UNIVERSITY OF BIRMINGHAM

Diversification of colonization factors in a multidrug-resistant Escherichia coli lineage evolving under negative frequency-dependent selection

McNally, Alan; Kallonen, Teemu; Connor, Christopher; Abudahab, Khalil; Aanensen, David M; Horner, Carolyne; Peacock, Sharon J; Parkhill, Julian; Croucher, Nicholas J; Corander, Jukka

Link to publication on Research at Birmingham portal

Check for updates

Diversification of Colonization Factors in a Multidrug-Resistant *Escherichia coli* Lineage Evolving under Negative Frequency-Dependent Selection

Alan McNally,[a] Teemu Kallonen,[b,c] Christopher Connor,[a] Khalil Abudahab,[b] David M. Aanensen,[b] Carolyne Horner,[d] Sharon J. Peacock,[b,e,f] Julian Parkhill,[b] Nicholas J. Croucher,[g] Jukka Corander[b,c,h]

[a]Institute of Microbiology and Infection, University of Birmingham, Birmingham, United Kingdom
[b]Infection Genomics, Wellcome Sanger Institute, Cambridge, United Kingdom
[c]Department of Biostatistics, University of Oslo, Oslo, Norway
[d]British Society of Antimicrobial Chemotherapy, Birmingham, United Kingdom
[e]Department of Medicine, University of Cambridge, Cambridge, United Kingdom
[f]London School of Hygiene and Tropical Medicine, London, United Kingdom
[g]Faculty of Medicine, School of Public Health, Imperial College, London, United Kingdom
[h]Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

ABSTRACT *Escherichia coli* is a major cause of bloodstream and urinary tract infections globally. The wide dissemination of multidrug-resistant (MDR) strains of extraintestinal pathogenic *E. coli* (ExPEC) poses a rapidly increasing public health burden due to narrowed treatment options and increased risk of failure to clear an infection. Here, we present a detailed population genomic analysis of the ExPEC ST131 clone, in which we seek explanations for its success as an emerging pathogenic strain beyond the acquisition of antimicrobial resistance (AMR) genes. We show evidence for evolution toward separate ecological niches for the main clades of ST131 and differential evolution of anaerobic metabolism, key colonization, and virulence factors. We further demonstrate that negative frequency-dependent selection acting across accessory loci is a major mechanism that has shaped the population evolution of this pathogen.

IMPORTANCE Infections with multidrug-resistant (MDR) strains of *Escherichia coli* are a significant global public health concern. To combat these pathogens, we need a deeper understanding of how they evolved from their background populations. By understanding the processes that underpin their emergence, we can design new strategies to limit evolution of new clones and combat existing clones. By combining population genomics with modelling approaches, we show that dominant MDR clones of *E. coli* are under the influence of negative frequency-dependent selection, preventing them from rising to fixation in a population. Furthermore, we show that this selection acts on genes involved in anaerobic metabolism, suggesting that this key trait, and the ability to colonize human intestinal tracts, is a key step in the evolution of MDR clones of *E. coli*.

KEYWORDS AMR, *Escherichia coli*, evolutionary genomics, negative frequency-dependent selection

Address correspondence to Alan McNally, a.mcnally.1@bham.ac.uk, or Jukka Corander, jukka.corander@helsinki.fi.

A.M., T.K., and C.C. as well as N.J.C. and. J.C contributed equally to this article.

*E*scherichia coli* is now the most common cause of bloodstream infections in the developed world, outnumbering cases of *Staphylococcus aureus* bacteremia by 2:1 (1). *E. coli* is also the most common cause of urinary tract infections (UTI), which in turn are among the most common bacterial infections in the world (2). Bacteremia and UTI are caused by a subset of *E. coli* termed extraintestinal pathogenic *E. coli* (ExPEC). ExPEC

are not a phylogenetically distinct group of *E. coli* but rather represent strains which have acquired virulence-associated genes that confer the ability to invade and cause disease in extraintestinal sites (3). Genes associated with virulence that confer the ability to adhere to extraintestinal tissues, to sequester extracellular iron, and to evade the nonspecific immune response and toxins resulting in localized tissue destruction have all been described as essential in the process of ExPEC pathogenesis (4).

The problem presented by the scale of ExPEC infections is exacerbated by the number of cases involving multidrug-resistant (MDR) strains (1, 5, 6). Epidemiological surveys report as many as 60% of UTI ExPEC isolates as being resistant to three or more classes of antibiotics and as many as 50% of bacteremia isolates (5, 6). The increase in MDR ExPEC prevalence has been rapid and primarily attributable to a small number of ExPEC lineages (5). The most common of these is the *E. coli* ST131 lineage, which has rapidly become a dominant cause of ExPEC UTI and bacteremia globally (5–7). *E. coli* ST131 is particularly associated with carriage of the CTX-M class of extended-spectrum β-lactamase (ESBL) which confers resistance to 3rd-generation cephalosporins (7), and there have been a small number of reports of *E. coli* ST131 isolates carrying metallo-β-lactamases conferring resistance to carbapenems (8). The carriage of these resistance genes is driven by the acquisition and stable maintenance of large MDR plasmids (9).

The phylogenetic structure of *E. coli* ST131 is well characterized (10–14) and shows the emergence of a globally disseminated MDR-associated clade C from primarily drug-susceptible clades A and B. The lack of phylogeographic signal and phylogenetic structure based on host source suggests rapid global dispersal and frequent host transitions within clade C (14). Research has suggested that the acquisition of fluoro-quinolone resistance via point mutations in DNA gyrase and DNA topoisomerase genes was the primary driver in the rapid emergence of clade C, alongside the predated acquisition of well-defined ExPEC virulence factors (11, 12). Later work also suggested that clade C *E. coli* ST131 may dominate as a successful MDR clade due to the ability to offset the fitness cost of MDR plasmid acquisition and maintenance via compensatory mutations in gene regulatory regions (14). Genome-wide association studies (GWAS) have been used to identify loci and lineage-specific alleles significantly associated with clade C *E. coli* ST131, which suggested a secondary flagellum locus encoding lateral flagella (Flag-2 [15]), and a number of hypothetical proteins and promoter regions as being clade C *E. coli* ST131-associated loci (14).

Recent work on *E. coli* causing bacteremia provided compelling evidence that resistance to antimicrobials has not been the major driver of the success of ST131 (16). An analysis of a large 11-year population survey across the United Kingdom showed that ST131 rapidly stabilized at a level of approximately 20% after its emergence around 2002 in the United Kingdom. This was far in excess of already-resident MDR clones, such as ST88 or ST405. Nevertheless, the overall prevalence of resistance phenotypes remained approximately constant in the population. Furthermore, most currently known major ExPEC clones (primarily ST12, ST73, ST95, and ST69, the last of which also rapidly emerged in 2002) show a similar stable population frequency across the 10 years following the introduction of ST131, despite exhibiting far less extensive resistance profiles. These observations suggested the distribution of ExPEC strains was shaped by negative frequency-dependent selection (NFDS) (16). NFDS describes the situation in which a given phenotype is most beneficial to a population when it is rare, such as the emergence of a new antigen or resource-use strategy. This is because as the phenotype becomes common it either becomes costly, such as when antigens are recognized by an increased proportion of hosts, or less beneficial, such as when strains compete more fiercely for the same resources.

Recently, a multilocus NFDS model of postvaccination *Streptococcus pneumoniae* population dynamics was described (17). Frequencies of accessory genes were found to be highly conserved across multiple populations on different continents, despite these populations themselves being composed of different strains, as defined by core genome sequences. Detailed modeling and functional analysis indicated changes in strain prevalence could be explained by NFDS driving accessory loci toward equilibrium

frequencies, through mechanisms involving interactions with other bacteria, hosts, or mobile elements (17). The levels of the selective force were estimated to be similar across the populations and manifested in the maintenance of stable population frequencies of accessory loci, despite a substantial perturbation of the population by the introduction of the pneumococcal vaccine (17).

Here, we synthesized different genomic analyses to provide a detailed view of the ecology and evolution of ST131. We provide evidence of NFDS shaping the species-wide carried *E. coli* population using 1,094 systematically sampled bacteremia isolates from the British Society for Antimicrobial Chemotherapy (BSAC) collection. These genomes, and those collated from previous large-scale phylogenomic studies (11–14, 16, 18), allowed a high-resolution analysis of *E. coli* ST131 using a total of 862 genomes, revealing the steps in the clades' diversification. This found clade C to have accumulated significantly elevated allelic diversity, particularly enriched for genes involved in anaerobic metabolism and other loci important for colonization of the human host by ExPEC. Our data suggest the evolution of the MDR phenotype is part of a wider ongoing adaptation toward prolonged human colonization that permits expansion despite NFDS due to an adaptive radiation through diversification of metabolic and antigenic loci.

RESULTS

Species-wide distribution of the *E. coli* accessory genome. The multilocus NFDS model devised to explain the impact of vaccination on the *S. pneumoniae* population was applied to the 1,094 systematically sampled *E. coli* isolates from the BSAC collection (16). Collected between 2001 and 2011, this collection revealed that the sequence clusters, or strains, corresponding primarily to sequence types ST69 and ST131 rose rapidly at the start of the sampling period, prior to stabilizing at frequencies of 8% and 18%, respectively. The rise of ST131 was primarily driven by increases in MDR clade C isolates, although isolates from ST69 and ST131 clade B rose in frequency despite them lacking such a resistance profile. It is possible this represents drift in a neutral population. However, this seemed unlikely, given that the invading genotypes rose quickly and displaced some strains, such as ST10, ST14, and ST144, while leaving the common ST73 and ST95 at quite stable frequencies. Alternatively, these strains could each be filling one distinct separate niche. However, these two strains displace multiple other genotypes, suggesting no strict demarcation between niches. Additionally, there was little evidence of any strain being associated with a range of unique traits. Using the previous analysis of gene content with Roary, the 18 strains with at least ten representatives in the population had a mean of 16.7 private genes (range, 1 to 49), defined as those loci present at >95% in one strain and <5% in all others.

However, this does not mean the association between sequence clusters and accessory loci was random. The same Roary analysis identified 8,311 intermediate-frequency genes, present in between 5% and 95% of the overall population analyzed by Kallonen et al. (16) (see Fig. S2 in the supplemental material). Their distribution exhibited a strong association with individual sequence clusters, as defined by an analysis of the core genome (see Materials and Methods), in agreement with recent *k*-mer-based analyses of this population that showed divergence in the core and accessory genome was strongly correlated (19). Hence, as in *S. pneumoniae*, strains appear to be defined by a characteristic combination of common accessory loci, rather than distinctive private sequences (14, 20). This suggests sequence clusters may be ecologically distinct but with overlapping niches.

We therefore tested whether the population-level frequencies of these intermediate frequency genes were conserved, as following perturbation by vaccination in *S. pneumoniae*. Comparisons between the pre-ST131 2001 samples and subsequent data from up to 2011 found strong linear correlations between the prevalences of their intermediate-frequency genes (Fig. 1A, see also Fig. S3A). This is consistent with these loci existing at "equilibrium" frequencies, determined by their costs and frequency-dependent benefits. These correlations with the first sample, in 2001, weakened with
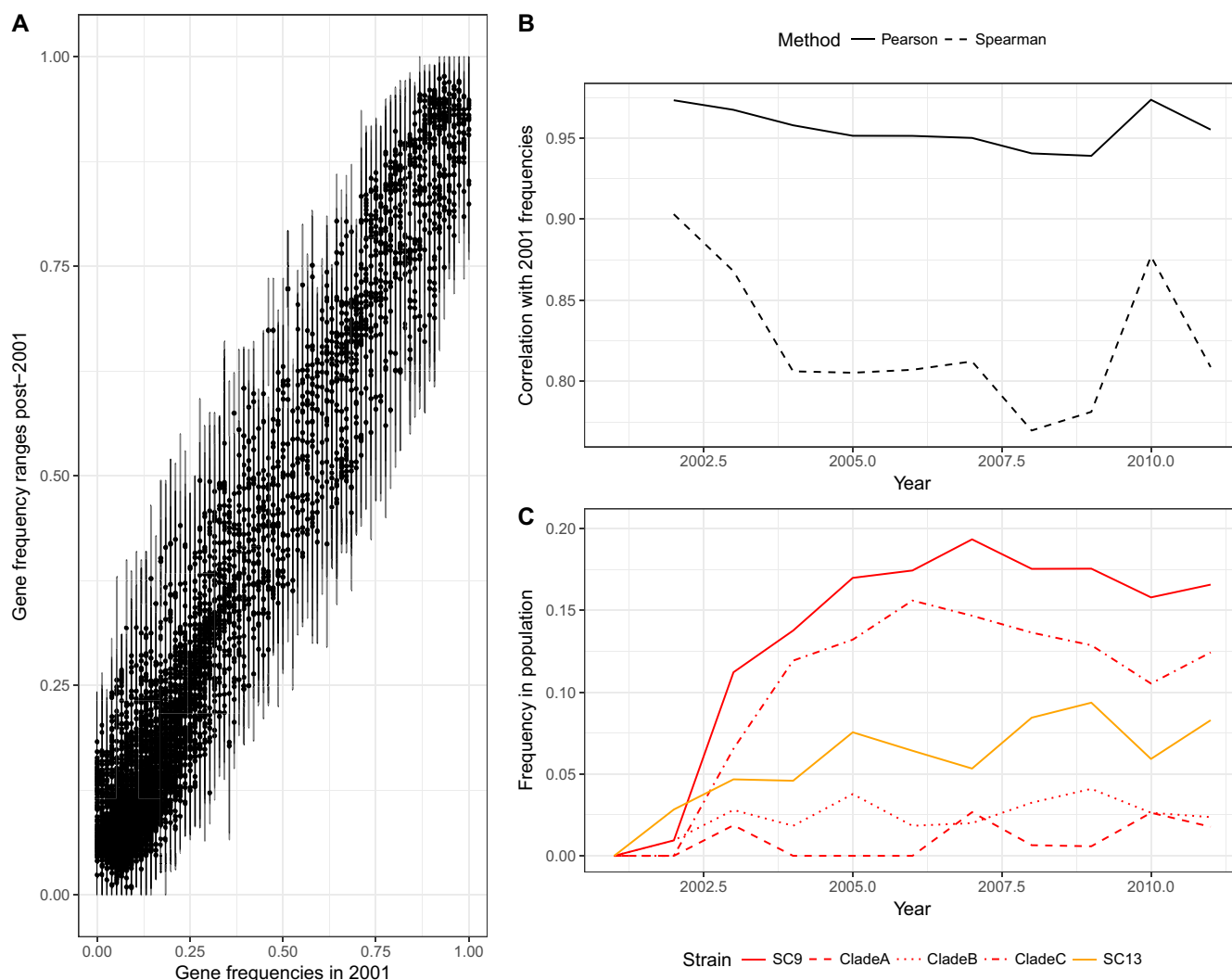
FIG 1 Summarizing the population dynamics of the British Society for Antimicrobial Chemotherapy extraintestinal pathogenic *E. coli* collection. These isolates were collected from bacteremia cases around the United Kingdom between 2001 and 2011. (A) Conservation of gene frequencies. Each point corresponds to one of the 6,824 genes identified by Roary in the BSAC collection with mean frequencies between 0.05 and 0.95 across all years. Error bars indicate the full range observed across annual samples. (B) Correlation of gene frequencies with those observed in 2001. This shows the changing correlation of gene frequencies, calculated by both the Pearson and Spearman methods, in each year relative to those observed in 2001. Both measures indicate a divergence in gene frequencies as ST69 and ST131 emerge, until 2010, at which point there is a reversion to the frequencies seen in the original population. (C) Emergence of ST69 (SC9, in orange) and ST131 (SC13, red). The frequencies of the subclades of ST131 are shown by the red dashed lines.

samples from until 2008, as might be expected with neutral drift (Fig. 1B), as ST131 and ST69 became more prevalent (Fig. 1C). However, the gene frequencies "bounced back" toward their original frequencies in 2008 (21), as indicated by the increased correlation with the 2001 frequencies (Fig. 1C). This elevated correlation was sustained in subsequent years, despite the persistently high frequencies of ST131 or ST69, suggesting a reconfiguration of other lineages in the population.

Multilocus NFDS modeling of the *E. coli* population. To obtain a population-wide view of these dynamics, the previously described multilocus NFDS model was applied to this data set to test whether these strain dynamics were consistent with selection of the phenotypes encoded by a set of accessory loci, L. L was primarily composed of the 7,204 genes, identified by Roary, present at intermediate frequencies (between 5% and 95%) in the 2001 sample. These prevalences were assumed to represent the equilibrium frequencies, e_l, determined by NFDS in an unperturbed population (17). As only a single sample was available that predated the emergence of ST69 and ST131, these e_l estimates are likely to be associated with considerable noise; however, the bounce back

of accessory locus frequencies toward these e_l suggests they are likely to be accurate for a high proportion of the modeled loci. Additionally, L included seven antibiotic resistance phenotypes present in the same intermediate frequency range. Although mechanistic models disagree as to whether antibiotic resistance is under NFDS directly (22) or linked to other loci under balancing selection (23), the stability of individual phenotype frequencies in this population (16) suggests this simplifying assumption is reasonable. These were then simulated as evolving under NFDS; a fraction p_f evolved under strong NFDS, determined by the parameter σ_f, while the rest evolved under weak NFDS, according to parameter σ_w (see Materials and Methods).

The model was initialized with the 2001 population, which was seeded with genotypes observed in later years at a low level, representing the possibility they were present in the population but unsampled. Subsequent simulation with a Wright-Fisher framework included these post-2001 genotypes migrating into the population at a rate m. Each isolate was assigned to a sequence cluster, or strain, through identifying clusters in the previously published hierBAPS analysis that approximately corresponded to a multilocus sequence typing clonal complex (see Materials and Methods). The invasion of the resident population by the sequence clusters corresponding to ST131 and ST69 was driven by an increased reproductive fitness relative to the rest of the population, parameterized in the model as r. Fitting this model through approximate Bayesian computation (ABC) using BOLFI estimated the parameters listed in Table S2a, which identified significant evidence for NFDS (σ_f and p_f greater than their respective lower bounds used in fitting). The 95% credibility interval σ_w was not higher than the lower bound used in fitting, which might represent a genuine absence of NFDS acting on these loci or a lack of power to detect NFDS in this population over this interval. The conclusion that NFDS acting on at least a subset of accessory loci accounts for the observed population dynamics better than equivalent simulations in which NFDS is negligible provides a quantitative gene-level model that explains the previous strain-level observations of Kallonen et al. (16).

These simulations successfully reproduced several aspects of the observed data (Fig. 2 and S3B). Both ST131 and ST69 rapidly spread through the population, before stabilizing at an equilibrium frequency of ~20%. This does not occur at the expense of the established common clones, such as ST73 and ST95. Instead, in accordance with the genomic data, the displaced sequence clusters include ST10, ST14, ST144, and ST405. These patterns are qualitatively distinct from an equivalent neutral model fit (Fig. 2, bottom). In these simulations lacking NFDS, both ST131 and ST69 are predicted to exponentially increase in frequency rather than rapidly stabilize at a frequency that permits coexistence with the already-resident strains. All other strains commensurately decrease at accelerating rates, proportionate to their original prevalence, meaning the biggest falls are observed for ST73 and ST95, which were actually unperturbed by the invading strains. The greater invasion rate of ST131 relative to ST69 is an artifact of its higher prevalence in the overall data set, meaning it is seeded at a higher level, rather than a true ecological difference.

Factors affecting the emergence of ST131. The estimate of the fitness advantage driving the invasion of ST69 and ST131, r, is contingent on aspects of the model structure. The absence of population genomic data predating 2001, or from other locations, makes it difficult to model the initial invasion of ST131. Therefore, isolates from 2002 onwards were seeded in the initial population at a frequency of 10^{-3} to replicate the likely scenario in which these genotypes were already present in the United Kingdom or Ireland in 2001 but not captured by the first sample. Increasing the level at which these strains were seeded by 10-fold resulted in a slightly increased final frequency (see Fig. S4). This frequency should represent an approximate upper bound, beyond which it would be unlikely that the strains would have remained unsampled in 2001. Decreasing the seeding level had an almost negligible effect on the simulated dynamics. Hence, the inferred parameters are not markedly sensitive to the seeding level used in the first time step of the simulations.
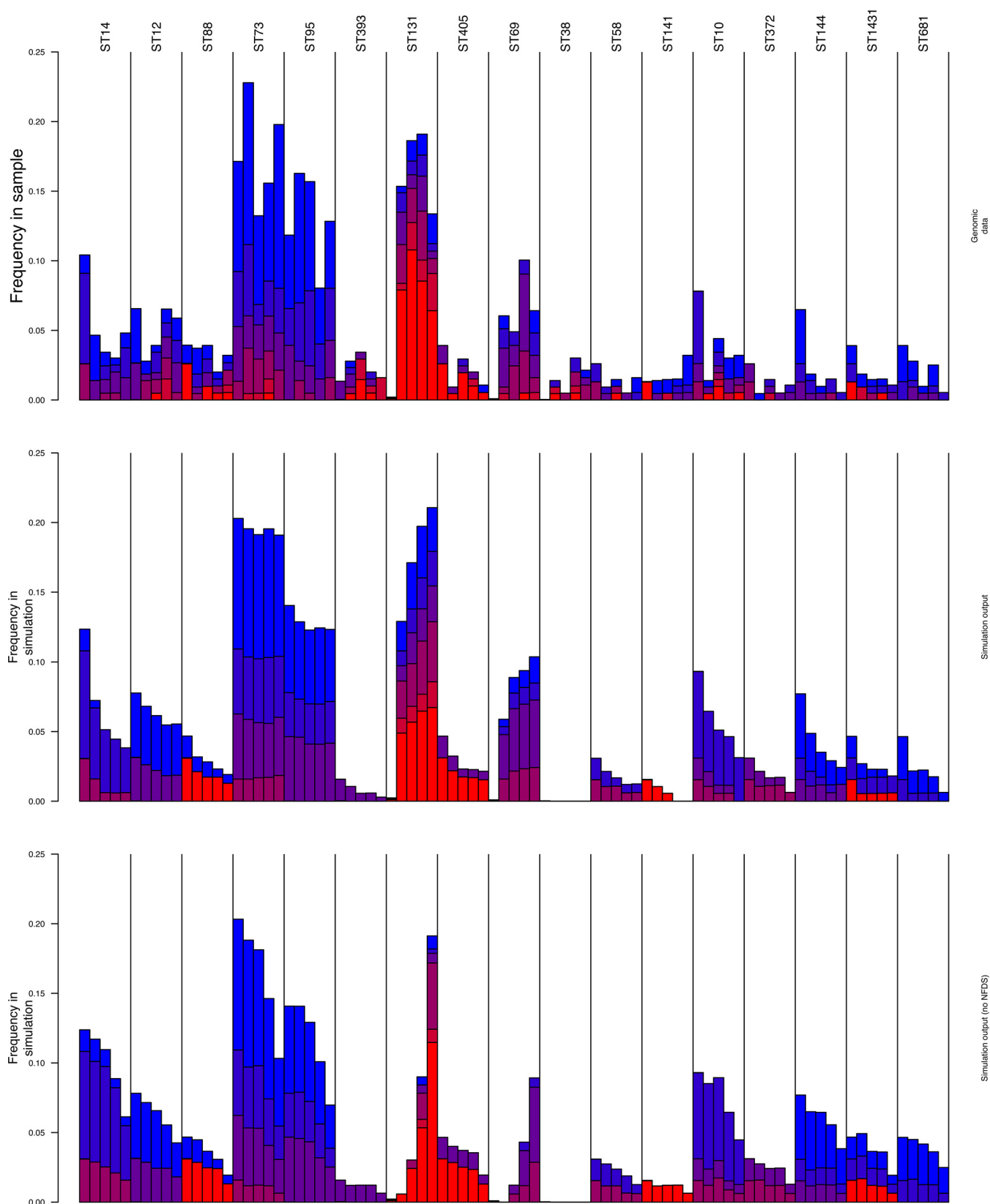
FIG 2 Simulations of changes in the BSAC extraintestinal pathogenic *E. coli* population evolving under multilocus NFDS. Genomic data (top) and median frequencies (middle) observed from 100 simulations run with the best-matching parameter set identified by fitting the model with BOLFI. This corresponded to $\sigma_f = 0.029$, $r = 0.179$, $m = 0.001$, $p_f = 0.425$, and $\sigma_w = 0.0048$. Each column corresponds to a sequence cluster identified by hierBAPS (see Materials and Methods) and is annotated with the predominant sequence type with which it is associated. Each bar indicates the frequency of the sequence cluster in

(Continued on next page)

The estimate of *r* was also affected by the simulated population size. To overcome the greater stochasticity inherent to smaller populations, *r* must be higher for ST69 and ST131 for them to consistently rise from their low initial frequencies in the early phases of the simulations. Hence, the same *r* with a 10-fold smaller population than was computationally feasible for fitting (5×10^4) results in lower median frequencies for ST69 and ST131, whereas maintaining *r* with a 20-fold increased population size results in the emerging strains reaching a higher prevalence by the end of the simulations (see Fig. S5A). Hence, as the effective population size of *E. coli* across the United Kingdom and Ireland is likely to be greater than 5×10^4, the fitness advantage driving these emerging strains to invade is probably lower than the *r* estimated here.

The emergence driven by *r* is limited by NFDS, which generally constrains the invasion of new clusters of genetically coherent isolates, as the traits with which they are associated have a smaller net benefit as they increase in frequency. Simulations in which ST69 and ST131 have the same fitness benefit, *r*, in the absence of NFDS confirm this advantage would be sufficient to dominate the population in a neutrally evolving population (Fig. S5B).

This inhibition would be expected to most strongly affect lineages with very similar sets of accessory loci. As the MDR isolates of clade C share the most recent common ancestor of the three clades, around 25 years ago (11), their expansion would be expected to be subject to the strongest inhibition by NFDS. Their rapid rise suggests the effects of NFDS may have been ameliorated by the diversification of accessory loci, such that the per locus constraint of the equilibrium frequencies does not limit the strain's overall prevalence. Extensive variation suggestive of this diversity has previously been noted with the capsules expressed within clade C (23). Therefore, the distributions of intermediate frequency loci within each sequence cluster and the clades of ST131 were analyzed by estimating the overall gene content through a Heap's law analysis of pangenome curves (24) and through pairwise analysis of gene content similarity by estimating genomic fluidity (25). Comparing the metrics across the sets of isolates revealed that ST131 was one of a small number of clades that had an open pangenome but low genomic fluidity (Fig. S4B). This implies each pair of isolates is similar, reflecting their recent shared ancestry; but, many isolates have undergone different diversifying recombinations at one or more loci present at an intermediate frequency in the wider population. This can account for the overall clade containing many loci in total, despite its recent origin. This is consistent with the adaptive radiation of a successful strain expanding in its population size.

To better understand this diversification and the basis for the high fitness of ST131 represented by *r*, a comprehensive genomic data set encompassing all known ST131 genome sequences was created to understand the unique characteristics of the ST131 lineage, with particular focus on the successful clades B and C.

Core and accessory genomic structure of the ST131 population. A maximum likelihood phylogeny generated from an alignment of concatenated core coding DNA sequences (CDS) from all 862 genomes confirmed the earlier consensus three-clade structure of the lineage (Fig. 3A), and in agreement with previous studies, there was no strong phylogeographic signal or host source clustering evident in the phylogeny (https://microreact.org/project/BJKoeBt2b). To confirm that the collation of the 862 genomes was consistent with previous descriptions of the accessory genome distribution in ST131, isolate relatedness based on shared accessory gene content was visualized as a two-dimensional projection using PANINI (Fig. 3B) (26). Clades A and B largely resided in dense clusters at the periphery of the projection. In contrast, clade C isolates were more diffuse, overlapping with some clade B isolates, forming a cloud with

FIG 2 Legend (Continued)
consecutive time periods, from left to right. The bars are colored according to the number of antibiotic resistance phenotypes associated with the isolates within the sequence cluster at different time points. (Bottom) The equivalent best fit in the absence of NFDS. Only sequence clusters reaching a frequency of at least 2.5% at one time point in the genomic sample are shown; the full results of the simulation, including measures of between-simulation variation, are shown in Fig. S3.
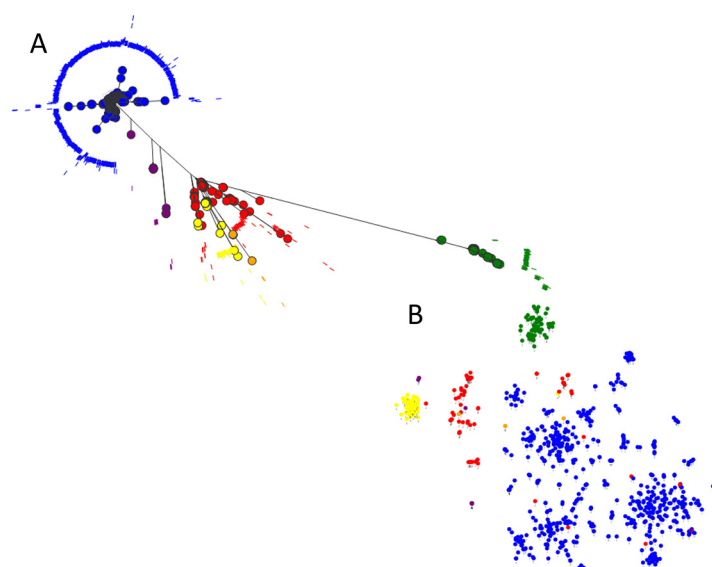
FIG 3 (A) Maximum likelihood phylogeny of 862 *E. coli* ST131 strains. The phylogeny was inferred using RAxML with a GTR Gamma model of substitution, on an alignment of concatenated core CDS as determined by Roary. (B) PANINI plot of the accessory genome content of all 862 strains based on a tSNE plot. The plot is a diagrammatical representation of the relatedness of each strain based on the presence/absence of accessory genes and is presented as a two dimensional representation. The taxa are color coded by BAPS grouping (Table S1) and show clade A (green, BAPS-3), clade B (red, yellow, and purple, BAPS-2, -4, and -5, respectively), and clade C (blue, BAPS-1).

discernible substructuring into distinct groups. This concurs with the previous analysis of the gene content of clade C and the previous finding of multiple accessory genome subclusters within this set of isolates (14).

Low-frequency accessory genes suggest clade A and clade B/C *E. coli* ST131 rarely cohabit the same ecological spaces. To identify which aspects of the accessory genome differed between the clades of ST131, the distributions of the 32,631 sets of orthologous genes identified by Roary were analyzed. Characterizing the full set of loci present at intermediate frequencies was not feasible, as even focusing on the 3,354 present at between 5% and 95% frequency revealed that the majority of these were present at a frequency <20% (see Fig. S1). Therefore, the search was refined to clade-specific genes occurring at a frequency of >95% in one clade but at <5% in the other two clades.

In both clade A and clade B/C, the overwhelming majority of low-frequency accessory genes encode hypothetical proteins (64.4% clade A, 58% clade B/C). Excluding the hypothetical proteins from the analysis showed an unexpected bias in functional gene categories differentially observed in the lineages (Fig. 4). The most common gene types were functional phage, plasmid, and other mobile genetic element (MGE) genes, with more private phage genes present in clade B/C than in clade A. Conversely, there were more private plasmid genes in clade A than in clade B/C, despite the presence of a diverse number of MDR plasmids within clade C (14). Together, this suggests that clade A strains of *E. coli* ST131 and clade B/C strains of *E. coli* ST131 are exposed to different plasmid and phage pools, an observation which is most parsimoniously explained by them rarely sharing the same ecological habitats at the same time.

Clade A contained the highest number of loci exclusive to a lineage (27) despite constituting the least-sampled clade. This is likely explained by the fact that clade A is on the longest branch of the ST131 phylogeny, and as such, has had most evolutionary time to accumulate differences in the gene repertoire. Clade B had only 2 exclusive loci and clade C had 18. When clades B and C were combined against clade A, there were 60 loci exclusively present in the B/C combination. The majority of clade A private genes encode hypothetical proteins, while those private to clade C encode DNA
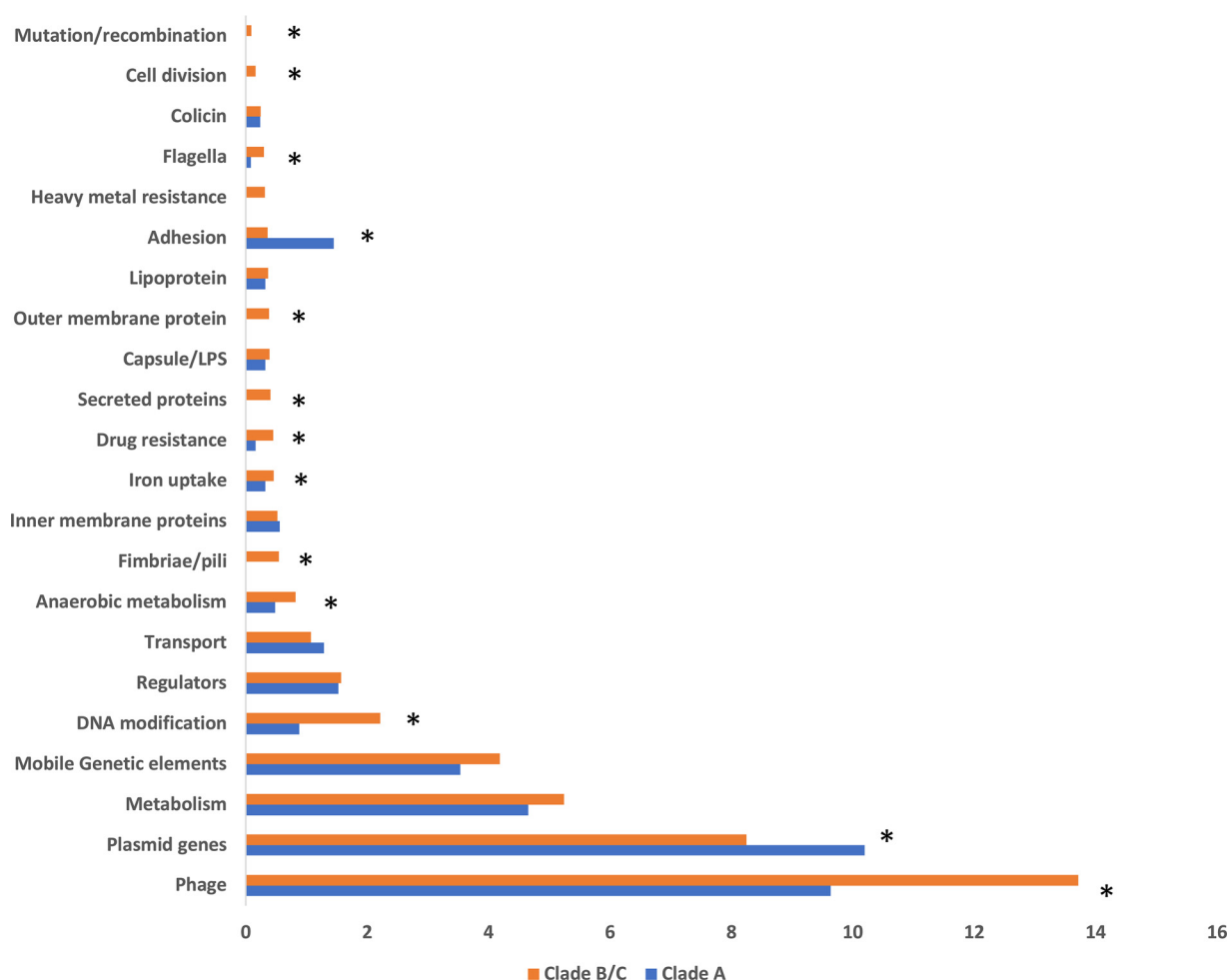
FIG 4 Bar chart depicting functional classes of accessory genes differentially present in clade A and clade B/C *E. coli* ST131. Functional classes are based on GO classes as described in Materials and Methods. *, significant difference exists between clade A and clade C as determined by *t* test.

modification proteins and metabolic functions. The genes private to clade B/C combined also encode hypothetical proteins and metabolic functions, notably, five dehydrogenase enzymes involved in anaerobic metabolism labeled *yihV*, *garR_3*, *fadJ*, *fdhD*, and *gnd* in our data set. A BLAST analysis against the NCBI nonredundant database suggested that the dehydrogenase enzyme gene annotated as *pdxA* in our Roary data set was confined to clade C ST131 strains. These dehydrogenase enzyme genes were found to be present across phylogroup B2 *E. coli* strains (of which ST131 is a member) through blastn searches of the NCBI nonredundant database. Therefore, these loci are not unique to clade C ST131 and were either acquired by an ancestral clade B/C strain or have been lost by clade A.

High diversity in core anaerobic metabolism genes unique to clade B/C. An analysis of accessory loci private to clade B/C (present in >95% of that population) identified two separate loci encoding 3-hydroxyisobutyrate dehydrogenase enzymes and loci encoding 3-hydroxyacyl-coenzyme A (3-hydroxyacyl-CoA) dehydrogenase, 6-phosphogluconate dehydrogenase, and formate dehydrogenase. An analysis of clade B/C loci circulating at a low frequency of <20% also identified a significant overrepresentation of genes encoding dehydrogenase enzymes involved in anaerobic metabolism (a total of 64 loci), including seven variants of formate dehydrogenase. There were also seven variants of the *eutA* gene found in the ethanolamine utilization pathway (the *eut* operon) and a distinct version the *cobW* gene, which encodes the
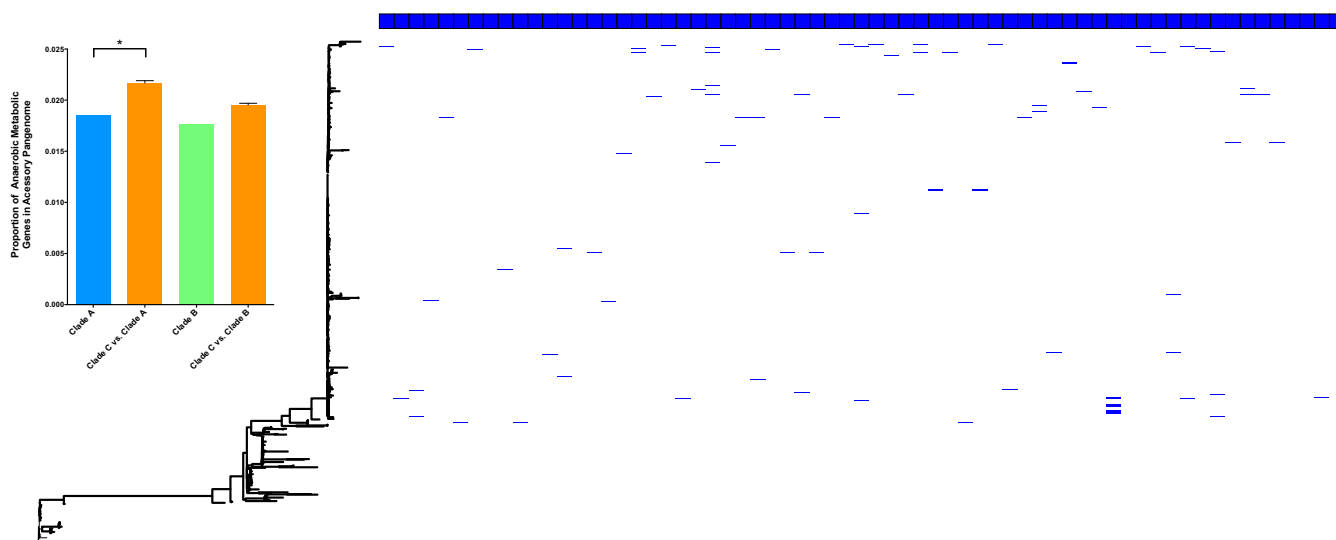
FIG 5 Annotation of a maximum likelihood phylogeny of *E. coli* ST131, based on concatenated core CDS, with the presence of alternative alleles of 64 loci involved in anaerobic metabolism. Each blue box along the top of the tree annotation represents an individual anaerobic metabolism gene, and its presence in the ST131 population is indicated by a blue line. The inset is a bar chart displaying the proportions of the accessory pangenome that are occupied by genes involved in anaerobic metabolism for ST131 clade A, clade B, subsampled clade C versus clade A, and subsampled clade C versus clade B. $P = 0.042$ for clade C versus clade A and $P = 0.086$ for clade C versus clade B. Error bars represent standard errors of the means. Significance was determined using the median value P value from chi-square tests performed on random subsamples of the C clade.

sensor kinase for activation of the cobalamin biosynthesis operon. These anaerobic metabolism genes would normally be considered essential to *E. coli*. Interrogation of the pangenome matrix for the data set showed that versions of these genes existed as core genome loci. Comparison of the nucleotide sequences of the clade-specific loci with their core genome counterparts showed that these were not genes private to clade B/C *per se* but rather represented multiple unique alleles of genes that are core to the ST131 population, which differ at the nucleotide sequence level by more than 5%. This suggests a unique selection pressure is acting on these core genes in clade B/C compared to those in clade A. The discovery of these core genes that display allelic variation is explained by the fact that the percent nucleotide identity cutoff applied in the formation of our pangenome matrix is 95%, meaning that any alternative alleles of core genes differing by 5% or more will appear in our matrix as accessory genes.

Further scrutiny of low-frequency loci in clade B/C also identified alternative alleles of a large number of well-characterized extraintestinal pathogenic *E. coli* virulence-associated genes: antigen 43 (7 alternative alleles); heavy metal resistance, such as arsenic (5 loci), copper (4 loci), and mercury (5 loci); capsule biosynthesis (20 loci); cell division and septation (14 loci); antibiotic resistance to chloramphenicol (3 loci), macrolides (2 loci), rifampin (1 locus), and MDR efflux pumps (21 loci); iron acquisition (39 loci); curli and type I fimbriae and P pili (42 loci); lateral and classical flagella (26 loci); and LPS synthesis (9 loci). These variant loci represent alternative alleles of genes found widely across the *E. coli* phylogeny, indicating there are multiple allelic variants of important genes that are confined to clade B/C of the *E. coli* ST131 lineage.

We sought to determine the distribution of this allelic diversity across the *E. coli* ST131 phylogeny by annotating the tips of the phylogenetic tree with the presence/absence of each of the anaerobic metabolism (Fig. 5) and capsule, cell division, MDR efflux, iron acquisition, pili, and flagellum divergent loci (Fig. 6). Our analysis shows that each alternative allele occurs at very low frequency, but that alleles are randomly distributed throughout the phylogeny of the C clade and are exclusive to clade C. An analysis of the source attribution of isolates displaying this allelic diversity showed that 80% of isolates displaying allelic diversity were human isolates (32% of those bloodstream isolates, 20% UTI isolates, 2% commensal, and the remainder of an unknown source), 5% were avian isolates, 10% companion animal isolates, and the remainder
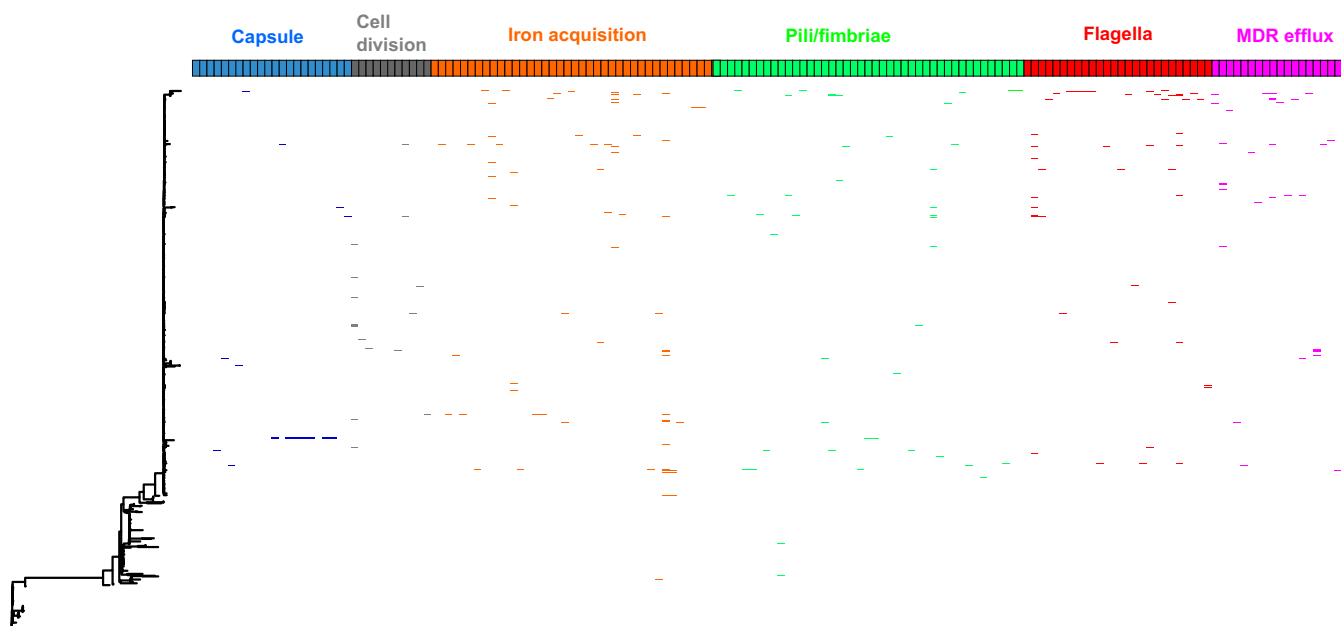
FIG 6 Annotation of a maximum likelihood phylogeny of *E. coli* ST131, based on concatenated core CDS, with the presence of alternative alleles of loci involved in capsule production, cell division, iron acquisition, pili/fimbriae production, flagella, and MDR efflux pumps. Each box represents an individual gene, and its presence in the ST131 population is indicated by an appropriately colored line.

from other sources. Given that these alleles differ from the normal conserved versions of genes by >5% at the nucleotide level, it is implausible that these alleles would be arising repeatedly and independently via mutation. Instead, the most parsimonious explanation is that the minor frequency alternative alleles are being distributed through the population via recombination. This conclusion is supported by the fact that every one of the allele variants identified in our analysis has 100% nucleotide identity matches with genes present in other *E. coli* in the NCBI nonredundant database.

Given that our data set is biased toward clade C genomes, we performed comparative analyses of the frequency with which allelic diversity occurs in anaerobic metabolism genes. We randomly subsampled clade C 100 times and compared equal numbers of clade A, B, and C genomes for allelic diversity. Our data show that even when randomly subsampling clade C, the levels of diversity observed in anaerobic metabolism genes is significantly higher than in clade A, providing evidence that the accumulation of sequence diversity is specific to the MDR clade C, despite this clade having the most recent common ancestor of the three (Fig. 5).

Finally, we sought to exclude the possibility that the presence of these allelic variants was skewed by some form of geographically localized expansion of variants. To do this, we compared the relative frequencies of all accessory genes, highlighting the allele variants in anaerobic metabolism, capsule, cell division, MDR efflux, iron acquisition, fimbriae, and flagella present in U.K. versus non-U.K. isolate genomes (see Fig. S6). Our data showed a strong linear relationship between the frequency of genes in the two populations, indicating that the data were not biased by expansion of alleles in a given geographical location, and that this accumulated diversity was equally as likely to happen in any given strain independent of its geographical origin.

Allelic diversity of anaerobic metabolism genes in clade C ST131 is not observed in other dominant ExPEC lineages. The possibility exists that the above observations made for clade C of *E. coli* ST131 simply reflect the general evolutionary path of a successful commensal and extraintestinal pathogen. To test this, we performed an identical analysis on the pangenome of 261 ST73 isolates and of 160 ST95 isolates from the U.K. BSAC population survey (16). *E. coli* ST73 and ST95 represent two
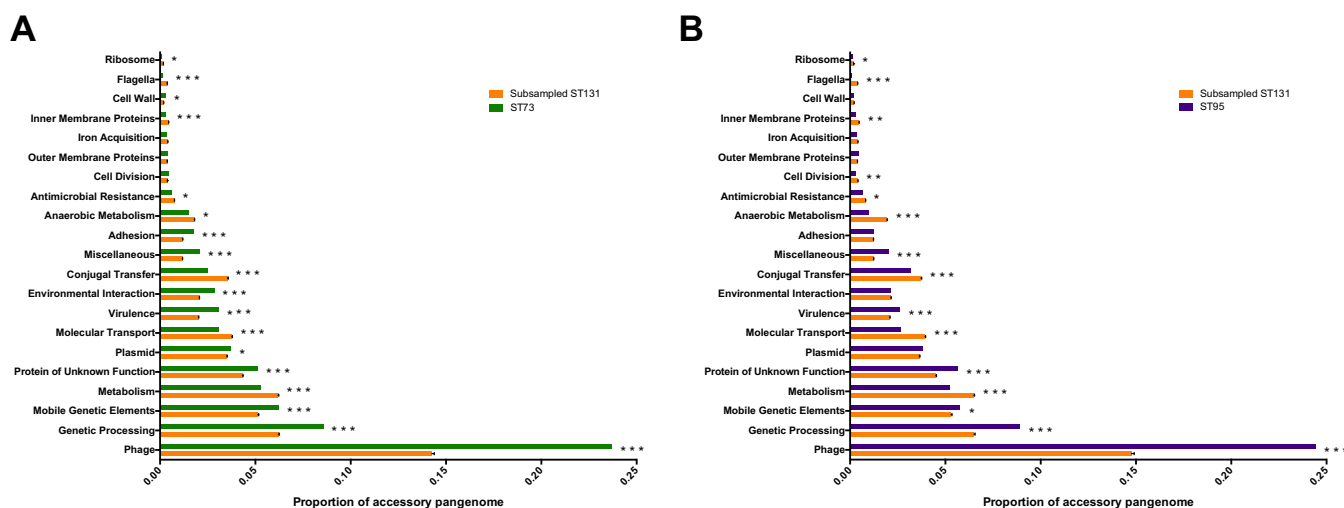
FIG 7 Bar charts depicting the compositions of the accessory genomes of ST73 (A) and ST95 (B) compared to a repetitively sampled clade C ST131. The proportions of the accessory genome are plotted against manually assigned functional categories. Hypothetical proteins are responsible for the majority of the accessory pangenome and are omitted from the graphs. Error bars are standard errors of the means. Iterative chi-square tests were performed to assess significance, as described in Materials and Methods. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

of the most dominant lineages associated with clinical extraintestinal disease alongside ST131 (5, 16) but are predominantly non-MDR lineages and rarely associated with MDR plasmids (16). As with our interclade comparisons, we randomly subsampled clade C ST131 100 times to allow equal numbers of genomes per lineage to be compared. Our analysis showed higher levels of plasmid, phage, and hypothetical proteins in the accessory genomes of both ST73 and ST95 than in ST131 (Fig. 7). ST73 and ST95 displayed similar ratios of alternative alleles in adhesins, cell division, and septation genes and multiple iron acquisition genes as observed in ST131 but a higher diversity in virulence-associated genes. However, enrichment in allelic variation in anaerobic metabolism genes was significantly higher in any given subsampled set of clade C ST131 genomes than in both lineages. This supports the hypothesis that the observation of increased diversity accumulating in anaerobic metabolism genes is not a more general extraintestinal pathogenic *E. coli* trait but is particularly enriched in the ST131 lineage. It also suggests that this is independent of diversification of adhesins or virulence factors, which was enhanced in the ST73 and ST95 lineages.

The accumulation of nucleotide diversity in a given set of loci can often be interpreted as a signature of some form of selection occurring on those genes. However, the low levels of frequency of any given allele across clade C strains contradict a hypothesis for positive selection, where one would expect successful or beneficial alleles to sweep to a high frequency or fixation. Indeed, a comparison of the sequences of each of anaerobic metabolism loci in which three or more alleles were observed identified just three loci which showed a mixture of weakly positive and highly negative Tajima's D scores, indicating no common selective mechanism in these loci (Table S2b).

However, these results can be reconciled with a lineage evolving under NFDS. Different resource use strategies can facilitate coexistence between competing strains, such those cocolonizing a host, resulting in frequency-dependent selection (28, 29). This would explain the sustained intermediate frequencies of genes encoding dehydrogenases over multiple years (see Fig. S7). Hence, this diversification of metabolic loci could represent the adaptive radiation of particular traits within a successful genetic background, able to efficiently compete with the resident *E. coli* population through a diverse panel of metabolic capacities suited to exploiting resources under anaerobic conditions.

DISCUSSION

The evolutionary events that led to the emergence of *E. coli* ST131 have been an intense focus of research, with consensus opinion suggesting that following the acquisition of key ExPEC virulence factors, the acquisition of fluoroquinolone resistance in the 1980s by the clade C sublineage of ST131 was a key event in that emergence (11, 12). However, a recent nationwide U.K. population survey rejected this hypothesis and suggested that success of the major ExPEC clones is not dictated by resistance traits (16). Here, we combined an analysis of population frequency dynamics with an analysis of the pangenome of the ST131 lineage, strongly suggesting that this species' population structure and dynamics are shaped by NFDS acting on genomic islands. Such multilocus NFDS is able to account for how an otherwise stable population was disrupted by the invasion of ST131 and ST69, displacing some lineages while leaving other, largely antibiotic-susceptible, genotypes at almost untouched prevalence.

Previous work has suggested that clade C strains of *E. coli* ST131 undergo reduced levels of detectable core genome recombination compared to those of other phylogroup B2 *E. coli* (30) or ST131 clade A strains (14). We have previously postulated that this may be a result of ecological separation between clade C strains and other common ExPEC (14, 30). Our analysis of nearly 900 genomes has allowed us to interrogate accessory gene movement to a far greater resolution than previously possible. From the analysis of the accessory genome, we identified thousands of plasmid, phage, and other mobile genetic element genes which are private to clade A and the combined clade B/C. Such an observation is a classic signature of ecological separation of the two populations (31, 32), particularly given that the genetic distance between clade A and clade B/C is much smaller than between other lineages and species from which the circulating genes are also found in the NCBI nonredundant database.

Our analysis also identified a significantly increased level of sequence diversity in genes involved in key host colonization processes in clade C. This diversity was uncovered through our pangenome analysis as allelic variants of core genes. Primary among these is a large number of genes involved in anaerobic metabolism, including seven allelic variants of the formate dehydrogenase gene, as well as allelic variants of genes involved in ethanolamine utilization and cobalamin biosynthesis. The pivotal role of ethanolamine production and cobalamin biosynthesis in the ability of Gram-negative pathogens to outcompete bacteria in the human intestine is well documented (33, 34), and this phenomenon only occurs when supported by an increased ability to perform anaerobic respiration in the presence of inflammation (33). It has been shown that MDR *E. coli* ST131 is able to colonize the gastrointestinal tract of humans for months or years in the absence of antibiotic selection (35, 36) and that this colonization results in a displacement of the *E. coli* colonizing the host prior to exposure to the MDR strain (35).

While this diversity in anaerobic metabolism genes was unique to clade C ST131, the allelic variation observed in other human colonization and virulence factors, such as iron acquisition, fimbriae, and cell division, was also observed in two of the other most commonly isolated lineages of *E. coli* from extraintestinal infections, ST73 and ST95. This diversity likely reflects selection occurring on genes important for ExPEC pathogenesis. Iron acquisition is well characterized as a key virulence determinant in ExPEC, with the ability to initiate successful UTI completely abrogated in the absence of functional iron acquisition systems (37). Recent experimental vaccine work exploiting siderophore production by ExPEC has shown it to be highly effective in rodent models on ExPEC UTI (38). The importance of iron acquisition can also explain many of the MDR efflux allele variants seen in this data set, with half occurring in the *acrD* gene, which has been experimentally shown to play a role in iron acquisition in *E. coli* (39). We identified multiple alleles of genes in the type 1 fimbria operon and in genes in the P pilus operon, which are classical virulence determinants in UTI (40), and multiple genes involved in capsule biosynthesis, which we have previously reported as being a hot spot for recombination in *E. coli* ST131 (13, 41). We also identified multiple alleles of

genes involved in controlling incomplete septation and filamentous growth, which is a crucial process in the formation of the filamentous intracellular bacterial communities (IBCs) which are thought to be fundamental in the ability of ExPEC to survive inside bladder epithelial cells and cause UTI (42). There are a small number of allelic variants in anaerobic metabolism genes also present in ST73 and ST95, possibly reflecting recent experimental studies suggesting a crucial role for the cytochrome *bd* oxidase system in the ability to cause urinary tract infection (43). Also, previous studies using saturated mutagenesis techniques and studying global transcriptional patterns during urinary tract infection of ExPEC strains have suggested a key role for dehydrogenase enzymes involved in anaerobic metabolism in the ability to cause pathology in the mammalian urinary tract (27, 44, 45).

Recent modeling data on why drug-resistant and drug-susceptible populations of bacteria coexist highlighted that any factors which increase the duration of colonization in a human host will also increase the selective pressure for them to evolve antibiotic resistance (23). Hence, both the success of ST131 in invading the population and the association of many isolates in this lineage with an MDR phenotype would be consistent with the accumulating diversity in anaerobic metabolism loci facilitating enhanced persistence within its host, perhaps through an improved ability to outcompete resident commensal *E. coli* strains. The fact that this selection is only seen in clade C of ST131 suggests that this occurred around the time of the emergence of the lineage as a human clinical threat (13) alongside the development of fluoroquinolone resistance. Subsequent acquisition of MDR plasmids, and the consequent selection for an ability to offset the fitness costs of long-term MDR plasmid maintenance (14), is likely to have occurred as a result of prolonged exposure to selective antibiotic environments during colonization of humans. Nevertheless, neither anaerobic metabolism genes nor antibiotic resistance loci have swept to fixation in ST131, reflecting their fluctuating but stable prevalence in the broader *E. coli* population.

This diversification can instead be explained by NFDS, under which these genes are beneficial when rare, because they provide an advantage over cocolonizing strains which will typically lack the same metabolic capacities. However, as these traits become more common as ST131 expands, representatives of this lineage will more commonly encounter one another, therefore necessitating further diversification for different clade C representatives to sustain their advantage over competitors. Similarly, the capsule locus diversification previously observed within clade C, resulting in the capsule synthesis locus corresponding to a "hot spot" of recombination (23), could result from NFDS on variable antigens (46), with the host immune system selecting for a diversity of capsule structures as the dominant type becomes more common following ST131's emergence (16).

This study presents evidence for two complementary aspects of diversification within the *E. coli* ST131 lineage. The first is ecological niche separation across a lineage, whereby the different clades of ST131 find themselves competing for similar niches, resulting in them very rarely cohabiting the same habitat. This then results in the formation of distinct subclades within ST131. The second is NFDS across a species within the specific environment of the human gut and the diversification of key genes involved in the colonization of the human gut niche to escape the constraints imposed by NFDS. Further studies are required to fully determine the extent to which niche separation and NFDS are either separate or linked processes. Determining whether loci subject to NFDS are also those that determine niche adaptation will be integral to this process. Understanding the processes that govern the epidemiological dynamics of dominant *E. coli* lineages, and those of similar pathogens causing bloodstream infections, is critical for addressing the public health threat of antibiotic resistance.

MATERIALS AND METHODS

Genome data. For the NFDS modeling study, the 1,094 isolates from the BSAC resistance surveillance bacteremia collection from the United Kingdom and Ireland between 2001 and 2011 were used. These were described by Kallonen et al. (16).

For the ST131-specific study, a collection of 862 *E. coli* ST131 genomes (see Table S1 in the supplemental material), of which 684 were previously sequenced as part of phylogenomic investigations

of the ST131 lineage (10, 11, 13, 14, 16, 47), was analyzed. An additional 184 ST131 isolates from the BSAC bacteremia resistance surveillance project conducted in the United Kingdom and Ireland between 2001 and 2011 (indicated in Table S1) were included. Together, this collection represents bacteria isolated from invasive disease (bloodstream infections, 504 isolates; 58.5%), human asymptomatic carriage (21 isolates, 2.4%), disease resulting from intestinal carriage (UTI, 163 isolates; 18.9%), and bacteria isolated from a range of veterinary livestock, pets, wild birds, and the wider environment (67 isolates, 7.8%). This provides a data set isolated between 1967 and 2013, providing as broad a view of isolate origin as possible within existing data sets.

In an attempt to avoid any issues arising from different assembly or annotation metrics employed in the previous projects, only raw sequence data were downloaded in fastq format using the previously published accession data. *De novo* assembly was then performed on all the genomes using Velvet (48) and annotation using Prokka (49) as previously described (16). A pangenome of the entire data set was constructed using Roary with 95% identity cutoff (50). A concatenated core CDS alignment was made from the Roary output, and a maximum likelihood phylogenetic tree was constructed from the alignment using RaxML version 8.2.8 (51) and the GTR model with Gamma rate heterogeneity.

For comparative lineage analysis, 264 ST73 genomes and 162 ST95 genomes that were sequenced and fully characterized as part of the U.K. BSAC genome study (16) were utilized.

Multilocus NFDS modeling. Multilocus NFDS modeling used genomic data from the previous publication analyzing the population dynamic of bloodstream infection *E. coli* isolates in the United Kingdom (16). This first required splitting the population into genetically coherent clusters of isolates. This is necessary, as the sequence cluster (or "strain") frequencies are used for model fitting; therefore, the clusters must be coherent enough to behave reproducibly and large enough to overcome the stochasticity of the model. The previous hierBAPS analysis of the core genome provided three levels of hierarchical clustering (52). The highest-level clusters corresponded to individual common strains in some cases but polyphyletic collections of diverse genotypes in others. Therefore, to ensure the sequence clusters used in the model were of similar levels of diversity, available sequence typing data were used to identify the level of clustering that corresponded to clonal complexes. Each of the highest-level hierBAPS clusters was tested by constructing links between single- and double-locus variants. If the isolates formed a single clonal complex, the cluster was defined as a sequence cluster. Otherwise, the second level of clustering was tested in the same way; if these groups represented a clonal complex, they were used as a sequence cluster. Otherwise, the third most fine-grained level of clustering was used to define a sequence cluster. This identified 56 sequence clusters across the population.

The multilocus NFDS model also requires the definition of a set of L loci, each of which is associated with an equilibrium population-wide frequency, e_l. The majority of these correspond to accessory genetic loci, identified as sets of orthologous sequences by a previous Roary analysis (16). Only the 7,204 present at between 5% and 95% frequency in the first sample, from 2001, were included in L and therefore modelled as evolving toward e_l under NFDS.

The set of L loci was supplemented with seven loci associated with resistance phenotypes if they were present within this frequency range in 2001: amoxicillin, clavulanic acid, ciprofloxacin, cefuroxime, gentamicin, piperacillin-tazobactam, and trimethoprim. The first six of these were directly inferred from the previously published analysis. Trimethoprim was instead inferred from the *sul* and *dfrA* alleles identified by Roary; data from the Cambridge University Hospitals collection (16) was used to train a model constructed with the randomForest R library (https://cran.r-project.org/web/packages/randomForest/), which had 93% accuracy when applied back to the training data set. This was used to infer resistance phenotypes for the BSAC collection.

Analysis used the heterogeneous multilocus NFDS model described previously (17), modified to treat a vaccine cost, v, as a fitness advantage, r. To simulate the invasion of the resident population by ST131 and ST69, all individuals, i, of the two sequence clusters containing ST131 (and related sequence types ST5640, ST5387, ST5494, ST5432, and ST5630) and ST69 (and related sequence type ST106) were assigned the same fitness advantage, $r_i = r$. For all other individuals in the population, $r_i = 0$. Hence, the function defining the number of progeny, $X_{i,t}$ produced by i at time t was

$$X_{i,t} \sim Pois\left(\left(\frac{\kappa}{N_t}\right)(1 + r_i)(1 - m)\left((1 + \sigma_f)^{\pi_{i,t}} + (1 + \sigma_w)^{\omega_{i,t}}\right)\right)$$

In this formula, density-dependent competition is parameterized by the carrying capacity κ, set at 50,000 to represent a large population that is still computationally feasible, and the total number of cells in the simulated population at t, N_t. The strength of NFDS was determined by the parameters p_f, σ_f, and σ_w. As previously, the set of L accessory loci and resistance phenotypes were ordered according to the statistic Δ_l:

$$\Delta_l = \frac{(f_{l,t>0} - e_l)^2}{(1 - e_l(1 - e_l))}$$

where $f_{l,t>0}$ is the mean post-2001 locus frequency. If the L loci and phenotypes considered to be under NFDS were ordered by ascending values of Δ_l, then l_f was the highest ranking locus meeting the criterion $\frac{l_f}{L} \leq p_f$. This determined the strength of NFDS acting on each locus. Therefore, the effect of NFDS on the reproductive fitness of individual i depended on which loci were encoded in its genome, as represented by the binary variable $g_{i,l}$, and on the deviation of their simulated frequencies at time t, $f_{l,t}$, from their corresponding equilibrium frequencies:

$$\pi_{i,t} = \sum_{l=1}^{l_f} g_{i,l} \left(e_l - f_{l,t} \right)$$

and

$$\omega_{i,t} = \sum_{l=l_f+1}^{L} g_{i,l} \left(e_l - f_{l,t} \right)$$

These summed deviations served as the exponents for the NFDS terms of the reproductive fitness, with $\pi_{i,t}$ and σ_f corresponding to those loci under stronger NFDS, and $\omega_{i,t}$ and σ_w corresponding to those loci under weaker NFDS.

The simulations were initialized with a random selection of κ genotypes from the genomic data, which were biased such that those isolates observed in 2001 were represented at 1,000-fold greater frequency than genotypes collected in later years. This was necessary to "seed" the initial population with ST131 and ST69, to facilitate their expansion in a realistic manner in subsequent years. This represents the likely scenario in which they were present in the sampled region in 2001 but were not observed in the necessarily finite set of sequenced isolates. The parameter m represented the rate at which all isolates entered the population through migration; this was biased to import all sequence clusters at the same rate, to avoid any fits in which high rates of migration would artifactually replicate the population observed in the later years of the collection (17).

Model fitting to genomic data. As in Corander et al. (17), the simulation model was fitted through approximate Bayesian computation (ABC) using the BOLFI algorithm, which has been shown to accelerate ABC inference 1,000 to 10,000 times without loss of accuracy (53). The prior constraints placed on the parameter values were as follows: the lower bound on all parameters was set to 0.0009; the upper bounds were 0.99 for r, 0.2 for m, 0.99 for p_f, 0.03 for σ_f, and 0.005 for σ_w. We used 500 iterations of the BOLFI algorithm to minimize the Jensen-Shannon divergence of the sequence cluster frequencies in the genomic data and in the simulations, as ascertained through randomly sampling discrete sets of isolates in accordance with the size and timings of the genomes selected for sequencing from the original collection. Convergence of BOLFI was monitored after every 100 iterations and the approximate likelihood estimate was assessed to have been stabilized after 500 iterations (53). The 95% posterior credible intervals for the parameters were obtained using three generations of sequential Monte Carlo sampling with the same default settings as used in Corander et al. (17). The neutral model was fitted by fixing p_f, σ_f, and σ_w at zero and estimating r and m through 500 iterations of the BOLFI algorithm, followed by sequential Monte Carlo sampling, as with the full model.

Accessory genome analysis. The pangenome matrix from Roary was utilized to investigate the presence of clade-specific loci. The PANINI tool was used with the default setting to visualize the accessory gene-sharing patterns in the population available at https://microreact.org/project/BJKoeBt2b (26). PANINI has been demonstrated to provide efficient complementary visual means to phylogenetic trees to accurately extract both distinct lineages present in a population-wide genomic data set and to highlight clusters within lineages, which are explained by rapidly occurring, homoplasic alterations, such as phage infection. Roary was run on the entire data set using the default 95% sequence identity threshold to cluster genes, allowing the separation of genes based on allelic as well functional differences. Based on a frequency distribution histogram (Fig. S1), a locus as assigned as being clade specific if it occurred at a frequency >95% in one clade and at <5% in the other two clades. Loci identified as clade specific were functionally annotated by performing a tblastn analysis of the nucleotide sequence of the loci against the NCBI nonredundant database.

Functional categorization of pangenomes. To assess the functional composition of the accessory pangenome, Gene Ontology (GO) terms were assigned to gene sequences from the pangenome. Briefly, representative sequences from the pan genome of ST131 were mapped to orthologous groups in the bactNOG database using the eggNOG emapper utility (54) Mapping was performed using the diamond search algorithm. The output from eggNOG was filtered to remove orthologous groups with no GO terms. The number of genes mapping to an orthologous group was used as a score for GO term enrichment analysis.

Comparisons of lineage and clade-specific loci. To compare lineage pangenomes while accounting for differences in the number of genomes, a sampling approach was utilized. Specifically, a subset with a size equal either to the number of ST73 or ST95 genomes was selected at random from the ST131 clade C, while ensuring a cross representation of the phylogenetic diversity of the clade. The functional enrichment of genes in the subset was quantified and statistically compared to the ST73 or ST95 pangenome using a chi-square test. This process was repeated 100 times to produce 100 P values, from which the median P value was calculated. Utilizing the same subsampling approach, the pangenome composition of clade C ST131 genomes was compared to both the clade A and clade B pangenomes.

Chi-square statistical tests were performed to assess the significance of the observed differences in functional enrichment. Briefly, with each iteration of the sampling procedure, a chi-square test was performed using the functional proportion of the subsampled pangenome as the observed value and the proportion for ST73 or ST95 as the expected value. This generated 100 P values from which one can use the average, maximum, or median to assess significance of the observed differences. In addition, proportional Z statistic tests were also performed to assess the significance of the observed difference. The measurements from the 100 replicates of the subsampling procedure were used to generate an average for the proportions as well as to estimate the variance. The tests were conducted using the proportional measurements from ST73 and ST95 as the "true" means and quantifying how distinct the ST131 subsamples were from these reference values.

Analysis of diversity observed in genes involved in anaerobic metabolism. The sequences of 64 anaerobic metabolic genes in which allelic diversity was observed were extracted from individual genomes. The nucleotide sequences were then clustered at 80% identity and 80% length using CD-HIT, which was run using the accurate flag and "word size" of 5 (55). An additional CD-HIT script was used to extract gene sequences for clusters with more than 3 genes, the minimum required by MEGA-CC for analysis. The sequences were then aligned using Muscle with default settings (56). The resulting alignment files were analyzed in MEGA-CC to produce measurements of Tajima's D (57).

ST131 clade-specific SNPs. To visualize the ST131 clades A, B, C, C1, and C2 within the ML tree and the PANINI clustering we identified clade-specific single-nucleotide polymorphisms (SNPs) (Table S1) as previously described (16).

Data availability. Accession numbers for the reads used in this study are listed in Table S1 with information of year and place of isolation and the results of the *in silico* PCR for clade-specific SNPs.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mBio .00644-19.

FIG S1, PDF file, 0.3 MB.
FIG S2, PDF file, 14.5 MB.
FIG S3, PDF file, 0.3 MB.
FIG S4, PDF file, 0.1 MB.
FIG S5, PDF file, 0.1 MB.
FIG S6, JPG file, 0.1 MB.
FIG S7, PDF file, 0.1 MB.
TABLE S1, CSV file, 0.1 MB.
TABLE S2, PDF file, 0.3 MB.

ACKNOWLEDGMENTS

REFERENCES

1. de Kraker MEA, Jarlier V, Monen JCM, Heuer OE, van de Sande N, Grundmann H. 2013. The changing epidemiology of bacteraemias in Europe: trends from the European Antimicrobial Resistance Surveillance System. Clin Microbiol Infect 19:860–868. https://doi.org/10.1111/1469 -0691.12028.

2. Foxman B. 2010. The epidemiology of urinary tract infection. Nat Rev Urol 7:653–660. https://doi.org/10.1038/nrurol.2010.190.

3. Kohler C-D, Dobrindt U. 2011. What defines extraintestinal pathogenic *Escherichia coli*? Int J Med Microbiol 301:642–647. https://doi.org/10 .1016/j.ijmm.2011.09.006.

4. Dobrindt U, Hochhut B, Hentschel U, Hacker JJ. 2004. Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol 2:414–424. https://doi.org/10.1038/nrmicro884.

5. Alhashash F, Weston V, Diggle M, McNally A. 2013. Multidrug-resistant *Escherichia coli* bacteremia. Emerg Infect Dis 19:1699–1701. https://doi .org/10.3201/eid1910.130309.

6. Croxall G, Hale J, Weston V, Manning G, Cheetham P, Achtman M, McNally A. 2011. Molecular epidemiology of extraintestinal pathogenic *Escherichia coli* isolates from a regional cohort of elderly patients highlights the prevalence of ST131 strains with increased antimicrobial resistance in both community and hospital care settings. J Antimicrob Chemother 66:2501–2508. https://doi.org/10.1093/jac/dkr349.

7. Banerjee R, Johnson JR. 2014. A new clone sweeps clean: the enigmatic emergence of *Escherichia coli* sequence type 131. Antimicrob Agents Chemother 58:4997–5004. https://doi.org/10.1128/AAC.02824-14.

8. Peirano G, Schreckenberger PC, Pitout J. 2011. Characteristics of NDM-

1-producing *Escherichia coli* isolates that belong to the successful and virulent clone ST131. Antimicrob Agents Chemother 55:2986–2988. https://doi.org/10.1128/AAC.01763-10.

9. Mathers AJ, Peirano G, Pitout J. 2015. The role of epidemic resistance plasmids and international high-risk clones in the spread of multidrug-resistant *Enterobacteriaceae*. Clin Microbiol Rev 28:565–591. https://doi .org/10.1128/CMR.00116-14.

10. 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-β-lactamase-producing *Escherichia coli* ST131 is driven by a single highly pathogenic subclone, H30-Rx. mBio 4:e00377-13. https://doi.org/10.1128/mBio.00377-13.

11. Stoesser N, Sheppard AE, Pankhurst L, de Maio N, Moore CE, Sebra R, Turner P, Anson LW, Kasarskis A, Batty EM, Kos V, Wilson DJ, Phetsouvanh R, Wyllie D, Sokurenko E, Manges AR, Johnson TJ, Price LB, Peto TEA, Johnson JR, Didelot X, Walker AS, Crook DW, Modernizing Medical Microbiology Informatics Group (MMMIG). 2016. Evolutionary history of the global emergence of the *Escherichia coli* epidemic clone ST131. mBio 7:e02162. https://doi.org/10.1128/mBio.02162-15.

12. Ben Zakour NL, Alsheikh-Hussain AS, Ashcroft MM, Nhu NTK, Roberts LW, Stanton-Cook M, Schembri MA, Beatson SA, Khanh Nhu NT, Roberts LW, Stanton-Cook M, Schembri MA, Beatson SA. 2016. Sequential acquisition of virulence and fluoroquinolone resistance has shaped the evolution of *Escherichia coli* ST131. mBio 7:e00347-16. https://doi.org/10.1128/mBio .00347-16.

13. Petty NK, Ben Zakour ZN, Stanton-Cook M, Skippington E, Totsika M, Forde BM, Phan MD, Gomes Moriel D, Peters KM, Davies M, Rogers BA, Dougan G, Rodriguez-Baño 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 111:5645–5649. https://doi.org/10.1073/pnas.1322678111.

14. McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, Sreecharan T, Vehkala M, Valimaki N, Prentice MB, Ashour A, Avram O, Pupko T, Dobrindt U, Literak I, Guenther S, Schaufler K, Wieler LH, Zhiyong Z, Sheppard SK, McInerney JO, Corander J. 2016. Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. PLoS Genet 12: e1006280. https://doi.org/10.1371/journal.pgen.1006280.

15. Ren C-P, 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 187:1430–1440. https://doi.org/10.1128/JB.187.4.1430-1440.2005.

16. Kallonen T, Brodrick HJ, Harris SR, Corander J, Brown NM, Martin V, Peacock SJ, Parkhill J. 2017. Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. Genome Res 27:1437–1449. https://doi.org/10.1101/gr.216606.116.

17. Corander J, Fraser C, Gutmann MU, Arnold B, Hanage WP, Bentley SD, Lipsitch M, Croucher NJ. 2017. Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. Nat Ecol Evol 1:1950–1960. https://doi.org/10.1038/s41559-017-0337-x.

18. Clark G, Paszkiewicz K, Hale J, Weston V, Constantinidou C, Penn C, Achtman M, McNally A. 2012. Genomic analysis uncovers a phenotypically diverse but genetically homogeneous *Escherichia coli* ST131 clone circulating in unrelated urinary tract infections. J Antimicrob Chemother 67:868–877. https://doi.org/10.1093/jac/dkr585.

19. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo S, Weiser JN, Corander J, Bentley SD, Croucher NJ. 2019. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Res 29:304–316. https://doi.org/10.1101/gr.241455.118.

20. Croucher NJ, Coupland PG, Stevenson AE, Callendrello A, Bentley SD, Hanage WP. 2014. Diversification of bacterial genome content through distinct mechanisms over different timescales. Nat Commun 5:5471. https://doi.org/10.1038/ncomms6471.

21. Azarian T, Grant LR, Arnold BJ, Hammitt LL, Reid R, Santosham M, Weatherholtz R, Goklish N, Thompson CM, Bentley SD, O'Brien KL, Hanage WP, Lipsitch M. 2018. The impact of serotype-specific vaccination on phylodynamic parameters of *Streptococcus pneumoniae* and the pneumococcal pan-genome. PLoS Pathog 14:e1006966. https://doi.org/10.1371/journal.ppat.1006966.

22. Davies NG, Flasche S, Jit M, Atkins KE. 2017. Within-host dynamics explain the coexistence of antibiotic-sensitive and resistant bacteria. bioRxiv https://doi.org/10.1101/217232

23. Lehtinen S, Blanquart F, Croucher NJ, Turner P, Lipsitch M, Fraser C. 2017. Evolution of antibiotic resistance is linked to any genetic mechanism affecting bacterial duration of carriage. Proc Natl Acad Sci U S A 114: 1075–1080. https://doi.org/10.1073/pnas.1617849114.

24. Tettelin H, Riley D, Cattuto C, Medini D. 2008. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11:472–477. https://doi.org/10.1016/j.mib.2008.09.006.

25. Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. 2011. Genomic fluidity: an integrative view of gene diversity within microbial populations. BMC Genomics 12:32. https://doi.org/10.1186/1471-2164-12-32.

26. Abudahab K, Prada JM, Yang Z, Bentley SD, Croucher NJ, Corander J, Aanensen DM. 2014. PANINI: pangenome neighbor identification for bacterial populations. Microb Genom 2018:4. https://doi.org/10.1099/mgen.0.000220.

27. Wiles TJ, Norton JP, Russell CW, Dalley BK, Fischer KF, Mulvey MA. 2013. Combining quantitative genetic footprinting and trait enrichment analysis to identify fitness determinants of a bacterial pathogen. PLoS Genet 9:e1003716. https://doi.org/10.1371/journal.pgen.1003716.

28. Stewart FM, Levin BR. 1973. Partitioning of resources and the outcome of interspecific competition: a model and some general considerations. Am Nat 107:171–198. https://doi.org/10.1086/282825.

29. Friesen ML, Saxer G, Travisano M, Doebeli M. 2004. Experimental evidence for sympatric ecological diversification due to frequency–dependent competition in *Escherichia* coli. Evolution 58:245–260. https://doi.org/10.1111/j.0014-3820.2004.tb01642.x.

30. McNally A, Cheng L, Harris SR, Corander J. 2013. The evolutionary path to extraintestinal pathogenic, drug-resistant *Escherichia coli* is marked by drastic reduction in detectable recombination within the core genome. Genome Biol Evol 5:699–710. https://doi.org/10.1093/gbe/evt038.

31. Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabo G, Polz MF, Alm EJ. 2012. Population genomics of early events in the ecological differentiation of bacteria. Science 336:48–51. https://doi.org/10.1126/science.1218198.

32. Reuter S, Corander J, de Been M, Harris S, Cheng L, Hall M, Thomson NR, McNally A. 2015. Directional gene flow and ecological separation in *Yersinia enterocolitica*. Microb Genom 1:e000030. https://doi.org/10.1099/mgen.0.000030.

33. Winter SE, Thiennimitr P, Winter MG, Butler BP, Huseby DL, Crawford RW, Russell JM, Bevins CL, Adams LG, Tsolis RM, Roth JR, Baumler AJ. 2010. Gut inflammation provides a respiratory electron acceptor for *Salmonella*. Nature 467:426–429. https://doi.org/10.1038/nature09415.

34. McNally A, Thomson NR, Reuter S, Wren BW. 2016. 'Add, stir and reduce': *Yersinia* spp. as model bacteria for pathogen evolution. Nat Rev Microbiol 14:177–190. https://doi.org/10.1038/nrmicro.2015.29.

35. Arcilla MS, van Hattem JM, Haverkate MR, Bootsma MCJ, van Genderen PJJ, Goorhuis A, Grobusch MP, Lashof AMO, Molhoek N, Schultsz C, Stobberingh EE, Verbrugh HA, de Jong MD, Melles DC, Penders J. 2017. Import and spread of extended-spectrum beta-lactamase-producing *Enterobacteriaceae* by international travellers (COMBAT study): a prospective, multicentre cohort study. Lancet Infect Dis 17:78–85. https://doi.org/10.1016/S1473-3099(16)30319-X.

36. Overdevest I, Haverkate M, Veenemans J, Hendriks Y, Verhulst C, Mulders A, Couprie W, Bootsma M, Johnson J, Kluytmans J. 2016. Prolonged colonisation with *Escherichia coli* O25:ST131 versus other extended-spectrum beta-lactamase-producing *E. coli* in a long-term care facility with high endemic level of rectal colonisation, the Netherlands, 2013 to 2014. Euro Surveill 21:30376. https://doi.org/10.2807/1560-7917.ES.2016.21.42.30376.

37. Reigstad CS, Hultgren SJ, Gordon JI. 2007. Functional genomic studies of uropathogenic *Escherichia coli* and host urothelial cells when intracellular bacterial communities are assembled. J Biol Chem 282:21259–21267. https://doi.org/10.1074/jbc.M611502200.

38. Mike LA, Smith SN, Sumner CA, Eaton KA, Mobley H. 2016. Siderophore vaccine conjugates protect against uropathogenic *Escherichia coli* urinary tract infection. Proc Natl Acad Sci U S A 113:13468–13473. https://doi.org/10.1073/pnas.1606324113.

39. Horiyama T, Nishino K. 2014. AcrB, AcrD, and MdtABC multidrug efflux systems are involved in enterobactin export in *Escherichia coli*. PLoS One 9:e108642. https://doi.org/10.1371/journal.pone.0108642.

40. Wright KJ, Seed PC, Hultgren SJ. 2007. Development of intracellular bacterial communities of uropathogenic *Escherichia coli* depends on type 1 pili. Cell Microbiol 9:2230–2241. https://doi.org/10.1111/j.1462-5822.2007.00952.x.

41. Alqasim A, Scheutz F, Zong Z, McNally A. 2014. Comparative genome analysis identifies few traits unique to the *Escherichia coli* ST131 H30Rx clade and extensive mosaicism at the capsule locus. BMC Genomics 15:830. https://doi.org/10.1186/1471-2164-15-830.

42. Anderson GG, Palermo JJ, Schilling JD, Roth R, Heuser J, Hultgren SJ. 2003. Intracellular bacterial biofilm-like pods in urinary tract infections. Science 301:105–107. https://doi.org/10.1126/science.1084550.

43. Shepherd M, Achard MES, Idris A, Totsika M, Phan M-D, Peters KM, Sarkar S, Ribeiro CA, Holyoake LV, Ladakis D, Ulett GC, Sweet MJ, Poole RK, McEwan AG, Schembri MA. 2016. The cytochrome bd-I respiratory oxidase augments survival of multidrug-resistant *Escherichia coli* during infection. Sci Rep 6:35285. https://doi.org/10.1038/srep35285.

44. Subashchandrabose S, Smith SN, Spurbeck RR, Kole MM, Mobley H. 2013. Genome-wide detection of fitness genes in uropathogenic *Escherichia coli* during systemic infection. PLoS Pathog 9:e1003788. https://doi.org/10.1371/journal.ppat.1003788.

45. Subashchandrabose S, Hazen TH, Brumbaugh AR, Himpsl SD, Smith SN, Ernst RD, Rasko DA, Mobley H. 2014. Host-specific induction of *Escherichia coli* fitness genes during human urinary tract infection. Proc Natl Acad Sci U S A 111:18327–18332. https://doi.org/10.1073/pnas.1415959112.

46. Gupta S, Ferguson N, Anderson R. 1998. Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents. Science 280:912–915. https://doi.org/10.1126/science.280.5365.912.

47. Salipante SJ, Roach DJ, Kitzman JO, Snyder MW, Stackhouse B, Butler-Wu SM, Lee C, Cookson BT, Shendure J. 2015. Large-scale genomic sequenc-

ing of extraintestinal pathogenic *Escherichia coli* strains. Genome Res 25:119–128. https://doi.org/10.1101/gr.180190.114.

48. Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829. https://doi.org/10.1101/gr.074492.107.

49. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.

50. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693. https://doi.org/10.1093/bioinformatics/btv421.

51. Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21:456. https://doi.org/10.1093/bioinformatics/bti191.

52. Corander J, Marttinen P, Sirén J. 2008. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics 16:539. https://doi.org/10.1186/1471-2105-9-539.

53. Gutmann MU, Corander J. 2016. Bayesian optimization for likelihood-free inference of simulator-based statistical models. J Mach Learn Res 17:1–47.

54. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. Mol Biol Evol 34:2115–2122. https://doi.org/10.1093/molbev/msx148.

55. Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

56. Edgar R. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340.

57. Kumar S, Stecher G, Peterson D, Tamura K. 2012. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. Bioinformatics 28:2685–2686. https://doi.org/10.1093/bioinformatics/bts507.