

# Genomic surveillance of *Escherichia coli* ST131 identifies local expansion and serial replacement of subclones

Catherine Ludden<sup>1,2,\*</sup>, Arun Gonzales Decano<sup>3†</sup>, Dorota Jamroz<sup>2</sup>, Derek Pickard<sup>4</sup>, Dearbhaile Morris<sup>5,6</sup>, Julian Parkhill<sup>7</sup>, Sharon J. Peacock<sup>4,8</sup>, Martin Cormican<sup>5</sup> and Tim Downing<sup>3</sup>

## Abstract

*Escherichia coli* sequence type 131 (ST131) is a pandemic clone that is evolving rapidly with increasing levels of antimicrobial resistance. Here, we investigated an outbreak of *E. coli* ST131 producing extended spectrum  $\beta$ -lactamases (ESBLs) in a long-term care facility (LTCF) in Ireland by combining data from this LTCF ( $n=69$ ) with other Irish ( $n=35$ ) and global ( $n=690$ ) ST131 genomes to reconstruct the evolutionary history and understand changes in population structure and genome architecture over time. This required a combination of short- and long-read genome sequencing, *de novo* assembly, read mapping, ESBL gene screening, plasmid alignment and temporal phylogenetics. We found that Clade C was the most prevalent (686 out of 794 isolates, 86%) of the three major ST131 clades circulating worldwide (A with *fimH41*, B with *fimH22*, C with *fimH30*), and was associated with the presence of different ESBL alleles, diverse plasmids and transposable elements. Clade C was estimated to have emerged in c. 1985 and subsequently acquired different ESBL gene variants (*bla*<sub>CTX-M-14</sub> vs *bla*<sub>CTX-M-15</sub>). An ISEcp1-mediated transposition of the *bla*<sub>CTX-M-15</sub> gene further increased the diversity within Clade C. We discovered a local clonal expansion of a rare C2 lineage (C2\_8) with a chromosomal insertion of *bla*<sub>CTX-M-15</sub> at the *mppA* gene. This was acquired from an IncFIA plasmid. The C2\_8 lineage clonally expanded in the Irish LTCF from 2006, displacing the existing C1 strain (C1\_10), highlighting the potential for novel ESBL-producing ST131 with a distinct genetic profile to cause outbreaks strongly associated with specific healthcare environments.

## DATA SUMMARY

The study sequences are available in the European Nucleotide Archive (<https://www.ebi.ac.uk/ena>) under project numbers PRJEB2974. A complete list of ENA accession numbers is available in Table S1 (available in the online version of this article).

## INTRODUCTION

*Escherichia coli* is the leading cause of urinary tract infections and bloodstream infections (BSIs) [1, 2], with the number of *E. coli* BSIs continuing to increase in Europe and

the USA since the early 2000s [3–7]. This has been associated with the emergence and dissemination of antibiotic-resistant *E. coli* producing extended-spectrum  $\beta$ -lactamases (ESBL-*E. coli*) conferring resistance to many beta-lactam antibiotics, including cephalosporins [6, 7]. Infections caused by ESBL-*E. coli* are associated with higher morbidity and mortality, longer hospital stays and higher healthcare costs compared to infections with antibiotic-susceptible *E. coli* [1, 8–10].

Received 31 October 2019; Accepted 23 February 2020; Published 26 March 2020

**Author affiliations:** <sup>1</sup>London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK; <sup>2</sup>Wellcome Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK; <sup>3</sup>School of Biotechnology, Dublin City University, Dublin 9, Ireland; <sup>4</sup>Department of Medicine, University of Cambridge, Box 157 Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK; <sup>5</sup>Discipline of Bacteriology, School of Medicine, National University of Ireland, Galway, Ireland; <sup>6</sup>Ryan Institute Centre for One Health, National University of Ireland Galway, Ireland; <sup>7</sup>Department of Veterinary Medicine, University of Cambridge, Madingley Road, Cambridge CB3 0ES, UK; <sup>8</sup>Cambridge University Hospitals NHS Foundation Trust, Hills Road, Cambridge CB2 0QQ, UK.

\*Correspondence: Catherine Ludden, [catherine.ludden@lshtm.ac.uk](mailto:catherine.ludden@lshtm.ac.uk)

**Keywords:** *Escherichia coli*; ST131; genomics; antimicrobial resistance; long-term care facilities.

**Abbreviations:** BSI, bloodstream infection; ESBL, extended-spectrum  $\beta$ -lactamase; ExPEC, extraintestinal pathogenic *E. coli*; FQ, fluoroquinolone; HPD, highest posterior density; LTCF, long-term care facility; MDR, multi-drug resistance; PFGE, pulsed field gel electrophoresis; rMLST, ribosomal multilocus sequence typing; rST, ribosomal sequence typing; ST, sequence type; TMRCA, time to the most recent common ancestor; TU, transposon unit.

†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. Six supplementary tables and five supplementary figures are available with the online version of this article.

000352 © 2020 The Authors



This is an open-access article distributed under the terms of the Creative Commons Attribution License.

The global spread of ESBL-*E. coli* is largely attributed to the dissemination of *E. coli* strains carrying the *bla*<sub>CTX-M-15</sub> gene, especially *E. coli* O25b:H4-ST131. Initial studies elucidated the complex clonal structure of ST131 [11, 12] by allocating isolates to subclades based on alleles of the type 1 fimbriae adhesin *fimH* gene: H41 to Clade A, H22 to Clade B and H30 to Clade C [13]. *FimH* allele type 30 (H30) is the most prevalent, followed by H22 and then H41 [13]. Although these three are the most frequent classifications [14–16], other alleles such as *fimH*35, H27, H31 and H94 have been observed in B subclades, B1–B5 [15–17]. Genomic analyses have estimated that ST131 emerged in North America over 30 years ago, coinciding with the first use of fluoroquinolone (FQ) in 1986 [15, 16]. Clade C has predominated since the 2000s, corresponding with the rapid dissemination of the *bla*<sub>CTX-M-15</sub> allele [18, 19]. Clade B also contains the subclade B0, which differs phylogenetically from the remaining B isolates by carrying *fimH*27 and is considered ancestral to Clade C [18, 20]. Clade C consists of three subclades termed C0, C1 and C2. Clade C0 has been reported as ancestral and is composed of FQ-susceptible isolates. In contrast, clades C1 (also known as H30R) and C2 (also known as H30Rx) are characterized by a double mutation at the *gyrA* and *parC* genes conferring high-level resistance to FQ [11, 15, 18]. Clade C2 is subdivided from C1 based on specific SNPs at *fimH*30 as previously described and is associated with the *bla*<sub>CTX-M-15</sub> gene [11].

ST131 has principally been associated with the hospital setting, although in recent years it has also been reported at high prevalence in the community [21–23]. There is increasing evidence that ST131 is common in the elderly and that long-term care facilities (LTCFs) are important reservoirs for ESBL-producing ST131. Reported rates of multidrug-resistant (MDR) *E. coli* ST131 carriage in residents of LTCFs include 55% in Ireland, 36% in the UK and 24% in the USA [24–26]. It is projected that the proportion of the European Union population aged ≥65 years and ≥80 years will increase to 29 and 11.5% by 2060, respectively [27]. This will probably lead to a rise in the number of people residing in LTCFs, potentially expanding the reservoir of ESBL-producing ST131. Infection control measures targeting *E. coli* have focused primarily on hospitals, and there is still a limited understanding of *E. coli* transmission dynamics within LTCFs, and between hospitals and LTCFs [25, 28]. To develop effective strategies for containment and prevention of infections, it is necessary to improve our ability to detect transmission events and to monitor the emergence of new clones. Here, we used short- and long-read genome sequencing to investigate an ESBL-*E. coli* ST131 outbreak in an LTCF in Ireland. We describe the genetic basis of antibiotic resistance and the evolution of ESBL-*E. coli* ST131 over a 7-year period. We focused our analyses on ST131 clade C because of its high frequency in this LTCF, and its MDR profile. We analysed the population structure and inferred the evolutionary history of the LTCF isolates in the context of a local hospital and global collections of *E. coli* ST131 to further our understanding of its epidemiology.

## Outcome

Extraintestinal pathogenic *E. coli* (ExPEC) ST131 is adapting in the context of antibiotic exposure, resulting in a pandemic with distinct genetic subtypes. Here, we track the evolution of antibiotic-resistance gene variants originally discovered in an ExPEC ST131 outbreak that was identified in an LTCF in Ireland. Analyses of 794 global ST131 genomes show that subclade C1 was associated with the initial infection outbreak, but that a new lineage from subclade C2 successfully displaced C1. This genetically distinct C2 subclade with a chromosomal insertion of a key antibiotic-resistance gene had clonally expanded within the LTCF. We provide new insights into the timing of genetic events driving the diversification of C2 subclades to show that that outbreak C2 strain probably evolved elsewhere before spreading to the LTCF. This study highlights the scope of antibiotic-resistance gene rearrangement within ST131, reinforcing the need to integrate genomic, epidemiological and microbiological approaches to understand ST131 transmission.

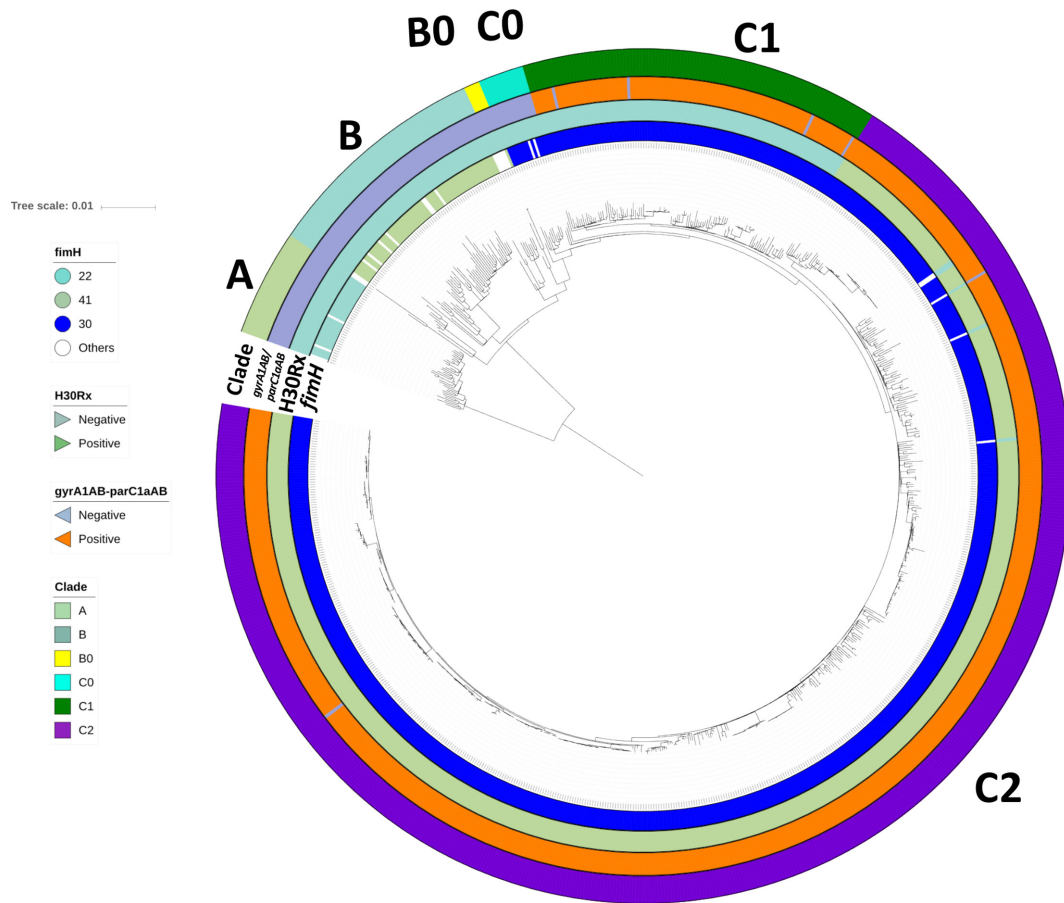
## RESULTS

### ESBL gene profiles among an *E. coli* ST131 outbreak in Ireland

In this study, we focused on the genetic profiles of 90 *E. coli* ST131 (local collection) isolated between 2005 and 2011 in Ireland, of which 69 were from one LTCF where an outbreak of ESBL-*E. coli* was first detected in 2006 [29]. The other isolates were from other LTCFs ( $n=9$ ), the referral hospital ( $n=10$ ) and the community ( $n=2$ ) (Table S1). Initial screening of the 90 isolates indicated that 64 were *bla*<sub>CTX-M-15</sub>-positive, 17 were *bla*<sub>CTX-M-14</sub>-positive, one was *bla*<sub>CTX-M-27</sub>-positive, and four were positive for both *bla*<sub>CTX-M-15</sub> and *bla*<sub>CTX-M-14</sub> (Table S1). Resistance to meropenem and ertapenem was not detected. Ribosomal sequence typing (rST) demonstrated a high incidence of rST1850 (44/90, 49%) (Table S1), suggesting emergence of a unique local epidemic clone.

### ST131 clade C predominates in Ireland and elsewhere

We analysed the 90 isolates from the local collection in the context of a global collection of 704 *E. coli* ST131 genomes that contained four additional isolates from the referral hospital described in the local collection and 10 isolates from other hospitals in Ireland. To better understand the global population structure of *E. coli* ST131, we reconstructed the phylogeny of all 794 isolates based on a core genome alignment containing 12518 SNPs (Fig. 1). This recapitulated the three established ST131 clades (A, B and C) [29] and showed that most isolates were from C ( $n=686$ , 86.4%) followed by B ( $n=75$ , 9.4%) and A ( $n=33$ , 4.2%). The clade classification was supported by previously described *fimH* allelic differences [11]: clade A was largely *fimH*41 (30 out of 33), clade B



**Fig. 1.** Phylogenetic reconstruction of  $n=794$  global ST131 strains. Maximum-likelihood phylogeny of  $n=794$  global ST131 showed three main clades, A ( $n=33$ ), B ( $n=70$ ), B0 ( $n=5$ ) and C ( $n=686$ ), with three common subclades in C: C0 ( $n=14$ ), C1 ( $n=111$ ) and C2 ( $n=561$ ). The mid-point rooted phylogram was constructed with RAxML from the chromosome-wide SNPs arising by mutation and visualized with iTol. Allelic profiling of *fimH*, *gyrA-parC*, the H30Rx phenotype and clade classification are represented in coloured strips around the phylogenetic tree.

**Table 1.** Distribution of *fimH* alleles across the entire collection of ST131 ( $n=794$ )

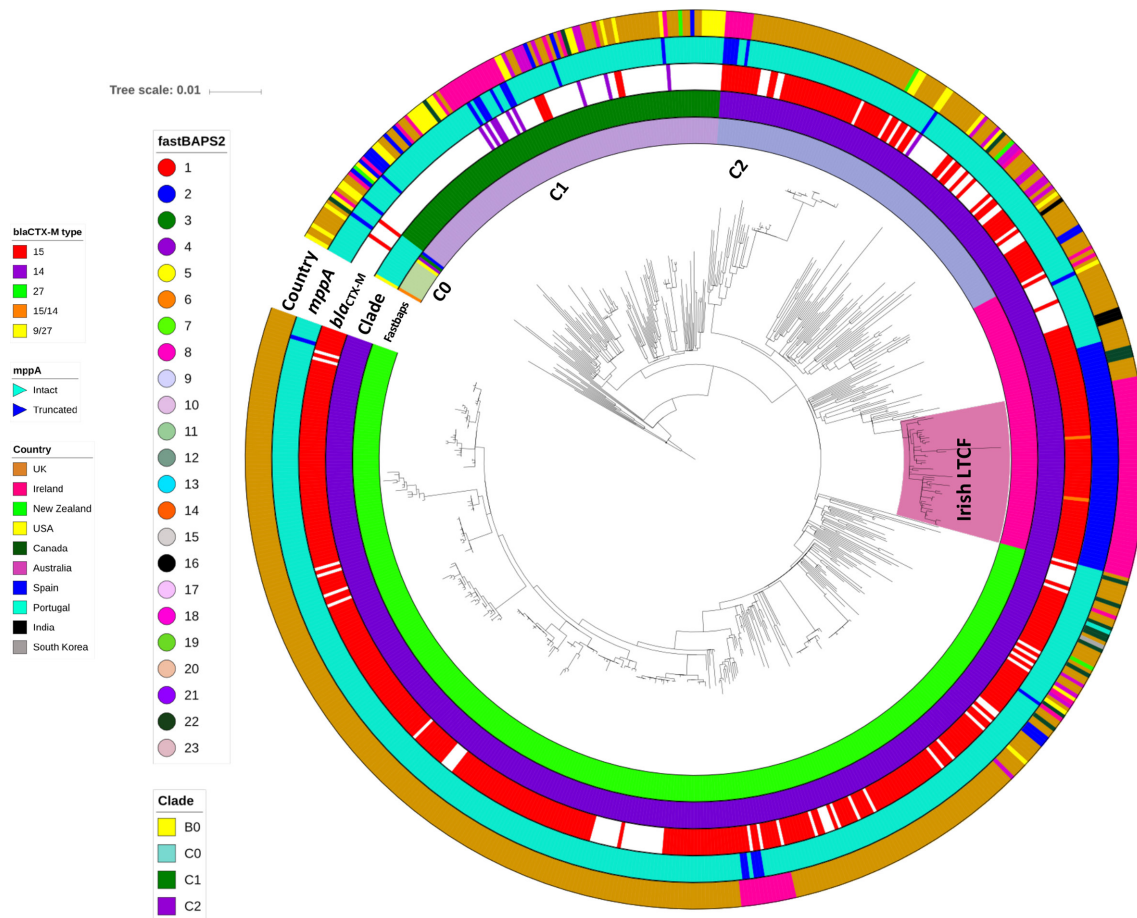
The collection consisted of three main clades subdivided into six subclades: A ( $n=33$ ), B ( $n=70$ ), B0 ( $n=5$ ), C0 ( $n=14$ ), C1 ( $n=111$ ) and C2 ( $n=561$ ). The frequencies of the four most common *fimH* allele types are shown: H41, H22, H27 and H30; the rest are classified as 'other'. No FQ-resistance mutations were detected in *fimH*22/27/41.

<i>fimH</i> allele	A	B	B0	C	C1	C2	Total
H41	30						30
H22		60	1				61
H27			4				4
H30 (non-Rx)				12	111		123
H30Rx						556	556
Other	3	10		2	1	4	20
<b>Total</b>	<b>33</b>	<b>70</b>	<b>5</b>	<b>14</b>	<b>112</b>	<b>560</b>	<b>794</b>

*fimH*22 (60 out of 70), subclade B0 *fimH*27 (4 out of 5) and clade C *fimH*30 (679 out of 686) (Table 1).

FQ-resistance alleles *gyrA1AB* and *parC1aAB* [16] were present in nearly all C1 (96%) and C2 (99.7%) isolates, along with the *fimH*30 allele, contrasting with their absence from clades A, B and B0 (Table S1). This indicated that the Clade C ancestor acquired the *fimH*30 allele and then differentiated into subclades FQ-S (*H30S* or C0) and FQ-R (*H30R* or C1, *H30Rx* or C2). A limited number of C1 ( $n=1$ ) and C2 ( $n=4$ ) isolates had lost the FQ-R *gyrA1AB-parC1aAB* genotype, consistent with intermittent recombination at these and the *fimH* genes [11].

Considerable diversity within Clade C was demonstrated by the genetic clusters identified by Fastbaps (Fig. 2, Table 2): C0 ( $n=14$ , Fastbaps clusters 2–5 and 11), C1 ( $n=111$ , Fastbaps cluster 10) and C2 ( $n=560$ , Fastbaps clusters 7–9). All 104 Irish ST131 from the National Collection (local=90, additional Irish isolates=14, see Methods) were from clade C and there were no major differences in the rates of C0, C1 and



**Fig. 2.** Maximum-likelihood phylogeny of Clade C strains from the global ST131 collection. Phylogenetic reconstruction of 686 strains from Clade C with B0 as the outgroup. This shows three common subclades in C, C0 ( $n=14$ ), C1 ( $n=111$ ) and C2 ( $n=561$ ), where the last had three distinct subgroups: C2\_7 ( $n=362$ , Fastbaps cluster 7), C2\_8 ( $n=86$ , Fastbaps cluster 8) and C2\_9 ( $n=113$ , Fastbaps cluster 9). Coloured strips surrounding the phylogram represent the clade classification, Fastbaps clusters, *bla*<sub>CTX-M</sub> allelic profile, *mppA* state (intact or truncated) and the country of origin of each strain. The highlighted 'Irish LTCF' clade was in C2\_8.

C2 in the national collection (1, 23 and 75%, respectively) compared to the global isolate collection (2, 12 and 70%, respectively).

### Phylogenetic reconstruction of three genetically distinct ST131 subclade C2 groups

Subclade C2 was structured into three Fastbaps clusters: 7 ( $n=362$ , named C2\_7), 8 ( $n=86$ , C2\_8) and 9 ( $n=113$ , C2\_9) (Fig. 2, Table 2). Most of the isolates in the National Collection ( $n=104$ ) were represented by C2\_8 ( $n=53$ , 51%), followed by C2\_7 ( $n=17$ , 16%) and C2\_9 ( $n=8$ , 8%). Within the global collection most isolates were C2\_7 ( $n=345$ , 50%), with fewer in C2\_8 ( $n=33$ , 5%) and C2\_9 ( $n=105$ , 15%) (Fig. 3). This showed C2\_7 was more common globally than in Ireland (odds ratio=3.1,  $P<6.5\times 10^{-6}$ ), and C2\_8 was more widespread in Ireland than elsewhere (odds ratio=10.6,  $P<2.2\times 10^{-16}$ ) (Fig. 3).

This difference was paralleled by the rST results, which showed that rST1503 was highly predictive of C2\_7 globally

(319 out of 345, 92.5%) and in Ireland (16 out of 17, 94%). Similarly, rST1850 was highly associated with C2\_8 in Ireland ( $n=45$ , 85%), but less so for the global collection (11 out of 33, 33%; Table 2). This limited resolution suggests rMLST (ribosomal multilocus sequence typing) has insufficient discrimination to accurately reflect the evolutionary history of clonal pathogens such as ST131, and that core genome analysis was more informative.

The entire ST131 set ( $n=794$ ) was largely composed of isolates from clade C ( $n=686$ , 86% of total) that was categorized into five subclades by Fastbaps clustering: C0 ( $n=14$ , clusters 2–5 and 11), C1 ( $n=111$ , cluster 10), C2\_7 ( $n=362$ , cluster 7), C2\_8 ( $n=86$ , cluster 8) and C2\_9 ( $n=113$ , cluster 9). The national ( $n=104$ ) and global ( $n=690$ ) ST131 collections had two main rSTs: rST1850 associated with the Irish C2\_8 LTCF set (85%), and rST1503 that often corresponded to C2\_7 (92.5%). Fastbaps clusters 2, 3, 4 and 5 in C0 represented one isolate each – only cluster 3 was *bla*<sub>CTX-M-15</sub>-positive.



**Table 2.** Distribution of ribosomal sequence types (rST) and Fastbaps clusters across clade C ( $n=686$ )

The entire ST131 set ( $n=794$ ) was largely composed of isolates from clade C ( $n=686$ , 86 % of the total) that was categorized into five subclades by Fastbaps clustering: C0 ( $n=14$ , clusters 2–5 and 11), C1 ( $n=111$ , cluster 10), C2\_7 ( $n=362$ , cluster 7), C2\_8 ( $n=86$ , cluster 8) and C2\_9 ( $n=113$ , cluster 9). The national ( $n=104$ ) and global ( $n=690$ ) ST131 collections had two main ribosomal sequence types (rSTs): rST1850 associated with the Irish C2\_8 LTCF set (85 %), and rST1503 that often corresponded to C2\_7 (92.5 %). Fastbaps clusters 2, 3, 4 and 5 in C0 represented one isolate each – only cluster 3 was *bla*<sub>CTX-M-15</sub>-positive.

	Subclade	C0	C1	C2_7	C2_8	C2_9	Total
<b>National collection</b>	Fastbaps	2–5, 11	10	7	8	9	
	rST1503	1	24	16	1	7	49
	rST1850				45		45
	Other rSTs		1	1	7	1	10
	Total	1	25	17	53	8	104
<b>Global collection</b>	rST1503	11	82	319	21	101	535
	rST1850				11		11
	Other rSTs	2	4	26	1	4	37
	Total	13	86	345	33	105	582
	Total	14	111	362	86	113	686

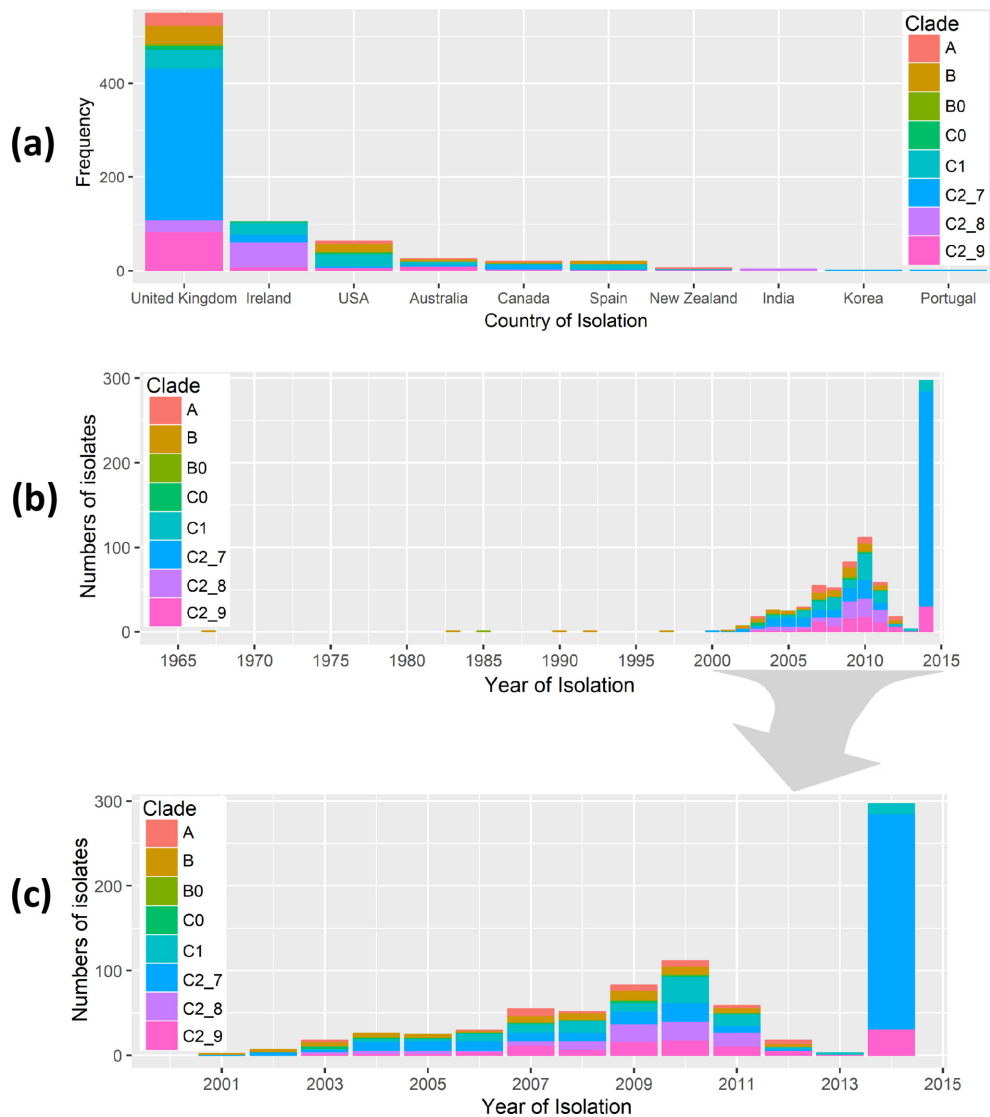
### Long read sequencing uncovers chromosomal transposition of *bla*<sub>CTX-M</sub> genes

Five isolates from the Irish collection were selected for long-read sequencing to more accurately determine the location and genomic environment of the *bla*<sub>CTX-M-14</sub> and *bla*<sub>CTX-M-15</sub> genes. Four of five samples selected were *bla*<sub>CTX-M-15</sub>-positive and members of Clade C2, of which three belonged to the predominant LTCF subclade (C2\_8) and one from the predominant global clade (C2\_7). The remaining long-read sequenced isolate was from Clade C1 and was *bla*<sub>CTX-M-14</sub>-positive. Each of the PacBio assemblies were used as references for Illumina read mapping for the collection of 794 isolates (see Methods). The three C2\_8 PacBio genomes (ERR191646, ERR191657, ERR191663) demonstrated chromosomal insertion of a 2971 bp ISEcp1-*bla*<sub>CTX-M-15</sub>-orf477Δ-Tn2 transposon unit (TU) (Fig. S1), similar to integration sites described previously [30, 31]. This TU was transposed into the 1617 bp *mppA* gene (encoding murein peptide permease A), which was split into 327 and 1290 bp segments (at NCTC13441 genome coordinates 2522100–2523713 bp). No direct repeats flanking the *bla*<sub>CTX-M-15</sub> element were observed. The *bla*<sub>CTX-M-15</sub> was separated upstream by a 48 bp spacer sequence from a fragmented ISEcp1 upstream adjacent to IS26, and downstream *bla*<sub>CTX-M-15</sub> was separated by a 46 bp spacer from an orf477 segment, which was flanked by an incomplete Tn2 and IS26 elements at the 3' and 5' ends (Table S2; Fig. S1), suggesting one-ended transposition or a deletion following transposition [30, 31]. The fourth assembly from C2\_7 (ERR191697) contained a *bla*<sub>CTX-M-15</sub> gene on an IncFII/FIA plasmid with an incomplete Tn2 element and a fragmented ISEcp1 (*p\_bla*<sub>CTX-M-15</sub>-orf477Δ-Tn2) flanked by IS26

elements (Table S2; Fig. S1). The fifth assembly was from C1 (ERR191724) and had a *bla*<sub>CTX-M-14</sub>-positive pV130-like IncFII plasmid (100% identity) with an intact ISEcp1 at the 5' end and an incomplete copy of IS903B at the 3' end (*p\_ISEcp1-bla*<sub>CTX-M-14</sub>-IS903B) (Table S2; Fig. S1).

### Genomic context of *bla*<sub>CTX-M-15</sub> in the Irish collection highlight genetically diverse C subclades

Our findings indicated that the chromosomal *bla*<sub>CTX-M-15</sub> TU inserted into the chromosome was a potentially unique characteristic of the Irish LTCF C2\_8 isolates, in contrast to the plasmid-associated *bla*<sub>CTX-M-15</sub> in other C2 isolates, and plasmid-associated *bla*<sub>CTX-M-14</sub> in C1 identified by the PacBio sequencing (Fig. S1). This was tested in 54 Clade C isolates from the Irish LTCF by resolving the exact genomic architecture of regions with the *bla*<sub>CTX-M</sub> by genome assembly and mapping reads to construct a phylogeny (Fig. S2). Assemblies of the 54 isolates were compared with the PacBio references and NCTC13441 (ERR718783), and the *bla*<sub>CTX-M-15</sub> ISEcp1, Tn2, IS903B and *mppA* copy numbers were inferred from read mapping distributions, including verification of reads spanning the genetic elements and TU boundaries (Fig. S3). Of the 54, 38 were *bla*<sub>CTX-M-15</sub>-positive (all C2), nine were *bla*<sub>CTX-M-14</sub>-positive (all C1), five had no *bla*<sub>CTX-M</sub> gene ( $n=3$  from C2,  $n=2$  from C1), and two had both *bla*<sub>CTX-M-15</sub> and *bla*<sub>CTX-M-14</sub> genes (ERR191646 and ERR191657 from C2\_8) (Fig. S3). C2\_8 isolates ( $n=29$ ) had a chromosomal insertion of *bla*<sub>CTX-M-15</sub> (Fig. S4), contrasting with C2\_7 ( $n=9$ ) that typically had a fragmented ISEcp1 with a plasmid-associated *bla*<sub>CTX-M-15</sub> gene like the C2\_8 and C2\_7 PacBio reference strains (Fig. S5). The C2\_9 ( $n=5$ ) isolates had a plasmid-bound *bla*<sub>CTX-M-15</sub> gene adjacent to a 496 bp

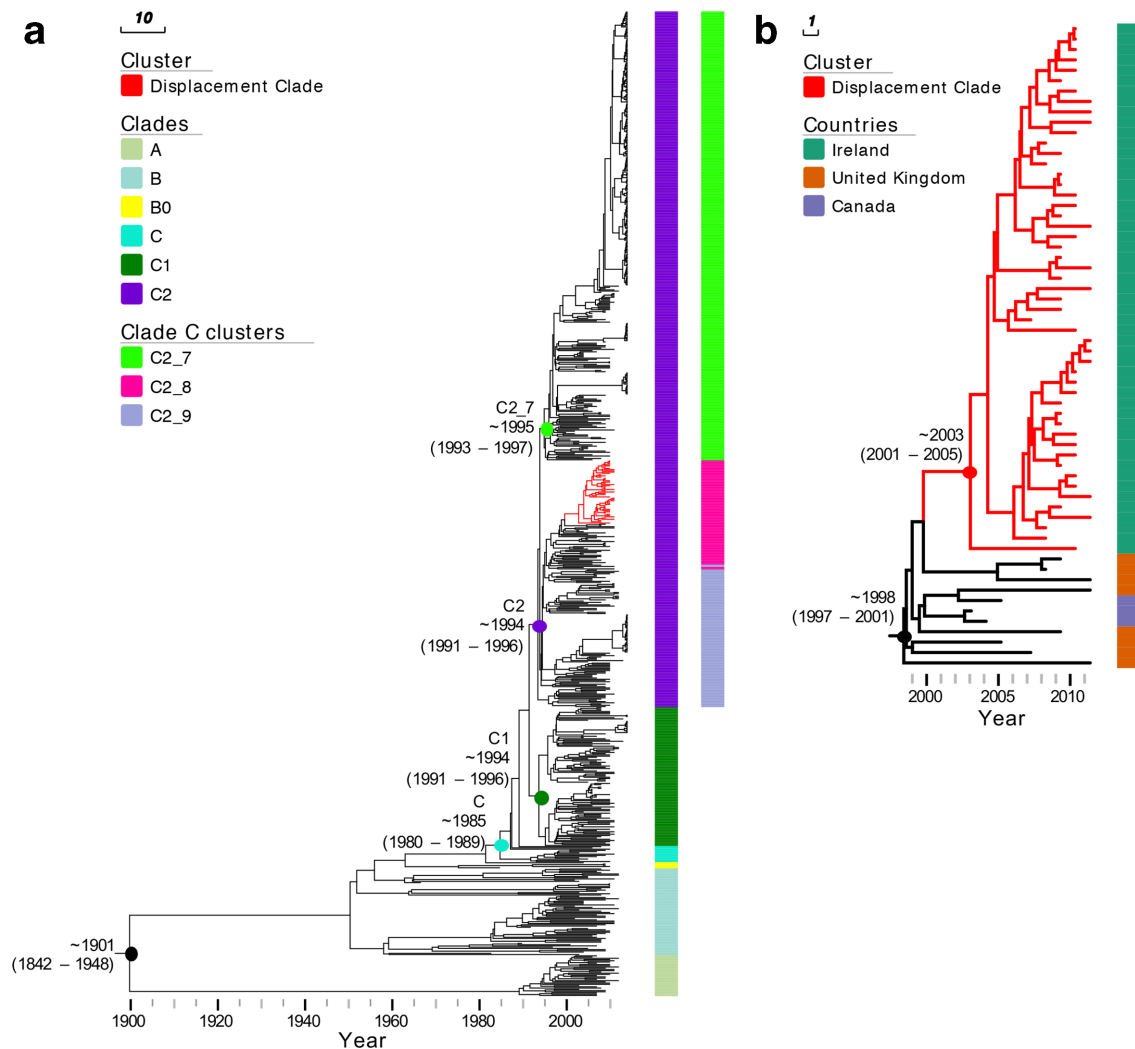


**Fig. 3.** Geographical and temporal distribution of global ST131 isolates. ST131 from the eight subclades ( $n=794$ ) showed differing frequencies across country of origin (a) and year of isolation (b and c). The subclades were A ( $n=33$ ), B ( $n=70$ ), B0 ( $n=5$ ), C0 ( $n=14$ ), C1 ( $n=111$ ), C2\_7 ( $n=362$ ), C2\_8 ( $n=86$ ) and C2\_9 ( $n=113$ ). The ST131 isolates were sampled during 1967–2014. The figures were generated using the ggplot2 and gjoy packages in R v.3.5.2.

ISEc1 fragment ( $p_{\text{shortISEc1}}\text{-}bla_{\text{CTX-M-15}}\text{-orf477}\Delta\text{-Tn2}$ , Fig. S5). Like the PacBio C1 assembly above, the C1 ( $n=11$ ) isolates had a plasmid-associated ISEc1- $bla_{\text{CTX-M-14}}$ -IS903B TU with three ISEc1 copies along with a duplicated  $bla_{\text{CTX-M-14}}$  gene, although two were  $bla_{\text{CTX-M}}$ -negative.

Examining the rest of the collection in the same way showed that the  $mppA$  TU insertion was unique to the 41 Irish LTCF isolates in Clade C2\_8 and this mutation was not found among any of the other 63 isolates from Ireland either in LTCFs, the community or hospitals. This is consistent with a pattern of clonal expansion in the LTCF. Of the 690 global isolates, 11 of the 19 with a disrupted  $mppA$  gene were  $bla_{\text{CTX-M-15}}$ -positive and clustered within the clonally expanded C2\_8  $mppA$ -insertion lineage, although the

Irish LTCF isolates with the  $mppA$  CTX-M-15 insertion formed a distinct subclade, proving that this mutation was a unique genetic feature of the LTCF lineage. Similar results have previously been observed when investigating a larger collection of 4071 globally distributed ST131 genomes from Enterobase whereby the 11 non-Irish LTCF isolates were the only isolates identified that had the CTX-M-15 insertion at the  $mppA$  gene [32]. Data for the source/origin type, the place and year of collection of the 11 non-Irish LTCF samples with CTX-M-15 in between a truncated  $mppA$  gene are given in Table S1. The remaining eight were independent events: six had no  $bla_{\text{CTX-M}}$  gene and one had a  $bla_{\text{CTX-M-19}}$  gene. Across all 794 genomes, C2 had a high rate of  $bla_{\text{CTX-M-15}}$ -positive isolates, reiterating the correlation of



**Fig. 4.** Bayesian maximum clade credibility tree of *E. coli* ST131 isolates. (a) Phylogeny of 794 isolates analysed in this study. The tree is annotated with columns representing major phylogenetic clades (Clades) as well as subclades within clade C (clade C clusters). The estimated TMRCA for major clades is shown on the tree. Branches of the cluster representing isolates from the Irish LTCF displacement clone are shown in red. (b) A higher resolution view of the Irish LTCF displacement clone, annotated with colour strips representing the isolate's country of origin.

*bla*<sub>CTX-M-15</sub> with the expansion of C2, with incidences of 84% in C2\_7 isolates (303 out of 362), 83% in C2\_8 (71 out of 86) and 67% in C2\_9 (76 out of 113).

### Time of origin of the ST131 clones

The estimated time of the most recent common ancestor (TMRCA) of different phylogenetic groups was investigated with BEAST. We estimated a mutation rate of  $4.14 \times 10^{-7}$  SNPs per site per year [95% highest posterior density (HPD) intervals  $3.74\text{--}4.57 \times 10^{-7}$ ], equivalent to 1.858 mutations per genome per year. A dated phylogeny (Fig. 4) of all 794 isolates estimated a TMRCA for ST131 of around 1901 (95% HPD intervals 1842–1948). Clade C originated in 1985 (95% HPD 1980–1989). The FQ-R C1/C2 ancestor originated in 1992 (95% HPD 1989–1994), more recently than previous estimates of 1987 [12] and 1986 [20]. Following this event, C1

and C2 diversified in parallel around 1994 (95% HPD 1991–1996, 95% HPD 1992–1995, respectively). C2 is composed of divergent subclades C2\_7, C2\_8 and C2\_9. C2\_7 diversified from the C2\_8/9 lineage in 1995 (95% HPD 1993–1997). Finally, a group of strains radiated within C2\_8 and formed a 'displacement clade' (Fig. S5a).

The TMRCA of the C2\_8 clade was estimated at around 2003 (95% HPD 2001–2005) and all isolates in this clade contained a chromosomal *bla*<sub>CTX-M-15</sub> inserted between a truncated *mppA* gene. The 'displacement clade' within the C2\_8 subcluster comprised 41 Irish LTCF isolates from the local collection which had a unique TU insertion in the *mppA*. This is in addition to 10 other Irish isolates from clinical or community sources ( $n=51$  in total) with a mutant *mppA* but showed a different TU insertion. The 11 *bla*<sub>CTX-M-15</sub>-positive isolates

clustering with the Irish C2\_8 isolates also had a disrupted *mppA* gene and were from the UK ( $n=8$ ) and Canada ( $n=3$ ). Together, these 62 shared a TMRCA of around 1998 (95% HPD 1997–2001), indicating that the *mppA* insertion may have occurred in the ancestral branch dating to 1996–1998 in the UK or North America (Fig. 5Sb).

This evidence highlighted a single genetic origin of the ancestral C2\_8 lineage in the Irish LTCF (Fig. 4), although it was rare until 2009 (Table 3), potentially presenting opportunities for multiple introductions of C2\_8. Prior to 2008, C1\_10 was most common, consistent with a pattern of replacement by C2\_8 with the mutant *mppA* insertion that clonally expanded. Nine out of 12 isolates from this facility detected between 2005 and 2007 belonged to C1\_10. This was the group of isolates corresponding to the outbreak identified in 2006. Conversely, all 57 samples isolated from 2008 to 2011 were *bla*<sub>CTX-M-15</sub>-positive, and 36 of these were classified in C2\_8 (four of which were also both *bla*<sub>CTX-M-14</sub>-positive), along with seven in C2\_9 and eight in C2\_7. The global isolates contrasted with the LTCF because C2\_8 accounted for only 5% of isolates, whereas C1\_10 and C2\_7 accounted for 14 and 49% (respectively), with no evidence of this clonal displacement outside the LTCF. Colonization and transmission within the LTCF was the most likely origin for C2\_8 given that all three rectal swabs from the index patient were ESBL-negative prior to its first detection 8 months later in 2007, and the two C2\_8 samples from 2008 had no records of recent hospitalizations.

Table 3. The numbers of isolates from the LTCF in Ireland ( $n=69$ ) across the ST131 clades showed that C1\_10 was most common at the outset of the study, and that C2\_8 became more prevalent after 2008, suggesting a possible replacement and clonal expansion of this lineage.

## DISCUSSION

Here, we traced the genomic background of ESBL-*E. coli* ST131 isolates collected from residents of an LTCF in Ireland where an outbreak was recognized in 2006. The relationship between the isolates was first identified based on indistinguishable PFGE patterns among 18 patients [29]. No point-of-source was identified during the initial investigation of the original outbreak and patients were advised accordingly regards cleaning and hand hygiene before sample collection to avoid contamination with other non-uropathogenic bacteria.

Since the outbreak was detected in 2006, there has been extensive progress in the higher discriminatory power of genome-sequencing compared to PFGE and other typing tools, such as MLST [33–35]. To gain further understanding of the origins of the outbreak and to observe changes in *E. coli* population structure in LTCF residents, we performed whole genome sequencing of all ST131 ESBL-*E. coli* isolates submitted from the LTCF over seven years. We compared these to 35 other ST131 isolated in Ireland: nine from other LTCFs, two from the community and 24 from hospitals (including 14 from the referral hospital and 10 from three other hospitals), in addition to 690 ST131 from global datasets.

We identified distinct genetic clusters within this set of 794 closely related isolates based on core genome phylogenetic signals, and as in previous studies [15], we identified subclade C2 as the most abundant ST131 group, accounting for 71% of the entire collection. Four genetic subgroups were common in the specific LTCF, one from subclade C1 (C1\_10) and three from C2 (C2\_7, C2\_8, C2\_9). The resident ST131 lineage (C1\_10) in the LTCF in the period 2005–2007 was the cause of the initial outbreak investigation, but surprisingly a newly introduced ST131 variant (C2\_8) was much more common by 2009, indicating displacement of *bla*<sub>CTX-M-14</sub>-positive C1 isolates and clonal expansion by a genetically distinct *bla*<sub>CTX-M-15</sub>-positive C2 lineage within the LTCF from 2007. This pattern of clonal displacement has not yet been published for *E. coli*, but is common in other species such as methicillin-resistant *Staphylococcus aureus* (MRSA) where it can be driven by inter-hospital transfer of patients [36].

In this study, we analysed the largest global collection of whole genome data on ST131 *E. coli* and estimated the emergence of ST131 in *c.* 1901. The clonal expansion of C2 in 1994 identified here was similar to Kallonen *et al.* [20] and Ben Zakour *et al.* [16], who reported 1990 and 1987, respectively. We dated the C2\_8 LTCF lineage to have emerged in 2001–2005 and we postulate that the clone originated in the UK or North America in 1996–1998. This was consistent with the first observation of *bla*<sub>CTX-M-15</sub>-positive cephalosporin-resistant *E. coli* isolated in 2001 in three locations in Britain and Northern Ireland [37, 38]. However, C2\_8 was generally not as successful as C2\_7, which emerged around the same time (1995) and disseminated globally. It has been suggested that the evolution of C2 subclades has been shaped by the acquisition of IncFII plasmids encoding *bla*<sub>CTX-M-15</sub> [15], which was also observed here for C2\_7. We extend this by showing that *bla*<sub>CTX-M-15</sub> in C2\_8 was mobilized from IncFII plasmids by *ISEcp1*-mediated transposition to the chromosome as *mppA* in a TU structured as *ISEcp1*-*bla*<sub>CTX-M-15</sub>-orf477Δ-Tn2. The high copy number and fragmented pattern of *ISEcp1*, which enabled a chromosomal insertion, was found for different *bla*<sub>CTX-M</sub> alleles in *E. coli* and may be linked to altered expression of the gene on the chromosome relative to the plasmid [18, 19]. Our work shows that although ST131 is disseminated globally, evolutionary events have resulted in the clonal expansion of new lineages, such as C2\_7 globally and C2\_8 locally in one LTCF. This has coincided not only with the horizontal gene transfer of plasmids encoding *bla*<sub>CTX-M-15</sub> or *bla*<sub>CTX-M-14</sub>, but also the chromosomal insertions such as *bla*<sub>CTX-M-15</sub> in C2\_8 followed by vertical transmission, and also *bla*<sub>CTX-M-14</sub> 5' of the chromosomal *rlmL* gene in one C1 isolate (ERR191666). Clonal expansion globally or locally implies that C2\_7 and C2\_8 have properties that favour their dissemination and survival in the global or local context, respectively. Factors that may support global dissemination include global travel and trade, including trade in food products. Local expansion in an LTCF may be supported by the facilities and practices in the LTCF, and antimicrobial use patterns in the LTCF are also likely to be relevant. Correlating the microbial characteristics with the environmental factors



that support expansion is a major challenge in understanding this phenomenon. Although experimental evaluation of fitness was not performed here, the most probable explanation for the displacement was that C2\_8 was biologically fit in the context of all of the conditions operating in that LTCF during the outbreak.

In conclusion, we have investigated an outbreak of ESBL-*E. coli* ST131 in an LTCF in Ireland and observed changes in this LTCF that differ from the global pattern. We found that the outbreak began with a Clade C1 strain encoding the *bla*<sub>CTX-M-14</sub> gene on a plasmid, and that this lineage was displaced by a Clade C2 strain with a chromosomally-encoded *bla*<sub>CTX-M-15</sub> gene. Both lineages associated with the LTCF are resistant to broad-spectrum cephalosporins, and the selective forces in this specific niche driving lineage displacement are unclear.

This highlighted the importance of long-read sequencing to resolve plasmids and to decipher plasmid and chromosomal spread of ESBL genes. The ability of long-read sequencing to identify novel plasmids and extra-chromosomal elements such as bacteriophages that do not integrate but replicate chromosomally should become a new standard. The sustained use of ciprofloxacin and third-generation cephalosporins will continue to enrich for Clade C2 lineages and mobile genetic elements, highlighting the global need to reduce the selective pressure from these antimicrobials. The diversity of ST131 lineages and resistance elements indicates a need for surveillance strategies to identify ST131 subclones, plasmids and transposable elements. The characterization of those specific properties that make specific lineages successful in particular contexts remains one of the key challenges in understanding the dynamics of emergence and spread of new variants of common bacterial species. Focused attention on successful strains could help to explore these interactions and control the epidemic of *E. coli* resistance.

## METHODS

### Irish bacterial isolate collection and short read genome sequencing

A total of 90 *E. coli* ST131 isolates from Ireland were isolated and sequenced. Among these, 69 were sampled from 63 residents during 2005–2011 from a single LTCF with an outbreak of ESBL-producing *E. coli* in 2006 [29] and 21 were clinical isolates from the referral hospital (Galway University Hospital:  $n=8$  hospitalized patients,  $n=11$  residents of other Irish LTCFs and  $n=2$  community isolates submitted from general practitioners).

Bacterial genomic DNA for the 90 isolates was extracted using the QIAxtractor (Qiagen) according to the manufacturer's instructions. Library preparation was conducted according to the Illumina protocol, and sequencing (96-plex) was on an Illumina HiSeq 2000 platform (Illumina) using 100 bp paired-end reads. On average, 5014175 (range 3489126–8166084) raw sequence reads were generated per isolate, with a mean insert size of 260 (range 244–280).

### Complementary datasets

For context, DNA read libraries and associated metadata were retrieved for 704 *E. coli* ST131 isolates, 14 of which were BSIs from four referral hospitals in Ireland and 4/14 isolates were obtained from the referral hospital (Galway University Hospital). The remaining global 690 were isolated between 1967 and 2014 and included 167 (clinical=155, environmental=7, unknown=5) isolates obtained from global collections [11, 12], 297 from a UK LTCF [25], and 226 were associated with BSI in the UK [20, 25] (Table S1).

### Long read sequencing, assembly and annotation

DNA was extracted using the phenol/chloroform method [39] and sequencing was done using a PacBio RSII Instrument (Pacific Biosciences) for five isolates (ERR191646, ERR191657, ERR191663, ERR191724 and ERR191697). Sequence reads were assembled using HGAP v3 [40] of the SMRT analysis software v2.3.0 (<https://github.com/PacificBiosciences/SMRT-Analysis>), circularized using Circlator v1.1.3 [41] and Minimus 2 [42], and polished using the PacBio RS\_Resequencing protocol and Quiver v1 (<https://github.com/PacificBiosciences/SMRT-Analysis>). This assembled the plasmids for each of the isolates used as references for short read mapping. NCTC13441's HDF5 files were converted to FASTQ with 308854 reads using pbh5 tools (smrtanalysis v2.3.0p4). These reads were screened for PacBio adapter sequence using Cutadapt v1.9.1 and corrected using BayesHammer from SPAdes v3.0.0 with a seed k-mer of 127, yielding a total of 41813 reads.

### Genome assembly, read mapping, AMR gene identification and plasmid typing

*De novo* assembly of short read data for the 794 libraries was performed using VelvetOptimiser v2.2.5 [43] and Velvet v1.2 [44]. An assembly improvement step was applied to the assembly with the best N50, whose contigs were scaffolded using SSPACE [45] and contig gaps reduced using GapFiller [46]. The assembly pipeline generated an average total length of 5166846 bp (range 4697700–5460279 bp) from 97 contigs (range 31–486) with an average contig length of 59340 bp (range 11186–1661401 bp) and an N50 of 227849 bp (range 30788–763538 bp) (Table S5). Assemblies were annotated using Prokka v1.5 [47] and a genus-specific database from RefSeq [48].

The 794 short read libraries were mapped to the NCTC13441 genome (accession ERS530440) [25], PacBio assemblies and reference plasmids using SMALT v7.6 (<http://www.sanger.ac.uk/resources/software/smalt/>). The genomic locations of the *bla*<sub>CTX-M</sub> genes and nearby mobile genetic elements (MGE)s were examined by aligning the short and long read assemblies using BLAST to the *bla*<sub>CTX-M</sub>-positive TU isoforms, including one with a split *mppA* gene containing the TU (Fig. S3). The two observed *mppA* isoforms were recorded as T for truncated (separated by 327 and 1290 bp segments) or I for intact (Table S1). SNP screening at *mppA* across the 794 libraries showed limited variation: just one doubleton and four singleton SNPs.

Antimicrobial resistance (AMR) genes in the 794 libraries were identified by alignment with the 2158 gene homologue subset of the Comprehensive Antibiotic Resistance Database (CARD) v1.1.5. Plasmid incompatibility group and replicon types were identified (Table S6) by comparing the genomes against the PlasmidFinder database (accessed 16 March 2017) [49] with a 95% identity threshold.

### Quality control, genome assembly and read mapping of 54 Irish read libraries

Adapter sequences in the libraries of the 54 Irish Clade C reads were trimmed with Trimmomatic v0.36 [50] using a Phred score threshold of 30 (Q30), a 10 bp sliding window and a minimum read length of 50 bp. On average, these had 2400763 reads initially, of which 7.8% were removed by trimming. These were corrected using BayesHammer in SPAdes v3.9. The effects of removing low-quality bases and reads was quantified using FastQC v0.11.5 with MultiQC v1.3, which showed that base correction removed an additional 14.3% of reads on average, leaving a mean of 1898990 per library. This showed that levels of base quality and potential contaminants were consistent across the libraries.

Read libraries of the Irish 54 libraries were assembled into contigs using SPAdes v3.9 with a k-mer of 77 [51]. This optimal k-mer maximized the N50 value determined by Quast v5.0 [52]. The contigs were ordered and scaffolded based on the NCTC13441 reference chromosome, plasmid and annotation using ProgressiveMauve [53], producing an average scaffold N50 of 177758±12199 bp (mean±SD) with a mean assembly length of 5434674±153210 bp and an average of 234 contigs per library.

A total of 59536 bases at low-complexity repeats, homopolymers, sites within 1 kb of chromosome edges, bases within 100 bp of a contig edge, or at tandem repeats were masked from the NCTC13441 reference chromosome using Tantan v0.13 ([www.cbrc.jp/tantan/](http://www.cbrc.jp/tantan/)), which was indexed using SMALT v7.6 using a k-mer of 19 with a skip of 1, as were all reference sequences here. The short read libraries were mapped to reference sequences using SMALT v7.6, and the resulting SAM files were converted to BAM format, sorted and PCR duplicates removed using SAMtools v1.19. The MGE, *mppA* and *bla<sub>CTX-M</sub>* gene structures were examined by alignment as above so that local copy number changes, mapping breakpoints and read pileups could be screened by mapping Illumina reads to the PacBio and contig references. The local gene structure was visualized with R v3.5.2 and the MARA Galileo AMR database [54, 55].

### Phylogenetic analysis of 794 isolates

To construct phylogenies reflecting the genealogical relationships and evolutionary changes, SNPs were identified using Gubbins v2.3.4. The SNPs arising by mutation were used to create a maximum-likelihood midpoint-rooted phylogeny using RAxML v8.0.19 [56] using a General Time Reversible +gamma (GTR+G) substitution model with 100 bootstraps across 362009 sites. Phylogenetic trees were

visualized with iTOL (<http://itol.embl.de>) [57] and FigTree v1.4.3 (<http://tree.bio.ed.ac.uk/software/figtree/>) [58]. For the 54 Irish Clade C collection, a phylogeny was created as above with RAxML with 100 bootstraps, and a network was constructed using uncorrected p-distances with Splitstree v4.14.2 [59], and visualized with FigTree.

### Inference of subclade common ancestry and historical population size changes

To reconstruct a time-calibrated phylogeny for ST131, we used a core genome alignment of 794 isolates that contained 8567 SNPs after the exclusion of regions representing MGEs, recombinant tracts and sites with an uncalled genotype across >1% of sequences. Each sequence in the alignment was annotated with the year of isolation. The strength of the molecular clock signal was measured by linear regression of the root-to-tip genetic distance against year of sampling using TempEst [58], which revealed a correlation coefficient of  $R^2=0.4$ . Bayesian inference of phylogeny was performed with BEAST v2.4.7 [60] based on a GTR+G nucleotide substitution model. To optimize computing efficiency in a large dataset, model selection was implemented on a subset of isolates ( $n=205$ ) that tested two clock rates (strict versus relaxed uncorrelated lognormal) across three population models (constant, exponential and Bayesian skyline). Five replicates for each of the six models were tested. The Markov chain Monte Carlo chain was run for 50 million generations, sampling every 1000 states. Log files from the five independent runs per model option were assessed for convergence using Tracer v1.5, and combined after removal of the burn-in (10% of samples) using LogCombiner. The relaxed lognormal clock with Bayesian skyline model was the best fit, consistent with previous work [61], so this was used to model the evolutionary history across all 794 isolates with 15 replicates. The maximum clade credibility tree was generated with TreeAnnotator.

#### Funding information

This project was funded by a Dublin City University (DCU) O'Hare PhD fellowship, a DCU Enhancing Performance grant, and the Health Innovation Challenge Fund (WT098600, HICF-T5-342), a parallel funding partnership between the Department of Health and Wellcome Trust. C.L. was a Wellcome Trust Sir Henry Wellcome postdoctoral fellow (110243/Z/15/Z). The views expressed in this publication are those of the authors and not necessarily those of the Department of Health or Wellcome Trust.

#### Author contributions

Conceptualization, C.L., D.M., D.P., G.D., J.P., M.C. and S.J.P.; methodology, A.G.D., C.L., D.J. and T.D.; formal analysis, A.G.D., C.L., D.J. and T.D.; investigation, A.G.D., C.L. and T.D.; visualization, A.G.D., C.L., D.J. and T.D.; resources, D.M., D.P., G.D., J.P., M.C., S.J.P. and T.D.; writing of original draft, A.G.D., C.L., D.J., J.P. and T.D.; writing, reviewing and editing, A.G.D., C.L., D.J., J.P., M.C., S.J.P. and T.D.; funding acquisition, C.L., D.M., G.D., J.P., M.C. and S.J.P. We acknowledge Mark Achtman (Warwick Medical School, UK) for initial project development and the Wellcome Sanger Institute (UK) for their assistance with sequencing.

#### Conflicts of interest

J.P. is a paid consultant of Next Gen Diagnostics.

## Data Bibliography

1. Kallonen T, Brodrick HJ, Harris SR, Corander J, Brown NM, et al. European Nucleotide Archive, PRJEB4681 (2017).
2. Brodrick HJ, Raven KE, Kallonen H, Jamroz D, Blane B, et al. European Nucleotide Archive, PRJEB7657 (2017).
3. Petty NK, Ben Zakour NL, Stanton-Cook M, Skippington E, Totsika M, et al. European Nucleotide Archive, PRJEB2968 (2014).
4. Price LB, Johnson JR, Aziz M, Clabots C, Johnston J, et al. European Nucleotide Archive, PRJNA211153 (2013).

## References

1. Tumbarello M, Spanu T, Di Bidino R, Marchetti M, Ruggeri M et al. Costs of bloodstream infections caused by *Escherichia coli* and influence of extended-spectrum-beta-lactamase production and inadequate initial antibiotic therapy. *Antimicrob Agents Chemother* 2010;54:4085–4091.
2. Burns K, Foley M, Donlon S. Point prevalence survey of hospital acquired infections & antimicrobial use in European acute care hospitals: May 2012 Republic of Ireland National Report. *HPSC Ireland, Dublin* 2012. <https://www.hpsc.ie/a-z/microbiologyanti-microbialresistancemicrobiologiantimicrobialresistance/infectioncontrolandhaiinfectioncontrolandhai/surveillance/hospitalpointprevalencesurveys/hospitalpointprevalencesurveys/2012/paps2012reportsforirelandpaps2012reportsforireland/File,13788,en.pdf>
3. Public Health England. English Surveillance Programme for Antimicrobial Utilisation and Resistance (ESPAUR) Report 2017. Report: <https://webarchive.nationalarchives.gov.uk/20191003132022/> <https://www.gov.uk/government/publications/english-surveillance-programme-antimicrobial-utilisation-and-resistance-espaur-report>
4. Gagliotti C, Balode A, Baquero F, Degener J, Grundmann H et al. *Escherichia coli* and *Staphylococcus aureus*: bad news and good news from the European Antimicrobial Resistance Surveillance Network (EARS-Net, formerly EARSS), 2002 to 2009. *Euro Surveill* 2011;16:19819.
5. Poolman JT, Wacker M. Extraintestinal pathogenic *Escherichia coli*, a common human pathogen: Challenges for vaccine development and progress in the field. *J Infect Dis* 2016;213:6–13.
6. Thaden JT, Fowler VG, Sexton DJ, Anderson DJ. Increasing incidence of extended-spectrum  $\beta$ -lactamase-producing *Escherichia coli* in community hospitals throughout the southeastern United States. *Infect Control Hosp Epidemiol* 2016;37:49–54.
7. European Centre for Disease Prevention and Control. *Antimicrobial resistance surveillance in Europe 2015. Annual Report of the European Antimicrobial Resistance Surveillance Network (EARS-Net)*. Stockholm: ECDC; 2017. <https://www.ecdc.europa.eu/sites/default/files/media/en/publications/antimicrobial-resistance-europe-2015.pdf>
8. Schwaber MJ, Carmeli Y. Mortality and delay in effective therapy associated with extended-spectrum beta-lactamase production in *Enterobacteriaceae* bacteraemia: a systematic review and meta-analysis. *J Antimicrob Chemother* 2007;60:913–920.
9. Rottier WC, Ammerlaan HSM, Bonten MJM. Effects of confounders and intermediates on the association of bacteraemia caused by extended-spectrum  $\beta$ -lactamase-producing *Enterobacteriaceae* and patient outcome: a meta-analysis. *J Antimicrob Chemother* 2012;67:1311–1320.
10. Roberts RR, Hota B, Ahmad I, Scott RD, Foster SD et al. Hospital and societal costs of antimicrobial-resistant infections in a Chicago teaching hospital: implications for antibiotic stewardship. *Clin Infect Dis* 2009;49:1175–1184.
11. Price LB, Johnson JR, Aziz M, Clabots C, Johnston B et al. The epidemic of extended-spectrum- $\beta$ -Lactamase-producing *Escherichia coli* ST131 Is driven by a single highly pathogenic subclone, H30-Rx. *MBio* 2013;4.
12. Petty NK, Ben Zakour NL, Stanton-Cook M, Skippington E, Totsika M et al. Global dissemination of a multidrug resistant *Escherichia coli* clone. *Proc Natl Acad Sci U S A* 2014;111:5694–5699.
13. Johnson JR, Tchesnokova V, Johnston B, Clabots C, Roberts PL et al. Abrupt emergence of a single dominant multidrug-resistant strain of *Escherichia coli*. *J Infect Dis* 2013;207:919–928.
14. Adams-Sapper S, Diep BA, Perdreau-Remington F, Riley LW. Clonal composition and community clustering of drug-susceptible and -resistant *Escherichia coli* isolates from bloodstream infections. *Antimicrob Agents Chemother* 2013;57:490–497.
15. Stoesser N, Sheppard AE, Pankhurst L, De Maio N, Moore CE et al. Evolutionary history of the global emergence of the *Escherichia coli* epidemic clone ST131. *mBio* 2016;7:e02162.
16. Ben Zakour NL, Alsheikh-Hussain AS, Ashcroft MM, Khanh Nhu NT, Roberts LW et al. Sequential acquisition of virulence and fluoroquinolone resistance has shaped the evolution of *Escherichia coli* ST131. *mBio* 2016;7:e00347–16.
17. Nicolas-Chanoine M-H, Robert J, Vigan M, Laouénan C, Brisse S et al. Different factors associated with CTX-M-producing ST131 and Non-ST131 *Escherichia coli* clinical isolates. *PLoS One* 2013;8:e72191.
18. Matsumura Y, Pitout JDD, Peirano G, DeVinney R, Noguchi T et al. Rapid identification of different *Escherichia coli* sequence type 131 clades. *Antimicrob Agents Chemother* 2017;61:e00179–00117.
19. Cantón R, González-Alba JM, Galán JC. CTX-M enzymes: origin and diffusion. *Front Microbiol* 2012;3:110.
20. Kallonen T, Brodrick HJ, Harris SR, Corander J, Brown NM et al. Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Res* 2017;27:1437–1449.
21. Vidal-Navarro L, Pfeiffer C, Bouziges N, Sotto A, Lavigne J-P. Faecal carriage of multidrug-resistant Gram-negative bacilli during a non-outbreak situation in a French university hospital. *J Antimicrob Chemother* 2010;65:2455–2458.
22. Rogers BA, Sidjabat HE, Paterson DL. *Escherichia coli* O25b-ST131: a pandemic, multiresistant, community-associated strain. *J Antimicrob Chemother* 2011;66:1–14.
23. Tchesnokova V, Riddell K, Scholes D, Johnson JR, Sokurenko EV. The uropathogenic *Escherichia coli* subclone sequence type 131-H responsible for most antibiotic prescription errors at an urgent care clinic. *Clinical Infectious Diseases* 2019;68:781–787.
24. Ludden C, Cormican M, Vellinga A, Johnson JR, Austin B et al. Colonisation with ESBL-producing and carbapenemase-producing *Enterobacteriaceae*, vancomycin-resistant enterococci, and methicillin-resistant *Staphylococcus aureus* in a long-term care facility over one year. *BMC Infect Dis* 2015;15:168.
25. Brodrick HJ, Raven KE, Kallonen T, Jamroz D, Blane B et al. Longitudinal genomic surveillance of multidrug-resistant *Escherichia coli* carriage in a long-term care facility in the United Kingdom. *Genome Med* 2017;9:70.
26. Burgess MJ, Johnson JR, Porter SB, Johnston B, Clabots C et al. Long-term care facilities are reservoirs for antimicrobial-resistant sequence type 131 *Escherichia coli*. *Open Forum Infect Dis* 2015;2:ofv011.
27. Suetens C. Healthcare-associated infections in European long-term care facilities: how big is the challenge? *Euro Surveill* 2012;17:pii=20259.
28. Burke L, Humphreys H, Fitzgerald-Hughes D. The revolving door between hospital and community: extended-spectrum beta-lactamase-producing *Escherichia coli* in Dublin. *J Hosp Infect* 2012;81:192–198.
29. Pelly H, Morris D, O'Connell E, Hanahoe B, Chambers C et al. Outbreak of extended spectrum beta-lactamase producing *E. coli* in a nursing home in Ireland, May 2006. *Weekly releases* 2006;11:3036.
30. Johnson JR, Johnston B, Clabots C, Kuskowski MA, Castanheira M. *Escherichia coli* sequence type ST131 as the major cause of serious multidrug-resistant *E. coli* infections in the United States. *Clin Infect Dis* 2010;51:286–294.



31. Johnson JR, Porter S, Thuras P, Castanheira M. The pandemic H30 subclone of sequence type 131 (ST131) as the leading cause of multidrug-resistant *Escherichia coli* infections in the United States (2011–2012). *Open Forum Infect Dis* 2017;4:ofx089.
32. Decano AG, Downing T. An *Escherichia coli* ST131 pangenome atlas reveals population structure and evolution across 4,071 isolates. *Sci Rep* 2019;9:17394.
33. Salipante SJ, SenGupta DJ, Cummings LA, Land TA, Hoogestraat DR et al. Application of whole-genome sequencing for bacterial strain typing in molecular epidemiology. *J Clin Microbiol* 2015;53:1072–1079.
34. Rumore J, Tschetter L, Kearney A, Kandar R, McCormick R et al. Evaluation of whole-genome sequencing for outbreak detection of Verotoxigenic *Escherichia coli* O157:H7 from the Canadian perspective. *BMC Genomics* 2018;19:870.
35. Ludden C, Raven KE, Jamrozny D, Gouliouris T, Blane B et al. One health genomic surveillance of *Escherichia coli* demonstrates distinct lineages and mobile genetic elements in isolates from humans versus livestock. *mBio* 2019;10:e02693-18 [Epub ahead of print 22 01 2019].
36. Hsu L-Y, Harris SR, Chlebowicz MA, Lindsay JA, Koh T-H et al. Evolutionary dynamics of methicillin-resistant *Staphylococcus aureus* within a healthcare system. *Genome Biol* 2015;16:81.
37. Mushtaq S, Woodford N, Potz N, Livermore DM. Detection of CTX-M-15 extended-spectrum beta-lactamase in the United Kingdom. *J Antimicrob Chemother* 2003;52:528–529.
38. Livermore DM, Mushtaq S, James D, Potz N, Walker RA et al. In vitro activity of piperacillin/tazobactam and other broad-spectrum antibiotics against bacteria from hospitalised patients in the British Isles. *Int J Antimicrob Agents* 2003;22:14–27.
39. Hull RA, Gill RE, Hsu P, Minshew BH, Falkow S. Construction and expression of recombinant plasmids encoding type 1 or D-mannose-resistant pili from a urinary tract infection *Escherichia coli* isolate. *Infect Immun* 1981;33:933–938.
40. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 2013;10:563–569.
41. Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA et al. Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol* 2015;16:294.
42. Sommer DD, Delcher AL, Salzberg SL, Pop M. Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* 2007;8:64.
43. Gladman S, Seemann T, Victorian Bioinformatics Consortium. Velvet Optimiser: for automatically optimising the primary parameter options for the velvet de novo sequence assembler 2008. <http://www.vicbioinformatics.com/software.velvetoptimiser.shtml>
44. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 2008;18:821–829.
45. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 2011;27:578–579.
46. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. *Genome Biol* 2012;13:R56.
47. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
48. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* 2012;40:D130–D135.
49. Carattoli A, Bertini A, Villa L, Falbo V, Hopkins KL et al. Identification of plasmids by PCR-based replicon typing. *J Microbiol Methods* 2005;63:219–228.
50. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
51. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
52. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;29:1072–1075.
53. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 2010;5:e11147.
54. Partridge SR, Tsafnat G, Coiera E, Iredell JR. Gene cassettes and cassette arrays in mobile resistance integrons. *FEMS Microbiol Rev* 2009;33:757–784.
55. Tsafnat G, Coiera E, Partridge SR, Schaeffer J, Iredell JR. Context-driven discovery of gene cassettes in mobile integrons using a computational grammar. *BMC Bioinformatics* 2009;10:281.
56. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–1313.
57. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* 2016;44:W242–.
58. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution* 2016;2:vev007.
59. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 2006;23:254–267.
60. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H et al. Beast 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* 2014;10:e1003537.
61. von Mentzer A, Connor TR, Wieler LH, Semmler T, Iguchi A et al. Identification of enterotoxigenic *Escherichia coli* (ETEC) clades with long-term global distribution. *Nat Genet* 2014;46:1321–1326.

### Five reasons to publish your next article with a Microbiology Society journal

1. The Microbiology Society is a not-for-profit organization.
2. We offer fast and rigorous peer review – average time to first decision is 4–6 weeks.
3. Our journals have a global readership with subscriptions held in research institutions around the world.
4. 80% of our authors rate our submission process as 'excellent' or 'very good'.
5. Your article will be published on an interactive journal platform with advanced metrics.

Find out more and submit your article at [microbiologyresearch.org](http://microbiologyresearch.org).