Whole-genome analysis of extraintestinal *Escherichia coli* sequence type 73 from a single hospital over a 2 year period identified different circulating clonal groups


**Abstract**

Sequence type (ST)73 has emerged as one of the most frequently isolated extraintestinal pathogenic *Escherichia coli*. To examine the localized diversity of ST73 clonal groups, including their mobile genetic element profile, we sequenced the genomes of 16 multiple-drug resistant ST73 isolates from patients with urinary tract infection from a single hospital in Sydney, Australia, between 2009 and 2011. Genome sequences were used to generate a SNP-based phylogenetic tree to determine the relationship of these isolates in a global context with ST73 sequences (*n*=210) from public databases. There was no evidence of a dominant outbreak strain of ST73 in patients from this hospital, rather we identified at least eight separate groups, several of which reoccurred, over a 2 year period. The inferred phylogeny of all ST73 strains (*n*=226) including the ST73 clone D12 reference genome shows high bootstrap support and clusters into four major groups that correlate with serotype. The Sydney ST73 strains carry a wide variety of virulence-associated genes, but the presence of *iss*, *pic* and several iron-acquisition operons was notable.

**DATA SUMMARY**

1. All sequencing reads and assemblies for isolates sequenced in this study have been submitted to the ENA Sequence Read Archive (SRA) and GenBank, respectively. GenBank, SRA accession numbers and URLs are included in Table S2, available in the online version of this article. which has been deposited in Figshare; DOI: 10.6084/m9.figshare.5477461 (URL – https://doi.org/10.6084/m9.figshare.5477461).

2. Scripts used for the analysis of SNP phylogeny have been deposited in Github; (URL – https://github.com/bogemad/snp_phylogeny).

3. Fig. S1 has been deposited in Figshare; DOI: 10.6084/m9.figshare.5477449 (URL – https://doi.org/10.6084/m9.figshare.5477449).

4. Fig. S2 has been deposited in Figshare; DOI: 10.6084/m9.figshare.5477485 (URL – https://doi.org/10.6084/m9.figshare.5477485).

5. Table S1 has been deposited in Figshare; DOI: 10.6084/m9.figshare.5477464 (URL – https://doi.org/10.6084/m9.figshare.5477464).

6. Table S3 has been deposited in Figshare; DOI: 10.6084/m9.figshare.5477485 (URL – https://doi.org/10.6084/m9.figshare.5477485).

7. Table S4 has been deposited in Figshare; DOI: 10.6084/m9.figshare.5477473 (URL – https://doi.org/10.6084/m9.figshare.5477473).

8. Table S5 has been deposited in Figshare; DOI: 10.6084/m9.figshare.5477479 (URL – https://doi.org/10.6084/m9.figshare.5477479).
INTRODUCTION

Extraintestinal pathogenic *Escherichia coli* (ExPEC) are phylogenetically diverse and comprise uropathogenic *E. coli* (UPEC), neonatal meningitis-causing *E. coli* (NMEC) and avian pathogenic *E. coli* (APEC). ExPEC account for ~75–95% of urinary tract infections. A proportion of these infections can spread from the urinary tract with invasion of epithelial cells in the bladder (cystitis) and kidney cells (pyelonephritis) and transmission to systemic circulation (blood sepsis), posing a serious threat to human health. ExPEC are enteric bacteria, but their capacity to capture a wide array of virulence-associated genes (VAGs) by lateral gene transfer has expanded the repertoire of niches they colonize. ExPEC may carry diverse and often redundant combinations of VAGs whose impact on human health remains ill-defined. Epidemiological studies indicate that a subset of pathogenic *E. coli* lineages, including sequence type (ST)73, ST131, ST405, ST393, ST69, ST95, ST10, ST38 and ST127 [1–5], are responsible for most ExPEC infections [3, 6–8]. Carriage of combinations of virulence genes enhances virulence [9]; however, carriage of antimicrobial-resistance genes, particularly those encoding extended-spectrum β-lactamases and fluoroquinolones, as well as an ability to cause opportunistic infections in vulnerable (elderly) hosts, may also contribute to virulence. It is notable that none of these hypotheses have been experimentally validated [3, 6].

ExPEC have become the leading cause of blood sepsis in Europe [10]. Notable in this regard is the alarming rise in the incidence of ST73, now one of the most frequently isolated UPEC globally and the leading cause of bacteraemia in the East Midlands region of the UK [1, 11–14]. ST73 belongs to Clermont phylogroup B2 and is known to display different serogroups, with serotype O6 predominating (ST73-O6-B2) [15, 16]. It has recently been suggested that the rise in the incidence of multiple-drug resistant (MDR) ST73 in the UK is not due to the emergence of a dominant clone, because they are genetically diverse and carry a different array of plasmids encoding resistance to multiple antimicrobials. Many recently described isolates of ST73 carry genes that encode extended-spectrum β-lactamases and resistance to antimicrobials used in veterinary medicine [14]. This seems to be a recent adaptation in this ST, as previously characterized ST73 isolates from cases of uncomplicated urinary tract infection sourced from Greece, Portugal, Sweden and the UK were susceptible to most clinically relevant antimicrobials and most (75%) did not carry plasmids, classic vehicles of multiple-drug resistance [17]. These data combined with the most recent findings seem to suggest that the rise in the carriage of MDR plasmids in ST73 may be a recent concerning event [14, 18].

Here, we have characterized whole-genome sequences of 15 class 1 integrase (*intI1*)-containing ST73 strains from a hospital in Sydney. To determine whether these highly localized strains were from a limited number of clonal lineages, phylogenetic inferences were made by comparing SNP differences in core genomes shared by ST73 strains from our Sydney collection with those from six high-quality reference genome sequences and ST73 strains (n=204) from seven countries sourced from global sequence read archives. We also examined mobile and chromosomal genetic content within this localized isolate cohort to further examine their accessory genome diversity. We compiled the repertoire of antimicrobial genes and virulence genes, and mapped the class 1 integron structures carried by these isolates. As carriage of the class 1 integrase is considered a reliable proxy for multiple-drug resistance [19], we used *S1* nuclease-PFGE (*S1*-PFGE), followed by Southern hybridization with an *intI1* probe, to examine plasmid content and carriage of the class 1 integrase on plasmids.

**METHODS**

Isolate source and culture conditions

Clinical samples in this project were from a larger collection obtained from the Sydney Adventist Hospital from 2009 to 2011. Bacterial species were identified by the VITEK 2 (bioMérieux) system at the Sydney Adventist Hospital. For DNA extraction, strains were first grown on a lysogeny broth (LB) agar plate to isolate single colonies, of which one was used to inoculate 2 ml LB, followed by shaking for 16 h at 37°C. Antibiotic-susceptibility testing for ampicillin, cefotaxime, chloramphenicol, streptomycin and sulfadiazole was performed via the Calibrated Dichotomous Sensitivity (CDS) method [20]. These antibiotics were
selected based on antibiotic-resistance gene content inferred from genome sequencing.

**Nucleic acid purification and whole-genome sequencing**

*E. coli* DNA was extracted using the Isolate II genomic DNA extraction kit (Bioline), according to the manufacturer’s instructions. For each sample, ‘fragmentation’ of genomic DNA and PCR amplification of tagged DNA were performed in triplicate using the Nextera system (Illumina). Sequencing libraries were pooled, then cleaned and size selected using SPRI beads (Beckman Coulter). Normalization was guided by read counts obtained from a Nano flow cell run on a MiSeq instrument. An Agilent 2100 Bioanalyzer, with a High Sensitivity DNA kit, was used to quantitate the pooled library before loading onto an Illumina HiSeq. Paired-end 150 bp reads were generated using the HiSeq 2500 v4 system.

**Genome assembly and gene presence**

Genome assembly was achieved with raw reads using the A5-MiSeq pipeline [21] and checked for consistency by additional assembly with SPAdes 3.9.0 [22]. Antibiotic-resistance genes and VAGs were identified from assembled genomes using BLASTN and SRST2 [23]. Searches were performed against antibiotic-resistance genes sourced from the ARG-ANNOT v3 database and a panel of VAGs identified from the Virulence Factors Database (VFDB) and literature searches [24, 25]. Serotyping was performed *in silico* with SRST2 using EcOH sequences supplied with this package. Draft genome reads obtained from the Sequence Read Archive (SRA) were searched using SRST2 for a minimal set of marker genes derived from integron structures and commonly associated transposons characterized in the Sydney strains. Low-quality alignments based on SRST2 output were discounted ($n=3$).

**Archived sequence read selection**

All additional ST73 sequences not generated by this study were obtained from complete whole-genome assemblies ($n=6$) [26–30] and public sequence read archives (NCBI/EMBL/DDBJ). Raw Illumina reads sourced from ST73 isolates ($n=284$) were considered for SNP-based phylogenetic analysis, including strains sequenced in this study from the Sydney Adventist Hospital ($n=16$), isolates with host, source and isolation location meta-data identified from the Enterobase database ($n=246$; http://enterobase.warwick.ac.uk/; accessed 5/12/2016) and a previous ST73-focused study from the UK ($n=22$) [11]. Samples were excluded if the ST could not be confirmed as ST73 using SRST2 ($n=4$). Further samples were excluded ($n=30$) if isolate status could not be confirmed by BioProject meta-data or where the description of methods could not be identified by an associated publication [31–38]. Samples were additionally excluded ($n=30$) if they produced low reference genome coverage (>90%) in whole-genome alignments. Additional sample filtering of ST73 reads is described in Table S1.

**S1-PFGE analysis**

The complement of large (>20 kb) plasmids in each bacterial isolate was determined by S1 nuclease (Promega) digestion and PFGE, as described previously [39, 40]. Southern blot hybridization was used to determine the genomic location of the *intI1* gene. PCR amplicons for *intI1* were obtained using published primers (int1F and int1R [41]) and labelled using a PCR DIG probe synthesis kit (Roche). DNA was transferred from the S1-PFGE gel to a nylon membrane (GE Health) using a VacuAid vacuum transfer apparatus (HybAid) and hybridization was performed using a DIG filter hybridization system (Roche), following the manufacturer’s instructions. Images were acquired on a ChemiDoc MP System (Bio-Rad Laboratories).

**SNP-based phylogenetic analyses**

Our initial attempts to examine ST73 phylogeny with our 16 genome sequences and 6 complete whole-genome sequences (CFT073, ABU83972, ATCC25922, Nissile 1917, clone D i2 and clone D i14) using marker gene approaches [42] provided limited resolving power (Fig. S1). Consequently, to more appropriately examine these ST73 strains, we employed SNP-based phylogenetic methods.

For SNP-based phylogenetic trees, core genome alignments were generated with Snippy v3.1.0 (https://github.com/tseemann/snippy) using default options. Briefly, reads were mapped using BWA MEM v0.7 to an ST73 reference genome. Raw alignments were processed with samtools v1.3.1 and variants called using freebayes v1.1.0. SNP-derived genomes were reconstructed using vcftools v0.1.14, with low-coverage (<10×) and degenerate reference positions filtered. Recombinant regions were removed using Gubbins v2.20 (option -i 10) [43], yielding aligned, SNP-derived, recombination-filtered core genomes. From this alignment, core genome phylogenetic trees were inferred by maximum-likelihood using RAxML v8.2.9 [44]. Branch support was estimated by bootstrap analysis employing 100 replicate trees. Trees were rooted using the ultrametric tree method included with RAxML. Scripts used for the analysis of SNP phylogeny can be found online (at https://github.com/bogema/snpy_phylgeny).

Initially, we performed this analysis using only isolates sequenced in this study and a high-quality published ST73 reference (Fig. 1). To identify the most suitable reference genome for this purpose, we aligned reads individually to the six complete reference genomes (above) and examined reference sequence quality, core alignment lengths and final tree support values. Using this methodology, the most suitable reference was identified as clone D i2, which with archived public sequence reads generated a core genome of 3 818 344 bp, representing 75.8% of the ST73 clone D i2 sequence.
RESULTS

Assembly information and statistics

The genome sequences of 16 ST73 strains from the Sydney Adventist Hospital were determined here. The whole-genome shotgun project has been deposited at GenBank/ENA/DDBJ and the SRA. Assembly statistics, as well as accession numbers, number of sequencing reads and the amount of sequencing data used to generate assemblies can be found in Table S2.

Public read high-throughput sequencing analysis

The Sydney isolates separated into several groups that closely aligned with E. coli serotype, including O22:H1, O25: H1 and O6:H1. However, most of the Sydney isolates clustered within the larger O6:H1 group. To further interrogate observed diversity within the O6:H1 group of isolates and to place Sydney strains within a broader global context, we expanded the SNP-based phylogenetic analysis to include additional ST73 sequence reads obtained from public sequence read archives (NCBI/EMBL/DDBJ) and the six complete ExPEC genomes with the 16 Sydney genomes. The inferred phylogeny of the 226 strains (Figs 2 and S2) for all strain labels and branch support values shows strong major branch support. Analysis of the SNP-derived phylogenetic tree shows correlation with observed serotypes O6: H1, O25:H1 and O22:H1, with most strains observed within the O6:H1 cluster. Strain 2011_82 could not be assigned an O-type from in silico serotyping, but was identified as H1 and clustered most closely with O6:H1 isolates (Fig. 2).

Examination of ST73 phylogenetic structure reveals four significant clades (Figs 2 and S2): one exclusively associated with serotype O6:H1 (group A), one exclusively associated with serotype O2:H1 (group B), one primarily serotype O6: H1 with a subclade of O18:H1 (group C), and finally a polyphyletic group consisting of O2:H1, O6:H1, O22:H1, O25: H1 and O120:H31 serotypes (group D). Strains sequenced from Sydney separated into eight distinct groups correlated with serotype (Fig. 2 – red labels). Isolates with O22:H1 and O25:H1 serotypes formed their own clusters, while serotype O6:H1 separated into five distinct clusters (O6-1: 2009_48/45; O6-2: 2009_6/2011_69; O6-3: 2009_44; O6-4: 2009_5/14/16/2010_64; O6-5: 2009_8). Strain 2011_82 clustered with other O6:H1 strains and formed the final group (Ox).

Virulence profiles of Sydney strains

We identified differences in virulence gene profiles of E. coli strains examined in this study (Fig. 1). Significantly, we found that virulence gene profiles were largely consistent in strains from the same phylogenetic groups O25, O22, O6-1 – O6-5, Ox (Fig. 1, see Tables S3 and S4 for
greater detail). For adhesion-related genes, critical components of P fimbriae (papACDEFGHK) were absent in six strains (Fig. 1). Additionally, in strains of group O25, genes encoding the type I fimbriae major subunit (fimA), periplasmic chaperone (fimC), regulatory subunit (fimE) and the fimbriae-associated fimI were missing in BLASTN searches. Other genes encoding F1C fimbriae, curli fibres, type IV pili, E. coli common pili and the fdeC adhesion genes were present in all strains. The importance of FdeC as a putative virulence factor is underpinned by the observation that it is: (i) a broadly conserved E. coli adhesin whose expression is upregulated on the surface of UPEC when it contacts host cells; and (ii) a major target during humoral immune responses that significantly reduced kidney colonization in mice challenged transurethrally with UPEC strain 536 [45].

Fig. 2. SNP-based maximum-likelihood phylogram of 226 ST73 strains. A more detailed tree with branch support values and tip labels can be found in Fig. S2. ST73 isolates separate into four distinct groups, labelled A–D, which correlate well with in silico serotyping (inset). ST73 isolates sequenced in this study cluster into eight distinct groups, shown in red, high-quality complete ST73 genomes are shown in blue. Trees were reconstructed using 18 426 SNPs identified by read mapping to the clone D i2 reference sequence, reduced from 27 568 SNPs by filtering of recombination regions.
Iron acquisition is critical for the growth of ExPEC in low-iron environments in vivo, and it is not uncommon to identify genes linked to siderophore production and processing in UPEC. Complete enterobactin, salmochelin and yersiniabactin gene clusters were identified in all ST73 strains, while aerobactin genes were identified in all strains except those belonging to group O22. Genes for haem uptake, including the chu operon and hma gene, were present in all strains, as were those related to iron uptake such as the sit ABC transporter operon and ferric I Yersinia uptake (fyuA) gene. In contrast, the putative iron-uptake gene cluster eitABCD and adhesion/iron-uptake gene treA were only identified in a subset of strains. In addition to iron uptake, genes encoding copper resistance have also been linked to virulence [46] and antimicrobial resistance [47]. The cus system, encoding a four-component copper efflux pump, was present and complete in all strains. However, in strain 2009_45, cueR, an important regulator controlling copper detoxification and efflux copA and cueO genes, was not located in all searches.

Larger differences were observed in the presence of toxin genes. Strains from group O25 and strain 2009_8 contained the highest number of toxin genes, including cytotoxic necrotizing factor 1 (cnf1), the haemolysin (hlyABCD) cluster, haemolysin E (hlyE) and secreted autotransporter toxin (sat). Genes that have been previously shown to promote propagation of E. coli in blood, such as proteases pic and tsh and the increased serum survival (iss) gene, were present in all strains, as well as the cellular invasion promoting ibe gene cluster. Closely related hek and tia genes, associated with epithelial cell invasion in neonatal meningitis-causing and enterotoxigenic E. coli, respectively, are both found in separate strains. Furthermore, tcpC associated with immune modulation via inhibition of Toll/IL-1 receptor signalling was only found in groups O6-1-5 and Ox.

### Antibiotic resistance

All intI1-positive isolates were tested for resistance to ampicillin, cefotaxime, chloramphenicol, streptomycin, sulfafurazole and trimethoprim using the CDS method. Strain 2011_82 did not have a class 1 integrase gene and was not tested. All strains were resistant to ampicillin, streptomycin and sulfafurazole (Table 1). Genes encoding resistance to these antibiotics were all accounted for in the genome sequence data by the class 1 integron-associated genes aadA1 and sul1, as well as one of three bla gene variants (Table S5). Only strain 2009_45 was resistant to the third-generation cephalosporin cefotaxime, likely due to the presence of the blaOXA1 gene. However, this resistance was not observed in strain 2009_38, which contained an almost identical antimicrobial-resistance region, suggesting this gene is not expressed in this strain. Interestingly, both of these strains also showed phenotypic resistance to chloramphenicol despite only 2009_45 containing a complete copy of the catA1 gene. The full repertoire of antibiotic-resistance genes found in the 16 Sydney ST73 strains is presented in Fig. 1.

### Structure of class 1 integrons in ST73 strains from Sydney

All locally sourced strains in this study, excepting 2011_82, were positive for a complete copy of the sulfonamide-resistance gene sul1, a structural marker of the 3’-conserved segment (3’-CS) of class 1 integrons. Similarly, all strains contained the aminoglycoside-resistance gene cassette aadA1. Strain 2011_82 was found to contain only a class 2 integron carrying the standard dfrA1-sat2-aadA1 cassette array, resulting in trimethoprim resistance as tested by the hospital upon initial isolation.

---

**Table 1. Phenotypic antimicrobial resistance**

<table>
<thead>
<tr>
<th>Patient</th>
<th>AMP</th>
<th>AMC</th>
<th>LEX</th>
<th>CIP</th>
<th>GEN</th>
<th>NIT</th>
<th>NOR</th>
<th>TMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009_3</td>
<td>R</td>
<td>R</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>2009_5</td>
<td>R</td>
<td>R</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>2009_6</td>
<td>R</td>
<td>R</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>–</td>
</tr>
<tr>
<td>2009_8</td>
<td>R</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>–</td>
</tr>
<tr>
<td>2009_13</td>
<td>R</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>2009_14/16</td>
<td>R</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>–</td>
</tr>
<tr>
<td>2009_21</td>
<td>R</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>–</td>
</tr>
<tr>
<td>2009_38</td>
<td>R</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>2009_44</td>
<td>R</td>
<td>–</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>2009_45</td>
<td>R</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>2010_64</td>
<td>R</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>–</td>
</tr>
<tr>
<td>2011_69</td>
<td>R</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>R</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>2011_82</td>
<td>S</td>
<td>–</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>R</td>
</tr>
<tr>
<td>2011_93/98</td>
<td>R</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>–</td>
</tr>
</tbody>
</table>

AMC, amoxicillin-clavulanic acid; AMP, ampicillin; CIP, ciprofloxacin; GEN, gentamicin; LEX, cephalixin; NIT, nitrofurantoin; NOR, norfloxacin; R, resistant; S, susceptible; TMP, trimethoprim.
There were three class 1 integron-containing resistance regions represented within our collection (Fig. 3), all containing the same base structures with minor variations. The first structure was identified in 11 out of 15 class 1 integron-containing isolates (Fig. 3a). It consisted of an In-2 type class 1 integron with and aadA1 gene cassette housed within an incomplete Tn21 transposon, matching (99% sequence identity) the sequence in the R100 plasmid identified in Japan in the 1950s from Shigella flexneri (accession no. NC_002134.1) [48]. However, our structure bears an IS26-mediated partial deletion of the Tn21 tnpR gene, which is a signature that has been reported previously twice within a UPEC strain from Australia, and in association with a different class 1 integron structure [49]. A Tn3 transposon has inserted within the mer module of Tn21 with partial deletion of merA and merT, and complete deletion of merC and merP. The transposon is abutted downstream of merR by an inward facing IS1 insertion element. One strain, 2009–64, housed this exact structure apart from the Tn21 tnpM, which appears to have been lost due to an IS26-mediated deletion event.

The complex resistance locus (CRL) shown in Fig. 3(b) was identified in isolates 2009–6 and 2011–69, and shares homology with the structure in Fig. 3(a). It bore identical IS26 and Tn3 insertion points, with the only major difference being a crossover event where the standard Tn3 tnpA gene and terminal inverted repeat have been replaced by that of Tn1000, a transposon originally identified in a cosmid clone of a human DNA sequence in 1995 [50]. This signature was recently identified in the sequence of an unannotated plasmid of a Salmonella enterica serovar Typhi strain sequenced as part of a larger study of Typhi from typhoid-endemic regions of Asia and Africa (accession no. LT904889.1). This is, therefore, the first report of this hybrid transposon and its presence in an E. coli isolated in Australia, to the best of our knowledge. Due to the nature of Illumina sequence technology, we have no confirmed sequence information downstream of Tn3/Tn1000.

Structure 3 (Fig. 3c) shares homology with the previously discussed structures. However, this CRL, present in strains 2009–38 and 2009–45, has a blaOXA-1 gene cassette within the integron cassette array in addition to aadA1. Here, the Tn21 transposon housing the class 1 integron is complete, with both the initial and terminal inverted repeats intact, and has an inward facing IS1 flanking its mer end. There are two variants of this CRL in our collection, one of which
contains a complete catA1 gene downstream of Tn21 followed by a second IS1 element in the same orientation as the first, with the intergenic ORF identified as an acetyltransferase. This appears to be an established insertion event, with numerous reports in GenBank. In the second variant, the terminal IS1 is inverted with consequent deletion of 476 bp of the catA1 gene, forming a signature unique to isolates 2009–38 and 2009–43. This integron has been reported in its entirety in Shigella dysenteriae 1 plasmid p3099-85 (KT754164.1), Salmonella enterica serovar Typhimurium plasmid pUO-StVR2 (AM991977.1) and Salmonella enterica serovar Typhimurium strain T000240 (AP011957.1). Less than 10 SNPs were identified in comparative BLASTN alignments spanning the integron.

Sixty genomes from the SRA cohort returned adequate alignments to integron marker genes, with 25 of these appearing to possibly have only the base class 1 integron with an aadA1 cassette, but no indication of a bordering Tn21 transposon. Eight contained an intI1 gene but no aadA1, suggesting the likely presence of a class 1 integron with a different cassette array. Sixteen contained aadA1, but no class 1 integrase; this could indicate a deletion event or more likely the presence of aadA1 in a class 2 integron, though the aadA1 gene can also exist independent of integron context. Five genomes contained an unidentifiable integron structure, possibly variants of those described in the Sydney collection, although it is impossible to say this definitively from read alignments against the abridged gene database used here.

Only three genomes, HVH_93_4–5851025, MOD1-EC6690 and MOD1-EC6783, contained all marker genes necessary to potentially contain integron C (Fig. 3). However, within the SRA cohort, the presence of integrons A and B could not be confirmed.

All 16 strains of ST73 that were sourced from Sydney were shown to carry one or more plasmids (up to five) that ranged in size from 15 to 180 kb. Only one plasmid in each strain hybridized with the intI1 probe (data not shown). The sizes varied greatly between 80 kb and >200 kb.

**DISCUSSION**

This study forms a part of wider global efforts to further understand the structure of disease-causing ST73 clones. Whole-genome sequencing and maximum-likelihood phylogenetic analyses of these clones is providing important information on the community structure of ExPEC. Here, we examined 16 ST73 isolates sourced from a single hospital and used sequence data sourced from the SRA to place these isolates into a broader global context and aid in identifying clonal lineages. Phylogenetic trees from this combined dataset, when overlaid with geographical and temporal data sourced from EnteroBase (data not shown), indicate that ST73 is globally disseminated in a manner similar to ST131 [51, 52], which is currently the most studied pandemic ExPEC lineage due to the frequency of CTX-M gene carriage. However, while ST131 tends to be relatively conserved in terms of core genome, ST73 appears more variable. Analysis of locally sequenced strains and comparison to globally sourced reads from public databases can provide context that can allow the identification of outbreak clusters with more confidence than using total SNP counts alone, and may help elucidate key outbreak groups and improve public health control of disease. This is valuable as the identification of clonal groups associated with outbreaks within larger bacterial populations remains a challenge.

Characterization of molecular signatures can also assist in the identification of outbreaks as their transfer requires physical proximity of cells. CRL including integrons and transposons are common sites of genetic rearrangement and frequently carry unique molecular signatures due to insertion elements such as IS26 [53–55]. While all class 1 integrons in the Sydney collection are not necessarily novel, there are IS-mediated signature deletions that do not appear to have been widely reported based on the current literature, such as that of the catA1 gene. This suggests that these are
Local integron variants, an idea consistent with the lack of these structures in the global SRA cohort. The major representative class 1 integron described here has been reported in its entirety once within an Australian *E. coli* O2:K1:H7 ST95 strain isolated from a bloodstream infection in 2010 (K. G. K. Goh *et al*, unpublished data; GenBank accession no. CP021289.1). This integron also shares an IS26-mediated deletion of the Tn21 resolvase gene *tnpR* with plasmid pUO-SeVR1 from a Spanish *Salmonella enterica* serovar Enteritidis strain sourced from a child with gastroenteritis [49, 56]. This is significant as this precise signature is likely the product of a single event. As such, a lateral transfer event is a likely explanation for the occurrence of this signature in disparate and geographically separate strains, followed by changes in class 1 integron cassette content. Based on the plasmid typing and PFGE data, it is likely that transfer of these integrons is being facilitated by IncF plasmids similar to pUO-SeVR1, as this is the major plasmid incompatibility type within our ST73 collection and our S1-PFGE data confirm that the class 1 integrons described here are plasmid-borne. Plasmids appear to increasingly play an important role in the mobilization of drug-resistance genes in ExPEC ST73, and their characterization relies heavily on the use of whole-genome sequencing (ideally long-read) and read-mapping technologies such as those described here.

Whole-genome sequencing allows for the analysis of gene presence/absence in clinical isolates, which will provide data on the importance of virulence genes in pathogenesis. The virulence profiles of strains sequenced in this study are consistent with other examinations of virulence in ST73 and in ExPEC more broadly. Genes encoding P fimbrial adhesins, the aerobactin siderophore (*iuc/iut*), and toxins haemolysin A and cytotoxic necrotizing factor 1 are not universally identified in worldwide ExPEC populations sourced from humans and animals [1, 57–59]. In previous work on ST73 isolates sourced from the UK, the prevalence of these genes/genetic families was also non-universal; however, *hlyA* and *cnf1* showed a substantially higher prevalence in ST73 compared with ExPEC-associated ST10, ST69 and ST95 [1]. In isolates sourced from Sydney, a relatively clear association could be identified between phylogenetic groups and virulence profiles. Further in silico categorization of virulence profiles using global ST73 reads would provide insight into virulence patterns/groups within ST73 and ExPEC, which could potentially lead to improved response, prevention and treatment of ExPEC-linked disease.

In endemic pathogens like *E. coli*, genetic comparisons of clonal group and mobile genetic element diversity can be difficult to perform with localized populations, as high numbers of closely related isolates are required for robust SNP-phylogenetic analysis and this may require the long-term collection of bacterial isolates to isolate a sufficient number of representatives. Here, we used Illumina sequencing combined with SNP-phylogenetic methods to identify at least eight distinct clonal lineages in a pilot sample of 16 ST73 isolates collected from a single hospital, indicating
NCBI Sequence Read Archive. SRR784936, SRR958320, SRR770136. 
[SRR958319 [Sample name: HVH 172 (4-3248542); BioSample: 
SAMN01885814]]. (2013).

NCBI Sequence Read Archive. SRR800420, SRR784980, SRR784979. 
[SRR800461 [Sample name: HVH 2 (4-6943160); BioSample: 
SAMN01885669]]. (2013).

NCBI Sequence Read Archive. SRR784678, SRR958401, SRR958400, SRR770170 [Sample name: HVH 207 (4-3113221); BioSample: 
SAMN01885846]]. (2013).

NCBI Sequence Read Archive. SRR784747, SRR784706, SRR784863 [Sample name: HVH 216 (4-3042952); BioSample: 

NCBI Sequence Read Archive. SRR784935, SRR784707. 
[SRR784863 [Sample name: HVH 172 (4-3248542); BioSample: 
SAMN01885814]]. (2013).

NCBI Sequence Read Archive. SRR800491, SRR784652, SRR784678. 
[SRR800461 [Sample name: HVH 2 (4-6943160); BioSample: 
SAMN01885669]]. (2013).

NCBI Sequence Read Archive. SRR800420, SRR784980, SRR784979. 
[SRR800461 [Sample name: HVH 2 (4-6943160); BioSample: 
SAMN01885669]]. (2013).

NCBI Sequence Read Archive. SRR800491, SRR784652, SRR784678. 
[SRR800461 [Sample name: HVH 2 (4-6943160); BioSample: 
SAMN01885669]]. (2013).

NCBI Sequence Read Archive. SRR784927, SRR958459, SRR958460, SRR784928 [Sample name: HVH 227 (4-2277667); BioSample: 
SAMN01885863]]. (2013).

NCBI Sequence Read Archive. SRR784927, SRR958459, SRR958460, SRR784928 [Sample name: HVH 227 (4-2277667); BioSample: 
SAMN01885863]]. (2013).
ERR1618956 (Sample name: SCK22-23; BioSample: SAMEA4429064). (2016).


References


