Sperandio V, Da Silveira WD. 2010. The type VI secretion system plays a role in type 1 fimbria expression and pathogenesis of an avian pathogenic *Escherichia coli* strain. *Infection and Immunity* **78**:4990–4998 DOI 10.1128/IAI.00531-10.

Delmont TO, Eren EM. 2018. Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. *PeerJ* **2018**:1–23 DOI 10.7717/peerj.4320.

Devanga Ragupathi NK, Muthuirulandi Sethuvel DP, Inbanathan FY, Veeraraghavan B. 2018. Accurate differentiation of Escherichia coli and Shigella serogroups: challenges and strategies. *New Microbes and New Infections* **21**:58–62 DOI 10.1016/j.nmni.2017.09.003.

Dodson KW, Jacob-Dubuisson F, Striker RT, Hultgren SJ. 1993. Outer-membrane PapC molecular usher discriminately recognizes periplasmic chaperone-pilus subunit complexes. *Proceedings of the National Academy of Sciences of the United States of America* **90**:3670–3674 DOI 10.1073/pnas.90.8.3670.

Dziva F, Stevens MP. 2008. Colibacillosis in poultry: unravelling the molecular basis of virulence of avian pathogenic Escherichia coli in their natural hosts. *Avian Pathology* **37(4)**:355–366 DOI 10.1080/03079450802216652.

Eren AM, Esen OC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015. Anvi'o: an advanced analysis and visualization platformfor 'omics data. *PeerJ* **2015(358)**:1–29 DOI 10.7717/peerj.1319.

Foster-Nyarko E, Alikhan N, Ikumapayi UN, Sarwar G, Okoi C, Tientcheu PM, Defernez M, O'Grady J, Antonio M, Pallen MJ. 2021a. Genomic diversity of Escherichia coli from healthy children in rural Gambia. *PeerJ* **9(Suppl 4)**:e10572 DOI 10.7717/peerj.10572.

Foster-Nyarko E, Alikhan NF, Ravi A, Thilliez G, Thomson NM, Baker D, Kay G, Cramer JD, O'grady J, Antonio M, Pallen MJ. 2020. Genomic diversity of Escherichia coli isolates from non-human primates in the Gambia. *Microbial Genomics* **6(9)**:1–15 DOI 10.1099/mgen.0.000428.

Foster-Nyarko E, Alikhan N-F, Ravi A, Thomson NM, Jarju S, Kwambana-Adams BA, Secka A, O'grady J, Antonio M, Pallen MJ. 2021b. Genomic diversity of escherichia coli isolates from backyard chickens and guinea fowl in the gambia. *Microbial Genomics* **7(1)**:292 DOI 10.1099/mgen.0.000484.

Fox S, Goswami C, Holden M, Connolly JPR, Mordue J, O'Boyle N, Roe A, Connor M, Leanord A, Evans TJ. 2020. A highly conserved complete accessory Escherichia coli type III secretion system 2 is widespread in bloodstream isolates of the ST69 lineage. *Scientific Reports* **10(1)**:1–11 DOI 10.1038/s41598-020-61026-x.

Francetic O, Pugsley AP. 1996. The cryptic general secretory pathway (gsp) operon of Escherichia coli K-12 encodes functional proteins. *Journal of Bacteriology* **178(12)**:3544–3549 DOI 10.1128/jb.178.12.3544-3549.1996.

Franzin FM, Sircili MP. 2015. Locus of enterocyte effacement: a pathogenicity island involved in the virulence of enteropathogenic and enterohemorragic Escherichia coli subjected to a complex network of gene regulation. *BioMed Research International* **2015(3)**:1–10 DOI 10.1155/2015/534738.

Fratamico PM, DebRoy C, Liu Y, Needleman DS, Baranzoni GM, Feng P. 2016. Advances in molecular serotyping and subtyping of Escherichia coli. *Frontiers in Microbiology* **7(e0126749)**:1–8 DOI 10.3389/fmicb.2016.00644.

Gilroy R, Ravi A, Getino M, Pursley I, Horton DL, Alikhan N-F, Baker D, Gharbi K, Hall N, Watson M, Adriaenssens EM, Foster-Nyarko E, Jarju S, Secka A, Antonio M, Oren A, Chaudhuri RR, La Ragione R, Hildebrand F, Pallen MJ. 2021. Extensive microbial diversity within the chicken gut microbiome revealed by metagenomics and culture. *PeerJ* **9(7753)**:1–142 DOI 10.7717/peerj.10941.

Gomes TAT, Ooka T, Hernandes RT, Yamamoto D, Hayashi T. 2020. Escherichia albertii Pathogenesis. *EcoSal Plus* **9(1)** DOI 10.1128/ecosalplus.ESP-0015-2019.

Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11(1)**:673 DOI 10.1186/1471-2105-11-119.

Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* **9(1)**:1–8 DOI 10.1038/s41467-018-07641-9.

Jang J, Hur HG, Sadowsky MJ, Byappanahalli MN, Yan T, Ishii S. 2017. Environmental Escherichia coli: ecology and public health implications—a review. *Journal of Applied Microbiology* **123(3)**:570–581 DOI 10.1111/jam.13468.

Johnson JR, Brown JJ. 1996. A novel multiply primed polymerase chain reaction assay for identification of variant papG genes encoding the Gal(α1-4)Gal-binding PapG adhesins of *Escherichia coli*. *Journal of Infectious Diseases* **173**:920–926 DOI 10.1093/infdis/173.4.920.

Johnson JR, Murray AC, Gajewski A, Sullivan M, Snippes P, Kuskowski MA, Smith KE. 2003. Isolation and molecular characterization of nalidixic acid-resistant extraintestinal pathogenic *Escherichia coli* from retail chicken products. *Antimicrobial Agents and Chemotherapy* **47**:2161–2168 DOI 10.1128/AAC.47.7.2161-2168.2003.

Johnson TJ, Wannemuehler Y, Doetkott C, Johnson SJ, Rosenberger SC, Nolan LK. 2008. Identification of minimal predictors of avian pathogenic Escherichia coli virulence for use as a rapid diagnostic tool. *Journal of Clinical Microbiology* **46(12)**:3987–3996 DOI 10.1128/JCM.00816-08.

Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14(6)**:587–589 DOI 10.1038/nmeth.4285.

Kaper JB, Nataro JP, Mobley HLT. 2004. Pathogenic Escherichia coli. *Nature Reviews Microbiology* **2(2)**:123–140 DOI 10.1038/nrmicro818.

Kariyawasam S, Nolan LK. 2011. papA gene of avian pathogenic *Escherichia coli*. *Avian Diseases* **55**:532–538 DOI 10.1637/9663-011911-Reg.1.

Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America* **102(7)**:2567–2572 DOI 10.1073/pnas.0409727102.

Kubelová M, Koláčková I, Karpíšková R, Gelbíčová T, Florianová M, Kalová A. 2021. Virulence properties of mcr-1-positive Escherichia coli isolated from retail poultry meat. *Microorganisms* **9(2)**:1–9 DOI 10.3390/microorganisms9020308.

Lagerstrom KM, Hadly EA. 2021. The under-investigated wild side of Escherichia coli: genetic diversity, pathogenicity and antimicrobial resistance in wild animals. *Proceedings of the Royal Society B: Biological Sciences* **288(1948)**:20210399 DOI 10.1098/rspb.2021.0399.

Letunic I, Bork P. 2021. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* **49(W1)**:W293–W296 DOI 10.1093/nar/gkab301.

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. *International Journal of Systematic and Evolutionary Microbiology* **65(Pt_7)**:2130–2134 DOI 10.1099/ijs.0.000228.

Lu S, Jin D, Wu S, Yang J, Lan R, Bai X, Liu S, Meng Q, Yuan X, Zhou J, Pu J, Chen Q, Dai H, Hu Y, Xiong Y, Ye C, Xu J. 2016. Insights into the evolution of pathogenicity of Escherichia coli from genomic analysis of intestinal E. coli of Marmota himalayana in Qinghai-Tibet plateau of China. *Emerging Microbes and Infections* **5(1)**:1–9 DOI 10.1038/emi.2016.122.

Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America* **95(6)**:3140–3145 DOI 10.1073/pnas.95.6.3140.

Mehat JW, van Vliet AHM, La Ragione RM. 2021. The Avian Pathogenic Escherichia coli (APEC) pathotype is comprised of multiple distinct, independent genotypes. *Avian Pathology* **50(5)**:402–416 DOI 10.1080/03079457.2021.1915960.

Mohapatra BR, Broersma K, Mazumder A. 2007. Comparison of five rep-PCR genomic fingerprinting methods for differentiation of fecal Escherichia coli from humans, poultry and wild birds. *FEMS Microbiology Letters* **277(1)**:98–106 DOI 10.1111/j.1574-6968.2007.00948.x.

Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32(1)**:268–274 DOI 10.1093/molbev/msu300.

Nicoletti M, Superti F, Conti C, Calconi A, Zagaglia C. 1988. Virulence factors of lactose-negative Escherichia coli strains isolated from children with diarrhea in Somalia. *Journal of Clinical Microbiology* **26(3)**:524–529 DOI 10.1128/jcm.26.3.524-529.1988.

Olm MR, Brown CT, Brooks B, Banfield JF. 2017. DRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME Journal* **11(12)**:2864–2868 DOI 10.1038/ismej.2017.126.

Papouskova A, Papouskova A, Masarikova M, Masarikova M, Valcek A, Valcek A, Senk D, Cejkova D, Jahodarova E, Cizek A, Cizek A. 2020. Genomic analysis of Escherichia coli strains isolated from diseased chicken in the Czech Republic. *BMC Veterinary Research* **16(1)**:1–10 DOI 10.1186/s12917-020-02407-2.

Priest FG, Barker M. 2010. Gram-negative bacteria associated with brewery yeasts: reclassification of Obesumbacterium proteus biogroup 2 as Shimwellia pseudoproteus gen. nov., sp. nov., and transfer of Escherichia blattae to Shimwellia blattae comb. nov. *International Journal of Systematic and Evolutionary Microbiology* **60(4)**:828–833 DOI 10.1099/ijs.0.013458-0.

Roer L, Tchesnokova V, Allesøe R, Muradova M, Chattopadhyay S, Ahrenfeldt J, Thomsen MCF, Lund O, Hansen F, Hammerum AM, Sokurenko E, Hasman H. 2017. Development of a web tool for Escherichia coli subtyping based on fimH alleles. *Journal of Clinical Microbiology* **55(8)**:2538–2543 DOI 10.1128/JCM.00737-17.

Rouquet G, Porcheron G, Barra C, Répérant M, Chanteloup NK, Schouler C, Gilot P. 2009. A metabolic operon in extraintestinal pathogenic *Escherichia coli* promotes fitness under stressful conditions and invasion of eukaryotic cells. *Journal of Bacteriology* **191**:4427–4440 DOI 10.1128/JB.00103-09.

Sarowska J, Futoma-Koloch B, Jama-Kmiecik A, Frej-Madrzak M, Ksiazczyk M, Bugla-Ploskonska G, Choroszy-Krol I. 2019. Virulence factors, prevalence and potential transmission of extraintestinal pathogenic Escherichia coli isolated from different sources: recent reports. *Gut Pathogens* **11(1)**:1–16 DOI 10.1186/s13099-019-0290-0.

Schouler C, Schaeffer B, Brée A, Mora A, Dahbi G, Biet F, Oswald E, Mainil J, Blanco J, Moulin-Schouleur M. 2012. Diagnostic strategy for identifying avian pathogenic Escherichia coli based on four patterns of virulence genes. *Journal of Clinical Microbiology* **50(5)**:1673–1678 DOI 10.1128/JCM.05057-11.

Schwengers O, Barth P, Falgenhauer L, Hain T, Chakraborty T, Goesmann A. 2020. Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microbial Genomics* **6(10)**:1–12 DOI 10.1099/mgen.0.000398.

Schürch AC, Arredondo-Alonso S, Willems RJL, Goering RV. 2018. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene-based approaches. *Clinical Microbiology and Infection* **24(4)**:350–354 DOI 10.1016/j.cmi.2017.12.016.

Selander RK, Musser JM, Caugant DA, Gilmour MN, Whittam TS. 1987. Population genetics of pathogenic bacteria. *Microbial Pathogenesis* **3(1)**:1–7 DOI 10.1016/0882-4010(87)90032-5.

Shaw LP, Chau KK, Kavanagh J, AbuOun M, Stubberfield E, Gweon HS, Barker L, Rodger G, Bowes MJ, Hubbard ATM, Pickford H, Swann J, Gilson D, Smith RP, Hoosdally SJ, Sebra R, Brett H, Peto TEA, Bailey MJ, Crook DW, Read DS, Anjum MF, Walker AS, Stoesser N. 2021. Niche and local geography shape the pangenome of wastewater- and livestock-associated Enterobacteriaceae. *Science Advances* **7(15)**:eabe3868 DOI 10.1126/sciadv.abe3868.

Shi R, Yang X, Chen L, Chang HT, Liu HY, Zhao J, Wang XW, Wang CQ. 2014. Pathogenicity of Shigella in chickens. *PLOS ONE* **9(6)**:1–7 DOI 10.1371/journal.pone.0100264.

Smati M, Clermont O, Bleibtreu A, Fourreau F, David A, Daubié AS, Hignard C, Loison O, Picard B, Denamur E. 2015. Quantitative analysis of commensal Escherichia coli populations reveals host-specific enterotypes at the intra-species level. *MicrobiologyOpen* **4(4)**:604–615 DOI 10.1002/mbo3.266.

Spurbeck RR, Dinh PC, Walk ST, Stapleton AE, Hooton TM, Nolan LK, Kim KS, Johnson JR, Mobley HLT. 2012. *Escherichia coli* isolates that carry *vat*, *fyua*, *chua*, and *yfcv* efficiently colonize the urinary tract. *Infection and Immunity* **80**:4115–4122 DOI 10.1128/IAI.00752-12.

Stromberg ZR, Johnson JR, Fairbrother JM, Kilbourne J, Van Goor A, Curtiss R, Mellata M. 2017. Evaluation of *Escherichia coli* isolates from healthy chickens to determine their potential risk to poultry and human health. *PLOS ONE* **12**:e0180599 DOI 10.1371/journal.pone.0180599.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Smirnov S, Nikolskaya AN, Rao BS, Mekhedov SL, Sverlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. 2003. The COG database: an updated vesion includes eukaryotes. *BMC Bioinformatics* **4(1)**:1–14 DOI 10.1186/1471-2105-4-41.

Tenaillon O, Skurnik D, Picard B, Denamur E. 2010. The population genetics of commensal Escherichia coli. *Nature Reviews Microbiology* **8(3)**:207–217 DOI 10.1038/nrmicro2298.

Touchon M, Perrin A, De Sousa JAM, Vangchhia B, Burn S, O'Brien CL, Denamur E, Gordon D, Rocha EPC. 2020. Phylogenetic background and habitat drive the genetic diversification of Escherichia coli. *PLOS Genetics* **16(6)**:e1008866 DOI 10.1371/journal.pgen.1008866.

Uelze L, Grützke J, Borowiak M, Hammerl JA, Juraschek K, Deneke C, Tausch SH, Malorny B. 2020. Typing methods based on whole genome sequencing data. *One Health Outlook* **2(1)**:1–19 DOI 10.1186/s42522-020-0010-1.

van der Putten BCL, Matamoros S, Mende DR, Scholl ER, Consortium† C, Schultsz C. 2021. Escherichia ruysiae sp. nov., a novel Gram-stain-negative bacterium, isolated from a faecal sample of an international traveller. *International Journal of Systematic and Evolutionary Microbiology* **71(2)**:4609 DOI 10.1099/ijsem.0.004609.

Vounba P, Arsenault J, Bada-Alambédji R, Fairbrother JM. 2019a. Antimicrobial resistance and potential pathogenicity of Escherichia coli isolates from healthy broilers in Québec, Canada. *Microbial Drug Resistance* **25(7)**:1111–1121 DOI 10.1089/mdr.2018.0403.

Vounba P, Arsenault J, Bada-Alambédji R, Fairbrother JM. 2019b. Pathogenic potential and the role of clones and plasmids in beta-lactamase-producing E. coli from chicken faeces in Vietnam. *BMC Veterinary Research* **15(1)**:1–13 DOI 10.1186/s12917-019-1849-1.

Walk ST, Alm EW, Gordon DM, Ram JL, Toranzos GA, Tiedje JM, Whittam TS. 2009. Cryptic lineages of the genus Escherichia. *Applied and Environmental Microbiology* **75(20)**:6534–6544 DOI 10.1128/AEM.01262-09.

Wigley P. 2015. Blurred lines: pathogens, commensals, and the healthy gut. *Frontiers in Veterinary Science* **2(15)**:10–13 DOI 10.3389/fvets.2015.00040.

Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biology* **20(1)**:257 DOI 10.1186/s13059-019-1891-0.

Yang F, Yang J, Zhang X, Chen L, Jiang Y, Yan Y, Tang X, Wang J, Xiong Z, Dong J, Xue Y, Zhu Y, Xu X, Sun L, Chen S, Nie H, Peng J, Xu J, Wang Y, Yuan Z, Wen Y, Yao Z, Shen Y, Qiang B, Hou Y, Yu J, Jin Q. 2005. Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. *Nucleic Acids Research* **33(19)**:6445–6458 DOI 10.1093/nar/gki954.

Zhou Z, Alikhan NF, Mohamed K, Fan Y, Achtman M. 2020. The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. *Genome Research* **30(1)**:138–152 DOI 10.1101/gr.251678.119.

Zhou Z, Alikhan NF, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, Carriço JA, Achtman M. 2018. Grapetree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Research* **28(9)**:1395–1404 DOI 10.1101/gr.232397.117.