Escherichia coli plasmidome maps the game of clones

Sergio Arredondo-Alonso[1]‡, Anna K. Pöntinen[1,2]‡, João Alves Gama[3]‡, Rebecca A. Gladstone[1]‡, Klaus Harms[3], Gerry Tonkin-Hill[1], Harry A. Thorpe[1], Gunnar S. Simonsen[2,4], Norwegian *E. coli* BSI Study Group[#], Ørjan Samuelsen[2,3]†, Pål J. Johnsen[3]†, Jukka Corander[1,5,6]†*

‡Equal contributions
†Equal contributions

Corresponding authors: Sergio Arredondo-Alonso, Jukka Corander*

[#]Collaborators: Nina Handal, Nils Olav Hermansen, Anita Kanestrøm, Hege Elisabeth Larsen, Paul Christoffer Lindemann, Iren Høyland Löhr, Åshild Marvik, Einar Nilsen, Marcela Pino, Elisabeth Sirnes, Ståle Tofteland, Kyriakos Zaragkoulias.

[1]Department of Biostatistics, University of Oslo, Oslo, Norway

[2]Norwegian National Advisory Unit on Detection of Antimicrobial Resistance, Department of Microbiology and Infection Control, University Hospital of North Norway, Tromsø, Norway

[3]Department of Pharmacy, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway

[4]Department of Medical Biology, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway

[5]Parasites and Microbes, Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

[6]Helsinki Institute for Information Technology HIIT, Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

Abstract

Escherichia coli is a major human pathogen and the most widely studied microbe in history, but its extrachromosomal elements known as plasmids remain poorly delineated. Here we used long-read technology to high-resolution sequence the entire plasmidome and the corresponding host chromosomes from a longitudinal survey covering two decades and over 2,000 *E. coli* isolates. Separation of chromosomal and extrachromosomal DNA

enabled us to reconstruct co-evolutionary trajectories of host lineages and their plasmids on a population-wide scale, demonstrating that plasmid evolution is markedly constrained contrary to the established dogma. We find that some plasmids have persisted in lineages for centuries, and provide a high-resolution map of recent vertical and horizontal evolutionary events in plasmids with key antibiotic resistance, competition and virulence determinants. We present genomic evidence of both chromosomal and plasmid-driven success strategies that represent convergent phenotypic evolution in distant lineages, and use *in vitro* experiments to verify the importance of bacteriocin-producing plasmids for clone success. Our study has general implications for understanding plasmid biology and bacterial evolutionary strategies, and informing development of future interventions against pathogens.

Introduction

The history of *Escherichia coli* as a defined bacterial species is rich with fundamental microbiological discoveries lining the path from the late 19th century when it was first isolated and named, until today, where it has a ubiquitous role in bioengineering applications and still occupies a center stage in public health burden globally. The extra-intestinal pathogenic *E. coli* (ExPEC) have been of particular interest to both eco-evolutionary and epidemiological research communities since they universally colonize the healthy human gut and play the roles of both Dr. Jekyll and Mr. Hyde inside us (*1*). The ExPEC population is composed of a high diversity of clones, but a discrete set of lineages contribute to the majority of infections (*2*), and the problem is exacerbated by the emergence and increase of antibiotic resistance. Epidemiological investigations have uncovered an intriguing feature of ExPEC, showing a stable maintenance of pandemic lineages over longer timescales and the emergence of novel successful clones in clinical surveillance, many with foodborne associations (*2*–*5*). The new clones typically expand rapidly initially, but are then constrained within only a few years such that the population transits to another equilibrium still maintaining a considerable diversity of lineages. Unbiased longitudinal genomic surveys have demonstrated that the transient disruption of the equilibrium bears the hallmarks of negative frequency-dependent selection (NFDS) (*6*–*8*), and further modeling work suggested this selection is acting on accessory genomic elements rather than on core genome variation (*9*).

Plasmids, initially described as transmissible agents of antibiotic resistance and competitor-inhibiting determinants in *E. coli* in the 1950s (*10–13*), play a key role in accessory genome dynamics. Despite the strong evidence that plasmids have played a critical role in the emergence and dissemination of successful clones (*14–17*), the structure of the plasmidome, and the horizontal and vertical evolutionary dynamics of plasmids and ExPEC hosts remain largely unknown. This is due to paucity of scalable computational approaches to type complete plasmid sequences, and the lack of large-scale longitudinal long-read sequencing studies has hindered developing a firm understanding of the plasmidome structure and prevalence of particular plasmids. This motivated us to employ long-read sequencing technology on a large representative longitudinal collection of 2,045 *E. coli* isolates (*7*) to elucidate the entire plasmidome of ExPEC and to uncover the evolutionary and epidemiological trends associated with it. This collection comprises isolates included regardless of their clonal background, antimicrobial resistance profile or any other bacterial genotypic or phenotypic characteristic, making it ideal for the study of evolution, expansion and persistence of the plasmidome.

Our approaches resulted in a total of 4,485 circularized plasmid sequences and through inferred evolutionary rates and dated phylogenies we estimated the timing of plasmid emergence/acquisition into major ExPEC clones. We demonstrate multiple cases of plasmid-driven convergent evolution in distant genetic backgrounds, supporting the conclusion that plasmids play key roles in initial clonal establishment as well as later endemic maintenance in the host population. We further built a map of bacteriocin variation, proposed to play a role in shaping bacterial population structures (*12*, *13*) and through combined population genomics and experimental approaches identified the plasmid-encoded microcin V as a contributor to clonal success and maintenance of stable population equilibrium.

Results

Resolving the plasmidome

We inferred the plasmidome from a long-read sequenced collection of >2,000 isolates representing 226 different sequence types (STs) where the four most common pandemic clones (ST69, ST73, ST95 and ST131) had caused nearly half of the bloodstream infections (1,438/3,254, 44.19%) (Fig. 1A, Fig. S1A, Methods). A hybrid assembly approach resulted in highly contiguous assemblies for 1,999 genomes with median N50 chromosome length 4.98 Mbp and an average L50 chromosome count 1.09 (Fig. S1B). Chromosome lengths differed significantly between the major clones (Fig. S1C, p-value < 0.05), where ST69 and ST73 showed a larger chromosome size compared to ST95 and ST131. A total of 5,417 contigs were identified as plasmid-derived, collectively here referred to as the plasmidome. The total length and the number of plasmid sequences observed per isolate was significantly lower for ST73 (p-value < 0.05) compared to the other major clones (Fig. S1D). This finding confirms previous reports showing that ST73 frequently carries lower plasmid load, demonstrated by lower conjugation frequencies and higher fitness costs (*18*, *19*).

To type the circular plasmid sequences (n = 4,485) identified from the hybrid assemblies, we applied our recently developed method, mge-cluster (*20*). Of the 4,485 plasmids, mge-cluster assigned the majority (3,734; 83.26%) into 23 non-overlapping groups, referred to as plasmid types (pTs) (Fig. 1B). Of the remaining plasmids, 413 (9.21%) were left unassigned to a type (Fig. 1B), and 338 sequences (7.54%) were filtered out after quality control (Methods). The non-circular plasmid sequences (n = 932) were then assigned to the defined types resulting in 394 additional sequences assigned to the 23 pTs (Table S1). When possible, types were annotated based on existing plasmid labeling schemes (summarized in Table S2) and further contextualized with reference plasmids.

Strikingly, all except one of the 13 plasmid types with larger sizes discovered in total (average length >70 kb) were associated with either antimicrobial resistance (AMR), competition or virulence genes, highlighting the definite role of large plasmids as vehicles of important traits for the host cell (Table S2). Comparison of the plasmidome structure (Fig. 1B) with the published plasmid sequences of *E. coli* as delineated recently (*20*) revealed a

very high degree of congruence (adjusted Rand Index 0.94). Furthermore, the only major discrepancy between the two arose from a subdivision of pT11 and pT14 which were merged in the analysis of published plasmids, apparently due to lack of resolution caused by the limited number of sequences available. Taken together, these results suggest that plasmid evolution is far more constrained than previously thought, and that there are only a limited number of successful plasmid backbones which can serve as stable vehicles for the dissemination of advantageous genes.

Plasmid types corresponding to long (>70 kb) plasmid sequences tended to display high degrees of gene sharing (Fig. S2) as measured by the containment index (*21*), while some smaller (<10 kb) plasmid types only shared the genes involved in the replication machinery (Fig. S2). To investigate gene sharing between plasmid types in detail, we performed a gene synteny analysis (minimum identity 0.8) among representative sequences of most types with large gene content (Figure 2). A subset of plasmid types, including pT6, pT14 and pT21, had large genomic blocks in common, corresponding to siderophores, genes required for plasmid conjugation, or AMR/virulence gene cassettes. By increasing the stringency to create a link between genes in synteny (minimum identity 0.99), we observed that the IncF plasmid transfer region diversified overall more than other genomic blocks, and that the plasmids in pairs pT6/pT14 and pT10/pT11 most likely evolved from the same ancestral plasmid within each pair, since the transfer region remains highly conserved (Fig. S3). We confirmed this by reconstructing a recombination-free phylogeny using the aligned plasmid sequences from each pair of plasmid types (Fig. S4), which showed that pT6 arose from pT14, and pT10 correspondingly from pT11.

Some of the plasmid types (pT17, pT18, pT19, pT20) with smaller average sizes often carried bacteriocin genes and were found in a wide diversity of genetic backgrounds (Fig. 1B, Fig. S5). Notably, none of these genes/plasmids showed evidence of clonal expansion in contrast with the larger plasmids discussed in detail below.

Evolutionary history and acquisition timeline of central plasmids

The typing analysis permitted us to investigate whether particular plasmid types were highly prevalent and potentially contributed to the success of the four major clones. To understand

when these plasmids were acquired, we inferred the evolutionary rates and dated the phylogenies of the four most common STs using BactDating (*22*), and further detected probable expansions using CaveDive (*23*). We focused on medium to large sized plasmids (>10 kbp, Table S2), since those plasmids tend to encode features that can enhance the success of their bacterial hosts.

ST95 showed a high prevalence of pT14 (38%), pT3 (24%) and pT11 (22%), of which the latter two are predominantly non-overlapping in the phylogeny (Fig. 3A). pT3 corresponds to a conjugative IncB/O/K/Z plasmid belonging to the I-plasmid complex (*24*), often encoding for aminoglycoside (*aph(6)-Id*, *aph(3'')-Ib*), sulfonamide (*sul2)* and beta-lactam (bla_{TEM-1}) resistance genes (Table 1), which is frequently reported in other bacterial species from *Enterobacteriaceae* (*24*). pT14 corresponds to a conjugative IncF plasmid (F24:A-:B1) characterized by: i) the presence of the *cvac* gene encoding microcin colV, together with another colicin (colicin Ia, *cia* gene) (*25*, *26*), ii) multiple siderophore systems involved in iron acquisition including *iroBCDEN* (salmochelin operon), *iucABCD, iutA* (aerobactin operon), and the iron-transport system *sitABCD* (*27*), iii) the outer membrane protein T-encoding gene *ompT* and the hemolysin-encoding gene *hlyF*, iv) the ABC transport system *etsABC*, and v) the *iss* gene conferring resistance to the complement system (*28*). pT14 was contextualized as a pcolV-like plasmid (reference plasmid pECOS88). As extensively reviewed by Johnson et. al (*29*), pcolV-like plasmids have been associated with urinary tract infection (UTI) and meningitis, and linked to iron acquisition mechanisms, serum resistance, growth in human urine or bacteriophage resistance among other phenotypes. Both pT3 and pT14 were widely distributed across clades in ST95 suggesting they were introduced at the most recent common ancestor (MRCA) of ST95, dating back to 1768 (95% CI, 1721–1806) (Fig. 3A, Table S3). The major branch (n=436/442 isolates) leading from the root was detected as an expansion event estimated to have occurred in 1823 (95% CI, 1789-1851).

The phylogeny of ST95 has been previously split into five distinct clades that are correlated with the strain serotype (O-H typing and *fimH* allele combination) (*30*). The clade of ST95 represented by *fimH41*:O1/O2:H7 (Fig. 3A) carries a distinct IncF plasmid corresponding to pT11, which was likely acquired around 1884 (95% CI 1817-1937) and displaced pT14 (Table S3, Fig. 3A). This is a conjugative plasmid (F29:A-:B10) that possesses the *senB* gene cluster contributing to uropathogenesis. It has been previously termed as pUTI89-like

and closely resembles the reference plasmids pUTI89 and pRS218 (*31*, *32*). pUTI89-like plasmids have been shown to increase the fitness of *E. coli* during the acute stages of infection in a murine UTI infection model enhancing the binding, invasion of bladder epithelial cells and providing immune evasion properties (*31*). In addition, this plasmid has been proven to play an important role in the pathogenesis of *E. coli* causing neonatal meningitis using both *in vitro* and *in vivo* models (*32*).

The basal part of the ST69 phylogeny represents isolates notably void of larger plasmid carriage, suggesting that such plasmids did not play a central role in the early evolution of this lineage (Fig. 3B). However, one of the ST69 clades (*fimH27*:017/015:H18/H4) acquired pT10 around 1989-1990 (95% CI, 1986-1992) (Table S3, Fig. 3B); this acquisition was nested within a major expansion event occurring two nodes earlier in the tree (1987, 95% CI, 1984-1990). pT10, arose from pT11 (Fig. S4), by addition of an AMR cassette providing resistance up to four distinct antimicrobial groups and to heavy-metals (mercury) (Fig. 2). These are both pUTI89-like plasmids, including the reference plasmid p1ESCUM, linking this subpopulation of ST69 with an extended ability to cause UTI, similar to the subpopulation of ST95 discussed above in the context of plasmids belonging to pT11.

Consistent with its generally smaller inferred plasmidome size, ST73 only carries large plasmid types in a minority of isolates (181/560, 32%). Out of these, the dominant pT4 (prevalence 18%, F51:A-:B10) (Fig. S6) has been introduced multiple times into ST73. An expansion event in 1983 (95% CI, 1977-1989) occurred in one introduction of pT4 after the plasmid had acquired bla_{SHV-1}, displaying increased ampicillin resistance, around 80 years (1900, 95% CI 1877-1919) after the initial pT4 acquisition into the *fimH10:*O6:H1 clade (Table S3, Fig. 3C). This plasmid type shared a similar gene synteny and content with pT10 (pUTI89-like) (Fig. 2, Fig. S2), encoding the same virulence traits (*senB* gene cluster) and thus most likely contributing to a heightened uropathogenicity of the isolates. pT4 also included several AMR genes including *aadA1* (aminoglycoside), *sul1* (sulfonamide) and in some cases bla_{SHV-1} (beta-lactam). Despite the general trend of susceptibility for this clone, this demonstrates that particular subclades of ST73 can become multi-drug resistant (MDR) by incorporating specific plasmid types.

For ST131, three major clades have been widely described in the literature; clade ST131-A (*fimH41*:O16:H5), clade ST131-B (*fimH22*:025:H4) and clade ST131-C (*fimH30*:025:H4) which is further subdivided into clades C1 (also termed as H30-R) and C2 (H30-Rx) (Fig. 3D). In agreement with previous studies based on short-reads (*15*, *33*), for each clade we observed clear associations with specific plasmid types (Fig. 3D, Fig. S6).

ST131-A showed a high prevalence (49%) of pT10 (F29:A-:B10, pUTI89-like) and pT13 (prevalence 22%) (Fig. S6). As shown previously, pT10 was associated with UTI phenotype and an expansion of a ST69 subpopulation, which is recapitulated in the data for ST131-A. The distribution of these plasmids suggests that pT13 was present at the MRCA of ST131-A (1984, 95% CI 1978-1989) and that it was later replaced by pT10 (pUTI89-like plasmid) around 1988 (95% CI, 1980-1993) (Table S3, Fig. 3D). The branch corresponding to the replacement of pT13 with pT10 was indicated as a clonal expansion event estimated to have occurred around 1992 (95% CI, 1988-1996).

ST131-B followed a similar plasmid distribution as observed for ST95. One of the sublineages carried the pT6 (F2:A-:B58/B1) which arose from pT14 (Fig. S6). pT6 corresponds to a pcolV-like plasmid characterized by the presence of two colicin genes (microcin V, *cvaC* gene*;* colicin Ia, *cia* gene), multiple siderophore systems (salmochelin, aerobactin) and multiple virulence genes (*iss, hlyF, ompT*). In addition, pT6 encodes multiple AMR genes, which is a feature that distinguishes it from pT14. We estimated that this plasmid was introduced in ST131-B around 1955 (95% CI, 1944-1963) (Table S3, Fig. 3D). The other sublineage of ST131-B harbored pT11 (F29:A-:B10) (prevalence 45%) which was also found in ST95, and was acquired between 1957 and 1981 (95% CI, 1946-1986) (Table S3).

pT12 was present in ST131-C2 (F36:A4:B1/B58) (prevalence 35%) (Fig. S6) and was also present in its basal subclade (C0) (Fig. 3D). This plasmid type contained an AMR cassette encoding for tetracycline resistance (*tetA*) and in some cases an additional cassette harboring *dfrA17* (trimethoprim), *aadA5* (aminoglycoside), *sul1* (sulfonamide), and *mph(A)* (macrolide) (Fig. 2). Notably, this plasmid never harbored bacteriocin producing genes unlike pT6 and pT14 which were found in ST131-B. Plasmid sequences labeled as F2:A1:B- were also present in ST131-C2 at high prevalence (39%) and included the CTX-

M15 allele. These plasmid sequences were closely related (Fig. 1) but not assigned a type as they were not present in the minimum number of samples required to define a cluster. The reference plasmid pJJ1886_5 was embedded by the mge-cluster algorithm in this grouping. These sequences were present in the ancestral lineage of ST131-B (clade B0) as F2:A1:B63, indicating that the IncFIB replicon was lost later in ST131-C2 (Fig. 3D). In contrast, ST131-C1 showed a high prevalence (90%) of pT13 (F1:A2:B20) (Fig. S6) which is associated with the reference plasmid pG150. Its distribution suggests that this plasmid was present in the MRCA of ST131-C1 (Table S3). This plasmid encodes for the *senB* gene cluster, and frequently harbors multiple non-fixed AMR genes, including $bla_{\text{TEM-1}}$ and $bla_{\text{CTX-M-27}}$ (beta-lactam), or *aac(3)-IId* (aminoglycoside) among others.

The role of plasmids in bacterial competition

The majority of the detected large plasmid types encode beneficial traits such as AMR and bacterial competitiveness. Because AMR does not seem to be the major determinant for clonal success (*7*), we decided to investigate the role of plasmid-encoded bacteriocins in bacterial competition to shed further light on the NFDS dynamics (*9*). We screened genes coding for bacteriocins and found that they are most often present in plasmids (Fig. S5, Fig. S7), the exception was *mchB* encoding microcin H47 typically found in the chromosome of ST73 isolates (Fig. S6). The most frequently found plasmid-encoded bacteriocins were *cvaC* (microcin V) and colicin Ia (also termed as pECS88_04) (Fig. S5), typically co-occurring in pT6 and pT14 (Fig. 2, Fig. S7) that are associated with ST131-B and ST95, respectively. To elucidate the role of microcin V and colicin Ia in clone competition, we examined their activity experimentally.

The microcin V gene cluster is composed of four genes (Fig. 4A): *cvaC* encodes the toxin, the immunity gene *cvi* encodes a peptide that protects the toxin-producing cells, and the genes *cvaA* and *cvaB* that encode a transport system to secret the toxin (*13*). Production of microcin V is induced by iron limitation, and in target cells it is recognized by the Cir siderophore receptor (*13*). When grown in iron-limiting conditions, the culture supernatants of three ST95 and one ST131-B plasmid-harboring isolates strongly inhibited the growth of the *E. coli* lab strain MG1655 (Fig. 4B). The killing effect was dose-dependent which is

consistent with bacteriocin activity, while the absence of visible plaques ruled out bacteriophage-mediated killing. In the absence of induction, the ST95 isolate 27-61 (Table S4) produced a weaker inhibition zone, possibly due to some basal expression of one of the two bacteriocins. As a further control, the ST131-B isolate 27-56 that does not harbor a bacteriocinogenic plasmid was unable to produce inhibition despite growing under iron-limiting conditions (Fig. 4B). The strains 31-16 and 31-17 (Fig. 4) served as further controls (see Methods). Having proven that the microcin V is functional, we tested its range of activity against 51 *E. coli* isolates belonging to the main four disease-associated STs, and ST10 as a commensal model. As expected, the four bacteriocin-producing isolates mentioned above were insensitive to any of the corresponding four batches of microcin V, while 39 of the remaining 47 isolates were sensitive (Fig. S8, Table S4). Sensitive isolates were generally inhibited by all four batches, except in a few cases where inhibition was only observed for the more concentrated bacteriocin batches (produced by isolates 23-46 and 27-61). Overall, ST69 and ST131 isolates were sensitive, while isolates belonging to the other STs display a more heterogeneous behavior. The supernatants of uninduced 27-61 and induced 27-56 (lacking colicinogenic plasmid) did not manifest inhibitory effects on the clinical isolates, as expected. Altogether, these results show that pTs 6 and 14 encode a functional microcin V gene cluster whose product inhibits the growth of a wide range of *E. coli* isolates from different genetic backgrounds.

The colicin Ia gene cluster is composed of two genes (Fig. 4A): *ciaA* (termed as that encodes the toxin, and *iia* that encodes immunity. This colicin is expressed upon SOS induction (a response to genotoxic stress), and its receptor is also Cir (*26*). The same four plasmid-harboring isolates used above were exposed to UV irradiation to induce the SOS response and consequently production of colicin Ia. The four culture supernatants inhibited the growth of *E. coli* MG1655 showing a dose-dependent effect and no observable phage plaques (Fig. 4C). As shown for microcin V, the uninduced ST95 isolate 27-61 produces a weaker inhibition zone, while the induced ST131-B isolate 27-56 not harboring a colicinogenic plasmid does not produce inhibition. When the four supernatants were tested against our isolates only six out of 51 displayed sensitivity (Fig. S8, Table S4). Supernatants of isolates 27-56 (induced, no plasmid) and 27-61 (uninduced) were not inhibitory. Thus pTs 6 and 14 encode a functional Ia gene cluster but the bacteriocin has a limited activity compared with microcin V.

Discussion

Plasmids are generally considered as the ultimate vehicles of rapid evolution in bacterial genomes, while also sometimes regarded as parasitic elements to their host cells (*10–12*, *34*). Our analysis indicated that larger plasmids in *E. coli* are almost always associated with traits falling under bacterial competition, pathogenesis and antibiotic resistance, thus providing useful services to the host cell rather than representing parasitism. Even many of the small plasmids were found to incorporate likely beneficial bacteriocin-producing systems that were scattered around the population phylogeny, indicating their repeated horizontal acquisition, but lacking evidence of further clonal expansion.

Only two co-occurring plasmid-encoded bacteriocins (microcin V and colicin Ia) were frequent enough and showed evidence of clonal expansion in different genetic backgrounds to indicate a major role in colonization competition and contribution towards maintaining stable population equilibrium. To further examine their role in strain competition, we experimentally confirmed the ability of microcin V to inhibit the majority of frequently observed clones in the population that lack this system. Similarly, only a single dominant chromosomal bacteriocin, the microcin H47, was identified in our collection, and the characterics of these bacteriocins stand in stark contrast with the substantial bacteriocin diversity previously found in *S. pneumoniae* population to contribute to the maintenance of a stable population equilibrium over time (*35*). Of note, two ST10 isolates were found resistant to microcin V, despite lacking this bacteriocin-producing system. This is well aligned with the frequently observed commensal colonization ability of ST10 and its lower virulence potential (*36*) that was recently quantified at the population level by systematically comparing neonatal colonization rates with those observed in bloodstream infections (*37*). Why a subset of ST10 population would remain resistant to microcin V is currently unclear and warrants further investigation into evasion of the effects of these inhibition mechanisms.

The unique characteristics of our dataset allowed detection of multiple events of convergent phenotypic evolution across distant lineages, associated with acquisition and stable maintenance of certain plasmid types over time. In addition to the same plasmid-based bacteriocin-producing systems observed in ST95 and ST131-B discussed above, these two

distant lineages harbor other clades hosting plasmids from pT11, which corresponds to the widely studied reference plasmid pUTI89 encoding virulence traits archetypal for uropathogenesis (*31*). In both major lineages, these clades were non-overlapping with the clades hosting colV type plasmids (pTs 6, 14), which have been found to be strongly associated with avian niche and the avian pathogenic *E. coli* (APEC) strains (*25*, *38*). ST131-B has been previously widely indicated as a foodborne uropathogen, in particular related to consumption of poultry meat (*39*, *40*). Combined, this evidence indicates that both ST95 and ST131-B lineages circulating in human hosts are stably split into human vs. avian adapted subclades assisted by their plasmid-associated traits, and that the latter likely reflects a constant spillover from the avian niche via foodborne transmission. Interestingly, the estimated timing of colV acquisition in ST131-B during 1950-60s coincides with a rapidly intensifying poultry production in many countries after WW2, which may have provided an opportunity for *E. coli* equipped with particular plasmid-derived traits to quickly expand in this niche.

Similar to the previously discussed observations, ST69 and ST131-A display clades defined by a shared plasmid content, which we showed arose from the pUTI89 virulence plasmids and acquired an additional resistance cassette. This exemplifies plasmid-associated convergent phenotypic evolution in two lineages belonging to different phylogroups (D, B2). ST69 was originally detected in an outbreak in California (*41*), and later epidemiological investigation suggested it emerged in the late 1990s and then became endemic globally within 10 years (*42*). Our dating analysis gives the narrow time interval between 1987-1992 for the acquisition of the plasmid which ultimately led to a global dissemination of this clone with the specific virulence and AMR characteristics. Further, as shown by our results, this evolutionary process closely reflects the parallel acquisition of the same plasmid by ST131-A subclade, which is again associated with remarkably similar epidemiological characteristics with a rapid expansion (*7*, *8*) and global endemicity over time. These two evolutionary trajectories across the successful lineages of *E. coli* bear the same hallmarks as the convergence of hypervirulence and multi-drug resistance in *Klebsiella pneumoniae*, which were first identified as non-overlapping traits (*43*), but later studies pinpointed to plasmid-driven merging of the two in particular genetic backgrounds (*44*, *45*).

Our observations of multiple cases of convergent evolution in distant genetic backgrounds by plasmids as a vehicle, suggests that they function as a significant means for new clones to first establish, and then later endemically sustain themselves in a host population. However, the evidence emerging from the comparison of publicly available *E. coli* plasmidome with the additional >4,000 plasmids sequenced as part of our investigation, further suggests that plasmid evolution within the species is relatively constrained, given that hardly any novel plasmid types were discovered. This would further imply that novel clones actually have a limited freedom in choosing what plasmid backbones to rely on for acquisition of traits on their road to eventual success, as reflected for example in the observed highly constrained variation in bacteriocin-producing systems found to be prevalent. Future studies including mechanistic modeling of plasmid evolution and maintenance, could shed further light on the factors that clone success ultimately depends on.

Materials and Methods

Isolate selection for long-read sequencing

To resolve the accessory genome and mobile genetic elements (MGEs) of the extensive Norwegian Surveillance System for Resistant Microbes (NORM) ExPEC population (*7*), we selected 2,045 isolates (2,045/3,254, 62.85%) for long-read Oxford Nanopore Technologies (ONT) sequencing and, subsequently, for hybrid assembly. This selection was performed following a two-step strategy. First, to ensure that the resulting genomes represented the accessory genome diversity inherent in the NORM collection, 1,085 isolates (1,085/2,045, 53.06%) were selected regardless of their clonal complex using an unbiased statistical approach (*46*) (Fig S1A). Briefly, we considered the presence/absence matrix of the orthologous genes computed by Panaroo (version 1.2.3) (*47*) on the 3,254 Illumina assemblies (*7*). From these, we used the k-means algorithm to select 1,085 isolates on the dimensionally reduced matrix computed by t-sne considering Jaccard distances.

Second, we focused on the four major ExPEC clones (ST73, ST95, ST131 and ST69) and sequenced all their remaining isolates not selected within the first step (960/2,045, 46.94%).

This permitted us to directly estimate the prevalence of particular plasmid types among the four major ExPEC clones.

DNA isolation and ONT sequencing

Long-read sequencing of the 2,045 isolates was performed using a high-throughput multiplexing approach based on ONT reads (*46*). All the isolates were separately grown on MacConkey agar No. 3 (Oxoid Ltd., Thermo Fisher Scientific Inc., Waltham, MA, USA) at 37˚C overnight. Individual colonies were picked for overnight growth at 37˚C in 1.6mL LB (Miller) broth (BD, Franklin Lakes,NJ, USA) each. Genomic high-molecular-weight DNA was extracted from cell pellets using MagAttract R HMW DNA Kit (Qiagen, Hilden, Germany) to a final elution volume of 100 µL. Output concentration and DNA integrity were measured using NanoDropOne spectrophotometer (Thermo Scientific) and the Qubit dsDNA HS assay kit (Thermo Fisher Scientific) on a CLARIOstar microplate reader (BMG Labtech, Ortenberg, Germany). The samples were then adjusted to 400 ng for long-read sequencing. The ONT libraries were prepared using SQK-LSK109(-XL) or SQK-NBD110-96 barcoding kit for 24- and 96-barcoding runs, respectively, and 40 fmol per sample was loaded onto FLO-MIN106 flow cells. Sequencing was run for 72 hours on GridION (Oxford Nanopore Technologies, Oxford, UK) using MinKNOW Core versions 3.6.0 to 4.2.5. Basecalling and demultiplexing were performed using Guppy versions 3.2.8 to 4.3.4, with fast basecalling for the 24-barcoding and high-accuracy basecalling for the 96-barcoding runs.

Hybrid assemblies

ONT long reads were combined with the existing short-read data (*7*) using a previously published hybrid assembly pipeline (*48*) publicly available at https://github.com/arredondo23/hybrid_assembly_slurm. This pipeline is largely based on Unicycler (version 0.4.7) and was designed to automate the generation of (near-)complete genomes.

In total, we could obtain a hybrid assembly for 1,999 genomes (1,999/2,045, 97.75%). The information derived from Unicycler regarding the length, circularity and depth of each contig was extracted and considered for the downstream analyses. The hybrid assemblies were split into individual contigs and circlator (version 1.5.5) (*49*) was used with the command fixstart to rotate the starting position of each contig.

Each contig was classified either as chromosome-derived, plasmid-derived or virus-derived (bacteriophage) considering both the predictions of mlplasmids (species 'Escherichia coli') (version 2.1.0) and geNomad (command end-to-end) (version 1.5.0) (*50*). First, contigs were labelled as chromosomal if they either had a size larger than 500 kbp or a mlplasmids chromosome probability higher than 0.7 and geNomad chromosome score higher than 0.7. Second, contigs were labeled as plasmids if they had a minimum mlplasmids plasmid probability of 0.3 and geNomad plasmid score of 0.7. These settings were used to compensate for the number of false negative predictions reported previously for mlplasmids using the *E. coli* model (*51*). Third, contigs not meeting any of the previous criteria were classified as virus-derived if they had a minimum geNomad virus score of 0.7. Fourth, the remaining contigs were considered as unclassified.

Prokka (version 1.14.6) (*52*) using the genus *Escherichia* and species *coli* was considered to annotate each of the contigs resulting from the hybrid assemblies. Abricate (version 1.0.1) (https://github.com/tseemann/abricate) coupled with the databases of plasmidfinder (*53*), ecoli_vf (https://github.com/phac-nml/ecoli_vf), and ecoh (*54*) was used to assign the presence (minimum identity and coverage of 80%) of plasmid replicon sequences, *E. coli* known virulence factors (including colicin genes) and *E. coli* O-H serotype predictions respectively. AMRFinderPlus (version 3.10.18) (*55*) using the flag --plus and *Escherichia* as organism was also considered to identify AMR, stress-response and virulence genes. The presence of plasmid replicon sequences, relaxases, mate-pair formation types and the predicted mobility of the contig was computed using the module mob_typer (version 3.0.3) of MOB-suite (*56*, *57*). Plasmid multi-locus sequence typing (pMLST) was computed using the IncF, IncA/C, IncHI1, IncHI2, IncI1, IncN and pbssb1-family schemes using the pmlst.py script available at https://bitbucket.org/genomicepidemiology/pmlst (*53*). For the plasmids with an IncF scheme annotation, we reported the FAB formula to contextualize them with previous literature.

Typing the plasmid-derived sequences with mge-cluster

The circular plasmid sequences were considered to develop a plasmid typing scheme using mge-cluster (version 1.1.0) (*20*). First, we removed the redundancy within the set of circular plasmid sequences using cd-hit-est (version 4.8.1) (*58*, *59*) using a sequence identity threshold of 0.99, alignment coverage of 0.9 and length difference cutoff of 0.9. Cd-hit-est computed a total of 2,560 clusters from which a single representative sequence was chosen. This set of non-redundant plasmid sequences (n=2,560) was considered as input for the --create operational mode of mge-cluster using a unitig filtering variance of 0.01, a perplexity value of 30 and a minimum threshold of 30 sequences to define a cluster. The resulting mge-cluster typing model is available at https://doi.org/10.6084/m9.figshare.24305392.v1 allowing reuse of the clustering system presented with new plasmid sequences. The set of redundant plasmid sequences (n=1,925) filtered by cd-hit-est was embedded and assigned to the resulting clusters using the operational mode --existing of mge-cluster. For each mge-cluster, we discarded any sequences differing by two standard-deviations of the average plasmid length. This criteria was included to ensure that only plasmids with a similar size were part of the same mge-cluster referred to as plasmid types (pTs).

The circular plasmid sequences included in the 23 mge-clusters were sketched (k=31,scaled=1000,noabund) with sourmash (version 4.8.2) (*21*) and a containment matrix computed using the command compare and the flag --containment. This matrix contains containment values corresponding to the fraction of a particular sketch found in a second sketch. The containment values were averaged within and between mge-clusters.

To contextualize the described plasmid clusters with well-studied ExPEC plasmids, we used the --existing operational mode of mge-cluster to assign the following reference plasmids with an mge-cluster: CP000244 (pUTI89), CP007150 (pRS218), CU928146 (pECOS88), CU928148 (p1ESCUM), DQ381420 (pAPEC-O1-ColBM), EU330199 (pVM01), LQHK01000003 (pG150_1) and NC_022651 (pJJ1886_5). These reference plasmids were used to further identify the presence of virulence genes previously studied and reported in literature (e.g *hlyF* and *ompT* genes).

Furthermore, the set of non-circular sequences (n=932) was also assigned to the resulting mge-cluster model using the operational mode --existing but we again only considered the assignment given by mge-cluster if the contig size was within two standard-deviations of the cluster average plasmid length. The estimated structure of the public domain *E. coli* plasmidome in (*20*) was finally compared with the structure of the plasmidome from the current study by merging the two using the –existing operational mode. The two resulting partitions of plasmid sequences were compared using adjusted Rand Index, which is a standard measure of congruence between two partitions of data. This resulted in the value 0.94 (maximum 1.0) and the only major discrepancy between the two partitions arose from a subdivision of pT11 and pT14 which remained merged in the previous analysis of published plasmid sequences, apparently due to lack of resolution caused by the limited number of sequences available.

Plasmid synteny and phylogenetic analyses

For the most relevant pTs(4, 6, 10, 11, 12, 13, 14, 21, 22), we selected a plasmid sequence (Fig. 2) with a membership probability value reported by mge-cluster of 1.0 and representative of the most common pMLST annotation and AMR/virulence gene content of the cluster (Table S2). The starting position of these sequences was changed to their IncFIB replicon using circlator (version 1.5.5) with the command fixstart. Clinker (version 0.0.21) with default settings was used to perform a gene synteny analysis and visualize which genomic blocks were shared among pTsconsidering a minimum sequence identity of 0.8 (Fig. 2) and 0.99 (Fig. S3) to draw a link between genes.

We performed recombination-free phylogenies using the circular plasmid sequences from pTs 6,14 and 10,11. To create a core-genome alignment from these pairs of pTs, we used snippy (version 4.6.0) (*60*) which mapped the sequences from pTs 6, 14 and 10, 11 against the reference genomes 30134_6#129_2 (pT 14) and 30224_1#245_5 (pT 11), respectively. The resulting core genome alignment was polished using the module from snippy (snippy-clean_full_aln). Gubbins (version 3.1.3) (*61*) was used to compute a recombination-free phylogeny of the pTs using 100 bootstrap replicates, 50 algorithm iterations, and RAxML (*62*) as the application for model fitting and other default settings (GTRGAMMA as nucleotide substitution model).

Dating the phylogenies of the four major ExPEC clones

The four major ExPEC clones were each mapped to a reference genome belonging to that lineage and recombination was removed using Gubbins v2.4.1 (*61*). Gubbins output was supplied to the R package BactDating v1.1.1 in three replicates and one with randomised tip dates. These ran through Markov chain Monte Carlo (MCMC) chains of 100,000,000 generations sampled every 1000 states with a 10,000,000 burn-in using the Additive Relaxed Clock (ARC) model (*22*). The three replicate MCMC chains were deemed to have converged with Gelman diagnostic of approximately 1 for mu, sigma and alpha using the coda R package (*63*). We assessed whether the effective sample size (ESS) on the first replicate model was greater than 200 using the effectiveSize function of the coda R package (*63*) and determined that the true dates model was better than the randomised dates model. The R CaveDive package v0.1.1 was used to detect probable expansion events in the dated phylogeny (*23*). To report the dates of the distinct plasmid acquisition or introduction dates, we indicate in the text the lower and upper bounds of the confidence intervals (CI). For instance, pT11 was acquired by ST95 from 1846 (95% CI 1817-1870) to 1922 (95% CI 1904-1937) (Table S2). This is reported in the text as acquired around ~1884 (95% CI 1817-1937).

Determination of bacteriocin activity

We produced four independent batches of microcin V by growing plasmid-harboring isolates 23-46, 27-20, 27-61 (ST95) and 28-33 (ST131 clade B) for 20 h at 37°C with aeration in 10 mL Lysogeny Broth (LB; LB Broth, Miller, Difco) with 0.2 mM 2,2'-bipyridyl (to limit iron availability; Sigma-Aldrich). The cultures were then centrifuged (4000 rpm, 10 min), filtered (0.2 µm polyethersulfone filter; VWR), and stored at 4°C. As controls, a plasmid-free ST131 clade B isolate 27-56 was treated in the same conditions, while isolate 27-61 was further treated without induction (no 2,2'-bipyridyl added).

To produce batches of colicin Ia, the four plasmid-harboring isolates mentioned above were grown in 10 mL LB overnight (15-16 h) at 37°C with aeration, centrifuged (4000 rpm, 10 min) and then the pellets were resuspended in 10 mL PBS. The suspensions were UV-

irradiated (254 nm wavelength) with a dose of 36 mW/cm^2/second for 10 seconds. 500 µL from these irradiated suspensions were inoculated into 9.5 mL LB and incubated 4 h at 37°C with aeration (covered in aluminum foil to prevent photoreactivation). The cultures were then centrifuged (4000 rpm, 10 min), filtered (0.2 µm polyethersulfone filter; VWR), and stored at 4°C. As controls, the plasmid-free isolate 27-56 was treated in the same conditions, while isolate 27-61 was further treated without induction (no UV-irradiation).

51 NORM isolates (Table S4), as well as *E. coli* MG1655, were grown in 1 mL LB in 96-deep-well plates overnight at 37°C with aeration. Indicator plates were prepared by inoculating 15 µL of each overnight culture into 15 mL LB top agar (LB with 0.75% agar; Sigma-Aldrich) with 0.1 mM 2,2'-bipyridyl (to limit iron availability and promote expression of the colicin receptor Cir). 10 µL of each bacteriocin and control batch were spotted on indicator plates, which were incubated overnight at 30°C after drying. These assays were performed three independent times. *E. coli* MG1655 was further exposed to 10-fold dilutions of the bacteriocin batches to verify the characteristic dose-dependent inhibition and exclude phage activity.

Most of the sensitive NORM isolates displayed sensitivity only against microcin V, while the six sensitive to colicin Ia were sensitive to both bacteriocins. Because the supernatants could contain both bacteriocins (since at least one seems to be spontaneously produced at low concentrations), and they have a common receptor in target cells, we added a further control to unequivocally test the sensitivity of these strains to each of the bacteriocins. From genomic data, a few isolates were predicted to encode only one of the bacteriocins and we used two of those (isolates 31-16 and 31-17) for further controls. These two isolates were treated, with and without specific induction for each bacteriocin, as described above. The ST95 isolate 31-17 encodes only microcin V, and produces inhibition only when grown under iron-limiting conditions (Fig. 4B). Alternatively, ST131-B isolate 31-16 encodes only colicin Ia and always produces inhibition, albeit at stronger levels upon SOS induction (Fig. 4C). Combined, this shows that the spontaneously produced bactericion is colicin Ia. This strain is sensitive against the supernatant of 31-17, verifying the loss of the microcin V genes (lack of the immunity genes). Each of these two supernatants was tested against the

six strains mentioned above, and both were inhibitory to all six (as displayed in Fig. S10, Table S4).

Availability of data

The 4,759 nucleotide plasmid sequences indicated in Table S1 are publicly available in the permanent figshare link https://doi.org/10.6084/m9.figshare.24302884.v1

Acknowledgement

Collaborators forming The Norwegian *E. coli* BSI Study Group: Nina Handal (Akershus University Hospital), Nils Olav Hermansen (Oslo University Hospital, Ullevål), Anita Kanestrøm (Østfold Hospital), Hege Elisabeth Larsen (Nordland Hospital), Paul Christoffer Lindemann (Haukeland University Hospital), Iren Høyland Löhr (Stavanger University Hospital), Åshild Marvik (Vestfold Hospital), Einar Nilsen (Molde Hospital and Ålesund Hospital), Marcela Pino (Oslo University Hospital, Rikshospitalet), Elisabeth Sirnes (Førde Hospital), Ståle Tofteland (Sørlandet Hospital), Kyriakos Zaragkoulias (Nord-Trøndelag Hospital Trust).
We acknowledge the support from the Genomic Support Centre Tromsø, UiT The Arctic University of Norway for Oxford Nanopore sequencing and François Cléon for excellent technical assistance.

References

1. M. Donnenberg, *Escherichia coli: Pathotypes and Principles of Pathogenesis* (Academic Press, 2013).

2. L. W. Riley, Pandemic lineages of extraintestinal pathogenic Escherichia coli. *Clin. Microbiol. Infect.* **20**, 380–390 (2014).

3. E. A. Cummins, A. E. Snaith, A. McNally, R. J. Hall, The role of potentiating mutations in the evolution of pandemic Escherichia coli clones. *Eur. J. Clin. Microbiol. Infect. Dis.* (2021), doi:10.1007/s10096-021-04359-3.

4. S. J. Dunn, C. Connor, A. McNally, The evolution and transmission of multi-drug resistant Escherichia coli and Klebsiella pneumoniae: the complexity of clones and plasmids. *Curr. Opin. Microbiol.* **51**, 51–56 (2019).

5. A. R. Manges, J. R. Johnson, Food-borne origins of Escherichia coli causing extraintestinal infections. *Clin. Infect. Dis.* **55** (2012), pp. 712–719.

6. T. Kallonen, H. J. Brodrick, S. R. Harris, J. Corander, N. M. Brown, V. Martin, S. J. Peacock, J. Parkhill, Systematic longitudinal survey of invasive Escherichia coli in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Res.* **27**, 1437–1449 (2017).

7. R. A. Gladstone, A. McNally, A. K. Pöntinen, G. Tonkin-Hill, J. A. Lees, K. Skytén, F. Cléon, M. O. K. Christensen, B. C. Haldorsen, K. K. Bye, K. W. Gammelsrud, R. Hjetland, A. Kümmel, H. E. Larsen, P. C. Lindemann, I. H. Löhr, Å. Marvik, E. Nilsen, M. T. Noer, G. S. Simonsen, M. Steinbakk, S. Tofteland, M. Vattøy, S. D. Bentley, N. J. Croucher, J. Parkhill, P. J. Johnsen, Ø. Samuelsen, J. Corander, 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*. **2**, e331–e341 (2021).

8. A. K. Pöntinen, R. A. Gladstone, Pesonen Henri, M. Pesonen, F. Cléon, B. J. Parcell, T. Kallonen, G. Skov Simonsen, N. J. Croucher, A. McNally, J. Parkhill, P. J. Johnsen, Ø. Samuelsen, J. Corander, Modulation of multi-drug resistant clone success in Escherichia coli populations: a longitudinal multi-country genomic and antibiotic usage cohort study. *Lancet Microbe*. **in press**.

9. A. McNally, T. Kallonen, C. Connor, K. Abudahab, D. M. Aanensen, C. Horner, S. J. Peacock, J. Parkhill, N. J. Croucher, J. Corander, Diversification of Colonization Factors in a Multidrug-Resistant Escherichia coli Lineage Evolving under Negative Frequency-Dependent Selection. *MBio*. **10** (2019), doi:10.1128/mBio.00644-19.

10. T. Watanabe, Infective heredity of multiple drug resistance in bacteria. *Bacteriol. Rev.* **27**, 87–115 (1963).

11. D. R. Helinski, A Brief History of Plasmids. *EcoSal Plus*. **10**, eESP00282021 (2022).

12. D. J. Rankin, E. P. C. Rocha, S. P. Brown, What traits are carried on mobile genetic elements, and why? *Heredity* . **106**, 1–10 (2011).

13. F. Baquero, V. F. Lanza, M.-R. Baquero, R. Del Campo, D. A. Bravo-Vázquez, Microcins in Enterobacteriaceae: Peptide Antimicrobials in the Eco-Active Intestinal Chemosphere. *Front. Microbiol.* **10**, 2261 (2019).

14. Johnson Timothy J., Role of Plasmids in the Ecology and Evolution of “High-Risk” Extraintestinal Pathogenic Escherichia coli Clones. *EcoSal Plus*. **9**, eESP–0013–2020 (2021).

15. T. J. Johnson, J. L. Danzeisen, B. Youmans, K. Case, K. Llop, J. Munoz-Aguayo, C. Flores-Figueroa, M. Aziz, N. Stoesser, E. Sokurenko, L. B. Price, J. R. Johnson, Separate F-Type Plasmids Have Shaped the Evolution of the H30 Subclone of Escherichia coli Sequence Type 131. *mSphere*. **1** (2016), doi:10.1128/mSphere.00121-16.

16. M. L. Cummins, C. J. Reid, S. P. Djordjevic, F Plasmid Lineages in Escherichia coli ST95: Implications for Host Range, Antibiotic Resistance, and Zoonoses. *mSystems*. **7**, e0121221 (2022).

17. J. Rodríguez-Beltrán, J. DelaFuente, R. León-Sampedro, R. C. MacLean, Á. San Millán, Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nat. Rev. Microbiol.* **19**, 347–359 (2021).

18. S. Bengtsson, U. Naseer, A. Sundsfjord, G. Kahlmeter, M. Sundqvist, Sequence types and plasmid carriage of uropathogenic Escherichia coli devoid of phenotypically detectable resistance. *J. Antimicrob. Chemother.* **67**, 69–73 (2012).

19. J. A. Gama, J. Kloos, P. J. Johnsen, Ø. Samuelsen, Host dependent maintenance of a blaNDM-1-encoding plasmid in clinical Escherichia coli isolates. *Sci. Rep.* **10**, 9332 (2020).

20. S. Arredondo-Alonso, R. A. Gladstone, A. K. Pöntinen, J. A. Gama, A. C. Schürch, V. F. Lanza, P. J. Johnsen, Ø. Samuelsen, G. Tonkin-Hill, J. Corander, Mge-cluster: a reference-free approach for typing bacterial plasmids. *NAR Genom Bioinform.* **5**, lqad066 (2023).

21. N. T. Pierce, L. Irber, T. Reiter, P. Brooks, C. T. Brown, Large-scale sequence comparisons with sourmash. *F1000Res.* **8**, 1006 (2019).

22. X. Didelot, N. J. Croucher, S. D. Bentley, S. R. Harris, D. J. Wilson, Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res.* **46**, e134 (2018).

23. D. Helekal, A. Ledda, E. Volz, D. Wyllie, X. Didelot, Bayesian Inference of Clonal Expansions in a Dated Phylogeny. *Syst. Biol.* **71**, 1073–1087 (2022).

24. M. Rozwandowicz, M. S. M. Brouwer, J. Fischer, J. A. Wagenaar, B. Gonzalez-Zorn, B. Guerra, D. J. Mevius, J. Hordijk, Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae. *J. Antimicrob. Chemother.* **73**, 1121–1137 (2018).

25. T. J. Johnson, K. E. Siek, S. J. Johnson, L. K. Nolan, DNA sequence of a ColV plasmid and prevalence of selected plasmid-encoded virulence genes among avian Escherichia coli strains. *J. Bacteriol.* **188**, 745–758 (2006).

26. A. Jeziorowski, D. M. Gordon, Evolution of microcin V and colicin Ia plasmids in Escherichia coli. *J. Bacteriol.* **189**, 7045–7052 (2007).

27. M. Sabri, S. Léveillé, C. M. Dozois, A SitABCD homologue from an avian pathogenic Escherichia coli strain mediates transport of iron and manganese and resistance to hydrogen peroxide. *Microbiology*. **152**, 745–758 (2006).

28. L. K. Nolan, S. M. Horne, C. W. Giddings, S. L. Foley, T. J. Johnson, A. M. Lynne, J. Skyberg, Resistance to serum complement, iss, and virulence of avian Escherichia coli. *Vet. Res. Commun.* **27**, 101–110 (2003).

29. Johnson Timothy J., Nolan Lisa K., Pathogenomics of the Virulence Plasmids of Escherichia coli. *Microbiol. Mol. Biol. Rev.* **73**, 750–774 (2009).

30. D. M. Gordon, S. Geyik, O. Clermont, C. L. O'Brien, S. Huang, C. Abayasekara, A. Rajesh, K. Kennedy, P. Collignon, P. Pavli, C. Rodriguez, B. D. Johnston, J. R. Johnson, J.-W. Decousser, E. Denamur, Fine-Scale Structure Analysis Shows Epidemic Patterns of Clonal Complex 95, a Cosmopolitan Escherichia coli Lineage Responsible for Extraintestinal Infection. *mSphere*. **2** (2017), doi:10.1128/mSphere.00168-17.

31. C. K. Cusumano, C. S. Hung, S. L. Chen, S. J. Hultgren, Virulence plasmid harbored by

uropathogenic Escherichia coli functions in acute stages of pathogenesis. *Infect. Immun.* **78**, 1457–1467 (2010).

32. D. S. S. Wijetunge, K. H. E. M. Karunathilake, A. Chaudhari, R. Katani, E. G. Dudley, V. Kapur, C. DebRoy, S. Kariyawasam, Complete nucleotide sequence of pRS218, a large virulence plasmid, that augments pathogenic potential of meningitis-associated Escherichia coli strain RS218. *BMC Microbiol.* **14**, 203 (2014).

33. K. Kondratyeva, M. Salmon-Divon, S. Navon-Venezia, Meta-analysis of Pandemic Escherichia coli ST131 Plasmidome Proves Restricted Plasmid-clade Associations. *Sci. Rep.* **10**, 36 (2020).

34. C. T. Bergstrom, M. Lipsitch, B. R. Levin, Natural selection, infectious transfer and the existence conditions for bacterial plasmids. *Genetics*. **155**, 1505–1519 (2000).

35. J. Corander, C. Fraser, M. U. Gutmann, B. Arnold, W. P. Hanage, S. D. Bentley, M. Lipsitch, N. J. Croucher, Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. *Nature Ecology & Evolution*, 1 (2017).

36. C.-D. Köhler, U. Dobrindt, What defines extraintestinal pathogenic Escherichia coli? *Int. J. Med. Microbiol.* **301**, 642–647 (2011).

37. T. Mäklin, H. A. Thorpe, A. K. Pöntinen, R. A. Gladstone, Y. Shao, M. Pesonen, A. McNally, P. J. Johnsen, Ø. Samuelsen, T. D. Lawley, A. Honkela, J. Corander, Strong pathogen competition in neonatal gut colonisation. *Nat. Commun.* **13**, 7417 (2022).

38. T. J. Johnson, S. J. Johnson, L. K. Nolan, Complete DNA sequence of a ColBM plasmid from avian pathogenic Escherichia coli suggests that it evolved from closely related ColV virulence plasmids. *J. Bacteriol.* **188**, 5975–5983 (2006).

39. C. M. Liu, M. Stegger, M. Aziz, T. J. Johnson, K. Waits, L. Nordstrom, L. Gauld, B. Weaver, D. Rolland, S. Statham, J. Horwinski, S. Sariya, G. S. Davis, E. Sokurenko, P. Keim, J. R. Johnson, L. B. Price, Escherichia coli ST131-H22 as a Foodborne Uropathogen. *MBio*. **9** (2018), doi:10.1128/mBio.00470-18.

40. C. M. Liu, M. Aziz, D. E. Park, Z. Wu, M. Stegger, M. Li, Y. Wang, K. Schmidlin, T. J. Johnson, B. J. Koch, B. A. Hungate, L. Nordstrom, L. Gauld, B. Weaver, D. Rolland, S. Statham, B. Hall, S. Sariya, G. S. Davis, P. S. Keim, J. R. Johnson, L. B. Price, Using source-associated mobile genetic elements to identify zoonotic extraintestinal E. coli infections. *One Health*. **16**, 100518 (2023).

41. A. R. Manges, J. R. Johnson, B. Foxman, T. T. O'Bryan, K. E. Fullerton, L. W. Riley, Widespread distribution of urinary tract infections caused by a multidrug-resistant Escherichia coli clonal group. *N. Engl. J. Med.* **345**, 1007–1013 (2001).

42. J. R. Johnson, M. E. Menard, T.-L. Lauderdale, C. Kosmidis, D. Gordon, P. Collignon, J. N. Maslow, A. T. Andrasević, M. A. Kuskowski, Trans-Global Initiative for Antimicrobial Resistance Analysis Investigators, Global distribution and epidemiologic associations of Escherichia coli clonal group A, 1998-2007. *Emerg. Infect. Dis.* **17**, 2001–2009 (2011).

43. K. E. Holt, H. Wertheim, R. N. Zadoks, S. Baker, C. A. Whitehouse, D. Dance, A. Jenney, T. R. Connor, L. Y. Hsu, J. Severin, S. Brisse, H. Cao, J. Wilksch, C. Gorrie, M. B. Schultz,

D. J. Edwards, K. Van Nguyen, T. V. Nguyen, T. T. Dao, M. Mensink, V. L. Minh, N. T. K. Nhu, C. Schultsz, K. Kuntaman, P. N. Newton, C. E. Moore, R. A. Strugnell, N. R. Thomson, Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E3574–81 (2015).

44. M. M. C. Lam, K. L. Wyres, R. R. Wick, L. M. Judd, A. Fostervold, K. E. Holt, I. H. Löhr, Convergence of virulence and MDR in a single plasmid vector in MDR Klebsiella pneumoniae ST15. *J. Antimicrob. Chemother.* **74**, 1218–1222 (2019).

45. K. L. Wyres, T. N. T. Nguyen, M. M. C. Lam, L. M. Judd, N. van Vinh Chau, D. A. B. Dance, M. Ip, A. Karkey, C. L. Ling, T. Miliya, P. N. Newton, N. P. H. Lan, A. Sengduangphachanh, P. Turner, B. Veeraraghavan, P. V. Vinh, M. Vongsouvath, N. R. Thomson, S. Baker, K. E. Holt, Genomic surveillance for hypervirulence and multi-drug resistance in invasive Klebsiella pneumoniae from South and Southeast Asia. *Genome Med.* **12**, 11 (2020).

46. S. Arredondo-Alonso, A. K. Pöntinen, F. Cléon, R. A. Gladstone, A. C. Schürch, P. J. Johnsen, Ø. Samuelsen, J. Corander, A high-throughput multiplexing and selection strategy to complete bacterial genomes. *Gigascience*. **10** (2021), doi:10.1093/gigascience/giab079.

47. G. Tonkin-Hill, N. MacAlasdair, C. Ruis, A. Weimann, G. Horesh, J. A. Lees, R. A. Gladstone, S. Lo, C. Beaudoin, R. A. Floto, S. D. W. Frost, J. Corander, S. D. Bentley, J. Parkhill, Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020).

48. A. K. Pöntinen, J. Top, S. Arredondo-Alonso, G. Tonkin-Hill, A. R. Freitas, C. Novais, R. A. Gladstone, M. Pesonen, R. Meneses, H. Pesonen, J. A. Lees, D. Jamrozy, S. D. Bentley, V. F. Lanza, C. Torres, L. Peixe, T. M. Coque, J. Parkhill, A. C. Schürch, R. J. L. Willems, J. Corander, Apparent nosocomial adaptation of Enterococcus faecalis predates the modern hospital era. *Nat. Commun.* **12**, 1523 (2021).

49. M. Hunt, N. D. Silva, T. D. Otto, J. Parkhill, J. A. Keane, S. R. Harris, Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol.* **16**, 294 (2015).

50. A. P. Camargo, S. Roux, F. Schulz, M. Babinski, Y. Xu, B. Hu, P. S. G. Chain, S. Nayfach, N. C. Kyrpides, Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.*, 1–10 (2023).

51. J. A. Paganini, N. L. Plantinga, S. Arredondo-Alonso, R. J. L. Willems, A. C. Schürch, Recovering Escherichia coli Plasmids in the Absence of Long-Read Sequencing Data. *Microorganisms*. **9** (2021), doi:10.3390/microorganisms9081613.

52. T. Seemann, Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*. **30**, 2068–2069 (2014).

53. A. Carattoli, E. Zankari, A. Garciá-Fernández, M. V. Larsen, O. Lund, L. Villa, F. M. Aarestrup, H. Hasman, In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* **58**, 3895–3903 (2014).

54. D. J. Ingle, M. Valcanis, A. Kuzevski, M. Tauschek, M. Inouye, T. Stinear, M. M. Levine, R.

M. Robins-Browne, K. E. Holt, 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*. **2**, e000064 (2016).

55. M. Feldgarden, V. Brover, N. Gonzalez-Escalona, J. G. Frye, J. Haendiges, D. H. Haft, M. Hoffmann, J. B. Pettengill, A. B. Prasad, G. E. Tillman, G. H. Tyson, W. Klimke, AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Sci. Rep.* **11**, 12728 (2021).

56. J. Robertson, J. H. E. Nash, MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom*. **4** (2018), doi:10.1099/mgen.0.000206.

57. J. Robertson, K. Bessonov, J. Schonfeld, J. H. E. Nash, Universal whole-sequence-based plasmid typing and its utility to prediction of host range and epidemiological surveillance. *Microb Genom*. **6** (2020), doi:10.1099/mgen.0.000435.

58. W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. **22**, 1658–1659 (2006).

59. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. **28**, 3150–3152 (2012).

60. T. Seemann, Snippy: fast bacterial variant calling from NGS reads. 2015. *All rights reserved. No reuse allowed without permission* (2017).

61. N. J. Croucher, A. J. Page, T. R. Connor, A. J. Delaney, J. A. Keane, S. D. Bentley, J. Parkhill, S. R. Harris, Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).

62. A. Stamatakis, RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. **22**, 2688–2690 (2006).

63. M. Plummer, N. Best, K. Cowles, K. Vines, CODA: convergence diagnosis and output analysis for MCMC. *R News*. **6**, 7–11 (2006).

Main Figures

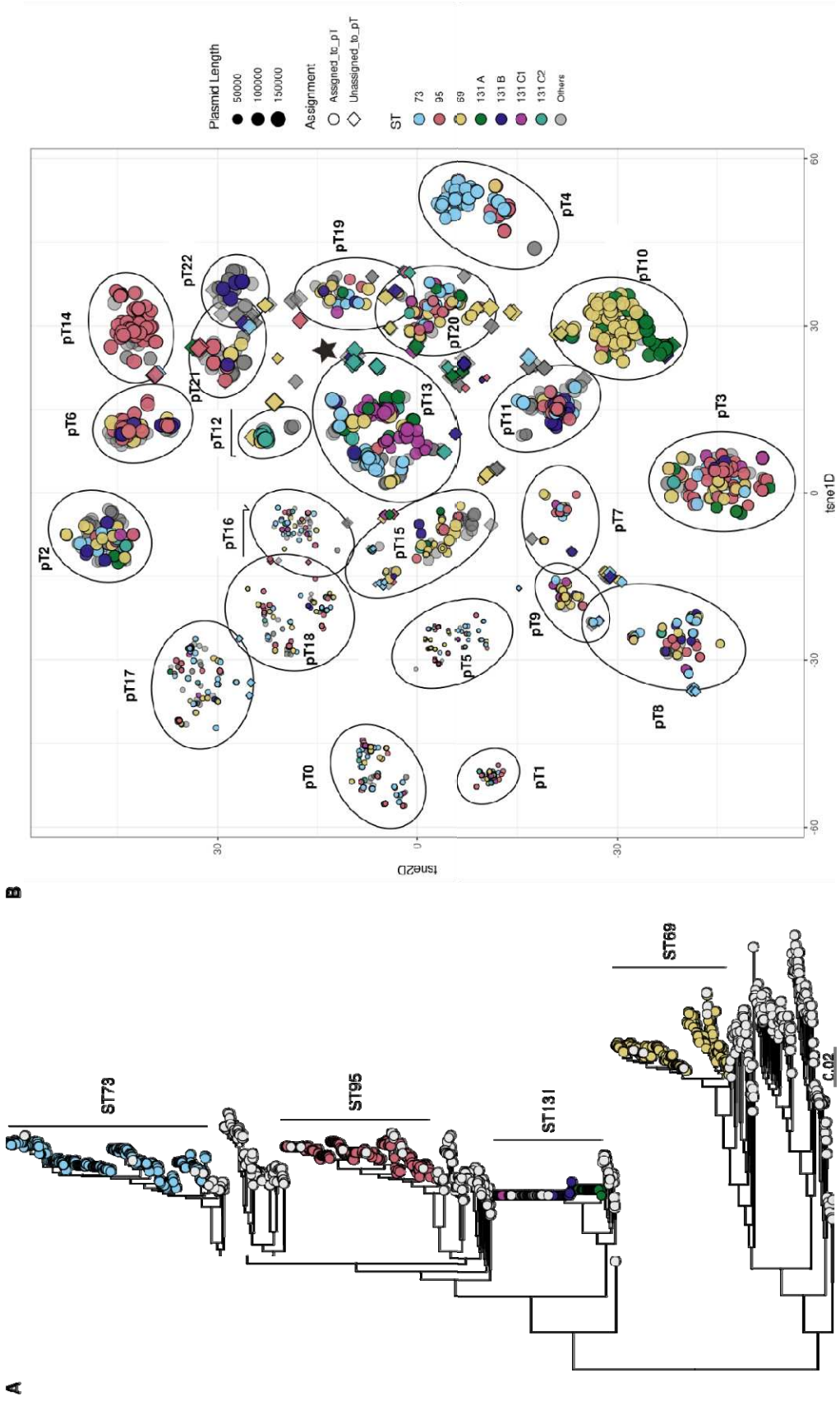
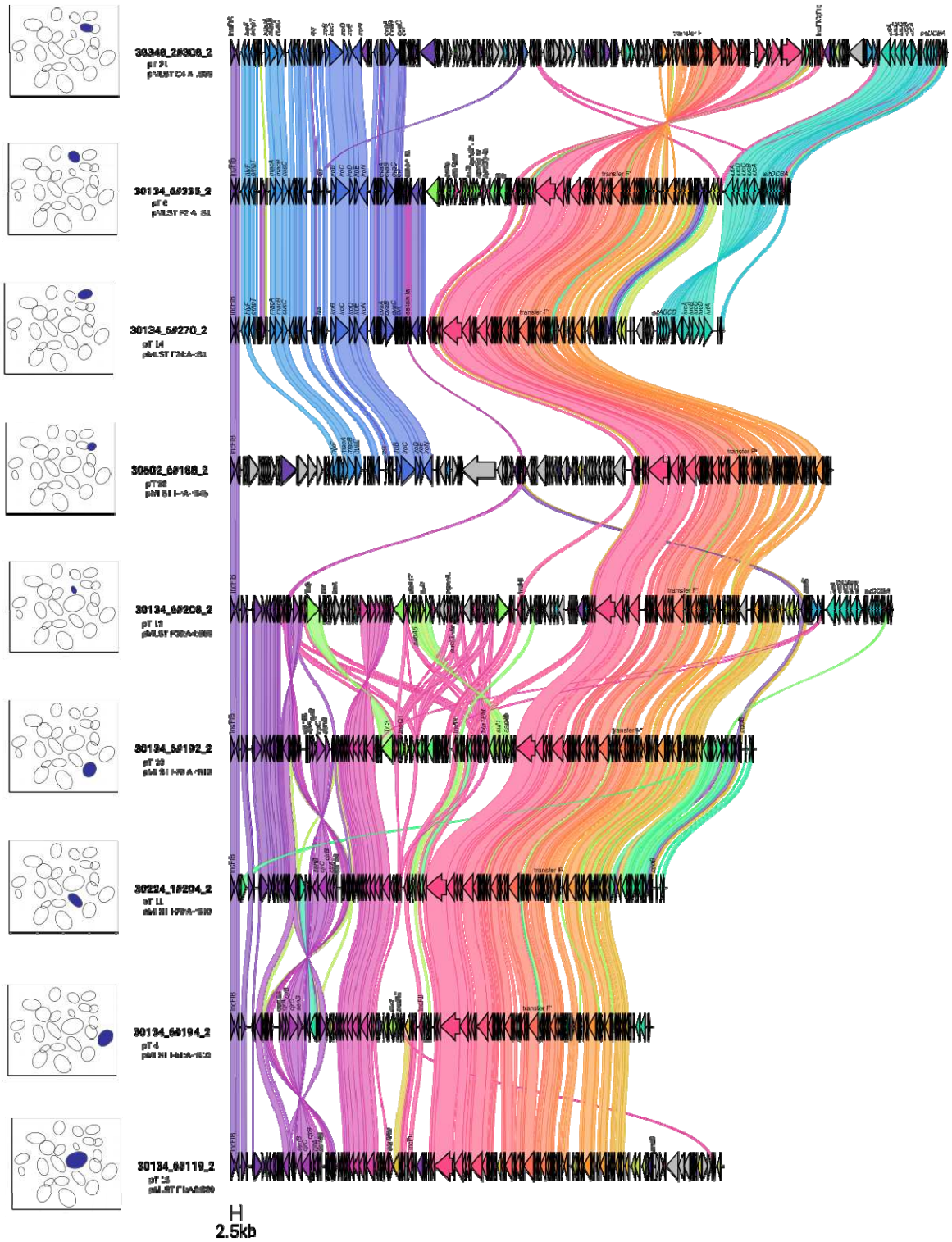


Figure 1

Figure 2



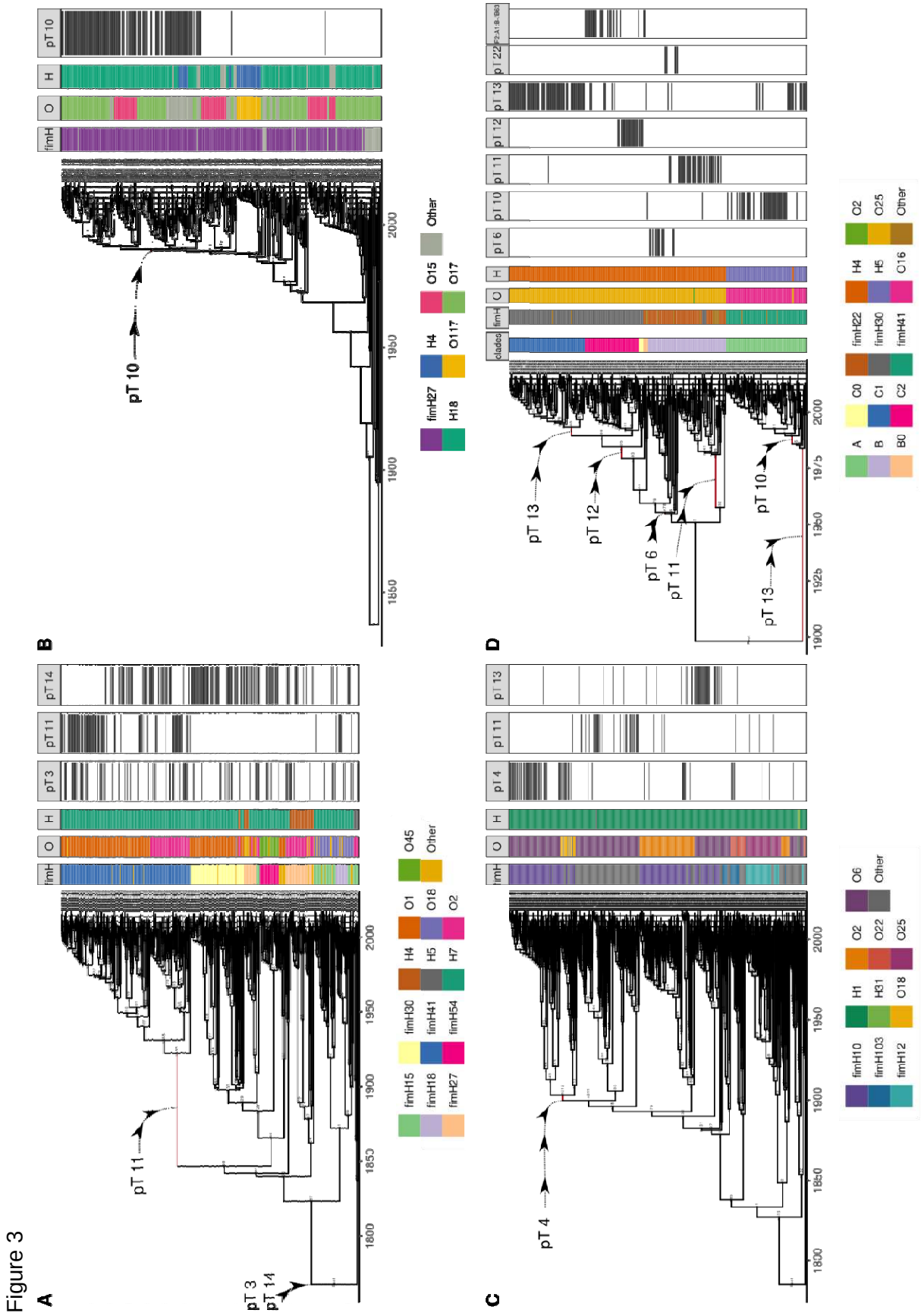


Figure 3

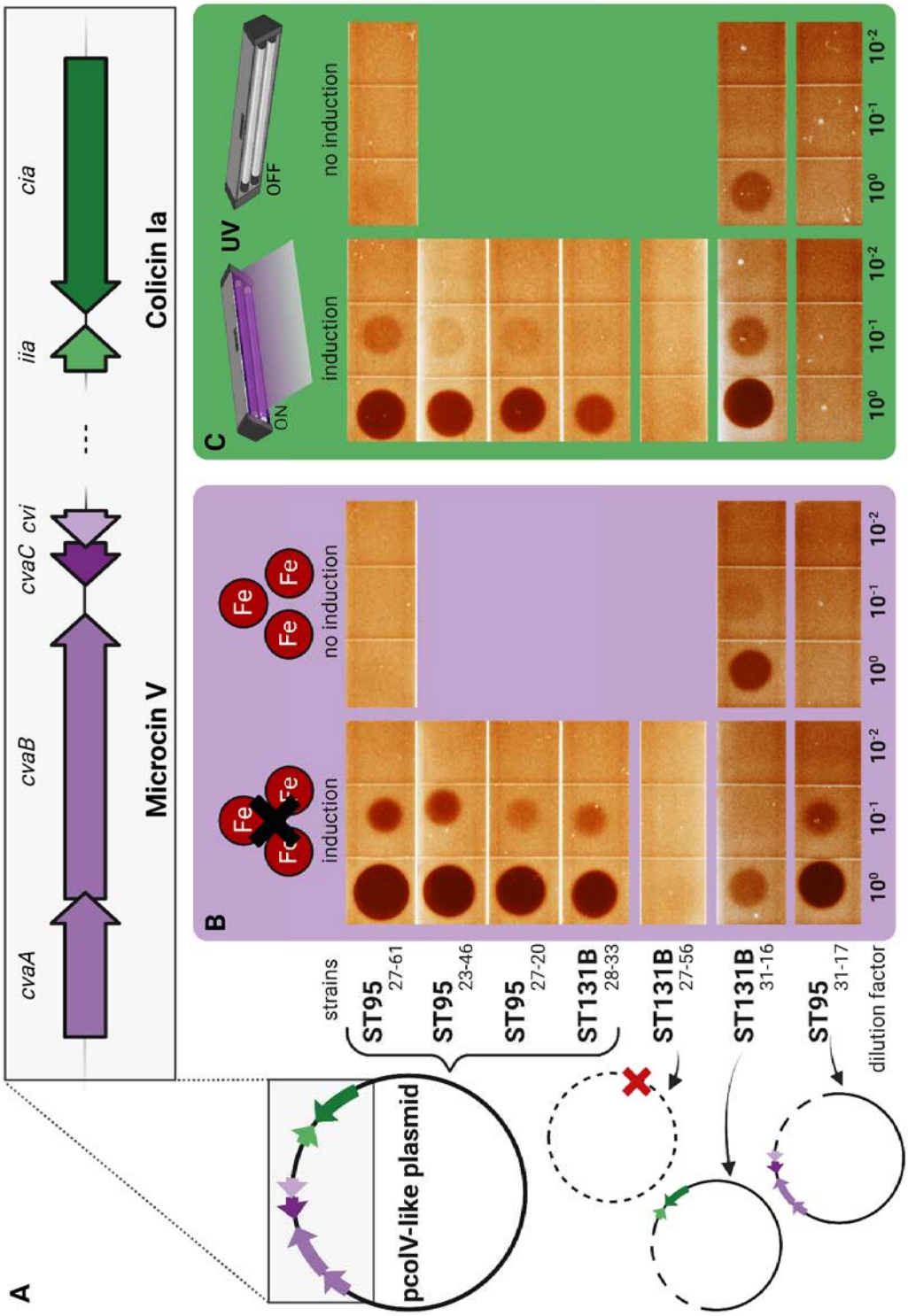


Figure 4

Figures and Tables Legends

Figure 1. Typing and visualization in the embedding space computed by mge-cluster of the plasmid sequences present in the four most common pandemic ExPEC clones (ST69, ST73, ST95 and ST131). A) Maximum-likelihood phylogeny of the 1,999 ExPEC isolates with an associated hybrid assembly. Isolates corresponding to the four pandemic ExPEC clones are indicated with a distinct color and their position marked in the phylogeny. In the case of ST131, the isolates were differentiated according to their associated clade. B) Non-linear embedding space created by mge-cluster considered to type and visualize 4,147 circular plasmid sequences. From these, 3,734 were assigned to a plasmid type (pT) (round shape) while 413 remained unassigned (diamond shape). Each point is coloured according to the ST (or ST131 clade) of the isolate carrying the plasmid and its size is proportional to the plasmid length. We indicated with a black star the position of a group of unassigned plasmid sequences which did not constitute an independent mge-cluster (pT) but had a distinct pMLST type corresponding to F2:A1:B-/B63.

Figure 2. Synteny analysis and visualization of the main large plasmid types (pT). For each pT, we selected a plasmid sequence (tagged with its sequence ID) representing the most predominant features present in the cluster (as summarized in Table S2). The gene synteny analysis was performed using clinker considering a minimum identity threshold of 0.8 to draw a link between genes. Virulence genes were identified in the synteny plot and their names indicated and curated according to literature or reference plasmids. Next to each pT, we indicated the position of the pT in the embedding space created by mge-cluster.

Figure 3. Dated phylogenies of the four main ExPEC clones and most likely acquisitions/introductions of their associated prevalent plasmid types (pT). A) Dated phylogeny of ST95 with clades indicated based on fimH/O/H typing and presence/absence of the pTs 3, 11 and 14. B)). B) Dated phylogeny of ST69 with clades indicated based on fimH/O/H typing and presence/absence of the pT 10. C) Dated phylogeny of ST73 with clades indicated based on fimH/O/H typing and presence/absence of the pTs 4, 11 and 13. C) Dated phylogeny of ST73 with clades indicated based on existing nomenclature (A, B, B0, C0, C1, C2), fimH/O/H typing and presence/absence of the pTs 6, 10, 11, 12, 13, 22. In addition, we indicated the presence/absence of the plasmids with pMLST F2:A1:B-/B63 (marked with a star in Figure 1).

Figure 4. Schematic representation of the bacteriocin susceptibility assay. A) Description of bacteriocinogenic plasmids from plasmid type 6 (ST131 clade B) and plasmid type 14 (ST95). These archetypical pcolV-like plasmids encode two bacteriocin gene clusters: microcin V and colicin Ia. Darker and lighter arrows indicate respectively the toxin (cvaC, cia) and immunity genes (cvi, iia) of each cluster. Control strains either harbor no plasmid (ST131 clade B strain 27-56) or harbor plasmid versions encoding only a single bacteriocin (colicin Ia for ST131 B strain 31-16, microcin V for ST95 strain 31-17). B) Susceptibility of E. coli MG1655 to microcin V. Production of microcin V is induced by iron limitation (simulated experimentally with the addition of the chelating agent 2,2'-bipyridyl). Bacteriocinogenic activity is demonstrated by growth inhibition in the spotted area (in a dilution-dependent way and without formation of individual plaques). C) Susceptibility of E. coli MG1655 to colicin Ia. Production of colicin Ia is induced through SOS induction (stimulated experimentally with UV irradiation). Bacteriocinogenic activity is demonstrated as in B.

Figure S1. Genome statistics of the hybrid assemblies obtained for 1,999 ExPEC genomes. A) Barplot indicating the total number of assemblies available per ST. B) Boxplot of the N50 metric (in base-pairs) only considering contigs classified as chromosomal. C). Boxplot of the sum length (in base-pairs) considering contigs classified as chromosomal. D) Boxplot of the sum length (in base-pairs) of all contigs classified as plasmid (referred as plasmidome).

Figure S2. Sourmash containment analysis of the 23 plasmid types (pT) defined by mge-cluster. The containment value can range from 0 (pTs share no genomic region in common) to 1 (the entire genomic region of the pT indicated the y-axis is present in the pT indicated in the x-axis).

Figure S3. Clinker synteny visualization of the main large plasmid types (pT) considering a stringent minimum identity threshold of 0.99 to draw links between genes. Virulence genes were identified in the synteny plot and their names indicated and curated according to literature or reference plasmids. Next to each pT, we indicated the position of the pT in the embedding space created by mge-cluster.

Figure S4. Recombination-free phylogeny based on plasmid sequences of pairs of plasmid types (pT) 6, 14 (panel A) and 10, 11 (panel B).

Figure S5. Distribution of the presence/absence of colicin genes in the ExPEC phylogeny (as shown in Figure 1A). Each colicin was coloured according if the gene was encoded in the chromosome, plasmid, virus (bacteriophage) or an unclassified contig.

Figure S6. Barplot distribution of the large plasmid types (pT) present in the four pandemic ExPEC clones. In the case of ST131, the plasmid distribution was split according to their clades (A, B, C1 or C2).

Figure S7. Stacked barplot showing the association of plasmid types (pT) and colicin genes. In some cases, the colicin gene was present on plasmid sequences unassigned by mge-cluster and represented with a -1 label.

Figure S8. Bacteriocin susceptibility of NORM isolates. Pruned phylogeny of the 51 E. coli NORM isolates tested experimentally for bacteriocin susceptibility (right). Susceptibility to microcin V and colicin Ia is individually indicated in the heatmap in blue, while yellow indicates bacteriocin insensitivity (no effect detected). Strains harboring the archetypical pcolV-like bacteriocinogenic plasmids are indicated in grey.

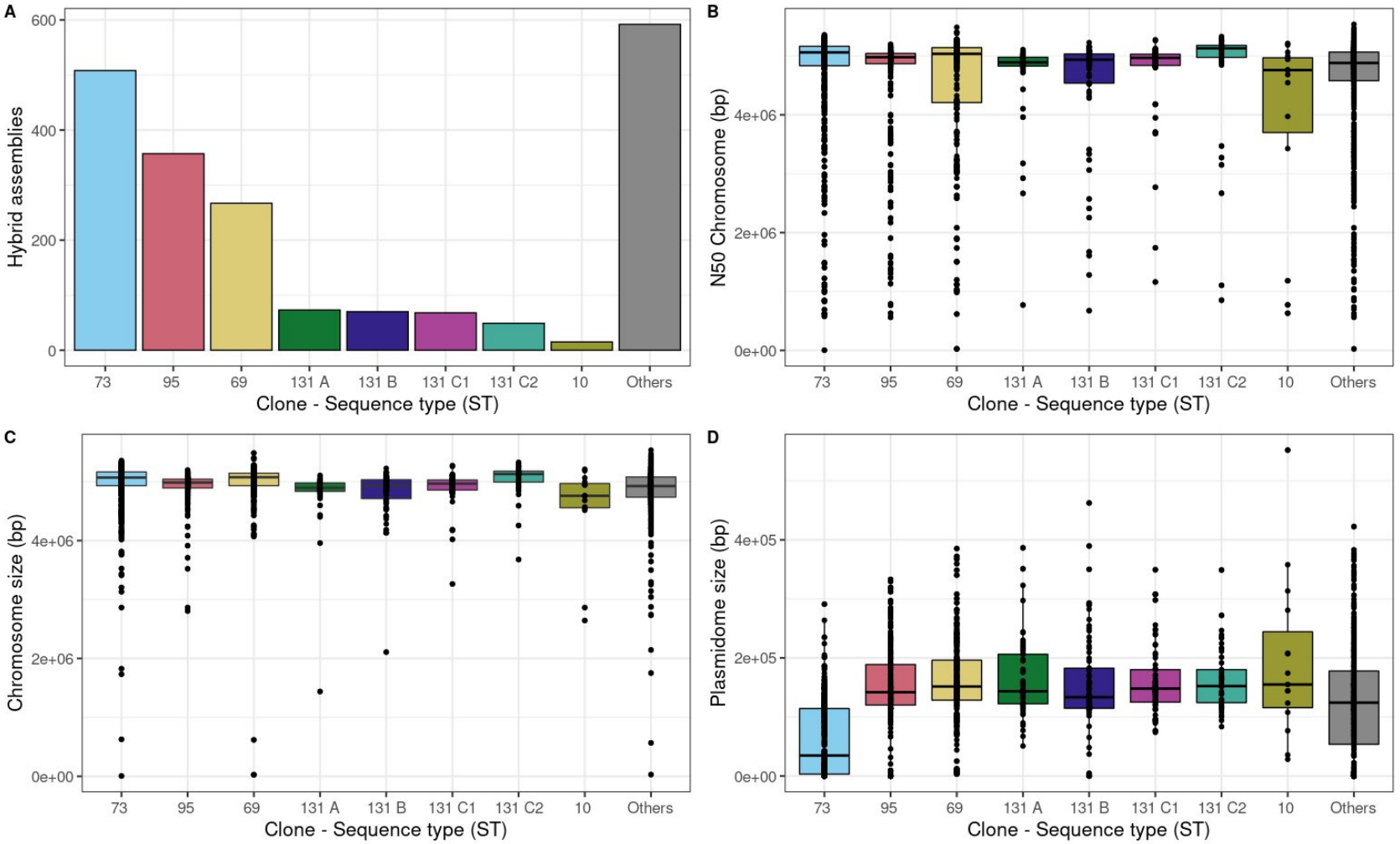
Supplementary Table S1. Report of the plasmid sequences (n=4,759) (4,174 circular and 585 non-circular) considered in this manuscript after quality control filtering. Each sequence is uniquely identified based on the ‘Sequence’ column. We provide the strain identifier carrying the plasmid (‘Genome’) together with its associated year of isolation, ST and (if applicable) ST131 clade. For each plasmid, we provide the operational mode of mge-cluster to type the sequences into the 23 non-overlapping groups (‘mge_cluster(pT)’) and their associated tsne coordinates (‘mge_cluster_tsne1D’ and ‘mge_cluster_tsne2D’). In addition, we report the replicons found in each plasmid by Plasmidfinder, their pmlst annotation (if applicable), the AMR genes and their class as reported by AMRFinderPlus, and the virulence genes found reported by Abricate using the ecoli_vf database (curated database of known virulence factors of *E. coli*).

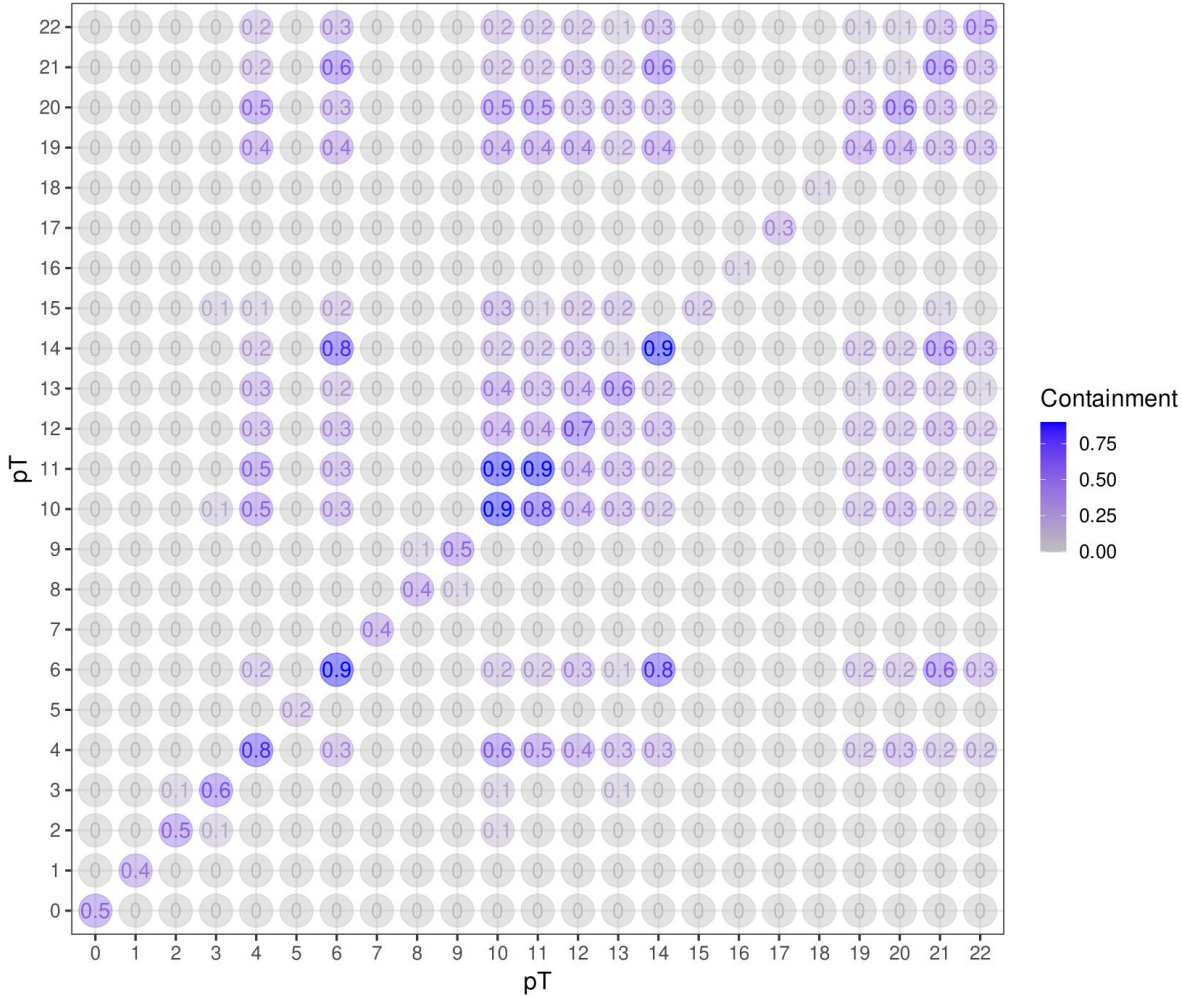
Supplementary Table S2. Summary of the plasmid characteristics inferred for each plasmid type (pT). The number of sequences assigned to each type is reported after filtering out sequences initially predicted to belong to it, but differing by more than 2 standard-deviation of the average plasmid length reported. The mean length reported was computed only considering the circular sequences for each type. The predominant replicon(s), pMLST and predicted mobility are

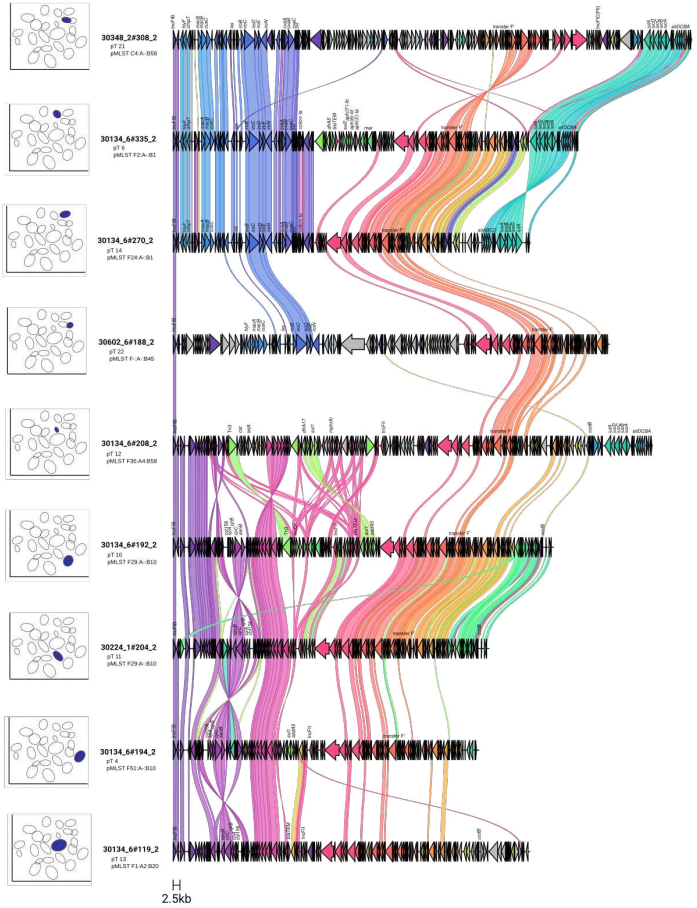
reported corresponding to a majority vote for the plasmids assigned to a given type. The list of virulence genes reported was uniquely based on the genes reported in the ecoli_vf database present in Abricate. For the complete overview and diversity analysis, we refer to Supplementary Table S1.

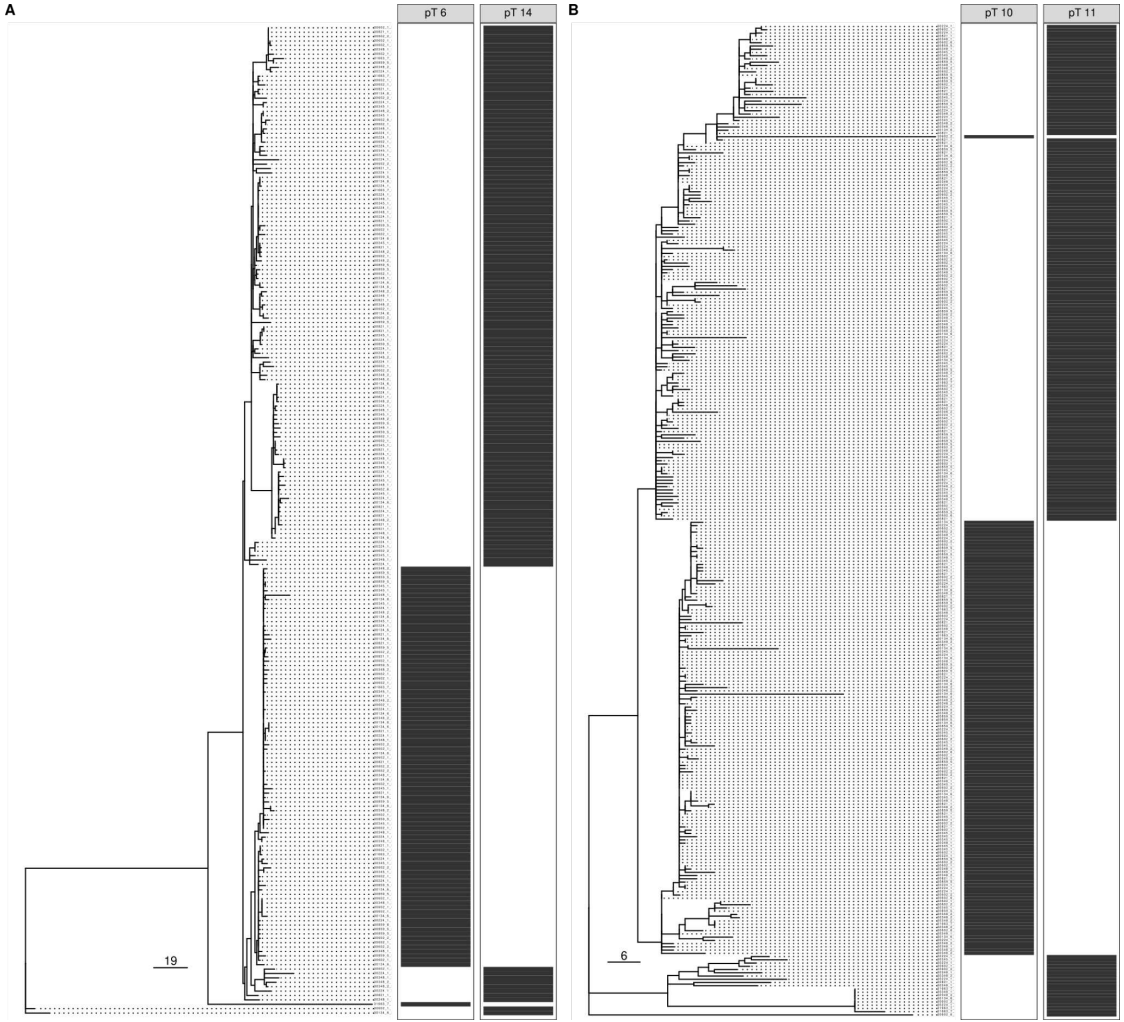
Supplementary Table S3. Acquisition dates of the most prevalent plasmid types among the four ExPEC clones. In the case of pT3 and pT14 (ST95), a single acquisition date is given since those were introduced by the MRCA of the clone. For each reported date, we provide the 95% confidence interval (CI) given by BactDating.

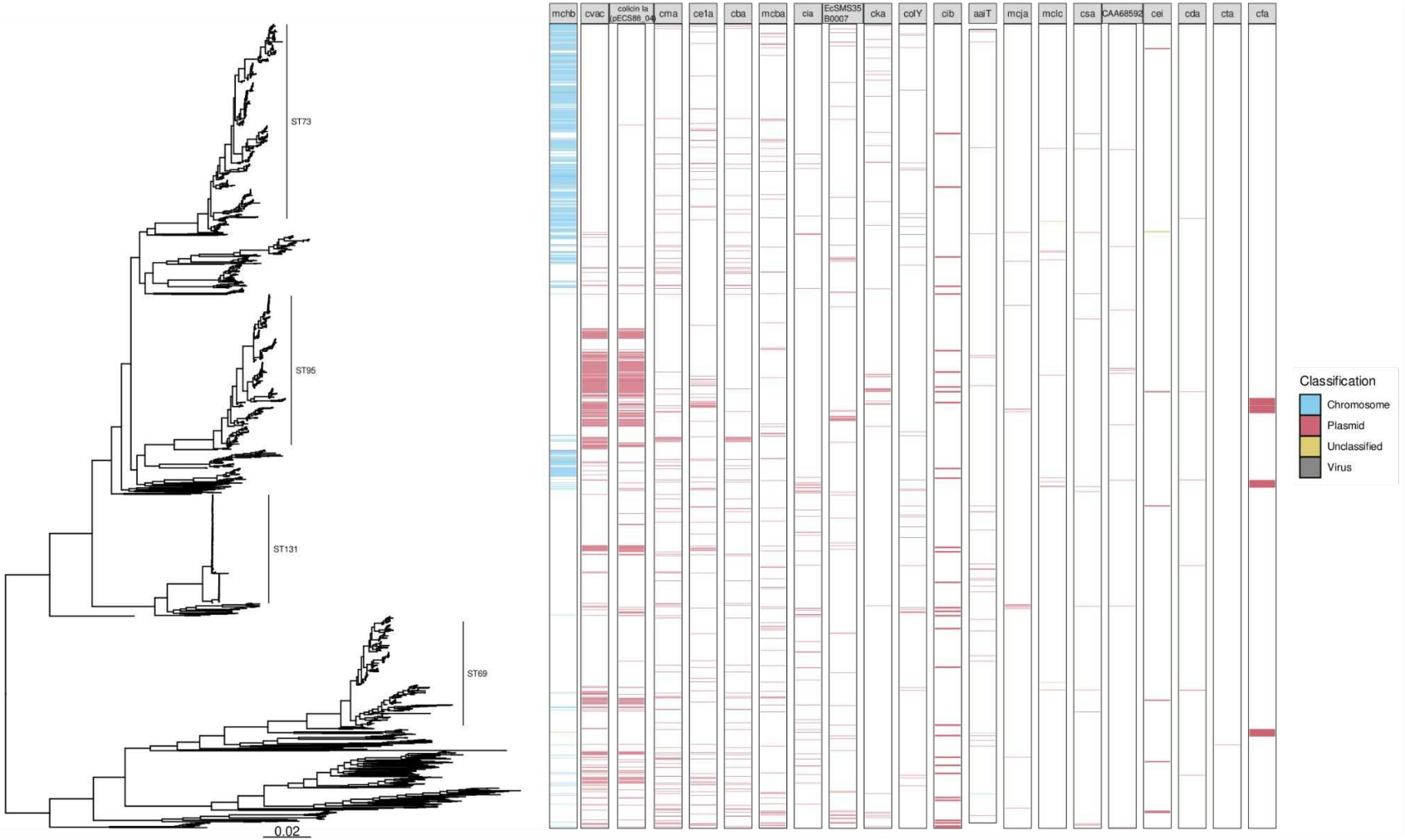
Supplementary Table S4. Strains used in experiments. Internal collection numbers and NORM sequence identifiers (or alternative identifiers for lab strains) are given for each strain. Features identified in NORM genome sequences (plasmid pColV, bacteriocin genes, siderophore systems) are indicated as presence (+) or absence (-). Experimentally tested sensitivity to microcin V (col V) and colicin Ia (col Ia) is indicated as sensitivity (S) or insensitivity (I).

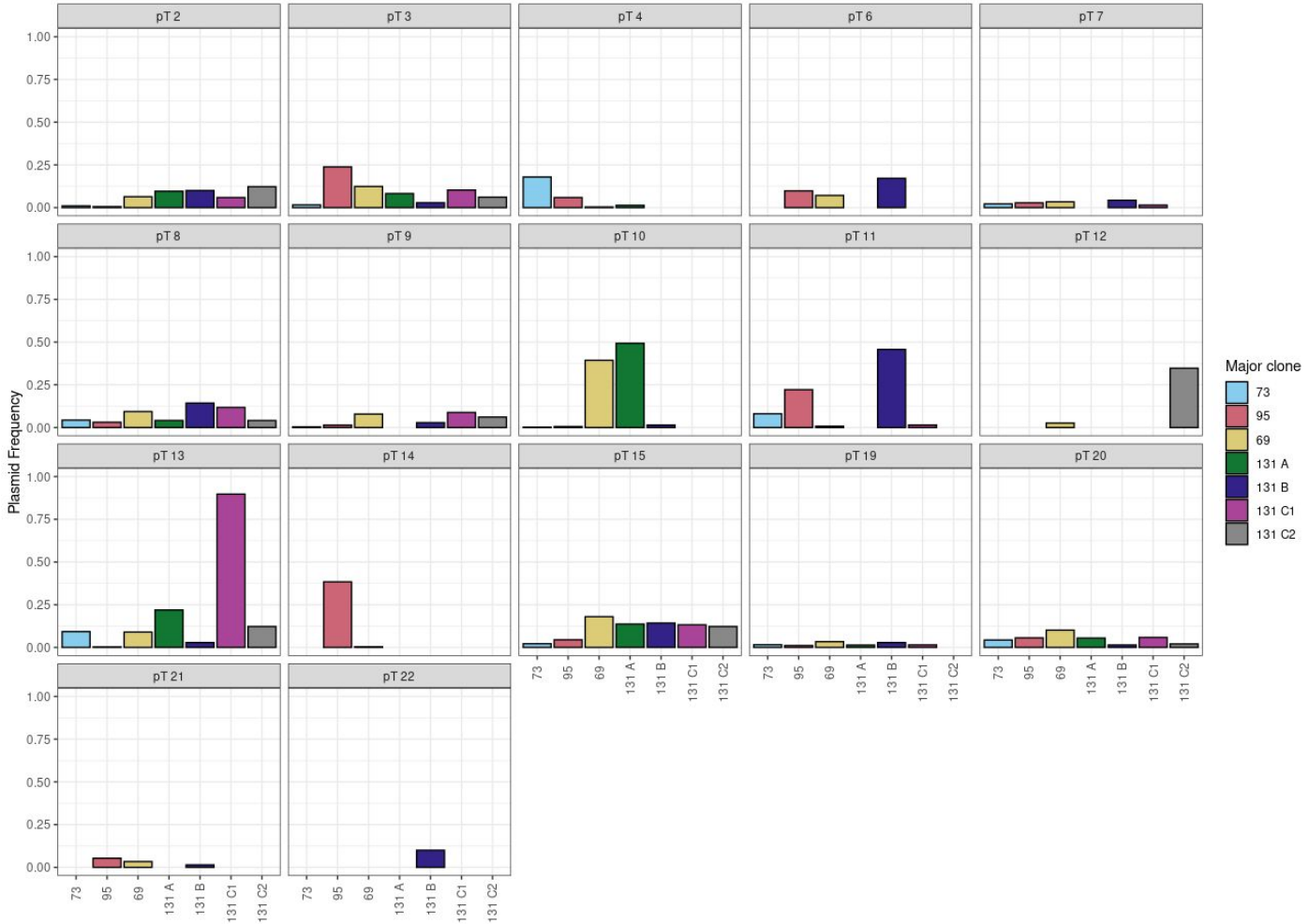

A
Hybrid assemblies
600
400
200
0
73
95
69
131 A
131 B
131 C1
131 C2
10
Others
Clone - Sequence type (ST)
B
N50 Chromosome (bp)
4e+06
2e+06
0e+00
73
95
69
131 A
131 B
131 C1
131 C2
10
Others
Clone - Sequence type (ST)
C
Chromosome size (bp)
4e+06
2e+06
0e+00
73
95
69
131 A
131 B
131 C1
131 C2
10
Others
Clone - Sequence type (ST)
D
Plasmidome size (bp)
4e+05
2e+05
0e+00
73
95
69
131 A
131 B
131 C1
131 C2
10
Others
Clone - Sequence type (ST)

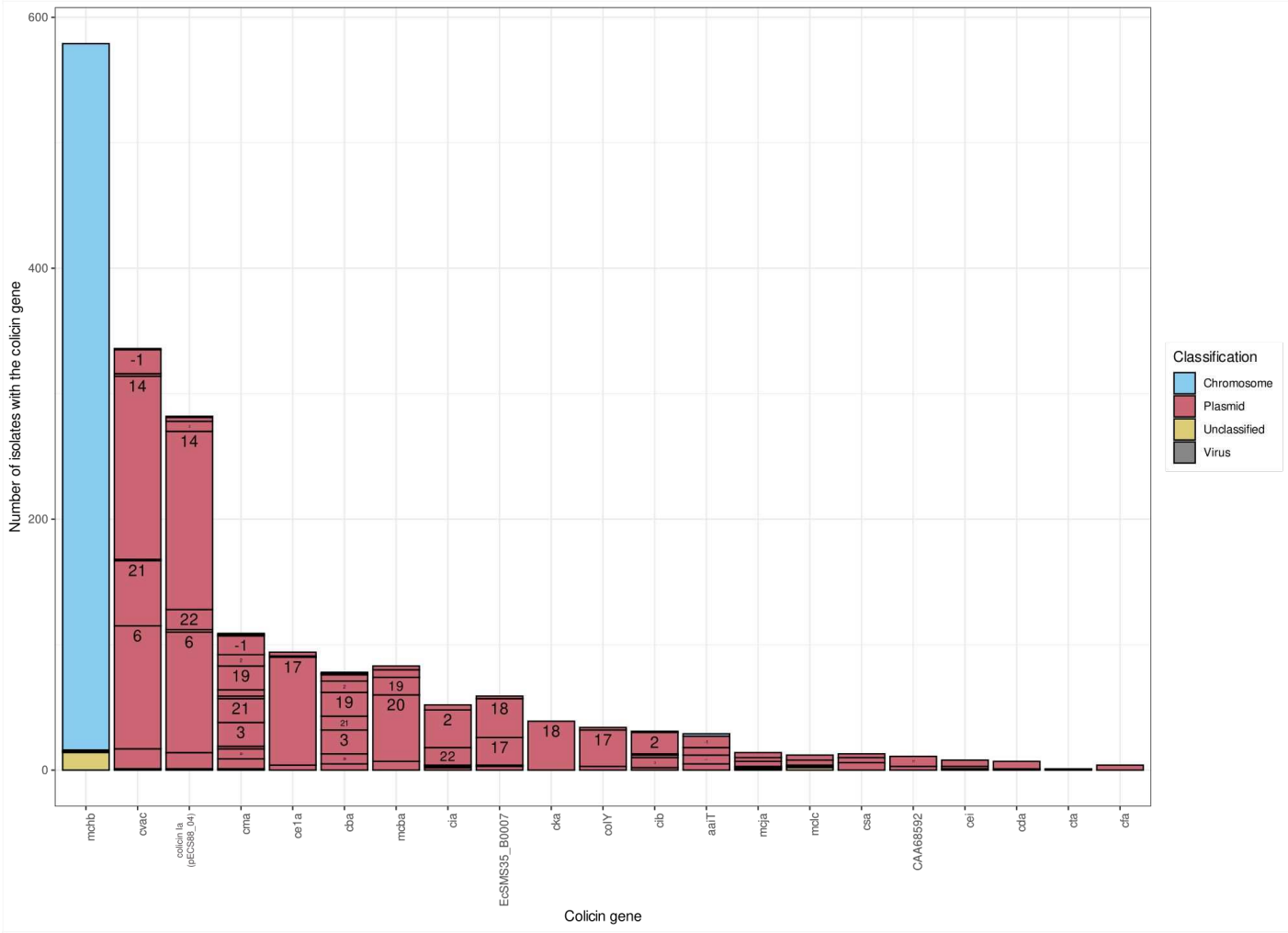

pT
pT
Containment
0.75
0.50
0.25
0.00

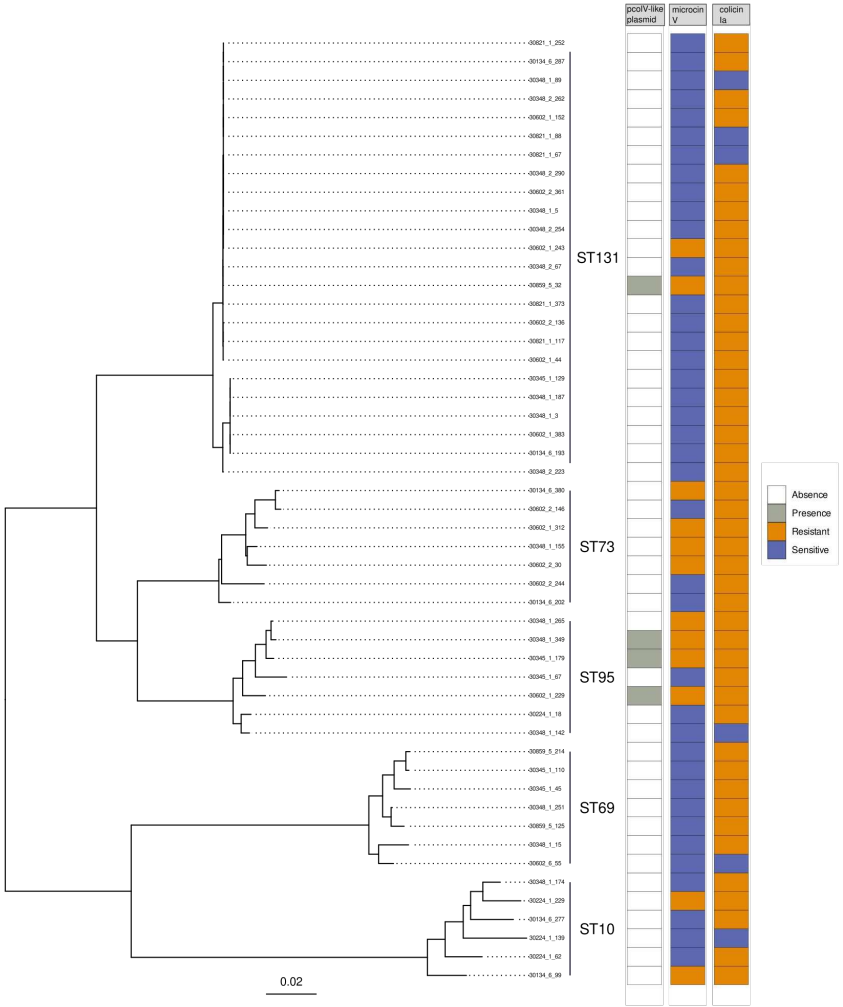

30348_2#308_2
pT 21
pMLST C4:A:-B58
30134_6#335_2
pT 6
pMLST F2:A:-B1
30134_6#270_2
pT 14
pMLST F24:A:-B1
30602_6#188_2
pT 22
pMLST F:-A:-B45
30134_6#208_2
pT 12
pMLST F36:A4:B58
30134_6#192_2
pT 10
pMLST F29:A:-B10
30224_1#204_2
pT 11
pMLST F29:A:-B10
30134_6#194_2
pT 4
pMLST F51:A:-B10
30134_6#119_2
pT 13
pMLST F1:A2:B20
H
2.5kb


A
pT 6
pT 14
19
B
pT 10
pT 11
6


mchb
cvac
colicin Ia (pECS88_04)
cma
ce1a
cba
mcba
cia
EcSMS35 B0007
cka
colY
cib
aaiT
mcja
mclc
csa
CAA68592
cei
cda
cta
cfa
ST73
ST95
ST131
ST69
0.02
Classification
Chromosome
Plasmid
Unclassified
Virus


pT 2
pT 3
pT 4
pT 6
pT 7
pT 8
pT 9
pT 10
pT 11
pT 12
pT 13
pT 14
pT 15
pT 19
pT 20
pT 21
pT 22
Plasmid Frequency
1.00
0.75
0.50
0.25
0.00
73
95
69
131 A
131 B
131 C1
131 C2
Major clone
73
95
69
131 A
131 B
131 C1
131 C2


Number of isolates with the colicin gene
600
400
200
0
Colicin gene
mchb
cvac
colicin Ia (pECS88_04)
cma
ce1a
cba
mcba
cia
EcSMS35_B0007
cka
colY
cib
aaiT
mcja
mdc
csa
CAA68592
cei
cda
cta
cfa
Classification
Chromosome
Plasmid
Unclassified
Virus


pcolV-like plasmid
microcin V
colicin Ia
ST131
ST73
ST95
ST69
ST10
Absence
Presence
Resistant
Sensitive
0.02