MICROBIAL GENOMICS

RESEARCH ARTICLE

Luo *et al.*, *Microbial Genomics* 2021;7:000710

DOI 10.1099/mgen.0.000710

Comparative genomics of Chinese and international isolates of *Escherichia albertii*: population structure and evolution of virulence and antimicrobial resistance

Lijuan Luo[1]†, Hong Wang[2]†, Michael J. Payne[1], Chelsea Liang[1], Li Bai[3], Han Zheng[4], Zhengdong Zhang[2], Ling Zhang[2], Xiaomei Zhang[1], Guodong Yan[2], Nianli Zou[2], Xi Chen[2], Ziting Wan[2], Yanwen Xiong[4], Ruiting Lan[1,*] and Qun Li[2,*]

Abstract

Escherichia albertii is a recently recognized species in the genus *Escherichia* that causes diarrhoea. The population structure, genetic diversity and genomic features have not been fully examined. Here, 169 *E. albertii* isolates from different sources and regions in China were sequenced and combined with 312 publicly available genomes (from additional 14 countries) for genomic analyses. The *E. albertii* population was divided into two clades and eight lineages, with lineage 3 (L3), L5 and L8 more common in China. Clinical isolates were observed in all clades/lineages. Virulence genes were found to be distributed differently among lineages: subtypes of the intimin encoding gene *eae* and the cytolethal distending toxin gene *cdtB* were lineage associated, and the second type three secretion system (ETT2) island was truncated in L3 and L6. Seven new *eae* subtypes and one new *cdtB* subtype (*cdtB*-VI) were identified. Alarmingly, 85.9 % of the Chinese *E. albertii* isolates were predicted to be multidrug-resistant (MDR) with 35.9 % harbouring genes capable of conferring resistance to 10 to 14 different drug classes. The majority of the MDR isolates were of poultry source from China and belonged to four sequence types (STs) [ST4638, ST4479, ST4633 and ST4488]. Thirty-four plasmids with some carrying MDR and virulence genes, and 130 prophages were identified from 17 complete *E. albertii* genomes. The 130 intact prophages were clustered into five groups, with group five prophages harbouring more virulence genes. We further identified three *E. albertii* specific genes as markers for the identification of this species. Our findings provided fundamental insights into the population structure, virulence variation and drug resistance of *E. albertii*.

DATA SUMMARY

(1) All newly sequenced data in this work were deposited in National Centre for Biotechnology Information (NCBI) under the BioProject of PRJNA693666.

(2) All accession numbers of the public available genomes were available in Table S1 (available in the online version of this article). Table S2–S9 are the supporting tables of the main results.

(3) The *eae* gene types including the seven newly defined types, as well as the three newly identified specific genes for *E. albertii* were deposited in Figshare (https://doi.org/10.6084/m9.figshare.14846994.v3).

Received 16 February 2021; Accepted 08 October 2021; Published 09 December 2021

Author affiliations: [1]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, New South Wales, Australia; [2]Zigong Center for Disease Control and Prevention, Zigong, Sichuan, PR China; [3]Division I of Risk Assessment, National Health Commission Key Laboratory of Food Safety Risk Assessment, Food Safety Research Unit (2019RU014) of Chinese Academy of Medical Science, China National Center for Food Safety Risk Assessment, Beijing, PR China; [4]State Key Laboratory of Infectious Disease Prevention and Control, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, PR China.

*Correspondence: Ruiting Lan, r.lan@unsw.edu.au; Qun Li, lqzgcdc@163.com

Keywords: *Escherichia albertii*; population structure; virulence; multidrug resistance; transmissible elements; species-specific marker genes.

Abbreviations: AR, antimicrobial resistance; CC, clonal complex; Cdt, cytolethal distending toxin; EHEC, enterohemorrhagic *Escherichia coli*; EPEC, enteropathogenic *E. coli*; ETT2, type III secretion system 2; HPI, high pathogenicity island; LEE, enterocyte effacement; MDR, multidrug resistance; MLST, multi-locus sequence typing; MVP, microbe versus phage; NCBI, National Center for Biotechnology Information; ST, sequence type; Stx, Shiga toxin; T3SS, type III secretion system; VFDB, virulence factor database.

All newly sequenced data in this work were deposited in National Center for Biotechnology Information (NCBI) under the BioProject of PRJNA693666.

†These authors contributed equally to this work

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Five supplementary figures and nine supplementary tables are available with the online version of this article.

(4) Figure S1–S5 are the supporting figures of the main results.

INTRODUCTION

Escherichia albertii is a recently defined species and a recognised foodborne human pathogen [1–3]. *E. albertii* mainly causes diarrhoea [3, 4], while bacteraemic human infections were also reported [5]. *E. albertii* has historically been misidentified as various pathogens such as enterohemorrhagic *Escherichia coli* (EHEC), enteropathogenic *E. coli* (EPEC), *Shigella boydii* serotype 13, and *Hafnia alvei* [1, 6]. In 2003, it was confirmed to be a novel species of the genus *Escherichia* and named as *E. albertii* [2, 6]. Through retrospective studies, *E. albertii* was found to be responsible for six human diarrhoea outbreaks in Japan from 2003 to 2015 [7, 8]. *E. albertii* can also cause infections in other animals. An outbreak of *E. albertii* infection in common redpoll finches in Alaska led to deaths of hundreds of birds in 2004 [9]. Furthermore, *E. albertii* has also been isolated from a variety of sources (including food products) and from wide geographic regions [9–12].

The pathogenicity of *E. albertii* was mainly attributed to a type III secretion system (T3SS) encoded by the locus of enterocyte effacement (LEE) and the cytolethal distending toxin (Cdt) encoded by the *cdtABC* operon, both of which were commonly found in *E. albertii* [1, 10, 13]. Based on the presence of the intimin *eae* gene, the LEE locus was found to be widely present in *E. albertii* [1, 10]. Non-LEE effector genes, which were mainly acquired through prophages in *E. coli* [14], were also observed in three *E. albertii* complete genomes [13]. Another *E. coli* type III secretion system 2 (ETT2), which has major effects on the surface proteins associated with serum survival (as a prerequisite for bloodstream infections) and motility of *E. coli,* has also been found in *E. albertii* [15]. ETT2 was predicted to be common in *E. albertii* based on the representative *eivG* gene [1, 13]. Shiga toxin (Stx) genes, stx_{2f} and stx_{2a}, are also sporadically observed in *E. albertii* [1]. However, the detailed distribution of these genes in *E. albertii* remained unclear, and other virulence factors reported in *E. coli* have not been systematically investigated in *E. albertii*.

Antimicrobial resistance (AR), especially multidrug resistance (MDR) which is defined as resistance to three or more drug classes, is an increasing global challenge [16]. Phenotypic AR and MDR of *E. albertii* strains were observed in Brazil and China [17, 18]. Poultry source *E. albertii* isolates in China were phenotypically resistant to up to 11 drug classes, some of which were commonly used in clinical treatment such as cephalosporins, aminoglycosides, fluoroquinolones, and beta-lactam antibiotics [17]. However, the overall presence of AR genes in *E. albertii* isolates from different geographic regions and sources remains unclear.

It is well known that transmissible elements, especially plasmids and phages, are associated with the acquisition of virulence and AR genes [19]. They are key transmissible elements for the acquisition of *stx* genes, T3SS effector genes, and other virulence genes in *E. coli* [19]. Multiple intact plasmids of *E. albertii* carrying virulence and MDR genes were reported [1, 17, 20]. However, plasmids in draft genomes of *E. albertii* and their association with the acquisition of AR and virulence genes remain to be characterised [1, 13]. Prophages have been found in *E. albertii* with 4–7 prophages per genome from three complete genomes analysed [1]. However, their carriage of virulence and AR genes has not been examined.

Two clades of *E. albertii* have previously been defined based on whole genome sequencing analysis [1, 21], with no isolates from China. In this work, *E. albertii* from different sources and regions of China were isolated and sequenced, including 163 draft and six complete genomes. Publicly available complete genomes and draft genomes of *E. albertii* were analysed together to elucidate the population structure, virulence and resistance of *E. albertii* and the relationships of Chinese and international isolates.

Impact Statement

E. albertii is an emerging foodborne pathogen causing diarrhoea. Elucidation of its genomic features is important for the surveillance and control of *E. albertii* infections. In this work, 169 *E. albertii* genomes from different sources and regions in China were collected and sequenced, which contributed to the currently limited genomic data pool of *E. albertii*. In combination with publicly available genomes, the population structure of *E. albertii* was defined. The presence and subtypes of virulence genes in different lineages were significantly different, indicating potential pathogenicity variation. Additionally, the presence of MDR genes was alarmingly high in the Chinese dominated lineages. MDR related STs and plasmid subtypes were identified, which could be used as sentinels for MDR surveillance. Moreover, the subtypes of plasmids and prophages were distributed differently across lineages, and were found to contribute to the acquisition of virulence and MDR genes in *E. albertii*. Altogether, this work revealed the diversity of *E. albertii* and characterised its genomic features in unprecedented detail. The three *E. albertii* specific genes found would facilitate the identification of this emerging foodborne pathogen.

METHODS

Genomic sequences

A total of 169 *E. albertii* isolates from different sources and regions in China were collected and sequenced. The *E. albertii* type strain LMG20976 was also sequenced in this study. All of the isolates were sequenced using Illumina sequencing [22], except for six isolates that were sequenced using Pacbio to obtain complete genomes [23].

Raw reads and assemblies of publicly available *E. albertii* isolates were downloaded. To identify *E. albertii* isolates that were potentially misidentified as *E. coli,* one reported specific gene (EAKF1_ch4033) of *E. albertii* [24], was searched against a total of 30 021 representative *E. coli* and *Shigella* genome assemblies using BLASTN, using coverage and identity thresholds of 50 and 70%, respectively.

In summary, there were a total of 482 genomic sequences of *E. albertii* included in this study (Table S1). For draft genome sequences, 164 were from this study and 296 were from public databases (255 raw reads from European Nucleotide Archive and 41 assemblies from NCBI). For complete genomes, there were six genomes from this study, and 16 from NCBI (ten of which were sequenced using PacBio). Raw reads of Illumina sequencing were assembled using Skesa v2.4.0 [25].

Phylogenetic analysis and *in silico* multilocus sequence typing (MLST) of *E. albertii*

In an initial analysis, 38 representative isolates were selected to represent *E. albertii* diversity to obtain the overall picture and to identify the root of the *E. albertii* phylogeny. Using *E. coli* (Accession No. NZ_CP014583.1) as a reference, SNPs were called by snippy v4.4.0 (https://github.com/tseemann/snippy), and recombinant SNPs were detected and removed by Gubbins v2.0.0 [26]. A maximum parsimony tree based on SNPs of the 38 isolates using *E. coli* as an outgroup was constructed by Mega X with 1000 bootstraps [27].

To elucidate the phylogenetic relationship of the 482 *E. albertii* isolates, a phylogenetic tree was constructed using quicktree.pl (which is a pipeline of SaRTree v1.2.2) with ASM287245v1 as reference [28]. The recombination sites of the SNPs were removed using RecDetect v6.0 [28]. The SNP alignment of the genomes was analysed with Fastbaps v1.0.4 to identify lineages of *E. albertii* [29]. The lineages defined were mapped onto the phylogenetic tree using iTOL v4 [30].

The *in silico* MLST which is based on the seven housekeeping genes of *E. coli*, was performed on *E. albertii* with sequence types (STs) assigned [31]. Clonal complexes (CCs) of the STs were called based on one allele difference using the eBURST algorithm [32].

Screening for *E. albertii* specific gene markers

To screen for *E. albertii* specific gene markers, a total of 243 *E. albertii* genomes were randomly sampled from different lineages and 1898 representative genomes of *E. coli* and *Shigella* from the 'identification dataset' defined by Xiaomei *et al.* [33] were used. Those 2141 genomes were annotated with Prokka v1.13.3 [34]. The pangenome was defined using Roary v3.12.0 with a nucleotide identity threshold of 80% [35]. Using Scoary, genes that were significantly associated with *E. albertii* were identified [36]. Using BLASTN, the candidate-specific gene markers were further searched against 482 *E. albertii* genomes using coverage >=50% and identity >=70% as cutoffs, and 30 021 representative *E. coli* and *Shigella* genomes to evaluate sensitivity and specificity as used by Xiaomei *et al.* [33].

Virulence and antibiotic resistance analysis of *E. albertii*

Predicted virulence and AR genes from the *E. albertii* genomes were identified by Abricate v0.8.13 (https://github.com/tseemann/abricate): virulence genes were screened against the *E. coli* virulence factors database (Ecoli_VF) and the virulence factor database (VFDB) with an identity of >=70% and coverage of >=50% [37]; AR genes were screened through the NCBI AMRFinder database with the identity of >=90% and coverage of >=90% [16].

The pangenome of *E. albertii* and phylogeny of the *eae* and *cdtB* genes

To predict the subtypes of the *eae* and *cdtB* genes harboured by each *E. albertii* isolate, representative sequences for each *eae* and *cdtB* type were used to search the collection of *E. albertii* genomes using BLASTN with an identity of >=97% and coverage of >=50% [38]. The new *eae* and *cdtB* subtypes were defined based on the tree structure and BLASTN results. A new subtype was defined, if it was phylogenetically distant from the known subtypes and was present in >=5 isolates (with identity >=97%).

To construct the phylogenetic tree of *eae* and *cdtB* genes from different isolates of *E. albertii*, the pangenome of *E. albertii* was defined. High-quality assemblies were identified using the cutoffs of N50 >29 Kb, total length between 4.3 Mb and 5.7 Mb, and the number of contigs <=450 contigs based on the output of Quast v5.0.2 [39]. A total of 422 high-quality assemblies were included and were annotated using Prokka v1.12 [34]. The pangenome of *E. albertii* was defined using Roary v3.11.2 with the identity threshold of 70% [35]. The *eae* and *cdtB* genes for each *E. albertii* isolate were identified from the output of Roary and Prokka. MEGA X was used to align the nucleotide sequences of *eae* and *cdtB* genes using MUSCLE [27]. Neighbour-joining trees were constructed with partial deletion (90%) and 1000 bootstrap repeats using MEGA X [27].

Plasmid and prophage analysis based on complete genomes of *E. albertii*

For intact plasmids and prophages of *E. albertii*, 16 complete genomes by PacBio and one reference genome GCA_001549955.1 (sequenced by 454 GS-FLX) were selected for the prophage and plasmid analyses.

To identify the plasmids in the draft genomes, we used both PlasmidFinder and MOB-suite [32, 40]. Plasmid replicon genes were screened against the PlasmidFinder database with an identity of >=65% and coverage of >=50% using Abricate v0.8.13 (https://github.com/tseemann/abricate). MOB-suite was able to identify the potential plasmid sequences in draft genomes. MOB types were assigned if the predicted plasmids were known. To evaluate if the presence of the invasive plasmid

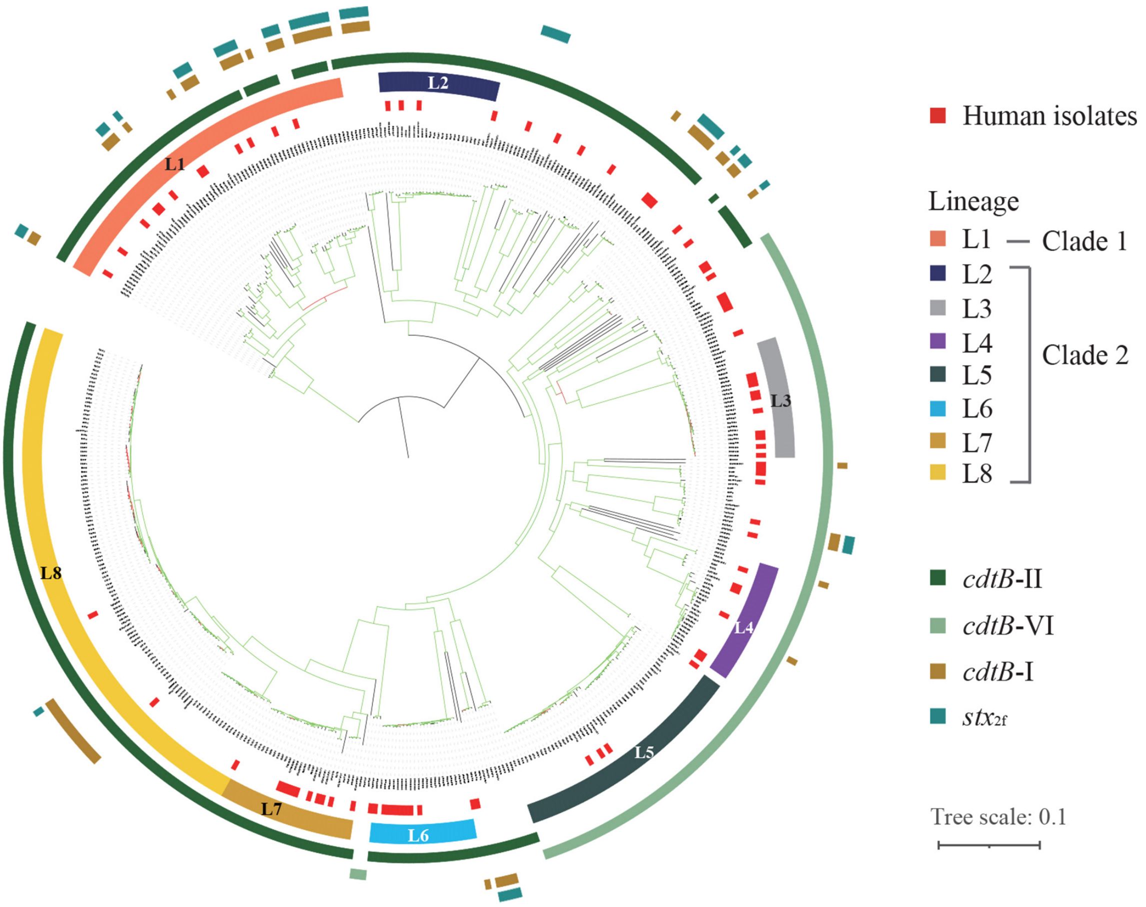

Fig. 1. Phylogenetic structure of *E. albertii*. The phylogenetic tree of the 482 *E. albertii* isolates was constructed using Quicktree with bootstrap replicates of 1000 [28]. The colour of the branches represented the percentage of bootstrap supporting from 10–100% (from red to green). The innermost ring marks the isolates from human source. The next ring marks the lineages by colour as shown in the colour legend. The outer four rings represented the *cdtB* subtypes and the stx_{2f} gene, which were represented with different colours as shown in the colour legend.

pINV of *Shigella* present in *E. albertii*, the pINV specific gene *ipaH* and 38 plasmid-borne virulence genes were screened in the raw reads of *E. albertii* using ShigEiFinder [33]. AR genes and virulence genes present on the intact plasmids and MOB-suite predicted plasmids were screened using the aforementioned criteria.

The complete genomes were submitted to Phaster for prophage prediction [41]. To define the groups of the intact prophages, the genomic sequences of prophages were annotated with Prokka v1.12 [34]. The gff files of the intact prophages were clustered by Roary v3.11.2 with an identity of >=70%, and a binary gene presence and absence tree was generated [35]. The concatenated prophage sequences in the order of binary clustering were visualized in similarity plots by Gepard v1.3 [42]. Genes whose presence was significantly associated with prophage groups (*P*<=0.001) were identified using Scoary [36]. The top three to five genes that were of 100% specificity and sensitivity for each prophage group were identified as potential prophage specific genes. These prophage specific gene candidates were searched against the 482 genomes with identity >=70% and coverage >=50% using BLASTN. The distribution of the prophage specific genes was visualized in Phandango [43]. AR genes, plasmid replicon genes and virulence genes present on the intact prophages were screened using the aforementioned criteria.

To compare the prophages of *E. albertii* with public phage clusters from the Microbe Versus Phage (MVP) database, the representative phage sequences of different phage clusters were downloaded [44]. Each prophage sequence of *E. albertii* was searched against the MVP reference phage cluster sequences with an identity of 80% and coverage of 50% using BLASTN [44].

RESULTS

A dataset representing *E. albertii* distribution in different source types and geographic regions

A total of 169 *eae* gene-positive *E. albertii* isolates from different regions of China were collected from 2014 to

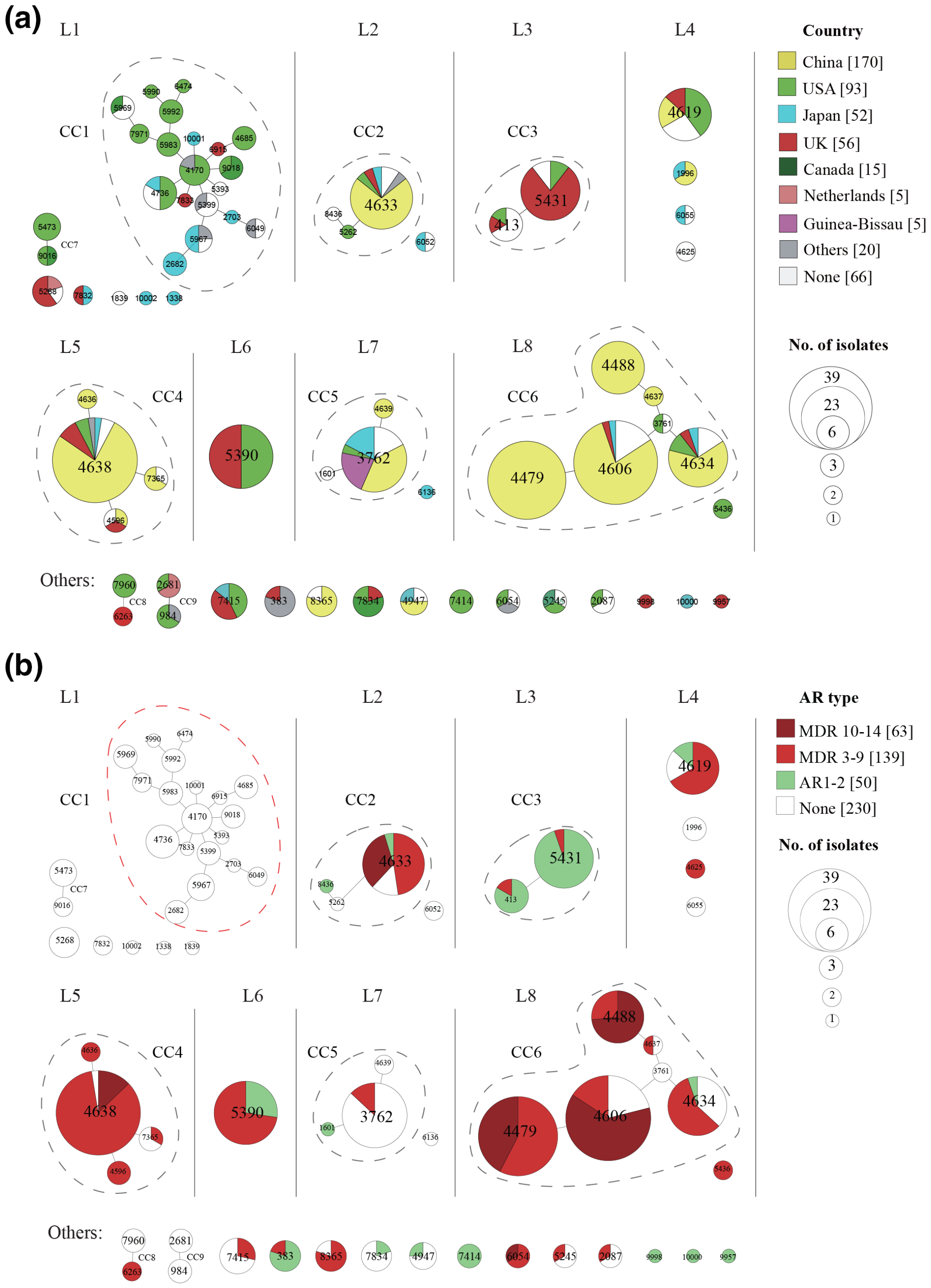

Fig. 2. Region distribution and resistance profiles of clonal complex (CC) and sequence type (ST) of *E. albertii* isolates based on the seven-gene multi-locus sequence typing (MLST). (a) Region distribution of STs and CCs. (b) Drug resistance profiles of STs and CCs. Each circle represented an ST and the size of the circles reflected the number of isolations. STs and CCs belonging to different lineages were separated. STs with one allele difference were linked with solid lines as one CC. Singleton STs were shown for each lineage. While for the 42 singleton STs belonging to none of the eight lineages, only 12 STs with AR genes were shown. The top seven countries with five or more isolates were highlighted in different colours as shown in the colour legend. The predicted antibiotic resistance of different STs is denoted by different colours of different levels of resistance (by the number of predicted drug classes as indicated) as shown in the colour legend. The numbers in the square brackets were the number of isolates. The pie chart within an ST denotes different proportions of isolates displaying a particular characteristic.

Table 1. Evaluation of *E. albertii* specific gene markers

Locus tag of reference strain KF1	Location (Strand)	Length	Sensitivity* (No. of false negatives)	Specificity† (No. of false positives)	NCBI BLASTN searches‡	Ref
EAKF1_ch4033	4 243 592…4 243 984 (-)	393	99.2% (4)	99.9% (33)	Some hits to *E. coli*	[24]
EAKF1_ch3804	3 999 536…3 999 946 (+)	411	100% (0)	100% (0)	One *E. coli*‡	
EAKF1_ch4075c	4 276 067…4 276 220 (-)	154	100% (0)	100% (0)	Specific	
EAKF1_ch0408c	429 916…430 443 (-)	528	100 % (0)	100 % (0)	Specific	

*Sensitivity=1-no. of false-negative genomes/total no. of *E. albertii* genomes.

†Specificity=1-no. of false-positive genomes/total no. of non-*E. albertii* genomes.

‡One complete genome of *E. coli* (Accession No. CP053258.1) from guinea fowl of ST9286.

2019, and sequenced in this study. The *E. albertii* isolates were from five provinces in China, the majority of which were from Sichuan province in Southern China and Shandong province in Northern China (Table S1). The Chinese *E. albertii* isolates belonged to seven different source types, with 90.5% from poultry intestine (with 110 isolates from chicken intestines and 43 from duck intestines). There were six human source isolates from China (Table S2). Three isolates were from patients with diarrhoea, including one patient with bloody diarrhoea. Three *E. albertii* isolates were from poultry butchers and retailers who were asymptomatic. Two *E. albertii* isolates were from the faecal samples of bats in Yunnan, China. Notably, as only *eae* positive and lactose nonfermenting samples were cultured for *E. albertii* in this study, any *eae* negative or lactose fermenting *E. albertii* isolates would not have been isolated.

To compare the genomic characteristics of *E. albertii* globally, a total of 312 publicly available *E. albertii* genome sequences were included in this study. Based on the metadata available, these isolates were from six continents and 12 different source types including humans, birds, bovine, swine, cats, water mammals, camelid, plants, soil and water. Humans (76 isolates) and birds (30 isolates) were the dominant sources (Table S3).

E. albertii lineages and their distribution in different geographic regions and source types

Previous studies showed that *E. albertii* is divided into two clades [1, 21]. To better define the phylogenetic lineages, we used Fastbaps to analyse the population divisions of the 482 *E. albertii* isolates using alignment of non-recombinant SNPs (with recombinant SNPs removed) as input. Eight lineages of *E. albertii* were defined containing 353 of the 482 isolates while 129 did not belong to any lineage (Fig. 1) [29]. Lineage 1 (L1) corresponds to previously defined clade 1, and L2 to L7 belonged to the previously defined clade 2 [1, 21]. It is noteworthy that the *E. albertii* isolates which were previously identified as *S. boydii* serotype 13 belonged to L3. Each lineage includes isolates from multiple continents. L5 and L8 were more common in Asia, while L1 (or clade 1), L3 and L6 were more common in Europe and North America (Fig. S1).

The 85 human isolates were distributed among the eight lineages indicating all of these lineages were potentially pathogenic to humans (Fig. 1). For Chinese *E. albertii* isolates, the six human source isolates belonged to L4 (2), L7 (1), and L8 (1), with two not falling into any lineages (Table S2). The two bat source isolates did not belong to any of the lineages but were most related to L3. There were 158 poultry source isolates from China, 55.7% (88/158) of which belonged to L8 followed by L5 (22.8%, 36/158) (Table S3), and there were two isolates of L8 from wild birds. By contrast, the majority of the bird source isolates from other countries came from wild birds, 51.6% (16/31) of which did not belong to any of the eight lineages while 32.3% (10/31) were from L1. These findings demonstrated that the bird source *E. albertii* isolates from the other countries were phylogenetically different from the wild birds and poultry source isolates in China.

In silico MLST of *E. albertii* isolates

We performed *in silico* MLST on the isolates using the established *E. coli* scheme [31]. The 482 *E. albertii* isolates were subtyped into 98 STs, among which 53 STs contained >=2 isolates. By lineage, with the exception of L1 and L8, each lineage was dominated by one ST. ST4633 accounted for 84.0% of the total number of isolates in L2, ST5431 for 76.0% of L3, ST4619 for 60.0% of L4, ST4638 for 81.3% of L5, ST5390 for 100% of L6 and ST3762 for 82.1% of L7. And 94.6% of L8 belonged to four STs (ST4488, ST4634, ST4479 and ST4606). We further grouped closely related STs as CC using one allele difference [45]. Nearly half of the STs (43 of 98) were grouped into nine CCs while the remaining 55 STs were singletons (Fig. 2a). With the exception of L4 and L6 which only contained STs, the other lineages were dominated by one CC. CC1 represented 68.1% of the L1 isolates. CC2 to CC6 were representative of more than 90% of the isolates in L2, L3, L5, L7 and L8, respectively. The majority of the singletons (42 of 55) belonged to none of

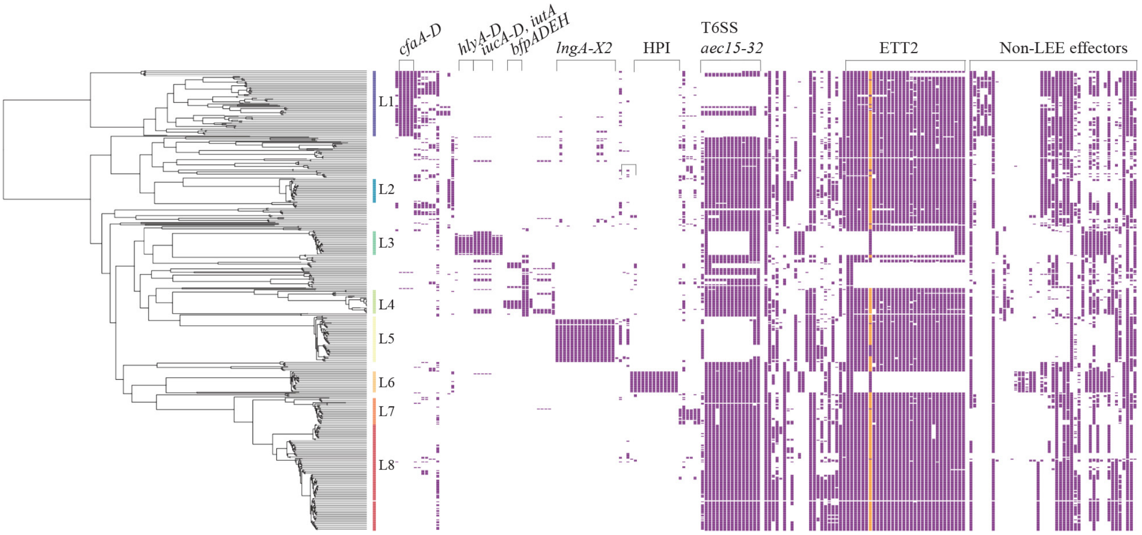

Fig. 3. Virulence genes that were significantly associated with different lineages of *E. albertii*. The distribution of different virulence genes in *E. albertii* was visualized using Phandango [43]. The lineages of *E. albertii* were labelled with different colours. The presence of a gene was marked with a coloured box. Only genes or gene clusters significantly associated with lineages were shown.

the eight lineages and were classified as other in the lineage division above.

Thirty-three STs were found in more than one country while 57 STs were only found in one country. The six largest CCs were found in more than one country. However, individual STs or CCs were predominant in different countries or regions. ST5390 was the most common ST in both USA and UK, and ST5431 was the second most common ST in the UK. In China, ST4479, ST4638 and ST4606 were the main STs, representing 54.7 % of the Chinese isolates. CC1 and CC3 were predominant in the USA and UK while CC2, CC4, and CC6 were predominantly found in China.

E. albertii specific gene markers

The reported specific gene of *E. albertii* EAKF1_ch4033 [24] was missing in four isolates (three ST5268 and one ST10002), all of which belonged to L1. Additionally, EAKF1_ch4033 was found in 33 *E. coli* genomes of ST378 and some *E. coli* assemblies in NCBI, although the estimated specificity is still 99.9% (Table 1). Therefore, we searched the *E. albertii* genomes as described in the methods and found three candidate specific genes (EAKF1_ch3804, EAKF1_ch4075c and EAKF1_ch0408c), which were positive in all of the *E. albertii* genomes and negative in all 30 021 *E. coli* and *Shigella* genomes (Table 1). We further searched the NCBI web database using BLASTN, EAKF1_ch3804 was present in one complete genome of *E. coli* ST9286 from guinea fowl [46] while the other two markers were only found in *E. albertii*. Therefore, EAKF1_ch4075c and EAKF1_ch0408c have 100% sensitivity and specificity, while EAKF1_ch3804 has 100% sensitivity and ~99.99% specificity. The DNA sequences of the three *E. albertii* specific genes are deposited in Figshare.

Virulence genes and their distribution in *E. albertii* lineages

Virulence genes from *E. coli*_VF database were screened to evaluate the potential pathogenicity of *E. albertii*. The LEE island from LEE1 to LEE7 contains 41 genes [47]. The 41 genes were present in slightly different proportions ranging from 91.1–99.8%, with the *espF* gene the lowest presenting in 439 of the 482 isolates (Table S4). The *eae* gene on LEE5 was harboured by 99.4% (479/482) of the isolates. Thirteen previously defined *eae* subtypes were observed in 387 (80.3%) of the 482 isolates, and seven new *eae* subtypes were identified (which were observed in >=5 isolates each) among the remaining 92 isolates (Fig. S2). Subtype sigma was the dominant type (37.9%), followed by rho (10.4%), itota2 (6.6%) and epsilon3 (6.2%) (Fig. S2). The *eae* subtypes were associated with specific lineages: epsilon3, iota2 and rho were the predominant subtypes in L2, L3 and L5, respectively, and subtype sigma was dominant in L6, L7 and L8. However, L1, L4, L5 and L7 harboured multiple *eae* subtypes. L1 (or clade 1), possessed eight *eae* subtypes, with beta3, alpha8 and the newly defined sigma2 and alpha9 as the main subtypes (Fig. S2).

Cdt facilitates bacterial survival and enhances pathogenicity [48] and is encoded by the *cdtABC* genes which were widely distributed in *E. albertii* [1, 49]. In this study, *cdtABC* genes were present in 99.4 % (479/482) of the isolates. The *cdtB* gene had been previously divided into five subtypes (*cdtB*-I to *cdtB*-V), with *cdtB*-II/III/V as one group, and *cdtB*-I/IV as another group [50]. By phylogenetic analysis of the *cdtB* genes in *E. albertii*, a new *cdtB* subtype was identified and named as *cdtB*-VI. *CdtB*-VI was phylogenetically closer to *cdtB* group II/III/V (Fig. S3). Notably, almost all *cdtB*-VI

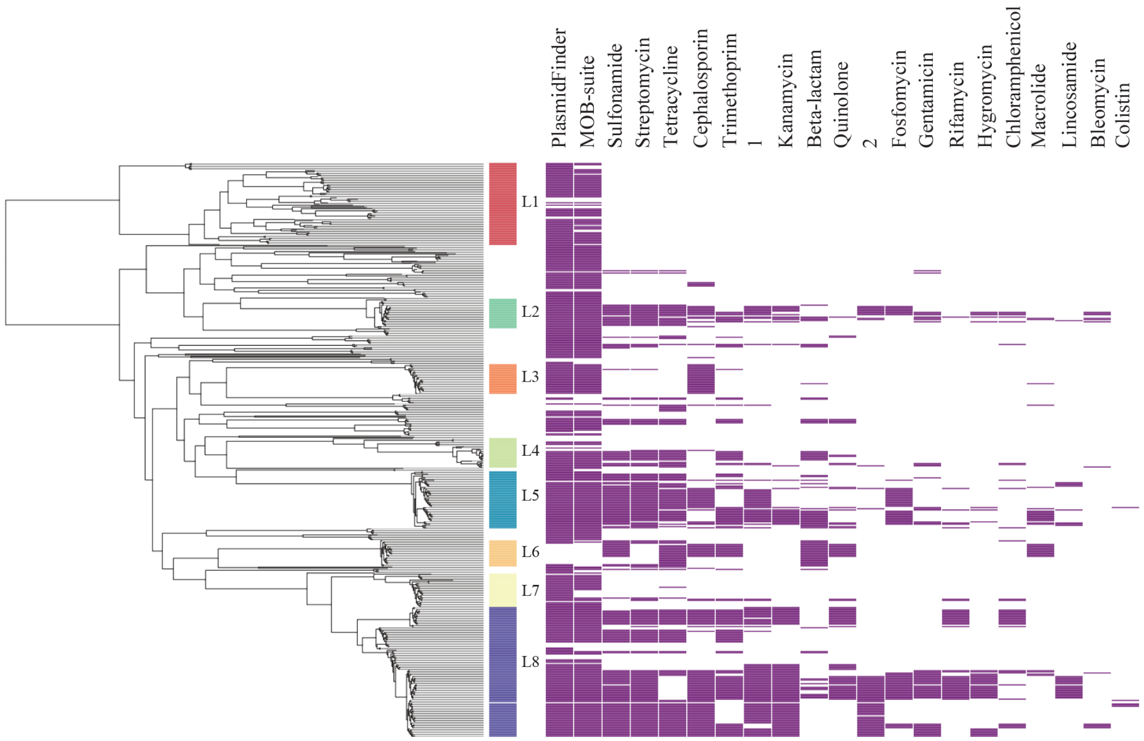

Fig. 4. Predicted resistance to drug classes in *E. albertii*. *E. albertii* isolates that harboured genes conferring resistance to different drug classes are shown in purple. The two columns headed with one and two denote the combination of two drugs as follows: 1 = chloramphenicol and florfenicol, 2 = phenicol and quinolone. Isolates with predicted plasmids by PlasmidFinder and MOB-suite (respectively) were also highlighted.

positive *E. albertii* isolates (30.1%, 145/482) were located on the same branch that includes L3, L4 and L5 isolates (Fig. 1). *CdtB*-II, as the dominant type, was present in 68.3% (329/482) of *E. albertii* isolates across five lineages (L1, L2, L6, L7 and L8). *CdtB*-I was found in 65 (13.5%) *E. albertii* isolates, 89.2% (58/65) of which were also positive for either *cdtB*-II or VI. There were 49 isolates positive for stx_{2f} (10.2%, 49/482), 44 of which possessed *cdtB*-I (Fig. 1). *E. albertii* isolates with *cdtB*-I were significantly more likely to harbour stx_{2f} gene (Chi-Square test, *P*<0.001). Both *cdtB*-I and stx_{2f} were observed on the same intact prophage of two complete genomes (ASM331252v2_PF4 and ASM386038v1_PF5). None of the Chinese *E. albertii* isolates were positive for stx_{2f}.

ETT2, which plays a role in motility and serum resistance in *E. coli* [15], was found to be nearly intact in 61.4 % (296/482) of the isolates, except for the *ygeF* gene which was absent in all *E. albertii* isolates [13]. Eighty-eight isolates (18.3%) harboured 29 to 31 ETT2 genes with two to four genes missing. Interestingly, ETT2 genes were mostly deleted in L3 and L6 with only four and three genes remaining, respectively (Fig. 3). Other virulence genes were also lineage-restricted, such as the type VI secretion system (T6SS) *aec* genes, which were present in most of the lineages except for L1, L3 and L5. The haemolysin genes *hlyABCD* were present only in L3 isolates (Fig. 3). The *iuc* gene cluster (*iuc-ABCD* and *iutA*) which encodes the siderophore aerobactin and the aerobactin receptor [51] was mainly present in L3 and L4 and one isolate of L6. The *Yersinia* high pathogenicity island (HPI), which encodes the yersiniabactin (Ybt) [52], was only found in L6 isolates (100%). The *lng* gene cluster that encodes the CS21 pilus (class b type IV) [53–55] was mainly observed in L5.

There were other *E. coli* virulence genes including *paa, efa1* and the bundle forming pilus encoding *bfp* genes that were found to be variably present in *E. albertii*, which are summarised in Table S4. One genome assembly (ERR1953722) from L5 was found to harbour *Shigella* invasive plasmid pINV genes [56]. However, further investigation by reads mapping found that it was most likely due to contamination (data not shown).

Drug resistance genes and their high prevalence in some STs of *E. albertii*

Presence of AR genes was screened using the NCBI AMRFinder database [16]. Among the 482 isolates, 52.3% (252/482) harboured AR genes, 41.9% (202/482) were MDR (harbouring AR genes conferring resistance to >=3 different drug classes), and 13.1% (63/482) harboured genes conferring resistance to 10 to 14 different drug classes that were regarded as highly resistant. Notably, 72.3% (146/202) of the predicted MDR isolates were from China with an AR rate of 88.2% and an MDR rate of 85.9%, with 61 isolates (35.9%)

being predicted to be highly resistant. The predicted AR drug classes were shown in Fig. 4, including sulfamethoxazole-trimethoprim, cephalosporin, streptomycin, beta-lactam antibiotics, etc. The AR genes observed in each isolate were shown in Table S5. We determined resistance profiles by STs and found that some STs contained a high proportion of MDR isolates. The predicted MDR rates in ST4638, ST4479, ST4633 and ST4488 were >=80% (Fig. 2b). Additionally, 63.2% of the isolates in ST4606 were highly resistant. For the top six STs in China representing 84.7% (144/170) of the Chinese isolates, 93.8% (135/144) of the isolates were predicted to be MDR, and 41.7% (60/144) were highly resistant. In contrast, isolates from the USA and UK had relatively lower predicted MDR rate (26.2%, 39/149) and were mainly observed in ST5390, ST4619 and ST4638, with only one highly resistant isolate (Fig. 2). By CCs, CC2, CC4 and CC6 had high MDR rates. CC1 carried hardly any resistance genes while CC3 and CC5 had low levels of carriage of resistance genes.

Plasmids and plasmid associated drug resistance and virulence genes

We firstly analysed the 17 complete *E. albertii* genomes for the carriage of plasmids. There were 34 intact plasmids ranging from 19 118 bp to 266 043 bp (Table S6). Nineteen plasmids were previously reported [1, 17, 20], while 15 plasmids were newly identified in this study.

We further performed plasmid typing using PlasmidFinder and MOB-suite [32, 40]. PlasmidFinder identifies plasmids by replicon types [32]. However, it should be noted that a plasmid may carry more than one replicon type. MOB-suite

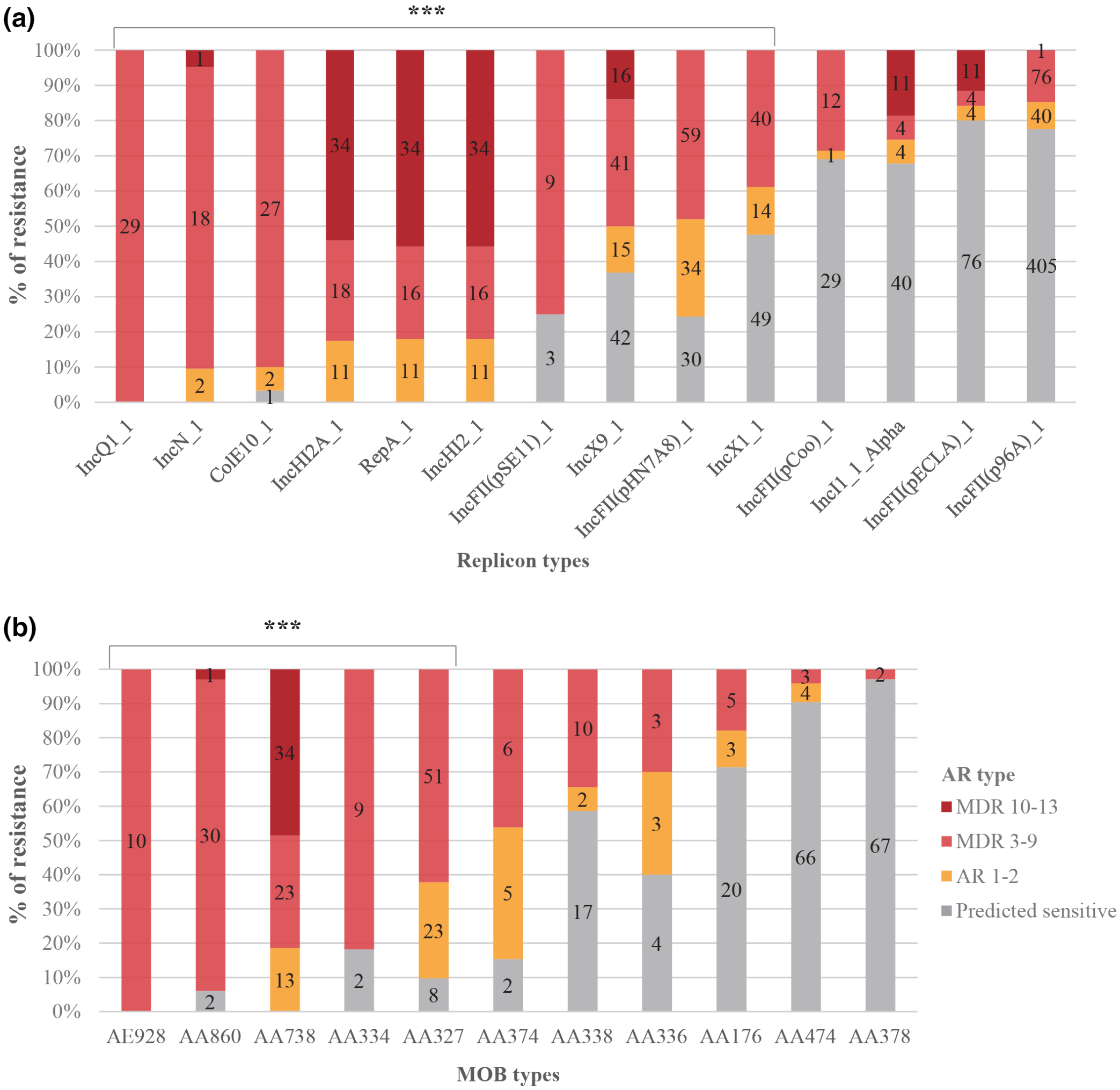

Fig. 5. Multidrug resistance (MDR) associated plasmid subtypes. (a) Replicon types detected. (b) MOB types detected. Those types significantly associated with MDR are marked with *** (*P*-value<0.001). The proportion of drug resistance (%) for each replicon or MOB type was shown as a colour legend.

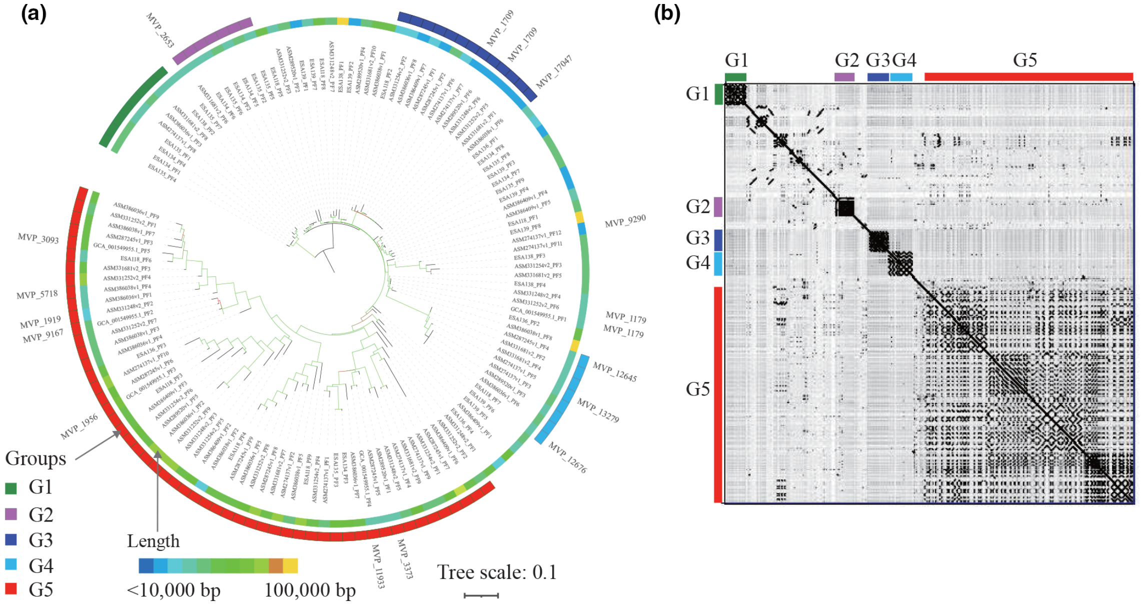

Fig. 6. Clustering of the intact prophages of *E. albertii*. (a) Accessory binary gene presence tree of the prophages constructed using Roary v3.11.2 [35]. The five main groups of prophages were labelled with different strip colours. There were 15 prophages of *E. albertii* with phage cluster types in the Microbe Versus Phage (MVP) database, the 15 MVP phage cluster types were labelled. (b) Dot plot of similarity of prophages using the nucleotide dot plot tool Gepard [42] and the five prophage groups were marked.

predicts plasmids using the relaxase gene and groups those predicted plasmids into different MOB types [40]. However, some plasmids have no relaxase gene. Thus, both methods were used to predict and identify plasmids in all *E. albertii* isolates. Among the 482 *E. albertii* isolates, PlasmidFinder found that 86.7% (418/482) of the isolates harboured plasmids, with a total of 54 replicon types detected. There were 34 replicon types that each was present in more than ten isolates. And 26 replicon types were found to be significantly associated with lineages (P<0.001) (Table S7): for example, IncFII(29)_pUTI89 type with L2, Col156 with L3, and IncFII (pSE11) with L4, IncX1, IncX9, IncHI2, IncHI2A and RepA with L5 and L8. By MOB-suite, a total of 1854 plasmid sequences were predicted in 427 of the 482 isolates with an average of 4.3 plasmids per genome while 55 isolates had no plasmids predicted. The vast majority (90.3%, 1674/1854) of the predicted plasmids were grouped into 170 MOB types with the remaining 9.7% (180/1854) being novel with no MOB types. There were 47 MOB types each of which was present in >=10 isolates, 36 of which were significantly associated with different lineages, which is concordant with findings from replicon types (Table S7). Additionally, there were 64 isolates with neither replicon types nor MOB types observed, including 77.3% (17/22) of L6 isolates (Fig. 3). However, 35.9% (23/64) of these isolates harboured AR genes, especially 72.3% of L6 were predicted to be MDR.

Plasmids are known to be responsible for the acquisition of MDR genes. Among the 34 intact plasmids, nine were found to harbour AR genes (Table S6). One newly identified MDR plasmid, ESA136_plas1 (Accession No. CP070297.2) which is of MOB type AA738 and replicon types IncHI2, IncHI2A and RepA, contained 15 AR genes resistant against 13 drug classes.

Since plasmids were not fully assembled in the draft genomes, we used the statistical association to determine which plasmid types were likely to carry the MDR genes. Note that this analysis was not aimed to determine whether these plasmid types were more likely to be associated with MDR in general. By PlasmidFinder, 13 replicon types were found to be significantly associated with MDR genes (P<0.001, Chi-square test) (Fig. S4). However, this analysis may be biassed when the MDR genes were not located on the same plasmid as the replicon genes. This bias can be resolved by MOB-suite, which offers the predicted plasmid sequences from the draft genomes. We screened the plasmid replicon genes and MDR genes on the MOB-suite predicted plasmids. Ten replicon types were confirmed to be significantly more likely to be observed in MDR isolates (P<0.001) including IncQ, IncN, ColE10, IncHI2A, IncHI2, RepA, IncFII(pSE11), IncX9, IncFII(pHN7A8), and IncX1. The predicted odds ratio (OR) values ranged from 6.1 to infinity (Fig. 5a). Further, each MOB type possessed one to eight plasmid replicon genes, indicating MOB typing is of higher resolution than replicon typing (Table S8). Five MOB types AE928, AA860, AA738, AA334 and AA327 were significantly associated with MDR genes (P<0.001, OR 15.0 to infinity) (Fig. 5b). Moreover, the MDR associated replicon types and MOB types were mainly observed in L4, L5 and L8, which had a high proportion of MDR isolates, which was consistent with the inference that the MDR genes were carried by these plasmid types.

Lastly, we also assessed whether any virulence genes were carried by plasmids. Among the 34 intact plasmids, 27 harboured virulence genes. Two plasmids from bat source isolates harboured the Type II secretion system and the putative heat-stable enterotoxin gene *astA* [57] (Table S6). Moreover, some lineage-restricted virulence genes were observed in the MOB-suite predicted plasmids, including the *LngA-lngX* gene cluster, the *iucA-iucD* gene cluster, and the *hlyABCD* gene cluster.

Prophages and carriage of resistance and virulence genes

PHASTER was used to search for prophages in the 17 complete genomes [41]. A total of 207 prophages were identified: 130 were intact, 50 were incomplete and 27 were indeterminant (Table S9). The size of the intact prophage genomes ranged from 11.163 to 98.311 Kb. Most of the intact prophages were integrated on the chromosomes with 11 (8.5%) being on plasmids.

We grouped the 130 intact prophages based on a tree generated using the presence/absence of prophage genes using Roary v3.11.2 [35], and a nucleotide dotplot generated using Gepard v1.3 [42]. Gepard was a useful method for grouping diverse prophages [58]. As seen in Fig. 6, the darker the colour in the dot plots, the more similar the sequences were. There were five main squares with dense dots corresponding to five main groups of prophages (G1-G5). G5 was more diverse and potentially can be further subdivided into subgroups. Of prophages in G1 and G2, 50% (4/8) and 85.7% (6/7) (respectively) were from the two bat source isolates.

Based on the annotation of the 130 intact prophages, genes that were present only in one prophage group were identified using Scoary [36], and were designated as group-specific gene markers for each of the prophage groups. By screening the group-specific genes among the draft genomes, G1 was predicted to be present in 34.4% (166/482) of the *E. albertii* isolates, with at least two specific genes of G1 identified in these genomes. G2 was predicted to be in 3.7%, G3 in 46.7%, G4 in 59.1% and G5 in 96.1% of the 482 *E. albertii* isolates (Fig. S5). In terms of lineage distribution, G3 prophage specific genes were more likely to be observed in L5 and L8, and G4 prophages in L3, L4 and L8 ($P<0.001$, OR value >3.9). G1 prophage specific genes were negatively associated with L3 and L6, G3 prophages with L2, L3, L6 and L7, and G4 prophages with L2, L5, L6 and L7 ($P<0.001$, OR value <0).

There were 27 T3SS non-LEE effector genes present in 59 of the 130 intact prophages, 64.7% of which were in G5 prophages (Table S9). Two intact G5 prophages were positive for both stx_{2f} and *cdtABC* genes. Additionally, there were three intact prophages harbouring AR genes and all three prophages were located on plasmids.

The MVP database collected viral genomes and prophage sequences from bacterial and archaeal genomes [44]. Those virus and prophage genomes were clustered based on their sequence similarity, with unified cluster types assigned [44]. By nucleotide comparison with the MVP representative phage clusters database using BLASTN, only 13.1% (17/130) of the intact prophage sequences were previously recorded in the MVP database, belonging to 15 phage cluster types (Fig. 6a), indicating high diversity of prophages in *E. albertii* which have not been recorded in the database. Interspecies transmissions of prophages were observed: among the 15 MVP phage clusters, 11 prophages were previously observed in *E. coli*; cluster 12645 was previously observed in both *E. coli* and *Salmonella enterica*; and cluster 17047 from *Salmonella enterica*, while five phage clusters were only observed in *E. albertii.* In the five groups of prophages, MVP phage clusters were observed in G1, G3, G4 and G5, indicating G2 is a new prophage group specific for *E. albertii.*

DISCUSSION

E. albertii is a recently defined species of *Escherichia*, with infections previously wrongly attributed to *E. coli* and *Shigella* owing to the lack of sufficient subtyping techniques [1, 2, 21]. The *eae* gene and the *cdtB* gene have since been used for *E. albertii* identification [10, 24, 59]. However, both genes were not present in all *E. albertii* isolates or unique to *E. albertii*. In this study, only *eae* positive samples were cultured for *E. albertii*, which would have missed any potential *eae* negative *E. albertii* isolates. However, this bias is also reflected in the global collection of *E. albertii* isolates. Among the 312 publicly available *E. albertii* genomes, only three (1.0%) were *eae* negative. *E. albertii* as a pathogen is defined by its attaching and effacing pathogenicity [7, 9], and thus it is unsurprising that nearly all *E. albertii* isolates carried the virulence determinant *eae*. Since there were no markers to identify *eae* negative *E. albertii*, it is possible that the *E. albertii* population diversity is much larger, and the studies up to date including this study only assessed the population structure of the attaching and effacing *E. albertii*.

Isolation and identification of *E. albertii* have been hampered by the difficulties of differentiating *E. albertii* from other *Escherichia* species biochemically. It is now possible to use genetic markers to identify *E. albertii*. Lindsey *et al.* reported that the gene EAKF1_ch4033 is specific to *E. albertii* [24]. However, our analysis found that four of the 482 isolates (0.83%) were negative for the marker as false negatives and some *E. coli* genomes (e.g. ST378) harboured fragments of this gene and thus were potential false positives. We found three specific gene markers that were present in all *E. albertii* genomes analysed in this study and absent in all 30 021 representative *E. coli* and *Shigella* genomes screened. By BLASTN in the NCBI web database, all markers are specific to *E. albertii*, except for EAKF1_ch3804 being found in one *E. coli* complete genome (ST9286). Therefore, there are now four markers available for the identification of *E. albertii* with two markers (EAKF1_ch4075c and EAKF1_ch0408c) offering 100% sensitivity and specificity, which will facilitate future studies of the population diversity of *E. albertii*.

E. albertii is phylogenetically diverse

A previous study defined two clades of *E. albertii* [21], which is supported by this study. Further, we defined eight lineages. The previously defined clade 1 corresponds to L1, and clade 2 was further divided into seven lineages (L2 to L8). Isolates causing human infection were observed in all eight lineages, indicating that all lineages are potentially pathogenic to humans. Ooka *et al.* defined five groups (G1 to G5) based on 34 genomes, of which 32 genomes were included in this study [13]. All their G1 isolates belonged to L1 or clade 1, and all G2 to G5 isolates to clade 2 (Table S1). G2 isolates were divided into L7 and L8 while all G4 isolates belonged to L2. The majority of the G3 isolates belonged to none of the lineages, except for two isolates to L4 and one isolate to L5. There is only one isolate in G5, which belonged to none of the lineages in clade 2.

Our lineages have also been mapped to STs and CCs by the seven-gene MLST [31], for example, ST4638 and ST5390 belonged to L5 and L6, respectively. STs or CCs can be used as hallmarks for different lineages of *E. albertii* when genomic information is not available, which would facilitate comparison between different studies and surveillance of global spread and MDR by MLST. Although the isolates sequenced may not be representative, lineages were of significantly different proportions in different geographic regions: L5 (represented by ST4638) and L8 (represented by four STs) were more common in China, and L3 and L6 were only observed in Europe and North America.

Hyma *et al.* found that some *S. boydii* serotype 13 isolates and one *S. boydii* serotype 7 isolate, K-1, belonged to *E. albertii* [60]. In this study, there were 20 *E. albertii* isolates with *S. boydii* 13 serotype according to ShigEiFinder [33]. All belonged to L3, with one to L1. The genome of MLST sequences of the *S. boydii* serotype 7 isolate K-1 is unavailable. However, based on its *eae* (Accession No. AY696839) and *cdtB* (Accession No. AY696753) gene sequences, the *eae* was subtype tau and the *cdtB* type was *cdtB*-VI. The combination of *eae* subtype tau and *cdtB*-VI was found in eight isolates belonging to a branch phylogenetically closer to L3, but in none of the seven lineages in clade 2. Thus, *S. boydii* serotype 7 isolate K-1 does not belong to any lineages but may be phylogenetically closer to L3. Overall, our study showed that the diversity of *E. albertii* is high and new lineages are likely to be identified with more isolates sequenced.

Virulence gene variation in different lineages of *E. albertii*

The T3SS and the Cdt are the main virulence factors present in the vast majority of the *E. albertii* isolates. However, the subtypes of *eae* and *cdtB* were phylogenetically diverse. The *eae* gene was more diverse than the *cdtB* gene, and different lineages were dominated by different *eae* subtypes. Thus, it is likely that multiple independent acquisitions of the *eae* subtypes have occurred in *E. albertii*. There were seven new *eae* subtypes identified, and these *eae* subtypes were phylogenetically distant from each other, indicating potential independent acquisition. It is also possible that these new *eae* subtypes evolved within *E. albertii*. For the *cdtB* gene, *cdtB*-II was dominant and present in all lineages except L3, L4 and L5 whereas the newly defined *cdtB*-VI was found in L3, L4 and L5. Given the phylogenetic relationship of the lineages, *cdtB*-VI must have replaced *cdtB*-II in L3-L5. However, it is unclear if the *cdtB*-VI evolved within *E. albertii* or was acquired from other species. Moreover, some subtypes of *eae* and *cdtB* were prevalent in *E. coli* but were rare in *E. albertii* and vice versa. For example, *cdtB*-III and *cdtB*-V were common in Shiga toxin-producing *E. coli*, but were not observed in *E. albertii* [49, 61]; the *E. coli* prevalent *eae* subtypes were not common in *E. albertii* [62]; and the *eae* iota2 was observed in *S. boydii* serovar 13 isolates, which are in fact *E. albertii* [60]. The *eae* and *cdt* genes seemed to have been acquired by *E. albertii* multiple times during its long evolutionary history. More studies are required to elucidate the intra- and inter-species transfer of *eae* and *cdt* genes in the genus *Escherichia*.

Some virulence genes and pathogenicity islands were found to be associated with certain lineages. ETT2, which contributes to motility and serum resistance (which is essential for invasive infections) in *E. coli* [15], was truncated in L3 and L6, while in the other lineages only the *yqeF* gene of ETT2 was absent. Experimental evaluation is required to determine whether ETT2 is functional without the *yqeF* gene in *E. albertii*. *Yersinia* HPI encodes the siderophore yersiniabactin (Ybt) for iron scavenging, which causes oxidative stress in host cells and contributes to invasive extra-intestinal infections [52]. HPI comprises 11 genes, all of which were only observed in L6 isolates of *E. albertii*. Moreover, the *iuc* gene cluster including *iucABCD* encoding siderophore aerobactin and *iutA* encoding ferric aerobactin receptor for iron acquisition [51, 52] was mainly present in L3, L4 and one isolate of L6. More studies are required to evaluate the pathogenicity of those lineages that were equipped with different iron uptake systems. There were other lineage-restricted virulence genes like T6SS, *hlyABCD* and the *lng* gene cluster. Although their expression remains unknown, these lineage-restricted virulence factors may result in variation of the pathogenicity and environmental survival in different lineages [15, 55, 63].

Plasmid-mediated acquisition of virulence genes was observed in *E. albertii*. The lineage-restricted *hlyABCD* genes, the *iuc* gene cluster and the *lng* gene cluster were observed in MOB-suite predicted plasmids, indicating plasmid-mediated acquisition, which was supported by previous studies in *E. coli* [51, 55, 63]. The two *E. albertii* isolates from bats harboured a plasmid with T2SS genes and the metalloprotease encoding *stcE* gene. T2SS genes are critical for the survival and pathogenicity of bacteria [64]. And *stcE* gene, which is located on pO157 plasmid, contributes to the intimate adherence of EHEC and atypical *S. boydii* 13 [65, 66]. Like plasmids, prophages were also found to have contributed to the acquisition of virulence genes in *E. albertii*. The non-LEE effector genes of the T3SS were observed in intact prophages, which were found to be significantly associated with G5 prophages defined in this study. A previous report that lambdoid prophages carried various T3SS secretion effectors supports this finding [14]. Altogether, plasmids and prophages play key roles in the transfer of virulence genes in *E. albertii* and may facilitate large changes in pathogenicity like those seen in the pathovars of *E. coli* [19].

Plasmid-mediated AR genes were associated with STs and geographic regions

The predicted MDR rate in Chinese *E. albertii* isolates is astonishingly high (85.9%, 146/170), with 35.9% highly resistant isolates. These results are supported by previous phenotypic results, which found isolates resistant to up to 14 clinically relevant drugs and 11 drug classes [17]. Importantly, resistance was observed in clinically relevant drug classes including sulfamethoxazole-trimethoprim, cephalosporin, streptomycin and beta-lactam antibiotics [67]. However, it should be noted that the MDR *E. albertii* isolates were mainly obtained from poultry in China, and more clinical isolates are required to evaluate their clinical significance. It is likely that poultry source MDR *E. albertii* passes down the food production chain to humans, posing a threat to human health. There is an urgent need for surveillance and control of the spread of MDR and using MLST, we identified some STs that were associated with MDR *E. albertii* in China. ST4638, ST4479, ST4633 and ST4488 carried proportionally more MDR isolates and were mainly from China, which should facilitate the surveillance of the MDR. MDR in North America and Europe is emerging and the MDR associated STs from these continents were different from those of China. This may be due to the different control strategies on the use of antibiotics in different countries. In this study, we identified plasmid types that were significantly associated with MDR using both PlasmidFinder and MOB-suite, suggesting that the drug resistance genes were carried by plasmids. As the genomes were mostly draft genomes with incomplete plasmid sequences, further studies are required to understand the structure of these plasmids and the carriage of the resistance genes. Moreover, most of the L6 isolates harboured AR/MDR genes without predicted plasmids observed, which indicates potential new plasmids or prophages, or other means of MDR acquisition in L6.

CONCLUSION

In this study, the population structure of *E. albertii* was elucidated based on 170 genomes from China and 382 genomes from other countries. There were eight lineages identified, seven of which (L2-L8) belonged to previously defined clade 2. Isolates causing human infection were found in all lineages suggesting that most *E. albertii* has some pathogenicity. However, the uneven distribution of many virulence factors suggests that the degree of pathogenicity may differ across the lineages. The predicted MDR rate and MDR gene profiles varied between regions, STs and CCs, with Chinese isolates and STs being predominantly MDR. Plasmid replicon and MOB types that were related to the acquisition of MDR genes were identified in Chinese isolates. *E. albertii* contained a large number of prophages which were divided into five groups, with G5 prophages found to have contributed to the acquisition of the T3SS non-LEE effector genes. Therefore, prophages and plasmids played key roles in creating the virulence and MDR repertoires of *E. albertii*. Our findings provided fundamental insights into the population structure, virulence variation and MDR of *E. albertii*. Moreover, three new *E. albertii* specific gene markers were identified to facilitate the identification of this emerging foodborne pathogen.

Funding information
This work was supported by the National Key Research and Development Program of China (2017YFC1601502), State Key Laboratory of Infectious Disease Prevention and Control, China CDC (2016SKLID309), and an Australian Research Council Discovery Grant (DP170101917). Lijuan Luo was supported by a UNSW scholarship (University International Postgraduate Award).

Acknowledgements
Zigong CDC and L.L. acknowledge and commemorate Mr Yimao Miao for the guidance in all research studies in Zigong CDC, China. The authors thank Duncan Smith and Robin Heron from UNSW Research Technology Services for technical assistance.

Author contributions

L.L., M.P., R.L., Y.X. H.W. and Q.L. designed the study. H.W., L.Z., L.B., G.Y., Z.Z, Z.W., Y.X. and Q.L. collected the isolates and sequenced genomes. L.L., C.L. and H.W. curated the data. L.L, R.L, M.P, C.L. and X.Z. analysed the results, R.L., M.P., Y.X. and L.B. provided critical analysis and discussions. L.L. wrote the first draft. M.P. and R.L revised the drafts. All authors approved the final manuscript.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

1. Gomes TAT, Ooka T, Hernandes RT, Yamamoto D, Hayashi T, *et al*. *Escherichia albertii* Pathogenesis. *EcoSal Plus* 2020;9:1–18.
2. Huys G, Cnockaert M, Janda JM, Swings J. *Escherichia albertii* sp. nov., a diarrhoeagenic species isolated from stool specimens of Bangladeshi children. *Int J Syst Evol Microbiol* 2003;53:807–810.
3. Albert MJ, Alam K, Islam M, Montanaro J, Rahaman AS, *et al*. *Hafnia alvei*, a probable cause of diarrhea in humans. *Infect Immun* 1991;59:1507–1513.
4. Ooka T, Seto K, Kawano K, Kobayashi H, Etoh Y, *et al*. Clinical significance of *Escherichia albertii*. *Emerg Infect Dis* 2012;18:488–492.
5. Inglis TJJ, Merritt AJ, Bzdyl N, Lansley S, Urosevic MN. First bacteraemic human infection with *Escherichia albertii*. *New Microbes New Infect* 2015;8:171–173.
6. Janda JM, Abbott SL, Albert MJ. Prototypal diarrheagenic strains of *Hafnia alvei* are actually members of the genus *Escherichia*. *J Clin Microbiol* 1999;37:2399–2401.
7. Ooka T, Tokuoka E, Furukawa M, Nagamura T, Ogura Y, *et al*. Human gastroenteritis outbreak associated with *Escherichia albertii*, Japan. *Emerg Infect Dis* 2013;19:144–146.
8. Masuda K, Ooka T, Akita H, Hiratsuka T, Takao S, *et al*. Epidemiological aspects of *Escherichia albertii* outbreaks in Japan and genetic characteristics of the causative pathogen. *Foodborne Pathog Dis* 2020;17:144–150.
9. Oaks JL, Besser TE, Walk ST, Gordon DM, Beckmen KB, *et al*. *Escherichia albertii* in wild and domestic birds. *Emerg Infect Dis* 2010;16:638–646.
10. Wang H, Li Q, Bai X, Xu Y, Zhao A, *et al*. Prevalence of eae-positive, lactose non-fermenting *Escherichia albertii* from retail raw meat in China. *Epidemiol Infect* 2016;144:45–52.
11. Asoshima N, Matsuda M, Shigemura K, Honda M, Yoshida H, *et al*. Isolation of *Escherichia albertii* from Raw Chicken Liver in Fukuoka City, Japan. *Jpn J Infect Dis* 2015;68:248–250.

12. Grillová L, Sedláček I, Páchníková G, Staňková E, Švec P, *et al*. Characterization of four *Escherichia albertii* isolates collected from animals living in Antarctica and Patagonia. *J Vet Med Sci* 2018;80:138–146.
13. Ooka T, Ogura Y, Katsura K, Seto K, Kobayashi H, *et al*. Defining the genome features of *Escherichia albertii*, an emerging enteropathogen closely related to *Escherichia coli*. *Genome Biol Evol* 2015;7:3170–3179.
14. Tobe T, Beatson SA, Taniguchi H, Abe H, Bailey CM, *et al*. An extensive repertoire of type III secretion effectors in *Escherichia coli* O157 and the role of lambdoid phages in their dissemination. *Proc Natl Acad Sci U S A* 2006;103:14941–14946.
15. Shulman A, Yair Y, Biran D, Sura T, Otto A, *et al*. The *Escherichia coli* Type III secretion system 2 has a global effect on cell surface. *mBio* 2018;9:e01070-01018.
16. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, *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* 2019;63:e00483-00419.
17. Li Q, Wang H, Xu Y, Bai X, Wang J, *et al*. Multidrug-Resistant *Escherichia albertii*: Co-occurrence of β-Lactamase and MCR-1 Encoding Genes. *Front Microbiol* 2018;9:258.
18. Lima MP, Yamamoto D, Santos AC de M, Ooka T, Hernandes RT, *et al*. Phenotypic characterization and virulence-related properties of *Escherichia albertii* strains isolated from children with diarrhea in Brazil. *Pathog Dis* 2019;77:1–13.
19. Nakamura K, Murase K, Sato MP, Toyoda A, Itoh T, *et al*. Differential dynamics and impacts of prophages and plasmids on the pangenome and virulence factor repertoires of Shiga toxin-producing *Escherichia coli* O145:H28. *Microb Genom* 2020;6:1–13.
20. Lindsey RL, Rowe LA, Batra D, Smith P, Strockbine NA. PacBio genome sequences of eight *Escherichia albertii* strains isolated from humans in the United States. *Microbiol Resour Announc* 2019;8:1–3.
21. Ooka T, Seto K, Ogura Y, Nakamura K, Iguchi A, *et al*. O-antigen biosynthesis gene clusters of *Escherichia albertii*: their diversity and similarity to *Escherichia coli* gene clusters and the development of an O-genotyping method. *Microb Genom* 2019;5:e000314:11.:.
22. Shen R, Fan J-B, Campbell D, Chang W, Chen J, *et al*. High-throughput SNP genotyping on universal bead arrays. *Mutat Res* 2005;573:70–82.
23. Eid J, Fehr A, Gray J, Luong K, Lyle J, *et al*. Real-time DNA sequencing from single polymerase molecules. *Science* 2009;323:133–138.
24. Lindsey RL, Garcia-Toledo L, Fasulo D, Gladney LM, Strockbine N. Multiplex polymerase chain reaction for identification of *Escherichia coli, Escherichia albertii* and *Escherichia fergusonii*. *J Microbiol Methods* 2017;140:1–4.
25. Souvorov A, Agarwala R, Lipman DJ. SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biol* 2018;19:153.
26. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, *et al*. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015;43:e15.
27. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* 2018;35:1547–1549.
28. Hu D, Liu B, Wang L, Reeves PR. Living trees: high-quality reproducible and reusable construction of bacterial phylogenetic trees. *Mol Biol Evol* 2020;37:563–575.
29. Tonkin-Hill G, Lees JA, Bentley SD, Frost SDW, Corander J. Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Res* 2019;47:5539–5549.
30. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 2019;47:W256–W259.
31. Wirth T, Falush D, Lan R, Colles F, Mensa P, *et al*. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol* 2006;60:1136–1151.
32. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, *et al*. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* 2014;58:3895–3903.
33. Zhang X, Payne M, Nguyen T, Kaur S, Lan R. Cluster-specific gene markers enhance shigella and enteroinvasive escherichia coli in silico serotyping. *Microb Genom* 2021.
34. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
35. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, *et al*. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;31:3691–3693.
36. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol* 2016;17:238.
37. Chen L, Zheng D, Liu B, Yang J, Jin Q. VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. *Nucleic Acids Res* 2016;44:D694-7.
38. Ito K, Iida M, Yamazaki M, Moriya K, Moroishi S, *et al*. Intimin types determined by heteroduplex mobility assay of intimin gene (eae)-positive Escherichia coli strains. *J Clin Microbiol* 2007;45:1038–1041.
39. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;29:1072–1075.
40. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom* 2018;4.
41. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, *et al*. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 2016;44:W16-21.
42. Krumsiek J, Arnold R, Rattei T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* 2007;23:1026–1028.
43. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, *et al*. Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics* 2018;34:292–293.
44. Gao NL, Zhang C, Zhang Z, Hu S, Lercher MJ, *et al*. MVP: a microbe-phage interaction database. *Nucleic Acids Res* 2018;46:D700–D707.
45. Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J Bacteriol* 2004;186:1518–1530.
46. Foster-Nyarko E, Alikhan N-F, Ravi A, Thomson NM, Jarju S, *et al*. Genomic diversity of Escherichia coli isolates from backyard chickens and guinea fowl in the Gambia. *Microb Genom* 2021;7.
47. Gaytán MO, Martínez-Santos VI, Soto E, González-Pedrajo B. Type Three Secretion System in Attaching and Effacing Pathogens. *Front Cell Infect Microbiol* 2016;6:129.
48. Pickett CL, Whitehouse CA. The cytolethal distending toxin family. *Trends Microbiol* 1999;7:292–297.
49. Hinenoya A, Yasuda N, Mukaizawa N, Sheikh S, Niwa Y, *et al*. Association of cytolethal distending toxin-II gene-positive *Escherichia coli* with *Escherichia albertii*, an emerging enteropathogen. *Int J Med Microbiol* 2017;307:564–571.
50. Tóth I, Nougayrède J-P, Dobrindt U, Ledger TN, Boury M, *et al*. Cytolethal distending toxin type I and type IV genes are framed with lambdoid prophage genes in extraintestinal pathogenic *Escherichia coli*. *Infect Immun* 2009;77:492–500.
51. Ling J, Pan H, Gao Q, Xiong L, Zhou Y, *et al*. Aerobactin synthesis genes *iucA* and *iucC* contribute to the pathogenicity of avian pathogenic *Escherichia coli* O2 strain E058. *PLoS One* 2013;8:e57794.
52. Galardini M, Clermont O, Baron A, Busby B, Dion S, *et al*. Major role of iron uptake systems in the intrinsic extra-intestinal virulence

of the genus *Escherichia* revealed by a genome-wide association study. *PLoS Genet* 2020;16:e1009065.

53. Girón JA, Gómez-Duarte OG, Jarvis KG, Kaper JB. Longus pilus of enterotoxigenic *Escherichia coli* and its relatedness to other type-4 pili--a minireview. *Gene* 1997;192:39–43.

54. Gomez-Duarte OG, Chattopadhyay S, Weissman SJ, Giron JA, Kaper JB, *et al*. Genetic diversity of the gene cluster encoding longus, a type IV pilus of enterotoxigenic *Escherichia coli*. *J Bacteriol* 2007;189:9145–9149.

55. Saldaña-Ahuactzi Z, Rodea GE, Cruz-Córdova A, Rodríguez-Ramírez V, Espinosa-Mazariego K, *et al*. Effects of lng Mutations on LngA Expression, Processing, and CS21 Assembly in Enterotoxigenic *Escherichia coli* E9034A. *Front Microbiol* 2016;7:1201.

56. Octavia S, Lan R. Shigella and Shigellosis. In: Tang Y-W, Sussman M, Liu D, Poxton I and Schwartzman J (eds). *Molecular Medical Microbiology*. Boston: Academic Press; 2015. pp. 1147–1168.

57. Savarino SJ, Fasano A, Watson J, Martin BM, Levine MM, *et al*. Enteroaggregative *Escherichia coli* heat-stable enterotoxin 1 represents another subfamily of *E. coli* heat-stable toxin. *Proc Natl Acad Sci U S A* 1993;90:3093–3097.

58. Bleriot I, Trastoy R, Blasco L, Fernández-Cuenca F, Ambroa A, *et al*. Genomic analysis of 40 prophages located in the genomes of 16 carbapenemase-producing clinical strains of *Klebsiella pneumoniae*. *Microb Genom* 2020;6:1–18.

59. Hinenoya A, Ichimura H, Yasuda N, Harada S, Yamada K, *et al*. Development of a specific cytolethal distending toxin (cdt) gene (Eacdt)-based PCR assay for the detection of *Escherichia albertii*. *Diagn Microbiol Infect Dis* 2019;95:119–124.

60. Hyma KE, Lacher DW, Nelson AM, Bumbaugh AC, Janda JM, *et al*. Evolutionary genetics of a new pathogenic *Escherichia* species: *Escherichia albertii* and related *Shigella boydii* strains. *J Bacteriol* 2005;187:619–628.

61. Hinenoya A, Shima K, Asakura M, Nishimura K, Tsukamoto T, *et al*. Molecular characterization of cytolethal distending toxin gene-positive *Escherichia coli* from healthy cattle and swine in Nara, Japan. *BMC Microbiol* 2014;14:97.

62. Yang K, Pagaling E, Yan T. Estimating the prevalence of potential enteropathogenic *Escherichia coli* and intimin gene diversity in a human community by monitoring sanitary sewage. *Appl Environ Microbiol* 2014;80:119–127.

63. Schwidder M, Heinisch L, Schmidt H. Genetics, toxicity, and distribution of enterohemorrhagic *Escherichia coli* Hemolysin. *Toxins* 2019;11:502.

64. Patrick M, Gray MD, Sandkvist M, Johnson TL. Type II Secretion in *Escherichia coli*. *EcoSal Plus* 2010;4.

65. Grys TE, Siegel MB, Lathem WW, Welch RA. The StcE protease contributes to intimate adherence of enterohemorrhagic *Escherichia coli* O157:H7 to host cells. *Infect Immun* 2005;73:1295–1303.

66. Walters LL, Raterman EL, Grys TE, Welch RA. Atypical *Shigella boydii* 13 encodes virulence factors seen in attaching and effacing *Escherichia coli*. *FEMS Microbiol Lett* 2012;328:20–25.

67. Eyler RF, Shvets K. Clinical pharmacology of antibiotics. *Clin J Am Soc Nephrol* 2019;14:1080–1090.