Molecular Biology and Evolution, 2025, **42**, 1–10
https://doi.org/10.1093/molbev/msae272
Advance access publication 3 January 2025
Discoveries

MBE OXFORD

Quantifying the Evolutionary Dynamics of Structure and Content in Closely Related *E. coli* Genomes

Marco Molari,[1,2] **Liam P. Shaw**,[3,4] **Richard A. Neher**[1,2]*

[1]Swiss Institute of Bioinformatics, Basel, Switzerland
[2]Biozentrum, University of Basel, Basel, Switzerland
[3]Department of Biology, University of Oxford, Oxford, UK
[4]Department of Biosciences, University of Durham, Durham, UK

***Corresponding author:** E-mail: richard.neher@unibas.ch.

Associate editor: Russell Corbett-Detig

Abstract

Bacterial genomes primarily diversify via gain, loss, and rearrangement of genetic material in their flexible accessory genome. Yet the dynamics of accessory genome evolution are very poorly understood, in contrast to the core genome where diversification is readily described by mutations and homologous recombination. Here, we tackle this problem for the case of very closely related genomes. We comprehensively describe genome evolution within $n = 222$ genomes of *Escherichia coli* ST131, which likely shared a common ancestor around 100 years ago. After removing putative recombinant diversity, the total length of the phylogeny is 6,000 core genome substitutions. Within this diversity, we find 22 modifications to core genome synteny and estimate around 2,000 structural changes within the accessory genome, i.e. one structural change for every three core genome substitutions. Sixty-three percent of loci with structural diversity could be resolved into individual gain and loss events with 10-fold more gains than losses, demonstrating a dominance of gains due to insertion sequences and prophage integration. Our results suggest the majority of synteny changes and insertions in our dataset are likely deleterious and only persist for a short time before being removed by purifying selection.

Keywords: horizontal gene transfer, genome evolution, mobile genetic elements

Introduction

Many microbial species are able to acquire new genes and advantageous alleles in a process known as horizontal gene transfer (HGT) (Arnold et al. 2022), in which genetic material is transferred from one microbe to the next. The horizontal movement of genes is often mediated by mobile genetic elements (MGEs) (Haudiquet et al. 2022; Tokuda and Shintani 2024) such as plasmids, transposons, and phages. These elements participate in a complex evolutionary interplay with their host and sometimes carry cargo that is beneficial to the host, such as antibiotic resistance genes or defense systems against other MGEs (Georjon and Bernheim 2023; Mayo-Muñoz et al. 2023). This makes HGT one of the major drivers of the spread of antimicrobial resistance (Von Wintersdorff et al. 2016; Partridge et al. 2018), and understanding its dynamics and effects on the microbial genome is of crucial importance to combat this spread.

As a consequence of HGT, bacteria in the same species can differ dramatically in the content of their genomes, with any given individual only possessing a fraction of the total collection of genes present in the species, named the *pangenome* (Brockhurst et al. 2019). The pangenome can be split in two components: the *core* genome, which refers to the set of genes that are common to all isolates, and the *accessory* genome, indicating genes that are specific to a subset of isolates, sometimes associated to adaptations to specific niches or lifestyles (Touchon et al. 2020).

Given this split into core and accessory genome, the similarity of genomes is often quantified by two different metrics: the number of substitutions in the alignable core regions and the overlap within the accessory genome—that is, how many accessory genes two genomes have in common. In many microbial species, genomes that differ by very few substitutions in their core genome can show large differences in accessory genome content (Doolittle and Zhaxybayeva 2009; Touchon et al. 2009, 2020). The horizontal exchange of accessory genes thus has to be very fast, but a quantitative understanding of this dynamics is lacking. Yet despite this flexibility the structural organization of the genome remains strongly conserved, with the order of core genes being maintained even across large evolutionary distances (Rocha 2006, 2008).

To quantify how genomic diversity is generated and accumulates over time, one would need to observe individual structural modifications on short evolutionary timescales. However, systematically identifying structural changes (gain, loss, or changes in gene order) and quantifying their rates is a difficult problem with two complementary challenges. First, detecting structural changes across a large set of isolates requires comparisons of only partially alignable and potentially shuffled portions of the genome, which is technically and computationally challenging. Second, homologous recombination in the core genome is sufficiently rapid that the genealogy of bacteria can often no longer be reconstructed (Sakoparnig et al. 2021), meaning one lacks the consistent estimates of divergence between strains which are necessary to calibrate rates of structural evolution.

Here, we aim to overcome these challenges and accurately quantify rates of structural evolution in *Escherichia coli*. To address the challenge of a reliable clonal phylogeny, we focus

Received: August 21, 2024. **Revised:** December 11, 2024. **Accepted:** December 27, 2024

on the recently expanded sequence type ST131 where homologous recombination is still a perturbation that can be controlled for. To identify individual modifications to genome structure and content, we use a *pangenome graph* representation for the genomes in our dataset, as provided by *PanGraph* (Noll et al. 2023). Within this gene-agnostic representation, genomes are encoded as paths through alignments of homologous regions. Any structural changes appear as deviations between paths. This approach allows us to identify changes in core genome order and find regions harboring structural changes in the accessory genome. We can then not only quantify their diversity and size but also infer the rates at which structural diversity accumulates, thereby providing a quantitative picture of the evolution of genome structure on short timescales.

Materials and Methods

A detailed description of the materials and methods used in this study is provided in the Supplementary material. Moreover, the results of our analysis can be fully reproduced with the pipeline available on GitHub at https://github.com/mmolari/pangraph_ecolist131_paper (v1.1) and archived on Zenodo at https://doi.org/10.5281/zenodo.14576202.

Results

Dataset Overview and Graph Construction

The multidrug-resistant sequence type ST131, first identified in 2008, is associated with the expression of extended-spectrum beta-lactamases and is a major cause of antibiotic-resistant urinary tract infections worldwide (Stoesser et al. 2016; Decano and Downing 2019; Pitout and Finn 2020). The most recent common ancestor of all ST131 is estimated to have lived around 1,900 (95% highest posterior density interval 1,842 to 1,948; Ludden et al. 2020) with most of the clonal expansion taking place in the last 35 years driven by the widespread use of beta-lactam antibiotics (Ben Zakour et al. 2016; Stoesser et al. 2016; Kallonen et al. 2017; Gladstone et al. 2021). Due to its importance for public health there are a large number of high-quality *E. coli* ST131 genomes in public databases, representing a set of closely related genomes that we can be confident diverged only recently.

We downloaded all complete *E. coli* assemblies from RefSeq (O'Leary et al. 2016), and selected the ones that were classified as ST131 by multilocus sequence typing and had a mash (Ondov et al. 2016) distance <0.008 from the ST131 reference `NC_013654.1` (see supplementary section S1, Supplementary Material online for more details). After quality control and filtering the dataset included $n = 222$ isolates (average genome size 5.1 ± 0.1 Mbp), with a large number of genomes from environmental sampling in Switzerland (Biggel et al. 2023) and hospital sampling in Sweden (Jaén-Luchoro et al. 2023; see supplementary fig. S1, Supplementary Material online). We used PanGraph (Noll et al. 2023) to losslessly encode all chromosomal sequences in the form of a pangenome graph (a "pangraph") that we could query for all downstream analyses (see supplementary section S2, Supplementary Material online). By summing the lengths of all block sequences in our graph we get a total pangenome size of 8 Mbp. Analogously, the core genome size is 3.6 Mbp and the soft-core (present in >95% of isolates) 4.2 Mbp (see supplementary fig. S3, Supplementary Material online). As expected, isolates from the same sequence type are extremely similar in their core genome (<0.01% divergence on average if we exclude recombined regions, see below). However, they can differ by hundreds of kbp in their accessory genome (see supplementary fig. S6, Supplementary Material online).

As a first step, we extracted all core blocks from the pangenome graph and built a core genome alignment. Due to the recent divergence of ST131 the density of SNPs in this alignment is very small, except for regions that underwent homologous recombination with isolates outside of ST131. These regions are visible as islands of high SNP density (Sakoparnig et al. 2021; see supplementary fig. S4, Supplementary Material online). Using a strategy similar to Gubbins (Croucher et al. 2015), we excluded these regions by filtering out from the alignment any 2 kbp window in which an isolate has more than three mutations with respect to the consensus (see supplementary section S3, Supplementary Material online). This filtering procedure reduced the size of the core genome alignment roughly by one-third (from 3.6 to 2.4 Mbp) while at the same time removing 70% of the polymorphic positions (from 22 to 6 k polymorphic sites in the alignment) that are clustered in the recombined regions.

After removing these putative recombinant regions, we used *FastTree* (Price et al. 2010) to infer a phylogenetic tree from the filtered alignment (Fig. 1). The resulting phylogeny is compatible with the known global diversity of *E. coli* ST131, with representatives from the three main clades A, B, and C (Pitout and Finn 2020). Locus typing, plasmid typing, and resistance gene presence are broadly consistent with the phylogeny. This phylogeny and the filtered alignment are mostly compatible and only a few homoplasies remained when inferring ancestral sequences on the phylogeny (2.8% of all polymorphic sites, compared to 35% before recombination filtering, see supplementary section S3B and supplementary fig. S5, Supplementary Material online). This suggests that our recombination filtering not only removed recombination with divergent genomes outside of our dataset, but that residual recombination within our dataset does not substantially distort the phylogeny.

Core Genome Synteny Mutations

Previous work has established that the order of core genes is conserved even across large evolutionary distances (Rocha 2006, 2008; Touchon et al. 2009). Yet it is unclear whether this conservation is due to a very low rate of introduction of synteny changes (synteny mutation rate), or because such changes to synteny are deleterious and never persist. We, therefore, first surveyed patterns of synteny variation in our *E. coli* ST131 dataset. Considering only alignable regions (blocks) larger than 500 bp, the 3.59 Mbp core genome of ST131 is divided into 32 maximal syntenic regions (range: 1 to 478 kbp; see supplementary section S3.C, Supplementary Material online).

The vast majority of isolates in our dataset (196/222) have these blocks in the same consensus order and the remaining 26 isolates show 22 different synteny patterns (Fig. 2; see supplementary fig. S7, Supplementary Material online for the consensus order mapped to the genome). Most of the nonconsensus patterns are observed in a single isolate, likely caused by events that occurred on the terminal branches of the phylogenetic tree. There are, however, four patterns that are shared by pairs of isolates in close phylogenetic proximity, indicating that the patterns likely originated from a single event on an internal branch of the tree. Most patterns deviate from the

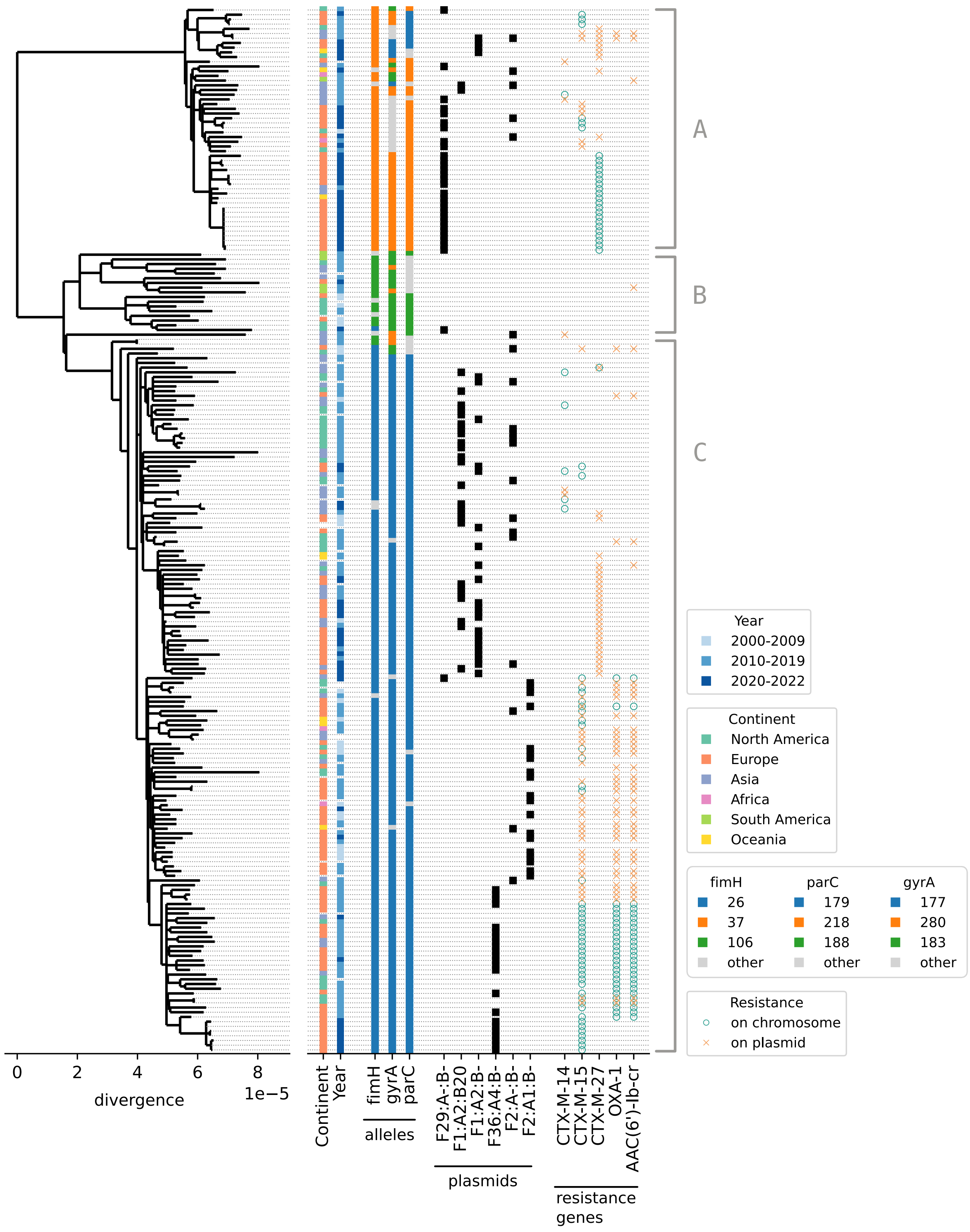

Fig. 1. Core genome tree reconstructed after removing regions with suspected recombination. Annotations on the right report geographic origin and isolation year, *fimH*, *gyrA*, and *parC* alleles, IncF plasmid type, and the presence of particular resistance genes, see supplementary section S1, Supplementary Material online for details. The tree recapitulates the known global diversity of *E. coli* ST131 with representatives from the three main clades A, B, and C.

consensus pattern by a single inversion, often centered around the origin or terminus of replication consistent with previous observations (Eisen et al. 2000). There are four exceptions: the patterns of isolates NZ_CP107184 and NZ_CP107182 feature two inversions, while in isolates NZ_CP107117 and NZ_CP124372 the inverted region is also translocated.

Isolates NZ_CP069583.1 and NZ_CP049085.2 are the only case of a synteny pattern shared by two isolates that are

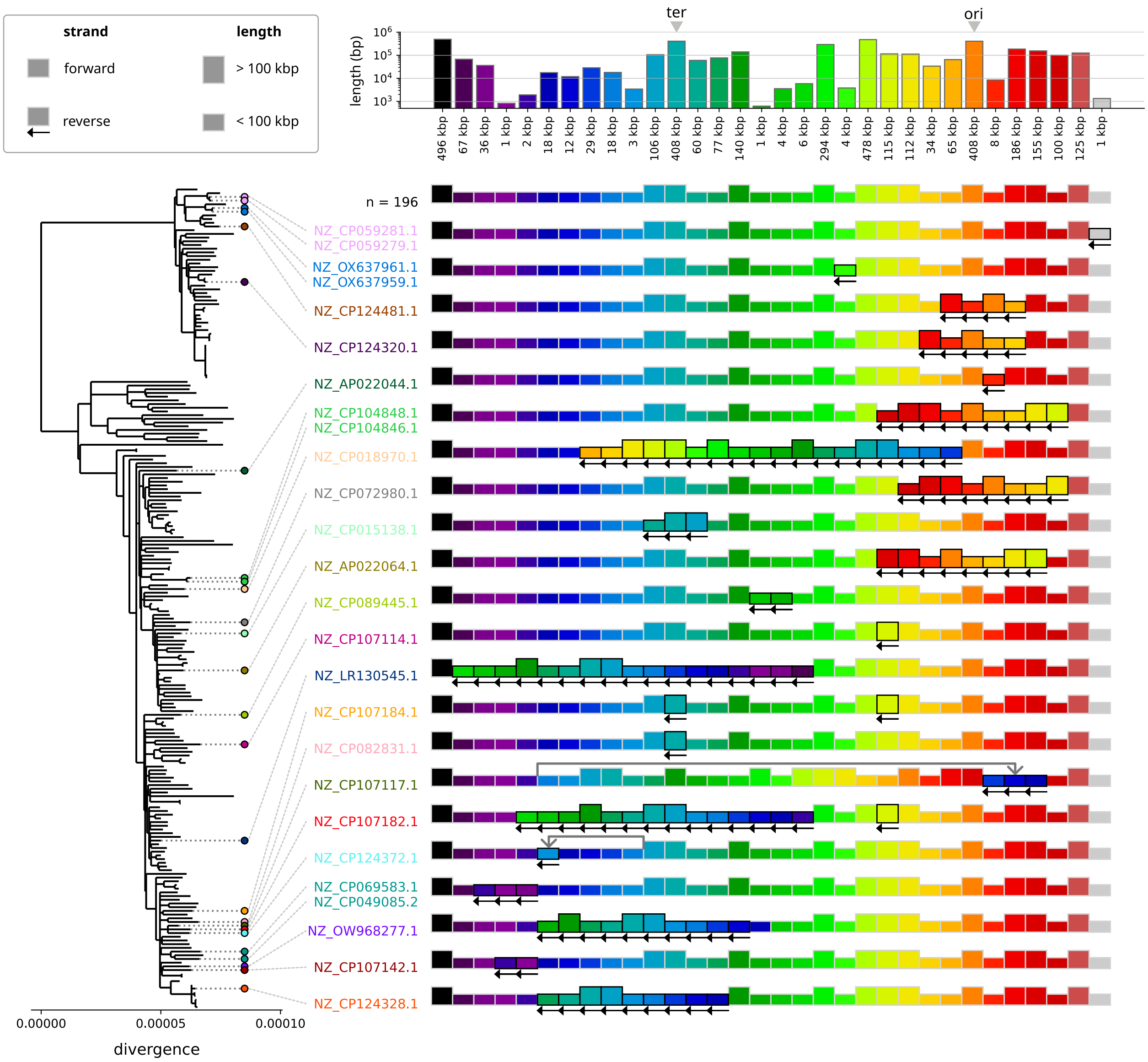

Fig. 2. Variation in core genome synteny across ST131. Each block represents a region of core sequence that is syntenic in all isolates, and each row a particular order of these regions in the chromosome. Most isolates (196/222) have the consensus order reported in the first row. Each subsequent row represents a different order present in other isolates, highlighted in corresponding color on the core genome tree. The height of a block is related to its length (see top), while left-pointing arrows indicate reverse-complemented blocks. Note that this analysis uses the entire core genome (3.59 Mbp) before filtering out suspected homologous recombination.

not nearest neighbors. These form a clade together with isolate NZ_CP103557.1, which does not feature the inversion. We investigated this case in greater detail (see supplementary fig. S8, Supplementary Material online) and found a high density of repetitive sequence flanking the inversion, compatible with a reversion of this inversion. Another interesting observation is the presence of the same 115 kb inversion in three different synteny patterns (isolates NZ_CP107182.1, NZ_CP107184.1, and NZ_CP107114.1). A detailed look reveals that the regions flanking the inversions are structurally different in all three isolates (see supplementary fig. S9, Supplementary Material online) which is compatible with independent inversion events in a structurally very dynamic region.

The total tree length corresponds to 6,040 substitutions in the core genome, implying that events that change core genome synteny happen at a rate of roughly one every 270 substitutions on the (filtered) core genome. Note that here we only consider clonally acquired substitutions and exclude the ones likely obtained in recombination events. The fact that the majority of synteny changes is observed on terminal branches suggest that many of them might be deleterious. However, as expected for populations that recently expanded, the core genome tree is dominated by terminal branches (80% of total branch length) such that there is little power to test for overrepresentation of inversions or translocations on terminal branches.

Accessory Genome Diversification

The strong conservation of synteny in ST131 provides a good frame of reference to study variation of the accessory genome. Any pair of core blocks defines a specific context that is present in all isolates with very few exceptions due to synteny breaks (see supplementary fig. S11, Supplementary Material online). Accessory genome variation between a pair of consecutive core blocks can thus be meaningfully compared across isolates

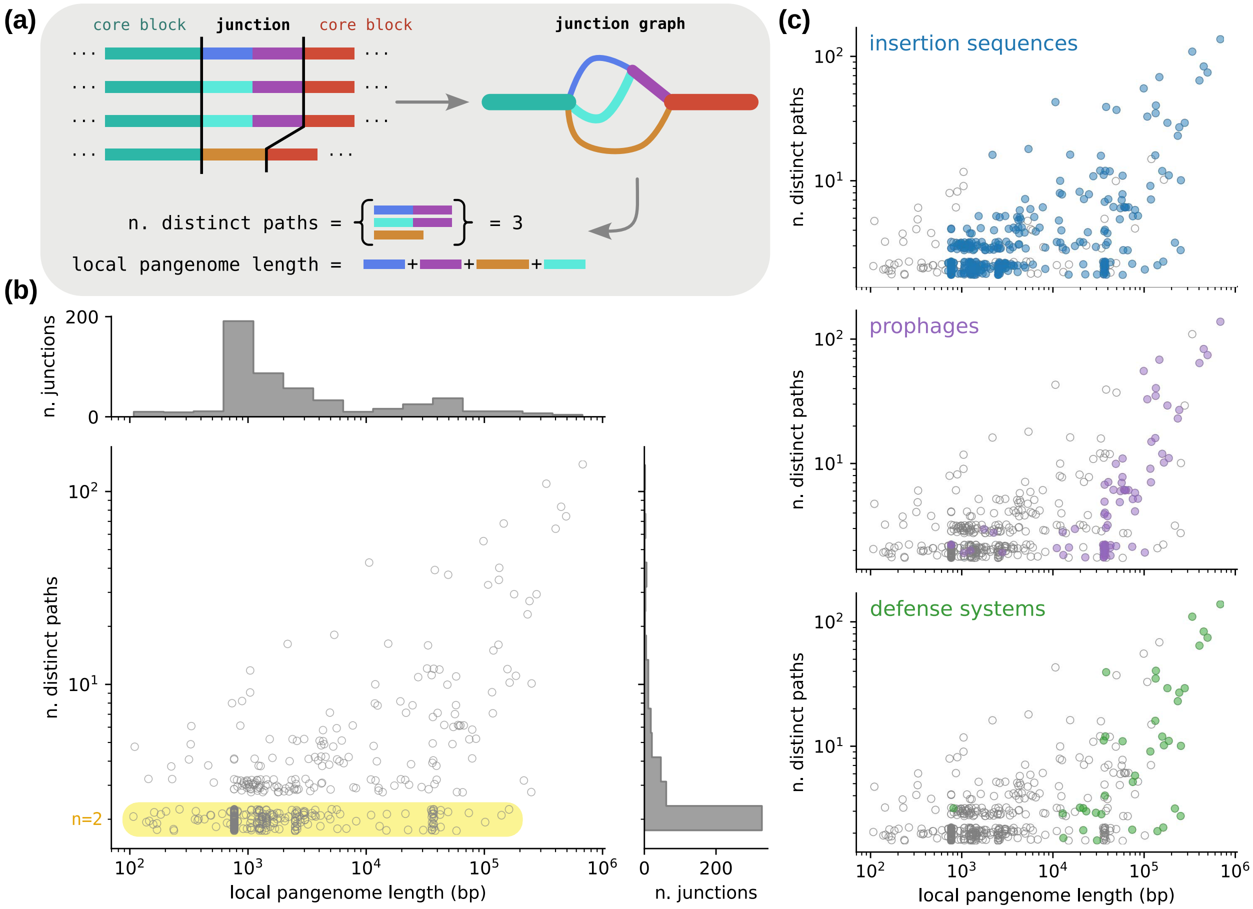

Fig. 3. Structural diversity in the accessory genome. a) Junction graphs are created by running PanGraph on a region spanning the interval between two consecutive core blocks, including the core blocks themselves. For each graph, we evaluate the number of different distinct paths and the total local accessory pangenome length. b) Joint distribution of junction size, measured as accessory pangenome content, and complexity, measured as total different number of paths, for the backbone junctions in our dataset. To better visualize regions of high density where many points are overlapping, we added a vertical uniform random displacement to each point of magnitude ±0.25. Binary junctions with only two alternative paths are highlighted in yellow. c) The same distribution as (b) and highlights junctions that contain ISs (top panel, blue), prophages (middle panel, purple), and defense systems (bottom panel, green). See main text for details.

in the dataset. This diversity likely arose through gains and loss since the most recent common ancestor of the genomes in the dataset.

We call a pair of consecutive core blocks and the enclosed accessory genome a core genome *junction*. We analyzed each junction in isolation by extracting the genetic region that spans the two flanking core blocks from every isolate and then use PanGraph to build a small local genome graph for this region (Fig. 3a), related to the idea of local graphs in Colquhoun et al. (2021). While this local graph is in principle embedded in the global graph, rebuilding the local graph increases reliability and avoids excessive graph fragmentation due to homology with other regions in the genome, e.g. for heavily duplicated elements. We found a total of $n = 519$ such core block pairs flanking accessory variation, for details see supplementary section S6, Supplementary Material online.

We characterized each junction by its *diversity*, measured in terms of number of distinct realized paths between the flanking core blocks, and its *size*, expressed as the total length (bp) of the accessory genome contained between the two core blocks (*local pangenome length*). The joint distribution of these two quantities (Fig. 3b) shows the large heterogeneity of structural diversity across the genome; 328 of the 519 junctions (63%) display only two different distinct paths, and often one of the two is an empty path (i.e. no accessory content) corresponding to an individual gain/loss event. Their size distribution is bimodal, with a peak around 1 kbp due mainly to insertion sequences (ISs), and another peak at 30 to 40 kbp corresponding to prophages (Fig. 3c). On the other end of the spectrum are a small number of junctions of very high diversity, with many different paths and more than 100 kbp of distinct genetic sequence.

We characterized junctions by the presence of specific features such as ISs, prophages, or defense systems (see Fig. 3c). IS are detected using *ISEScan* (Xie and Tang 2017). Each isolate has an average of 70 ISs on their chromosome. The three most represented families are IS3, IS66, and IS21, see supplementary fig. S12, Supplementary Material online. ISs are present in most junctions (451/519) and account for the large number of binary junctions of size 1 kbp. Prophages, detected using *geNomad* (Camargo et al. 2024), are present in many large and complex junctions, and they also account for a large fraction of binary junctions of size 30 to 40 kbp (90% of such junctions contain prophages). Each chromosome harbors on average 6 prophages (supplementary fig. 13, Supplementary Material online). Defense systems were detected using *DefenseFinder* (Tesson et al. 2022). As expected they are located mostly in large and diverse junctions and are often

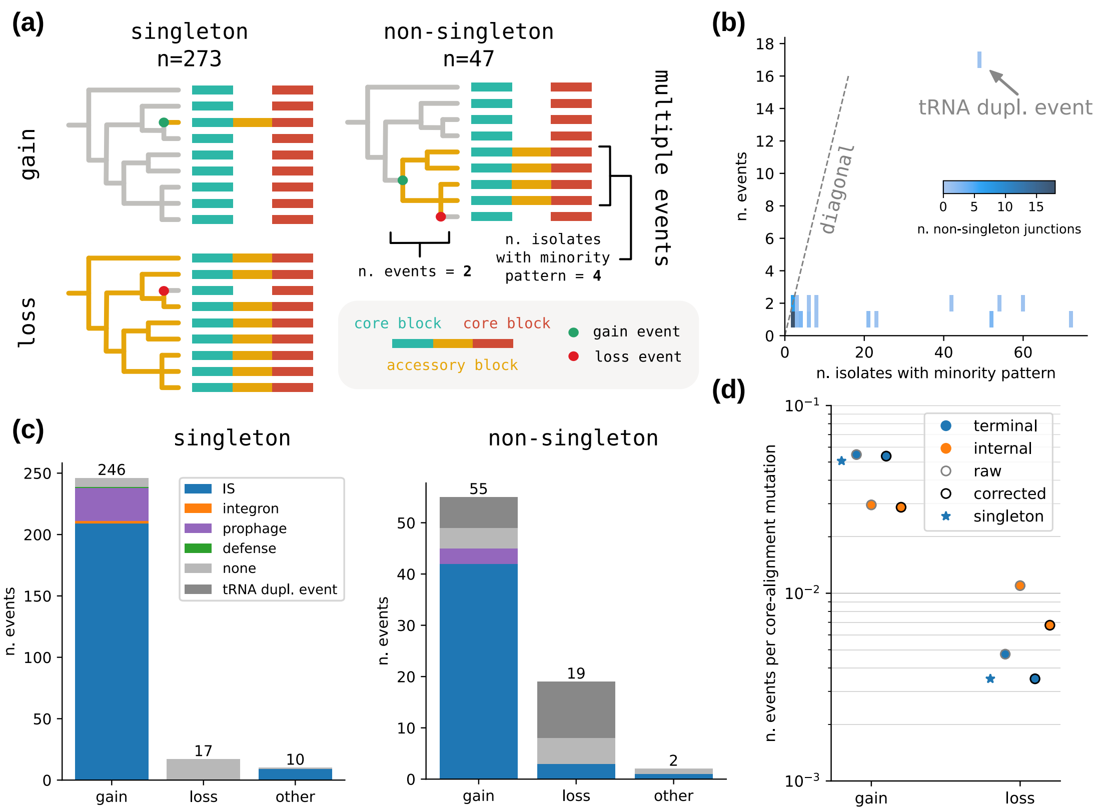

Fig. 4. Quantifying rates of genome structure evolution. a) Junctions with *singleton* structural diversity are trivially classified as single gain or loss events. *Nonsingleton* junctions, i.e. where more than one isolate follows the majority path, are often also explained by a single event, but can have more complex pattern and require multiple gain/loss events to be explained. b) The number of gain/loss events required (*y*-axis) to explain nonsingleton junctions is mostly one or two regardless of the prevalence of the minority path (*x*-axis), with a single exception that corresponds to a tRNA duplication event. c) number of gain/loss/other events for singleton and nonsingleton junctions, stratified by associated annotation (see main text). d) Estimates of gain and loss rates obtained by dividing the number of events by the corresponding length of the core genome tree measured as expected number of substitutions. The different estimate are for: just singleton events (star), all events on terminal branches (blue), and events on internal branches (orange). The correction (black edge) corresponds to removing events attributed to the single junction featuring the tRNA duplication.

carried by prophages or colocalized with ISs. Each isolate has around 10 different defense systems (supplementary fig. S14, Supplementary Material online). We also verified that most of the diverse junctions identified by our analysis are in the same core genome location as the defense system "hotspots" identified in Hochhauser et al. (2023) (see supplementary section S10 and supplementary fig. S20, Supplementary Material online). Finally, we used *IntegronFinder* (Néron et al. 2022) to locate integrons. However, we could detect only 11 integron-associated junctions in our graph.

Estimating Rates of Accessory Genome Evolution

The structure of many of the diverse junctions with their dozens of different paths is hard to decompose into individual gain, loss, or transposition events. We, therefore, first focus on the 63% of junctions with binary variation (Fig. 3b) whose diversity can usually be interpreted in terms of gain or loss of accessory segments.

We only considered binary *backbone* junctions ($n = 320$), i.e. junctions associated to core-edges that are present in all isolates. This only excluded few ($n = 8$) junctions flanking synteny breaks in some isolates. We further subdivided these junctions into *singletons* ($n = 273$, 85%), i.e. where the minority pattern is present only in one isolate, and *nonsingletons* ($n = 47$, 15%) (Fig. 4a). Determining whether a singleton is a gain or loss event amounts to verifying whether the accessory region in the minority pattern is a strict subset (*loss*) or superset (*gain*) of the accessory region of the majority pattern. In case none of the two conditions was met (e.g. a mixture of loss and gain, an inversion, or a translocation), we classified the event as *other*. We find that the majority of events are gains ($n = 246$, 90%), and only a minority ($n = 17$, 6%) losses (Fig. 4c).

For nonsingleton junctions, we inferred a parsimonious set of gain and loss events to explain the presence/absence pattern of the two alleles across the phylogeny using TreeTime (Sagulenko et al. 2018; Fig. 4a). The inferred events can be on internal or terminal branches. Figure 4b shows the distribution of number of events required to explain the pattern of each nonsingleton junction, along with the number of isolates with the minority pattern. While a random pattern would require a number of events roughly equal to the number of isolates with the minority pattern, we found that all patterns can be explained with just one ($n = 33$) or two ($n = 13$) events, with the exception of a single junction (see supplementary fig. S17, Supplementary Material online for the gainloss pattern of all 14 multiple-event junctions). The latter is due to a tRNA duplication event that appears to have been lost multiple times in clade A; the presence/absence pattern requires 17 independent events. High rates of *pheV* tRNA insertion have been previously described by Touchon et al. (2009), but a different tRNA is involved here (*Tyr*-tRNA). As before the great majority of events are gains, but losses are a 4-fold or 2-fold higher proportion for nonsingleton than for singletons depending on whether or not the tRNA duplication event is counted, respectively. With the exception of this tRNA duplication event, the majority of nonsingleton junctions are compatible with one (and sometimes two) events on the core genome tree, confirming that clonal frame is mostly intact and that masking of events through second order events (e.g. gains followed by immediate loss) is rare.

To gain insight into the dominant mobilization mechanisms, we tried to associate gain or loss events with particular MGEs using the annotations provided by the tools discussed above. When more than one signature MGE is present, we associated a junction with one category using assignment hierarchy: integrons > prophages > defense systems > ISs. This prioritizes larger or rarer elements. We found that the great majority of gains were associated with IS elements, followed by prophages (Fig. 4c). While most IS elements are present in a single copy on the genome, the majority of IS elements associated to these recent gain events are IS1 family elements present in multiple copies in the genome, suggesting that such gains are due to activity of elements that preexist in the affected genome, see supplementary fig. S15, Supplementary Material online. This is also confirmed by gene-based inference of gains and losses, which indicates that IS1-family elements are the gene cluster associated with the highest number of gain events, see supplementary section S11, Supplementary Material online.

Having mapped all the events on the phylogeny where a junction became structurally polymorphic with two alleles, we can estimate the rate of such events in comparison to the number of core genome SNPs on the tree like we did above for modifications of core genome synteny. We estimate these rates separately for terminal and internal tree branches, which account for 80% and 20% of the total branch length, respectively. Rates on terminal branches can be estimated either using junctions with *singleton* variation or including events from *nonsingleton* junctions inferred to have happened on terminal branches. The two estimates are very similar, especially if the tRNA duplication event is discarded, and indicate that a gain event is detected every roughly 20 core genome substitutions, while a loss event roughly every 300 (see Fig. 4d). For internal branches, gain rates are slightly lower (one event per 35 core substitutions) and loss rates are slightly higher (one event per 150 core substitutions). We further confirmed that there is a noisy but clear correlation between the length a branch and the number of events mapped to it (supplementary fig. S16, Supplementary Material online).

Our analysis reveals that ISs are responsible for most of the binary structural genomic diversity, and mostly in the form of insertions present in a minority of isolates. These insertions might have a fitness effect by disrupting functional genes. To investigate the patterns of gene disruption, we considered all binary junctions with an IS gain, and checked the gene annotations spanning the break point in genomes that do not contain the IS (see supplementary section S9 and supplementary fig. S18, Supplementary Material online). In 71% of cases the IS interrupted a gene, thus resulting in a loss of a gene. If insertion happened uniformly across the genome, one would expect the fraction of gene-disrupting insertions to be similar to the fraction of 88.4% of the *E. coli* core genome that is coding sequence. We assume that the underrepresentation of gene-disrupting insertions is due to purifying selection. Under the hypothesis that insertions in noncoding regions have small fitness effects, while insertions in coding regions are on average deleterious, the observed excess of insertions in noncoding regions suggests that roughly 60% of gene-disrupting insertions are removed by purifying selection before being observed, and that the rate of gain events before selection is higher than our estimate. Note, however, that insertions in noncoding regions can have fitness effects as well, for example by carrying promoters that can activate expression of downstream genes (Siguier et al. 2014).

In contrast to binary junctions, for junctions with more than two alternative paths it is often impossible to decompose the diversity into individual events. Inspecting the most complex junctions suggests that they have almost certainly exchanged genetic material with a diverse reservoir many times. While we cannot estimate separate rates for gain and loss, we can nevertheless estimate a lower bound for the overall rate of structural diversification. If a junction has n realized paths, there must have been at least $n - 1$ structural modifications to generate this diversity from a single ancestral form. We thus summed $n_i - 1$ for all junctions i and thereby estimate that at least ~2,000 modifications have taken place in the evolutionary history of *E. coli* ST131 ($\sum_i (n_i - 1) = 1,936$, see supplementary fig. S28, Supplementary Material online)—about one for every three core genome substitutions. One sixth of these structural modifications fall into the 320 binary junctions, and half of them into the 20 most complex junctions.

To illustrate the structural diversity found in complex junctions we picked two examples and created a linear representation for their structure and content. Supplementary figures S21 and S22, Supplementary Material online show the junction corresponding to *hotspot 18* in Hochhauser et al. (2023). This hotspot is empty in all but one isolate of clade A, which harbors a prophage. A different prophage is integrated in clade B and C, with nested diversity due to the movement of ISs. Overall this junction features 33 different paths and a total accessory genome length of 107 kbp. Supplementary figures S23 and S24, Supplementary Material online show the junction corresponding to Hochhauser *hotspot 11*. Diversity in clade A is generated mainly by the integration of a large segment containing several ISs and defense systems. In clade B and C structural diversity is generated by the integration of different prophages and the movement of ISs. This junction features 27 different paths and a total pangenome length of 240 kbp. Overall visual inspection of the structural diversity within these regions reveals both large-scale differences, that are broadly compatible with the phylogenetic structure, but also nested small-scale changes generated by the movement of small segments. Note that in both examples, even when empty the hotspot region contains tRNA genes, which are known targets for integrase enzymes (Williams 2002; Bellanger et al. 2014).

Discussion

Analyses in molecular evolution are typically based on SNPs in alignments of homologous sequences. From such data, it is straightforward to define distances and, in case of asexual inheritance, reconstruct phylogenetic trees that approximate the evolutionary history of the population. We have a good understanding of the elementary evolutionary process—mutation—which generates this diversity. Mutation can be described by substitution models that define rates for transitions between its discrete states. The maturity of this field and the associated analysis tools is one of the reason why bacterial genomes are often analyzed in a similar way: identify the core genome and analyze SNPs in this core genome. However, two bacterial genomes that differ only at a few hundred positions in the core genome can differ by hundreds of thousands of bases in the accessory genome, and by focusing on the core genome most of the evolutionary dynamics are missed.

Developing a quantitative description of evolution of the accessory genome, analogous to that of mutations in alignments of homologous sequences, is a hard problem. Instead of mutating between a fixed set of nucleotides, the accessory genome evolves by gain of genetic material of various length from a diverse pool, loss of segments, inversions, translocations, duplications, etc. Furthermore, the rates of these processes vary dramatically and are unknown to many orders of magnitude. Here, we systematically surveyed all structural changes in a population of recently diverged bacteria using pangenome graphs to begin to quantify these rates.

We chose *E. coli* ST131 because its recent divergence means horizontal processes are rare enough that they can mostly be identified as single events on a robust core genome phylogeny. This then allowed us to estimate rates of the processes that change genome content and structure, at least to within an order of magnitude, by comparing the number of events to the length of the core genome tree. We found that core genome synteny changes at a rate of one event for every 250 substitutions in the core genome, with the majority of changes being inversions. Extrapolation of this rate to typical core genome distances of two *E. coli* genomes results in around 100 synteny breaks, which is not observed (see supplementary fig. S10, Supplementary Material online). This is consistent with the notion that microbial genomes are under selection to maintain their rigid organization (Rocha 2008).

Our survey of variation in the accessory genome highlights the heterogeneity of genome structure evolution. We identified 519 loci with accessory genome variation with at least 2,000 structural modifications to the genome, suggesting that one structural change occurs every three core genome substitutions. At around 300 of these loci, we found only two distinct structural variants which allows to decompose the diversity into individual gain and loss events. In these region, we estimated that gain of genetic material (insertion) is about 10 times more frequent than loss. Despite overall low diversity of ST131, we found a multitude of different variants often consisting of multiple MGEs at about 200 other genomic locations. Such locations have been called “hotspots” by other authors, and many we have identified in this ST131 dataset were previously identified as region of high genetic turnover harboring defense systems (Oliveira et al. 2017; Hochhauser et al. 2023), indicating that diversification rates consistently and strongly depend on chromosomal location. This is consistent with the suggestion that the accumulation of defense systems into such “defense islands” arises from the repeated insertion of MGEs at these locations (Rocha and Bikard 2022).

Our results underscore the rapidity of genomic diversification: starting from the common ancestor of *E. coli* ST131 on the order of 100 years ago, large-scale genome modifications have produced more than 500 structurally diverse regions that often harbor tens of different variations in a sample of ~200 genomes. This diversity is likely to increase with the size of the dataset, given that a large fraction of the structural variants we observe are rare.

In contrast to the deletional bias described for bacterial genomes (Mira et al. 2001; Kuo and Ochman 2009), we found that when a modification of genome structure can be confidently decomposed into gains and losses (binary junction), gains of genetic material were far more common than losses. While this insight is limited to the minority of binary junctions, it raises the question why bacterial genomes are stable in size and compact. The excess of gains over losses is mostly attributable to IS or prophage integrations that are likely deleterious, especially when they disrupt functional genes. Even on the short time scale spanned by this dataset, two-thirds of insertions seem to have been removed by purifying selection. On the other hand, gene disruption or the introduction of new promoters by IS elements can also be adaptive, at least transiently (Siguier et al. 2014; Sastre-Dominguez et al. 2024). Such transient adaptive effects could contribute to an overrepresentation of gains on short time scales. In addition to purifying selection, the excess of insertions and their dominance on terminal branches (along with rearrangements) could also, in part, be due to recent bursts of IS activity in these genomes.

It is important to note that many of these gains of gene-disrupting IS elements would likely be counted as gene losses by traditional gene-based pangenome tools (Tonkin-Hill et al. 2023; Marin et al. 2024). Similarly, the nucleotide based pangenome graph interprets local indel variation like SNPs, while a frame shift producing indel often results in “gene loss” in gene-based pipelines (see supplementary fig. S25 and supplementary section S11, Supplementary Material online). Correcting for this often requires integration with a graph-based approach (Tonkin-Hill et al. 2020; Horsfield et al. 2023). Interestingly, a gene-based inference of gains and losses on our dataset also suggests an excess of gain over loss events. This excess is most pronounced for duplicated genes and more moderate for single copy genes. However, the total number of inferred gain and loss events is much higher than the corresponding estimate for the total number of structural changes from the pangenome graph. This discrepancy is likely due to the fact that one structural change can move several genes at once, see supplementary section S11, Supplementary Material online, which underscores the importance of analyzing the dynamics of genome structure using genome graph rather than relying solely on gene-based approaches.

Among the limitations of this work is the reliance on assembly accuracy. Assembly errors will introduce spurious structural diversity that is difficult to control for computationally. It is thus possible that some technical artifacts are interpreted as structural variation, leading to inflated rate estimates. By restricting the dataset to high-quality reference sequences we sought to minimize this problem. The concordance of the structurally diverse junctions we identified with previously identified hotspots in broader collections of *E. coli* suggests spurious diversity is a minor factor.

Another possible limitation is that our conclusions are based on *E. coli* ST131, a strain associated with hospital outbreaks and antimicrobial resistance. It is possible that ST131 represents a special case and that our findings do not generalize across all bacterial genomes or even all *E. coli*. A natural extension of this work would be to apply the same analysis to other datasets to verify how general our findings are.

A more systematic analysis of the rapid diversification of “hotspots” would be another natural extension of this work. For example, one could quantify whether the observed diversity of these regions is compatible with nested gain/loss/

translocation events on the phylogeny, or whether wholesale exchange with broader *E. coli* diversity is dominating over local diversification. Moreover, the analysis we conducted is well-suited to be extended to any recently diverged microbial clade for which a reliable recombination-free phylogeny can be inferred. It will be interesting to see whether patterns of genome evolution differ qualitatively between different bacterial species.

We have shown that in our dataset structural modifications of bacterial genomes happen at a rate comparable to that of individual nucleotide substitutions anywhere in the genome. This rapid differentiation in genome content as a function of divergence was observed across more diverse *E. coli* collections (Touchon et al. 2020; Shaw et al. 2021). This ST131 dataset allowed us to explore this rapid differentiation regime. Pairs of genomes have a core genome divergence of $\sim 10^{-4}$ on the clonal frame, but already share only roughly 90% of their genes (i.e. have 500 kbp of private sequence per isolate, see supplementary fig. S6, Supplementary Material online). Extrapolating this dynamics to longer time scales is an extremely interesting and challenging problem. Rates of accessory genome turn-over and structural genome evolution are probably very heterogeneous, such that some loci have close to saturated diversity in ST131, while other loci are under strong purifying selection and diversify on much longer scales. Systematically exploring this diversification at different levels of divergence and across different species will be crucial for a quantitative understanding of genomic and phenotypic diversification of bacteria.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors are grateful for stimulating discussions with Erik van Nimwegen and Boris Shraiman and thank Eduardo Rocha and Zamin Iqbal for their helpful feedback on previous versions of this manuscript.

Funding

This work was supported by the University of Basel (M.M. and R.A.N.). L.P.S. is a Sir Henry Wellcome Postdoctoral Fellow funded by Wellcome (grant 220422/Z/20/Z).

Data Availability

No new data were generated as part of this study. All analyzed genomes are publicly available, and a list of their accession numbers can be found at https://github.com/mmolari/ecoliST131-structural-evo/blob/main/config/datasets/ST131_ABC/chromosomes.txt.

References

Arnold BJ, Huang I-T, Hanage WP. Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev Microbiol*. 2022:20(4):206–218. https://doi.org/10.1038/s41579-021-00650-4.

Bellanger X, Payot S, Leblond-Bourget N, Guédon G. Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. *FEMS Microbiol Rev*. 2014:38(4):720–760. https://doi.org/10.1111/1574-6976.12058.

Ben Zakour NL, Alsheikh-Hussain AS, Ashcroft MM, Khanh Nhu NT, Roberts LW, Stanton-Cook M, Schembri MA, Beatson SA. Sequential acquisition of virulence and fluoroquinolone resistance has shaped the evolution of *Escherichia coli* ST131. *mBio*. 2016:7(2):e00347-16. https://doi.org/10.1128/mBio.00347-16.

Biggel M, Hoehn S, Frei A, Dassler K, Jans C, Stephan R. Dissemination of ESBL-producing *E. coli* ST131 through wastewater and environmental water in Switzerland. *Environ Pollut*. 2023:337:122476. https://doi.org/10.1016/j.envpol.2023.122476.

Brockhurst MA, Harrison E, Hall JPJ, Richards T, McNally A, MacLean C. The ecology and evolution of pangenomes. *Curr Biol*. 2019:29(20):R1094–R1103. https://doi.org/10.1016/j.cub.2019.08.012.

Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, Chain PSG, Nayfach S, Kyrpides NC. Identification of mobile genetic elements with geNomad. *Nat Biotechnol*. 2024:42(8):1303–1312. https://doi.org/10.1038/s41587-023-01953-y.

Colquhoun RM, Hall MB, Lima L, Roberts LW, Malone KM, Hunt M, Letcher B, Hawkey J, George S, Pankhurst L, *et al.* Pandora: nucleotide-resolution bacterial pan-genomics with reference graphs. *Genome Biol*. 2021:22(1):1–30. https://doi.org/10.1186/s13059-021-02473-1.

Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res*. 2015:43(3):e15. https://doi.org/10.1093/nar/gku1196.

Decano AG, Downing T. An *Escherichia coli* ST131 pangenome atlas reveals population structure and evolution across 4,071 isolates. *Sci Rep*. 2019:9(1):1–13. https://doi.org/10.1038/s41598-019-54004-5.

Doolittle WF, Zhaxybayeva O. On the origin of prokaryotic species. *Genome Res*. 2009:19(5):744–756. https://doi.org/10.1101/gr.086645.108.

Eisen JA, Heidelberg JF, White O, Salzberg SL. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol*. 2000:1(6):1–9. https://doi.org/10.1186/gb-2000-1-6-research0011.

Georjon H, Bernheim A. The highly diverse antiphage defence systems of bacteria. *Nat Rev Microbiol*. 2023:21(10):686–700. https://doi.org/10.1038/s41579-023-00934-x.

Gladstone RA, McNally A, Pöntinen AK, Tonkin-Hill G, Lees JA, Skytén K, Cléon F, Christensen MOK, Haldorsen BC, Bye KK, *et al.* Emergence and dissemination of antimicrobial resistance in *Escherichia coli* causing bloodstream infections in Norway in 2002–17: a nationwide, longitudinal, microbial population genomic study. *Lancet Microbe*. 2021:2(7):e331–e341. https://doi.org/10.1016/S2666-5247(21)00031-8.

Haudiquet M, de Sousa JM, Touchon M, Rocha EPC. Selfish, promiscuous and sometimes useful: how mobile genetic elements drive horizontal gene transfer in microbial populations. *Philos Trans R Soc B*. 2022:377(1861):20210234. https://doi.org/10.1098/rstb.2021.0234.

Hochhauser D, Millman A, Sorek R. The defense island repertoire of the *Escherichia coli* pan-genome. *PLoS Genet*. 2023:19(4):e1010694. https://doi.org/10.1371/journal.pgen.1010694.

Horsfield ST, Tonkin-Hill G, Croucher NJ, Lees JA. Accurate and fast graph-based pangenome annotation and clustering with ggCaller. *Genome Res*. 2023:33(9):1622–1637. https://doi.org/10.1101/gr.277733.123.

Jaén-Luchoro D, Kahnamouei A, Yazdanshenas S, Lindblom A, Samuelsson E, Åhrén C, Karami N. Comparative genomic analysis of ST131 subclade C2 of ESBL-producing *E. coli* isolates from patients with recurrent and sporadic urinary tract infections. *Microorganisms*. 2023:11(7):1622. https://doi.org/10.3390/microorganisms11071622.

Kallonen T, Brodrick HJ, Harris SR, Corander J, Brown NM, Martin V, Peacock SJ, Parkhill J. Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Res*. 2017:27(8):1437–1449. https://doi.org/10.1101/gr.216606.116.

Kuo C-H, Ochman H. Deletional bias across the three domains of life. *Genome Biol Evol*. 2009:1:145–152. https://doi.org/10.1093/gbe/evp016.

Ludden C, Decano AG, Jamrozy D, Pickard D, Morris D, Parkhill J, Peacock SJ, Cormican M, Downing T. Genomic surveillance of *Escherichia coli* ST131 identifies local expansion and serial replacement of subclones. *Microb Genom*. 2020:6(4):e000352. https://doi.org/10.1099/mgen.0.000352.

Marin MG, Wippel C, Quinones-Olvera N, Behruznia M, Jeffrey BM, Harris M, Mann BC, Rosenthal A, Jacobson KR, Warren RM, *et al*. Analysis of the limited *M. tuberculosis* accessory genome reveals potential pitfalls of pan-genome analysis approaches. bioRxiv 586149. https://doi.org/10.1101/2024.03.21.586149, March 2024, preprint: not peer reviewed.

Mayo-Muñoz D, Pinilla-Redondo R, Birkholz N, Fineran PC. A host of armor: prokaryotic immune strategies against mobile genetic elements. *Cell Rep*. 2023:42(7):112672. https://doi.org/10.1016/j.celrep.2023.112672.

Mira A, Ochman H, Moran NA. Deletional bias and the evolution of bacterial genomes. *Trends Genet*. 2001:17(10):589–596. https://doi.org/10.1016/S0168-9525(01)02447-7.

Néron B, Littner E, Haudiquet M, Perrin A, Cury J, Rocha EPC. IntegronFinder 2.0: identification and analysis of integrons across bacteria, with a focus on antibiotic resistance in Klebsiella. *Microorganisms*. 2022:10(4):700. https://doi.org/10.3390/microorganisms10040700.

Noll N, Molari M, Shaw LP, Neher RA. PanGraph: scalable bacterial pan-genome graph construction. *Microb Genom*. 2023:9(6):001034. https://doi.org/10.1099/mgen.0.001034.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, *et al*. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016:44(D1):D733–D745. https://doi.org/10.1093/nar/gkv1189.

Oliveira PH, Touchon M, Cury J, Rocha EPC. The chromosomal organization of horizontal gene transfer in bacteria. *Nat Commun*. 2017:8(1):841. https://doi.org/10.1038/s41467-017-00808-w.

Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol*. 2016:17(1):1–14. https://doi.org/10.1186/s13059-016-0997-x.

Partridge SR, Kwong SM, Firth N, Jensen SO. Mobile genetic elements associated with antimicrobial resistance. *Clin Microbiol Rev*. 2018:31(4):10–1128. https://doi.org/10.1128/CMR.00088-17.

Pitout JDD, Finn TJ. The evolutionary puzzle of *Escherichia coli* ST131. *Infect Genet Evol*. 2020:81:104265. https://doi.org/10.1016/j.meegid.2020.104265.

Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*. 2010:5(3):e9490. https://doi.org/10.1371/journal.pone.0009490.

Rocha EPC. Inference and analysis of the relative stability of bacterial chromosomes. *Mol Biol Evol*. 2006:23(3):513–522. https://doi.org/10.1093/molbev/msj052.

Rocha EPC. The organization of the bacterial genome. *Annu Rev Genet*. 2008:42(1):211–233. https://doi.org/10.1146/genet.2008.42.issue-1.

Rocha EPC, Bikard D. Microbial defenses against mobile genetic elements and viruses: who defends whom from what? *PLoS Biol*. 2022:20(1):1–18. https://doi.org/10.1371/journal.pbio.3001514.

Sagulenko P, Puller V, Neher RA. Treetime: maximum-likelihood phylodynamic analysis. *Virus Evol*. 2018:4(1):vex042. https://doi.org/10.1093/ve/vex042.

Sakoparnig T, Field C, van Nimwegen E. Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. *Elife*. 2021:10:e65366. https://doi.org/10.7554/eLife.65366.

Sastre-Dominguez J, DelaFuente J, Toribio-Celestino L, Herencias C, Herrador-Gomez P, Costas C, Hernandez-Garcia M, Canton R, Rodriguez-Beltran J, Santos-Lopez A, *et al*. Plasmid-encoded insertion sequences promote rapid adaptation in clinical enterobacteria. *Nat Ecol Evol*. 2024:8(11):2097–2112. https://doi.org/10.1038/s41559-024-02523-4.

Shaw LP, Chau KK, Kavanagh J, AbuOun M, Stubberfield E, Gweon HS, Barker L, Rodger G, Bowes MJ, Hubbard ATM, *et al*. Niche and local geography shape the pangenome of wastewater-and livestock-associated Enterobacteriaceae. *Sci Adv*. 2021:7(15):eabe3868. https://doi.org/10.1126/sciadv.abe3868.

Siguier P, Gourbeyre E, Chandler M. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol Rev*. 2014:38(5):865–891. https://doi.org/10.1111/1574-6976.12067.

Stoesser N, Sheppard AE, Pankhurst L, De Maio N, Moore CE, Sebra R, Turner P, Anson LW, Kasarskis A, Batty EM, *et al*. Evolutionary history of the global emergence of the *Escherichia coli* epidemic clone ST131. *mBio*. 2016:7(2):e02162. https://doi.org/10.1128/mBio.02162-15.

Tesson F, Hervé A, Mordret E, Touchon M, D'humières C, Cury J, Bernheim A. Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat Commun*. 2022:13(1):2561. https://doi.org/10.1038/s41467-022-30269-9.

Tokuda M, Shintani M. Microbial evolution through horizontal gene transfer by mobile genetic elements. *Microb Biotechnol*. 2024:17(1):e14408. https://doi.org/10.1111/1751-7915.14408.

Tonkin-Hill G, Corander J, Parkhill J. Challenges in prokaryote pangenomics. *Microb Genom*. 2023:9(5):001021. https://doi.org/10.1099/mgen.0.001021.

Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, Gladstone RA, Lo S, Beaudoin C, Floto RA, *et al*. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol*. 2020:21(1):1–21. https://doi.org/10.1186/s13059-020-02090-4.

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, *et al*. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet*. 2009:5(1):e1000344. https://doi.org/10.1371/journal.pgen.1000344.

Touchon M, Perrin A, De Sousa JAM, Vangchhia B, Burn S, O'Brien CL, Denamur E, Gordon D, Rocha EPC. Phylogenetic background and habitat drive the genetic diversification of *Escherichia coli*. *PLoS Genet*. 2020:16(6):e1008866. https://doi.org/10.1371/journal.pgen.1008866.

Von Wintersdorff CJH, Penders J, Van Niekerk JM, Mills ND, Majumder S, Van Alphen LB, Savelkoul PHM, Wolffs PFG. Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front Microbiol*. 2016:7:173. https://doi.org/10.3389/fmicb.2016.00173.

Williams KP. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res*. 2002:30(4):866–875. https://doi.org/10.1093/nar/30.4.866.

Xie Z, Tang H. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics*. 2017:33(21):3340–3347. https://doi.org/10.1093/bioinformatics/btx433.