Global dissemination of a multidrug resistant *Escherichia coli* clone

Nicola K. Petty[a,b,c,1], Nouri L. Ben Zakour[a,b,1], Mitchell Stanton-Cook[a,b], Elizabeth Skippington[a,b], Makrina Totsika[a,b], Brian M. Forde[a,b], Minh-Duy Phan[a,b], Danilo Gomes Moriel[a,b], Kate M. Peters[a,b], Mark Davies[a,b,d], Benjamin A. Rogers[e], Gordon Dougan[d], Jesús Rodriguez-Baño[f,g], Alvaro Pascual[f,g], Johann D. D. Pitout[h,i], Mathew Upton[j], David L. Paterson[a,e], Timothy R. Walsh[k], Mark A. Schembri[a,b,2], and Scott A. Beatson[a,b,2]

[a]Australian Infectious Diseases Research Centre, and [b]School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD 4072, Australia; [c]The ithree institute, University of Technology Sydney, Sydney, NSW 2007, Australia; [d]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, United Kingdom; [e]University of Queensland Centre for Clinical Research, Royal Brisbane and Women's Hospital Campus, The University of Queensland, Brisbane, QLD 4029, Australia; [f]Unidad Clínica de Enfermedades Infecciosas y Microbiología, Hospital Universitario Virgen Macarena, 41007 Seville, Spain; [g]Departmentos de Medicina y Microbiología, Universidad de Sevilla, 41009 Seville, Spain; [h]Division of Microbiology, Calgary Laboratory Services, Calgary, AB, Canada T2L 2KB; [i]Department of Pathology and Laboratory Medicine, and Department of Microbiology and Infectious Diseases, University of Calgary, Calgary, AB, Canada T2N 1N4; [j]School of Biomedical and Healthcare Sciences, Plymouth University, Plymouth PL4 8AA, United Kingdom; [k]School of Medicine, Cardiff University, Cardiff CF14 4XN, United Kingdom

Edited by Scott J. Hultgren, Washington University School of Medicine, St. Louis, MO, and approved March 4, 2014 (received for review December 5, 2013)

***Escherichia coli* sequence type 131 (ST131) is a globally disseminated, multidrug resistant (MDR) clone responsible for a high proportion of urinary tract and bloodstream infections. The rapid emergence and successful spread of *E. coli* ST131 is strongly associated with several factors, including resistance to fluoroquinolones, high virulence gene content, the possession of the type 1 fimbriae FimH30 allele, and the production of the CTX-M-15 extended spectrum β-lactamase (ESBL). Here, we used genome sequencing to examine the molecular epidemiology of a collection of *E. coli* ST131 strains isolated from six distinct geographical locations across the world spanning 2000–2011. The global phylogeny of *E. coli* ST131, determined from whole-genome sequence data, revealed a single lineage of *E. coli* ST131 distinct from other extraintestinal *E. coli* strains within the B2 phylogroup. Three closely related *E. coli* ST131 sublineages were identified, with little association to geographic origin. The majority of single-nucleotide variants associated with each of the sublineages were due to recombination in regions adjacent to mobile genetic elements (MGEs). The most prevalent sublineage of ST131 strains was characterized by fluoroquinolone resistance, and a distinct virulence factor and MGE profile. Four different variants of the CTX-M ESBL–resistance gene were identified in our ST131 strains, with acquisition of CTX-M-15 representing a defining feature of a discrete but geographically dispersed ST131 sublineage. This study confirms the global dispersal of a single *E. coli* ST131 clone and demonstrates the role of MGEs and recombination in the evolution of this important MDR pathogen.**

bacterial evolution | genomics | phylogeography | genomic epidemiology

Many multidrug-resistant (MDR) bacterial strains are now recognized as belonging to clones that originate in a specific locale, country, or even globally. *Escherichia coli* sequence type 131 (ST131) is one such recently emerged and globally disseminated MDR pandemic clone responsible for community and hospital-acquired urinary tract and bloodstream infections. *E. coli* ST131 was identified in 2008 as a major clone linked to the spread of the CTX-M-15 extended-spectrum β-lactamase (ESBL) resistance (1–3). Since then, *E. coli* ST131 has also been strongly associated with fluoroquinolone resistance, and coresistance to aminoglycosides and trimethoprim-sulfamethoxazole (4–6). Alarmingly, strains of *E. coli* ST131 resistant to carbapenems have also been reported (7, 8), further limiting treatment options for this clone.

E. coli ST131 belongs to the B2 phylogenetic subgroup I, with most isolates characterized as serotype O25b:H4 (1). Epidemiology studies using pulse-field gel electrophoresis (PFGE) have demonstrated that *E. coli* ST131 strains exhibit diversity, with some dominant PFGE pulsotypes including the UK epidemic strain A (9) and pulsotype 968 (10, 11) widely distributed across the globe. More recently, a typing scheme using the type 1 fimbriae *fimH* adhesin gene revealed that a large subclonal lineage of *E. coli* ST131 strains possess the FimH30 allele, which is also associated with specific mutations in the *gyrA* and *parC* genes that confer resistance to fluoroquinolones (12).

Several whole genome (13–16) and PCR (1, 17–20) studies have revealed that *E. coli* ST131 strains possess a variable complement of genes encoding established virulence factors commonly associated with extraintestinal pathogenic *E. coli* (ExPEC). Indeed, few virulence genes appear to be uniformly present in *E. coli* ST131 and, thus, it is likely that differences in virulence gene content contribute to the variable virulence potential that has been reported. For example, although some ST131 strains cause rapid death in a mouse sepsis infection model (21), this phenotype is not consistent among all strains (22). The *E. coli* ST131 strain EC958, which is a representative of the FimH30-fluoroquinolone resistant subgroup, has been characterized at the molecular level (15). *E. coli* EC958 contains an insertion in the type 1 fimbriae regulator gene *fimB* (15) that

Significance

***Escherichia coli* sequence type 131 (ST131) is a globally disseminated multidrug-resistant clone associated with human urinary tract and bloodstream infections. Here, we have used genome sequencing to map the temporal and spatial relationship of a large collection of *E. coli* ST131 strains isolated from six distinct geographical regions across the world. We show that *E. coli* ST131 strains are distinct from other extraintestinal pathogenic *E. coli* and arose from a single progenitor strain prior to the year 2000.**

Author contributions: N.K.P., N.L.B.Z., M.A.S., and S.A.B. designed research; N.K.P., N.L.B.Z., M.S.-C., E.S., M.T., B.M.F., M.-D.P., D.G.M., K.M.P., M.D., M.U., and S.A.B. performed research; M.S.-C., E.S., B.A.R., G.D., J.R.-B., A.P., J.D.D.P., M.U., D.L.P., and T.R.W. contributed new reagents/analytic tools; N.K.P., N.L.B.Z., M.A.S., and S.A.B. analyzed data; and N.K.P., N.L.B.Z., M.A.S., and S.A.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Database deposition: The sequences reported in this paper have been deposited in the European Nucleotide Archive under study nos. ERP001354 and ERP004358. For a list of accession numbers, see Dataset S1.

[1]N.K.P. and N.L.B.Z. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: s.beatson@uq.edu.au or m.schembri@uq.edu.au.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1322678111/-/DCSupplemental.

is also common to other strains in the FimH30 subgroup (23) and colonizes the mouse bladder in a type 1 fimbriae-dependent manner (15). In mice, *E. coli* EC958 establishes acute and chronic urinary tract infection (UTI), forms intracellular bacterial communities in the bladder (24), and causes impairment of ureter contractility (25). *E. coli* EC958 is also resistant to the bactericidal action of human serum (26).

The rapid global dissemination of *E. coli* ST131, combined with its MDR phenotype and the lack of new antimicrobial drugs in the developmental pipeline, highlights the urgent need to understand this pathogen and combat its spread. Here, we sequenced the genomes of 95 *E. coli* ST131 strains from six geographical regions (isolated from 2000 to 2011) to examine the spatial and temporal relationships of *E. coli* ST131. Our data supports the rapid and recent global dispersal of *E. coli* ST131 as a single clone.

Results and Discussion

A Global Collection of *E. coli* ST131 Strains. A collection of 99 *E. coli* strains defined as ST131, using a described PCR test specific for the O25b *rfb* gene and allele 3 of the *pabB* gene (27), were isolated between 2000 and 2011 from six countries (Australia, Canada, India, Spain, United Kingdom, New Zealand) (Dataset S1). The strains were obtained from several clinical sources and included isolates from urine ($n = 53$), blood ($n = 21$), peritoneal fluid ($n = 1$), abdominal abscess ($n = 1$), surgical wound ($n = 2$), and rectal swabs ($n = 11$). The strains were selected with an endeavor to encompass diversity with respect to geographic origin, date of isolation, and clinical source. The strains possessed a range of antibiograms, including variable resistance to aminoglycosides, second and third generation cephalosporins, fluoroquinolones, penicillins, and sulfonamides (Dataset S1). All strains were sequenced by using the Illumina HiSeq, assembled using Velvet, and *in silico* multilocus sequence typing (MLST) was performed to confirm the sequenced strains were ST131. Four strains originally defined as ST131 by *rfb* and *pabB* PCR actually belonged to ST95 (Dataset S1), thus reducing the final number of ST131 strains examined to 95.

Rapid Global Dispersal of *E. coli* ST131 as a Single Clone. Phylogenetic analysis of the 95 *E. coli* ST131 strains was carried out by using whole genome alignment and single-nucleotide polymorphism (SNP) analysis using the completely sequenced ST131 representative strain SE15 (13). A maximum likelihood (ML) tree built using all 142,750 SNPs confirmed that all ST131 strains belonged to phylogroup B2, subgroup I and showed that ST131 clustered into three well-supported clades that we refer to as A, B, and C (*SI Appendix*, Fig. S1*A*). ML trees based on the 3,186,979-bp core alignment of the assembled sequence data supported this topology (*SI Appendix*, Fig. S2*A*). Recombination is the primary contributor to interclade diversity with only 70 nucleotide substitutions found to distinguish clades B and C after removal of recombinant regions (Fig. 1*A* and Dataset S2). Neither temporal nor geographical clustering between the major clades could be observed (Fig. 1*A*); however, each clade is comprised of at least two well-supported sublineages and smaller clusters of closely-related strains that exhibit some geographical association (Fig. 1*B* and Dataset S2). This data suggested an evolutionary history more complex than a standard geographical clonal expansion, as exemplified by many occurrences of nearly identical strains isolated in different countries and continents and over different periods of time. Similar phylogeographic patterns have been observed for other successful MDR global lineages such as *Staphylococcus aureus* ST239 and the PMEN1 pneumococcal lineages (28, 29), whereas a contrasting example of clonal expansion with more defined geographical clustering has been reported for *Shigella sonnei* (30).

ST131 clade A contains the previously-sequenced SE15 strain and is the most divergent clade (~7,000 and ~8,900 SNPs from clades B and C, respectively) characterized by the *fimH41* allele and different *gyrA* and *parC* variants. ST131 clade B is very similar to clade C (distinguished by ~2,900 SNPs) and is characterized by

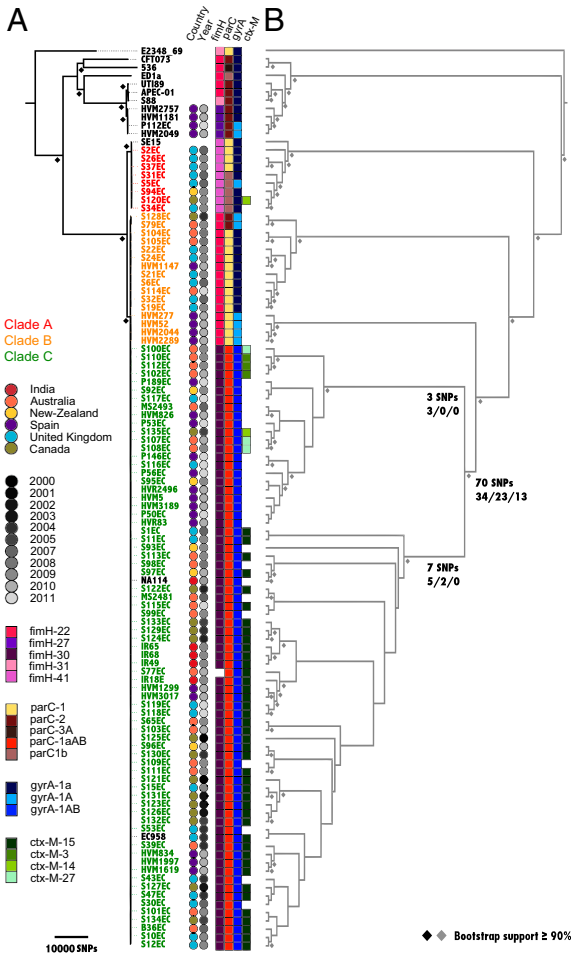

Fig. 1. Phylogenetic relationship of ST131 strains. (*A*) ML phylogram with triangles indicating bootstrap support of >90% from 1,000 replicates. The tree is rooted by using the outgroup phylogroup D strain UMN026; branch lengths correspond to the number of SNPs difference (scale bar bottom left). The phylogram was built from 119,514 substitution-only SNPs determined by read-mapping using *E. coli* SE15 as reference excluding recombinant regions, as defined by BRATNextGen analysis (34). The taxa labels for sequenced ST131 strains are colored red (clade A), orange (clade B) and green (clade C). Previously sequenced reference strains are colored black. Colored circles next to each strain correspond to country and year of isolation (see key). Squares indicate allelic profiling for *fimH*, *parC*, *gyrA*, and CTX-M (see key). A missing square indicates the gene is absent. (*B*) Several well-supported subclades are evident in the ST131 phylogeny, with the CTX-M-15 gene confined to the second subclade of clade C. The topology-only cladogram (not to scale) corresponding to the phylogram in *A* is shown in gray, with node support of >90% depicted as gray diamonds. The number of SNPs that define clade C and sublineages C1 (*Upper*) and C2 (*Lower*) are shown below relevant branches (nonsynonymous, synonymous, intergenic); refer to Dataset S2 for full list of SNPs and consequences.

an intact *fimB*, the *fimH*22 allele, and *gyrA* and *parC* variants that are consistent with their fluoroquinolone sensitivity (Fig. 1*A* and Dataset S1). ST131 clade C strains make up 79% of the ST131 strains sequenced in this study and are distinguished by possession of the *fimH*30 allele and the fluoroquinolone resistance alleles *gyrA*1AB and *parC*1aAB (Fig. 1*A*). All but one of the clade C strains contained an insertion within the *fimB* gene as we originally observed in the clade C strain EC958 (15). These isolates were collected in six countries from 2000 to 2011, indicating that the dominant clade C ST131 lineage originated from a single clone before the year 2000 (Fig. 1*A*). Although we cannot rule out the possibility of a bias in our strain collection, we note that the dominant group among another large collection of ST131 strains

MICROBIOLOGY

was also found to share the same *fimH*30-*gyrA*1AB-*parC*1aAB allelic profile (12).

Analysis of the density of all SNPs along the SE15 reference chromosome revealed a nonhomogeneous distribution, with many core-genome regions associated with a density ~8.5-fold higher than the expected average (Fig. 2). Because discrete regions with a high-density of SNPs may be the result of recombination events, as opposed to mutational hotspots (31, 32), we inferred the recombination across ST131 genomes by using a Bayesian clustering approach that was previously successfully applied to *S. aureus* and *Streptococcus pneumoniae* (33, 34). We found that recombination has introduced 76.6% of the 16,424 SNPs and 2,050 small indels that differentiate the strains within the ST131 lineage (*SI Appendix*, Fig. S3 and Dataset S3). Phylogenetic analysis using only SNPs found in recombinant regions also clustered the ST131 strains into the same three-clades structure (*SI Appendix*, Figs. S1*B* and S2*B*). Overall these results reflect the significant role that recombination has played in shaping the three major ST131 lineages with subsequent point mutations driving the fine-scale diversity within each clade.

Antibiotic Resistance Is Associated with ST131 Clade C. Besides the major contribution of recombination events to the between-clade diversity of ST131, we also observed differences in the distribution of SNPs between recombinant and nonrecombinant regions (Fig. 2). SNP density across all strains combined was higher in recombinant regions with an estimated 1.19×10^{-2} SNPs per site compared with 1.39×10^{-3} SNPs per site in nonrecombinant regions. Despite the lower density of SNPs, nonrecombinant regions were characterized by a relatively higher ratio of nonsynonymous to synonymous SNPs (0.05 and 0.07 SNPs per kilobase, respectively) compared with recombinant regions (0.2 and 0.89 SNPs per kilobase, respectively). This difference was significant ($\chi^2 = 1{,}045.8$, $P < 0.00001$) and is consistent with a pairwise comparison of ST131 clade A and clade C strains (23).

Fluoroquinolone resistance is one of the major determining features of the ST131 clone and is associated with point mutations in the *gyrA* and *parC* genes (12) (Fig. 1). The three major *gyrA* alleles found in our ST131 dataset were attributed to vertically transmitted point mutations, with unique *gyrA* mutations also found in clade A strain S5EC (A669T) and clade C strain B36EC (Q453R), respectively. In contrast, the *parC*1aAB allele was introduced into clade C via recombination, replacing the *parC*1 allele and surrounding Rec_089 region that is conserved in most clade A and B strains (Dataset S3). Multiple, overlapping recombination events continue to shape the ST131 lineage as evidenced by two independent replacements of Rec_089 in subgroups of clade A (encompassing *parC*2) and clade B (*parC*3A), with a further two partial replacements of a 1.8-kb Rec_089 subfragment immediately upstream of *parC* in two clade C strains (S101EC and S113EC). Among the 34 nonsynonymous and nonrecombinant substitutions that define clade C, we could map nine to crystal structures of homologs, several of which encode amino acid changes that may impact their function (*SI Appendix*, Fig. S5 and Dataset S2). For example, there is a mutation in the gene encoding the MukB chromosome partition protein, a known interacting partner of ParC (35). In addition to established *fimH*, *parC* and *gyrA* mutations in clade C strains, our identification of further genes with clade C-specific mutations paves the way for more targeted investigations to identify key evolutionary events that underpin the success of *E. coli* ST131.

Among the SNPs that have arisen in individual ST131 clade C strains or subgroups, there are a number within potential antibiotic resistance genes that may have been selected in response to antibiotic treatment (Dataset S2). Each ST131 clade C strain (minus NA114) has between 0 and 50 (mean = 13, SD = 11) unique, nonrecombinant SNPs, 49% of which are nonsynonymous. There are numerous examples of nonsynonymous SNPs within genes that encode homologs of multidrug resistance proteins or other putative transporters that may affect antimicrobial uptake or efflux (Dataset S2). There are also several SNPs in genes encoding penicillin-binding proteins (e.g., ECSF_2363/PBP1C, ECSF_0094/PBP3), other cell wall modifying enzymes (e.g., ECSF_2495 lytic murein transglycosylase B) and examples of cell division genes (e.g., ECSF_2198), or essential genes that may be important for intrinsic resistance development. Although the majority of ST131 clade C SNPs are unique to the strain in which they are found, or exhibit patterns of descent consistent with the inferred phylogeny, we identified genes in which the same mutation appeared to have been acquired independently (Dataset S2). For example, the dihydrofolate reductase gene (ESCF_0053) acquired the trimethoprim resistance L28R mutation in two phylogenetically separated clade C strains (S116EC and S11EC), with several other nonsynonymous mutations in this gene present in different strains.

The majority of clade C strains also possess the CTX-M-15 gene (36 of 42 strains in sublineage C2), with seven other clade C strains containing different CTX-M alleles (3, 14, or 27) (Fig. 1*A* and Dataset S1). The CTX-M-15–positive strains cluster within a discrete, but temporally and geographically dispersed, sublineage within clade C (Fig. 1*B*). Although the pattern of CTX-M-15 distribution within this sublineage is suggestive of an ancestral acquisition of the CTX-M-15 gene and subsequent loss by some individual strains, this allele does not associate with any particular plasmid incompatibility group defined by sequence-based typing (*SI Appendix*, Fig. S4). Furthermore, the CTX-M-15 gene is found on assembled contiguous fragments (contigs) ranging in size from 1.4 kb to 10 kb with variable adjacent gene content (many of which have been previously identified on plasmids), suggesting that the CTX-M-15 gene has

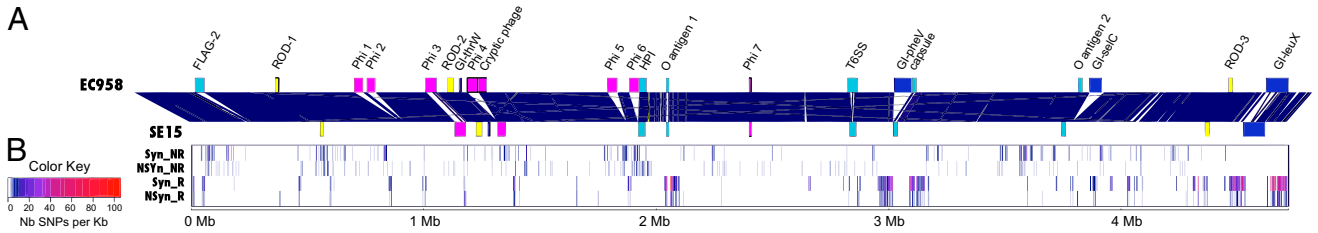

Fig. 2. Distribution of ST131-only core SNPs in recombinant versus nonrecombinant regions. (*A*) Comparison of the linear genome arrangement of the clade C strain EC958 (*Upper*) and the clade A strain SE15 (*Lower*). Solid dark-blue lines between EC958 and SE15 indicate BLAST match of ≥99% nucleotide identity between the two genomes. Genomic features of interest are highlighted for both strains as follows: prophages (pink); ST131 characteristic ROD1, ROD2, and ROD3 (yellow); previously characterized genomic islands (blue); and other regions of interest (turquoise). Labels refer to the ST131-characteristic regions defined in the genome of EC958 (15). (*B*) Heatmap showing the density of 16,424 ST131-only core SNPs along the SE15 chromosome: Syn_NR (synonymous, nonrecombinant); NSyn_NR (nonsynonymous, nonrecombinant); Syn_R (synonymous, recombinant); and NSyn_R (nonsynonymous, recombinant). ST131-only core SNPs were defined as bases called from the mapping data in all strains of the dataset with polymorphisms specific to the ST131 lineage. Recombinant region coordinates were delineated by using BratNextGen. The SNP density heatmap with (number of SNPs per 1 kbp nonoverlapping bin) is indicated by the color key. The *x* axis at the bottom of the figure represents the SE15 reference chromosome coordinates.

been independently acquired several times or that it has translocated between different plasmids or the chromosome. Both scenarios are consistent with previous reports of different types of ST131 plasmids that harbor CTX-M-15 (37).

***E. coli* ST131 Contains Many ExPEC-Associated Genes.** The complement of virulence-associated genes was determined in the 95 *E. coli* ST131 strains by examining for the presence of genes encoding chaperone-usher (CU) fimbriae (38), autotransporter (AT) proteins (39), siderophore receptors (40), toxins, colicins (41), and other genes often assessed by PCR in ExPEC (1) (*SI Appendix*, Fig. S6). The *E. coli* ST131 strains contained genes encoding type 1, Mat (ECP), Yde, ECSF_0166, EC958_4610, and Yeh fimbriae; other CU fimbriae genes including Afa and P fimbriae were variable. The complement of AT-encoding genes was highly conserved, with most strains containing genes encoding antigen 43, UpaB, UpaC, YfaL, and Sat. The ECSF_4014 AT gene was uniquely present in *E. coli* ST131 strains. Most *E. coli* ST131 strains contained a number of genes associated with iron acquisition; of note, the Yersiniabactin receptor (ECSF_1835) was found to be widely prevalent but highly diverse with 17 independent substitutions (14 nonsynonymous) confined to clade B and C strains, strongly suggesting that, like *fimH,* this gene may be under positive selection. Approximately 15% of *E. coli* ST131 strains contained genes encoding the HlyA and Cnf1 toxins. In all but clade C strain S115EC, these genes were colocated on the chromosome, which is consistent with their presence on the same genomic island in other ExPEC strains such as CFT073. In general, 131 UPEC-specific genes present in CFT073, UTI89, 536, and F11 (42) were also conserved, with only the gene encoding the putative regulator c0765 absent from all ST131 strains.

Diversity Within the ST131 Lineage Is Primarily due to Mobile Genetic Elements and Recombination of Associated Regions. *E. coli* ST131 strain EC958 contains several mobile genetic elements (MGEs) and other genomic regions not found in completely sequenced non-ST131 UPEC strains (i.e., CFT073, 536, UTI89, UMN026, IAI39), including seven prophage elements (Phi1-7), the Flag-2 lateral flagellar locus, the O-antigen loci, the *ratA*-like toxin encoding gene, the type VI secretion locus, the capsular locus, and four genomic islands (GI) in chromosomal integration hotspots (GI-*pheV*, GI-*selC*, GI-*leuX*, and GI-*thrW*) (15). The majority of these regions were highly conserved in strains from clade C but were fully or partly absent in strains from clades A and B (Fig. 3 and *SI Appendix*, Fig. S6). Exceptions included the Phi6, GI-*selC*, and capsular loci regions, which were not exclusively associated with a particular clade, suggesting a more complex evolutionary history. The Flag-2 locus was completely absent in strains from clade A and in four Spanish strains of clade B (HVM277, HVM52, HVM2044, HVM2289), replaced by the *fhiA-mbhA* scar found in *E. coli* K12 strains (43). Interestingly, these four Spanish strains form a discrete sublineage (Fig. 1*B*) and also lack prophage and genomic islands that are present in other clade B strains. In contrast to the O25b serotype of most clade B and C strains, the LPS core biosynthesis region (specifically *wbbJ-rfbE*) of clade A strains was the same as in SE15, which has been reported as serotype O150 (13). Three regions of difference (ROD) > 10 kb in length were shown to be unique in ST131 strains EC958 and SE15 (15). Although the functions of genes encoded by ROD1 are unclear, ROD1 is conserved in all ST131 strains. Similarly, ROD2 (which contains several sugar metabolism genes) was also ubiquitous but contained deletions in at least three ST131 strains (*SI Appendix*, Fig. S7). ROD3 is also conserved across all ST131 strains except for clade A. The absence of several regions within the NA114 genome that are otherwise present in closely related clade C strains such as S97EC (Fig. 3) is consistent with the assembly of this genome, which was performed by concatenating ordered contigs to produce a single pseudomolecule without gap closure and finishing (14).

Fig. 3. Selected regions of interest in ST131 strains. ST131-characteristic regions previously defined in the genome of EC958 (15) are shown along the *x* axis with strain identifiers listed on the *y* axis according to the phylogenetic tree order displayed in Fig. 1*A*. Regions A–M are shown to scale in order of their location relative to the SE15 chromosome (Fig. 2) and correspond to: A, flag2 flagellar region (38.1 kb); B, GI-ThrW genomic island; C, ROD1; D, Phi1 prophage; E, Phi2 prophage; F, Phi3 prophage; G, ROD2; H, Phi4 prophage; I, Phi5 prophage; J, Phi6 prophage; K, High-Pathogenicity Island; L, cryptic prophage; M, O-antigen 1 region (*wbbJ-rfbE*); N, Phi7 prophage; O, RatA-like region; P, T6SS region; Q, GI-PheV genomic island; R, capsule region; S, O-antigen2 region; T, GI-SelC genomic island; U, ROD3; V, GI-LeuX. Black shading indicates a match of ≥95% nucleotide identity in a minimum window of 200 bp calculated by comparing the query sequence to the assembled contigs or the consensus from mapped reads for each strain, as implemented in seqfindr (http://github.com/mscook/seqfindr).

A total of 137 regions were defined as recombinant within the ST131 lineage (Dataset S3), with a clear propensity to be located adjacent to predicted MGEs (*SI Appendix*, Fig. S3). These recombinant regions totalled 0.94 Mb, or nearly one-fifth of the entire *E. coli* SE15 genome and include the aforementioned *fimH* and *parC* genes, which are found on recombinant regions Rec_137 (92.3 kb) and Rec_089 (18.5 kb), respectively. Although the majority of regions are less than 1,000 bp in length, ~80% of the recombinant bases are contained within 24 large recombinant regions that range in size from 10.2 kb to 166.2 kb. We could define the lineage on which the recombination event occurred in the majority of cases; however, the larger fragments, such as Rec_088, Rec_089, and Rec_137, have a more complex evolutionary history with evidence for multiple blocks of different origin, reflecting sequential, overlapping recombination events within the same region (*SI Appendix*, Fig. S3 and Dataset S3). When considering the repertoire of recombinant regions that distinguish each clade, clade A was the most distant, with a total of 0.52 and 0.6 Mb differing from clade B and C, respectively.

MICROBIOLOGY

Fewer recombinant regions distinguish clade B from clade C, with the majority of differences contained in regions upstream of Phi3, and upstream and downstream of GI-*pheV* and GI-*leuX*, respectively (Dataset S3).

A striking feature of the recombination distribution along the chromosome is that the majority of large recombinant regions were associated with the sites of insertion of prophage and genomic island MGEs (*SI Appendix*, Fig. S3). Statistical evaluation of 10,000 replicates of the Kolmogorov–Smirnov test confirmed that the distribution of the observed distances between recombinant regions and MGEs was significantly negatively skewed compared with randomly selected regions (K–S test, mean D = 0.370, SD = 0.049, mean $P = 6.082 \times 10^{-6}$, SD = 8.004×10^{-5}) (*SI Appendix*, Fig. S8). This phenomenon has been observed in a comparison using the *E. coli* ST131 SE15 and NA114 genomes, for which our analysis agrees with 20 of 22 recombination regions (23), and in other comparisons of closely related *E. coli* genomes (44). In contrast, a reduced role for recombination was reported in a study comparing 12 ST131 genomes and 50 publicly available *E. coli* reference genomes (45).

Recombinant Regions Have Shaped the ST131 Lineage. Fimbrial adhesins and bacterial motility genes were significantly overrepresented in recombinant fragments (*SI Appendix*, Fig. S9). A prime example was the *fliC-fliY* flagellar locus encoded on the recombinant fragment Rec_051 (ECSF_1762 to ECSF_1776). In SE15 and other ST131 clade A strains, the *fliC* allele corresponds to the H5 serogroup. In contrast, clade B and C strains possess an H4 *fliC* allele within Rec_051, a 12.6-kb recombinant fragment that is adjacent to the Phi5 insertion in EC958. The *fim* operon containing the type 1 fimbrial *fimH* gene resides within Rec_137, which at 92.6 kb is one of the largest and most complex recombinant fragments within the ST131 lineage (Dataset S3). The subfragment of Rec_137 that encodes the region *fimC* to *uxuR* displayed characteristic recombination patterns, introducing a clade-specific *fimH allele* (*fimH*41 in clade A, *fimH*22 in clade B, and *fimH*30 in clade C). Interestingly, these three *fimH* alleles were also identified as the major signatures in a small collection of mainly American isolates, and the same recombinant region was deduced from the comparison of SE15 and NA114 (23). As observed in EC958 (15), an insertion within *fimB* was found in clade C strains, although it is not clear if this insertion was acquired by homologous recombination concomitant with the acquisition of the *fimH*30 subfragment, or subsequent to this event. The only exception in our collection was the ST131 clade C strain S77EC, which contained a large deletion encompassing part of the 3′ end of the adjacent GI-*leuX* island (Fig. 3) and the *fim* locus.

Several regions containing putative virulence genes, namely Rec_087 (ECSF_2626 to ECSF_2634) and part of Rec_088 (ECSF_2784 to ECSF_2804), which contain genes related to a Type 6 Secretion System (T6SS) and a Type 2 Secretion System (T2SS), respectively, have also undergone gene conversion. Clade B and C strains carry T6SS alleles that are distinct from clade A strains. In contrast, the T2SS locus in clade C strains appears to have been subjected to several independent recombination events, consistent with its location in a recombination hotspot downstream of the GI-*pheV* island (*SI Appendix*, Fig. S3). Between the T2SS region and GI-*pheV*, the Rec_088 recombinant fragment also encodes the group II capsule synthesis locus (ECSF_2771 to ECSF_2783). Several variant region 2 gene clusters were observed between region 1 (*kpsFEDUCS*) and region 3 (*kpsTM*) of ST131 genomes, consistent with multiple instances of replacement since divergence of ST131 clades A and C with corresponding differences in K-antigen serotype (46). As described above, differences in the LPS core biosynthesis locus within the 70.3-kb Rec_069 recombinant fragment suggest that the O25b serotype is also associated with divergence of clades B and C from clade A (13).

Several less-well characterized genomic regions that could differentiate clade C strains from other ST131 strains were also identified. Two regions with the most distinctive recombination profiles that clearly distinguished all three clades were Rec_131 (ECSF_4099 to ECSF_4159) upstream of GI-*leuX*, and part of Rec_137 (ECSF_4277 to ECSF_4338). The Rec_131 region contains the *tamAB* genes, which encode a recently described translocation and assembly module that contributes to the secretion of some AT proteins (47), whereas the Rec_137 region contains genes associated with salt resistance (*osmY*), siderophore-based iron transport (*fhuF*), and regulation (*creBC*). When the impact of recombination on major gene functions independent of virulence was considered, significant differences were observed in genes encoding transporters, fructose-mannose metabolism, histidine metabolism, and the pentose-phosphate pathway (*SI Appendix*, Fig. S9). The impact of these sequence changes remains to be determined.

Conclusion

Our whole-genome phylogenetic analysis indicates that ST131 has arisen from a single progenitor *E. coli* that diverged into three sublineages some time before the year 2000 with acquisition of multiple mobile genetic elements, associated recombination events, and point-mutations jointly responsible for the emergence of the most prevalent clade C strains. In addition to the known *fimH*, *fimB*, *parC*, and *gyrA* alleles that characterize ST131 clade C, we have defined several additional genes and regions that may be important for adaptive diversification in response to host or antibiotic resistance pressures. These results also provide a framework for future PCR-based assays to rapidly classify ST131 strains and monitor their evolution. Further molecular analysis of the clade defining variants and MGEs identified in this study will help to elucidate the mechanisms that have led to ST131 colonization of the urinary tract and other clinical sites, and the rapid global dispersal of this important group of ExPEC.

Materials and Methods

Genome Sequencing and Assembly. Draft genomes were generated by using 100-bp paired-end Illumina HiSeq 2000 reads and assembled with Velvet (48). Contigs ≥200 bp were ordered against the EC958 draft genome (BioProject: PRJEA61443) by using Mauve (49). Sequencing reads are available at the European Nucleotide Archive under study number ERP001354, accessions in Study ERP001354 (ERS126551–ERS126646) (see Dataset S1 for accession numbers) with draft genomes available at http://github.com/BeatsonLab-MicrobialGenomics/ST131_99. See also *SI Appendix, SI Materials and Methods*.

Genome Analysis. Alignment of the ST131 draft genome assemblies and three ST131 reference genomes (SE15, NA114, and EC958), plus completely sequenced non-ST131 genomes belonging to the *E. coli* B2 phylogenetic group (CFT073, UTI89, E2348/69, ED1a, 536, S88, APEC O1), was performed by using Mugsy (50) and GBLOCKS (51) with a minimum syntenic block of 5 kbp. Recombination in the ST131 sequences was estimated by using BratNextGen, which implements a Bayesian clustering algorithm for detection of recombinant fragments in closely related sequences (34). See also *SI Appendix, SI Materials and Methods*.

Read Mapping and SNP Analysis. Reads from each ST131 isolate and reads simulated *in silico* for the 10 complete genomes used in this study were mapped onto the reference genome SE15 (16) by using SHRIMP 2.0 (52). Nesoni (www.vicbioinformatics.com/software.nesoni.shtml) was used to perform SNP calling (conservative default parameters), small indel prediction, and coding effect SNP annotation. In addition, the Nesoni *n-way* pairwise comparison method was used to establish the list of all polymorphic positions conserved in all strains of the dataset. Polymorphic substitution-only sites were concatenated to produce an alignment that was used for phylogenetic tree construction. Analysis and visualization of SNP distribution across the collection were performed by using custom R scripts. See also *SI Appendix, SI Materials and Methods*.

Phylogeny. ML phylogenetic trees were estimated by using RAxML 7.2.8 (53) for the inferred core genome and the SNP alignments (prerecombination and postrecombination filtering) under the GTR nucleotide substitution model with a gamma correction for ASRV. Recombination filtering was performed by collapsing the recombinant segment boundaries predicted for each strain into a unique list of 137 nonoverlapping segments and subsequently

masking these regions from the alignment (Dataset S3). Support for nodes was assessed by using 1,000 random bootstrap replicates. See also *SI Appendix, SI Materials and Methods*.

Comparative Genomics. Virulence factor profiles, and the presence of other regions in the draft genomes, were visualized by using seqfindr (http://github.com/mscook/seqfindr). Query sequences and their source are listed in Dataset S1 and with sequences available at http://github.com/BeatsonLab-MicrobialGenomics/ST131_99/. Comparisons between individual genomes and verification of seqfindr results were performed by using BLAST (54), Artemis Comparison Tool (55), Easyfig (56), and BRIG (57). See also *SI Appendix, SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank Deborah Williamson, Karina Kennedy, and Hanna Sidjabat for their contribution of strains. We thank the Wellcome Trust Sanger Institute and Australian Genome Research Facility for sequencing. This work was supported by Australian National Health and Medical Research Council (NHMRC) Grants APP1012076 and APP1067455 (to M.A.S. and S.A.B.). N.K.P. and E.S. were supported by the Australian Infectious Diseases Research Centre. M.A.S. is supported by Australian Research Council (ARC) Future Fellowship FT100100662. M.T. is supported by ARC Discovery Early Career Researcher Award DE130101169. M.D. is supported by NHMRC Training Fellowship GNT0635250. J.R.-B. and A.P. are supported by the Ministerio de Economía y Competitividad, Instituto de Salud Carlos III, the Spanish Network for Research in Infectious Diseases RD12/0015; Fondo de Investigación Sanitaria Grants 070190, 10/02021, 10/0195, and 10/00795; and Junta de Andalucía Grants 0048/2008 and CTS-5259.

1. Nicolas-Chanoine MH, et al. (2008) Intercontinental emergence of *Escherichia coli* clone O25:H4-ST131 producing CTX-M-15. *J Antimicrob Chemother* 61(2):273–281.
2. Coque TM, et al. (2008) Dissemination of clonally related *Escherichia coli* strains expressing extended-spectrum beta-lactamase CTX-M-15. *Emerg Infect Dis* 14(2): 195–200.
3. Lau SH, et al. (2008) Major uropathogenic *Escherichia coli* strain isolated in the northwest of England identified by multilocus sequence typing. *J Clin Microbiol* 46(3): 1076–1080.
4. Rogers BA, Sidjabat HE, Paterson DL (2011) *Escherichia coli* O25b-ST131: A pandemic, multiresistant, community-associated strain. *J Antimicrob Chemother* 66(1):1–14.
5. Johnson JR, et al. (2010) *Escherichia coli* sequence type ST131 as an emerging fluoroquinolone-resistant uropathogen among renal transplant recipients. *Antimicrob Agents Chemother* 54(1):546–550.
6. Uchida Y, et al. (2010) Clonal spread in Eastern Asia of ciprofloxacin-resistant *Escherichia coli* serogroup O25 strains, and associated virulence factors. *Int J Antimicrob Agents* 35(5):444–450.
7. Peirano G, Schreckenberger PC, Pitout JD (2011) Characteristics of NDM-1-producing *Escherichia coli* isolates that belong to the successful and virulent clone ST131. *Antimicrob Agents Chemother* 55(6):2986–2988.
8. Morris D, et al. (2012) Detection of OXA-48 carbapenemase in the pandemic clone *Escherichia coli* O25b:H4-ST131 in the course of investigation of an outbreak of OXA-48-producing *Klebsiella pneumoniae*. *Antimicrob Agents Chemother* 56(7):4030–4031.
9. Lau SH, et al. (2008) UK epidemic *Escherichia coli* strains A-E, with CTX-M-15 beta-lactamase, all belong to the international O25:H4-ST131 clone. *J Antimicrob Chemother* 62(6):1241–1244.
10. Johnson JR, et al.; MASTER Investigators (2012) Comparison of *Escherichia coli* ST131 pulsotypes, by epidemiologic traits, 1967-2009. *Emerg Infect Dis* 18(4):598–607.
11. Banerjee R, et al. (2013) *Escherichia coli* sequence type 131 is a dominant, antimicrobial-resistant clonal group associated with healthcare and elderly hosts. *Infect Control Hosp Epidemiol* 34(4):361–369.
12. Johnson JR, et al. (2013) Abrupt emergence of a single dominant multidrug-resistant strain of *Escherichia coli*. *J Infect Dis* 207(6):919–928.
13. Toh H, et al. (2010) Complete genome sequence of the wild-type commensal *Escherichia coli* strain SE15, belonging to phylogenetic group B2. *J Bacteriol* 192(4):1165–1166.
14. Avasthi TS, et al. (2011) Genome of multidrug-resistant uropathogenic *Escherichia coli* strain NA114 from India. *J Bacteriol* 193(16):4272–4273.
15. Totsika M, et al. (2011) Insights into a multidrug resistant *Escherichia coli* pathogen of the globally disseminated ST131 lineage: Genome analysis and virulence mechanisms. *PLoS ONE* 6(10):e26578.
16. Clark G, et al. (2012) Genomic analysis uncovers a phenotypically diverse but genetically homogeneous *Escherichia coli* ST131 clone circulating in unrelated urinary tract infections. *J Antimicrob Chemother* 67(4):868–877.
17. Gibreel TM, et al. (2012) Population structure, virulence potential and antibiotic susceptibility of uropathogenic *Escherichia coli* from Northwest England. *J Antimicrob Chemother* 67(2):346–356.
18. Coelho A, et al. (2011) Spread of *Escherichia coli* O25b:H4-B2-ST131 producing CTX-M-15 and SHV-12 with high virulence gene content in Barcelona (Spain). *J Antimicrob Chemother* 66(3):517–526.
19. Johnson JR, Johnston B, Clabots C, Kuskowski MA, Castanheira M (2010) *Escherichia coli* sequence type ST131 as the major cause of serious multidrug-resistant *E. coli* infections in the United States. *Clin Infect Dis* 51(3):286–294.
20. Lavigne JP, et al. (2012) Virulence potential and genomic mapping of the worldwide clone *Escherichia coli* ST131. *PLoS ONE* 7(3):e34294.
21. Clermont O, et al. (2008) The CTX-M-15-producing *Escherichia coli* diffusing clone belongs to a highly virulent B2 phylogenetic subgroup. *J Antimicrob Chemother* 61(5):1024–1028.
22. Johnson JR, Porter SB, Zhanel G, Kuskowski MA, Denamur E (2012) Virulence of *Escherichia coli* clinical isolates in a murine sepsis model in relation to sequence type ST131 status, fluoroquinolone resistance, and virulence genotype. *Infect Immun* 80(4):1554–1562.
23. Paul S, et al. (2013) Role of homologous recombination in adaptive diversification of extraintestinal *Escherichia coli*. *J Bacteriol* 195(2):231–242.
24. Totsika M, et al. (2013) A FimH inhibitor prevents acute bladder infection and treats chronic cystitis caused by multidrug-resistant uropathogenic *Escherichia coli* ST131. *J Infect Dis* 208(6):921–928.
25. Floyd RV, et al. (2012) *Escherichia coli*-mediated impairment of ureteric contractility is uropathogenic *E. coli* specific. *J Infect Dis* 206(10):1589–1596.
26. Phan MD, et al. (2013) The serum resistome of a globally disseminated multidrug resistant uropathogenic *Escherichia coli* clone. *PLoS Genet* 9(10):e1003834.
27. Clermont O, et al. (2009) Rapid detection of the O25b-ST131 clone of *Escherichia coli* encompassing the CTX-M-15-producing strains. *J Antimicrob Chemother* 64(2):274–277.
28. Croucher NJ, et al. (2011) Rapid pneumococcal evolution in response to clinical interventions. *Science* 331(6016):430–434.
29. Harris SR, et al. (2010) Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 327(5964):469–474.
30. Holt KE, et al. (2012) *Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat Genet* 44(9):1056–1059.
31. Martincorena I, Seshasayee AS, Luscombe NM (2012) Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature* 485(7396):95–98.
32. Touchon M, et al. (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet* 5(1):e1000344.
33. Castillo-Ramírez S, et al. (2012) Phylogeographic variation in recombination rates within a global clone of methicillin-resistant *Staphylococcus aureus*. *Genome Biol* 13(12):R126.
34. Marttinen P, et al. (2012) Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res* 40(1):e6.
35. Li Y, et al. (2010) *Escherichia coli* condensin MukB stimulates topoisomerase IV activity by a direct physical interaction. *Proc Natl Acad Sci USA* 107(44):18832–18837.
36. Cantón R, Coque TM (2006) The CTX-M beta-lactamase pandemic. *Curr Opin Microbiol* 9(5):466–475.
37. Woodford N, et al. (2009) Complete nucleotide sequences of plasmids pEK204, pEK499, and pEK516, encoding CTX-M enzymes in three major *Escherichia coli* lineages from the United Kingdom, all belonging to the international O25:H4-ST131 clone. *Antimicrob Agents Chemother* 53(10):4472–4482.
38. Wurpel DJ, Beatson SA, Totsika M, Petty NK, Schembri MA (2013) Chaperone-usher fimbriae of *Escherichia coli*. *PLoS ONE* 8(1):e52835.
39. Wells TJ, Totsika M, Schembri MA (2010) Autotransporters of *Escherichia coli*: A sequence-based characterization. *Microbiology* 156(Pt 8):2459–2469.
40. Alteri CJ, Mobley HL (2007) Quantitative profile of the uropathogenic *Escherichia coli* outer membrane proteome during growth in human urine. *Infect Immun* 75(6):2679–2688.
41. Bruant G, et al. (2006) Development and validation of an oligonucleotide microarray for detection of multiple virulence and antimicrobial resistance genes in *Escherichia coli*. *Appl Environ Microbiol* 72(5):3780–3784.
42. Lloyd AL, Rasko DA, Mobley HL (2007) Defining genomic islands and uropathogen-specific genes in uropathogenic *Escherichia coli*. *J Bacteriol* 189(9):3532–3546.
43. Ren CP, Beatson SA, Parkhill J, Pallen MJ (2005) The Flag-2 locus, an ancestral gene cluster, is potentially associated with a novel flagellar system from *Escherichia coli*. *J Bacteriol* 187(4):1430–1440.
44. Didelot X, Méric G, Falush D, Darling AE (2012) Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genomics* 13:256.
45. McNally A, Cheng L, Harris SR, Corander J (2013) The evolutionary path to extra-intestinal pathogenic, drug-resistant *Escherichia coli* is marked by drastic reduction in detectable recombination within the core genome. *Genome Biol Evol* 5(4):699–710.
46. Whitfield C (2006) Biosynthesis and assembly of capsular polysaccharides in *Escherichia coli*. *Annu Rev Biochem* 75:39–68.
47. Selkrig J, et al. (2012) Discovery of an archetypal protein transport system in bacterial outer membranes. *Nat Struct Mol Biol* 19(5):506–510.
48. Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18(5):821–829.
49. Darling AE, Mau B, Perna NT (2010) progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* 5(6):e11147.
50. Angiuoli SV, Salzberg SL (2011) Mugsy: Fast multiple alignment of closely related whole genomes. *Bioinformatics* 27(3):334–342.
51. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17(4):540–552.
52. David M, Dzamba M, Lister D, Ilie L, Brudno M (2011) SHRiMP2: Sensitive yet practical SHort Read Mapping. *Bioinformatics* 27(7):1011–1012.
53. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688–2690.
54. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402.
55. Carver T, et al. (2008) Artemis and ACT: Viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* 24(23):2672–2676.
56. Sullivan MJ, Petty NK, Beatson SA (2011) Easyfig: A genome comparison visualizer. *Bioinformatics* 27(7):1009–1010.
57. Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA (2011) BLAST Ring Image Generator (BRIG): Simple prokaryote genome comparisons. *BMC Genomics* 12:402.