Frequent emergence of pathogenic lineages of *Klebsiella pneumoniae* via mobilisation of yersiniabactin and colibactin

Margaret M. C. Lam[1,2], Ryan R. Wick[1,2], Kelly L. Wyres[1,2], Claire Gorrie[1,2], Louise M. Judd[1,2], Sylvain Brisse[3], Adam Jenney[4], and Kathryn E. Holt[1,2].

Author affiliations:
[1]Centre for Systems Genomics, University of Melbourne, Parkville, Victoria, Australia, 3010.
[2]Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Parkville, Victoria, Australia, 3010.
[3]Microbial Evolutionary Genomics, Institut Pasteur, 75724, Paris, France.
[4]Department Infectious Diseases and Microbiology Unit, The Alfred Hospital, Melbourne, Victoria, Australia, 3004.

ABSTRACT:
Klebsiella pneumoniae (*Kp*) is a commensal bacterium that causes opportunistic infections. Evidence is mounting that *Kp* strains carrying acquired siderophores (yersiniabactin, salmochelin and aerobactin) and/or the genotoxin colibactin are highly pathogenic and can cause invasive disease. Here we explored the diversity of the *Kp* integrative conjugative element (ICE*Kp*), which mobilises the yersiniabactin locus *ybt*, by comparing 2499 diverse *Kp* genomes. We identified 17 distinct *ybt* lineages and 14 ICE*Kp* structural variants (some of which carry colibactin (*clb*) or salmochelin synthesis loci). Hundreds of ICE*Kp* transmission events were detected affecting hundreds of *Kp* lineages, including nearly >20 transfers into the globally-disseminated, carbapenem-resistant clonal group CG258. Additionally, we identify a plasmid-encoded lineage of *ybt*, representing a new mechanism for *ybt* dispersal in *Kp* populations. We introduce a novel sequence-based typing approach for identifying *ybt* and *clb* variants, to aid the identification of emerging pathogenic lineages and the convergence of antibiotic resistance and hypervirulence.

SIGNIFICANCE:
Klebsiella pneumoniae infections are increasingly difficult to treat with antibiotics. Some *K. pneumoniae* carry extra genes that allow them to synthesise yersiniabactin, an iron-scavenging molecule, which enhances their ability to cause disease. These genes are located on a genetic element that can easily transfer between strains. Here, we screened 2499 *K. pneumoniae* genome sequences and found substantial diversity in the yersiniabactin genes and the associated genetic elements, including a novel mechanism of transfer, and detected hundreds of distinct yersiniabactin acquisition events between *K. pneumoniae* strains. We also developed tools to identify and type yersiniabactin genes, to help track the evolution and spread of yersiniabactin in global *K. pneumoniae* populations and to monitor for acquisition of yersiniabactin in antibiotic-resistant strains.

INTRODUCTION

Multidrug-resistant *Klebsiella pneumoniae* (*Kp*) is one of the leading causes of healthcare-associated (HA) infections worldwide and poses significant treatment challenges. The production of carbapenemases and extended-spectrum beta-lactamases (ESBLs) are particularly problematic, and are often associated with clones such as sequence type (ST) 258 or 15 that cause hospital outbreaks (1–3). The prevalence of community-acquired (CA) invasive infections is also rising and is caused by *Kp* clones with enhanced pathogenicity such as ST23 (4–6) that are typically characterised by the presence of mobile genetic elements (MGEs) encoding siderophores, the genotoxin colibactin and the *rmpA* gene which contributes to the 'hypermucoid' phenotype by upregulating capsule production (4,7). CA *Kp* infections are also often associated with capsular serotypes that display greater serum resistance (K1, K2, K5), and are encoded by loci that can be transferred between *Kp* lineages via homologous recombination (8,9).

Siderophore systems comprised of iron-chelating molecules and associated receptors are common in *Kp* and other bacteria (10,11). They are considered integral to virulence as they allow bacteria to scavenge for iron – which is essential for growth – from host transport proteins, thereby enhancing the ability to survive and replicate within the host (12). Nearly all *Kp* produce the siderophore enterobactin. However its uptake mechanisms are inhibited by human lipocalin-2 (Lcn2), which has a strong binding affinity for ferric and aferric enterobactin (13) and induces an inflammatory response upon binding (14). The next most common siderophore in *Kp* is yersiniabactin, which escapes Lcn2 binding and also has iron-independent effects on virulence (14–16).

Yersiniabactin synthesis is encoded by the *ybt* locus. It was first described in the *Yersinia* 'high pathogenicity island', variants of which have since been identified in several Enterobacteriaceae species (17). In *Kp*, *ybt* is found on an integrative and conjugative element (ICE) that has been characterised in a few completely sequenced genomes (18–20). The ICE is self-transmissible, involving excision, formation of an extrachromosomal circular intermediate (requiring *int* and 17 bp direct repeats at the outer ends), mobilization to recipient cells (requiring *virB1, mobB* and *oriT*) and integration at *attO* sites present in any of four closely-located copies of tRNA-Asn in the *Kp* chromosome (18,21). The ICE sometimes includes loci for the synthesis of the siderophore salmochelin (*iro*) or the genotoxin colibactin (*clb*) (18,19,22). Colibactin has been shown to induce double strand breaks in human intestinal cells, and has been linked to colorectal cancer (23).

The mobility of the major virulence determinant *ybt* is highly concerning as it could theoretically be acquired by any *Kp* lineage, including those which are multidrug-resistant, leading to the emergence of new problematic clones. Yet detailed studies investigating the evolution, diversity and distribution of *ybt* in *Kp* has so far been limited. In this study, we address this lack by analysing 2499 *Kp* genomes to investigate the evolution of the ICEs (referred to hereafter as ICE*Kp*) responsible for

ybt and *clb* mobilization, and develop an approach for detecting and tracking these variants easily from genome data.

MATERIAL AND METHODS

Bacterial genome sequences.

We analysed a total of 2499 *Kp* genomes (2285 *Kp sensu stricto*, 63 *K. quasipneumoniae*, 146 *K. variicola,* 5 undefined or hybrid (4,7)) obtained from various sources representing a diverse geographical and clonal distribution (**Table 1**; see **Supplementary Table 1** for full list of isolates and their properties). Where available, Illumina short reads were analysed directly and assembled using SPAdes v3.6.1, storing the assembly graphs for further analysis of genetic context. Where reads were unavailable (n=921), publicly available pre-assembled contigs were used. These had been generated using various strategies and assembly graphs were not available for inspection.

An isolate from our collection (strain INF167, isolated from a patient at the Alfred Hospital, Melbourne, Australia in 2013) was subjected to further sequencing using a MinION Mk1B and R9 Mk1 flow cell (Oxford Nanopore Technologies). A 2D MinION library was generated from 1.5 µg purified genomic DNA using the Nanopore Sequencing Kit (SQK-NSK007). DNA was repaired (NEBNext FFPE RepairMix), prepared for ligation (NEBNextUltra II End-Repair/dA-tailing Module) and ligated with adapters (NEB Blunt/TA Ligase Master Mix). We sequenced the library for 48 hours, yielding 3862 reads (mean length 3049 bp, maximum 44026 bp) that were used to scaffold the SPAdes assembly graph using a novel hybrid assembly algorithm (http://github.com/rrwick/Unicycler). The resulting assembly included one circular plasmid, which was annotated using Prokka (35) and submitted to GenBank under accession TBA.

Multi-locus sequence typing (MLST) analysis

Genomes were assigned chromosomal sequence types by comparison to the *Kp* MLST scheme (36) in the *Kp* BIGSdb database (http://bigsdb.pasteur.fr/klebsiella/klebsiella.html) (24) using SRST2 to analyse reads (37) and BLAST+ to analyse assemblies. Alleles of genes belonging to the yersiniabactin (*ybtS, ybtX, ybtQ, ybtP, ybtA, irp2 irp1, ybtU, ybtT, ybtE, fyuA*) and colibactin (*clbABCDEFGHIJKLMNOPQR*) loci were determined by comparison to known alleles in the *Kp* BIGSdb database. Genomes were excluded from comparative analyses if at least one yersiniabactin allele could not be accurately determined due to data quality. Novel MLST schemes (38) were constructed for the yersiniabactin and colibactin loci, so that each observed combination of alleles was assigned a unique yersiniabactin sequence type (YbST, listed in **Supplementary Table 2**) or colibactin sequence type (CbST, listed in **Supplementary Table 3**). The schemes and allele sequences are available from the BIGSdb-*Kp* website and in the Kleborate repository

(https://github.com/katholt/Kleborate), which includes a command-line tool for genotyping new genomes.

Phylogenetic analysis of the siderophore loci and *K. pneumoniae* chromosome.

For each YbST, concatenated alignments of the corresponding allele sequences were produced using Muscle v3.8.31. Recombination events were identified and removed from the alignment using Gubbins v2.0.0 (39) and visualised using Phandango (https://github.com/jameshadfield/phandango/). Maximum likelihood (ML) trees were inferred from the post-Gubbins alignment by running RAxML v7.7.2 (40) five times with the generalised time-reversible (GTR) model and a Gamma distribution, selecting the final tree with the highest likelihood. The same approach was used to generate a colibactin ML tree.

A core genome SNP tree for clonal group (CG) 258 (which includes ST258, ST11, ST340 and ST512, among others) was inferred for a selection of representative isolates using the mapping pipeline RedDog v1b5 (https://github.com/katholt/reddog) to map short reads against *Kp* ST258 reference strain NJST258-1 (28) using Bowtie 2 v2.2.3, and identify core gene SNPs using SAMtools v1.1. The resulting SNP alignment was subjected to analysis with Gubbins v2.0.0 and RAxML v7.7.2 to infer a clonal phylogeny.

Chromosomal insertion sites and ICE structures.

For each *ybt*-positive (*ybt+)* genome, the annotated assembly was manually inspected to determine which of the four tRNA-Asn sites was occupied by ICE*Kp*. This was done with reference to the MGH78578 genome, which lacks any genomic islands at tRNA-Asn sites. The Artemis genome viewer was used to inspect the annotation of the region; BLAST+ was used for genome comparison; and when the region failed to assemble into a single contig, Bandage (41) was used to inspect the locus in the assembly graph where available. Once the insertion site was determined, the structure of the ICE*Kp* was inferred by extracting the sequence between the flanking direct 17 bp repeats 'CCAGTCAGAGGAGCCAA', either directly from the contigs using Artemis or from the assembly graph using Bandage. Representative sequences for each ICE*Kp* structure were annotated and deposited in GenBank (accessions TBA) and are included in the Kleborate repository (https://github.com/katholt/Kleborate).

RESULTS

Diversity of yersiniabactin genes in *K. pneumoniae*

A screen of 2499 genomes detected the *ybt* locus in 39.5% of *Kp* genomes, but only 2/146 *K. variicola* and 0/63 *K. quasipneumoniae*. Prevalence was 40.0% in CG258, 87.8% in the hypervirulent CG23, and 32.2% in the wider *Kp* population. Source information was available for 1341 human isolates and demonstrated a strong and statistically significant association between *ybt* and infection isolates (**Table 2**), particularly those from invasive infections (OR=33.4 for liver abscess, OR=5.6 for blood isolates), and also in respiratory (OR=3.4), urinary tract (OR=3.2) and wound infections (OR=3.3).

Next we explored the diversity of the eleven *ybt* locus genes using phylogenetic and MLST analyses. YbSTs defined by unique combinations of *ybt* gene alleles were successfully assigned to 834 *ybt*+ isolates (**Supplementary Table 1**). A total of 329 distinct YbSTs were identified; **Figure 1** shows their phylogenetic relationships (excluding a small number of recombination events, **Fig. S1**). The majority of YbSTs clustered into 17 lineages (referred to hereafter as *ybt* 1, *ybt* 2, etc) with 0.004 - 0.457% nucleotide divergence and a mean of eight shared loci within lineages compared to 0.032 - 1.127% nucleotide divergence and zero mean shared loci between lineages (**Fig. S2**).

ICE*Kp* structures and insertion sites

With the exception of *ybt* 4 (see below), the *ybt* locus was predominantly located within an ICE*Kp* structure that integrated into a chromosomal tRNA-Asn site. The four tRNA-Asn genes that serve as integration sites are located within a single chromosomal region, which is 16.4 kbp in size in strains that lack any MGE insertions at these sites (**Fig. 2**). Examples of ICE*Kp* integration were observed at all four sites (**Fig. 1**). Multiple ICE*Kp* integration sites were observed for most *ybt* lineages (**Fig. 1**); thus there is no evidence that distinct ICE*Kp*s preferentially integrate at specific sites. The frequencies of ICE*Kp* integration differed substantially by site: 35.7%, 44.7%, 19.5% for sites 1, 3 and 4 respectively, and just one integration at site 2.

The boundaries of each ICE*Kp* were identified by the 17 bp direct repeats formed upon integration. Each ICE*Kp* structure includes (i) a P4-like integrase *int* at the left end; (ii) the 29 kbp *ybt* locus; (iii) a 14 kbp sequence encoding the and *oriT* transfer origin, *virB*-type 4 secretion system (T4SS) and *mobBC* proteins (responsible for mobilisation) (18); and (iv) a distinct cluster of genes at the right end that we used to classify the ICE into 14 distinct structures (see **Fig. 3**, gene clusters are detailed in **Supplementary Table 4**). Further gene content variation within the ICE*Kp* can arise from transposases and other insertions or deletions. Most of the 14 ICE*Kp* structures were associated with a unique *ybt* lineage (**Figs. 1**, **3**; listed in **Supplementary Table 4**); the main exception was ICE*Kp10*, which carries a *clb* insertion and was associated with three *ybt* lineages (see below). BLAST searching NCBI for each ICE*Kp* structure detected only four occurring outside *Kp* (ICE*Kp3* and ICE*Kp4* in *E. coli*, ICE*Kp5* in

Enterobacter hormachei and ICE*Kp10* in *Citrobacter koseri* and *Enterobacter aerogenes*) and none outside Enterobacteriaceae.

All ICE*Kp* carried *int*, however we identified two variant forms of ICE*Kp* lacking the mobilisation genes. *Kp* subsp. *rhinoscleromatis* (ST67) genomes carried *ybt* 11 but lacked the entire mobilisation module. The 1979 strain NCTC11697 (novel ST) carried a highly divergent *ybt* locus (>2% nucleotide divergence from all other *ybt* sequences, **Fig. 1**) and its ICE lacked the *virB*-T4SS genes (**Fig. 3B**).

A ~34 kbp Zn^{2+} and Mn^{2+} metabolism module (KpZM) was found upstream of six different ICE*Kp* structures (most of ICE*Kp10*, *11* and 12; and a small subset of ICE*Kp2*, *4* and *5*; see **Fig. 3**). The KpZM module encodes a P4-like integrase at the left end that shares 97.5% amino acid identity with that of ICE*Kp*, and the same 17 bp direct repeat was found upstream of both integrases and downstream of ICE*Kp*. It is therefore likely that the entire sequence between the outer-most direct repeats (grey bars in **Fig. 3**) – including the KpZM module, *ybt* locus and variable region – can be mobilised together as a single ICE, and we refer to these structures as e.g. ICE*Kp2*-KpZM, to distinguish them from the forms that lack KpZM. Notably, the KpZM ICEs were clustered in the *ybt* sequence tree (*ybt* 1, *ybt* 12 – 13 and *ybt* 15 – 17; **Fig. 1**), suggesting that the KpZM was acquired in the ancestors of each of these three clusters, of which the latter two subsequently diversified into multiple ICE*Kp* structures by acquiring distinct gene modules at their right ends.

Two ICE*Kp* structures carried additional known *Kp* virulence factors. ICE*Kp1*, which was first described in ST23 strain NTUH-K2044 (18,42), carried *ybt* 2 genes and was one of only two ICE structures to have an additional gene cluster inserted between the *ybt* and mobilisation genes (see **Fig. 3**). As previously reported, this ~18 kbp insertion is homologous to a region on *Kp* plasmid pLVPK encoding *iro* as well as *rmpA* (which upregulates capsule production and is associated with hypermucoid phenotype) and other virulence determinants (18). ICE*Kp10* is characterised by the presence of the ~51 kbp colibactin (*clb*) module at its right end and associated with three distinct *ybt* lineages (1, 12 and 17; see **Fig. 1** and further details in the colibactin section below). The ICE*Kp10* structure corresponds to the genomic island described in ST23 strain 1084 as GM1-GM3 of genomic island KPHPI208 and in ST66 strain Kp52.145 as an ICE-Kp1-like region (19,20).

Plasmid-encoded yersiniabactin

No chromosomal insertion site could be identified in genomes carrying *ybt* 4. Inspection of the *de novo* assemblies of these genomes revealed that in all cases, contigs containing the *ybt* locus also harboured common *Kp* plasmid replicon sequences including FIB_K *repA,* FII_K *repB* and/or FIA *repE* (plasmid replication) and *sopAB* (plasmid partitioning) genes. It was not possible to resolve complete circular plasmid sequences from the short read assemblies, however inspection of the assembly graphs showed that the *ybt* 4-encoding contigs were disconnected from the chromosomal contigs, consistent with a plasmid location. To confirm this, we subjected one of the isolates (ST2370 strain INF167) to long-read sequencing using a MinION (Oxford Nanopore) device and resolved the complete sequence for a 165 kbp

FIB_K circular plasmid carrying *ybt* 4. Annotation of the complete *ybt*+ plasmid and the *ybt*+ contigs from the remaining isolates did not reveal any other genes with identifiable virulence or antimicrobial resistance (AMR)-related functions.

These results indicate that *ybt* 4 is typically plasmid-encoded in *Kp*, providing an alternative transfer mechanism between different *Kp* hosts. The *ybt* 4 sequences were distinct from those of other *ybt* lineages found in *Kp* (>0.28% nucleotide divergence; maximum 1 shared allele) (**Fig. 1**) and shared closer sequence identity with *ybt* genes found in *Yersinia* species (0.01% nucleotide divergence). The ICE*Kp* integrase and mobilisation genes were absent from the *ybt*+ plasmids and additional complete plasmid sequences will be required to resolve the mechanisms by which *ybt* 4 was acquired. The *ybt*+ plasmids were found in a variety of *Kp* hosts with distinct chromosomal backgrounds, indicative of plasmid transfer among Kp sublineages. Interestingly 20 (83%) came from isolates collected from patients at the Alfred Hospital (Melbourne, Australia; 2013-2014), belonging to 15 different chromosomal lineages, suggestive of transmission of the plasmid between locally co-circulating strains. The other *ybt*+ plasmids were found in isolates from Singapore (n=1, 2014), Australia (n=1, 2001) and the US (n=2, 2013 and 2014).

Colibactin diversity

Three distinct *ybt* lineages (1, 12 and 17) were associated with the *clb*-positive ICE*Kp*10 (**Fig. 1**), which was detected in 40% of ST258, 77% of ST23 and 4.0% of other *Kp* genomes including 25 other STs. Notably, all but three of the ICE*Kp10* strains carried the KpZM module at the left end, suggesting that the *clb* locus is usually mobilized within the larger ICE*Kp10*-KpZM. Sixty-five CbSTs were identified, similar to the number of YbSTs (n=86) detected in ICE*Kp10* (**Supplementary Table 3**). The *clbJ* and *clbK* genes were excluded from this analysis due to a common 4173 bp deletion, which results in a new open reading frame fusing the 5' end of *clbJ* with the 3' end of *clbK* (**Fig. 4A, Fig. S3**). Phylogenetic analysis of the *clb* locus revealed three lineages that were each associated with a different *ybt* lineage: *clb* 1 (*ybt* 12), *clb* 2A (*ybt* 1) and *clb* 2B (*ybt* 17) (**Fig. 4B**). The only exceptions were three isolates with *clb* 2B that had rare YbSTs not assigned to any lineage: ST258 strain UCI91 and ST48 strains WGLW1 and WGLW3. Two *ybt*- *clb*+ isolates were observed (both ST23). The corresponding *clb* loci clustered with those from the other ST23 ICE*Kp10* *ybt*+*clb*+ isolates and shared the same ICE*Kp10* integration site, suggesting a shared ancestral integration event in ST23 followed by subsequent loss of *ybt*. The *clbJ*/*clbK* deletion was detected sporadically in all *clb* lineages, suggesting it has arisen on multiple independent occasions and thus may be under positive selection (**Fig. 4B**).

ClbJ and *clbK* encode multi-domain proteins of 2166 and 2154 amino acids, respectively, whose functions are not yet characterised (**Fig. S3**). The deletion appears to be mediated by recombination between two copies of a 1480 bp stretch of homologous sequence that occurs with ~95% identity within the *clbJ* and *clbK* genes, which encodes an amino acid adenylation domain (A-domain) that is frequently a component of multi-domain non-ribosomal peptide synthetases. The fusion product

created by the *clbJ/clbK* deletion is a 2440 amino acid protein (**Fig. S3C**) that could potentially be functional, however its effect on colibactin synthesis is not yet known.

Frequency of yersiniabactin acquisition in *K. pneumoniae*

We identified at least 206 unique combinations of ICE*Kp* structure, insertion site and chromosomal ST, representing distinct *ybt* acquisition events. Twenty-six chromosomal STs showed evidence of multiple insertion sites and/or ICE*Kp* structures, indicating multiple independent acquisitions of ICE*Kp* within the evolutionary history of these clones (**Fig. 5**). Most unique acquisition events (65%) were identified in a single genome sequence. The greater the number of genomes observed per *Kp* ST, the greater the frequency of *ybt* carriage and unique *ybt* acquisitions per ST, suggesting that deeper sampling would continue to uncover further acquisitions (**Fig. 5**). Notably, of the 35 clonal groups that were represented by ≥10 genomes, 30 (86%) included at least one *ybt* acquisition (**Fig. 5**). Further, the five remaining clonal groups each consisted mostly of isolates from a localised hospital cluster (ST323, Melbourne; ST490, Oxford, ST512, Italy; ST681 Melbourne, ST874 Cambridge), so do not represent diverse sampling. Of the acquisition events that were detected in more than one genome, 68% (n=50/73) showed diversity in the YbST, consistent with clonal expansion of *ybt*-positive *Kp* strains and diversification of the *ybt* locus *in situ*. The greatest amount of YbST diversity within such groups was observed in hypervirulent clones ST23 (18 YbSTs of ICE*Kp10*/*ybt* 1 in site 1), ST86 (12 YbSTs of ICE*Kp3* in site 3) and ST67 *K. rhinoscleromatis* (5 YbSTs in site 1); followed by hospital outbreak-associated MDR clones ST15 (six YbSTs of ICE*Kp4* in site 1 and five in site 3), ST45 (five YbSTs of ICE*Kp4*), ST101 (five YbSTs of ICE*Kp3* in site 3) and ST258 (detailed below). This level of diversity suggests long-term maintenance of the ICE*Kp* in the genome, allowing time for diversification of the *ybt* genes.

Given the clinical significance of the carbapenemase-associated CG258, we explored ICE*Kp* acquisition in these genomes in greater detail. *Ybt* was detected in 269 isolates (40%) from 17 countries; 218 isolates also carried *clb* (nearly all from USA; see **Supplementary Table 1**). Fifty-eight YbSTs were identified amongst CG258 isolates and clustered into seven *ybt* lineages associated with six ICE*Kp* structures. Comparison of *ybt* lineage, ICE*Kp* structure and insertion site with a recombination-filtered core genome phylogeny for CG258 indicated dozens of independent acquisitions of ICE*Kp* sequence variants in this clonal complex (**Fig. 6**). Near-identical *clb* 2B (ICE*Kp10*/*ybt* 17) sequences were identified in 211 ST258, mostly at tRNA-Asn site 3, isolated from the USA during 2003–2014. Most of these isolates carried the *clbJ*/*clbK* deletion (n=175, 83%), and also transposase insertions within other *clb* genes (n=173) that may prevent colibactin production (**Fig. 6**). A total of 27 *ybt*+*clb*+ ST258 isolates had an apparently intact *clb* locus; two were isolated in Colombia in 2009 and the rest from USA during 2004–2010 (**Supplementary Table 1**), including KPNIH33 (43).

DISCUSSION

The yersiniabactin synthesis locus *ybt* was detected in over a third of all human *Kp* isolates, which is highly concerning given its role in virulence models (15,16,44). Our data strengthens the previously reported evidence that *ybt* is significantly associated with invasive infections in humans (7), such as liver abscess (OR=33.4, $p < 2x10^{-16}$) and bacteraemia (OR 5.6, $p < 4x10-^{15}$). The detection of significant associations with respiratory tract, urinary tract and wound infections (ORs 3.2-3.4, **Table 2**) indicate that even these classically opportunistic infections are more likely to occur if *ybt* is present.

While *ybt* was first identified in *Yersinia spp.* (45), the frequency and extensive sequence diversity of *ybt* and corresponding ICE*Kp* structures in the *Kp* population (**Figs. 1, 3**) reveals the locus is a long-standing and well-adapted component of the *Kp* accessory genome. The sheer number of distinct ICE*Kp* insertion events detected (n=214) is remarkable, and reveals that the high frequency of *ybt* in *Kp* is the result of highly dynamic processes. The benefits of gaining *ybt* are clear, as the ability to scavenge iron is essential to survival in iron-depleted conditions which are commonly encountered in a wide range of environmental and host-associated niches (12). However it appears that loss of the locus is also common, which could be due to the high-energy costs associated with synthesising the polyketide siderophore.

The identification of FIB_K plasmid-borne *ybt* constitutes an entirely novel mechanism for *ybt* transfer in *Kp*. The FIB_K plasmid replicon is very common (found in over half of all the *Kp* genomes we surveyed), and seems to be highly stable in *Kp* (46), suggesting these plasmids have the potential to rapidly transmit *ybt* within the *Kp* population. Indeed, the detection of *ybt* plasmids in 15 otherwise genetically-unrelated *Kp* isolates from a single hospital, as well as unrelated isolates from three other countries, shows it is a significant mechanism for *ybt* dissemination in *Kp*. Worryingly, FIB_K plasmids frequently carry AMR genes or virulence genes in *Kp* (47), suggesting there may be few barriers to convergence of AMR and virulence genes in a single FIB_K plasmid replicon, which could potentially pose a substantial public health threat and deserves careful monitoring.

The functional relevance of the genetic variation in the *ybt* and *clb* loci, as well as the cargo genes in the variable regions of the ICE*Kp*, remains to be explored. Experimental studies demonstrating the contribution of yersiniabactin and/or colibactin to virulence have been conducted with a limited number of strains (14,15,18–20), which we found to harbour either ICE*Kp1* or ICE*Kp10* and one of three *ybt* lineages (*ybt* 1, 2 or 12): NTUH-K2044 (ST23, ICE*Kp1*/*ybt* 2), ATCC 43816/KPPR1 (ST493, ICE*Kp1*/*ybt* 2), B5055/CIP 52.145 (ST66, ICE*Kp10*/*ybt* 12/*clb* 1) and 1084 (ST23, ICE*Kp10*/*ybt* 1/*clb* 2A). A systematic comparison of different *ybt* lineages, particularly the plasmid-borne lineage, in the same *in vivo* model might identify important differences in virulence potential. The genome collection analysed here is a convenience sample of available genome data that was originally generated for a variety of different purposes, however future genomic epidemiology studies of prospectively collected isolate collections may reveal

associations between distinct *ybt* or ICE*Kp* variants (identified using the tools developed here) and the extent of clinical risk.

Colibactin is known to be common in the hypervirulent liver abscess clone ST23 (5). The genotoxic property of colibactin has been experimentally demonstrated in strain 1084 (ST23, ICE*Kp10*/*ybt* 1/*clb* 2A) and may be associated with colorectal cancer (19,23,48). The presence of *clb* in 31 other *Kp* lineages, particularly the hospital-associated clone ST258, is therefore concerning and warrants further investigation. It remains to be determined whether colibactin is effectively synthesised by these strains, particularly those carrying the common *clbJ*/*K* deletion. Notably, most ST258 *clb*+ strains also carried transposase disruptions in the *clbB, clbH* and/or *clbO* genes, which likely interrupt colibactin synthesis and may represent selection against its production, which presumably carries a high metabolic cost to the host bacterium.

The extensive diversity uncovered amongst *ybt* and *clb* sequences and ICE*Kp* structures in this study provides several epidemiological markers with which to track their movements in the *Kp* population through analysis of whole genome sequence data, which is increasingly being generated for infection control and AMR surveillance purposes (49,50). The work presented here provides a clear framework for straightforward detection, typing and interpretation of *ybt* and *clb* sequences via the YbST and CbST schemes (**Figs. 1, 3**), which are publicly available and can be easily interrogated using our Kleborate package (https://github.com/katholt/Kleborate), the BIGSdb-*Kp* resource, or common tools such as BLAST or SRST2. Application of these tools in genomic surveillance will provide much-needed insights into the emergence and spread of pathogenic *Kp* lineages, and the convergence of virulence and AMR in this troublesome pathogen.

ACKNOWLEDGMENTS

This work was funded by the NHMRC of Australia (project #1043822 and Fellowship #1061409 to K. E. H).

REFERENCES

1. Arnold RS, Thom KA, Sharma S, Phillips M, Johnson JK, Morgan DJ (2012) Emergence of *Klebsiella pneumoniae* Carbapenemase (KPC)- Producing Bacteria. South Med J. 104(1):40–5.
2. Munoz-Price LS, Poirel L, Bonomo RA, Schwaber MJ, Daikos GL, Cormican M, et al (2013) Clinical epidemiology of the global expansion of *Klebsiella pneumoniae* carbapenemases. Lancet Infect Dis. 13:785–96.
3. Wyres KL, Holt KE (2016) *Klebsiella pneumoniae* Population Genomics and Antimicrobial-Resistant Clones. Trends Microbiol. 24(12):944–56.
4. Shon AS, Bajwa RPS, Russo TA (2013) Hypervirulent (hypermucoviscous) *Klebsiella pneumoniae*: a new and dangerous breed. Virulence. 4(2):107–18.
5. Struve C, Roe CC, Stegger M, Stahlhut SG, Hansen DS, Engelthaler DM, et al (2015) Mapping the evolution of hypervirulent *Klebsiella pneumoniae*. MBio. 6(4):1–12.
6. Brisse S, Fevre C, Passet V, Issenhuth-Jeanjean S, Tournebize R, Diancourt L, et al (2009) Virulent clones of *Klebsiella pneumoniae*: Identification and evolutionary scenario based on genomic and phenotypic characterization. PLoS One. 4(3).
7. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, et al (2015) Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. Proc Natl Acad Sci USA. 112(27):e3574-81.
8. Fung C, Hu B, Chang F, Lee S, Kuo BI, Ho M, et al (2000) A 5-Year Study of the Seroepidemiology of *Klebsiella pneumoniae*: High Prevalence of Capsular Serotype K1 in Taiwan and Implication for Vaccine Efficacy. J Infect Dis. 181(6):2075–9.
9. Lee IR, Molton JS, Wyres KL, Gorrie C, Wong J, Hoh CH, et al (2016) Differential host susceptibility and bacterial virulence factors driving *Klebsiella* liver abscess in an ethnically diverse population. Sci Rep. 6:29316.
10. Woodridge KG, Williams PH (1993) Iron uptake mechanisms of pathogenic bacteria. FEMS Microbiol Rev. 12(4).
11. Koczura R, Kaznowski A (2003) Occurrence of the Yersinia high-pathogenicity island and iron uptake systems in clinical isolates of *Klebsiella pneumoniae*. Microb Pathog. 35:197–202.
12. Holden VI, Bachman MA, Holden VI (2015) Diverging roles of bacterial siderophores during infection. Metallomics. Royal Soc Chemistry. 7:986–95.
13. Goetz DH, Holmes MA, Borregaard N, Bluhm ME, Raymond KN, Strong RK (2002) The neutrophil lipocalin NGAL is a bacteriostatic agent that interferes with siderophore-mediated iron acquisition. Mol Cell. 10(5):1033–43.
14. Bachman MA, Miller VL, Weiser JN (2009) Mucosal lipocalin 2 has pro-inflammatory and iron-sequestering effects in response to bacterial enterobactin. PLoS Pathog. 5(10).
15. Bachman MA, Oyler JE, Burns SH, Caza M, Lépine F, Dozois CM, et al (2011) *Klebsiella pneumoniae* yersiniabactin promotes respiratory tract infection through evasion of lipocalin 2. Infect Immun. 79(8):3309–16.
16. Holden VI, Breen P, Houle S, Dozois CM, Bachman MA (2016) *Klebsiella pneumoniae* Siderophores Induce Inflammation, Bacterial Dissemination, and HIF-1α Stablization during Pneumonia. MBio. 7(5):e01397-16-10.
17. Bach S, De Almeida A, Carniel E (2000) The Yersinia high-pathogenicity

island is present in different members of the family Enterobacteriaceae. FEMS Microbiol Lett. 183(2):289–94.

18. Lin TL, Lee CZ, Hsieh PF, Tsai SF, Wang JT (2008) Characterization of integrative and conjugative element ICE*Kp1*-associated genomic heterogeneity in a *Klebsiella pneumoniae* strain isolated from a primary liver abscess. J Bacteriol. 190(2):515–26.
19. Lai YC, Lin AC, Chiang MK, Dai YH, Hsu CC, Lu MC, et al (2014) Genotoxic *Klebsiella pneumoniae* in Taiwan. PLoS One. 9(5).
20. Lery LM, Frangeul L, Tomas A, Passet V, Almeida AS, Bialek-Davenet S, et al (2014) Comparative analysis of *Klebsiella pneumoniae* genomes identifies a phospholipase D family protein as a novel virulence factor. BMC Biol. 12(1):41.
21. Marcoleta AE, Berríos-Pastén C, Nuñez G, Monasterio O, Lagos R (2016) *Klebsiella pneumoniae* Asparagine tDNAs Are Integration Hotspots for Different Genomic Islands Encoding Microcin E492 Production Determinants and Other Putative Virulence Factors Present in Hypervirulent Strains. Front Microbiol. 7:1–17.
22. Putze J, Hennequin C, Nougayrède JP, Zhang W, Homburg S, Karch H, et al (2009) Genetic structure and distribution of the colibactin genomic island among members of the family Enterobacteriaceae. Infect Immun. 77(11):4696–703.
23. Vizcaino MI, Crawford JM (2015) The colibactin warhead crosslinks DNA. Nat Chem. 7(5):411–7.
24. Bialek-davenet S, Criscuolo A, Ailloud F, Passet V, Jones L, Garin B, et al (2014) Genomic Definition of Hypervirulent and Multidrug-Resistant *Klebsiella pneumoniae* Clonal Groups. Emerg Infect Dis. 20(11):1812–20.
25. Bowers JR, Kitchel B, Driebe EM, MacCannell DR, Roe C, Lemmer D, et al (2015) Genomic analysis of the emergence and rapid global dissemination of the clonal group 258 *Klebsiella pneumoniae* pandemic. PLoS One. 10(7):1–24.
26. Chung The H, Karkey A, Pham Thanh D, Boinett CJ, Cain AK, Ellington M, et al (2015) A high-resolution genomic analysis of multidrug-resistant hospital outbreaks of *Klebsiella pneumoniae*. EMBO Mol Med. 7(3):227–39.
27. Davis GS, Waits K, Nordstrom L, Weaver B, Aziz M, Gauld L, et al. (2015) Intermingled *Klebsiella pneumoniae* Populations between Retail Meats and Human Urinary Tract Infections. Vol. 61, Clin Infect Dis. 61:892–9.
28. Deleo FR, Chen L, Porcella SF, Martens C a, Kobayashi SD, Porter AR, et al (2014) Molecular dissection of the evolution of carbapenem-resistant multilocus sequence type 258 *Klebsiella pneumoniae*. Proc Natl Acad Sci USA. 111(13):4988–93.
29. Follador R, Heinz E, Wyres KL, Ellington MJ, Kowarik M, Holt KE, et al (2016) The diversity of *Klebsiella pneumoniae* surface polysaccharides. Microb Genomics. 2.
30. Onori R, Gaiarsa S, Comandatore F, Pongolini S, Brisse S, Colombo A, et al (2015) Tracking nosocomial *Klebsiella pneumoniae* infections and outbreaks by whole-genome analysis: Small-scale Italian scenario within a single hospital. J Clin Microbiol. 53(9):2861–8.
31. Stoesser N, Batty EM, Eyre DW, Morgan M, Wyllie DH, Del Ojo Elias C, et al (2013) Predicting antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole genomic sequence data. J Antimicrob Chemother. 68(10):2234–44.

32. Stoesser N, Giess A, Batty EM, Sheppard AE, Walker AS, Wilson DJ, et al (2014) Genome sequencing of an extended series of NDM-producing *Klebsiella pneumoniae* isolates from neonatal infections in a Nepali hospital characterizes the extent of community- versus hospital- associated transmission in an endemic setting. Antimicrob Agents Chemother. 58(12):7347–57.
33. Wand ME, Baker KS, Benthall G, McGregor H, McCowen JWI, Deheer-Graham A, et al (2015) Characterization of pre-antibiotic era *Klebsiella pneumoniae* isolates with respect to antibiotic/disinfectant susceptibility and virulence in Galleria mellonella. Antimicrob Agents Chemother. 59(7):3966–72.
34. Wyres KL, Wick RR, Gorrie C, Jenney A, Follador R, Thomson NR, et al. (2016) Identification of *Klebsiella* capsule synthesis loci from whole genome data. Microb Genomics.
35. Seemann T (2014) Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 30(14):2068–9.
36. Diancourt L, Passet V, Verhoef J, Grimont PAD, Brisse S (2005) Multilocus Sequence Typing of *Klebsiella pneumoniae* Nosocomial Isolates. 43(8):4178–82.
37. Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T, et al. (2014) SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Med. 6(11):90.
38. Maiden MC (2006) Multilocus sequence typing of bacteria. Annu Rev Microbiol. 60:561–88.
39. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al (2015) Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 43(3):e15.
40. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22(21):2688–90.
41. Wick RR, Schultz MB, Zobel J, Holt KE (2015) Bandage: interactive visualization of *de novo* genome assemblies. Bioinformatics. 31(June):3350–2.
42. Wu KM, Li NH, Yan JJ, Tsao N, Liao TL, Tsai HC, et al (2009) Genome sequencing and comparative analysis of *Klebsiella pneumoniae* NTUH-K2044, a strain causing liver abscess and meningitis. J Bacteriol. 191(14):4492–501.
43. Conlan S, Deming C, Tsai Y, Lau AF, Dekker JP, Korlach J, et al (2014) Complete Genome Sequence of a *Klebsiella pneumoniae* Isolate with Chromosomally Encoded Carbapenem Resistance and Colibactin Synthesis Loci. Genome Announc. 2(6):e01332-14.
44. Lawlor MS, O'Connor C, Miller VL (2007) Yersiniabactin is a virulence factor for *Klebsiella pneumoniae* during pulmonary infection. Infect Immun. 75(3):1463–72.
45. Carniel E (2001) The Yersinia high-pathogenicity island: an iron-uptake island. Microbes Infect. 3:561–9.
46. Lohr IH, Hülter N, Bernhoff E, Johnsen PJ, Sundsfjord A, Naseer U (2015) Persistence of a pKPN3-Like CTX-M-15- Encoding IncFII K Plasmid in a *Klebsiella pneumoniae* ST17 Host during Two Years of Intestinal Colonization. PLoS One. 10(3):e0116516.
47. Villa L, Garcıa-Fernandez A, Fortini D, Carattoli A (2010) Replicon sequence typing of IncF plasmids carrying virulence and resistance determinants. J Antimicrob Chemother. 65:2518–29.

48. Huang W, Chang JW, See L, Tu H, Chen J, Liaw C, et al (2012) Higher rate of colorectal cancer among patients with pyogenic liver abscess with *Klebsiella pneumoniae* than those without: an 11-year follow-up study. Color Dis Off J Assoc Coloproctology Gt Britain Irel. 14(12):e794-801.
49. Kwong JC, McCallum N, Sintchenko V, Howden BP (2015) Whole genome sequencing in clinical and public health microbiology. Pathology. 47(3):199–210.
50. Robilotti E, Kamboj M (2015) Integration of Whole-Genome Sequencing into Infection Control Practices: the Potential and the Hurdles. J Clin Microbiol. 53(4):1054–5.

Table 1. Description and sources of genome data used in this study

DATASET	N	NOTES	REF
Data from individual *Kp* genomics studies:			
Bialek *et al.*	36	K1/K2 study (multiple sites)	(24)
Bowers *et al.*	157	CG258 study (US)	(25)
Chung *et al.*	76	Outbreak in Patan Hospital (Kathmandu, Nepal)	(26)
Davis *et al.*	63	Isolates from retail meats and human UTIs (US)	(27)
Deleo *et al.*	69	CG258 study (US)	(28)
Ellington *et al.*	193	Addrenbrookes Hospital (Cambridge, UK)	(29)
Holt *et al.*	273	Global diversity study (multiple sites)	(7)
Lee *et al.*	26	PLA study (Singapore)	(9)
Onori *et al.*	16	ST258 study at the Circolo Hospital and Macchai Foundation hospital (Varese, Italy)	(30)
Stoesser *et al.* (2013)	69	Isolates from John Radcliffe Hospital (Oxford, UK)	(31)
Stoesser *et al.* (2014)	54	Outbreak in Patan Hospital (Kathmandu, Nepal)	(32)
Struve *et al.*	67	ST23 study (multiple sites)	(5)
Wand *et al.*	33	Pre-antibiotic era strains isolated1917-1949 (UK, Murray Collection)	(33)
Wyres *et al.*	487	Diverse hospital isolates (Australia)	(34)
Data from genome databases:			
NCTC3000	83	Public Health England's National Collection of Type Cultures (NCTC)	phe-culturecollections.org.uk/collections/nctc.aspx
PATRIC	797	Accessed Feb 1, 2016	patricbrc.org

Table 2. Frequency of yersiniabactin locus (*ybt*) in *K. pneumoniae* isolated from humans. Carriage, isolates recorded as associated with asymptomatic carriage as opposed to infection; N, number of isolates; *ybt*, number carrying *ybt*. Statistics reported are odds ratio (OR), 95% confidence interval (CI) and p-value for association between *ybt* and infections of different types vs asymptomatic carriage, calculated using Fisher's exact test.

				Infection vs Carriage		
	N	***ybt***	**(%)**	**OR**	**95% CI**	**p-value**
Carriage	190	25	13%	-	-	-
Infection						
Liver abscess	68	57	84%	33.4	15.0 – 80.8	2×10^{-16}
Blood	325	149	46%	5.6	3.4 – 9.4	4×10^{-15}
Respiratory	237	80	34%	3.4	2.0 – 5.8	9×10^{-7}
Urine	497	164	33%	3.2	2.0 – 5.4	6×10^{-8}
Wound	51	17	33%	3.3	1.5 – 7.1	2×10^{-3}

FIGURES:

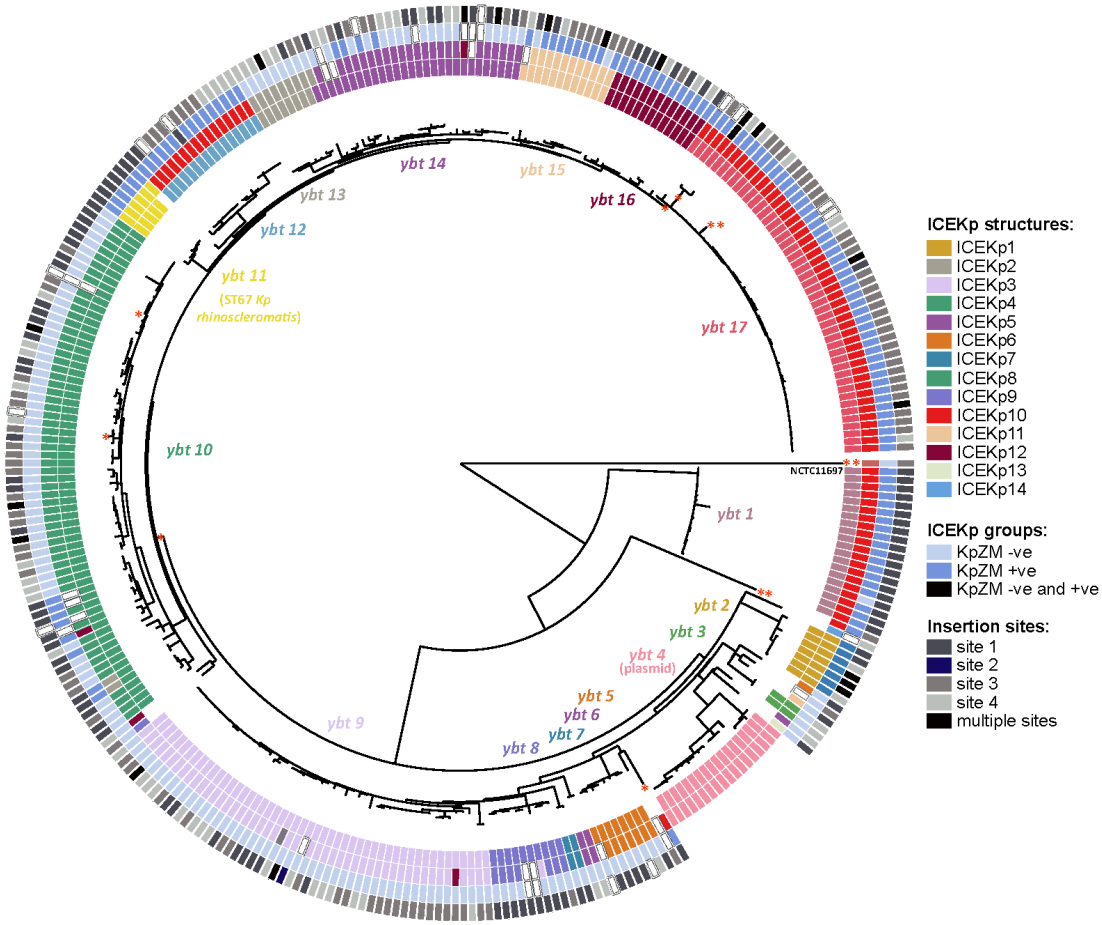

Figure 1. Recombination-filtered phylogenetic analysis of 329 yersiniabactin sequence types identified across 834 genomes. Each leaf represents a single yersiniabactin sequence type (YbST) and these YbST sequences cluster into 17 lineages, as labelled. Tracks (from inner to outer): (1) lineage (key as labelled above the tree nodes; white = unassigned), (2) ICE*Kp* structure (white = undetermined), (3) presence or absence of KpZM module, and (4) tRNA-Asn insertion site (white = undetermined). Recombination events (see **Fig. S1**) are depicted with a red asterisk next to the relevant branches or a single YbST.

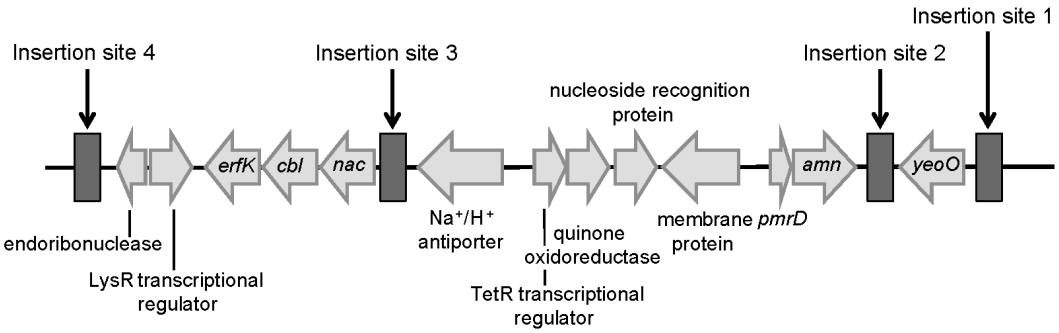

Figure 2. Chromosome region in *K. pneumoniae* containing tRNA-Asn sites that are targeted by yersiniabactin ICE*Kp* and other genomic islands. The hotspots for ICE*Kp* insertion occur within four tRNA-Asn sites, represented by the rectangular blocks, and are marked in the figure. Grey arrows represent coding sequences, labelled by gene symbol or the encoded protein product.

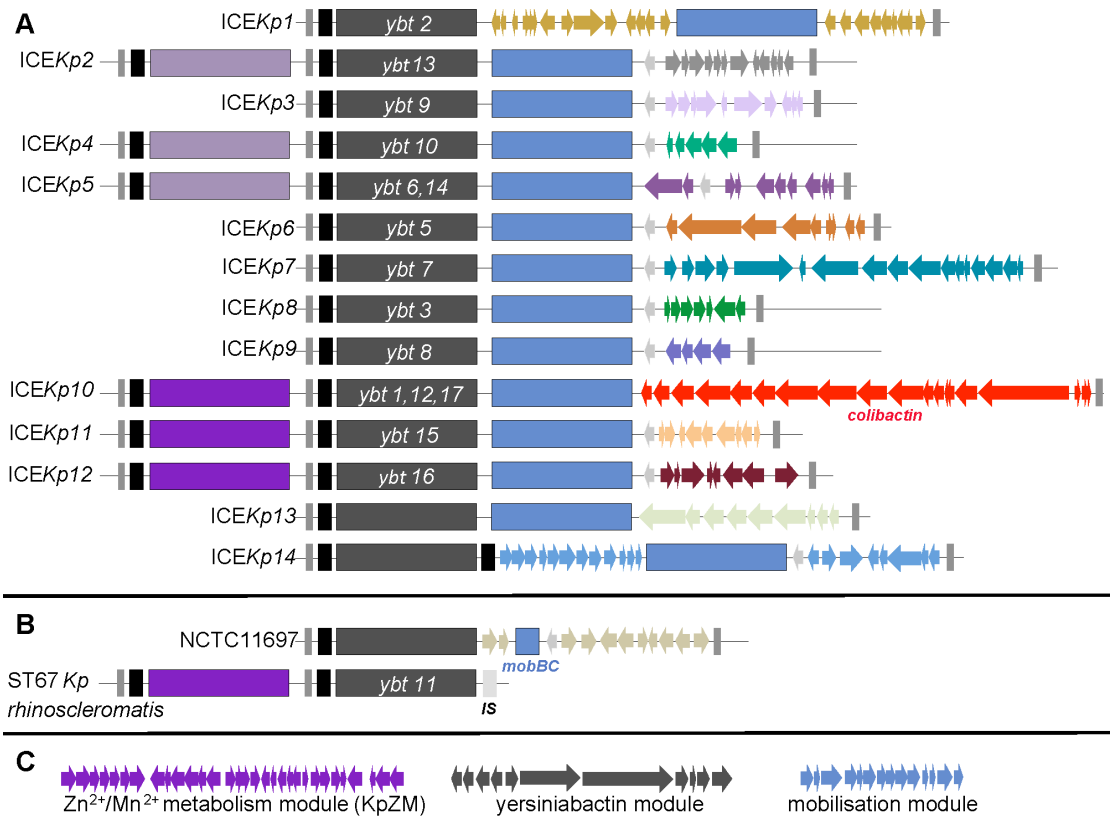

Figure 3. ICE*Kp* structures. (A) Genetic structures of apparently intact ICE*Kp* variants (see Supplementary Table 4 and GenBank deposited sequences for details of specific genes). **(B)** Disrupted ICE*Kp* loci. **(C)** Gene structures for core modules, which are shown in **A-B** as coloured blocks: yersiniabactin synthesis locus *ybt* (dark grey, labelled with the most commonly associated *ybt* lineage if one exists), mobilisation module (blue) and Zn^{2+}/Mn^{2+} module (purple = usually present, light purple = rarely present). In panels A-B, the variable gene content unique to each ICE*Kp* structure, which is typically separated from the mobilisation module by an antirestriction protein (light grey arrow), is shown in a unique colour as per **Figure 1**. Grey rectangles represent direct repeats; black rectangles, P4-like integrase genes.

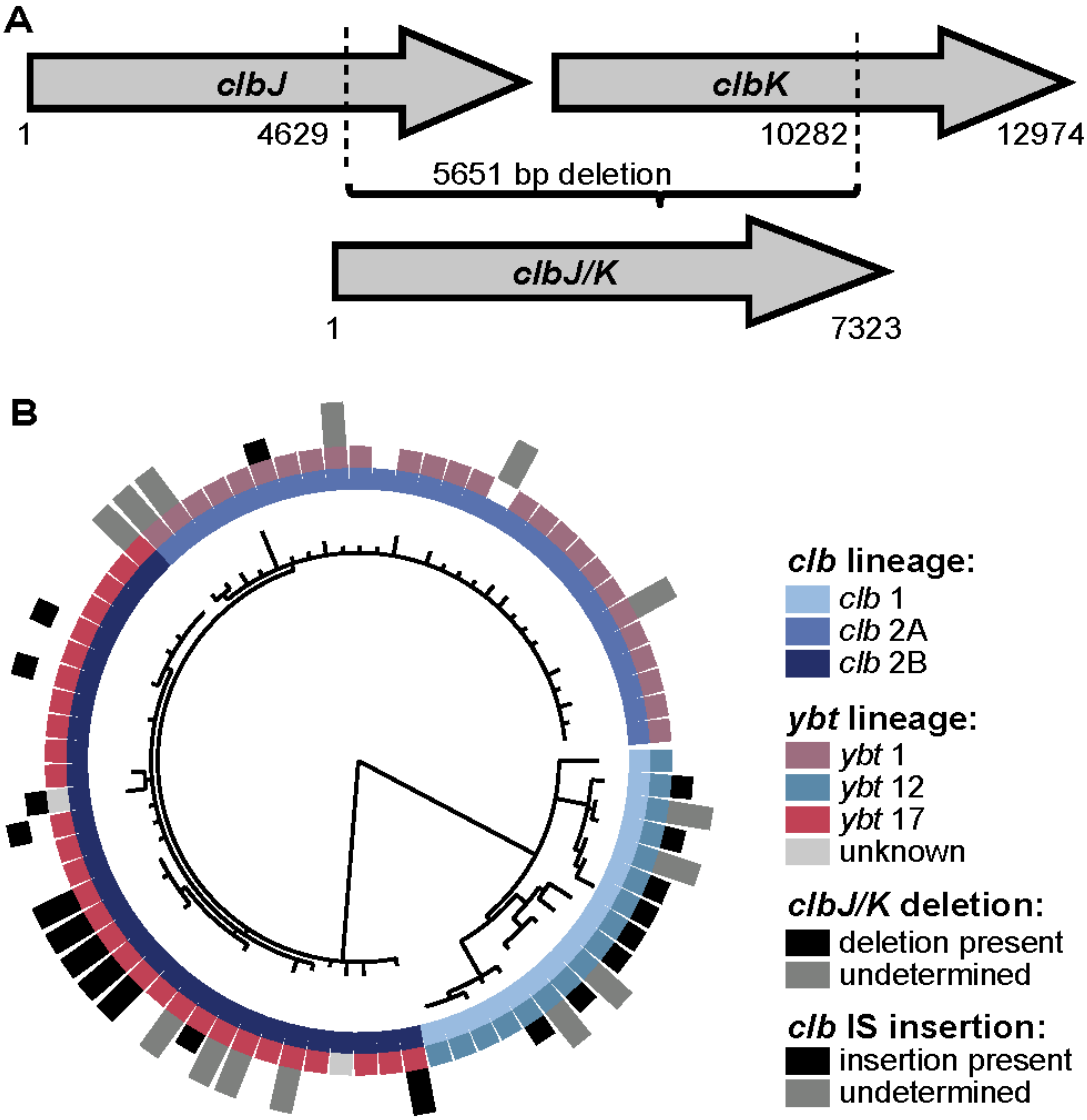

Figure 4. Colibactin diversity. (A) *ClbJ/K* deletion. A 5.6 kbp in-frame deletion between the *clbJ* (6501 kbp) and *clbK* (6465 kbp) genes of the colibactin locus observed in a large number of ICE*Kp10* structures. **(B)** Recombination-free phylogeny of genes in the colibactin locus (excluding *clbJ* and *clbK*). Each leaf represents a unique colibactin locus found across 314 isolates. Tracks (from inner to outer): (1) *clb* lineage, (2) *ybt* lineage (grey = not part of any main *ybt* lineage, white = *ybt*-negative), (3) presence of *clbJ/K* deletion (white = no deletion), and (4) occurrence of intragenic transposase insertions within *clb* locus (white = no insertion).

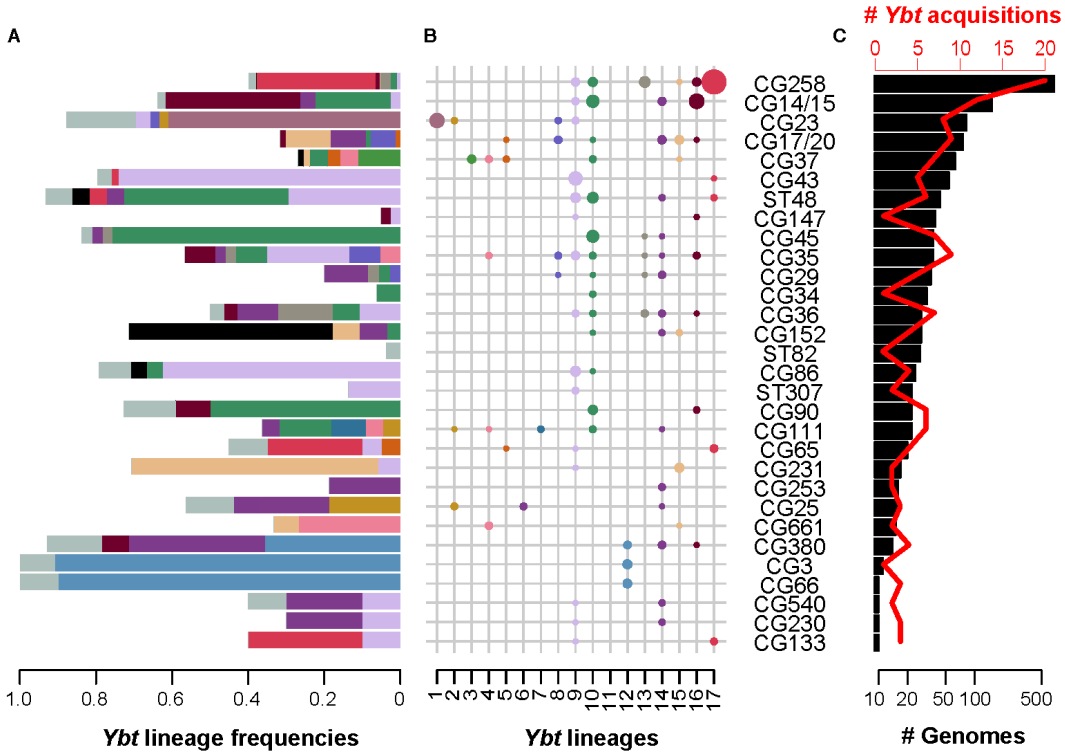

Figure 5. Frequency and diversity of *ybt* sequences and acquisition events in common *Kp* clonal groups. This analysis includes all clonal groups (labelled on the left) for which there were ≥10 genomes available. **(A)** Frequency of *ybt* presence within each clonal group, coloured by *ybt* lineage according to the scheme in **Figures 1, 3** and panel **B** (white = proportion of genomes without *ybt*). **(B)** Number of *ybt*-positive *Kp* genomes designated to a particular *ybt* lineage. The sizes correspond to relative sample size of genomes. **(C)** Number of genomes (black bars, bottom axis) and number of independent *ybt* acquisition events (red, top axis) per clonal group. Independent *ybt* acquisition events were defined as unique combinations of ICE*Kp* structure and insertion site in each ST.

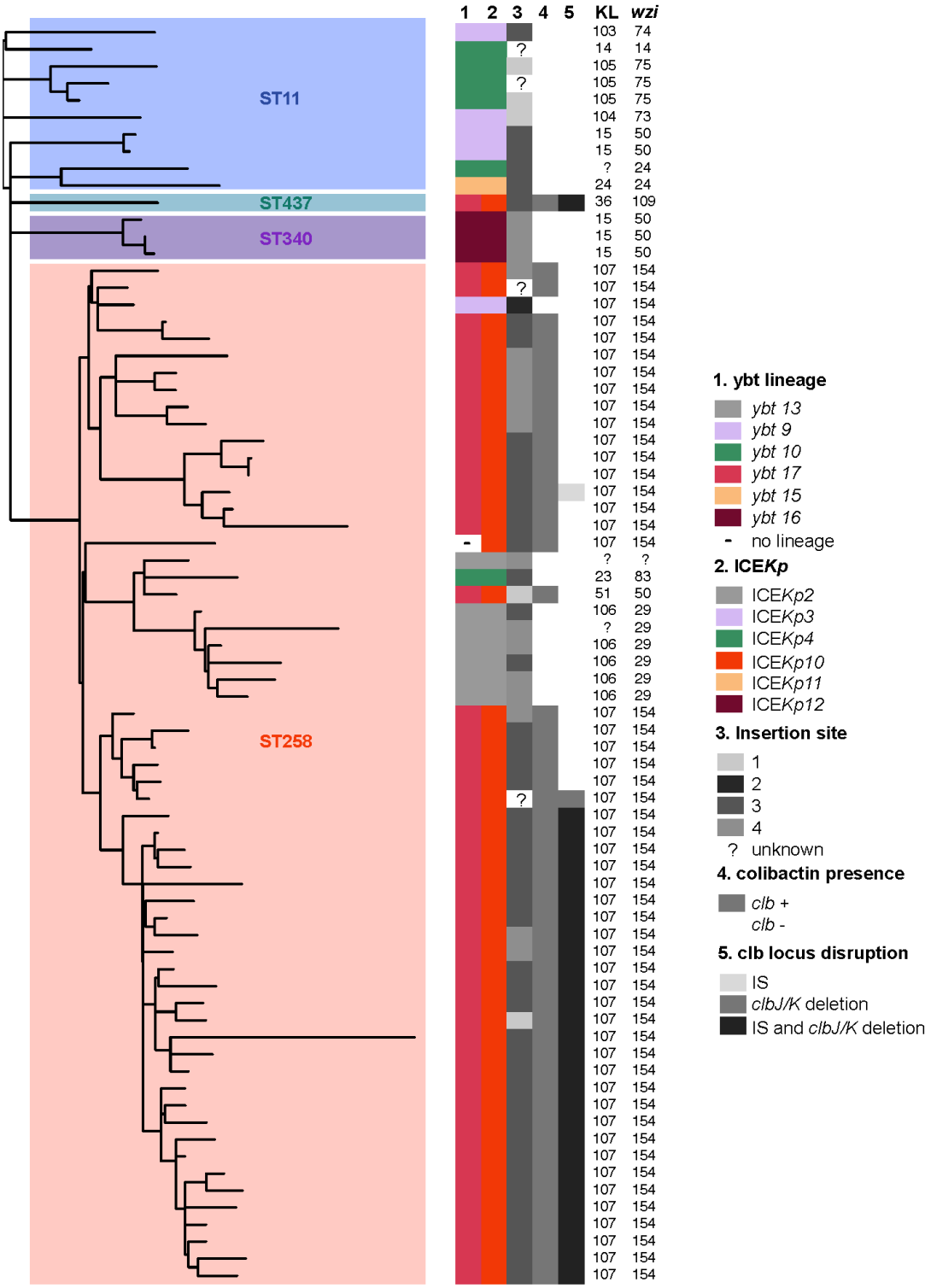

Figure 6. Diversity of yersiniabactin loci acquired by CG258. The tree shown is an outgroup-rooted maximum likelihood phylogeny inferred from recombination-filtered core gene SNPs for all ybt-positive CG258 isolates. Clades are labelled by chromosomal MLST; tracks indicate (1) *ybt* lineage, (2) ICE*Kp* structure, (3) ICE*Kp* insertion site, (4) *clb* presence/absence and (5) *clb* locus disruptions, coloured according to the inset legend.

SUPPLEMENTARY MATERIAL:

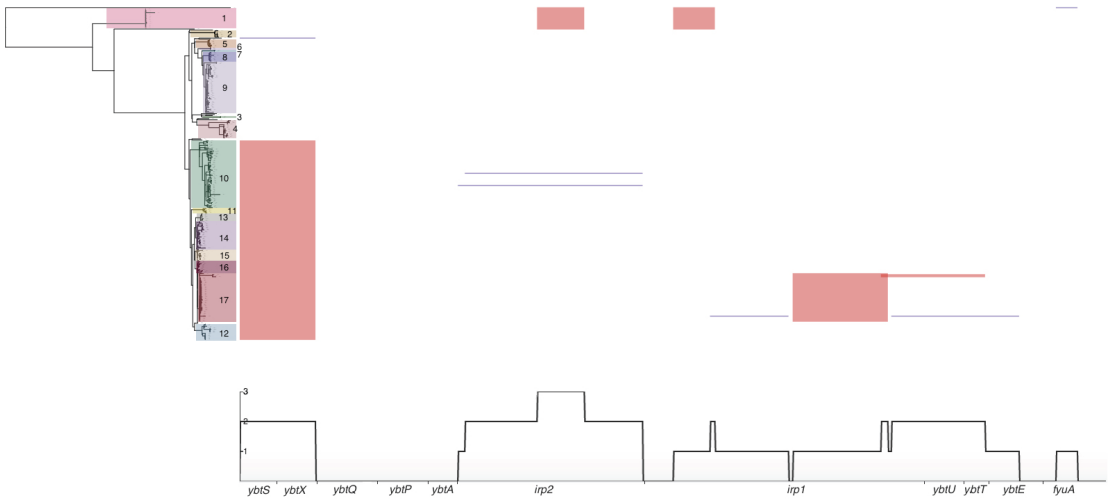

Figure S1. Predicted recombination events in the yersiniabactin locus. Recombination events were predicted by Gubbins and are shown as coloured blocks (visualised using Phandango). Coordinates along the *ybt* locus and gene boundaries are indicated on the x-axis. Each row in the plotting area represents a YbST. Phylogenetic relationships between the YbSTs are shown in the tree to the left, which is a midpoint-rooted, recombination-free YbST phylogeny reproduced from **Figure 1**. Colours and numbers on the tree indicate *ybt* lineages as detailed in the text and **Figure 1**.

Figure S2. Minimum spanning tree of YbSTs, visualised using PhyloViz. Each node represents a YbST, connections between the nodes indicate allele sharing between YbSTs; nodes are coloured by *ybt* lineages (as defined in **Figure 1**; black indicates no lineage assigned) and labelled with ICE*Kp* structures (as defined in **Figure 2**; the *clb*+ ICEKp10 structure, boxed, is associated with three *ybt* lineages).

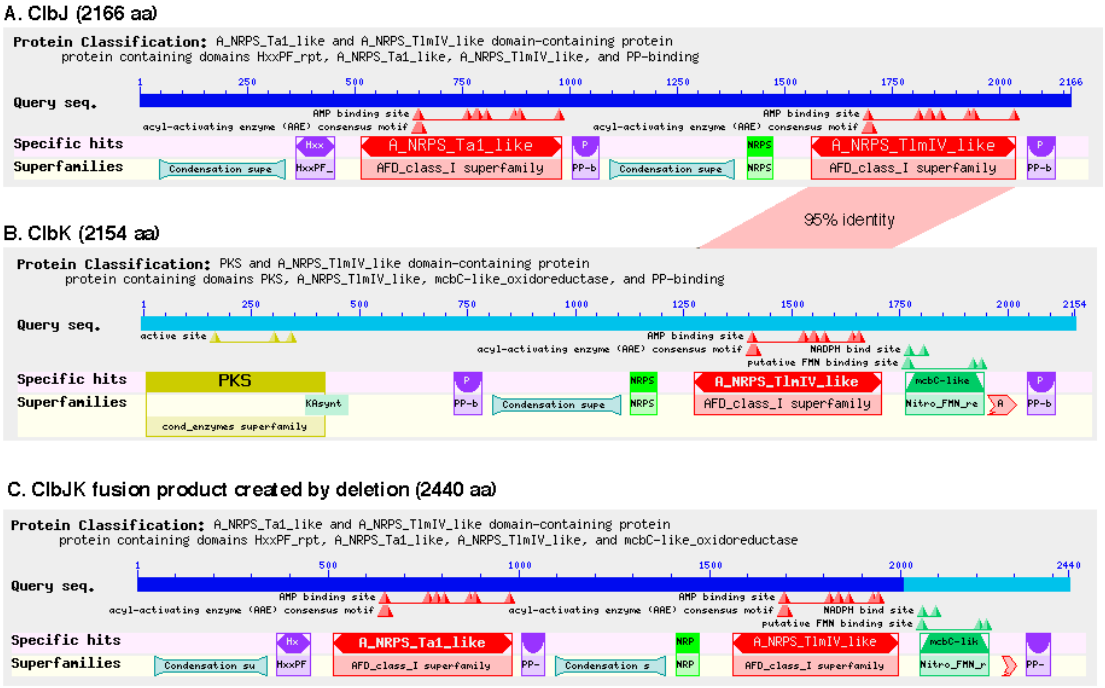

Figure S3. Conserved domains present in the predicted proteins encoded by (A) *clbJ*, (B) *clbK* and (C) the *clbJ/K* deletion. The homologous region shared between the amino acid adenylation domains of *clbJ* and *clbK* is shown. The left hand side and a large portion of the fusion product matches to *clbJ* (dark blue) while the right hand side matches to *clbK* (light blue).

Supplementary Table 1. Isolates used in this study.
[see supplementary document titled Supplementary_Table_1_isolates.csv]

Supplementary Table 2. Yersiniabactin sequence types (YbSTs) and corresponding alleles.
[see supplementary document titled Supplementary_Table_2_YbSTs.csv]

Supplementary Table 3. Colibactin sequence types (CbSTs) and corresponding alleles.
[see supplementary document titled Supplementary_Table_3_CbSTs.csv]

Supplementary Table 4. Description of ICE*Kp* variants.

ICE*Kp*	YbST lineages	Length (kbp)	Unique genes in variable regions*	Comments	Accession number
ICE*Kp1*	2	76	**Middle region:** Virulence associated proteins VagC, VagD, Salmochelin genes *iroN, iroB, iroC, iroD*, Drug/metabolite transporter permease. LuxR family transcriptional regulator. Regulator of mucoid phenotype *rmpA*. SAM-dependent methyltransferase. 3 Transposases. 2 Hypothetical proteins. **3' region:** Thiamine biosynthesis protein ThiF, DNA binding protein, 4 Hypothetical proteins	- First described in *Kp* strain NTUH-K2044 (Lin *et al.* 2008) - Referred to as a 'Group IV' genomic island (Marcoleta *et al.* 2016)	TBA
ICE*Kp2*	10, 13, 14	62	Thymidylate synthase, Adenylate kinase, TIR domain protein, 9 Hypothetical proteins		TBA
ICE*Kp3*	8, 9	65	Restriction endonuclease, DUF4917 domain containing protein, ATP/GTP	- BLAST: 99% identity to *E. coli* Co6114 - Referred to as a 'Group VI' genomic	TBA

			phosphatase, Reverse transcriptase, DDE endonuclease, 5 Hypothetical proteins	island (Marcoleta *et al*. 2016)	
ICE*Kp4*	10	58	Transposase, ABC transporter, Type I restriction endonuclease, DNA methyltransferase, Hypothetical protein	- BLAST: 99% identity to *E. coli* ED1a	TBA
ICE*Kp5*	6, 14	66	DEAD/DEAH box helicase, Thiamine biosynthesis protein ThiF, 2 Patatin-like phospholipases, 6 Hypothetical proteins	- BLAST: 99% identity to *Enterobacter hormachei* 05-545 - Referred to as a 'Group V' genomic island (Marcoleta *et al*. 2016)	TBA
ICE*Kp6*	5	69	Helicase, Kinetochore protein, DNA cytosine methylase, Low calcium response locus protein S, Transposase, 4 Hypothetical proteins		TBA
ICE*Kp7*	7	87	mRNA endoribonuclease LS, Chromosome partition protein ParA, ATPase, LPS kinase, Tellurite resistance protein TerY, 12 Hypothetical proteins		TBA
ICE*Kp8*	3	58	Nucleotidyltransferase, Helicase, 2 Hypothetical proteins		TBA
ICE*Kp9*	8	57	DNA methyltransferase, 3 Hypothetical proteins		TBA
ICE*Kp10*	1, 12 and 17	138	Colibactin synthesis locus (*clbQ, clbP,*	- BLAST: 99% identity to	TBA

			clbO, clbN, clbM, clbL, clbK, clbJ, clbI, clbH, clbG, clbF, clbE, clbD, clbC, clbB, clbA). Transposase IS3/IS911 family, Integrase catalytic subunit	*Citrobacter koseri* ATCC BAA-898 and a number of *Enterobacter aerogenes* strains - Referred to as KPHPI208 in *Kp* strain 1084 (Lai *et al.* 2014) - Referred to as GI-I in *Kp* strain Kp52.145 (Lery *et al.* 2014) - Referred to as a 'Group III' genomic island (Marcoleta *et al.* 2016)	
ICE*Kp11*	15	92	DNA protecting protein dprA, 8 Hypothetical proteins		TBA
ICE*Kp12*	9, 10, 14, 16	97	Exonuclease SbcC, DNA helicase UvrD, ATP-dependent endonuclease, 5 Hypothetical proteins		TBA
ICE*Kp13*	Not in main lineages	65	DEAD/DEAH box helicase, Chromosome partition protein Smc, Serine protease, Metallobetalactamase superfamily protein, 5 Hypothetical proteins		TBA
ICE*Kp14*	Not in main lineages		**Middle region:** Taurine catabolism dioxygenase TfdA, Histidine ammonia lyase, Histidinol-phosphate aminotransferase, Threonyl and alanyl tRNA synthetase domain protein, Lactate dehydrogenase,		TBA

			ATP-grasp domain-containing protein, MFS transporter, DNA-binding protein, Phage conjugal plasmid C4 type zinc finger protein, 3 Hypothetical proteins **3' region:** LysR family transcriptional regulator, Alcohol dehydrogenase, ATP-binding protein, DNA helicase UvrD, ATPase, 3 Hypothetical proteins		

* unique genes in the 3' variable region unless otherwise specified