MICROBIAL GENOMICS

RESEARCH ARTICLE

Auvray *et al.*, *Microbial Genomics* 2021;7:000579

DOI 10.1099/mgen.0.000579

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Insights into the acquisition of the *pks* island and production of colibactin in the *Escherichia coli* population

Frédéric Auvray[1,*], Alexandre Perrat[1], Yoko Arimizu[2], Camille V. Chagneau[1], Nadège Bossuet-Greif[1], Clémence Massip[1,3], Hubert Brugère[1], Jean-Philippe Nougayrède[1], Tetsuya Hayashi[2], Priscilla Branchu[1], Yoshitoshi Ogura[4] and Eric Oswald[1,3,*]

Abstract

The *pks* island codes for the enzymes necessary for synthesis of the genotoxin colibactin, which contributes to the virulence of *Escherichia coli* strains and is suspected of promoting colorectal cancer. From a collection of 785 human and bovine *E. coli* isolates, we identified 109 strains carrying a highly conserved *pks* island, mostly from phylogroup B2, but also from phylogroups A, B1 and D. Different scenarios of *pks* acquisition were deduced from whole genome sequence and phylogenetic analysis. In the main scenario, *pks* was introduced and stabilized into certain sequence types (STs) of the B2 phylogroup, such as ST73 and ST95, at the *asnW* tRNA locus located in the vicinity of the yersiniabactin-encoding High Pathogenicity Island (HPI). In a few B2 strains, *pks* inserted at the *asnU* or *asnV* tRNA loci close to the HPI and occasionally was located next to the remnant of an integrative and conjugative element. In a last scenario specific to B1/A strains, *pks* was acquired, independently of the HPI, at a non-tRNA locus. All the *pks*-positive strains except 18 produced colibactin. Sixteen strains contained mutations in *clbB* or *clbD,* or a fusion of *clbJ* and *clbK* and were no longer genotoxic but most of them still produced low amounts of potentially active metabolites associated with the *pks* island. One strain was fully metabolically inactive without *pks* alteration, but colibactin production was restored by overexpressing the ClbR regulator. In conclusion, the *pks* island is not restricted to human pathogenic B2 strains and is more widely distributed in the *E. coli* population, while preserving its functionality.

DATA SUMMARY

All sequence data of the 785 *E. coli* used in this study are freely available from the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJDB5579. This database was updated to include the sequence data obtained using ONT MinION for the *E. coli* reference strain SP15 and for *E. coli* strains ECSC054, JML285, KS-NP019, NS-NP030 and SI-NP020. The sequence data of *E. coli* strain UPEC129 obtained using the PacBio instrument were deposited in the NCBI BioProject database and are available at https://www.ncbi.nlm.nih.gov/Traces/study/ under accession number PRJNA669570. Hybrid MinION-Illumina and PacBio-Illumina assemblies are available at the NCBI nucleotide database. The genome sequences of 36 other *E. coli* reference strains and seven non-*E. coli* strains were retrieved from NCBI.

INTRODUCTION

Escherichia coli is not only a commensal resident of the human and animal gut, but also a pathogen responsible for intestinal or extra-intestinal infections. *E. coli* is characterized by a high genetic and phenotypic diversity, with a population distributed into at least eight major phylogenetic groups (A, B1, B2, C, D, E, F and G) [1]. *E. coli* strains from phylogroup

Received 07 December 2020; Accepted 11 April 2021; Published 07 May 2021

Author affiliations: [1]IRSD, INSERM, Université de Toulouse, INRAE, ENVT, UPS, Toulouse, France; [2]Department of Bacteriology, Kyushu University, Fukuoka, Japan; [3]CHU Toulouse, Hôpital Purpan, Service de Bactériologie-Hygiène, Toulouse, France; [4]Division of Microbiology, Department of Infectious Medicine, Kurume University School of Medicine, Kurume, Fukuoka, Japan.

***Correspondence:** Frédéric Auvray, frederic.auvray@envt.fr; Eric Oswald, eric.oswald@inserm.fr

Keywords: colibactin; *Escherichia coli*; enterobacteria; genotoxin; genetic diversity; pathogenicity island; *pks*.

Abbreviations: asn, asparagine; CC, clonal complex; CFU, colony forming unit; DR, Direct repeats; ExPEC, extra-intestinal pathogenic *E. coli*; HGT, horizontal gene transfer; HPI, high pathogenicity island; ICE, integrative and conjugative element; ICL, interstrand cross-link; Int, Integrase; IS, insertion sequence; ML, maximum likelihood; MRCA, most recent common ancestor; NJ, neighbour joining; NRPKS, NRP synthetase; PKS, PK synthase; SNP, Single nucleotide polymorphism; ST, sequence type; tRNA, transfer ribonucleic acid; VNTR, variable number of tandem repeat.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Six supplementary figures and three supplementary tables are available with the online version of this article.

B2 are increasingly being found in the faeces of healthy humans in high-income countries and also responsible for extra-intestinal diseases, including urinary tract infections, sepsis, pneumonia and neonatal meningitis [2]. By enabling the exchange of genetic material between bacterial cells, horizontal gene transfer (HGT) is a major driving force in the evolution of bacteria, including adaptation to their host and expansion of their ecological niche [3]. HGT-mediated acquisition of large genomic islands (GIs) or pathogenicity islands (PAIs) is recognized as a major contributor to the emergence of the various *E. coli* pathotypes [4]. The *E. coli pks* pathogenicity island consists of a *clbA-clbS* gene cluster enabling the biosynthesis of a polyketide (PK) – non-ribosomal peptide (NRP) hybrid genotoxin known as colibactin [5]. This island exhibits typical features of horizontally acquired genomic elements: (i) it is a large (i.e. 54 kb) region with a distinct GC content compared to that of the chromosomal backbone, (ii) it is physically associated with a phage-type integrase gene that probably mediated its insertion into the chromosome and (iii) it is located at a tRNA locus and is flanked by two short (i.e. 17 bp) direct repeats (DRs) reminiscent of those generated upon integrase-mediated insertion of mobile genetic elements [5, 6]. The *pks* island can be found in other members of the *Enterobacteriaceae* such as *Klebsiella pneumoniae*, *Citrobacter koseri* and *Enterobacter aerogenes* [6], and in the honeybee gut commensal *Frischella perrara* [7] and the marine sponge commensal *Pseudovibrio* sp. [8].

Colibactin is a virulence factor for extra-intestinal pathogenic *E. coli* (ExPEC) [9–11] and is also a suspected procarcinogenic factor [12–14]. Colibactin induces DNA interstrand cross-links (ICLs) [15] and double-strand breaks [5] in host eukaryotic cells. Its production involves the sequential action of the Clb proteins, including PK synthases (PKSs), NRP synthetases (NRPSs), hybrid PKS-NRPS, and accessory, editing and maturation enzymes [16]. Colibactin was first synthetized as a prodrug called precolibactin, carrying an *N*-myristoyl-D-asparagine (C14-Asn) side chain that is then cleaved in the periplasm to release the active genotoxin, whose translocation across the bacterial outer membrane remains unknown [17]. The production of colibactin is positively regulated by ClbR [18]. The multi-modular PKS-NRPS assembly line not only produces colibactin but also a set of numerous secondary metabolites with varying modes of action [19, 20]. These include analgesic lipopeptides, such as C12-Asn-GABA, with the ability to diffuse across the epithelial barrier and act on sensory neurons to decrease visceral pain in the host [21]. The *pks* island also contributes to the production of siderophores (enterobactin, salmochelin and yersiniabactin), via its promiscuous phosphopantetheinyl transferase ClbA [10], and siderophore-microcins via its ClbP peptidase [22].

To date, the presence of the *pks* island has been investigated mostly in *E. coli* strains isolated from humans with extra-intestinal infections [5, 6, 23, 24]. Here we explored the distribution, conservation and functionality of the *pks* island in a large collection of non-clinical *E. coli* strains originating from human and bovine hosts [25]. We found that the *pks* island was not only present in phylogroup B2 but also in other genetic phylogroups. We identified different scenarios for its integration into the *E. coli* genome. The sequence of the *pks* island is highly conserved and *pks*-positive strains were overwhelmingly capable of producing colibactin, suggesting that the *pks* island is under selective pressure for the adaptation of *E. coli* to various ecological niches, through the production of colibactin or other metabolites or *pks*-encoded enzymatic activities.

Impact Statement

Colibactin, a genotoxin associated with the carcinogenicity of certain strains of *Escherichia coli*, is encoded by a pathogenicity island called *pks*. We took advantage of a large collection of non-clinical *E. coli* strains originating from human and bovine hosts to explore the distribution, conservation and functionality of the *pks* island. We found that the *pks* island was not only present in phylogroup B2 (and more specifically in certain B2 sublineages), but also in other genetic phylogroups, highlighting its capacity to disseminate though horizontal gene transfer. We identified various genetic *pks* configurations indicative of an introduction of the *pks* island into *E. coli* on multiple independent occasions. Despite the existence of various acquisition scenarios, we found that the *pks* sequences were highly conserved and *pks*-carrying strains were overwhelmingly capable of producing colibactin, suggesting that the *pks* island is under selective pressure, through the production of colibactin or other secondary metabolites. Future implications include the identification of such metabolites and their biological activities that could be advantageous to *E. coli* and enable its adaptation to various ecological niches.

METHODS

Bacterial strains used in the study

The *E. coli* strains were collected in Japan from 418 healthy bovines in 2013 and 2014, 278 healthy humans in 2008, 2009 and 2015, and 89 humans with extra-intestinal infections, either bacteraemia (*n*=67) in 2002–2008 or urinary tract infection (*n*=22) in 2006 and 2011. They were described recently [25] and corresponded each to a single isolate; duplicates showing less than five SNPs difference in their whole genomes were excluded from this study. A list of the 109 *pks*-positive isolates is provided in Table S1 (available in the online version of this article). An additional set of 37 *E. coli* reference strains (Table S2) and seven non-*E. coli* strains (Table S3) were included in this study; their genome sequences were downloaded from NCBI, except for *E. coli* SP15 which was not available and was obtained here (see below).

Whole genome sequencing

The whole genome sequences of the 785 *E. coli* isolates were determined by Illumina sequencing [25]. Among these, the genomes of five *E. coli* strains (ECSC054, JML285, KS-NP019, NS-NP030 and SI-NP020) were further subjected here to long-read sequencing using an Oxford Nanopore Technologies (ONT) MinION device. The DNA libraries were prepared using the rapid barcoding kit (ONT) and sequenced using MinION R9.4.1 flow cells. Long-read sequencing of *E. coli* strain UPEC129 was also performed using a Pacific Biosciences (PacBio) RSII sequencer (Genoscreen). The DNA was extracted using Gentra Puregen Yeast/Bact (Qiagen) and the DNA libraries were prepared using the SMRTbell Template Prep kit (PacBio). Hybrid assembly of Illumina paired-end reads and MinION or PacBio reads was performed using Unicycler (v.0.4.8) [26]. The whole genome sequence of *E. coli* reference strain SP15 was obtained using Illumina and ONT MinION instruments and assembled as described above.

Sequence and phylogenetic analysis

The core gene-based phylogenetic tree was reconstructed as described previously [25]. Briefly, core genes were determined using Roary [27] and SNP sites were extracted from the core gene alignment using SNP-sites [28]. The maximum-likelihood (ML) tree was reconstructed using RAxML [29] with the GTR-GAMMA model and displayed using iTOL [30].

For the phylogenetic analysis of the entire *pks* island, the genome sequences of *pks*-positive strains were aligned with the entire *pks* island sequence of strain IHE3034 using MUMmer [31] and the SNP sites located therein were identified. After removing SNP sites on the VNTR region, a neighbour-joining (NJ) tree was reconstructed by MEGA7 [32] using the Tamura–Nei evolutionary model.

Co-phylogenetic analysis of the core-gene-based ML tree and the *pks*-based NJ tree was performed using the 'cophylo' function of the R package Phytools [33].

Sequence type and phylogroup determination was performed as described previously [25].

The *pks* sequences from four *E. coli* strains belonging to distinct phylogroups (i.e. SI-NP020, KS-NP019, UPEC129 and ECSC054 from phylogroups A, B1, B2 and D, respectively) were extracted from hybrid assemblies and compared at the nucleotide level with that of the reference *E. coli* strain IHE3034. In addition, the amino acid sequences were obtained for the 19 *clb* genes of each strain and aligned by MUSCLE with MEGA7 [32]. The alignment file was analysed with the sequence identity and similarity online software (http://imed.med.ucm.es/Tools/sias.html; accessed in July 2020).

The comparison of *pks* sequences from *E. coli* and other bacterial species was performed with BLASTN (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch). Each *pks* region was defined from *clbA* to *clbS* and used as the query nucleotide sequence against each *pks* region as the subject. Then, the alignment was visualized with the Artemis Comparison Tool (v13.0.0) [34].

The integrase nucleotide and amino acid sequences were aligned using MUSCLE (v3.8.31) and the phylogeny was analysed with PhyML (v3.1/3.0 aLRT) prior to tree visualization with TreeDyn (v198.3) (http://www.phylogeny.fr; accessed in September 2020).

The CC95 strains were typed for their *fimH* allele using FimTyper (v1.0) (https://cge.cbs.dtu.dk/services/FimTyper/; accessed in November 2020) and were further assigned to subgroups A–E by analysis of the presence of either of the five subgroup-specific genes described previously [35].

PCR analysis of the *clbJK* fusion gene

The 5651 bp deletion in the *clbJ-clbK* region resulting in the *clbJK* fusion gene was tested using a duplex PCR assay, with two primer pairs. The first primer pair (clbK-F, 5′-GACTGCCCAACATACGCTCCG-3′; clbK-R, 5′-TTGTGTCGTTGTACTCTCGGC-3′) was used to amplify a 722 bp long DNA fragment that is located within the deleted region and is thus only present in strains with an intact *clbJ-clbK* region. The second primer pair consisted of primers clbJK-F (5′-AGAATTACCCACTGCCACCA-3′) and clbJK-R (5′-GGCGCTAATGGATCAGATGT-3′) flanking the deleted region, and was used to amplify a 1441 bp long DNA fragment only present in strains with a *clbJK* fusion gene. The strains with an intact *clbJ-clbK* region or a *clbJK* fusion gene yielded a 722 or a 1441 bp long amplification product, respectively. The final reaction mixture volume of 50 µl contained 2 µl template DNA, 1× GoTaq Reaction buffer, 200 µM of each dNTP, 4 mM of $MgCl_2$, 1.25 U of GoTaq DNA polymerase (Promega) and 0.2 µM of each primer (Eurofins Genomics). Amplification was done in a GeneAmp 9700 thermal cycler (Applied Biosystems), with the following programme: initial denaturation at 95 °C for 2 min; 30 cycles of denaturation at 95 °C for 30 s, annealing at 56 °C for 45 s and extension at 72 °C for 1 min 30 s; and final extension at 72 °C for 5 min. Electrophoresis was carried out in 1% agarose gel and the PCR products were visualized after Gel Red (Biotium) staining using a Bio-Rad Chemidoc XRS system (Bio-Rad).

In vitro DNA interstrand crosslinking assay

ICL activity was assessed as described previously [15]. Briefly, 3×10^6 *E. coli* cells or 6×10^6 *Erwinia oleae* cells pre-grown for 3.5 h in Dulbecco's modified Eagle medium (DMEM) with 25 mM HEPES (Invitrogen) were mixed with EDTA (1 mM) and 400 ng of linearized plasmid pUC19 DNA and the mixtures were incubated for 40 min at 37 °C. After pelleting the bacteria, the DNA was purified from the supernatant and analysed by electrophoresis on denaturing (40 mM NaOH – 1 mM EDTA) 1% agarose gels. ICL activity of *Erwinia oleae* was also tested in the presence of 400 nM 6-histidine-ClbS, which was purified with HisPur nickel-nitrilotriacetic acid (Ni-NTA) agarose (Thermo Scientific) from a culture of

Table 1. Occurrence of *pks* in *E. coli* strains from healthy humans or bovines, and human patients with extra-intestinal infection.

Origin	Phylogroup (no. of *pks*$^+$ strains/no. of strains tested)								
	A	B1	B2	C	D	E	F	Uncl*	Total
Healthy bovines†	1/48	7/314	4/11	0/20	0/12	0/8	0/4	0/1	12/418
Healthy humans†	0/14	1/29	61/163	0/11	0/37	0/1	0/20	0/3	62/278
Human patients‡	0/3	0/5	34/59	0/4	1/10	0/0	0/8	0/0	35/89
Total	1/65	8/348	99/233	0/35	1/59	0/9	0/32	0/4	109/785

*Unclassified.
†Isolates were collected from faeces of healthy individuals.
‡Isolates were collected from blood (*n*=67) or urine (*n*=22) from human patients with extra-intestinal infection.

BL21(DE3) strain hosting the plasmid pET28a-ClbS-His, as described previously [15].

Megalocytosis assay

Non-haemolytic *pks*-positive strains were tested for megalocytosis on infected HeLa cells as described previously [5, 36]. Briefly, HeLa cells grown to 50% confluence in cell culture 96-well plates were inoculated with 5 µl of an overnight culture of bacteria in infection medium (DMEM with 25 mM HEPES) and incubated for 4 h at 37 °C in a 5% CO_2 atmosphere. Cells were then washed and incubated for 48–72 h in cell culture medium supplemented with gentamicin at 200 µg ml^{-1}, and then stained with methylene blue for microscopy examination.

H2AX phosphorylation assay

HeLa cells were infected as described above and H2AX phosphorylation was quantified immediately after the 4 h infection step by immunofluorescence as described elsewhere [37].

C14-asn quantification

E. coli strains were grown for 24 h at 37 °C in 10 ml DMEM-HEPES (Gibco), resuspended in 500 µl Hanks' balanced salt solution (HBSS; Invitrogen) and then crushed with a Precellys instrument (Ozyme). After addition of an internal standard mixture (deuterium-labelled compounds; 400 ng ml^{-1}), cold methanol (MeOH) was added and samples were solid-phase extracted on HLB plates (OASIS HLB 2 mg, 96-well plate; Waters). Lipids were eluted with MeOH, evaporated under N_2, resuspended in MeOH and analysed by HPLC/MS-MS analysis (LC-MS/MS) (Meta-ToulLipidomics Facility), as described previously [21].

RESULTS

The *pks* island was mainly found in specific *E. coli* lineages from phylogroup B2

The presence of the *pks* island was investigated in a collection of 785 *E. coli* strains [25] belonging to at least 296 different sequence types (STs) and originating mostly from faecal samples of healthy bovines and humans. Clinical isolates recovered from urine or blood samples of human patients with extra-intestinal infection were also included for comparison. We detected the *pks* island in 109 *E. coli* strains, including 62 (22.3%) out of 278 healthy human faecal isolates and 12 (2.9%) out of 418 healthy bovine faecal isolates (Table 1, Fig. 1). As expected, a higher proportion of *pks*-positive strains were found among ExPEC, i.e. 35 (39%) out of 89 strains, including 14 (63.6%) out of 22 strains from urinary tract infection and 21 (31.3%) out of 67 strains from bacteraemia. The vast majority of the 109 *pks*-positive strains corresponded to B2 isolates (Table 1, Fig. 1) and the *pks* island was mainly present in specific lineages or STs of the B2 phylogroup (Fig. 1).

Strikingly, the *pks* island was found in (nearly) 100% of strains belonging to ST12, ST73, ST95 and ST550, while it was excluded from other STs, such as ST131 and ST357 (Fig. 1, Table 2). Interestingly, these *pks*-positive and *pks*-negative STs are found in distinct clusters in the core-genomebased phylogenetic tree (Fig. 1), suggesting that *pks* acquisition occurred after the divergence of these clusters from a common ancestor. We further characterized the 54 *pks*-positive strains of ST95 for their *fimH* allele and affiliation to CC95 subgroups A–E defined previously [35]. We could assign 35 of them to subgroup A (*n*=22), B (*n*=12) or E (*n*=1) (Table S1). The remaining 19 strains, including 15 of serotype O1:H1, did not belong to any of these five subgroups. No *pks*-positive strain was assigned to CC95 subgroups C or D, in agreement with previous results [35].

Except for four B2 strains originating from healthy bovines, the *pks*-positive B2 isolates originated from humans, either patients with extra-intestinal infection (*n*=34) or healthy individuals (*n*=61) (Fig. 1, Table 1). The low occurrence of *pks* among bovine isolates probably reflected the low prevalence of B2 strains in cattle [25]. Interestingly, 10 non-B2 *pks*-positive strains were identified corresponding to one human blood isolate from phylogroup D, one healthy human faecal isolate from group B1, and eight healthy bovine faecal isolates from groups A (*n*=1) and B1 (*n*=7) (Fig. 1, Table 1). In contrast to the B2 *pks*-positive isolates, these strains were scattered throughout the core

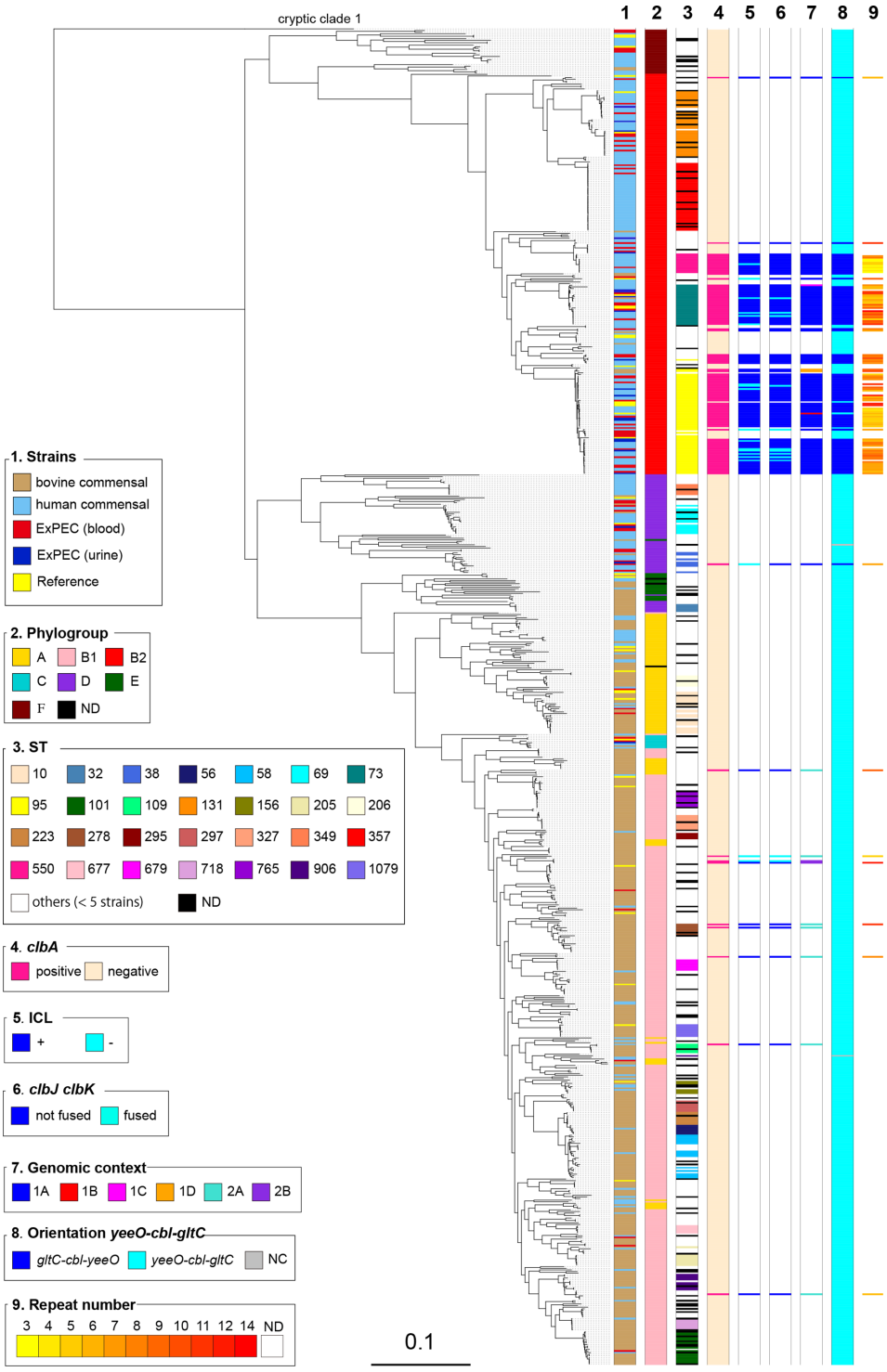

Fig. 1. Phylogenetic relationship and distribution of *pks*-positive/negative *E. coli* isolates among 696 human and bovine commensal *E. coli*, 89 ExPEC and 37 completely sequenced reference *E. coli* strains. A core gene-based ML tree was reconstructed based on 271403 SNPs located on 2000 core genes and rooted on cryptic *Escherichia* clade 1 strains as outgroups. Origin (column 1), phylogroup (column 2), major sequence type (ST) (i.e. ST identified for at least five strains) (column 3), presence of *pks* (*clbA*) (column 4), colibactin activity (ICL) (column 5), presence of the *clbJK* fusion gene (column 6), genetic *pks* configuration (see Fig. 3) (column 7), orientation of the *asnV-asnU-asnW* region situated downstream of the *asnT* tRNA gene (column 8) and the number of 5′-ACAGATAC-3′ repeats found in the *clbB-clbR* intergenic region (see Fig. 2) (column 9) are shown for each strain. The scale bar shows the number of substitutions per site. ND, not determined.

Table 2. Distribution of the *pks* island in the predominant sequence types (ST) among *E. coli* strains isolated from healthy bovines, healthy humans and human patients with extra-intestinal infection

Phylogroup	ST*	*pks*⁺/no. of strains
A	10	0/22
	206	0/5
	6126	0/4
B1	20	0/5
	29	0/4
	56	0/6
	58	0/23
	101	0/16
	109	1/6
	154	0/6
	155	0/4
	156	0/4
	164	0/6
	205	0/9
	223	0/5
	278	2/8
	295	0/4
	297	0/7
	300	0/4
	327	0/9
	332	0/4
	446	0/4
	677	0/6
	679	0/7
	718	0/6
	765	0/9
	795	0/4
	906	0/12
	1079	0/7
	1423	0/5
	5487	0/4
B2	12	4/4
	73	20/20
	95	54/56
	131	0/25
	357	0/35
	550	12/12
	1193	0/4
C	88	0/4
D	32	0/5
	38	1/8
	69	0/14
	349	0/6

*Only STs including at least four strains are listed.

genome phylogenetic tree and were not representative of any particular lineage or ST (Fig. 1, Table 2).

High level of genetic conservation of the *pks* island among *E. coli* phylogroups and other enterobacteria

The *pks* sequence from the B2 reference *E. coli* strain IHE3034 was compared to that of three non-B2 *E. coli* isolates, including the single group A isolate (i.e. SI-NP020), the single group D isolate (i.e. ECSC054) and one of the eight B1 isolates (i.e. KS-NP019). An additional B2 isolate (i.e. UPEC129) was also selected for this analysis. To perform this comparison, the whole genomes of these four isolates were assembled from a combination of short and long reads. At the amino acid level, over 99% identity was observed for each of the 19 *clb* gene products (Fig. 2). At the nucleotide level, the only variation observed in the *pks* sequence was the size of the region located between *clbB* and *clbR* which contains a variable number of tandem repeats (VNTRs) of the motif 5′-ACAGATAC-3′ [6]. This VNTR locus contained between three and 14 repeat units when the whole collection of *pks*-positive strains was analysed (except for 17 isolates for which the VNTR length could not be calculated), with no apparent correlation with the STs (Fig. 2). Therefore, apart from the size of the VNTR, the *pks* island was highly conserved among the strains, irrespective of their phylogroup or ST.

Comparison of the *pks* island nucleotide sequence from B2 reference *E. coli* strain IHE3034 with that of other *pks*-positive bacterial species confirmed that it was conserved in other members of the *Enterobacteriaceae* (Fig. S1) such as *K. pneumoniae*, *Enterobacter aerogenes*, *C. koseri, Serratia marcescens* and *Erwinia oleae*. A similar *pks* island was present, although less conserved, in *F. perrara* and *Pseudovibrio* sp. (Fig. S1).

The *pks* islands in *E. coli* from phylogenetic groups B2 and D share a similar genomic environment

To gain insights into the events leading to the acquisition of the *pks* island into the *E. coli* population, we analysed

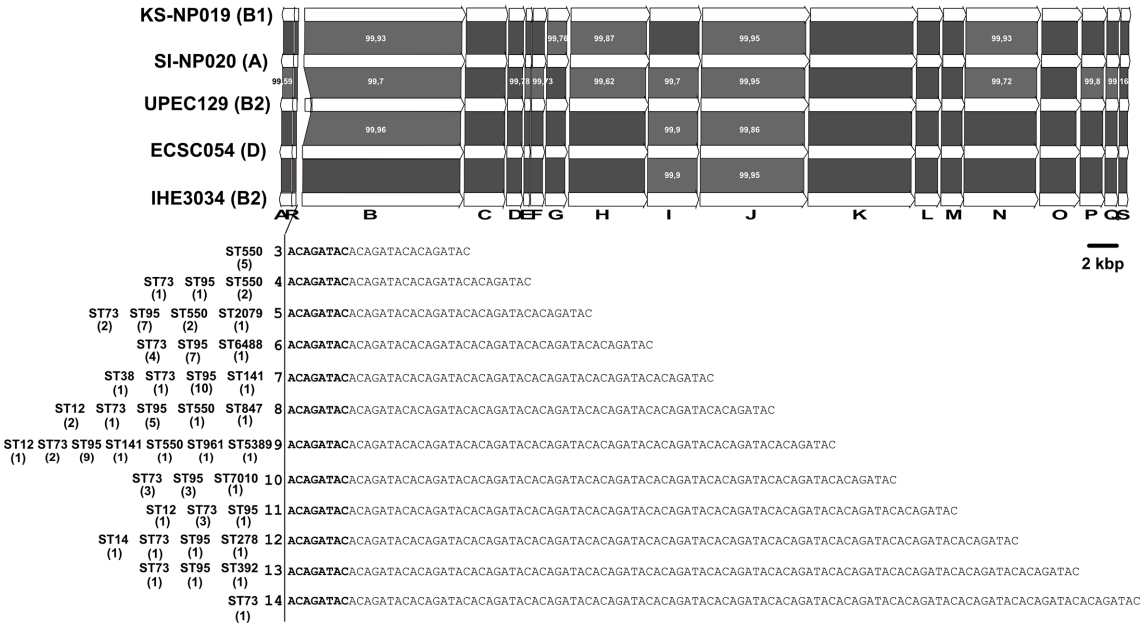

Fig. 2. Comparison of the *pks* islands of *E. coli* strains belonging to phylogroups A, B1, B2 and D. The 19 ORFs of the *clbA-clbS* gene cluster from the reference *E. coli* strain IHE3034 sequence (group B2) and MinION- or PacBio-derived sequences of *E. coli* strains KS-NP019 (group B1), SI-NP020 (group A), UPEC129 (group B2) and ECSC054 (group D) are represented by arrows with the arrowhead representing the direction of transcription. The areas between the corresponding genetic maps shaded in dark and light grey indicate 100% amino acid identity and *ca*. 99% amino acid similarity, respectively. The number of repeated 5′-ACAGATAC-3′ motifs found in the *clbB-clbR* nucleotide intergenic region of *pks*-positive *E. coli* strains and the sequence types (ST) of the corresponding strains are indicated below the *clbA-clbS* gene cluster, with the number of strains given in parentheses.

the genomic environment of the *pks* island in the 109 *pks*-positive strains and in seven *pks*-positive *E. coli* reference strains (i.e. 536, ABU83972, CFT073, Nissle 1917, UTI89, IHE3034 and SP15). Various configurations were found for the *pks* island genomic environment, suggesting two main scenarios of *pks* acquisition, depending on the presence or absence of an integrase gene (Fig. 3). The genetic configuration typical of B2 strains, named 1A, which is characterized by a *pks* island carrying an integrase gene and inserted into the *asnW* tRNA gene in the vicinity of the *asnT*-located high pathogenicity island (HPI) [5, 6], was found in 96 B2 strains of our collection and in one phylogroup D strain, ECSC054 (Figs 1 and 3).

A similar configuration was found in a few other B2 strains but with variations in the location of the *pks* island, which was inserted into either the *asnV* (corresponding to configurations 1B and 1C found in ST95 reference strain UTI89 and in ST73 strain JML226, respectively) or the *asnU* tRNA gene (corresponding to configuration 1D found in strains KS-P003 and KS-P027, both belonging to ST95) (Figs 1 and 3) . Besides the difference in the tRNA insertion site, the configuration 1B found in reference strain UTI89 differed from the major configuration 1A by the orientation of the 4309 bp *asnW-asnU-asnV* tRNA region upstream of the *pks* island. This region contains three other genes, namely *gltC* and *cbl*, encoding two LysR-family transcriptional regulators, and *yeeO*, encoding a flavin mononucleotide (FMN) and flavin adenine dinucleotide (FAD) exporter. Configurations 1C and 1D possessed the same *asnW-asnU-asnV* orientation as in UTI89 and carried a 25 kb region between the *pks* integrase gene and *clbS*. A 14 kb section from this region exhibited high sequence similarity (>99%) to integrative and conjugative elements (ICEs) identified in *E. coli* (ICE*Ec1*) and *K. pneumoniae* (ICE*Kp1*), in particular to the DNA regions I and II from ICE*Ec1* involved in mating-pair formation (Mpf) and DNA mobilization, respectively (Fig. 4) [6, 38]. This 14 kb section could therefore be considered as an ICE-like element, although it is most probably non-functional given the lack of a complete region II (Fig. 4). The remaining 11 kb section was not homologous to ICE*Ec1*, ICE*Kp1* or any other ICE, and its role could not be predicted.

The phage-type *pks* integrase is a tyrosine site-specific recombinase with similarity to the phage P4 integrase C-terminal catalytic domain (INT_P4_C). The integrase genes located at the *asnW* (configuration 1A) or *asnV* loci (configurations 1B and 1C) and their gene products were highly conserved and grouped into the integrase family 1 (Fig. S2), whereas the integrase genes located at the *asnU* locus (configuration 1D) and their gene products shared 94% nucleotide and 94.6% amino acid sequence similarity, respectively, with those of family 1 and were thus grouped into the integrase family 2 (Fig. S2).

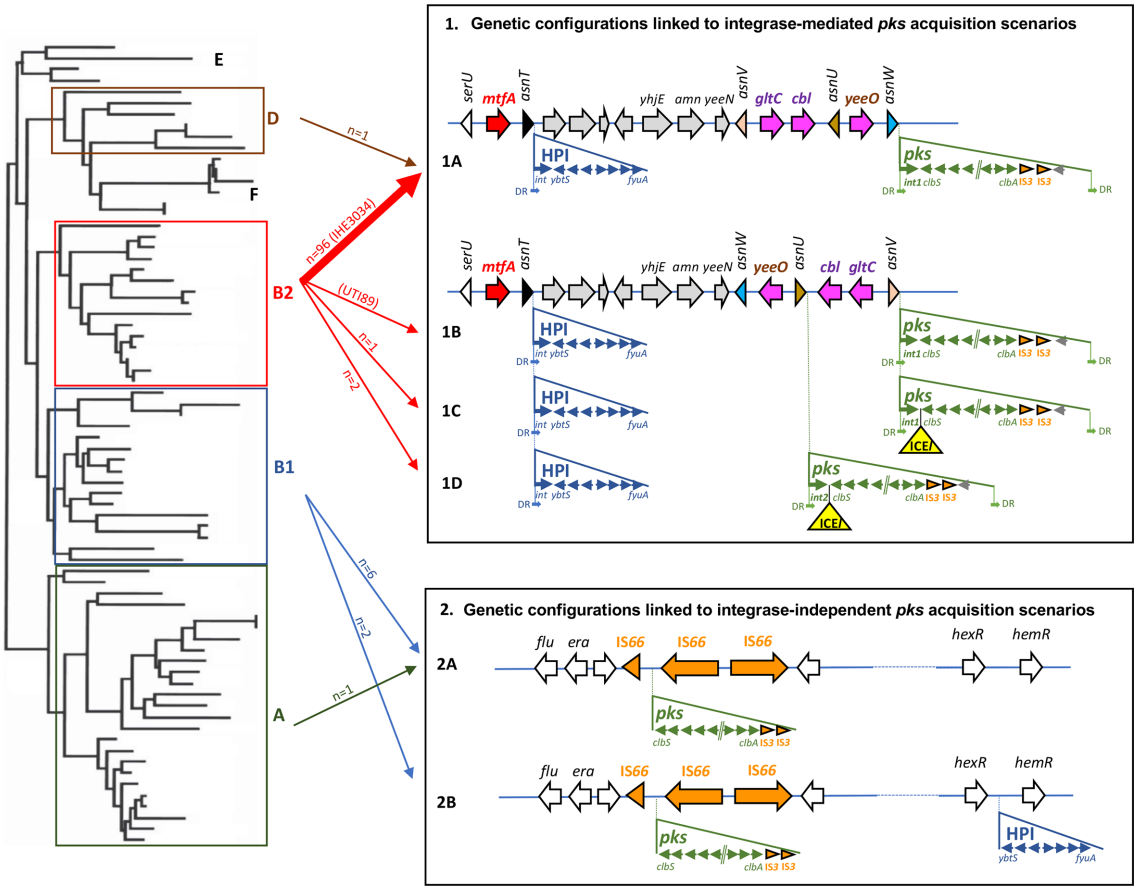

Fig. 3. Genetic configurations of the *pks* island and HPI in *E. coli* strains and proposed scenarios for their acquisition. Left: schematic phylogenetic tree showing the distribution of *E. coli* into the main phylogroups. Right: *E. coli* genetic *pks* and HPI configurations resulting from proposed acquisition scenarios involving site-specific recombination (configurations 1A, 1B, 1C and 1D) or not (configurations 2A and 2B). The location and orientation of the tRNA genes and the ORFs of the chromosomal regions, including the integrase and genes from *pks* and the HPI, are indicated by the arrows. Partial and complete insertion sequence (IS) elements are represented by orange arrowheads and arrows, respectively. The ICE-like element (ICE*l*) found in configurations 1C and 1D is represented as a yellow triangle. DR, direct repeats located at the extremities of the islands (except for the HPI in configurations 1A–1D, one DR lacking at the right border). Middle: arrows connect phylogroups A, B1, B2 and D (left) with the *pks* and HPI configurations (right). The number of *E. coli* isolates belonging to the collection of 785 strains and corresponding to each configuration is indicated (except for configuration 1B which was only found in reference strain UTI89, indicated in parentheses). The thick arrow represents the most frequently found configuration (exemplified here by reference strain IHE3034, indicated in parentheses).

Atypical genomic environments of *pks* islands in *E. coli* from phylogenetic groups A and B1

In the nine *pks*-positive *E. coli* strains from phylogroups A and B1, two different configurations (named 2A and 2B) were observed that drastically differed from those found in B2/D strains. Their *pks* islands lacked an integrase gene, were not inserted into a tRNA gene and there were no direct repeats at their chromosomal boundaries (Fig. 3). The *pks* islands were located in the vicinity of the genes *flu* (or *agn43*) and *era* encoding the Ag43 autotransporter adhesin and a GTPase essential for cell growth and viability, respectively. They were flanked on one side by a truncated copy of the IS*66* insertion sequence (IS) and on the other side by two intact IS*66* copies. Two truncated copies of IS*3* were also found next to the *clbA* gene but this was also the case for configurations 1A–1D. Moreover, in these B1/A strains, the HPI was absent (Table 3, Fig. 3, configuration 2A), except for two isolates in which the HPI was present but not in the vicinity of the *pks* island and not into a tRNA locus (Table 3, Fig. 3, configuration 2B). Using PCR assays, it was shown previously that three *E. coli* strains from phylogroup B1 (namely U12633, U15156 and U19010) possessed a *pks* island that co-localized with the HPI and the DNA transfer and mobilization region of an ICE*Ec1*-like element [6], a situation that is reminiscent of that of configuration 1D. However, as the whole genome sequences of these three strains were not available, this could not be confirmed here.

Since the *asnW-asnU-asnV* tRNA region displayed distinct orientations in *pks*-positive B2 strains depending on *pks*

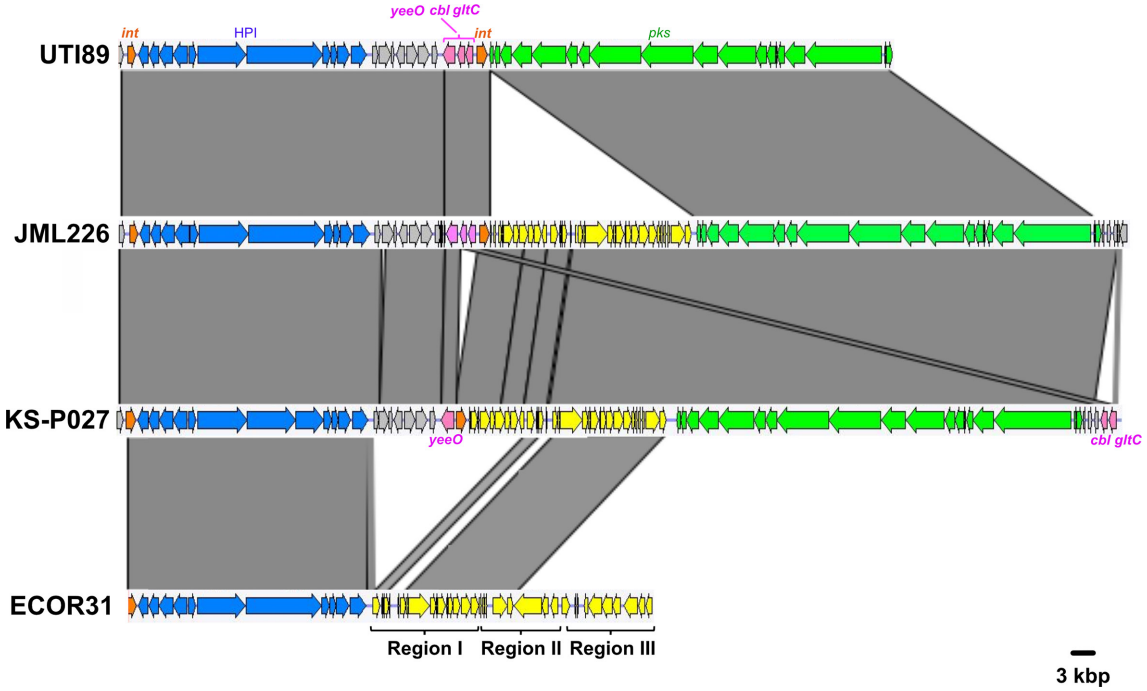

Fig. 4. Comparison of the chromosomal region covering the HPI and *pks* island between three atypical *E. coli* B2 strains (UTI89, JML226 and KS-P027, with genetic configurations 1B, 1C and 1D, respectively) and the integrative conjugative element ICE*Ec1* from *E. coli* strain ECOR31. Nucleotide sequence similarity (>99%) between different DNA regions is indicated by grey areas between the corresponding genetic maps. The *pks* island and the HPI are represented in green and blue, respectively, and the integrase genes in orange. The *yeeO*, *cbl* and *gltC* genes located between the *asnV* and *asnW* tRNA genes in UTI89 are represented in pink. The region between the HPI and the *yeeO* gene is represented in grey and the ICE-related region inserted either next to *pks* (JML226 and KS-P027) or next to the HPI (ECOR31) is represented in yellow. In strain ECOR31, the ICE is divided into three parts, including region I encoding a mating pair formation system, region II encoding a DNA-processing system, both involved in conjugative transfer, and region III comprising hypothetical genes.

configuration, we further analysed its orientation for the rest of the *E. coli* collection, i.e. in *pks*-positive B1/A strains and in *pks*-negative strains. The '*asnV-asnU-asnW*' orientation was uniquely found in typical *pks*-positive B2 strains with configuration 1A, suggesting that, in these strains, *pks* acquisition at the *asnW* locus was accompanied by an inversion of the upstream tRNA-encoding region (Figs 1 and 3).

The phylogeny of the *pks* island globally reflects that of the *E. coli* core genome

To shed further light on the different *pks* acquisition scenarios, we constructed a phylogenetic tree of the entire *pks* sequences (i.e. from *clbA* to *clbS*, except for the VNTR-containing region which was excluded from the analysis) from the 109 *pks*-positive strains. Globally, the *pks* sequences from the strains showing distinct *pks* genomic configurations formed distinct clusters (Fig. 5). The *pks* sequence of strain UTI89 with a unique configuration (configuration 1B) was clustered with those of strains with configuration 1A (Fig. 5). Remarkably, the *pks* sequences of B1/A strains segregated separately from those of B2/D strains. Moreover, the *pks* sequences with an insertion of an ICE-like element in the B2 (ST73) human strain JML226 (with configuration 1C) or in the pair of B2 (ST95) bovine strains KS-P003 and KS-P027 (with configuration 1D) also clustered separately and were closer to the *pks* sequences of B1/A strains than to those of the B2/D strains lacking this ICE-like element (Fig. 5). Finally, the *pks* sequences of *C. koseri*, *Enterobacter aerogenes*, *K. pneumoniae* and *S. marcescens* were close to those of *E. coli* B2/D strains with configuration 1A (Fig. 5) whereas that of *Erwinia oleae* was more phylogenetically distant and clustered separately (data not shown).

To further assess the evolutionary relationships between the *pks* sequences and the genetic background of the strains, a co-phylogenetic analysis was performed where the phylogenetic trees based on the *pks* sequence and the core genome were compared. Globally, congruence was observed between both trees (Fig. 6). It was noticeable that most of the typical *pks*-positive B2 strains whose core genomes clustered together into lineages of clonal complexes (CC) 12, CC14, CC73 and CC95 contained *pks* sequences that also clustered together in different subgroups of the main *pks* cluster (Fig. 6). In particular, the CC95 strains that clustered together into subgroups A and B (as defined by *fimH* typing) or O1:H1 subgroup in the core genome tree also clustered together in the *pks* tree (Fig. 6). These observations support the hypothesis of an introduction of the *pks* island into CC12, CC14,

Table 3. Characteristics of B2 and non-B2 *pks*-positive *E. coli* strains with atypical features regarding *pks* integrity, functionality or location

Group	Strain	Origin**	Sample	Year	ST	Serotype	*hlyA* (haemolysis)	*ybt* (locus)	*pks* (locus)	*clbJ-clbK**	Megal.†	ICL	H2AX†	C14-Asn †,‡
A	SI-NP020	b	Faeces	2014	7010	uncl:H14	– (–)	–	+ (not tRNA)	wt	+	+	+	+++
B1	JML285	H	Faeces	2015	109	Gp2:H8	– (–)	–	+ (not tRNA)	wt	+	+	+	+++
	HH-NP008	b	Faeces	2014	847	uncl:H2	– (–)	–	+ (not tRNA)	wt	+	+	NT	NT
	KK-NP025	b	Faeces	2014	6488	uncl:H8	– (–)	–	+ (not tRNA)	wt	NT	+	NT	NT
	KS-NP019	b	Faeces	2013	392	O8:H2	+ (+)	+ (not tRNA)	+ (not tRNA)	wt	NT	+	NT	+++
	SI-NP013	b	Faeces	2014	278	uncl:H7	– (–)	–	+ (not tRNA)	wt	+	+	NT	NT
	SI-NP017	b	Faeces	2014	278	Gp2:H21	– (–)	–	+ (not tRNA)	wt	+	+	NT	NT
	NS-NP014	b	Faeces	2014	2079	O8 :H19	– (–)	–	+ (not tRNA)	f	NT	–	NT	NT
	NS-NP030	b	Faeces	2014	392	uncl:H2	+ (+)	+ (not tRNA)	+ (not tRNA)	f	NT	–	NT	++
B2	JML114	H	Faeces	2015	73	O6:H1	– (–)	+ (asnT)	+ (asnW)	wt	–	–	–	++
	JML165	H	Faeces	2015	550	uncl:H5	– (–)	+ (asnT)	+ (asnW)	wt	–	–	NT	++
	JML201	H	Faeces	2015	95	O1:H1	– (–)	+ (asnT)	+ (asnW)	wt	–	–	NT	NT
	JML226	H	Faeces	2015	73	Gp7:H12	+ (+)	+ (asnT)	+ (asnV)	wt	NT	+	NT	NT
	KS-P003	b	Faeces	2013	95	Gp7:H5	+ (+)	+ (asnT)	+ (asnU)	wt	NT	+	NT	NT
	KS-P027	b	Faeces	2013	95	Gp7:H5	+ (+)	+ (asnT)	+ (asnU)	wt	NT	+	NT	NT
	JML008	H	Faeces	2015	95	Gp7:H4	– (–)	+ (asnT)	+ (asnW)	f	–	–	NT	NT
	JML102	H	Faeces	2015	73	O6:H1	– (–)	+ (asnT)	+ (asnW)	f	–	–	NT	NT
	JML282	H	Faeces	2015	95	Gp7:H7	– (–)	+ (asnT)	+ (asnW)	f	–	–	NT	NT
	JML288	H	Faeces	2015	95	O1:H7	– (–)	+ (asnT)	+ (asnW)	f	–	–	NT	NT
	JML291	H	Faeces	2015	95	O1:H12	– (–)	+ (asnT)	+ (asnW)	f	–	–	NT	NT
	JML296	H	faeces	2015	73	uncl:H1	– (–)	+ (asnT)	+ (asnW)	f	–	–	–	++
	SI-NP032	b	Faeces	2014	73	O25:H5	+ (+)	+ (asnT)	+ (asnW)	f	NT	–	NT	++
	ECSC09	H	Blood	2006	95	Gp7:H7	– (–)	+ (asnT)	+ (asnW)	f	–	–	NT	NT
	UPEC57	H	Urine	2011	95	Gp7:H7	– (–)	+ (asnT)	+ (asnW)	f	–	–	–	NT
	UPEC91	H	Urine	2011	95	O1:H7	– (–)	+ (asnT)	+ (asnW)	f	NT	–	NT	NT
	CM1	H	Urine	2006	95	O1:H1	– (–)	+ (asnT)	+ (asnW)	wt	–	–	NT	–
	UPEC129	H	Urine	2011	uncl	Gp7:H7	– (–)	+ (asnT)	+ (asnW)	wt	–	–	–	–
D	ECSC054	H	Blood	2004	38	O4:H30	– (–)	+ (asnT)	+ (asnW)	wt	–	–	–	+

*wt, full-length *clbJ* and *clbK* genes; f, *clbJK* fusion gene.
†Megal., megalocytosis; nt, not tested.
‡–, no C14-Asn detected; +, *ca.* 50–400 pg C14-Asn/10^8 c.f.u.; ++, *ca.* 400–600 pg C14-Asn/10^8 C.F.U.; +++, *ca.*650–1200 pg C14-Asn/10^8 c.f.u.
** b, bovine; H, human.

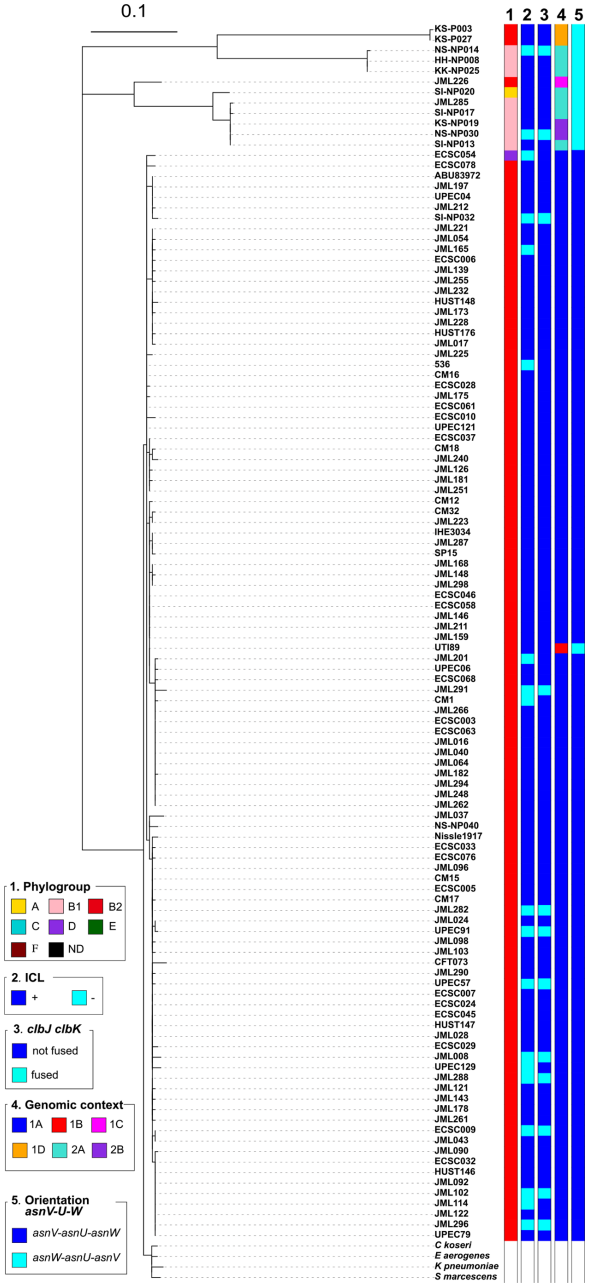

Fig. 5. Phylogenetic tree of the entire *pks* island. SNP analysis was performed with the *pks* sequences of the 109 *pks*-positive *E. coli* strains and an NJ phylogenetic tree was built. The *pks* sequences from seven reference *E. coli* strains (536, ABU83972, CFT073, Nissle 1917, UTI89, IHE3034 and SP15) and other enterobacteria (i.e. *C. koseri* ATCC BAA-895, *Enterobacter aerogenes* EA1509E, *K. pneumoniae* 1084 and *S. marcescens* AS012490) were also included in this tree. Phylogroup (column 1), colibactin activity (column 2), presence of a *clbJK* fusion gene (column 3), genetic *pks* configuration (see Fig. 3) (column 4) and orientation of the *asnV-asnU-asnW* region (see Fig. 3) (column 5) are indicated for each *pks*-positive strain. The scale bar shows the number of substitutions per site.

CC73 and CC95 through horizontal acquisition by their most recent common ancestor (MRCA) or by the MRCA of each of these lineages, followed by vertical transmission with subtle *pks* divergence overtime. Since CC95 subgroups C or D contain *pks*-negative strains [35] (strain ECSC026, subgroup C; this study), we further hypothesize that *pks* was lost during the evolution of these sublineages.

The fact that a single *pks*-positive strain from phylogroup D (i.e. ECSC054) possessed a *pks* island whose sequence clustered with that of B2 strains (Fig. 6) suggests that this strain acquired *pks* from a B2 strain through HGT. The *pks*-carrying B1/A strains were diverse based on their core genomes and their *pks* sequences clustered into two separate groups that were distantly related to the major *pks* cluster of B2 strains (Fig. 6), suggesting the existence of sporadic *pks* introduction within the B1 and A phylogroups, presumably through HGT from a donor strain different from typical *pks*-positive B2 strains. Finally, the co-phylogeny also confirmed that the two atypical B2 ST95 strains KS-P03 and KS-P027 clustered with the other B2 strains of ST95 based on the core genome but contained a divergent *pks* sequence which was closer to those of B1 or A strains (Fig. 6), suggesting that this pair of strains probably acquired their *pks* islands through HGT, possibly from a donor strain carrying a *pks* island with an ICE insertion. The same scenario also presumably occurred with the atypical B2 ST73 strain JML226, which clustered with the other B2 strains of ST73 in the core genome tree but carried a *pks* island characterized by an ICE-like insertion and a sequence closer to those of B1/A strains than to those of B2 ST73 strains.

The functionality of the cluster of genes of the *pks* island is conserved in the majority of the enterobacterial strains

We next investigated the functionality of the *pks* islands in *E. coli* strains belonging to various phylogroups and carrying phylogenetically distinct *pks* sequences, as well as in the *Erwinia oleae* strain DAPP-PG531. Production of the genotoxin colibactin was directly investigated through the formation of DNA ICLs (Fig. S3a). The vast majority of the *E. coli* strains carrying the *pks* island (i.e. 83.5%) produced ICLs (Fig. 1). DNA-crosslinking was also observed for the *Erwinia oleae* strain, and it was abrogated by adding purified colibactin self-resistance protein ClbS (Fig. S4), confirming the production of a *bona fide* colibactin by this strain carrying a less conserved sequence of the *pks* island. The *E. coli* genotoxic strains belonged to phylogroups B2, B1 and A (Fig. 1). Eighteen (16.5%) *pks*-positive *E. coli* isolates lacked a detectable interstrand crosslinking activity, including 15 strains from phylogroup B2, two strains from phylogroup B1 and the single *pks*-positive strain from phylogroup D (Table 3). These strains did not cluster together in the core genome phylogenetic tree but instead were intertwined among genotoxic strains (Figs 1 and 5). To confirm the absence of genotoxicity, we tested the ability of these ICL-negative strains to trigger megalocytosis in cultured HeLa cells (Fig. S3b) and phosphorylation of histone H2AX (Fig. S3c), a robust

Fig. 6. Co-phylogeny of *pks* sequences and *E. coli* host strains. A comparison generated with Phytools of the *E. coli* core gene-based ML tree and *pks*-based NJ tree is shown, including the links between *pks* and host strains (dashed lines). The phylogenetic groups (A, B1, B2 and D) are indicated. The strains belonging to the major clonal complexes (CCs) are shown with coloured names, including those of CC12 [containing four ST12 strains, one ST961 strain (ECSC078) and one ST5389 strain (ECSC078)], CC14 [containing one ST14 strain (ECSC010) and 12 ST550 strains], CC73 (containing only ST73 strains) and CC95 (containing only ST95 strains). Strains from CC95 that belong to subgroups A and B (as defined by *fimH* typing) or to serotype O1:H1 are boxed. Seven reference *pks*-positive *E. coli* strains (536, ABU83972, CFT073, Nissle 1917, UTI89, IHE3034 and SP15) were included in both trees, whereas the non-*E. coli* strains carrying a *pks* island (i.e. *C. koseri* ATCC BAA-895, *Enterobacter aerogenes* EA1509E, *K. pneumoniae* 1084 and *S. marcescens* AS012490) were included only in the *pks*-based tree.

marker for DNA damage in eukaryotic cells. To avoid cell lysis during infection, we assessed only non-haemolytic strains. No megalocytosis and no p-H2AX foci were detected in HeLa cells exposed to subsets of ICL-negative strains (Table 3; n=14 and n=5, respectively), even at a high multiplicity of infection, confirming the deficiency of these strains in colibactin production. These results showed that except for a few strains, *E. coli* strains carrying a *pks* island are overwhelmingly capable of producing the genotoxin colibactin, regardless of their phylogenies and the genomic configurations of their *pks* islands.

To examine the reasons for the lack of genotoxic activity of the 18 ICL-negative *E. coli* strains, we further analysed the

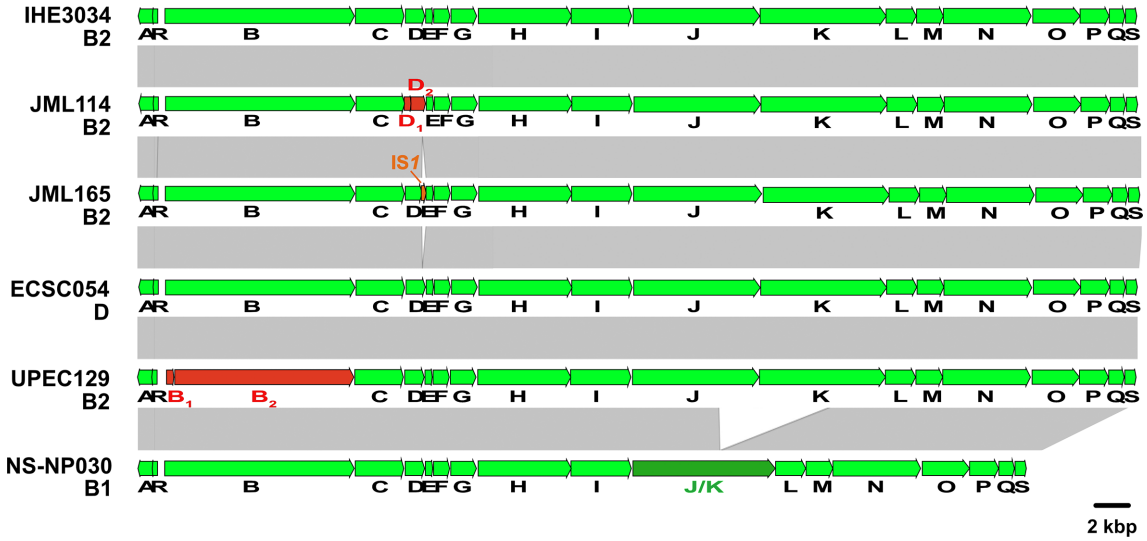

Fig. 7. Comparison of the *pks* island sequence from the *E. coli* reference strain IHE3034 with that of a selection of *pks*-positive but non-genotoxic *E. coli* isolates. Nucleotide sequence similarity (>99%) between different DNA regions is indicated by grey areas between the corresponding genetic maps. Fusion of two adjacent ORFs resulting from the deletion of a sequence overlapping the two ORFs is indicated in dark green. Adjacent ORF sequences resulting from the segregation of an original ORF following an insertion or deletion event are indicated in red. The IS*1* located in the *pks* island of strain JML165 is represented in orange.

sequence of the *pks* island from those strains. We identified genetic alterations of the *pks* island in 16 out of the 18 non-genotoxic isolates. Strain JML114 carried a single nucleotide deletion in *clbD* at position 172 (A), leading to the segregation of *clbD* into two ORFs (Fig. 7). Strain JML165 carried an IS*1* inserted at the 3′-end of *clbD* after position 838. In strains UPEC129 and JML201, *clbB* was segregated into two ORFs due to nucleotide substitutions at positions 452 and 453 (AC to GA) (UPEC129; Fig. 7), and at position 872 (G to A) (JML201; data not shown), respectively. The genetic alterations identified in these four non-genotoxic strains each resulted in premature stop codons in *clbB* or *clbD* genes coding enzymes that are essential for the production of colibactin.

Twelve ICL-negative strains carried a 5651 bp deletion resulting in a *clbJK* fusion gene, as shown for strain NS-NP030 in Fig. 7. This deletion presumably resulted from recombination between two copies of a 1480 bp homologous sequence located in *clbJ* and *clbK*. A PCR analysis of the corresponding region in the 109 *pks*-positive strains confirmed the presence of this deletion in the 12 strains, whereas the other 97 *pks*-positive strains contained full-length *clbJ* and *clbK* genes (Fig. 5). The 12 strains carrying the *clbJK* fusion were detected sporadically in the core genome phylogenetic tree (Fig. 1) and *pks* phylogenetic tree (Fig. 5), suggesting that occurrence of the deletion between *clbJ* and *clbK* arose from accidental recombination events. The predicted 2440 aa hybrid ClbJK protein encoded by the *clbJK* fusion gene lacks the PKS module of ClbK necessary for the formation of stable cross-links [19]. In agreement with this, the strains carrying this fusion were devoid of interstrand crosslinking activity and did not trigger megalocytosis or histone H2AX phosphorylation in infected eukaryotic cells (Table 3). It was reported however that rat *E. coli* isolates carrying a *clbJK* fusion gene caused DNA damage or displayed cytotoxicity to HeLa cells [39, 40]. This discrepancy with our results could be due to the use of distinct experimental conditions. The possibility that these rat isolates might produce additional genotoxins that would mask any colibactin deficiency caused by the *clbJK* fusion also cannot be excluded. Caution should also be observed during the assembly of sequencing reads as errors including deletions may be caused by the presence of tandem repeats.

For two other non-genotoxic isolates (i.e. CM1 and ECSC054), no mutation disrupting the *pks* genes was identified (Fig. 7; data not shown), suggesting that mutations located outside the *pks* island could negatively impact its expression. To test this hypothesis, we used plasmids pASK-clbR and pBAD-clbR, both overexpressing the *pks* regulator ClbR, and introduced either of them into strain CM1 which was susceptible to antibiotic, in contrast to ECSC054. In the resulting CM1 transformants, colibactin activity was restored, as seen by the formation of DNA ICLs (data not shown). Thus, in this strain, the lack of genotoxic activity probably resulted from a negative regulation of the *pks* island through an unknown mechanism.

The functionality of the *pks* island was also examined through analysis of the lipid metabolite profiles of selected genotoxic (n=3) and non-genotoxic (n=8) *pks*-positive strains, and in particular for production of C14-Asn, which was used as an indicator of the activity of the *pks* biosynthesis machinery. This lipopeptide is synthesized during the initial step of the biosynthesis process involving ClbN and ClbB, prior to elongation and final cleavage through the involvement of ClbC-H-I-J-K and ClbP, respectively [17]. The production of

C14-Asn was detected in all of the ICL-positive strains examined, SI-NP020, JML285 and KS-NP019 (*ca*. 650–1200 pg/10^8 c.f.u.) but not in the ICL-negative strain UPEC129 mutated in *clbB* (Fig. S5, Table 3). Interestingly, C14-Asn was detected (*ca*. 400–600 pg/10^8 c.f.u.) in three ICL-negative strains carrying a *clbJK* fusion gene (SI-NP032, JML296 and NS-NP030) and in two ICL-negative strains carrying a mutated *clbD* gene (JML114 and JML165). The two non-genotoxic strains carrying intact *clb* genes (ECSC054 and CM1) produced either a very low level or no detectable C14-Asn, respectively (Fig. S5, Table 3). For strain CM1 transformed with either plasmid pASK-clbR or pBAD-clbR (see above), overexpression of ClbR restored the production of C14-Asn (Fig. S5). These results suggest that even when the *pks* island does not allow production of active colibactin, enzymes from the *pks* pathway still produce metabolites with potential biological activities.

DISCUSSION

Acquisition of the *pks* island in the population of *E. coli* appears to have involved two distinct mechanisms differing by the presence or absence of a phage-type integrase. The integrase-mediated *pks* insertion pathway occurred mainly in B2 strains and resulted in *pks* insertion into either of three *asn* tRNA genes (i.e. *asnU*, *asnV* or *asnW*). This potential for integration into several DNA targets is consistent with the observed conservation and genetic integrity of the *pks* integrative module, i.e. the integrase gene and the two direct repeats flanking the island. The flexibility of *pks* insertion is reminiscent of what has been described for the HPI of *Yersinia pseudotuberculosis* which is also able to insert into either of the three *Y. pseudotuberculosis asn* tRNA genes [41], in contrast to the immobile truncated form of the HPI in *E. coli* whose right direct repeat is deleted and whose location is fixed at the *asnT* tRNA gene [42]. A divergent integrase sequence was found for the *pks* island inserted into the *asnU* tRNA gene compared to those inserted into the *asnV* or *asnW* tRNA gene. As the three *asn* tRNA sequences are 100% identical, the use of either of them as an attachment site by slightly different integrases probably reflects distinct histories of *pks* acquisition. After *pks* chromosomal integration, the endogenous *pks* integrase promoter is replaced by the promoter of the upstream *asn* tRNA gene (Fig. S6), a configuration similar to that found for the HPI integrase promoter [43]. Whether the site of integration influences the expression of the *pks* integrase and hence *pks* stability at the distinct *asn* tRNA loci is not known. In contrast to the integrase-mediated pathway, the *pks* chromosomal integration process in the B1 and A *E. coli* strains remains unclear as no site-specific recombinase-encoding gene was found near *pks* and chromosomal insertion occurred into a non-tRNA *locus*. In these strains, *pks* integration could have involved the participation of IS elements such as the IS*66* whose truncated or intact copies were found to flank the *pks* island.

The co-phylogeny analysis between the core genome- and *pks*-based phylogenetic trees shed further light on *pks* acquisition scenarios. In the case of the typical *pks*-positive B2 strains belonging to lineages from major CCs (i.e. CC12, CC14, CC73 and CC95), the congruence observed between both trees suggested that the *pks* island was horizontally acquired by the MRCA of these lineages, or by the MRCA of each of these, and then stably maintained in their descendants through vertical transmission. The fact that strains of certain CC95 subgroups lack the *pks* island probably suggests that *pks* was lost during the evolution of these sublineages. Such a loss might be closely linked to the change in the relative fitness of CC95 subgroups underlying the variations observed in their spatial and temporal distribution in several continents [35]. In the case of B1 or A strains, *pks* acquisition and dissemination probably occurred through sporadic lateral transfer events, as *pks*-positive B1 or A strains were scarce and not genetically related. The horizontal transferability of the *pks* island has previously been demonstrated using an *in vitro* approach where *pks* could be transferred together with the HPI via F' plasmid-mediated conjugation from a donor to a recipient *E. coli* strain [44]. We propose that *pks* acquisition by the single *pks*-positive D strain ECSC054 was mediated by HGT, presumably from a B2 donor strain given the *pks* sequence relatedness observed between the D and B2 strains. HGT was also probably involved in the exchange of the *pks* island between the three atypical B2 ST73 or ST95 strains and a (yet unknown) phylogenetically distant donor strain, since their *pks* sequences did not cluster with those from other B2 ST73 or ST95 strains. This hypothesis was further supported by the identification, in these three isolates, of an ICE-like element inserted in their *pks* island. Similar ICE-like elements have previously been identified in three B1 *E. coli* isolates and other members of the *Enterobacteriacae* such as *C. koseri*, *Enterobacter aerogenes* and *K. pneumoniae* [6]. They could therefore play a role in *pks* dissemination in enterobacteria, as proposed for the self-transmissible ICE linked to the HPI identified in the *E. coli* strain ECOR31 [38]. Due to its lack of a complete DNA mobilization region (region II), we assume that the ICE-linked *pks* island is no longer self-transferable. It might nevertheless correspond to a remnant of an ancient, complete and self-transmissible ICE-linked *pks* island that could have behaved as a large complex ICE and spread in enterobacteria before undergoing partial or entire deletion of the ICE region. To date, no bacterial strain carrying a complete ICE linked to *pks* has been identified and the origin of the *pks* island therefore remains elusive.

The ecological niche and/or genetic background of the bacterial strains probably had an impact on the acquisition and stable maintenance of *pks*. The high concentration of *pks*-positive strains in some CCs of the B2 group such as CC73 and CC95 suggests that *pks* might have contributed to their ecological and evolutionary success. CC73 and CC95 exhibit a similar phylogenetic history and are major ExPEC lineages, especially prior to the year 2000 where they were the most commonly detected [1, 45]. They are persistent intestinal colonizers and successful extra-intestinal pathogens with the particularity of exhibiting lower multidrug resistance levels compared to other ExPEC lineages. By contrast, as our collection contained *E. coli* B2 strains from

STs other than ST73 and ST95, it was interesting to note that *pks* was absent from STs corresponding to separate B2 lineages in the *E. coli* phylogenetic tree, including the ST131 clonal complex which is associated with multidrug resistance and is now the most predominantly isolated ExPEC lineage worldwide [45]. Consistent with the hypothesis of a *pks* acquisition by the MRCA(s) of CC73 and CC95 mentioned above, this finding suggests that such acquisition probably occurred after they diverged from the MRCA of CC131, i.e. before going through distinct evolutionary trajectories. The inversion of the upstream *asnW-asnU-asnV* tRNA-containing region which probably accompanied *pks* insertion into the *asnW* tRNA gene in the B2 group might have contributed to *pks* stabilization at this locus. Since the various *pks*-positive and *pks*-negative B2 lineages occupy the same ecological niche (i.e. primarily the intestinal tract of humans and animals), horizontal transfer of *pks* between them could have been expected, at least to some extent, but which was not revealed here. Several hypotheses can be proposed to explain this. First, some barriers to HGT might exist between members of distinct CCs, such as restriction-modification systems [46]. Second, *pks* might have been transferred to recipient strains without providing adaptive value, thus resulting in its rapid loss. Third, as crosstalk between virulence determinants and the chromosome backbone is required for the emergence of virulent clones [1], a specific chromosomal phylogenetic background might be required for appropriate *pks* expression and production of an adaptive value, thereby constituting a prerequisite for the stable maintenance of the island.

The structure of the *pks* island is very well conserved among the *E. coli* population, with more than 99% identity, suggesting that its integrity remains under strong structural and functional evolutionary constraints. We can speculate that transcription and translation of the 19 *pks* genes of this 54 kb long genomic island would be too high for the bacterial strains if the *pks* island did not bring a selective advantage to them. This is reinforced by the fact that only 18 out of 109 *pks*-positive strains lacked genotoxic activity. The importance of the biological role of *pks* is highlighted by the numerous activities associated with this genetic island, including genotoxicity, anti-inflammatory activity, antibiotic and analgesic effects. Given its interplay with siderophores (enterobactin, salmochelin and yersiniabactin) and siderophore-microcins (MccM and MccH47) [10, 22, 47], the *pks* island contributes to bacterial competition through the acquisition of iron or the production of inhibitory compounds, respectively. Protection of bacterial cells from genomic degradation through the production of ClbS could also be advantageous to *pks*-carrrying strains, as this multifunctional protein not only directly inactivates colibactin but also protects bacterial DNA from nucleolytic degradation by nucleases [48]. We also observed that non-genotoxic *E. coli* strains carrying an altered *pks* island still produced the prodrug motif C14-Asn synthesized at the early stage of the biosynthesis process, suggesting that yet-to-be-discovered bioactive compounds are produced by these strains. Given the high conservation observed for the *pks* island in *E. coli*, we can thus speculate that colibactin is a very important genotoxin but that *pks*-derived synthesis of other secondary metabolites could also be an advantage for *E. coli*.

Although our collection is characterized by a large diversity of *E. coli* strains from various phylogenetic groups and STs, one limitation of this study is that only strains from Japan were included, which may not be representative of the distribution of *pks* in a global collection of *E. coli* isolates from worldwide sources.

In conclusion, the various genetic configurations of the *pks* island and its distribution in the *E. coli* phylogenetic tree imply the existence of various scenarios for the introduction and spread of *pks* into the *E. coli* population. The presence of a functional *pks* island was demonstrated for the majority of the *pks*-positive strains, suggesting that the *pks* island is under selective pressure for the adaptation of *E. coli* to various ecological niches, through the production of colibactin or other secondary metabolites.

Funding information
This work was supported by funding from the National French Institute of Health and Medical Research (INSERM) to Camille Chagneau and the région Occitanie (grant ALDOCT-000610) and Ministère de l'Agriculture to Alexandre Perrat.

Acknowledgements
We thank Claire Hoede and Sarah Maman (SIGENAE group) and the GENOTOUL bioinformatics platform for providing computational resources. We also thank Pauline Le Faouder and the METATOUL lipidomic platform for their support in the analysis of the lipid metabolite profiles.

Author contributions
Conceptualization: F.A., T.H., P.B., Y.O., E.O. Methodology: F.A., T.H., P.B., Y.O., E.O. Validation: F.A., T.H., P.B., Y.O., E.O. Formal Analysis: F.A., A.P., C.C., Y.A., T.H., P.B., Y.O., E.O. Investigation: F.A., A.P., Y.A., C.C., N.B.G., C.M., J.P.N., P.B., Y.O., E.O. Resources: T.H., Y.O., E.O. Data Curation: F.A., A.P., Y.A., T.H., P.B., Y.O., E.O. Writing – Original Draft: F.A., T.H., Y.O., E.O. Writing - Review and Editing: F.A., A.P., C.C., H.B., J.P.N., T.H., P.B., Y.O., E.O. Visualization: F.A., Y.A., C.C., J.P.N., P.B., Y.O., E.O. Supervision: F.A., H.B., T.H., Y.O., E.O. Project Administration: F.A., E.O. Funding Acquisition: T.H., Y.O., E.O.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Denamur E, Clermont O, Bonacorsi S, Gordon D. The population genetics of pathogenic *Escherichia coli*. *Nat Rev Microbiol* 2021;19:37-54.
2. Tenaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal *Escherichia coli*. *Nat Rev Microbiol* 2010;8:207–217.
3. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. *Nature* 2000;405:299–304.
4. Kaper JB, Nataro JP, Mobley HL. Pathogenic *Escherichia coli*. *Nat Rev Microbiol* 2004;2:123–140.
5. Nougayrède HS, Homburg S, Taieb F, Boury M, Brzuszkiewicz E, Gottschalk G *et al*. *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. *Science* 2006;313:848–851.
6. Putze J, Hennequin C, Nougayrède JP, Zhang W, Homburg S *et al*. Genetic structure and distribution of the colibactin genomic island among members of the family Enterobacteriaceae. *Infect Immun* 2009;77:4696–4703.
7. Engel P, Vizcaino MI, Crawford JM. Gut symbionts from distinct hosts exhibit genotoxic activity via divergent colibactin biosynthesis pathways. *Appl Environ Microbiol* 2015;81:1502–1512.
8. Bondarev V, Richter M, Romano S, Piel J, Schwedt A *et al*. The genus Pseudovibrio contains metabolically versatile bacteria adapted for symbiosis. *Environ Microbiol* 2013;15:2095–2113.
9. Marcq I, Martin P, Payros D, Cuevas-Ramos G, Boury M *et al*. The genotoxin colibactin exacerbates lymphopenia and decreases

survival rate in mice infected with septicemic *Escherichia coli*. *J Infect Dis* 2014;210:285–294.

10. Martin P, Marcq I, Magistro G, Penary M, Garcie C *et al*. Interplay between siderophores and colibactin genotoxin biosynthetic pathways in *Escherichia coli*. *PLoS Pathog* 2013;9:e1003437.
11. McCarthy AJ, Martin P, Cloup E, Stabler RA, Oswald E *et al*. The genotoxin colibactin is a determinant of virulence in *Escherichia coli* K1 experimental neonatal systemic infection. *Infect Immun* 2015;83:3704–3711.
12. Arthur JC, Perez-Chanona E, Mühlbauer M, Tomkovich S, Uronis JM *et al*. Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* 2012;338:120–123.
13. Cougnoux A, Dalmasso G, Martinez R, Buc E, Delmas J *et al*. Bacterial genotoxin colibactin promotes colon tumour growth by inducing a senescence-associated secretory phenotype. *Gut* 2014;63:1932–1942.
14. Cuevas-Ramos G, Petit CR, Marcq I, Boury M, Oswald E *et al*. *Escherichia coli* induces DNA damage in vivo and triggers genomic instability in mammalian cells. *Proc Natl Acad Sci U S A* 2010;107:11537–11542.
15. Bossuet-Greif N, Vignard J, Taieb F, Mirey G, Dubois D *et al*. The colibactin genotoxin generates DNA interstrand cross-links in infected cells. *mBio* 2018;9:e02393-17 [Epub ahead of print 20 03 2018].
16. Taieb F, Petit C, Nougayrède JP, Oswald E. The enterobacterial genotoxins: cytolethal distending toxin and colibactin. *EcoSal Plus* 2016;7.
17. Brotherton CA, Balskus EP. A prodrug resistance mechanism is involved in colibactin biosynthesis and cytotoxicity. *J Am Chem Soc* 2013;135:3359–3362.
18. Wallenstein A, Rehm N, Brinkmann M, Selle M, Bossuet-Greif N *et al*. ClbR is the key transcriptional activator of colibactin gene expression in *Escherichia coli*. *mSphere* 2020;5.
19. Shine EE, Xue M, Patel JR, Healy AR, Surovtseva YV *et al*. Model Colibactins exhibit human cell genotoxicity in the absence of host bacteria. *ACS Chem Biol* 2018;13:3286–3293.
20. Vizcaino MI, Engel P, Trautman E, Crawford JM. Comparative metabolomics and structural characterizations illuminate colibactin pathway-dependent small molecules. *J Am Chem Soc* 2014;136:9244–9247.
21. Pérez-Berezo T, Pujo J, Martin P, Le Faouder P, Galano JM *et al*. Identification of an analgesic lipopeptide produced by the probiotic *Escherichia coli* strain Nissle 1917. *Nat Commun* 2017;8:1314.
22. Massip C, Branchu P, Bossuet-Greif N, Chagneau CV, Gaillard D *et al*. Deciphering the interplay between the genotoxic and probiotic activities of *Escherichia coli* Nissle 1917. *PLoS Pathog* 2019;15:e1008029.
23. Dubois D, Delmas J, Cady A, Robin F, Sivignon A *et al*. Cyclomodulins in urosepsis strains of *Escherichia coli*. *J Clin Microbiol* 2010;48:2122–2129.
24. Johnson JR, Johnston B, Kuskowski MA, Nougayrede JP, Oswald E. Molecular epidemiology and phylogenetic distribution of the *Escherichia coli* pks genomic island. *J Clin Microbiol* 2008;46:3906–3911.
25. Arimizu Y, Kirino Y, Sato MP, Uno K, Sato T *et al*. Large-scale genome analysis of bovine commensal *Escherichia coli* reveals that bovine-adapted *E. coli* lineages are serving as evolutionary sources of the emergence of human intestinal pathogenic strains. *Genome Res* 2019;29:1495–1505.
26. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;13:e1005595.
27. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S *et al*. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;31:3691–3693.
28. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T *et al*. *SNP-sites*: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* 2016;2:e000056.
29. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 2006;22:2688–2690.
30. Letunic I, Bork P. Interactive tree of life (iTOL) V3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* 2016;44:W242–W245.
31. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M *et al*. Versatile and open software for comparing large genomes. *Genome Biol* 2004;5:R12.
32. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* 2016;33:1870–1874.
33. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol* 2012;3:217–223.
34. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG *et al*. Act: the ARTEMIS comparison tool. *Bioinformatics* 2005;21:3422–3423.
35. Gordon DM, Geyik S, Clermont O, O'Brien CL, Huang S *et al*. Fine-Scale structure analysis shows epidemic patterns of clonal complex 95, a cosmopolitan *Escherichia coli* lineage responsible for extraintestinal infection. *mSphere* 2017;2
36. Bossuet-Greif N, Belloy M, Boury M, Oswald E, Nougayrede J-P. Protocol for HeLa cells infection with *Escherichia coli* strains producing colibactin and quantification of the induced DNA-damage. *BIO-PROTOCOL* 2017;7:e2520.
37. Tronnet S, Oswald E. Quantification of Colibactin-associated genotoxicity in HeLa cells by in cell Western (ICW) using γ-H2AX as a marker. *Bio Protoc* 2018;8:e2771.
38. Schubert S, Dufke S, Sorsa J, Heesemann J. A novel integrative and conjugative element (ICE) of *Escherichia coli:* the putative progenitor of the Yersinia high-pathogenicity island. *Mol Microbiol* 2004;51:837–848.
39. Fabian NJ, Mannion AJ, Feng Y, Madden CM, Fox JG. Intestinal colonization of genotoxic *Escherichia coli* strains encoding colibactin and cytotoxic necrotizing factor in small mammal pets. *Vet Microbiol* 2020;240:108506.
40. Kurnick SA, Mannion AJ, Feng Y, Madden CM, Chamberlain P *et al*. Genotoxic *Escherichia coli* strains encoding Colibactin, Cytolethal Distending Toxin, and Cytotoxic necrotizing factor in laboratory rats. *Comp Med* 2019;69:103–113.
41. Buchrieser C, Brosch R, Bach S, Guiyoule A, Carniel E. The high-pathogenicity island of *Yersinia pseudotuberculosis* can be inserted into any of the three chromosomal Asn tRNA genes. *Mol Microbiol* 1998;30:965–978.
42. Schubert S, Darlu P, Clermont O, Wieser A, Magistro G *et al*. Role of intraspecies recombination in the spread of pathogenicity islands within the *Escherichia coli* species. *PLoS Pathog* 2009;5:e1000257.
43. Rakin A, Noelting C, Schropp P, Heesemann J. Integrative module of the high-pathogenicity island of *Yersinia*. *Mol Microbiol* 2001;39:407–416.
44. Messerer M, Fischer W, Schubert S. Investigation of horizontal gene transfer of pathogenicity islands in *Escherichia coli* using next-generation sequencing. *PLoS One* 2017;12:e0179880.
45. Manges AR, Geum HM, Guo A, Edens TJ, Fibke CD *et al*. Global extraintestinal pathogenic *Escherichia coli* (ExPEC) lineages. *Clin Microbiol Rev* 2019;32
46. Oliveira PH, Touchon M, Rocha EP. Regulation of genetic flux between bacteria by restriction-modification systems. *Proc Natl Acad Sci U S A* 2016;113:5658–5663.
47. Massip C, Chagneau CV, Boury M, Oswald E. The synergistic triad between microcin, colibactin, and salmochelin gene clusters in uropathogenic *Escherichia coli*. *Microbes Infect* 2020;22:144–147.
48. Molan K, Podlesek Z, Hodnik V, Butala M, Oswald E *et al*. The *Escherichia coli* colibactin resistance protein ClbS is a novel DNA binding protein that protects DNA from nucleolytic degradation. *DNA Repair* 2019;79:50–54.