

Abstract

Escherichia coli ST131 is the most frequently isolated fluoroquinolone resistant (FQR) *E. coli* clone worldwide and a major cause of urinary tract and bloodstream infections. Although originally identified through its association with the CTX-M-15 extended-spectrum β -lactamase resistance gene, global genomic epidemiology studies have failed to resolve the geographical and temporal origin of the ST131 ancestor. Here, we developed a framework for the reanalysis of publicly available genomes from different sources and used this dataset to reconstruct the evolutionary steps that led to the emergence of FQR ST131. Using Bayesian estimation, we show that point mutations in chromosomal genes that confer FQR coincide with the first clinical use of fluoroquinolone in 1986, and illustrate the impact of this pivotal event in the rapid population expansion of ST131 worldwide from an apparent origin in North America. Furthermore, we identify key virulence factor acquisition events that predate the development of FQR, suggesting that the gain of virulence-associated genes followed by the tandem development of antibiotic resistance primed the successful global dissemination of ST131.

Keywords: Mobile Genetic Elements, Sequence Type 131, genomic islands, phylogenomic

INTRODUCTION:

Escherichia coli sequence type 131 (ST131) is a recently emerged multidrug resistant clone associated with urinary tract and bloodstream infections. *E. coli* ST131 was originally identified due to its strong association with the CTX-M-15-type extended-spectrum-beta-lactamase (ESBL) allele (1), and is now the predominant fluoroquinolone resistant (FQR) *E. coli* clone world-wide (2-4).

ST131 belongs to subgroup 1 from *E. coli* phylogroup B2, with most strains of serotype O25b:H4 (1-4). Two previous genomic studies have explored the ST131 clonal structure (2, 5) and identified a globally dominant FQR sublineage defined as clade C (2) or H30-R (5). Two additional well-supported ST131 sublineages, referred to as clades A and B, have also been described (2). Each of these clades contain a defined marker allele for the type 1 fimbriae *fimH* adhesin; H41 (clade A), H22 (clade B) and H30 (clade C) (6). Further analysis of clade C/H30-R ST131 identified a smaller subset of strains containing the *bla*_{CTX-M-15} ESBL allele referred to as clade C2 or H30-Rx (2, 5). The ST131 strain EC958 is a reference FQR clade C strain that has been well-characterised at the genomic and phenotypic level (3, 7-12).

Several early studies demonstrated variation in the complement of virulence genes in ST131, with only a few virulence factors consistently identified in all strains (1, 4, 13-15). Our comprehensive analysis of 95 *E. coli* ST131 genomes revealed that the virulence and mobile genetic element (MGE) profile was in fact consistent with the phylogenetic structure of the ST131 lineage, with clade C strains sharing a generally conserved set of genes. In contrast, the plasmid profile of ST131 is highly disparate

with multiple different replicons found in closely related strains and multiple genomic contexts for the clade C2 defining *bla*_{CTX-M-15} ESBL gene (16, 17).

Despite its successful dissemination globally, little information is available about evolution and emergence of ST131. Both recent independent genomic studies demonstrated that ST131 emerged from a single ancestor and that most strains belong to clade C/H30-R (2, 5). Notably, we found that recombination accounted for the majority of variation within the ST131 lineage and recombination events were associated with the positions of MGEs (2). However, despite the number of isolates in both studies, neither resolved the geographical or temporal origin of the ST131 ancestor. In contrast, studies of other large sets of bacteria with geographical or temporal separation have determined accurate dates of divergence of important clades using statistical analyses such as the Bayesian framework implemented in BEAST (Bayesian Evolutionary Analysis by Sampling Trees) (18). For example, Glaser identified tetracycline resistance as the major driver of diversification amongst the global population of Group B *Streptococci* (19). Similarly, a large study of Methicillin resistant *Staphylococcus aureus* (MRSA) was able to date the emergence of a FQR clade to the mid-1980's (20). These studies motivated us to combine data sets from our geographically diverse previous study (2) and from the temporally diverse Price et al study (5) to investigate the evolution of ST131 with the highest possible resolution.

RESULTS AND DISCUSSION:

Curation of a high-quality ST131 genome sequence dataset.

We first sought to obtain a high-quality set of data to carry out our analyses. 199 draft Illumina paired-end *E. coli* ST131 genomes were retrieved from public read data repositories (Dataset S1). Initial phylogenetic analyses of *de novo* and reference-guided assemblies of all 199 *E. coli* ST131 genomes indicated that several draft genomes were of low quality. Suboptimal genome data quality could interfere in subsequent phylogenetic analyses and may invalidate conclusions drawn from tree topologies. To ensure that only high quality sequences were included in our analyses, we removed 14 genome datasets that were determined to be outliers according to at least one of our assembly or mapping metrics, including number of uncalled bases, number of scaffolds and assembled genome size (Supplementary Fig. S1). This quality control filter is broadly applicable to reanalyzing public genomic data from multiple sources.

Phylogenomic analysis of ST131.

We next carried out phylogenetic reconstruction using our combined dataset of 185 Illumina paired end sequences, which represented strains from humans (n=167), animals (n=15) or other (n=3) sources isolated from USA, Canada, New Zealand, Australia, Spain, India and UK between 1967 and 2011 (Supplementary Table 1; Fig. 1). Sequence read mapping of these 185 high-quality ST131 genomes (and simulated reads from the SE15, EC958 and JJ1886 complete genomes) to the ST131 clade C reference genome *E. coli* EC958 defined 21,373 substitution Single Nucleotide Polymorphisms (SNPs) that were used to create an unrooted Maximum Likelihood phylogeny (Supplementary Fig. 2a). An independent phylogenetic tree produced with the kSNP alignment-free method was consistent with the overall topology of the ML trees (Supplementary Fig. 3). Using a Bayesian modeling algorithm, we identified 204

non-overlapping segments encompassing 1.542 Mb and containing 15,902 substitution SNPs that were introduced into the ST131 lineage by recombination (Supplementary Fig. 4 and 5; Supplementary Table 2). The length of recombinant sequence is higher than previously reported (2) as the larger dataset increases the probability that one strain will have a recombinant fragment not encountered before. However, the proportion of SNPs introduced by recombination (74.4%) is consistent with our previous study and highlights the important role recombination has played in shaping the ST131 lineage (2). Exclusion of these recombinant SNPs from phylogenetic analyses reduced the number of SNPs to 5,471 and resulted in a tree that maintained the original overall topology, albeit with substantially reduced branch lengths and some major within clade re-clustering of strains (Supplementary Fig. 2b).

Extending our phylogenomic analyses to include isolates from two large international collections provided a far greater resolution of the evolution within the ST131 lineage (Fig. 2). The global phylogeny of *E. coli* ST131 separated the strains into three distinct lineages (clades A, B and C). Congruent with our previous work, strains in clade C were characterized by the *fimH*-30 allele and the FQR conferring alleles *gyrA*-1AB and *parC*-1aAB (Fig. 2). Notable exceptions were strains JJ2643 and U004 in clade C that contain the *fimH*-35 allele. This appeared to be due to a recombination event encompassing *fimH* in these strains, and highlights why we have retained a neutral nomenclature (i.e. A, B, C) for our clade classifications. Likewise, the CTX-M-15 allele is not ubiquitous in all clade C2 strains, making this a more scalable classification than the *H30*-Rx designation originally suggested by Price et al (5) (Fig. 2). In addition to harbouring CTX-M-15 genes, clade C2 strains contain more resistance genes in total compared with other ST131 clades (Fig. 3), consistent with

co-localization of multiple plasmid-encoded resistance genes (Supplementary Fig. 6). Although the context of multidrug resistance cassettes can be resolved in some cases from draft genome data of ST131 isolates or transformants (16, 17), the full complexity of plasmid-mediated resistance in ST131 requires the generation of more complete genomes as per EC958 and JJ1886 (7, 21).

Combined dataset enables greater resolution of ST131 subclades.

The combined ST131 dataset enabled greater resolution of the differences between Clade B and C strains. We previously showed that Clade C strains can be further segregated into two distinct subclades, C1 and C2 (2). Our new analysis defined five discrete sub-clades in clade B (B1 to B5), each with distinct repertoires of selected marker genes *fimH*, *parC* and *gyrA* (Fig. 2). Notably, strains from the five clade B sublineages varied in their *parC* and *gyrA* allelic profile, with the vast majority of clade B strains containing allele combinations that are not associated with fluoroquinolone resistance. Additionally, while all strains from subclade B2 and B5 are associated with the USA, and B1 with Spain, strains within subclade B3 and B4 have a more diverse geographic origin. Each subclade showed a distinctive recombination profile (Supplementary Fig. S4 and S5) and MGE repertoire (Supplementary Fig. 7), indicative of independent evolutionary trajectories. In contrast, we found that the prevalence of virulence genes is largely conserved across all B subclades, with the absence of several UPEC-specific genes apparent in clade B3 (Supplementary Fig. 8). By comparison, our investigation of clade C MGEs and other regions of interest (as originally defined in the clade C reference strain EC958) showed a high degree of conservation across clade C, with the exception of the prophage Phi6, the capsule loci and *GI-selC* (Supplementary Fig. 7). For example, GI-

selC is only found in a geographically homogeneous cluster of clade C strains that include EC958 and excludes the reference strain JJ1886 (Supplementary Fig. S7, Fig. 2). Despite the general conservation of gene content within clade C genomes, it is apparent that genomic islands are hotspots for insertions, deletions (indels) and genome rearrangement (Supplementary Table 3). EC958 GI-*pheV* has several small indels relative to JJ1886 GI-*pheV*, and we have previously shown that the CMY-23 beta-lactamase gene that confers resistance to third generation cephalosporins has inserted within the EC958 GI-*leuX* whereas the JJ1886 GI-*leuX* element has a large duplication relative to EC958 (12) (Supplementary Table 3).

Temporal analysis of ST131 identifies major divergence dates.

Our initial studies of ST131 strains collected between 2001 and 2011 showed insufficient temporal depth to robustly date the emergence of clade C (2). By including 91 more strains from Price et al., (2013) (5), including 8 that predated 2000 (Fig. 1), we anticipated that we would be able to resolve this question using existing public data alone. We generated a linear regression of the genetic distance from the root-to-tip against time for the 172 ST131 isolates within clades B and C using Path-O-Gen (22). This analysis revealed a positive correlation ($R^2 = 0.3233$, $P < 0.0001$) confirming the molecular clock-like signal (Supplementary Fig. 9). To accurately estimate the date of divergence of clade C from clade B we employed BEAST (18). BEAST analysis rejected the strict clock and favored the uncorrelated log-normal clock model in combination with a Bayesian skyline population model (Supplementary Table 4). A mutation rate of 4.39×10^{-7} SNPs per site per year (95% highest posterior density (HPD) = $3.58 - 5.23 \times 10^{-7}$) was calculated, consistent with other large-scale phylogenomic analyses of the *E. coli/Shigella* lineage (6.0×10^{-7}

SNPs per site per year (95% HPD = $5.2 - 6.7 \times 10^{-7}$) (23) (Supplementary Fig. S10). Based on this approach we could estimate the divergence of the last common ancestor of clade B and C strains to have occurred between 1930 and 1958 (Fig. 4A), consistent with the Path-O-Gen prediction (Supplementary Fig. S9a). We could estimate the divergence of clade C from clade B to have occurred in 1980 (95% HPD 1973-1986), which was slightly earlier than the Path-O-Gen prediction (Supplementary Fig. S9a). Importantly, we identified that further diversification of clade C2 from clade C1 dated to 1987 (95% HPD 1983-1992), subsequent to all clade C1 and C2 strains acquiring *gyrA*-1AB and *parC*-1aAB alleles imparting elevated FQR (Fig. 4a). Bayesian skyline plots show a relatively constant population size over several decades, followed by a short recent expansion occurring in the late 1990's and early 2000's and subsequent stabilization (Fig. 2c). Interestingly, this pattern is consistent with the introduction of FQ for clinical use in 1986 (24) and the subsequent stabilization may reflect the improved stewardship of FQ (or its removal from general use). A similar phenomenon was observed for FQR amongst the MRSA ST22 global phylogeny (20). Remarkably, an identical date was identified in a recent pre-print report using 81 genomes from Price et al. (5), supplemented with ~100 newly sequenced genomes from more geographically dispersed strains (17), highlighting the value of careful analysis of existing datasets. Although acquisition of the CTX-M-15 gene within clade C2 may be a major factor in the diversification of C1 and C2, it is worth emphasizing that this alone does not explain the success of ST131 given that the population expansion identified in our study encompasses both clade C1 and clade C2 strains.

Phylogeography of ST131

To investigate the geographical context underlining the expansion of the multidrug-resistant ST131 O25b:H4 clone from clade B to C, we performed a discrete phylogeographical analysis as implemented in BEAST on the 172 ST131 isolates within clades B and C that included a variety of geographic sources (Fig. 1a) and dates (Fig. 1b). To reduce sampling biases due to the high number of strains isolated from USA, Canada, UK and Spain, we performed independent analyses on 10 randomly subsampled datasets containing 85 strains each (Supplementary Table 5). Under a BSSVS and symmetric diffusion models, results systematically supported the USA (74.31%, s.d. 12.1) as the most likely origin of clade B and C (Fig. 5a). The origin of clade C (C0, C1 and C2) was predicted to be associated either with USA (51.83%, s.d. 35.5) or Canada (45.59%, s.d. 36.1), overall North America (Fig. 5b). These results are consistent with the observation that the oldest reported ST131 strain was isolated in the USA in 1967⁵, and an independent BEAST analysis using a partly overlapping dataset (17). Although our resampling approach has minimized bias in strain origin, a dataset with greater diversity of strains from different geographical regions and pre-2000 isolation dates would be necessary to rule out a different origin (e.g. current datasets are under-represented in South America, Africa and many European and Asian countries). A greater number of strains would also help identify local outbreaks or clusters: with the exception of the *GI-selC* carrying clade C strains from the UK which cluster phylogenetically (Fig. 2), we did not observe other significant geographic clustering using this dataset alone.

Intermediate strains reveal key MGE acquisitions that define clade C.

Overall, excluding intermediate B0 and C0 strains, clade C differs from clade B by only 42 substitution SNPs (Supplementary Table 6). This list included the majority of

the 70 clade C defining SNPs reported in our earlier study (2), but was not identical due to the greater number of recombinant regions identified and removed in the present study. Intriguingly, we found that seven strains originating from the USA could be classified as “intermediate” on the basis of their SNP pattern (Supplementary Table 6). These strains showed progressive acquisition of clade C-defining point mutations, with the three isolates closer to clade B (classified B0) and four isolates closer to clade C (classified C0) illustrating the precise evolutionary events leading to the emergence of clade C (Fig. 2b, Supplementary Table 5). Closer examination of the recombination analysis identified intermediate patterns of recombination, primarily clustered around known MGEs, indicative of step-wise evolution among these intermediate strains (Supplementary Fig. S4 and S5). Most notably, we could trace the acquisition of GI-*pheV* and GI-*leuX* genomic islands to the most recent common ancestor of the C0 strains (C001, JJ2244, G213 and CD306), several years before the acquisition of the FQR mutations that define clade C (Fig. 4). The *pheV* genomic island acquired by clade C ST131 strains is known to carry the autotransporter genes *agn43* and *sat*, the ferric aerobactin biosynthesis gene cluster (*iucA-D*) and its cognate ferrisiderophore receptor gene *iutA* (3, 7). The clade-defining *fimH30* allele was acquired by recombination (25), possibly in conjunction with the acquisition of the nearby GI-*leuX* island; the same time-point is also predicted for the acquisition of the ISEc55 insertion sequence within the *fimB* recombinase gene that we have previously shown to affect the expression of type 1 fimbriae (3, 7). Thus, close scrutiny of these “intermediate” genomes enabled us to trace the acquisition of virulence-associated genes in ST131, which appears to have primed this clone for success prior to the acquisition of FQR mutations in the late 1980’s. Further molecular analysis is required to determine the contribution of these elements to ST131 colonization/fitness

in the gastrointestinal and urinary tracts. Notably, the role of virulence in the success of this clone may have been underappreciated in a recent report as these particular strains were distributed throughout clade B and clade C despite their inconsistent *parC*, *gyrA* and *fimH* alleles (17).

CONCLUSIONS

Overall, our work highlights how careful reanalysis of publicly available genomic datasets from heterogeneous sources can greatly improve the resolution of evolutionary history. Here, we have characterised the evolution of ST131 with unprecedented detail, from the acquisition of prophages and the O25b antigen region circa 1946, to the acquisition of GI-*pheV*, GI-*leuX*, *fimH*30 and ISEc55 around 1980, several years before the acquisition of mutations in *gyrA* and *parC* that led to FQR and the acquisition of the clade C2 defining CTX-M-15 ESBL gene. Whereas the development of FQR was accompanied by a large surge in the ST131 population globally, we propose that the acquisition of virulence factors by ST131 was a necessary precursor to this success. These events describe the ‘perfect storm’ for the evolution of a multidrug resistant pathogen; the acquisition of virulence-associated genes followed by the development of antibiotic resistance.

MATERIALS AND METHODS

Genome data

Two *E. coli* ST131 strain datasets from previously published work were used in this study, under the designation Dataset_1 and Dataset_2 (2, 5). Strain names, sources and available strain metadata are summarized in Supplementary Table 1. Dataset_1 comprised Illumina 101-basepair (bp) paired-end genome sequence data of 95 ST131 strains isolated from 2000 to 2011, mostly in Europe and Oceania (2) (study accessions ERP001354 and ERP004358). Dataset_2 comprised Illumina 101-bp (76 samples), 76-bp (19 samples) and 50-bp (10 samples) paired-end whole genome sequence data of 105 ST131 strains isolated from 1967 to 2011, mostly in North America (5) (study accession SRP027327). Additionally, reference strains of 11 published complete genomes were also used, namely *E. coli* ST131 strains SE15, JJ1886 and EC958, plus non-ST131 B2 phylogenetic group *E. coli* strains CFT073, UTI89, E2348, ED1a, 536, S88, APEC-01 and non-ST131 D phylogenetic group *E. coli* strain UMN026 (Supplementary Table 7). *E. coli* strain NA114 was excluded from the analysis due to poor assembly quality (2, 7, 26). To integrate reference genome data into our phylogenomic analyses, error-free 101-bp paired-end Illumina reads were simulated to 60 times coverage with an insert size of 340 bp +/- 40 bp as previously described (2).

Quality control, de novo genome assembly and variant detection.

Quality Control (QC) was performed for all raw read datasets. Briefly, raw reads were analyzed using PRINSEQ v0.20.3 (27) and trimmed with a mean base-pair quality score $Q \geq 20$ and a read length of $\geq 70\%$ of the original read length. Additionally, it

was necessary to correct 35 sets of raw read data from Dataset_2 that had heterogeneous Illumina encoding and/or erroneous paired-end length encoding (Supplementary Table 8). Quality control and assembly metrics for Dataset_1 have been previously reported in Petty et al. (2). Lastly, contaminant searches were performed for each sample using Kraken on a subset of 100,000 randomly chosen reads (28).

Quality filtered Illumina paired-end reads were assembled *de novo* using Velvet v1.2.07 (29) with a *k*-mer range of 45-85 for 101-bp reads, 29-61 for 76-bp reads, and 29-47 for 50-bp reads. An optimal *k*-mer value for each assembly was selected on the basis of best assembly metrics including N50 (50% of bases are incorporated in contigs of this length or above), number and size of contigs, number and continuity of uncalled bases and peak coverage. Contigs ≥ 200 bp at an optimal *k*-mer were then ordered against *E. coli* EC958 (7) using Mauve v2.3.1 (30). All QC and assembly statistics are summarized in Supplementary Table 8.

Quality filtered Illumina reads for Dataset_1 and Dataset_2, as well as error-free simulated reads of complete genomes, were mapped on the reference strain EC958 using SHRIMP v2.0 (31). Nesoni v0.108 (32) was used to call and annotate substitution-only SNPs, with a consensus cutoff and majority cutoff of 0.90 and 0.70, respectively. SNPs were also determined in parallel using the reference-free *k*-mer based approach developed in kSNP v2.0 (33). Default parameters as well as a *k*-mer value of 19 selected as the optimal value predicted by the kSNP associated Kchooser script were applied.

Exclusion of suboptimal genome datasets

We devised a statistical approach that excluded outliers based on several non-Gaussian metrics that could be determined from mapped and assembled Illumina genome data (summarized in Supplementary Table 8). Specifically we examined three mapping metrics: (i) sequence coverage, (ii) number of unmapped bases (in the mapping reference EC958 genome) and (iii) number of uncalled bases due to low coverage or mixed-base calls; and two assembly metrics: (iv) the number of scaffolds ≥ 200 bp and (v) estimated genome size. Sub-optimal genomes were discriminated quantitatively on the basis of metrics (iii), (iv) and (v), and a total of 14 outliers were identified (based on upper and lower cut-offs at the Quartile 3 + 1.5 Interquartile range and the Quartile 1 - 1.5 Interquartile range cut-offs, respectively). Metric (i) did not identify any outliers with low sequence coverage and outliers with high sequence coverage were not omitted, whereas metric (ii) did not discriminate any outliers. This additional QC process resulted in the exclusion of genome data from the following strains CD301, CD436, JJ1901, JJ1996, JJ2007, JJ2041, JJ2050, JJ2243, JJ2441, JJ2555, MH17102, QU300, QUC02 and ZH193. R scripts used are available on github at <https://github.com/BeatsonLab-MicrobialGenomics/ST131-200>. A final dataset of 188 ST131 genomes (including the complete genomes of EC958, JJ1886 and SE15) were chosen for further study after excluding the 14 genomes with suboptimal data quality.

Recombination detection

To avoid distortion of the phylogenetic signal caused by SNPs acquired through recombination, we used the Bayesian clustering algorithm BRATNextGen (34) to detect recombinant regions among the combined dataset. Similar to our previous work

(2), we used as an input a SNP-based multiple genome alignment composed of each strain-specific pseudo-genome built by integrating the SNPs predicted for each strain to the reference genome of EC958. To help identification of underlying clusters of strains, BRATNextGen initially computes a hierarchical clustering tree relative to the proportion of ancestral sequences shared between all strains. A segregation cut-off of 0.12 was then specified to separate each previously identified ST131 clade (clades A, B and C) and non-ST131 strains into distinct clusters. Recombination was then evaluated within and between each cluster with the convergence approximated using 20 iterations of the learning algorithm. Significance was estimated using 100 permutations with a statistical significance threshold of 0.05.

Phylogenetic analysis

SNPs identified through reference-based mapping for the 188 ST131 strains were used to build phylogenies using Maximum Likelihood (ML), prior to and after removal of SNPs associated with recombinant regions. Phylogenetic trees were generated with RAxML v7.2.8 (35) using the general time reversible (GTR) GAMMA model of among site rate variation (ASRV), and validated using 1000 bootstrap repetitions to assess nodal support. Additionally, reference-free k -mer based phylogenetic trees were constructed using kSNP v2.0 with default parameters (33) and genome assemblies as an input. A k -mer value of 19 was selected as the optimal value predicted by the kSNP associated Kchooser script. All trees were then viewed using Figtree v1.4.0 (36) or Evolview (37), and further compared using the Tanglegram algorithm of Dendroscope v3.2.10 (38), which generates two rectangular phylograms to allow comparison of bifurcating trees.

Bayesian temporal and geographical analysis

Preliminary estimation of the underlying temporal signal of our data was obtained by performing a regression analysis between the root-to-tip genetic distance extracted from the recombination-free maximum-likelihood tree, the isolation year and lineage information for each sequence, as implemented in Path-O-Gen v1.4 (22). To further investigate the divergence of clade C from clade B, we performed a temporal analysis on the 3,779 bp non-recombinant SNPs of the 172 clade B and C strains using BEAST 2.0 (18), a Bayesian phylogenetic inference software, which can estimate the dating of emergence of distinct lineages. We compared multiple combinations of molecular clock model (strict, constant relaxed lognormal, and exponential relaxed lognormal), substitution model (HKY, GTR) and population size change model (coalescent constant, exponential growth, Bayesian skyline, extended Bayesian skyline). Markov Chain Monte Carlo generations for each analysis were conducted in triplicate for 100 million steps, sampling every 1,000 steps, to ensure convergence. Replicate analyses were then combined with LogCombiner, with a 10% burn-in. The GTR nucleotide substitution model was preferred over the HKY model, and was used with four discrete gamma-distributed rate categories and a default gamma prior distribution of 1. The uncorrelated lognormal clock model consistently gave better support based on the Bayes Factor and AICM analyses, compared to a strict clock model. The Bayesian skyline population tree model was chosen as the best fitting tree model. Maximum clade credibility (MCC) trees reporting mean values with a posterior probability limit set at 0.5 were then created using TreeAnnotator.

In order to adequately investigate the biogeographical history of our ST131 collection, we evaluated potential bias in the geographical origin of strains, which could

negatively impact our predictions. Statistical significance of the geographical origin distribution in clade B, C1 and C2 was assessed by Chi-Square test with Bonferroni correction for multiple comparisons. Over-represented countries were randomly subsampled down to 15 representative sequences, while countries with fewer than 5 representatives had to be excluded from the analysis (Korea and Portugal). Overall, we constructed 10 independent randomly subsampled datasets with 85 isolates representing 7 countries, each with 5 to 15 representative sequences. Reconstruction of possible ancestral geographical states was then performed using BEAST 1.8.2 on each individual subsampled dataset. In addition to the previous parameters selected for the temporal analysis, a symmetric substitution model, a Bayesian stochastic search variable selection (BSSVS) model and a strict clock for discrete locations were chosen for the phylogeographical analysis. MCMC generations were conducted for 100,000,000 steps, sampling every 10,000 steps. MCC trees were then generated using TreeAnnotator for each run with a posterior probability limit set at 0.5. Location posterior probabilities of MRCA were then collated for clade B and C, and clade C only.

Genomic comparisons and in silico genotyping

Comparative genomic analyses were conducted using a combination of tools, namely Artemis, Artemis Comparison Tool (39) and Mauve (30). Graphical representations showing the presence, absence or variation of mobile genetic elements (MGE) or other regions of interest, virulence factor genes, and antibiotic resistance genes were carried out using BLASTn and read-mapping information as implemented in the Seqfndr visualization tool (40). Regions of interest previously described in the genome of ST131 reference strain EC958 (2, 7) and virulence factors including

429 autotransporters, fimbriae, iron uptake, toxins, UPEC-specific genes and other
430 virulence genes were screened in all ST131 strains with SeqFindr using a cut-off $\geq$
431 95% nucleotide identity over the whole length when compared to the assembly or the
432 consensus generated from mapping. Additionally, prevalence of antibiotic resistance-
433 associated genes was also investigated using Srst2 (41) against ARGannot database,
434 with a minimum of depth of 15X read coverage.

REFERENCES

1. **Nicolas-Chanoine MH, Blanco J, Leflon-Guibout V, Demarty R, Alonso MP, Canica MM, Park YJ, Lavigne JP, Pitout J, Johnson JR.** 2008. Intercontinental emergence of *Escherichia coli* clone O25:H4-ST131 producing CTX-M-15. *J Antimicrob Chemother* **61**:273-281.
2. **Petty NK, Ben Zakour NL, Stanton-Cook M, Skippington E, Totsika M, Forde BM, Phan MD, Gomes Moriel D, Peters KM, Davies M, Rogers BA, Dougan G, Rodriguez-Bano J, Pascual A, Pitout JD, Upton M, Paterson DL, Walsh TR, Schembri MA, Beatson SA.** 2014. Global dissemination of a multidrug resistant *Escherichia coli* clone. *Proc Natl Acad Sci U S A* **111**:5694-5699.
3. **Totsika M, Beatson SA, Sarkar S, Phan MD, Petty NK, Bachmann N, Szubert M, Sidjabat HE, Paterson DL, Upton M, Schembri MA.** 2011. Insights into a multidrug resistant *Escherichia coli* pathogen of the globally disseminated ST131 lineage: genome analysis and virulence mechanisms. *PLoS One* **6**:e26578.
4. **Totsika M, Moriel DG, Idris A, Rogers BA, Wurpel DJ, Phan MD, Paterson DL, Schembri MA.** 2012. Uropathogenic *Escherichia coli* mediated urinary tract infection. *Curr Drug Targets* **13**:1386-1399.
5. **Price LB, Johnson JR, Aziz M, Clabots C, Johnston B, Tchesnokova V, Nordstrom L, Billig M, Chattopadhyay S, Stegger M, Andersen PS, Pearson T, Riddell K, Rogers P, Scholes D, Kahl B, Keim P, Sokurenko EV.** 2013. The epidemic of extended-spectrum-beta-lactamase-producing *Escherichia coli* ST131 is driven by a single highly pathogenic subclone, H30-Rx. *MBio* **4**:e00377-00313.
6. **Johnson JR, Tchesnokova V, Johnston B, Clabots C, Roberts PL, Billig M, Riddell K, Rogers P, Qin X, Butler-Wu S, Price LB, Aziz M, Nicolas-Chanoine MH, Debroy C, Robicsek A, Hansen G, Urban C, Platell J, Trott DJ, Zhanel G, Weissman SJ, Cookson BT, Fang FC, Limaye AP, Scholes D, Chattopadhyay S, Hooper DC, Sokurenko EV.** 2013. Abrupt emergence of a single dominant multidrug-resistant strain of *Escherichia coli*. *J Infect Dis* **207**:919-928.

- 464 7. **Forde BM, Ben Zakour NL, Stanton-Cook M, Phan MD, Totsika M, Peters KM, Chan**
465 **KG, Schembri MA, Upton M, Beatson SA.** 2014. The complete genome sequence of
466 *Escherichia coli* EC958: a high quality reference sequence for the globally disseminated
467 multidrug resistant *E. coli* O25b:H4-ST131 clone. *PLoS One* **9**:e104400.
- 468 8. **Phan MD, Peters KM, Sarkar S, Lukowski SW, Allsopp LP, Gomes Moriel D, Achard**
469 **ME, Totsika M, Marshall VM, Upton M, Beatson SA, Schembri MA.** 2013. The serum
470 resistome of a globally disseminated multidrug resistant uropathogenic *Escherichia coli* clone.
471 *PLoS Genet* **9**:e1003834.
- 472 9. **Forde BM, Phan MD, Gawthorne JA, Ashcroft MM, Stanton-Cook M, Sarkar S, Peters**
473 **KM, Chan KG, Chong TM, Yin WF, Upton M, Schembri MA, Beatson SA.** 2015.
474 Lineage-Specific Methyltransferases Define the Methylome of the Globally Disseminated
475 *Escherichia coli* ST131 Clone. *MBio* **6**.
- 476 10. **Kakkanat A, Totsika M, Schaale K, Duell BL, Lo AW, Phan MD, Moriel DG, Beatson**
477 **SA, Sweet MJ, Ulett GC, Schembri MA.** 2015. The role of H4 flagella in *Escherichia coli*
478 ST131 virulence. *Sci Rep* **5**:16149.
- 479 11. **Phan MD, Forde BM, Peters KM, Sarkar S, Hancock S, Stanton-Cook M, Ben Zakour**
480 **NL, Upton M, Beatson SA, Schembri MA.** 2015. Molecular characterization of a multidrug
481 resistance IncF plasmid from the globally disseminated *Escherichia coli* ST131 clone. *PLoS*
482 *One* **10**:e0122369.
- 483 12. **Phan MD, Peters KM, Sarkar S, Forde BM, Lo AW, Stanton-Cook M, Roberts LW,**
484 **Upton M, Beatson SA, Schembri MA.** 2015. Third-generation cephalosporin resistance
485 conferred by a chromosomally encoded blaCMY-23 gene in the *Escherichia coli* ST131
486 reference strain EC958. *J Antimicrob Chemother* **70**:1969-1972.
- 487 13. **Coelho A, Mora A, Mamani R, Lopez C, Gonzalez-Lopez JJ, Larrosa MN, Quintero-**
488 **Zarate JN, Dahbi G, Herrera A, Blanco JE, Blanco M, Alonso MP, Prats G, Blanco J.**
489 2011. Spread of *Escherichia coli* O25b:H4-B2-ST131 producing CTX-M-15 and SHV-12 with
490 high virulence gene content in Barcelona (Spain). *J Antimicrob Chemother* **66**:517-526.
- 491 14. **Johnson JR, Johnston B, Clabots C, Kuskowski MA, Castanheira M.** 2010. *Escherichia*
492 *coli* sequence type ST131 as the major cause of serious multidrug-resistant *E. coli* infections
493 in the United States. *Clin Infect Dis* **51**:286-294.

- 494 15. **Johnson JR, Menard M, Johnston B, Kuskowski MA, Nichol K, Zhanel GG.** 2009.
495 Epidemic clonal groups of *Escherichia coli* as a cause of antimicrobial-resistant urinary tract
496 infections in Canada, 2002 to 2004. *Antimicrob Agents Chemother* **53**:2733-2739.
- 497 16. **Lanza VF, de Toro M, Garcillan-Barcia MP, Mora A, Blanco J, Coque TM, de la Cruz**
498 **F.** 2014. Plasmid flux in *Escherichia coli* ST131 sublineages, analyzed by plasmid
499 constellation network (PLACNET), a new method for plasmid reconstruction from whole
500 genome sequences. *PLoS Genet* **10**:e1004766.
- 501 17. **Stoesser N, Sheppard A, Pankhurst L, de Maio N, Moore CE, Sebra R, Turner P, Anson**
502 **LW, Kasarskis A, Batty EM, Kos V, Wilson DJ, Phetsouvanh R, Wyllie D, Sokurenko E,**
503 **Manges AR, Johnson TJ, Price LB, Peto TEA, Johnson JR, Didelot X, Walker AS,**
504 **Crook DW, .** 2015. Evolutionary history of the global emergence of the *Escherichia coli*
505 epidemic clone ST131. *bioRxiv*.
- 506 18. **Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut**
507 **A, Drummond AJ.** 2014. BEAST 2: a software platform for Bayesian evolutionary analysis.
508 *PLoS Comput Biol* **10**:e1003537.
- 509 19. **Da Cunha V, Davies MR, Douarre PE, Rosinski-Chupin I, Margarit I, Spinali S, Perkins**
510 **T, Lechat P, Dmytruk N, Sauvage E, Ma L, Romi B, Tichit M, Lopez-Sanchez MJ,**
511 **Descorps-Declere S, Souche E, Buchrieser C, Trieu-Cuot P, Moszer I, Clermont D,**
512 **Maione D, Bouchier C, McMillan DJ, Parkhill J, Telford JL, Dougan G, Walker MJ,**
513 **Consortium D, Holden MT, Poyart C, Glaser P.** 2014. *Streptococcus agalactiae* clones
514 infecting humans were selected and fixed through the extensive use of tetracycline. *Nat*
515 *Commun* **5**:4544.
- 516 20. **Holden MT, Hsu LY, Kurt K, Weinert LA, Mather AE, Harris SR, Strommenger B,**
517 **Layer F, Witte W, de Lencastre H, Skov R, Westh H, Zemlickova H, Coombs G, Kearns**
518 **AM, Hill RL, Edgeworth J, Gould I, Gant V, Cooke J, Edwards GF, McAdam PR,**
519 **Templeton KE, McCann A, Zhou Z, Castillo-Ramirez S, Feil EJ, Hudson LO, Enright**
520 **MC, Balloux F, Aanensen DM, Spratt BG, Fitzgerald JR, Parkhill J, Achtman M,**
521 **Bentley SD, Nubel U.** 2013. A genomic portrait of the emergence, evolution, and global
522 spread of a methicillin-resistant *Staphylococcus aureus* pandemic. *Genome Res* **23**:653-664.

- 523 21. **Andersen PS, Stegger M, Aziz M, Contente-Cuomo T, Gibbons HS, Keim P, Sokurenko**
524 **EV, Johnson JR, Price LB.** 2013. Complete Genome Sequence of the Epidemic and Highly
525 Virulent CTX-M-15-Producing H30-Rx Subclone of Escherichia coli ST131. Genome
526 Announc 1.
- 527 22. **Rambaut A.** 2013. Path-O-Gen. <http://tree.bio.ed.ac.uk/software/pathogen/>.
528 Accessed
- 529 23. **Holt KE, Thieu Nga TV, Thanh DP, Vinh H, Kim DW, Vu Tra MP, Campbell JI, Hoang**
530 **NV, Vinh NT, Minh PV, Thuy CT, Nga TT, Thompson C, Dung TT, Nhu NT, Vinh PV,**
531 **Tuyet PT, Phuc HL, Lien NT, Phu BD, Ai NT, Tien NM, Dong N, Parry CM, Hien TT,**
532 **Farrar JJ, Parkhill J, Dougan G, Thomson NR, Baker S.** 2013. Tracking the establishment
533 of local endemic populations of an emergent enteric pathogen. Proc Natl Acad Sci U S A
534 110:17522-17527.
- 535 24. **Kaitin KI, Richard BW, Lasagna L.** 1987. Trends in drug development: the 1985-86 new
536 drug approvals. J Clin Pharmacol 27:542-548.
- 537 25. **Paul S, Linardopoulou EV, Billig M, Tchesnokova V, Price LB, Johnson JR,**
538 **Chattopadhyay S, Sokurenko EV.** 2013. Role of homologous recombination in adaptive
539 diversification of extraintestinal Escherichia coli. J Bacteriol 195:231-242.
- 540 26. **Avasthi TS, Kumar N, Baddam R, Hussain A, Nandanwar N, Jadhav S, Ahmed N.** 2011.
541 Genome of Multidrug-Resistant Uropathogenic Escherichia coli Strain NA114 from India.
542 Journal of Bacteriology 193:4272-4273.
- 543 27. **Schmieder R, Edwards R.** 2011. Quality control and preprocessing of metagenomic datasets.
544 Bioinformatics 27:863-864.
- 545 28. **Davis MP, van Dongen S, Abreu-Goodger C, Bartonicek N, Enright AJ.** 2013. Kraken: a
546 set of tools for quality control and analysis of high-throughput sequence data. Methods 63:41-
547 49.
- 548 29. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de
549 Bruijn graphs. Genome Res 18:821-829.
- 550 30. **Darling AE, Mau B, Perna NT.** 2010. progressiveMauve: Multiple Genome Alignment with
551 Gene Gain, Loss and Rearrangement. PLoS ONE 5:e11147.

552 31. **David M, Dzamba M, Lister D, Ilie L, Brudno M.** 2011. SHRiMP2: sensitive yet practical
553 SHort Read Mapping. *Bioinformatics* **27**:1011-1012.

554 32. **Consortium VB.** 2013. Neson Downloads.
555 <http://www.vicbioinformatics.com/software/nesoni.shtml>. Accessed

556 33. **Gardner SN, Hall BG.** 2013. When whole-genome alignments just won't work: kSNP v2
557 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial
558 genomes. *PLoS One* **8**:e81760.

559 34. **Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, Corander**
560 **J.** 2011. Detection of recombination events in bacterial genomes from large population
561 samples. *Nucleic Acids Research* doi:10.1093/nar/gkr928.

562 35. **Stamatakis A.** 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses
563 with thousands of taxa and mixed models. *Bioinformatics* **22**:2688-2690.

564 36. **Rambaut A.** 2009. FigTree, a graphical viewer of phylogenetic trees.
565 <http://tree.bio.ed.ac.uk/software/figtree/>. Accessed

566 37. **Zhang H, Gao S, Lercher MJ, Hu S, Chen WH.** 2012. EvolView, an online tool for
567 visualizing, annotating and managing phylogenetic trees. *Nucleic Acids Res* **40**:W569-572.

568 38. **Huson DH, Scornavacca C.** 2012. Dendroscope 3: An Interactive Tool for Rooted
569 Phylogenetic Trees and Networks. *Systematic Biology* doi:10.1093/sysbio/sys062.

570 39. **Carver T, Berriman M, Tivey A, Patel C, Bohme U, Barrell BG, Parkhill J, Rajandream**
571 **MA.** 2008. Artemis and ACT: viewing, annotating and comparing sequences stored in a
572 relational database. *Bioinformatics* **24**:2672-2676.

573 40. **Stanton-Cook M, Ben Zakour NL, Alikhan NF, Beatson SA.** 2013.
574 <http://mscook.github.io/SeqFindR/>. Accessed

575 41. **Inouye M, Dashnow H, Raven LA, Schultz MB, Pope BJ, Tomita T, Zobel J, Holt KE.**
576 2014. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs.
577 *Genome Med* **6**:90.

FIGURE LEGENDS

Figure 1: Geographical diversity of the combined dataset across clades, and time.

a) Stacked histogram showing the number of strains in clades A, B, C1 and C2 according to their country of origin. The colour scheme is shown in the legend on the left along with abbreviated country names. b) Box-and-whiskers plot showing the distribution according to year of isolation for all strains based on their country of origin.

Figure 2: Maximum Likelihood phylogenetic tree of ST131 strains. The phylogram was built from 5,471 non-recombinant SNPs using maximum likelihood (ML). Branch support was performed by 1000 bootstrap replicates (see Supplementary Fig 2b). As shown by the scale, branch length corresponds to the number of SNP differences. Taxa labels for clades A, B and C are colored in red, orange, and green, respectively. Seven strains sharing intermediate characteristics between clades B and C are colored in pink. Of note, clade A strains were collapsed and the Clade A-specific branches shortened for display purpose. Metadata is represented as circles as follows: year of isolation in gray-gradient, and geographical region in assorted colours as depicted in the legend. Allelic profiling information is shown as coloured strips surrounding the phylogram (inner-to-outer) for *fimH*, *parC*, *gyrA*, and CTX-M. Two additional distinctions were made for some *fimH* variants: untypeable corresponds to a strain with a truncated or missing *fimH* gene; and pseudogene corresponds to a strain in which *fimH* is disrupted by an insertion sequence. Clade B0-5 and C0 sub-clades are shown as arcs in the outer-most ring with arrows and dotted lines denoting the division between sub-clade C1 and C2.

Figure 3: Prevalence of antibiotic resistance-associated genes by a) clade and b)

clade-country. Box-and-whiskers plot showing number of resistance-associated genes per a) clade, and b) per clade and country. Colours correspond to clades, namely: clade A, red; clade B, orange; clade C1, light green; and clade C2, dark green. Screening was done using Srst2 (41) against ARGannot, with a minimum of depth of 15X read coverage. *P*-value are indicated as follows: ‘***’<0.001, ‘**’<0.01, ‘*’<0.05.

Figure 4: Evolutionary scenario of the emergence of ST131 clades B and C. A

time-calibrated phylogeny was reconstructed using BEAST 2.0 based on 3,779 bp non-recombinant SNPs for the 172 clade B and C strains. Of all combinations tested (Table S4), the one combining the GTR substitution model, a constant relaxed clock model and the Bayesian Skyline population tree model was preferred. a) Maximum Clade Credibility tree colored according to clade origin as shown on the right with B, orange; Intermediate B(0) and C(0), in pink; C1, in light green; and C2, in dark green. X-axis indicates emergence time estimates of the corresponding strains. Major evolutionary events were also indicated by an arrow pointing at the branch onto which they are predicted to have occurred (position along the branch is arbitrary). Three categories of major events are displayed, namely: MGE and genomic island insertion events, in black; allelic change acquired through recombination events, in blue; and allelic change acquired through point mutation events, in purple. Of note, the two point mutations indicated by a purple arrow and pointing at the branch from which clade C1 and C2 originate, confer resistance to fluoroquinolone, for which the first introduction is indicated in the bottom timeline by a red arrow. b) Unrooted phylogenetic tree built on the same 3,779 bp non-recombinant set of SNPs using

maximum likelihood (ML). Branch support was performed by 1000 bootstrap replicates. Intermediate strain names and predicted acquisition of *fimH30* are indicated on the tree. C) The Bayesian Skyline plot illustrates the predicted demographic changes of the ST131 clade B and C population since the mid 1940's. The black curve indicates the effective population size (N_e) with 95% confidence intervals shown in grey.

Figure 5: Geographic location of the Most Recent Common Ancestor (MRCA) of ST131 major clades. Individual probabilities were predicted from 10 independent BEAST analyses of randomly subsampled data to limit bias related to over-representation of some locations. Mean probability of the geographic location of the MRCA for a) clades B and C, and b) clade C, are shown as a box-and-whiskers plot colored according to country using the scheme as described in Figure 1. Countries are labeled on the x-axis by abbreviation.



Figure 1

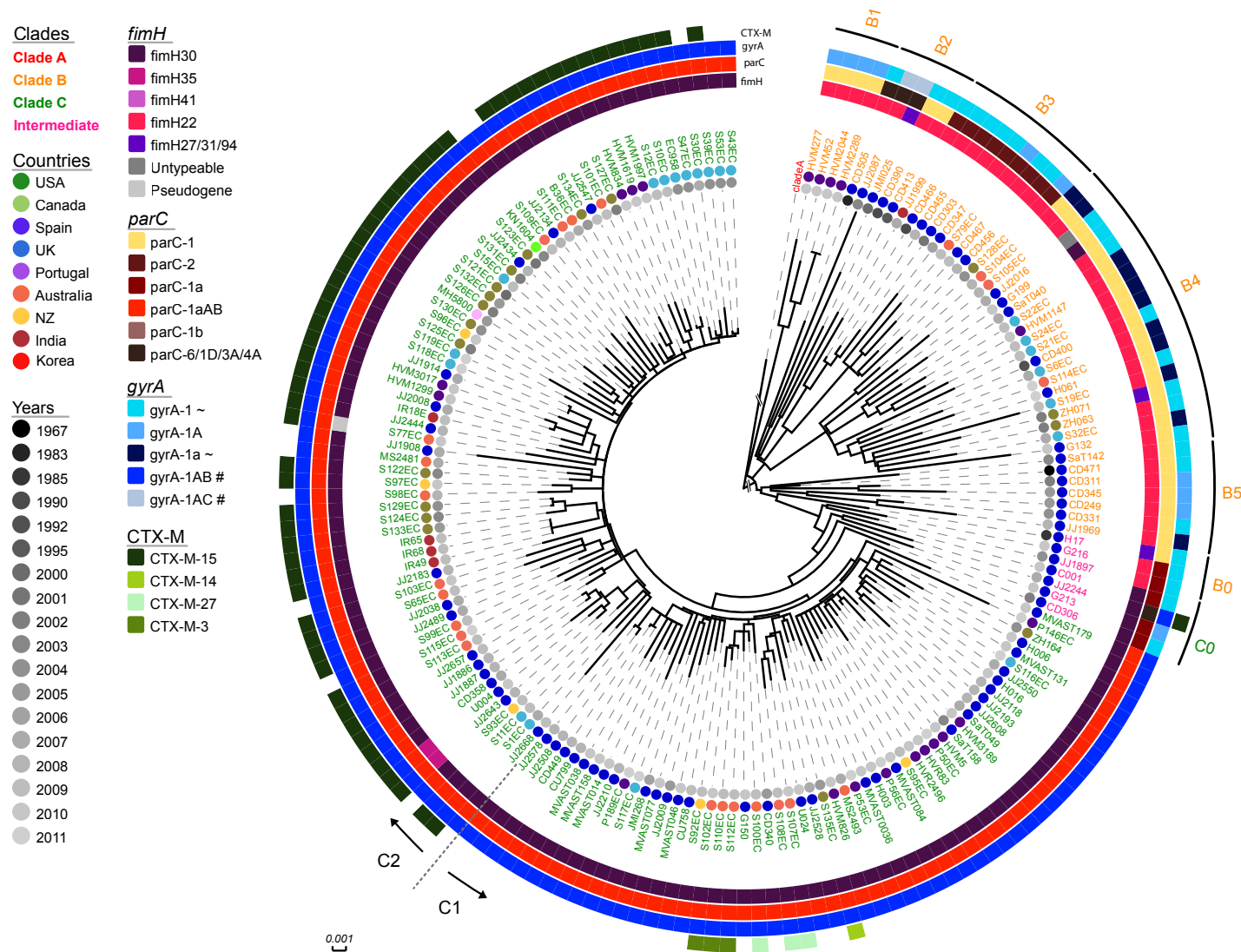


Figure 2

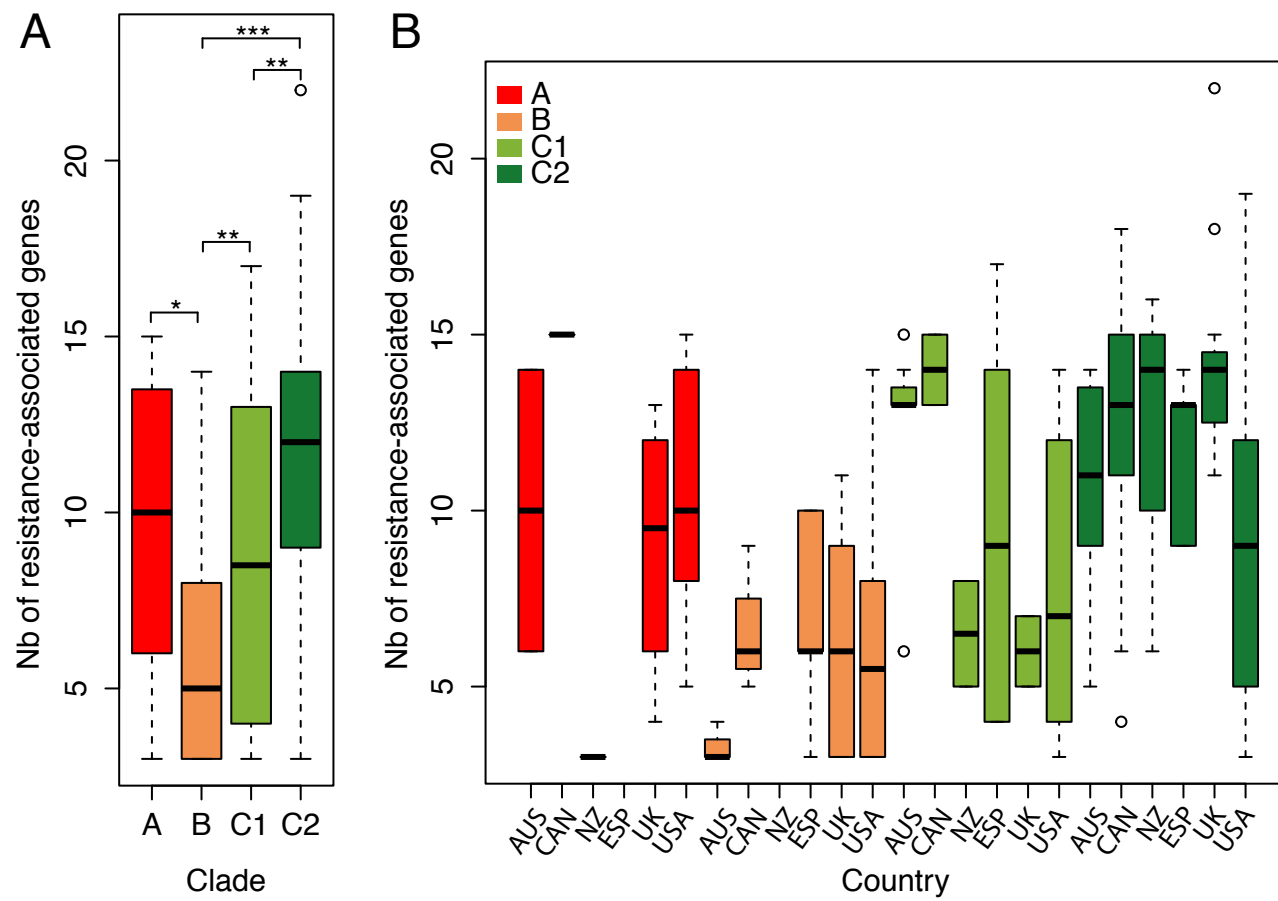


Figure 3

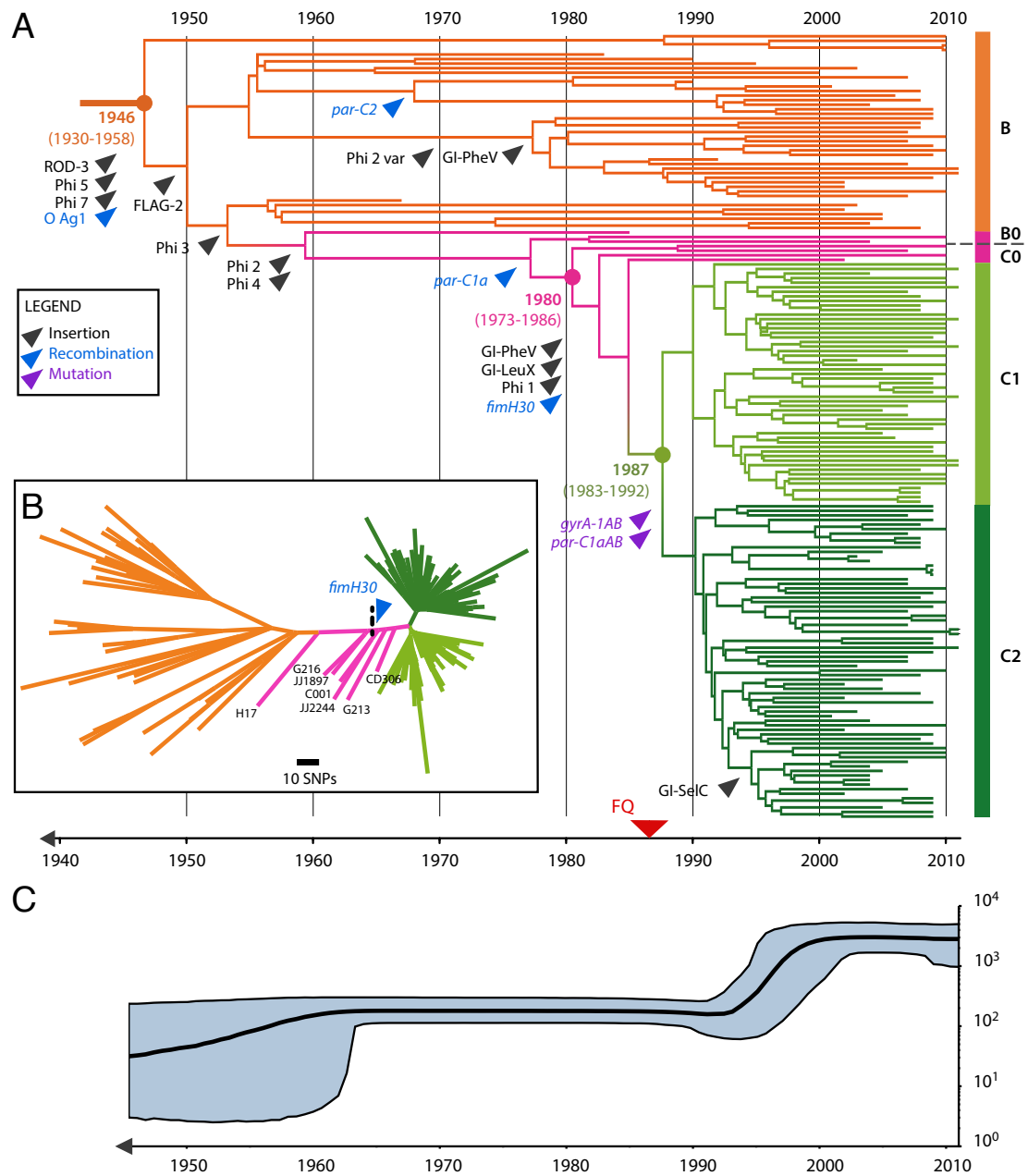


Figure 4

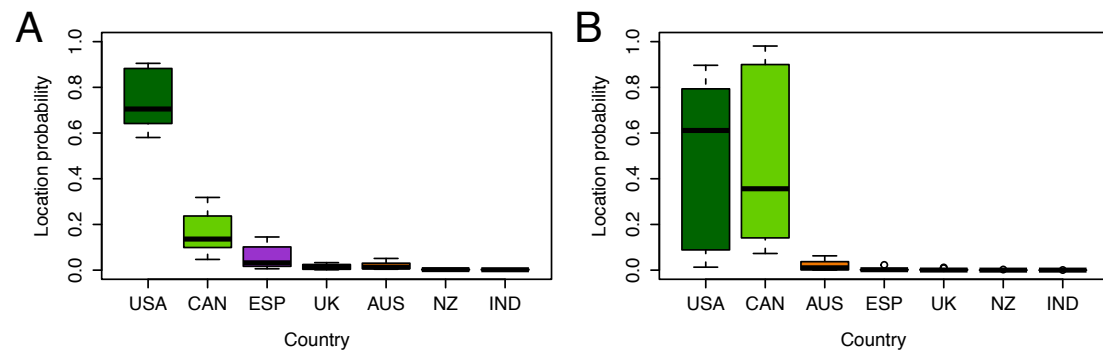


Figure 5