BMC Genomics

RESEARCH

Open Access

Lineage-specific variation in frequency and hotspots of recombination in invasive *Escherichia coli*

Kathryn R. Piper[1], Stephanie S. R. Souza[1], Odion O. Ikhimiukor[1], Adrienne A. Workman[2], Isabella W. Martin[2*] and Cheryl P. Andam[1*]

Abstract

Background The opportunistic bacterium *Escherichia coli* can invade normally sterile sites in the human body, potentially leading to life-threatening organ dysfunction and even death. However, our understanding of the evolutionary processes that shape its genetic diversity in this sterile environment remains limited. Here, we aim to quantify the frequency and characteristics of homologous recombination in *E. coli* from bloodstream infections.

Results Analysis of 557 short-read genome sequences revealed that the propensity to exchange DNA by homologous recombination varies within a distinct population (bloodstream) at narrow geographic (Dartmouth Hitchcock Medical Center, New Hampshire, USA) and temporal (years 2016 – 2022) scope. We identified the four largest monophyletic sequence clusters in the core genome phylogeny that are represented by prominent sequence types (ST): BAPS1 (mainly ST95), BAPS4 (mainly ST73), BAPS10 (mainly ST131), BAPS14 (mainly ST58). We show that the four dominant clusters vary in different characteristics of recombination: number of single nucleotide polymorphisms due to recombination, number of recombination blocks, cumulative bases in recombination blocks, ratio of probabilities that a given site was altered through recombination and mutation (r/m), and ratio of rates at which recombination and mutation occurred (ρ/θ). Each sequence cluster contains a unique set of antimicrobial resistance (AMR) and virulence genes that have experienced recombination. Common among the four sequence clusters were the recombined virulence genes with functions associated with the Curli secretion channel (*csgG*) and ferric enterobactin transport (*entEF*, *fepEG*). We did not identify any one recombined AMR gene that was present in all four sequence clusters. However, AMR genes *mdtABC*, *baeSR*, *emrKY* and *tolC* had experienced recombination in sequence clusters BAPS4, BAPS10, and BAPS14. These differences lie in part on the contributions of vertically inherited ancestral recombination and contemporary branch-specific recombination, with some genomes having relatively higher proportions of recombined DNA.

Conclusions Our results highlight the variation in the propensity to exchange DNA via homologous recombination within a distinct population at narrow geographic and temporal ranges. Understanding the sources of the genetic variation in invasive *E. coli* will help inform the implementation of effective strategies to reduce the burden of disease and AMR.

Keywords *Escherichia coli*, Bloodstream infection, Genome, Homologous recombination, Sequence types, Antimicrobial resistance

*Correspondence:
Isabella W. Martin
isabella.w.martin@hitchcock.org
Cheryl P. Andam
candam@albany.edu

Background

The commensal and opportunistic bacterium *Escherichia coli* is an ecologically versatile species. It inhabits a wide range of eukaryotic hosts as well as non-host environments such as water, soil and plants [1, 2]. *E. coli* is an important member of the normal intestinal microbiota of humans and other vertebrates [3, 4]. However, it also causes a plethora of diseases in humans, such as diarrhea, urinary tract infections and life-threatening diseases such as septicemia, meningitis and renal failure [5–7]. *E. coli* can invade normally sterile sites of the human body, such as the bloodstream, when defense barriers are breached, potentially leading to life-threatening organ dysfunction and even death [8–12]. *E. coli* transmission and disease outbreaks have been reported worldwide, often implicating contaminated food products and animal reservoirs [13–15]. The burden of antimicrobial resistant *E. coli* on the global human population is profound. The World Health Organization reported that *E. coli* is one of the six major pathogens causing deaths associated with or attributable to antimicrobial resistance (AMR) worldwide, with 219,000 deaths in 2019 due to *E. coli* resistant to one or more drugs [16]. The high mortality rates caused by cephalosporin-resistant *E. coli* and fluoroquinolone-resistant *E. coli* are particularly worrisome [16].

The tremendous diversity in the ecological niches of *E. coli* is reflected in its genetic diversity [17]. *E. coli* can be classified according to sequence types (ST; based on seven single-copy housekeeping genes [18]), serotypes (based on O/somatic, H/flagellar, and K/capsular surface antigens [19, 20]), pathotypes (based on clinical presentation and clinical site of infection [5, 17]), and phylogroups (*e.g.*, A, B1, B2, D, E, F [21, 22]). Diversification of the genome content of *E. coli* has been shaped by the recurrent ST-specific gene gain and loss [23], the co-occurrence of genes by function and mobility [24], and the specific ecological niches and local geography that the bacterium occupies [25]. A recent pan-genome analysis of 1,324 complete genomes of *E. coli* revealed that different ST and phylogroups harbor distinct accessory genes [26], which may underlie their unique pathogenic and epidemiological characteristics.

Homologous recombination, which refers to the non-reciprocal unidirectional transfer of a highly similar segment of DNA, is an important process in shaping *E. coli* evolution and diversity [27–29]. Acquisition of homologous tracts of exogenous DNA results to allelic substitution within conserved genomic regions and is thus more common in closely related organisms [30]. In *E. coli*, it is mediated by DNA repair systems that play a role in maintaining genomic integrity and in facilitating genetic exchanges between host chromosomes and foreign DNA during horizontal gene transfer [31]. The RecBCD, RecFOR, and RecBFI pathways are major players in enabling efficient DNA repair and homologous recombination in *E. coli* [32, 33]. Recombination is pivotal in the global success and pathogenicity of the *E. coli* sequence types ST131 [34], ST410 [35, 36] and ST1193 [37], which are well known to be multidrug resistant and cause severe diseases [38, 39]. The global dissemination of carbapenemase-producing *E. coli* was associated with the recombination of the mutated form of *ftsI* encoding penicillin-binding protein 3 (PBP3) and of the porin *ompC* gene among multiple lineages [40]. Extraintestinal pathogenic *E. coli* (ExPEC) has experienced highly variable frequencies of recombination but considerably higher compared to their commensal counterparts, with recombination linked to a higher number of virulence genes [41]. Although previous data have been informative, it is unclear whether such variation in recombination frequencies exists within a population occupying a specific niche (*e.g.*, bloodstream) and if it does, how it impacts the genetic diversity of the population. Hence, a more detailed quantification of the contributions of within-species recombination in pathogenic *E. coli* will be invaluable in understanding the behavior of distinct lineages in specific environments. This in turn should help us comprehend the genetic basis that governs the diversification, host adaptation, and pathogenicity of *E. coli* in causing invasive diseases.

Here, we aim to quantify the frequency and characteristics of homologous recombination to the population genetic diversity of *E. coli* derived from bloodstream infections from a medical center in New Hampshire, USA. Overall, our findings shed light on the importance of distinguishing the impacts of homologous recombination in shaping the diversification and pathogenicity features of individual *E. coli* lineages.

Results

Four lineages are dominant in the bloodstream *E. coli* population

We collected and sequenced short-read draft genomes of 557 *E. coli* isolates sampled from bloodstream infections in unique pediatric and adult patients at Dartmouth Hitchcock Medical Center (DHMC), New Hampshire, USA from November 2016 – May 2022 (Supplementary Table S1). The de novo assemblies resulted in genome sizes ranging from 4.59 – 5.54 Mb with an estimated 21,310 genes present in the entire dataset, or what is referred to as the pan-genome [42] (Supplementary Table S2). On average, each genome carried 4,694 genes. The core and soft-core genes together comprised 16.05% of the pan-genome. Across the entire dataset, the number of accessory genes per genome ranged from 4,221 to 5,247 and varied within and among the four major

clusters (Supplementary Table S1). The number of singleton genes in the pangenome was 4,431 genes, comprising 20.79% of the pan-genome, and ranged from 0 to 106 per genome (mean = 7.95, median = 1). Singleton genes are those genes that are present in only one genome and not in others. The number of singleton genes per genome varied within and among the four major clusters.

We retrieved a total of 480,947 single nucleotide polymorphisms (SNP) from the alignment of the combined core and soft-core genes (n = 3,422). The core genome SNP alignment was then used to construct a maximum likelihood phylogenetic tree (Fig. 1). Population structure analysis using Bayesian hierarchical clustering [43] identified 22 sequence clusters (Supplementary Table S1). The four largest sequence clusters were BAPS1 (n = 101 genomes), BAPS4 (n = 56 genomes), BAPS10 (n = 106 genomes), and BAPS14 (n = 52 genomes), which altogether represented 56.55% of the entire dataset. The remaining BAPS clusters contained 38 or fewer genomes. Multi-locus sequence typing (MLST) using seven single-copy housekeeping genes in *E. coli* [18, 44] identified 133 STs. The most prominent STs were ST95 (n = 96 representing 17.24% of the dataset), ST131(n = 75; 13.46%), ST73 (n = 49; 8.79%), ST69 (n = 34; 6.1%), and ST127 (n = 24; 4.31%). A total of 78 STs were represented by a single genome (Supplementary Table S1).

The largest sequence cluster BAPS10 (n = 106 genomes) consisted of 14 STs and 11 serotypes, of which ST131 (n = 75 genomes) and serotype O25:H4 (n = 66 genomes) were most frequently detected. The second largest BAPS cluster BAPS1 (n = 101) comprised five STs and six serotypes, of which ST95 (n = 96 genomes) and serotype O1:H7 (n = 53 genomes) were the most common. BAPS4 (n = 56 genomes) consisted of four STs and eight serotypes, of which ST73 (n = 49) and serotype O6:H1 (n = 25 genomes) were most frequently detected. BAPS14 (n = 54 genomes) contained 30 STs, of which ST58 (n = 10) was the most common and 41 serotypes were present.

Recombination frequencies vary between *E. coli* lineages

We calculated the recombination frequencies by examining regions of elevated SNP density across the core genome alignment using the program Gubbins [45]. These regions are inferred to have been generated by recombination events. Accurate estimation of recombination is strengthened by analysis of a large number of genomes being compared; hence, we selected four sequence clusters that each have ≥ 50 genomes (BAPS1,

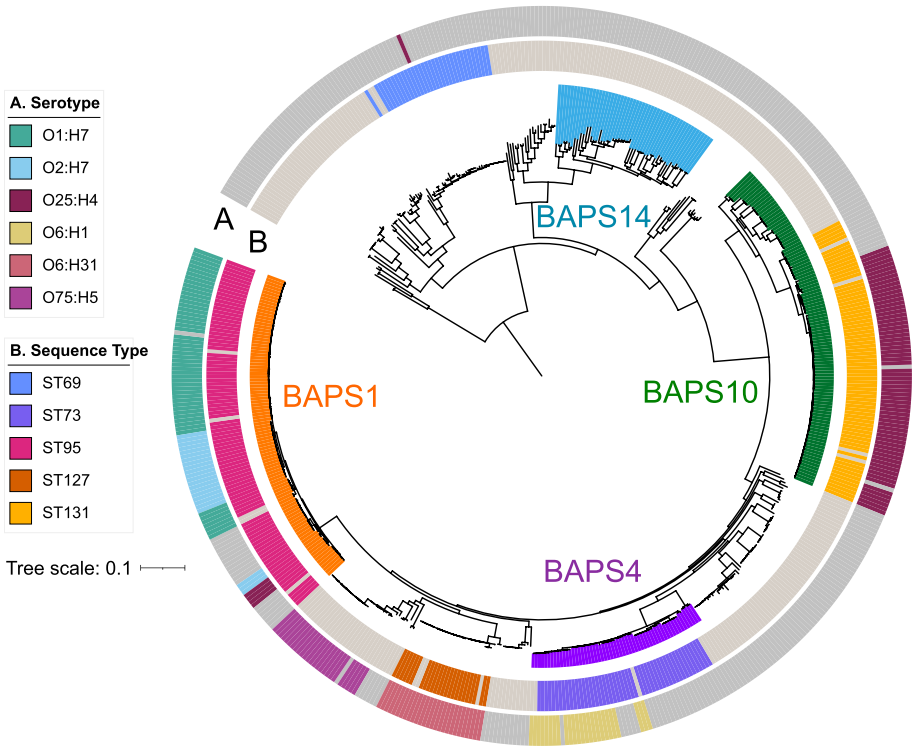

Fig. 1 Maximum likelihood phylogenetic tree of the 557 bloodstream *E. coli* isolates. The midpoint-rooted phylogeny is built from the sequence alignment of 3,422 core and soft-core genes. Tree scale represents the number of nucleotide substitutions per site. Outer rings show the most common serotypes and sequence types (ST). Colored lines extending from the branches show the sequence clusters inferred by BAPS. For visual clarity, only the four largest clusters (with at least 50 genomes each) are colored

BAPS4, BAPS10, BAPS14; Fig. 1). Based on the SNP densities, we calculated seven metrics to describe recombination among the four sequence clusters: number of SNPs inside inferred recombination events, number of SNPs outside of inferred recombination events, number of recombination blocks, number of bases detected in recombination events, number of bases detected in ancestral recombination events, ratio of recombination to mutation, and ratio of recombination to mutation of a branch (Supplementary Table S3).

We calculated the number of base substitutions reconstructed onto a branch on the phylogeny that fall within and outside a predicted recombination (Fig. 2A). We found significant difference in the four sequence clusters, with BAPS1, BAPS4, and BAPS10, having more SNPs outside of inferred recombination events and BAPS14 having more SNPs detected inside inferred recombination (all comparisons with $p < 0.05$, Wilcoxon test). The total number of recombination blocks reconstructed onto a branch varied among sequence clusters BAPS 1 and 4, 1 and 14, 4 and 14, and 10 and 14 (pairwise comparisons with $p < 0.001$, Games-Howell test) (Fig. 2B). BAPS14 carried the highest average number of recombination blocks per genome (92.04) compared to the other three sequence clusters.

In terms of the total number of nucleotide bases in predicted recombined regions, we found significant differences between sequence cluster BAPS14 and the other three clusters (pairwise comparisons with $p < 0.001$, Games-Howell test) (Fig. 2C). This metric refers to the total length of all recombination events reconstructed onto a branch. We also calculated the cumulative bases in predicted recombinations, which refers to the total number of nucleotide bases in the alignment affected by recombination on a branch and its ancestors (Fig. 2D). We found significant differences among all pairs of sequence clusters (pairwise comparisons with $p < 0.001$, Games-Howell test).

The ratio of the probabilities that a given site was altered through recombination and mutation (r/m) gives a measure of the relative impact of the two processes on the variation accumulated on the branch (Fig. 2E). Sequence cluster BAPS14 exhibited the highest r/m values (5.84), indicating that recombination has contributed on average 5X more than mutation to the overall genetic diversity of BAPS14 genomes. The r/m of BAPS14 is significantly different from those in sequence clusters BAPS1, 4, and 10 (pairwise comparisons with $p < 0.001$, Games-Howell test).

The rho/theta ratio refers to the number of recombination events to point mutations on a branch and provides a measure of how often recombination events happen relative to mutations (*i.e.*, relative rates) (Fig. 2F). We found significant differences in rho/theta values between sequence clusters BAPS1 and 10, 4 and 10, and 1 and 14. The rho/theta ratios of BAPS1, 4, 10, and 14 were 0.02, 0.03, 0.08, and 0.08, respectively. This means that, on average, recombination occurred less frequently than point mutations during strain evolution in all four sequence clusters.

Recombination in AMR and virulence genes

We sought to identify the genes that have experienced recombination in genomes of the four sequence clusters. For each of the four sequence clusters, we scanned the aligned genome sequences for regions of elevated SNP density that indicate the occurrence of recombination (Fig. 3). In total, we identified 4,146 genes across the four sequence clusters that experienced at least one instance of recombination. This was equivalent to 89.37% of the 4,639 genes annotated in the reference genome strain K-12 substrain MG-1655 (Accession NC_000913.3). Recombined regions were highly variable among the four lineages. Genomes from sequence cluster BAPS14 contained the highest number of recombined genes ($n = 3,883$), while BAPS10, BAPS4 and BAPS1 contained 2,432, 1,072, and 507 recombined genes, respectively (Supplementary Table S4). A total of 103 recombined genes were common across the four sequence clusters and these included genes with functions related to DNA and RNA binding, iron transport, starvation response,

(See figure on next page.)

Fig. 2 Comparison of the characteristics and frequencies of recombination among the four sequence clusters. **A** Number of base substitutions reconstructed onto the branch that fall within (In) and outside (Out) a predicted recombination for each sequence cluster. **B** Total number of recombination blocks reconstructed onto the branch. **C** Bases in recombinations refers to the total length of all recombination events reconstructed onto a branch. **D** Cumulative bases in recombinations refers to the total number of bases in the alignment affected by recombination on a branch and its ancestors. **E** The value r/m is the ratio of probabilities that a given site was altered through recombination and mutation. It gives a measure of the relative impact of recombination (r) and mutation (m) on the variation accumulated on the branch. **F** The rho/theta ratio refers to the number of recombination events to point mutations on a branch (*i.e.*, relative rates). It is a measure of how often recombination events happen relative to mutations. For visual clarity, only comparisons that are statistically significantly different are indicated with asterisk. Wilcoxon test was used for all comparisons in panel **A** and Games-Howell test in panels **B**-**F**. For all panels, colored circles represent genomes and the colors correspond to colors of the sequence clusters in Fig. 1. Red dot represents the mean. Details of the recombination metrics are presented in Supplementary Table S2

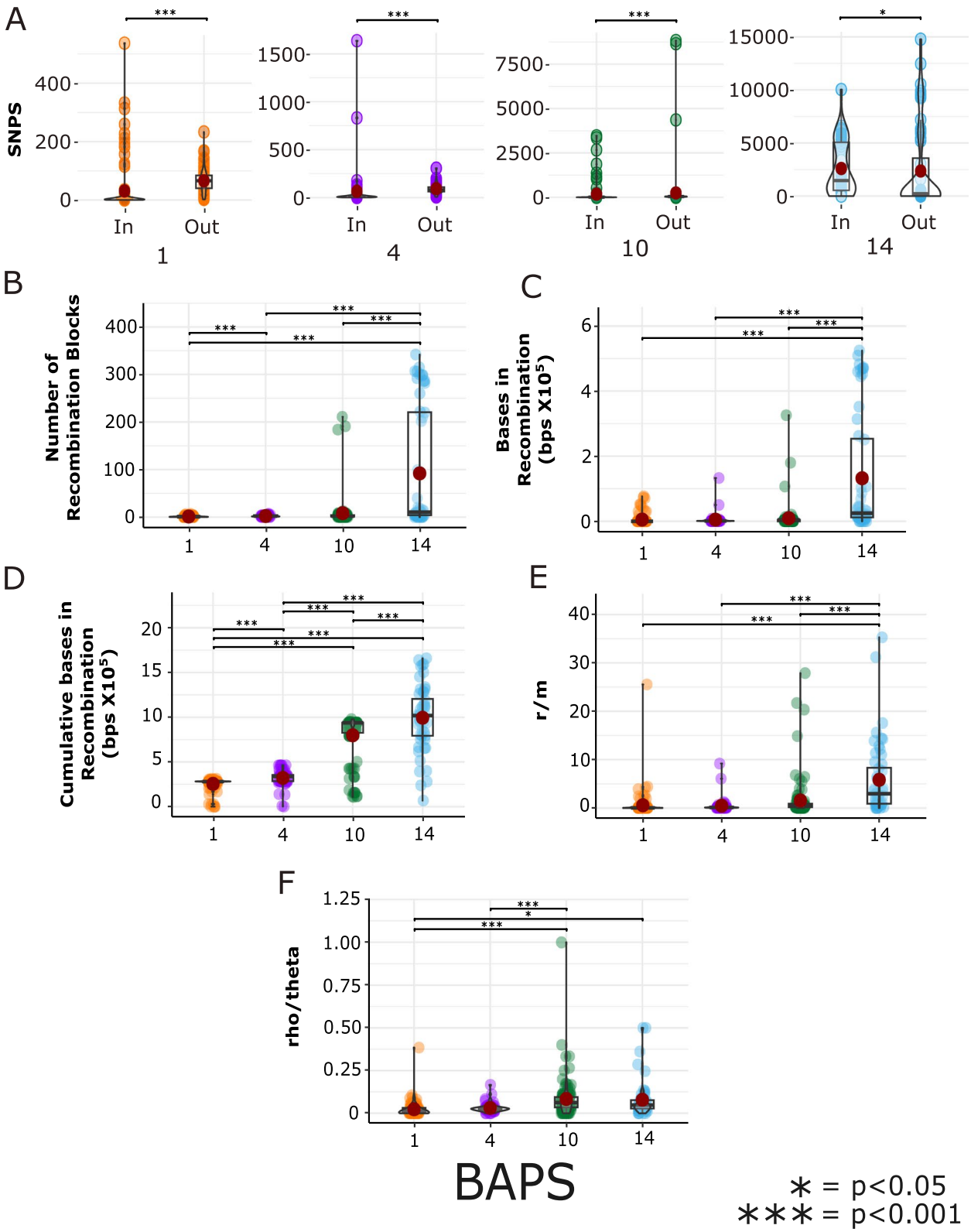

Fig. 2 (See legend on previous page.)

cell wall organization, and electron transport (Supplementary Table S5).

We next sought to identify genes associated with AMR and virulence that have experienced recombination. We detected recombination in 41, 9, 26, and 31 virulence genes in sequence clusters BAPS 1, 4, 10, and 14, respectively (Fig. 3A-E and Supplementary Table S6). Common among the four sequence clusters were the recombined virulence genes with functions associated with the Curli secretion channel (*csgG*) and ferric enterobactin transport (*entEF*, *fepEG*). *E. coli* and other enteric bacteria produce proteinaceous extracellular fibers called curli that are involved in adhesion to surfaces, cell aggregation, and biofilm formation [46]. Hence, they play a major role in mediating adhesion and invasion of host cells, and they are potent inducers of the host inflammatory response [46]. The siderophore enterobactin is a small iron chelator that facilitate cellular transport of iron, which is an essential nutrient for bacterial growth [47]. Excess of freely available iron is associated with enhance virulence in bacterial pathogens [48]. Flagellar genes were also frequently recombined, but the specific *flg* genes that showed evidence of recombination varied among the four clusters: *flgA-O* in BAPS1, *flgA* in BAPS4, *flgCDEHIJKLNO* in BAPS10, *flgAFGHIJKLMNO* in BAPS14. Flagella contribute to the infection of *E. coli* by mediating motility, surface adhesion, and invasion [49, 50]. Taken as a whole, each sequence cluster contained a unique set of virulence genes that had experienced frequent recombination.

We detected evidence of recombination in 1, 13, 17, and 25 AMR genes in sequence clusters BAPS 1, 4, 10, and 14, respectively (Fig. 3A-E and Supplementary Table S6). We did not identify any one recombined AMR gene that was common in all four sequence clusters. Nonetheless, we found AMR genes that were shared by at least three sequence clusters. The genes *mdtABC* and their regulator encoded by *baeSR* had experienced recombination in sequence clusters BAPS4, BAPS10, and BAPS14. MdtABC is a multidrug efflux system of the resistance-nodulation-cell division (RND) family and confers resistance to novobiocin and deoxycholate [51, 52]. The gene *mdtA* encodes a membrane fusion protein, while *mdtBC* encodes a transmembrane drug transporter [51]. BaeR is a response regulator, while BaeS is a sensor kinase [51, 52]. We also detected recombination in the genes *emrKY*, which encodes a tripartite multidrug efflux pump of the major facilitator superfamily (MFS) [53]. In *Shigella*, EmrKY functions in bacterial survival within macrophages [54]. We also detected recombination in the gene *tolC* encoding an outer membrane channel, which translocates an extremely broad spectrum of antimicrobial agents and other particles such as detergents, dyes, and organic solvents [55]. In addition, we detected recombination in *acrD* in sequence clusters BAPS10 and BAPS14. AcrD, along with the gene products of *acrB* and *acrF*, interact with the outer membrane channel TolC in multi-component efflux pumps [56, 57]. TolC partners with the transporters MdtABC and EmrKY [55]. TolC is also implicated in other functions beyond AMR, such as cell division, expression of outer membrane porins, acid sensitivity, virulence, aggregation, and cell envelope mechanics and integrity, mainly due to its extensive functional interactions with various genes [55, 58, 59]. Similar to the recombined virulence genes, each sequence cluster also contained a unique set of recombined AMR genes.

We next examined the distribution of synonymous and non-synonymous mutations on genes associated with AMR and virulence within each of the four sequence clusters (Supplementary Table S6). For each sequence cluster, we note that there is a far higher proportion of synonymous SNPs than non-synonymous SNPs in recombined AMR genes: BAPS1 (total = 102, 0, 1, 0 SNPs for synonymous, frameshift, missense, nonsense mutations, respectively), BAPS4 (total = 100, 0, 46, 0), BAPS10 (total = 859, 24, 193, 5), BAPS14 (total = 2566, 0, 510, 17). Similar results were observed for the recombined virulence-associated genes: BAPS1 (total = 1589, 6, 594,

(See figure on next page.)

Fig. 3 Recombination hotspots among the four *E. coli* sequence clusters. Panels **A** – **D** show the recombination hotspots in sequence clusters BAPS1, BAPS4, BAPS10, and BAPS14, respectively. The tree was built from non-recombinant regions in the core genome alignment. The reference genome is shown as an orange line with predicted coding sequences in the forward and reverse frames shown in light blue above and below this line. The length of the reference genome is 4.64 Mb. For visual clarity, only antimicrobial resistance (AMR) and virulence genes that were inferred to have experienced recombination are labeled. Each row represents an isolate and the columns relate to bases in the reference genome. Below the reference genome coordinates is a matrix showing all inferred recombination in each genome. The red columns are recombinations shared by multiple isolates and occuring in the internal branches, while the blue columns are recombinations in the terminal branch and represented by individual isolates. SNP density at each site of the sequence alignment is shown below the matrix. **E** Venn diagram showing the number of recombined AMR and virulence genes that are shared between sequence clusters or unique to each cluster. The percentage values in parentheses indicate the number of recombined AMR and virulence genes shared by the clusters or unique to the individual cluster divided by the total number of genes inferred to have recombined across all 4 BAPS clusters. Zoom-in view of each matrix is presented in Supplementary Figures S1-S4. Details of recombined genes in each sequence cluster are presented in Supplementary Tables S4-S6

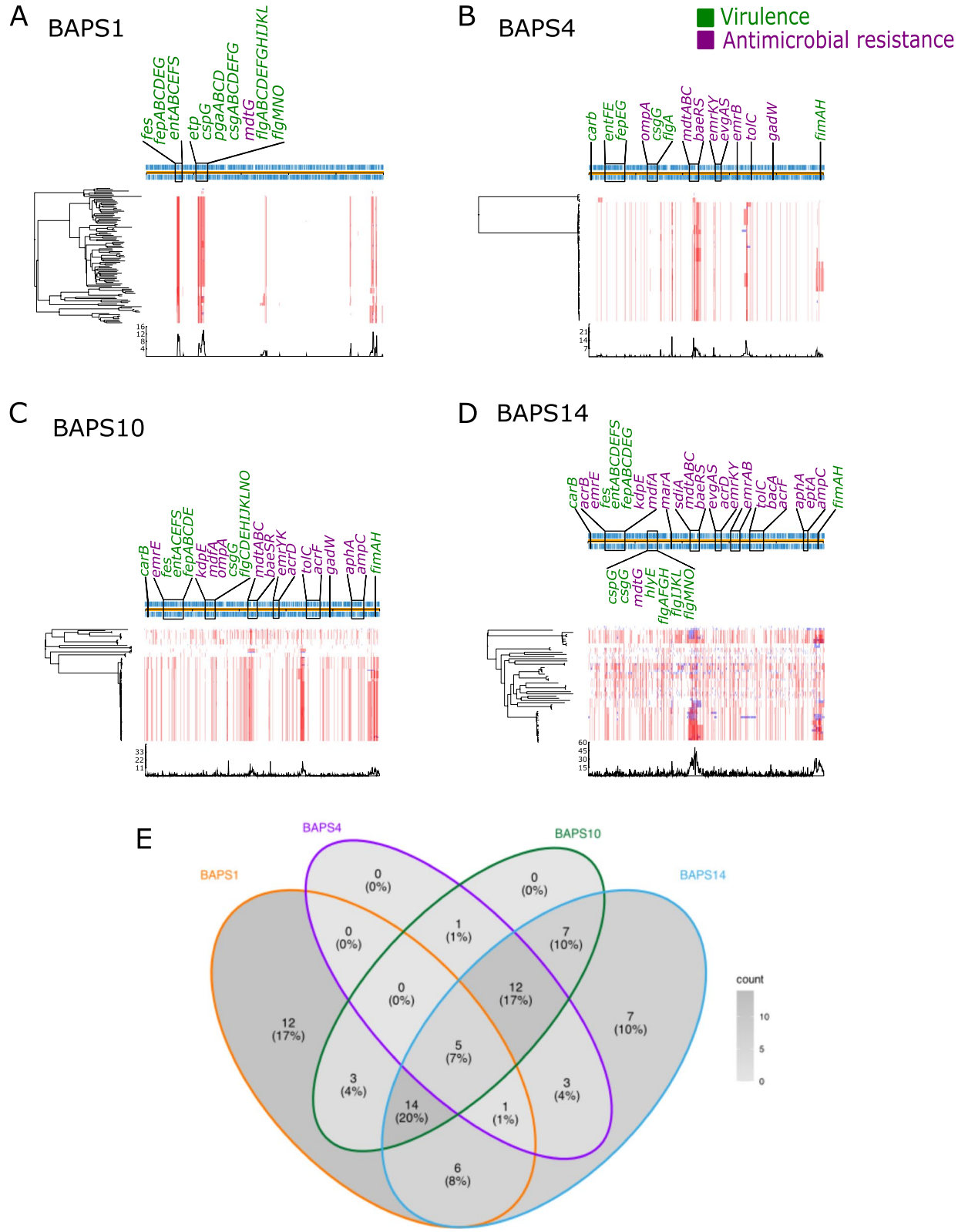

Fig. 3 (See legend on previous page.)

25), BAPS4 (total = 57, 0, 29, 1), BAPS10 (total = 1204, 42, 275, 8), BAPS14 (total = 1782, 0, 521, 20). These results demonstrate that the dominant selective force acting on the SNPs within the recombined genes associated with AMR and virulence has been purifying selection.

Contemporary and vertically inherited ancestral recombination

We sought to determine to what extent the genome of each *E. coli* isolate has been affected by recombination. We built recombination-free core genome phylogenies of each of the four sequence clusters (Fig. 4A-D). Two metrics were calculated for every genome: (a) Bases in recombinations, which refers to the total length of all recombination events reconstructed onto a branch; and (b) Cumulative bases in recombinations, which refers to the total number of bases in the alignment affected by recombination on a branch and its ancestors. The first value pertains only to contemporary recombination events and that are unique to a genome, while the second value also includes vertically inherited recombination events.

Results reveal that all or nearly all genomes from the four sequence clusters have inherited recombined DNA from ancestral lineages. However, we found that the contributions of vertically inherited ancestral recombination varied among the genomes of a sequence cluster as well as between clusters (Fig. 4A-D). Among members of sequence cluster BAPS1, the proportion of the genome with ancestral recombined DNA ranged from zero to 6.28%. In BAPS4 genomes, the proportion of the genome sequence with ancestral recombined DNA ranged from 0.04 to 10.28%. Among BAPS10 genomes, the proportion of the genome sequence with ancestral recombined DNA ranged from 2.45 to 22.07%. In BAPS14, the proportion of the genome sequence with ancestral recombined DNA ranged from 1.59 to 38.03%.

In terms of contemporary branch-specific recombination, we also found variation among members of each sequence cluster (Fig. 4A-D). In sequence cluster BAPS1, only 18 out of 101 isolates experienced recent recombination and the proportion of each genome acquired from recent recombination ranged from 0.0001 to 1.5%. In BAPS4, 51 out of 56 isolates experienced recent recombination and the proportion of each genome acquired from recent recombination ranged from 0.025 to 2.97%. In BAPS10, 93 out of 106 isolates experienced recent recombination and the proportion of each genome acquired from recent recombination ranged from 0.00027 to 7.37%. In BAPS14, 46 out of 52 isolates experienced recent recombination and the proportion of each genome acquired from recent recombination ranged from 0.1 to 12.09%. Notably, a few genomes within each cluster harbor relatively larger segments of recombined DNA acquired from contemporary recombination events. This is certainly evident in sequence clusters BAPS10 and BAPS14.

Discussion

The emergence of new lineages of *E. coli* with unique genetic and phenotypic features is a constant public health threat. Its propensity to acquire, exchange, and maintain alleles and genes through homologous recombination underlies the ever-growing challenge of controlling multidrug resistance, virulence, host colonization and adaptation, transmission, and other clinically relevant genetic traits [34, 40, 60]. In our study, we sought to quantify the frequency, characteristics, and impacts of homologous recombination on the population genetic diversity of bloodstream *E. coli*. Our primary findings were that (a) major phylogenetic lineages exhibit variable frequencies in genome-wide homologous recombination, (b) the suite of recombined AMR and virulence genes differ among lineages, and (c) these differences lie in part on vertically inherited ancestral recombination and contemporary branch-specific recombination. These results demonstrate that the impact of recombination is highly variable even among clonally related individuals, contributing to the remarkable genetic diversity and pathogenic potential that exists throughout the entire species. Our findings are consistent with previous studies that show within-species heterogeneity in recombination in other bacterial pathogens such as *Klebsiella pneumoniae* [61], *Neisseria meningitidis* [62], *Staphylococcus aureus* [63], and *Streptococcus pneumoniae* [64].

A notable feature of *E. coli* is its infinitely open accessory genome [65, 66] that have been subjected to frequent acquisition of exogenous DNA, creating genetically hybrid strains with novel features that blur the lines of existing classification designation [67, 68]. New combinations of virulence determinants and allelic variants that are horizontally acquired may manifest in altered pathogenic and drug resistant characteristics of the recipient strain, as in the case of sequence types ST131 [34], ST141 [69], ST410 [35, 36], and ST1193 [37]. The global success of these four lineages as a cause of many disease outbreaks may be attributed partly to their mosaic genomes. Our results revealed that each of the four major lineages of bloodstream *E. coli* differed in the composition of recombining AMR and virulence genes, suggesting unique lineage-specific patterns of convergence and evolutionary trajectories. Mosaic genomes with distinct combinations of genetic features can therefore arise from any clonal background. The flexibility of a genome to incorporate exogenous DNA via recombination is important in understanding lineage-specific

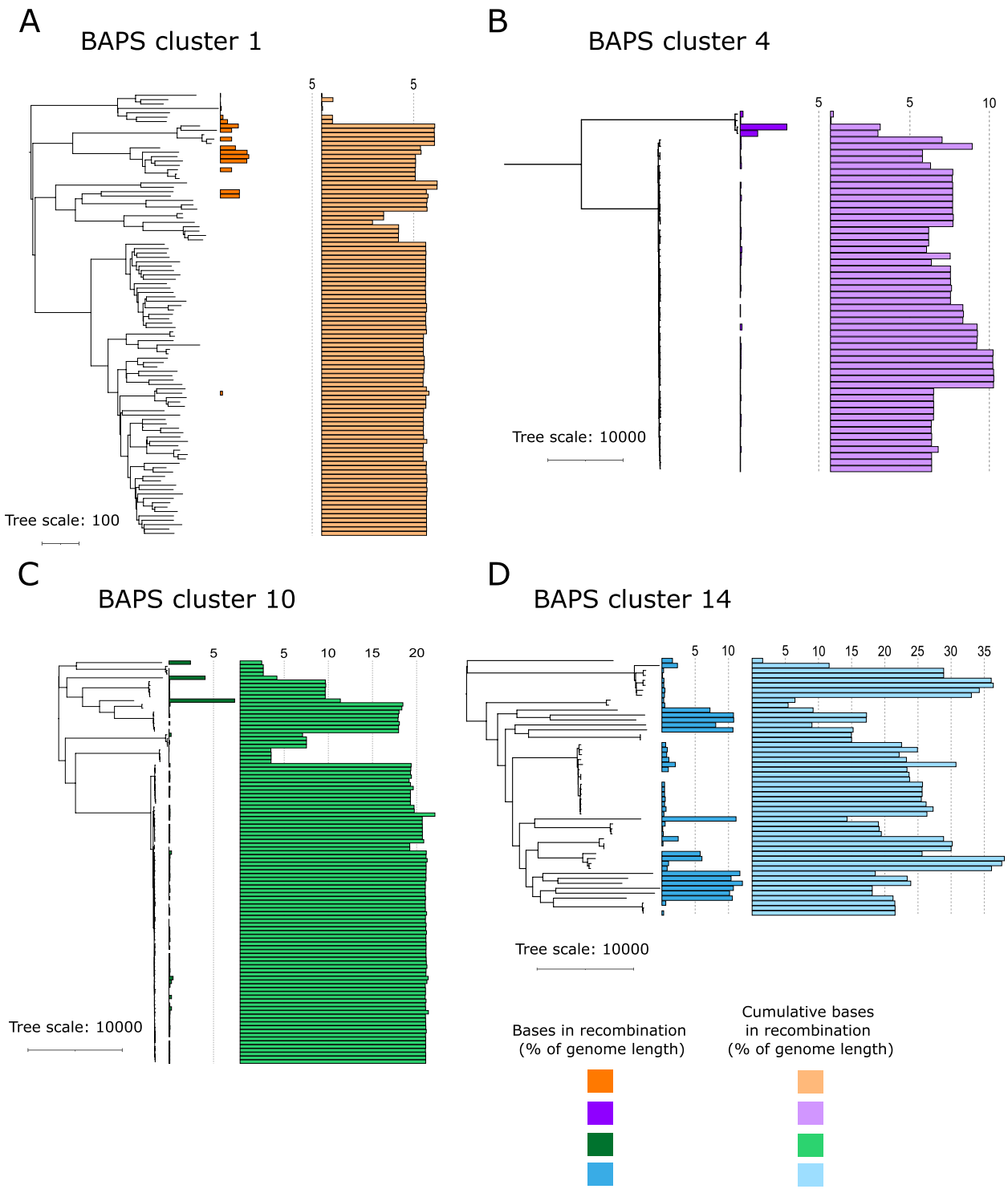

Fig. 4 Comparison of the total length of recombination events among members of each sequence cluster. Maximum likelihood phylogenetic trees of each sequence cluster were built from the core genome alignment and midpoint-rooted. Next to each phylogenetic tree are two sets of bar plots showing the bases in recombinations referring to the total length of all recombination events reconstructed onto a branch (dark-colored bar plots on the left), while cumulative bases in recombinations refers to the total number of bases in the alignment affected by recombination on a branch and its ancestors (light-colored bar plots on the right). For each genome, the two values were calculated as a proportion of the genome length

evolutionary trajectories and adaptive consequences. The frequent shuffling of AMR and virulence genes through recombination amplifies within-species diversity. This has important consequences for our ability to both foresee the outcomes of public health intervention programs and predict which lineages of invasive *E. coli* are likely to become a high-risk concern (*i.e.*, more virulent, more transmissible, multidrug resistant). Continuous long-term surveillance of the different lineages of bloodstream *E. coli* is therefore critical.

We observed lineage-specific patterns of homologous recombination, characterized by numerous contemporary branch-specific recombination events and relatively fewer vertically inherited ancestral events shared by multiple genomes. We also observed higher proportion of synonymous than non-synonymous SNPs in recombined AMR and virulence-associated genes in all four major lineages, similar to reports in the recombination regions of the marine pathogen *Vibrio anguillarum* [70]. While we were unable to trace the origins of every recombination event in our data, the contributions of both processes vary even among very closely related isolates that are members of the same lineage. Similar results have been reported in an ecologically diverse but geographically restricted population of the pathogen *Burkholderia pseudomallei* that inhabits tropical soils [71]. Our results provide evidence of ongoing recombination in invasive *E. coli* that is shaped mainly by the genetic background (lineage). Differences in recombination patterns and in the pool of recombining genes among lineages may suggest differences in DNA repair systems and defense systems (*e.g.*, restriction modification, CRISPR-Cas) against exogenous DNA entering the cell [72], although the presence of cryptic microscale niches may also be a possibility [73].

Differences in recombination among members of a species have important implications in understanding the evolution of evolvability, *i.e.*, how biological systems generate heritable adaptive variation [74, 75]. Environmental variability can modify rates of recombination (as well as mutation and migration) by increasing the genetic diversity on which selection can act on [76]. This occurs because recombination combines beneficial alleles or eliminates deleterious mutations [77]. It also reinforces advantageous phenotypes in the population [77]. Hence, elucidating the frequency and characteristics of recombination is critical when considering the adaptive potential of a species to environmental variability. Variation in the rates of genetic exchange within a species suggests that lineages respond to selective pressures in different ways and has implications in how rapidly they can adapt to new environments [64], including the shift from the gut to the blood and exposure to clinical interventions such as vaccines and antibiotic therapy. Lineage-specific genetic components that can alter the frequencies of recombination include saturation and variable efficiencies of the mismatch repair system [78] and bacterial immune systems against foreign DNA [79]. Future work should focus on within-species distribution and diversity of these repair and defense mechanisms against exogenous DNA.

Our study is not without limitations. First, we do not have clinical data about the source patients of these isolates, and as such, we are not able to examine if the differences in recombination frequencies have an impact on patient outcomes and treatment options. We also lacked data on the geographical location and social history of patients, as proximity may promote more opportunities for genetic exchange between bacteria. Second, we focused our estimation of recombination only on the four largest phylogenetic clusters. It is possible that less prevalent lineages have significantly different recombination frequencies and characteristics, but this information could not be reliably assessed with the current density of sampling. Future work investigating a broader range of *E. coli* lineages, including among the different STs, serotypes, pathotypes, and phylogroups, will be particularly informative. Third, we did not consider recombination from outside the population or from other species, even though *E. coli* can gain DNA from other members of *Enterobacterales* [80]. The microbial community that *E. coli* inhabits will likely influence the composition of potential recombination donors. The DNA sequences of *E. coli* strains will therefore carry a history of recombination events and recombination partners, and this information will be retained as *E. coli* invades the bloodstream. Lastly, recombination can occur between nearly identical DNA sequences and the same sequence may experience recombination multiple times throughout its evolution, and thus will remain invisible to current recombination detection methods. Missing strains due to incomplete sampling will also influence inference of recombination frequencies. These can lead to lower estimates of predicted recombination rates than the actual rates that occur in nature. Reconstruction and simulations of ancestral recombination events on each branch of a phylogeny may partly overcome such limitations. Nonetheless, our results provide important baseline estimates of lineage-specific recombination that can be extended to other phylogenetic groups, diseases, and environments of *E. coli*.

Conclusions

Our results highlight the variation in the propensity to exchange DNA via homologous recombination within a distinct population (bloodstream) at narrow geographic (DHMC, New Hampshire) and temporal (years 2016

– 2022) ranges. Understanding the sources of the genetic variation in invasive *E. coli* will help inform the implementation of effective strategies to reduce the burden of disease and AMR.

Materials and methods

Bacterial collection

We collected a total of 565 *E. coli* isolates from bloodstream infections in unique pediatric and adult patients. The archived isolates were grown from clinical blood culture specimens submitted to the Department of Pathology and Laboratory Medicine at DHMC, New Hampshire, USA from November 2016 to May 2022. The first significant blood culture isolate is routinely obtained from each patient and archived (freezer space permitting) in the event of future need for patient care, and epidemiologic, public health or laboratory quality studies. Ethical approval was granted by the Committee for the Protection of Human Subjects of DHMC and Dartmouth College. The study protocol was deemed not to be human subjects research. Samples used in the study were subcultured bacterial isolates that had been archived in the routine course of clinical laboratory operations. No patient specimens were used and patient protected health information was not collected. Therefore, informed consent was not required. Isolates were sub-cultured and assigned a study number with all patient identifiers removed. Only the date of collection was linked to the study number. All isolates were stored in DMSO solution at −80 °C degrees.

Genomic DNA extraction and whole genome sequencing

Isolates were sub-cultured from dimethyl sulfoxide (DMSO) stocks in brain heart infusion broth (BD Difco, Franklin Lakes, New Jersey) and incubated at 37 °C for 24 h. DNA was extracted and purified using the QuickDNA Fungal/Bacterial Miniprep Kit (Zymo Research, Irvine, California) following manufacturer's protocol. We used the Qubit fluorometer (Invitrogen, Grand Island, New York) to measure DNA concentrations. DNA libraries of each sample were prepared using the Illumina DNA prep tagmentation kit and unique dual indexes in accordance with manufacturer's protocols. Sequencing was carried out as multiplexed libraries on the Illumina NextSeq2000 platform using a 300-cycle flow cell kit to produce 2 × 150 bp paired reads. To support optimal base calling, 1 – 2% PhiX Control was spiked into the run. Sequencing was performed at the SeqCoast Genomics (Portsmouth, New Hampshire). Read demultiplexing, trimming, and run analytics were carried out using the DRAGEN v.3.10.12 on-board analysis software installed in NextSeq2000 (https://help.dragen.illumina.com/).

De novo genome assembly and sequence quality check

Sequence reads were assembled into contigs using Shovill v.1.1.0 (https://github.com/tseemann/shovill) with the setting -trim. The quality of the genomes was assessed using Quast [81] and CheckM [82]. We excluded those genomes with < 90% completeness and > 5% contamination. These thresholds were recommended by CheckM [82]. To ensure that we only include high quality sequences and minimize the inclusion of fragmented sequences in our analyses, we also excluded assemblies with > 350 contigs and an N50 < 40,000 bp. In all, we used 557 genomes that passed our filtering criteria for all downstream analyses. All genomes were compared to two *E. coli* reference genomes K-12 substrain MG-1655 (Accession NC_000913.3) and O157:H7 strain Sakai (Accession NC_002695.2) from the National Center for Biotechnology Information (NCBI) using fastANI v.1.32 [83].

Phylogenetic tree reconstruction and population clustering

The genomes were annotated using Prokka v.1.14.6 with default parameters [84]. The annotated genomes were used as input into Panaroo v.1.3.3 for pan-genome analysis [85]. Sequences of individual genes were aligned using MAFFT [86]. Aligned sequences of the core and soft core genes ($n = 3{,}422$ genes) were concatenated and used as input to construct a core genome phylogenetic tree. We used SNP-sites v.2.5.1 [87] to extract single nucleotide polymorphisms (480,947 SNPs) from the aligned core and soft core genes. We used RAxML v.8.2.12 to generate a maximum likelihood phylogenetic tree [88] with a generalized time reversible model of nucleotide substitution [89] with gamma distribution rate of heterogeneity (GTR + G) based on the results of ModelFinder [90]. The core genome phylogenetic tree was visualized using the Interactive Tree of Life (IToL) [91]. We used the Bayesian hierarchical clustering algorithm fastBAPS v.1.0.8 (fast Bayesian Analysis of Population Structure) to partition the genomes into sequence clusters consisting of genetically similar individuals [43].

In silico determination of ST and serotypes

ST designation was determined using MLST v.2.19.0 (https://github.com/tseemann/mlst) based on allelic variation in seven single copy housekeeping genes (*adk, fumC, gyrB, icd, mdh, purA, recA*; [18]) and compared against previously characterized STs in the *E. coli* database in PubMLST [44]. Novel ST designations were assigned by the MLST sequence archive at EnteroBase. Serotypes were determined based on antigenic variation in flagellar (H) and polysaccharide (O) groups, which

were compared to sequences in the EcOH database [19] using ABRicate v.1.0.1 (https://github.com/tseemann/abricate). The EnteroBASE module EBEis was used to assigned serotypes to genomes that could not be assigned above [92].

Recombination detection

We used Gubbins v.3.2.1 [45] to identify regions of recombination in the genome alignments. Gubbins uses a sliding window approach to identify regions in the genome containing elevated densities of SNPs. We ran Gubbins separately on each of the largest sequence clusters. Because inference of recombination is made difficult with small sample sizes, we selected those clusters with at least 50 genomes (BAPS1, BAPS4, BAPS10, BAPS14; Fig. 1) regardless of the phylogenetic distances within each cluster. Hence, recombination is inferred only on the core genes of each sequence cluster. To identify the specific genes that were inferred to have recombined, we used snippy v.4.6.0 (https://github.com/tseemann/snippy) to align our genomes to the reference genome K-12 substrain MG-1655 (Accession NC_000913.3). We used the default options in Gubbins: minimum number of SNPs to identify a recombination block = 3; minimum window size = 100 bp; and maximum window size = 10,000 bp. Recombination events were visualized using Phandango [93]. Using SnpEff v.5.2e with default parameters [94], we identified the types of SNPs identified in recombined AMR and virulence genes in each of the four sequence clusters: synonymous (no change in amino acid), frameshift (change in reading frame), missense (amino acid replacement), nonsense (variant causing a stop codon). We ran SnpEff on the vcf output file generated by Gubbins containing all the SNPs detected in a BAPS sequence cluster with the K-12 substrain MG-1655 as the reference genome.

Statistical analysis

All statistical analyses were carried out using the ggstatsplot v.0.12.1 [95] package in R v.4.3.1 [96]. To measure homoscedasticity, we used the Breusch-pagan test from the ncvTest() function in R. Low p-value (below 0.05) means not homoscedastic. To measure normality, we used the Shapiro–Wilk's test in the shapiro.test() function in R. Low p-value (below 0.05) indicates that it is not normally distributed. We used a Wilcoxon test to compare the number of SNPs inside recombined sequences and the number of SNPS outside of recombined sequences (Fig. 2A) and a Games-Howell test to compare the number of base pairs in recombined sequences, the number of cumulative bases in recombined sequences, the r/m ratio, and tho/theta (Fig. 2B-F). Results were considered significant when $p < 0.05$.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11367-6.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements

The authors thank the UAlbany Research Technology Services where all bioinformatics analyses were carried out. We are grateful to the staff of DHMC for laboratory support.

Authors' contributions

K.R.P. and S.S.R.S. carried out all bioinformatics analyses. I.W.M. and A.W. oversaw sampling, clinical bacterial culturing, metadata collection, and archiving. S.S.R.S.and O.O.I. carried out subculturing and DNA extraction. C.P.A. and K.R.P. wrote the initial manuscript. C.P.A. guided the work. All authors read and approved the final manuscript.

Funding

The study was supported by the National Institutes of Health Award no. R35GM142924 to C.P.A. The funders had no role in study design, data collection and analysis, decision to publish, and preparation of the manuscript and the findings do not necessarily reflect views and policies of the authors' institutions and funders.

Data availability

The dataset supporting the conclusions of this article is included within the article and its supplementary files. Genome sequence data of *E. coli* isolates are available in the NCBI Sequence Read Archive under BioProject accession number PRJNA1066820. BioSample accession numbers for each genome are listed in Supplementary Table S1. The codes for running the different software are available in the Zenodo repository (https://zenodo.org/records/14736203).

Declarations

Ethics approval and consent to participate

Ethical approval was granted by the Committee for the Protection of Human Subjects of Dartmouth-Hitchcock Medical Center and Dartmouth College. This study protocol was deemed not to be human subjects research. Samples used in the study were subcultured bacterial isolates that had been archived in the routine course of clinical laboratory operations. No patient specimens were used and patient protected health information was not collected. Therefore, informed consent was not required.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Biological Sciences, University at Albany, State University of New York, Albany, NY, USA. [2]Department of Pathology and Laboratory Medicine, Dartmouth Hitchcock Medical Center and Dartmouth College Geisel School of Medicine, Lebanon, NH, USA.

Received: 12 September 2024 Accepted: 14 February 2025
Published online: 24 February 2025

References

1. van Elsas JD, Semenov AV, Costa R, Trevors JT. Survival of Escherichia coli in the environment: fundamental and public health aspects. ISME J. 2011;5:173–83.
2. Blount ZD. The unexhausted potential of E. coli. Elife. 2015;4:e05826.

3. Muloi DM, Wee BA, McClean DMH, Ward MJ, Pankhurst L, Phan H, et al. Population genomics of Escherichia coli in livestock-keeping households across a rapidly developing urban landscape. Nat Microbiol. 2022;7:581–9.
4. Lagerstrom KM, Hadly EA. Under-appreciated phylogroup diversity of Escherichia coli within and between animals at the urban-wildland interface. Appl Environ Microbiol. 2023;89:e0014223.
5. Kaper JB, Nataro JP, Mobley HL. Pathogenic Escherichia coli. Nat Rev Microbiol. 2004;2:123–40.
6. Pokharel P, Dhakal S, Dozois CM. The diversity of Escherichia coli pathotypes and vaccination strategies against this versatile bacterial pathogen. Microorganisms. 2023;11:344.
7. Sora VM, Meroni G, Martino PA, Soggiu A, Bonizzi L, Zecconi A. Extraintestinal pathogenic Escherichia coli: virulence factors and antibiotic resistance. Pathogens. 2021;10:1355.
8. Daga AP, Koga VL, Soncini JGM, de Matos CM, Perugini MRE, Pelisson M, et al. Escherichia coli bloodstream infections in patients at a university hospital: virulence factors and clinical characteristics. Front Cell Infect Microbiol. 2019;9:191.
9. Begier E, Rosenthal NA, Gurtman A, Kartashov A, Donald RGK, Lockhart SP. Epidemiology of invasive Escherichia coli infection and antibiotic resistance status among patients treated in US hospitals: 2009–2016. Clin Infect Dis. 2021;73:565–74.
10. Mashau RC, Meiring ST, Dramowski A, Magobo RE, Quan VC, Perovic O, et al. Culture-confirmed neonatal bloodstream infections and meningitis in South Africa, 2014–19: a cross-sectional study. Lancet Glob Health. 2022;10:e1170–8.
11. Doua J, Geurtsen J, Rodriguez-Baño J, Cornely OA, Go O, Gomila-Grange A, et al. Epidemiology, clinical features, and antimicrobial resistance of invasive Escherichia coli disease in patients admitted in tertiary care hospitals. Open Forum Infect Dis. 2023;10:ofad026.
12. Nhu NTK, Phan MD, Hancock SJ, Peters KM, Alvarez-Fraga L, Forde BM, et al. High-risk Escherichia coli clones that cause neonatal meningitis and association with recrudescent infection. eLife. 2024;12:RP91853.
13. Franz E, Rotariu O, Lopes BS, MacRae M, Bono JL, Laing C, et al. Phylogeographic analysis reveals multiple international transmission events have driven the global emergence of Escherichia coli O157:H7. Clin Infect Dis. 2019;69:428–37.
14. Tack DM, Kisselburgh HM, Richardson LC, Geissler A, Griffin PM, Payne DC, et al. Shiga toxin-producing Escherichia coli outbreaks in the United States, 2010–2017. Microorganisms. 2021;9:1529.
15. Sarno E, Pezzutto D, Rossi M, Liebana E, Rizzi V. A review of significant European foodborne outbreaks in the last decade. J Food Prot. 2021;84:2059–70.
16. Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet. 2022;399:629–55.
17. Geurtsen J, de Been M, Weerdenburg E, Zomer A, McNally A, Poolman J. Genomics and pathotypes of the many faces of Escherichia coli. FEMS Microbiol Rev. 2022;46:fuac031.
18. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, et al. Sex and virulence in Escherichia coli: an evolutionary perspective. Mol Microbiol. 2006;60:1136–51.
19. Ingle DJ, Valcanis M, Kuzevski A, Tauschek M, Inouye M, Stinear T, et al. In silico serotyping of E. coli from short read data identifies limited novel O-loci but extensive diversity of O: H serotype combinations within and between pathogenic lineages. Microb Genom. 2016;2:e000064.
20. Liu B, Furevi A, Perepelov AV, Guo X, Cao H, Wang Q, et al. Structure and genetics of Escherichia coli O antigens. FEMS Microbiol Rev. 2020;44:655–83.
21. Gonzalez-Alba JM, Baquero F, Cantón R, Galán JC. Stratified reconstruction of ancestral Escherichia coli diversification. BMC Genomics. 2019;20:936.
22. Abram K, Udaondo Z, Bleker C, Wanchai V, Wassenaar TM, Robeson MS, et al. Mash-based analyses of Escherichia coli genomes reveal 14 distinct phylogroups. Commun Biol. 2021;4:117.
23. Cummins EA, Hall RJ, Connor C, McInerney JO, McNally A. Distinct evolutionary trajectories in the Escherichia coli pangenome occur within sequence types. Microb Genom. 2022;8:mgen000903.
24. Hall RJ, Whelan FJ, Cummins EA, Connor C, McNally A, McInerney JO. Gene-gene relationships in an Escherichia coli accessory genome are linked to function and mobility. Microb Genom. 2021;7:000650.
25. Shaw LP, Chau KK, Kavanagh J, AbuOun M, Stubberfield E, Gweon HS, et al. Niche and local geography shape the pangenome of wastewater- and livestock-associated Enterobacteriaceae. Sci Adv. 2021;7:eabe3868.
26. Tantoso E, Eisenhaber B, Kirsch M, Shitov V, Zhao Z, Eisenhaber F. To kill or to be killed: pangenome analysis of Escherichia coli strains reveals a tailocin specific for pandemic ST131. BMC Biol. 2022;20:146.
27. Didelot X, Méric G, Falush D, Darling AE. Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. BMC Genomics. 2012;13:256.
28. Dixit PD, Pang TY, Studier FW, Maslov S. Recombinant transfer in the basic genome of Escherichia coli. Proc Natl Acad Sci U S A. 2015;112:9070–5.
29. Cobo-Simón M, Hart R, Ochman H. Escherichia coli: what is and which are? Mol Biol Evol. 2023;40:msac273.
30. Conrad RE, Brink CE, Viver T, Rodriguez-R LM, Aldeguer-Riquelme B, Hatt JK, et al. Microbial species and intraspecies units exist and are maintained by ecological cohesiveness coupled to high homologous recombination. Nat Commun. 2024;15:9906.
31. Clark AJ. rec genes and homologous recombination proteins in Escherichia coli. Biochimie. 1991;73:523–32.
32. Spies M, Kowalczykowski SC. Homologous recombination by the RecBCD and RecF pathways. In: The bacterial chromosome. Washington, D.C: American Society for Microbiology Press; 2004. p. 389–403.
33. Buljubašić M, Hlevnjak A, Repar J, Đermić D, Filić V, Weber I, et al. RecBCD-RecFOR-independent pathway of homologous recombination in Escherichia coli. DNA Repair (Amst). 2019;83:102670.
34. Paul S, Linardopoulou EV, Billig M, Tchesnokova V, Price LB, Johnson JR, et al. Role of homologous recombination in adaptive diversification of extraintestinal Escherichia coli. J Bacteriol. 2013;195:231–42.
35. Chen L, Peirano G, Kreiswirth BN, Devinney R, Pitout JDD. Acquisition of genomic elements were pivotal for the success of Escherichia coli ST410. J Antimicrob Chemother. 2022;77:3399–407.
36. Ba X, Guo Y, Moran RA, Doughty EL, Liu B, Yao L, et al. Global emergence of a hypervirulent carbapenem-resistant Escherichia coli ST410 clone. Nat Commun. 2024;15:494.
37. Tchesnokova V, Radey M, Chattopadhyay S, Larson L, Weaver JL, Kisiela D, et al. Pandemic fluoroquinolone resistant Escherichia coli clone ST1193 emerged via simultaneous homologous recombinations in 11 gene loci. Proc Natl Acad Sci U S A. 2019;116:14740–8.
38. Riley LW. Pandemic lineages of extraintestinal pathogenic Escherichia coli. Clin Microbiol Infect. 2014;20:380–90.
39. Riley LW. Distinguishing pathovars from nonpathovars: Escherichia coli. Microbiol Spectr. 2020;8:8.
40. Patiño-Navarrete R, Rosinski-Chupin I, Cabanel N, Gauthier L, Takissian J, Madec J-Y, et al. Stepwise evolution and convergent recombination underlie the global dissemination of carbapenemase-producing Escherichia coli. Genome Med. 2020;12:10.
41. Rodríguez-Beltrán J, Tourret J, Tenaillon O, López E, Bourdelier E, Costas C, et al. High recombinant frequency in extraintestinal pathogenic Escherichia coli strains. Mol Biol Evol. 2015;32:1708–16.
42. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr Opin Genet Dev. 2005;15:589–94.
43. Tonkin-Hill G, Lees JA, Bentley SD, Frost SDW, Corander J. Fast hierarchical Bayesian analysis of population structure. Nucleic Acids Res. 2019;47:5539–49.
44. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res. 2018;3:124.
45. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015;43:e15.
46. Barnhart MM, Chapman MR. Curli biogenesis and function. Annu Rev Microbiol. 2006;60:131–47.
47. Raymond KN, Dertz EA, Kim SS. Enterobactin: an archetype for microbial iron transport. Proc Natl Acad Sci U S A. 2003;100:3584–8.
48. Bullen JJ, Rogers HJ, Spalding PB, Ward CG. Iron and infection: the heart of the matter. FEMS Immunol Med Microbiol. 2005;43:325–30.
49. Kakkanat A, Totsika M, Schaale K, Duell BL, Lo AW, Phan M-D, et al. The role of H4 flagella in Escherichia coli ST131 virulence. Sci Rep. 2015;5:16149.
50. Sevrin G, Massier S, Chassaing B, Agus A, Delmas J, Denizot J, et al. Adaptation of adherent-invasive E. coli to gut environment: impact on flagellum expression and bacterial colonization ability. Gut Microbes. 2020;11:364–80.

51. Nagakubo S, Nishino K, Hirata T, Yamaguchi A. The putative response regulator BaeR stimulates multidrug resistance of Escherichia coli via a novel multidrug exporter system, MdtABC. J Bacteriol. 2002;184:4161–7.
52. Baranova N, Nikaido H. The baeSR two-component regulatory system activates transcription of the yegMNOB (mdtABCD) transporter gene cluster in Escherichia coli and increases its resistance to novobiocin and deoxycholate. J Bacteriol. 2002;184:4168–76.
53. Tanabe H, Yamasak K, Furue M, Yamamoto K, Katoh A, Yamamoto M, et al. Growth phase-dependent transcription of emrKY, a homolog of multidrug efflux emrAB genes of Escherichia coli, is induced by tetracycline. J Gen Appl Microbiol. 1997;43:257–63.
54. Pasqua M, Grossi M, Scinicariello S, Aussel L, Barras F, Colonna B, et al. The MFS efflux pump EmrKY contributes to the survival of Shigella within macrophages. Sci Rep. 2019;9:2906.
55. Zgurskaya HI, Krishnamoorthy G, Ntreh A, Lu S. Mechanism and function of the outer membrane channel TolC in multidrug resistance and physiology of Enterobacteria. Front Microbiol. 2011;2:189.
56. Du D, Wang Z, James NR, Voss JE, Klimont E, Ohene-Agyei T, et al. Structure of the AcrAB-TolC multidrug efflux pump. Nature. 2014;509:512–5.
57. Zhang CZ, Chang MX, Yang L, Liu YY, Chen PX, Jiang HX. Upregulation of AcrEF in quinolone resistance development in Escherichia coli when AcrAB-TolC function is impaired. Microb Drug Resist. 2018;24:18–23.
58. Imuta N, Nishi J, Tokuda K, Fujiyama R, Manago K, Iwashita M, et al. The Escherichia coli efflux pump TolC promotes aggregation of enteroaggregative E. coli 042. Infect Immun. 2008;76:1247–56.
59. Zhu S, Alexander MK, Paiva TO, Rachwalski K, Miu A, Xu Y, et al. The inactivation of tolC sensitizes Escherichia coli to perturbations in lipopolysaccharide transport. iScience. 2024;27:109592.
60. Pang TY, Lercher MJ. Each of 3,323 metabolic innovations in the evolution of E. coli arose through the horizontal transfer of a single DNA segment. Proc Natl Acad Sci U S A. 2019;116:187–92.
61. Wyres KL, Wick RR, Judd LM, Froumine R, Tokolyi A, Gorrie CL, et al. Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of Klebsiella pneumoniae. PLoS Genet. 2019;15:e1008114.
62. MacAlasdair N, Pesonen M, Brynildsrud O, Eldholm V, Kristiansen PA, Corander J, et al. The effect of recombination on the evolution of a population of Neisseria meningitidis. Genome Res. 2021;31:1258–68.
63. Driebe EM, Sahl JW, Roe C, Bowers JR, Schupp JM, Gillece JD, et al. Using whole genome analysis to examine recombination across diverse sequence types of Staphylococcus aureus. PLoS One. 2015;10:e0130955.
64. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, et al. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet. 2014;46:305–9.
65. Park S-C, Lee K, Kim YO, Won S, Chun J. Large-scale genomics reveals the genetic characteristics of seven species and importance of phylogenetic distance for estimating pan-genome size. Front Microbiol. 2019;10:834.
66. Horesh G, Taylor-Brown A, McGimpsey S, Lassalle F, Corander J, Heinz E, et al. Different evolutionary trends form the twilight zone of the bacterial pangenome. Microb Genom. 2021;7:000670.
67. Bobay L-M, Traverse CC, Ochman H. Impermanence of bacterial clones. Proc Natl Acad Sci U S A. 2015;112:8893–900.
68. Santos ACDM, Santos FF, Silva RM, Gomes TAT. Diversity of hybrid- and hetero-pathogenic Escherichia coli and their potential implication in more severe diseases. Front Cell Infect Microbiol. 2020;10:339.
69. Gati NS, Middendorf-Bauchart B, Bletz S, Dobrindt U, Mellmann A. Origin and evolution of hybrid shiga toxin-producing and uropathogenic Escherichia coli strains of sequence type 141. J Clin Microbiol. 2019;58:e01309–19.
70. Coyle NM, Bartie KL, Bayliss SC, Bekaert M, Adams A, McMillan S, et al. A hopeful sea-monster: a very large homologous recombination event impacting the core genome of the marine pathogen Vibrio anguillarum. Front Microbiol. 2020;11:1430.
71. Nandi T, Holden MTG, Didelot X, Mehershahi K, Boddey JA, Beacham I, et al. Burkholderia pseudomallei sequencing identifies genomic clades with distinct recombination, accessory, and epigenetic profiles. Genome Res. 2015;25:129–41.
72. Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, et al. Systematic discovery of antiphage defense systems in the microbial pangenome. Science. 2018;359:eaar4120.
73. Sheppard SK, Cheng L, Méric G, de Haan CPA, Llarena A-K, Marttinen P, et al. Cryptic ecology among host generalist Campylobacter jejuni in domestic animals. Mol Ecol. 2014;23:2442–51.
74. Sniegowski PD, Murphy HA. Evolvability. Curr Biol. 2006;16:R831–834.
75. Lobkovsky AE, Wolf YI, Koonin EV. Evolvability of an optimal recombination rate. Genome Biol Evol. 2015;8:70–7.
76. Carja O, Liberman U, Feldman MW. Evolution in changing environments: modifiers of mutation, recombination, and migration. Proc Natl Acad Sci U S A. 2014;111:17935–40.
77. McDonald MJ, Rice DP, Desai MM. Sex speeds adaptation by altering the dynamics of molecular evolution. Nature. 2016;531:233–6.
78. Mostowy R, Croucher NJ, Hanage WP, Harris SR, Bentley S, Fraser C. Heterogeneity in the frequency and characteristics of homologous recombination in pneumococcal evolution. PLoS Genet. 2014;10:e1004300.
79. Sorek R, Lawrence CM, Wiedenheft B. CRISPR-mediated adaptive immune systems in bacteria and archaea. Annu Rev Biochem. 2013;82:237–66.
80. Holt KE, Lassalle F, Wyres KL, Wick R, Mostowy RJ. Diversity and evolution of surface polysaccharide synthesis loci in Enterobacteriales. ISME J. 2020;14:1713–30.
81. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–5.
82. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55.
83. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9:5114.
84. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.
85. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol. 2020;21:180.
86. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
87. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom. 2016;2:e000056.
88. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.
89. Tavaré S. Some probabilistic and statistical problems in the analysis of DNA sequences. American Mathematical Society Lect Math Life Sci. 1986;17:57–86.
90. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14:587–9.
91. Letunic I, Bork P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucleic Acids Res. 2024;2:W78–82.
92. Zhou Z, Alikhan N-F, Mohamed K, Fan Y, Agama Study Group, Achtman M. The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. Genome Res. 2020;30:138–52.
93. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR. Phandango: an interactive viewer for bacterial population genomics. Bioinformatics. 2018;34:292–3.
94. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6:80–92.
95. Patil I. Visualizations with statistical details: the "ggstatsplot" approach. Journal of Open Source Software. 2021;6:3167.
96. R Core Team. R: a language and environment for statistical computing. 2021. https://www.R-project.org/.

Publisher's Note