BMC Genomics

RESEARCH

Open Access

Genomic characterization of *Escherichia coli* with a polyketide synthase (*pks*) island isolated from ulcerative colitis patients

Chao Lv[1,2,3†], Mohd Abdullah[4†], Chun-Li Su[5], Weiye Chen[1,2], Nan Zhou[1,2], Zile Cheng[1,2], Yiwen Chen[1,2], Min Li[1,2], Kenneth W. Simpson[5], Ahmed Elsaadi[6], Yongzhang Zhu[1,2,7*], Steven M. Lipkin[6*] and Yung-Fu Chang[4*]

Abstract

The *E. coli* strains harboring the polyketide synthase (*pks*) island encode the genotoxin colibactin, a secondary metabolite reported to have severe implications for human health and for the progression of colorectal cancer. The present study involves whole-genome-wide comparison and phylogenetic analysis of *pks* harboring *E. coli* isolates to gain insight into the distribution and evolution of these organisms. Fifteen *E. coli* strains isolated from patients with ulcerative colitis (UC) were sequenced, 13 of which harbored *pks* islands. In addition, 2,654 genomes from the public database were also screened for *pks* harboring *E. coli* genomes, 158 of which were *pks*-positive (*pks*$^+$) isolates. Whole-genome-wide comparison and phylogenetic analysis revealed that 171 (158 + 13) *pks*$^+$ isolates belonged to phylogroup B2, and most of the isolates belong to sequence types ST73 and ST95. One isolate from a UC patient was of the sequence type ST8303. The maximum likelihood tree based on the core genome of *pks*$^+$ isolates revealed horizontal gene transfer across sequence types and serotypes. Virulome and resistome analyses revealed the0020preponderance of virulence genes and a reduced number of antimicrobial genes in *pks*$^+$ isolates. This study significantly contributes to understanding the evolution of *pks* islands in *E. coli*.

Keywords Colibactin, Colorectal cancer, Genome sequencing, Phylogenetics

†Chao Lv and Mohd Abdullah contribution and share first authorship.

*Correspondence:
Yongzhang Zhu
yzhzhu@126.com
Steven M. Lipkin
stl2012@med.cornell.edu
Yung-Fu Chang
yc42@cornell.edu
[1]School of Global Health, Chinese Center for Tropical Diseases Research, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[2]National Institute of Parasitic Diseases, Chinese Center for Disease Control and Prevention (Chinese Center for Tropical Diseases Research), National Health Commission Key Laboratory of Parasite and Vector Biology, WHO Collaborating Centre for Tropical Diseases, National Center for International Research on Tropical Diseases, Shanghai 200025, China
[3]School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[4]Department of Population Medicine and Diagnostic Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY, USA
[5]Graduate Program of Nutrition Science, School of Life Science, National Taiwan Normal University, Taipei, Taiwan
[6]Department of Clinical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY, USA
[7]Sanford and Joan Weill Department of Medicine, Weill Cornell Medical School, Cornell University, New York City, USA

Introduction

Specific *Escherichia coli* (*E. coli*) strains are pathogenic microbes that inhabit the gut of animals and humans and are associated with intestinal and extraintestinal infections [1, 2]. The diverse population of *E. coli* is widely distributed into eight major phylogenetic groups (A, B1, B2, C, D, E, F, and G) [3]. A major population of *E. coli* belongs to phylogroup B2, which causes severe infections such as urinary tract infections (UTI), sepsis, pneumonia, and neonatal meningitis [4]. There are multiple reasons for the evolution of virulence in *E. coli*, but a major role is played by horizontal gene transfer, point mutation, and inactivation of antivirulence genes [5, 6]. Virulence factors such as toxins, adhesins, capsules, and iron acquisition systems are often encoded by genes that can be mobilized through various methods, including mobile genetic elements, genomic islands, phages, and plasmids. Horizontal gene transfer allows the widespread distribution of these genes in extraintestinal pathogenic *E. coli* (ExPEC) strains [7, 8]. Genomic islands, large regions of more than 10 kb often bounded by repetitive structures and carry mobility factors such as integrases and transposes, exhibit associations with tRNA genes and have diverse G+C contents [9]. Pathogenicity islands (PAIs), a small subgroup of genomic islands, play pivotal roles in the evolution of bacterial virulence by incorporating virulence-associated factors and adaptive horizontal gene transfer [10, 11]. Colibactin, encoded by a PAI named pks, is recognized as a nonribosomal peptide-polyketide secondary metabolite and is observed in commensal strains of *E. coli* and strains associated with urinary tract infections and neonatal meningitis [12]. Colibactin can induce double-stranded DNA breaks in eukaryotic cells, leading to cell cycle arrest at the G_2-M phase and chromosomal aberrations [13, 14]. It significantly contributes to severe clinical manifestations such as meningitis [15] and sepsis [12].

Colibactin, known for inhibiting extraintestinal pathogen *E. coli* (ExPEC), is also suspected of being a procarcinogen factor [14, 16, 17]. Compared to healthy individuals, elevated colibactin-producing *E. coli* strains are found in colorectal cancer (CRC) patients [18]. Several recent studies suggest that certain strains of *E. coli* that possess the *pks* island may play a causal role in the development of human CRC [17, 19, 20]. The genes of the *pks*$^+$ island are significantly enriched in CRC patients, in familial adenomatous polyposis (FAP), and in DNA mismatch repair-deficient CRC patients [21, 22]. According to preclinical studies, pks+*E. coli* drives tumorigenesis and increases tumor burden in several CRC and FAP mouse models [21, 23, 24]. Notably, colibactin-induced DNA damage creates a specific mutational signature in CRC tumors that can be computationally monitored and used to measure the contribution of *pks*$^+$*E. coli* to tumor burden [25, 26].

The biosynthesis machinery of colibactin is located on the *pks* island, spanning a region of 54 kb and housing 19 genes. These genes included nonribosomal peptide mega synthases (NRPSs; *clb*H, *clb*J, and *clb*N), polyketide mega synthases (*PKSs*; *clb*C, *clb*I, and *clb*O), two hybrid NRPS-PKSs (*clb*B and *clb*K), and nine accessory and tailoring enzymes [13]. The presence of the *pks* island is not confined to pathogenic organisms; it has also been observed in commensal and probiotic bacterial strains [27]. Its presence extends beyond *E. coli*, encompassing members of the Enterobacteriaceae family, such as *Citrobacter koseri*, *Klebsiella pneumoniae*, and *K. aerogenes* [28]. The association between *pks*-positive (*pks*$^+$) *E. coli* and CRC is evident in biopsy samples, revealing an elevated prevalence of *pks* + island-harbouring *E. coli* [17, 29]. Notably, these isolates are found in more than half of patients with familial adenomatous polyps and contribute to carcinogenesis through mucus degradation, adherence, and enhanced colonization within colonic biofilms [30]. In addition to their speculated role in CRC progression, *pks* islands serve as virulence factors with clinical implications, contributing to systemic infection, neonatal meningitis, and lymphopenia, according to various studies [31–33].

In this study, we performed whole-genome sequencing (WGS) of 15 *E. coli* isolates from patients with ulcerative colitis (UC) and performed genome-wide comparisons and phylogenetic analysis of *pks* islands harboring *E. coli* isolates from 15 UC strains and 2654 datasets from the NCBI database. The study describes the distribution of *pks*$^+$*E. coli* among phylogroups, STs, and serogroups, followed by core and pangenome analysis. A phylogenomic study was also performed on the core genome to understand island acquisition and evolution. The antibiotic resistance genes and virulence genes were mined to understand the drug resistance and virulence characteristics of *pks* harboring *E. coli* isolates.

Materials and methods

The genomic DNA of the 15 strains

A total of 15 *E. coli* strains from UC patients were used in this study, and the strains were previously reported in 2004 [34]. The genomic DNA of the 15 strains was extracted using the TIANamp Bacteria D.N.A. Kit (TIANGEN, Beijing, China). Subsequently, the DNA libraries were prepared using the KAPA HyperPrep Kit (Roche, Basel, Switzerland) following the manufacturer's instructions and sequenced on the Illumina NovaSeq platform with a 150 bp paired-end strategy. Furthermore, the strain HM229 was subjected to long-read sequencing using an Oxford Nanopore Technology (ONT.) MinION device.

The draft genomes were assembled using the PGCGAP pipeline with the SPAdes v3.13.1 algorithm, and the long-read genome sequences were assembled using the Unicycler v0.5.0 algorithm (https://gitee.com/liaochenlanruo/pgcgap) [35]. To explore the genomic characteristics of *E. coli* isolates harboring the *pks* island, an additional dataset of 2654 complete genomes of *E. coli* was downloaded from the NCBI (deadline 2022.12.31, Table S1). Moreover, the strains in the downloaded dataset included information on the host, disease, geographic location, and collection date.

Identification of the *pks* island in *E. Coli* strains

The reference sequence of the *pks*$^+$ island (GenBank accession number: AM229678.1) was downloaded from NCBI and used as a query file to perform BLASTn searches against the genomes of *E. coli* strains from UC patients and genome sequences downloaded from NCBI. The identity and query coverage thresholds of BLASTn searches were 85%.

Phylogenetic analysis of *pks*$^+$*E. coli* strains

The phylogroups of the *E. coli* genomes (2654 downloaded strains and 15 strains from UC patients from this study) were determined using ClermonTyping (https://github.com/A-BN/ClermonTyping) [36]. Sequence typing (ST) was performed using MLST (https://github.com/tseemann/mlst) [37]. ECTyper (https://github.com/phac-nml/ecoli_) was used to perform in silico serotyping of the genomes [38]. Subsequently, the *pks*$^+$ strains were filtered, and a minimum spanning tree was generated based on the STs in PHYLOViZ 2.0 (https://www.phyloviz.net/) using the goeBURST algorithm [39]. After annotation by Prokka (http://vicbioinformatics.com/) [40], the software Roary 3.11.2 (http://sanger-pathogens.github.io/Roary/) [41] was used to determine the core genes of the *pks*$^+$ strains. The core genome-based phylogenetic tree was subsequently constructed using FastTree 2.1 (http://meta.microbesonline.org/fasttree/) [42], which infers an approximately maximum likelihood algorithm with generalized time-reversible (GTR) models. The core-gene-based phylogenetic tree was visualized using Interactive Tree Of Life (iTOL, https://itol.embl.de/).

Virulome and resistome profiling of *pks*$^+$ strains from UC patients

To explore the characteristics of the virulome and resistome of the *pks*$^+$ strains from UC patients, a total of 102 isolates (including *pks*$^+$ and *pks*$^-$; Table S2) were selected from among the 2669 strains following the established standards: (1) Based on the STs of UC patient strains, 10 *pks*$^-$ and *pks*$^+$ strains were selected randomly if the STs had more than 10 *pks*$^-$ and *pks*$^+$ strains, such as ST95; (2) Ten *pks*$^-$ and *pks*$^+$ strains were selected if the STs only had more than 10 *pks*$^+$ or *pks*$^-$ strains, such as ST127, ST73, ST453, and ST131; (3) The strains were all selected if the STs had fewer than 10 strains, such as ST141.

Then, 102 genomic core gene-based phylogenetic trees were constructed as above to display the phylogeny of the *pks*$^-$ and *pks*$^+$ strains. All the assemblies were screened for antimicrobial resistance genes (ARGs) and virulence genes (VGs) using Resfinder 4.0 [43] and the Virulence Factor Database (VFDB) [44] using Abricate (https://github.com/tseemann/ abricate). The numbers of ARGs and VGs in various comparison groups (*pks*$^+$ strains from UC patients versus (*VS*) *pks*$^+$ strains from others and *pks*$^-$ strains Table S1) were visualized using boxplots and dot plots generated with ggplot2 v3.3.2 in R 4.3.3.

Whole-genome alignment of eight *pks*$^+$ strains

A multiple genome alignment tool called Mauve (https://darlinglab.org/mauve/user-guide/screenshots.html) was used to construct and visualize the whole chromosomal alignment of the selected eight UCs strains belonging to different phylogroups. Mauve compares multiple genome sequences and finds regions of homology called locally collinear blocks (LCBs). The progressive Mauve algorithm was used with the default parameters.

Analysis *of the pks* island structure

The HM229 strain was compared to seven other selected strains (Table S3), which belonged to distinct phylogroups, based on *pks* island structure analysis, with the IHE3034 strain serving as the reference. First, the sequence of the *pks* island, along with 10k bp of its upstream and downstream regions, was extracted from the whole genome and then annotated by Prokka (http://vicbioinformatics.com/) to accrue the GBK format. The GBK files of the 8 strains were subsequently submitted to the software Easyfig2.2.5 (https://mjsull.github.io/Easyfig/) to create linear comparison figures of multiple genomic loci with an easy-to-use graphical user interface [45].

SNP analysis of whole genomes and *pks* genes

For SNP analysis, the 7 selected strains were mapped to the genome of HM229 by the snippy program (https://github.com/tseemann/snippy). The recombinant region was removed from the resulting alignment by the Gubbins program, and then core SNPs were extracted by the SNP-sites program. In addition to the SNP analysis of the whole genome of the 8 strains, the sequence of the *pks* island was also extracted after BLASTn searches to conduct SNP analysis of the other UC strains.

Statistical analysis

Categorical data were analyzed via the chi-square test. Continuous data with normal or nonnormal distributions

Table 1 The occurrence of the *pks* island in *E. Coli* strains from humans with different diseases**

Human diseases	No. pks$^-$	No. pks$^+$	Positive rate (%)	*P* value
Ulcerative colitis	2	13	86.67	
Urinary tract infection	71	12	14.46	< 0.001*
Bacteremia	63	5	7.35	< 0.001*
Diarrhea	62	1	1.59	< 0.001*
Sepsis	24	2	7.69	< 0.001*
Gastroenteritis	14	0	0.00	-
Hemolytic uremic syndrome	10	0	0.00	-
Hemorrhagic colitis	2	0	0.00	-
Healthy status	58	11	15.94	< 0.001*

*The positive rate was compared with that of ulcerative colitis

** The data came from the 546 strains with host disease information

were analyzed using a *t*-test or Mann–Whitney U test. For comparisons of multiple groups, an analysis of variance (ANOVA) or Kruskal–Wallis H test was used. All the statistical analyses were performed in IBM SPSS Statistics 25 (IBM, Armonk, USA).

Results

Genomic analysis of *pks* islands in *E. Coli* strains

Blastn revealed that 13 of the 15 (86.67%) *E. coli* strains from UC patients harbored the *pks* island, and 158 *pks*$^+$ strains (158/2654, 5.95%) were in the genomic dataset downloaded from the NCBI (Table S1); these strains included strains collected from humans (8.98%, 103/1147), food animals (2.07%, 9/435), wildlife (3.85%, 7/182), environmental samples (2.37%, 5/211), companion animals (6.67%, 4/60), food samples (0.82%, 1/122), marine organisms (14.29%, 1/7) and undefined sources (4.99%, 20/401). Moreover, the *pks*$^+$ percentage of strains from patients with urinary tract infections (12/83, 14.46%), bacteremia (5/68, 7.35%), diarrhea (1/63, 1.59%), sepsis (2/26, 7.69%), cystitis (2/14, 14.29%), gastroenteritis (0/14, 0.00%), hemolytic uremic syndrome (0/10, 0.00%), and healthy status (11/69, 15.94%) were significantly lower than that of UC patients (Table 1).

Distribution of *pks*$^+$*E. coli*

The phylogroup analysis showed that all 13 *pks*$^+$ strains from UC patients belonged to phylogroup B2 (100%). Additionally, the predominant occurrence of the *pks* island in phylogroup B2 was observed in 158 genomes downloaded (95.57%, 151/158; Table 2). The 13 *pks*$^+$ strains from UC. patients belonged to 6 different STs, namely ST12, ST73, ST95, ST127, ST141, and ST8303 (Table 2). Notably, the strains attributed to ST12 (30.77%, 4/13) exhibited the highest quantity. Among the 158 downloaded *pks*$^+$ strains, ST73 (29.75%, 47/158), ST95 (22.78%, 36/158) and ST127 (15.19%, 24/158) were the dominant STs. Notably, ST8303 represented a novel ST type among all 171 *pks*$^+$ strains. Interestingly, the strains belonging to ST73, ST127, and ST998 all harbored *pks* islands (Table 2). The minimum spanning tree showed that the 13 *pks*$^+$ strains from UC patients were mainly assigned to three clonal complexes (CC): CC12, CC95, and CC73, similar to that of the 158 *pks*$^+$ genomes downloaded (Fig. 1).

Table 2 Distribution of the *pks*$^+$ strains according to sequence type (ST.)

ST	Phylogroup					Total
	A	B1	B2	D	F	
UC patients (*n* = 13)						
12			4			4
73			3			3
95			2			2
127			2			2
141			1			1
8303			1			1
Downloaded genomes (*n* = 158)						
12			8			8
14			2			2
38				1		1
73			47			47
80			1			1
95			36			36
104			1			1
109		1				1
127			24			24
141			6			6
278		1				1
392		2				2
416			1			1
537			1			1
550			1			1
643			1			1
969			1			1
998			11			11
1193			1			1
1231			1			1
1262			2			2
1278			1			1
2604			1			1
6134			1			1
7010	1					1
10,522			1			1
Unknown			2		1	3
Total	**1**	**4**	**164**	**1**	**1**	**171**

The strains from the UC patients belonged to 10 different serotypes, with no predominant serotypes identified. Among the 158 downloaded strains, O6:H1 (25.95%, 41/158), O6:H31 (13.92%, 22/158), and O18:H7 (13.92%, 22/158) were the three predominant serotypes (Table S1, Fig. 2). Additionally, from the phylogenetic tree, the predominant serotype in CC73 was O6:H1 (74.55%, 41/55),

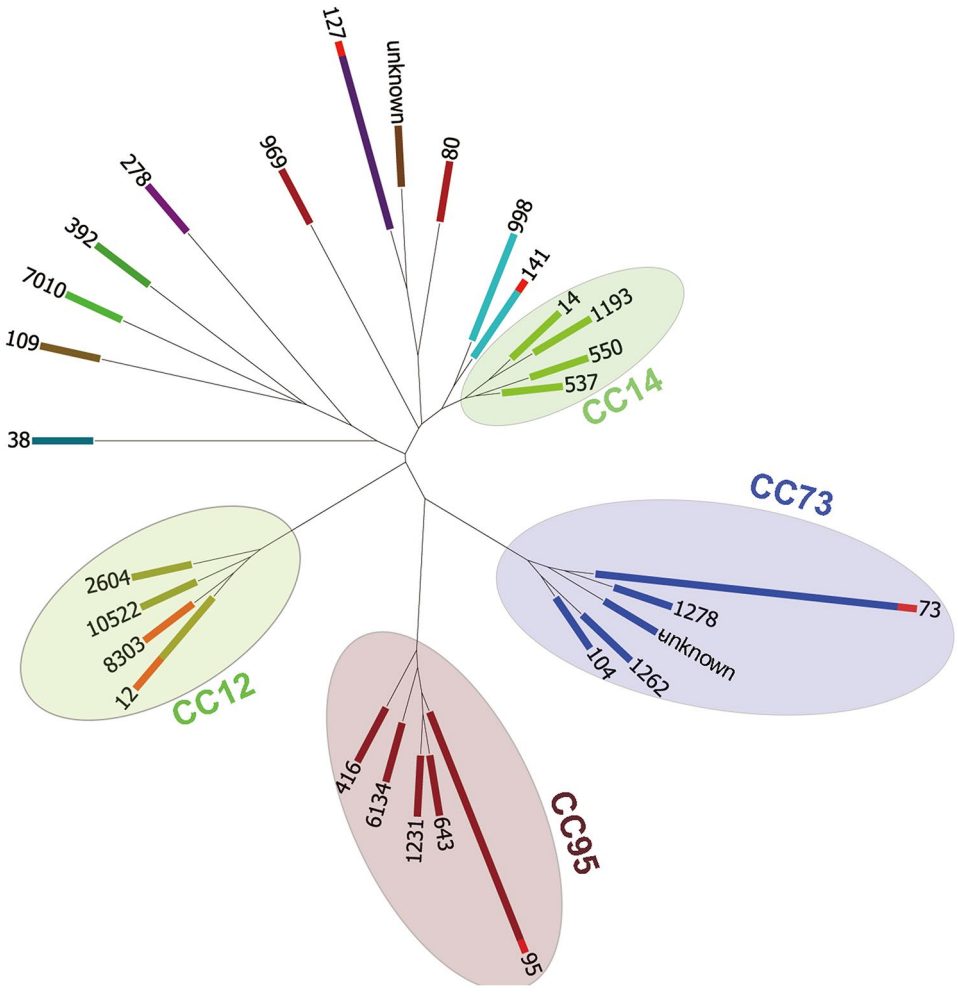

Fig. 1 The minimum spanning tree based on the STs of all *pks*$^+$ strains ($n=171$). The end of the bars were the corresponding STs. The four colored shaded circles represent different clonal complexes (CCs): CC12 (green), CC14 (light green), CC73 (blue), and CC95 (red). The length of the bar represents the quantity of strains for STs. The red parts of the bars of ST12, ST141, ST73, ST95, ST12 (orange), ST8303 (orange) represent the strains isolates from ulcerative colitis (UC) patients

the predominant serotype in CC95 was O18:H7 (51.22%, 21/41), and that in CC12 was O4:H5 (76.92%, 10/13).

Phylogenetic analysis of *pks*$^+$*E. coli*

The core genome maximum-likelihood phylogeny was obtained from Fast-Tree. The 13 *pks*$^+$ strains from UC patients were distributed into 5 different clades (we named clades 1–5). Five strains (5/13) were assigned to clade 3, which was the predominant branch of CC12 (Fig. 2). Interestingly, the clades of the 13 strains identified by core-gene-based phylogeny were restricted to the clusters shown in the minimum spanning tree based on the STs. Among the five clades, the strains from humans constituted the largest proportion of the strains, the strains from wildlife were mainly attributed to clade 4, and the strains from companion animals were primarily attributed to clade 3 (CC12). Among the serotypes, O6:H31, O18:H17, O4:H5, O2:H6, and O6:H1 were the predominant serotype from clade 1 to clade 5, respectively (Fig. 2). Moreover, 2 of the 3 *pks*$^+$ isolates from UC patients belonged to clade 5 (CC73), which contained another two serotypes (O25:H1, O2/O50:H1); the serotypes of UC patient-derived *pks*$^+$ strains were not the predominant serotype in clade 2 (CC12) or 3 (CC95) (Fig. 2).

Pangenome analysis of *pks*$^+$*E. coli* strains

The median number of core genes associated with the *pks*$^+$ strains from UC patients was 3470.00 (IQR: 3467.00-3473.00), which was slightly lower than that associated with the *pks*$^+$ strains downloaded from NCBI (3475.00, IQR: 3466.00-3479.00, $P=0.088$; Mann–Whitney U test; Fig. S1A). However, the median number of accessory genes associated with the *pks*$^+$ strains from UC patients (1333.00, IQR:1185.00-1395.00) was greater than that associated with the *pks*$^+$ strains downloaded from NCBI

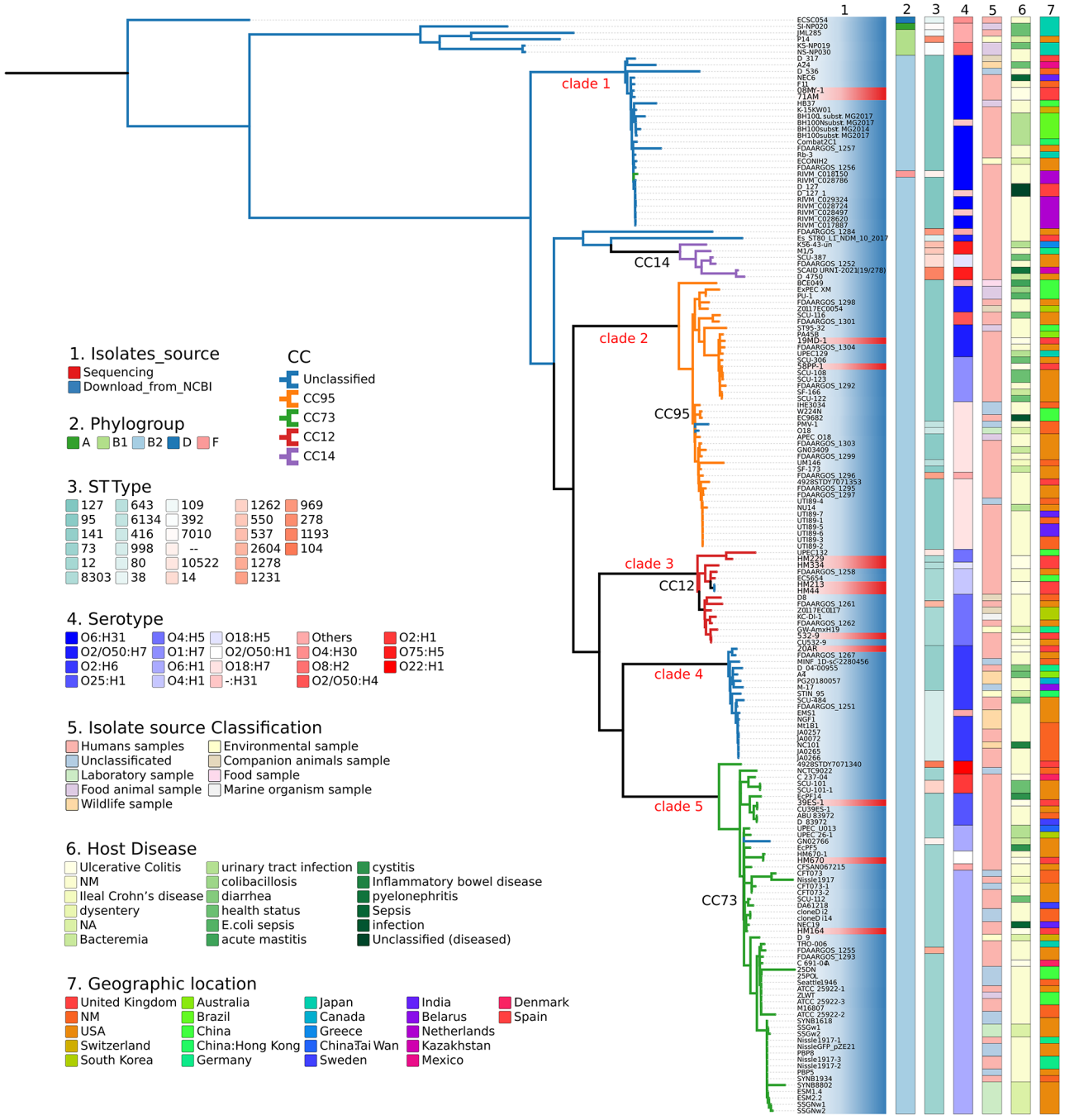

Fig. 2 Phylogenetic inference based on the analysis of core genes of 171 *pks*$^+$ strains. The core gene-based ML tree was constructed based on 3482 core genes which obtained from Roary. The branch colors were different clone complexes (CCs): unclassified (blue), CC95 (orange), CC73(green), CC12 (red), and CC14 (purple). The clade 1–5 was artificially labeled by obvious branch clusters containing the strains from ulcerative colitis (UC) patients. Column 1–7 were the strains' name with different colors referring to different sources (Legend 1), the phylogroups (Legend 2), the STs (Legend 3), the serotypes (Legend 4), the classification of isolate source (Legend 5), the host disease of the 171 strains (Legend 6) and the geographic source (Legend 7) of the 171 strains. NM (column 6): not mentioned; NA: not applicable

(1316.00, IQR: 1191.00-1453.00), and the distribution was also not different (P=0.877, Mann–Whitney U test; Fig. S1B).

The 54 kb *pks* island contains 19 genes (*clb*A to-S) encoding biosynthetic machinery. A heatmap of the pangenome analysis revealed four gene deletions, namely, *clb*J, *clb*H, *clb*M and *clb*I, in the *pks* island region of strains from UC patients (except for the HM229 strain) compared with most *pks*$^+$ isolates and the reference isolate IHE3034 (Fig. 3). Additionally, the UC

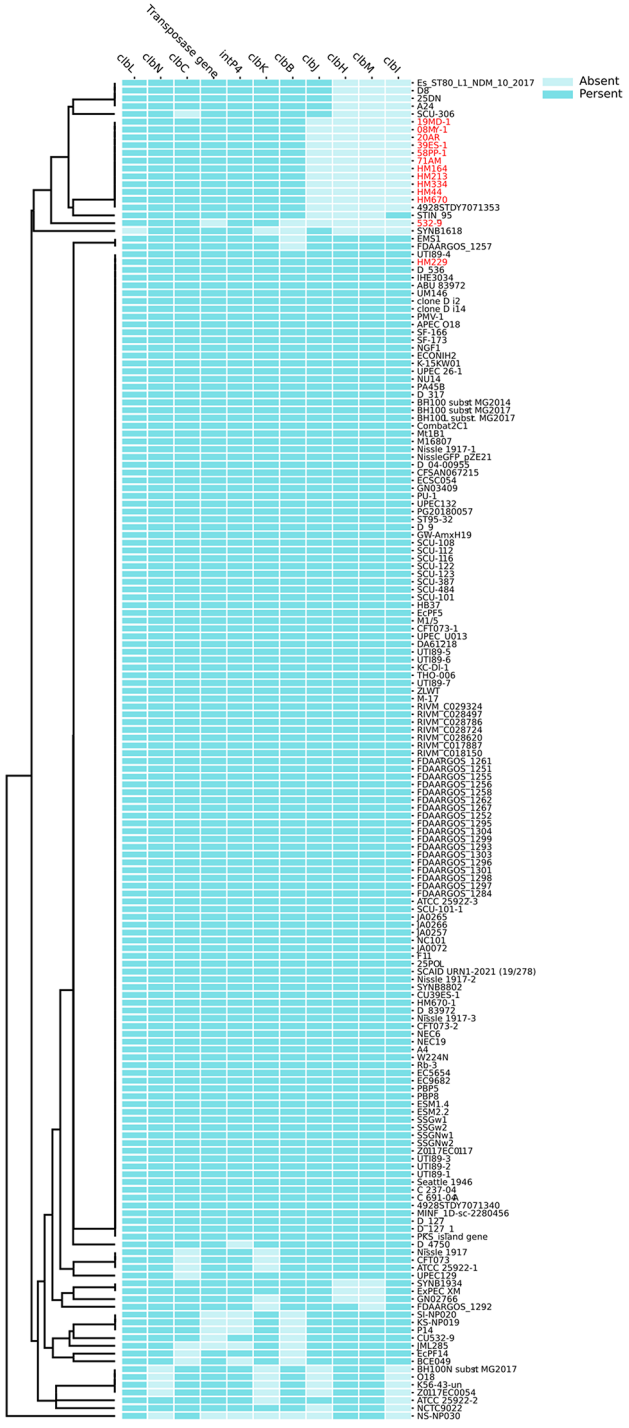

Fig. 3 The heatmap of the presence or absence of the genes of the *pks* island filter from the 171 *pks*$^{+}$*E. coli* isolates. The red color of the isolates name indicates the isolates from UC patients. The method of heatmap clustering is based on the complete Euclidean distance. The *pks* island genes of *clb*A, *clb*D, *clb*E, *clb*F, *clb*G, *clb*P, *clb*R, *clb*Q, two hypothetical protein genes, IS1400 and IS1351were presented in all 171 strains

patient-derived strain 532-9 had two additional deletions of *clb*B and the putative transposase gene. According to previous research, the *pks* island genes are involved in enzymatic interactions, and *clb*J, *clb*H, *clb*I, and *clb*B are involved in the assembly of mega synthase nonribosomal peptide synthetase (NRPS) and polyketide synthase (PKS) genes, while the *clb*M gene effluxes precolibactin, which is unloaded from the aforementioned assembly line through the periplasm.

Virulome and resistome analysis of *pks*$^{+}$*E. coli* strains

Based on the phylogenetic tree of the core genes of the 102 selected *E. coli* strains, the *pks*$^{+}$ or *pks*$^{-}$ strains demonstrated a clear clustering pattern (Fig. 4A). The number of VGs in the *pks*$^{+}$ strains, including strains from UC patients (median: 94, IQR:83.5-102.5) and other sources (median:99, IQR:95, IQR:87–108) were significantly greater than those of *pks*$^{-}$ strains (median: 80, IQR:80-88.25, $P=0.001$ and $P<0.001$, Mann–Whitney *U* test, Fig. 4B). The *pks*$^{+}$ strains from UC patients had significantly greater percentages of VGs encoding adherence factors (*papC*, *papD*, *papF*, *papH*, *papJ*, *papK*, *sfaB*, *sfaC*, *sfaD*, *sfaE*, *sfaF*, *sfaG*, *sfaX*, and *sfaY*); nutritional/metabolic *factors (ChuA*, *chuS*, *chuT*, *chuU*, *chuV*, *chuW*, *chuX*, and *chuY*); the effector delivery system (*vat*); invasion factor (*aslA*); exotoxin (*hlyA*, *hlyB*, *hlyC*, *hlyD*, and *cnf1*); and immune modulation factor (*tcpC*). Moreover, the percentages of effector delivery system genes *espX4*, *espX5*, and *espL1* in the *pks*$^{-}$ strains were greater than those in the *pks*$^{+}$ strains from the UC patients (27.78% vs. 0.00%, $P=0.045$; chi-square test; Fig. 4C; Table S4).

Among the 13 *pks*$^{+}$ strains from UC patients, the maximum ARGs count was 7, with a median of 3 (IQR:1–4). In contrast, the *pks*$^{-}$ strains ($n=36$)showed a maximum ARGs count of 29 and a median of 5 (IQR:1–9, $P=0.025$, Mann–Whitney *U* test, Fig S2). Detailed information regarding the quantity and types of ARGs in all strains can be found in Table S5.

Conserved genomic blocks in the chromosome of HM229

All of the eight *pks*$^{+}$*E. coli* strains selected shared a common core genome of approximately 3.54 Mb organized into linear conserved blocks (LCBs). Except for strain RIVM_C018150, the results of this comparative alignment showed that the chromosomes of the compared strains were organized into 20 LCBs (Fig. 5). Overall, the LCB of HM229 was identical to that of 4 of the 7 selected strains, including the strain ECSC054 from phylogroup D, highlighting the conserved genomic skeleton of the *pks*$^{+}$*E. coli* strains. The *pks* island was in LCB 6, although the location of LCB 6 in the genome underwent a translocation in certain strains, including NS_NP030, RIVM_C018150, and SI-NP020. NS_NP030 and SI-NP020 were isolated from bovines and belonged to phylogroups B1

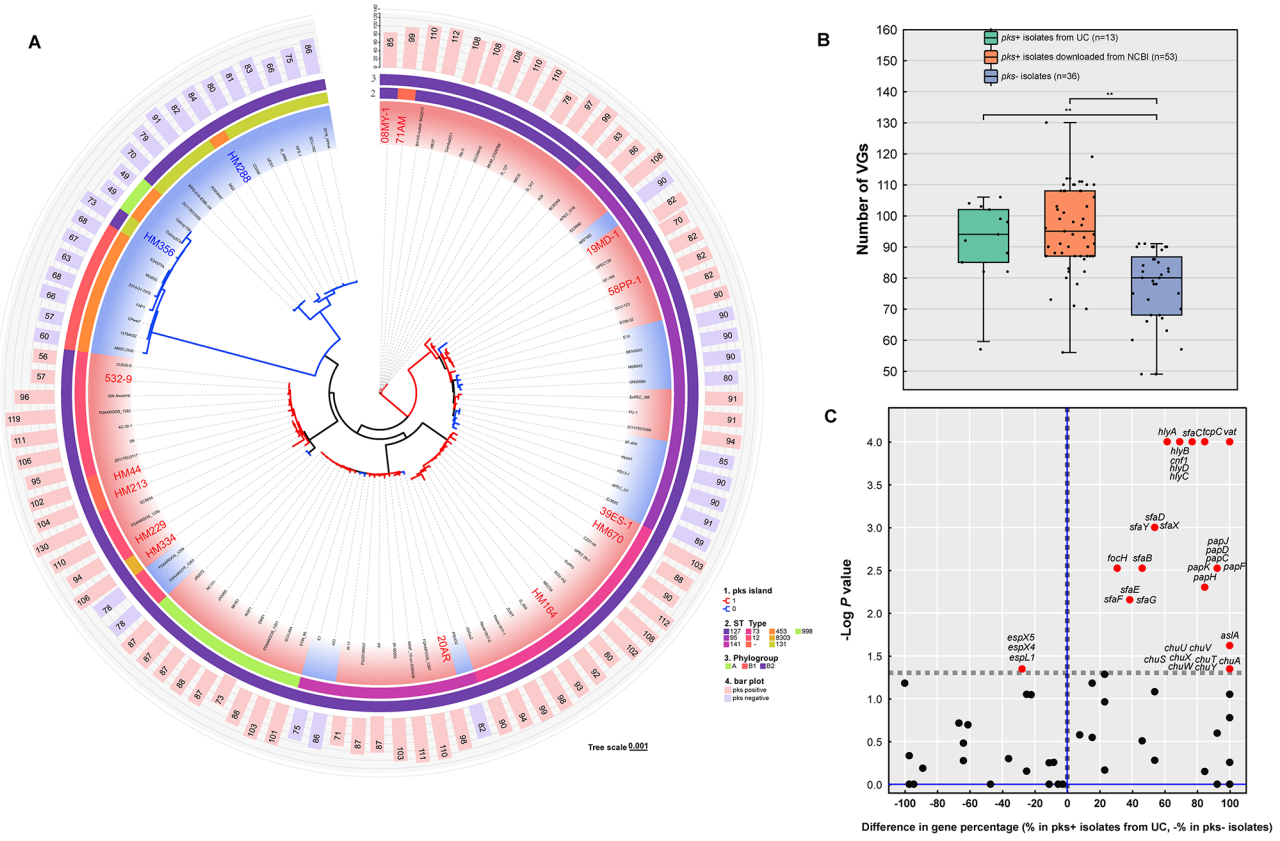

Fig. 4 Virulome profiling of UC patient pks + E. coli strains, the other pks + strains and pks- strains. The results were obtained from 102 selected E. coli strains. A: the core gene-based phylogenetic tree with the number of VGs. The tree branches with red color were pks + strains, the tree branches with blue color were pks- strains (Legend 1, 0: pks- strains, 1: pks + strains), Column 2 was ST, and Column 3 was phylogroup of those strains. The outer bar was the number of virulence genes (VGs); B: the box plot of the VGs, * P value < 0.05, ** P value < 0.001; C: the difference in VGs between the UC patient pks + E. coli strains ($n = 13$) and the pks- E. coli strains ($n = 36$), the dashed line indicates P value < 0.05 and the red plots was the gene with significant differences (P value < 0.05)

and A, respectively. RIVM_C018150 was from a clinical sample and belonged to phylogroup F. The varying locations, which underwent significant rearrangements among LCBs (NS_NP030, RIVM_C018150), suggests that the introduction of pks island may have led to substantial recombination in the genomic skeleton of non-phylogroup B2 strains.

Conserved *pks* islands in HM229

The fundamental structure of the *pks* island encompasses tRNA-Asn, *int*P4, *clb*A to -S, two *IS* (insertion sequence) of *IS*1400 and *IS*1351, and a hypothetical protein gene (Fig. 6A). The genetic backbone is highly conserved, with no differences observed in the 10 K bp regions upstream and downstream of certain strains. Importantly, these strains belong to distinct phylogenetic groups, including phylogroup B2 (IHE3034, HM229, STIN_95), D (ECSC054) and F (RIVM_C018150). *IS*1300 and *IS*1351 are members of the IS3 family transposase, indicating that this family likely plays a significant role in the transfer of the *pks* island. The 10 K bp upstream region of the *pks* island in HM229 encompasses genes of fosfomycin resistance protein (*aba*F), AMP nucleosidase (*amn*), transcriptional regulator (*yee*N, *glt*C and *cbl*), multidrug efflux transporters (*yee*O), and tRNA-Asn, while the 10 K bp downstream region encompasses genes of transpeptidase (*erf*K), ribazoletransferase (*cob*S), adenosylcobalamin biosynthesis protein (*cob*U) and colicin I receptor (*cir*A).

The *pks* island of SI-NP020 (phylogroup A, ST7010), NS_NP030 (B1, ST392) and k56-43-un (B2, ST537) exhibited certain differences compared to that of HM229. Consequently, we conducted further exploration of the *pks* islands in these three strains (Fig. 6B). The strains SI-NP020 and NS_NP030 had deletions of tRNA-Asn, the integrase gene of intP4 and a putative transposase gene. Additionally, the strains of K56_43_un and NS_NP030 were recognized as variants of the *clb*K-J fusion.

Single nucleotide polymorphisms (SNPs) in *pks* island

The snippy program identified SNPs in the 8 selected strain genomes and *pks* island sequences. There were

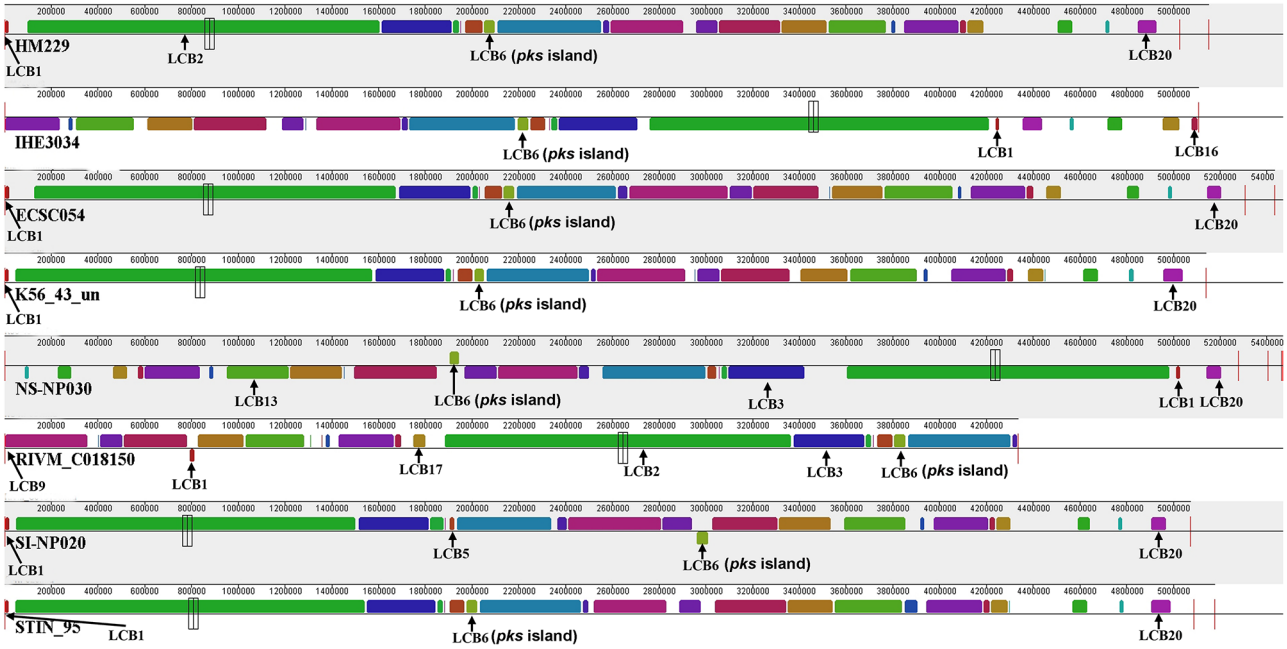

Fig. 5 Whole-genome alignments of the species created by Mauve. The colored rectangles represent LCBs. The sizes of the rectangles are proportional to the genomic extensions of the LCBs. The isolate HM229 was used as a reference, and the LCBs were ordered according to the reference

2722 core SNPs among the 8 strains, and the number of core SNPs among the 8 *pks* island sequences was 0. When comparing HM229 and the reference isolate IHE3034, 31,030 SNP sites were identified, and and only 8 SNP sites were found in the pks island sequence. The number of SNP sites in the pks islandof SI-NP020 (phylogroup A) and NS-NP030 (phylogroup B1) were 82 and 94, respectively, which is significantly greater than those of the other strains. Interestingly, the two strains were isolated from bovines.

Discussion

Colibactin, which is produced by pks islands, and has been identified in specific Enterobacteriaceae members, has emerged as an essential virulence factor implicated in the progression of CRC, meningitis, and septicaemia [12]. Many past studies have reported colibactin's involvement in CRC, as it plays an important role in the interaction between host cells and the microbiota during the progression of CRC, suggesting that colibactin is an important virulence factor with far-reaching implications [46].

In the present study, whole-genome-wide comparisons of *E. coli* isolates from an in-house culture collection and from a public database were performed to obtain insight into pks island acquisition and evolution. The in-house genome collection, derived from human UC patients (n=15), revealed that 13 of these genomes harbour the *pks* island. The scale of the study was further broadened by including 2654 *E. coli* genomes from the NCBI database, unveiling the presence of *pks* islands in 158 isolates. The subsequent phylogenomic analysis revealed that 13 *pks*$^+$ isolates from the in-house culture collection and 158 genomes from the public database belonged to phylogroup B2. This aligns seamlessly with earlier research findings [28, 47–49]. The 171 genomes exhibiting *pks* positivity were subjected to comprehensive in silico typing techniques to discern distribution patterns across diverse *E. coli* subtypes. Notably, the majority of the 13 *pks*$^+$ isolates were identified as belonging to ST12 (30.77%, 4/13) or ST73 (23.07%, 3/13). Conversely, the 158 *pks*$^+$ isolates sourced from NCBI showed dominance in ST73 (29.75%, 47/158), ST95 (22.78%, 36/158), and ST127 (15.19%, 24/158). These findings align consistently with outcomes reported in earlier investigations. We meticulously crafted a phylogenetic tree to determine the circumstances surrounding the acquisition of *pks* sequences by *pks*$^+$*E. coli*. Our findings illuminated a clustering of the core genome primarily within lineages of *the* CC12, CC14, CC95, and CC73 clonal complexes. This finding supports the hypothesis that the introduction of the *pks* island into CC12, CC14, CC73, and CC95 occurred through horizontal acquisition by their most recent common ancestor, subsequently followed by vertical transmission with gradual *pks* divergence over time [50]. Furthermore, a pangenome analysis of *pks*$^+$*E. coli* strains revealed that the core genome size was not significantly different between UC patients and those in the downloaded NCBI dataset. The heatmap shows that the *pks* island genes *clb*J, *clb*H, *clb*M, and *clb*I are missing

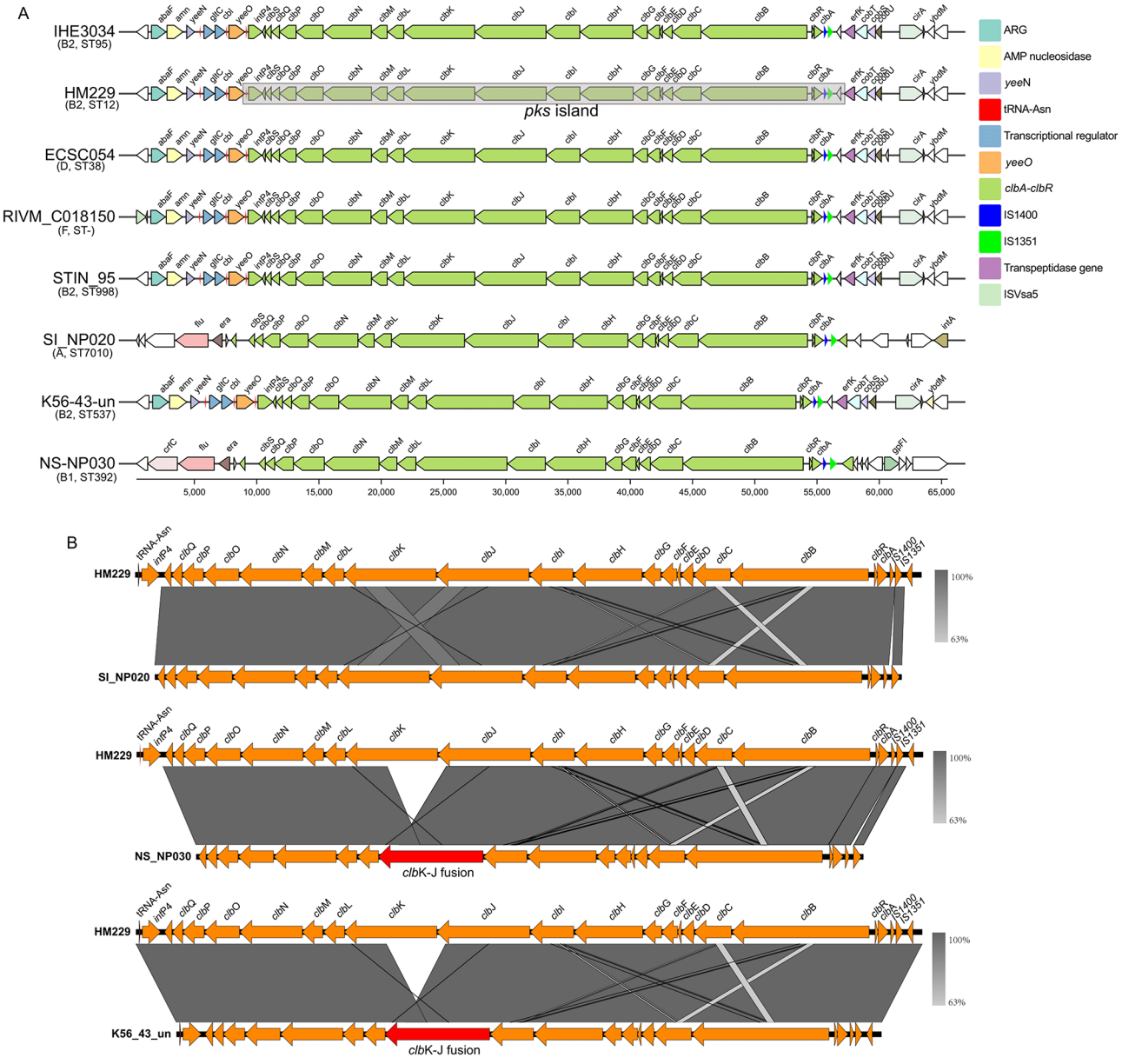

Fig. 6 The *pks* island structure of strains belonging to different phylogroups. **A** the *pks* island and the 10 K bp regions upstream and downstream of pks island; **B** the compariosn between HM229 and the other 3 *pks*$^+$ strains. IHE3034 was the reference isolate, HM229 was the isolate from UC patients, and ECSC054, C018150, SI-NP020, and NS_NP030 were the strains belonging to phylogroups D, F, A, and B1, respectively

from *pks*$^+$*E. coli* strains from UC patients, except for the HM229 strain. The absence of these genes may be due to genomic deletion or mutation. Random mutations or events such as recombination can lead to the loss of specific gene sequences. Further investigation is needed to determine the effect of these genes deletions within the pks island on colibactin production, *E. coli* toxicity, the presence of *pks*$^+$ strains in the microbiota, and the pathological processes of UC and CRC.

Whole-genome-based virulome and resistome analyses revealed that 102 *E. coli* strains contained 72 antibiotic resistance genes among the *pks*- strains and 130 virulence genes among the *pks*$^+$ strains (Supplementary Table S4). On the basis of our study, it was found that, compared with pks-negative isolates, *pks*$^+$ isolates contain fewer antibiotic resistance genes and a greater number of virulence genes. Our results are in line with those of previous studies, which showed low levels of antibiotic resistance and high numbers of virulence factors in pks+isolates [51, 52]. Among the virulence genes identified in the *pks*$^+$ isolates, a significantly higher number of genes were associated with several adherence factors, an invasion factor, an exotoxin, and an immune modulation factor. However, the role of these VGs, if any in the pathogenesis of UC,

requires further research to elucidate. The large number of virulence genes identified in the *pks*$^+$ isolates is consistent with the findings of previous reports based on PCR-based observations of bacteremia isolates [48]. The comparative genomic alignment of 8 strains revealed 20 linear conservative blocks (LCBs), with HM229 having an identical orientation to 4 of the other seven strains, highlighting the conserved genomic skeleton of the *pks*$^+$ isolates. The high conservation of the *pks* island suggested that colibactin is an important genotoxin that provides a selective advantage for these microorganisms.

UC is a chronic inflammatory bowel disease that primarily affects the inner lining of the colon and rectum. Persistent chronic inflammation is a key risk factor for CRC development in patients with UC. Persistent inflammation can lead to repeated injury and regeneration of epithelial cells in the colon, and this unstable cellular environment may lead to DNA damage, gene mutations, and abnormal cell proliferation, thereby increasing the risk of cancer [53]. Generally, the longer the duration of UC, the greater risk of developing CRC, and patients with a history of the condition exceeding 10 years face a significantly elevated risk [54]. Several studies have confirmed the association between UC and CRC. For example, cohort studies and retrospective analyses have shown that the incidence rate of colon cancer in patients with UC is several times higher than that in the general population [55–57]. The high population levels of *pks*$^+$*E. coli* in UC patients may lead to greater levels of colibactin and progression from UC to CRC. Interestingly, research has shown that the sole presence of *pks*$^+$*E. coli* in the intestine appears to be insufficient to induce CRC in mice models. This highlights the crucial function of both altered colonic microbiota [58] and intestinal inflammation [59] in the development of the disease. Intestinal inflammation may promote *pks*$^+$*E. coli* proliferation, enhances *pks* genes transcription, increases attachment of bacteria to the mucosa, and enhances the formation of bacterial biofilms in contact with precancerous lesions [60]. The findings of this study serve as supportive evidence for the involvement of the pks island and pks+E. coli in the progression of UC.

This work has the following limitations: (1) more clinical samples are required to validate our conclusions, though sample collection poses certain challenges; (2) Only one strain underwent long-read sequencing to obtain a complete genome, while the remaining strains were analyzed using contigs; and (3) Clinical and experimental studies on the role of the pks islands in the pathogenicity of UC, are urgently needed.

Conclusion

The prevalence of colibactin-producing *E. coli* isolates was found to be very high in UC patients, and most of the *pks*$^+$ isolates belonged to Phylogroup B2. Identification of the presence of the pks island in specific E. coli strains may help in the diagnosis of UC and, at the same time, increase our understanding of the role of *pks*$^+$ isolates in the pathogenesis of both UC and CRC. The *pks* island phylogeny indicates that the *pks* island spread through horizontal gene transfer. Finally, the pks+isolates demonstrated high virulence gene content and low antibiotic resistance gene content.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-11198-x.

Supplementary Material 1: Fig. S1: Total number of core and accessory genes in *pks+E. coli* isolates. (A) core genes. (B) Accessory genes. Significant levels are labeled with an asterisk using the Mann–Whitney U test

Supplementary Material 2: Fig. S2: Resistome profiles of UC patient *pks+E. coli* strains ($n=13$), the other *pks+* strains ($n=53$) and *pks−* strains strains ($n=36$). * P value < 0.05; **** P value < 0.001

Supplementary Material 3

Supplementary Material 4: Table S1: Details of the isolates obtained from 15 patients with ulcerative colitis and additional isolates downloaded from the NCBI database.

Supplementary Material 5: Table S2: Serotypes of 171 pks-positive isolates.

Supplementary Material 6: Table S3: In-depth analysis of the virulome and resistome profiles of the 102 *E. coli* isolates.

Supplementary Material 7: Table S4: The percentages of virulence genes (VGs) and antibiotic resistance genes (ARGs) in *pks*$^+$ isolates from patients with ulcerative colitis (UC) and *pks*$^-$ isolates.

Supplementary Material 8: Table S5: Eight selected isolates were used for genome mapping and SNP analysis.

Acknowledgements
N/A.

Author contributions
YFC, CLS, STL, and YZ designed the study, CL, CLS, MA, WC, NZ, Zc, YC, AE, and ML data analysis underthe supervision of Y.F.C., S.M.L., and K.W.S. MA and CL wrote the first draft of themanuscript, with contributions from YFC and YZ. All authors read andapproved the final manuscript.

Funding
We acknowledge funding from NIH, R01 CA231283 to support this study.

Data availability
Data will be made available upon request.The sequenced genome were submitted tro NCBI database under Bioproject No: PRJNA1064993.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 8 February 2024 / Accepted: 30 December 2024
Published online: 08 January 2025

References

1. Pakbin B, Bruck WM, Rossen JWA. Virulence factors of enteric pathogenic *Escherichia coli*: a review. Int J Mol Sci. 2021;22:9922.
2. Garcia A, Fox JG. A one health perspective for defining and deciphering *Escherichia coli* pathogenic potential in multiple hosts. Comp Med. 2021;71:3–45.
3. Takahashi T, Shigematsu H, Shivapurkar N, Reddy J, Zheng Y, Feng Z, Suzuki M, Nomura M, Augustus M, Yin J, Meltzer SJ, Gazdar AD. Aberrant promoter methylation of multiple genes during multistep pathogenesis of colorectal cancers. Int J Cancer. 2006;118:924–31.
4. Tenaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal *Escherichia coli*. Nat Rev Microbiol. 2010;8:207–17.
5. Bliven KA, Maurelli AT. Antivirulence genes: insights into pathogen evolution through gene loss. Infect Immun. 2012;80:4061–70.
6. Denamur E, Clermont O, Bonacorsi S, Gordon D. The population genetics of pathogenic *Escherichia coli*. Nat Rev Microbiol. 2021;19:37–54.
7. Clermont O, Bonacorsi S, Bingen E. Rapid and simple determination of the *Escherichia coli* phylogenetic group. Appl Environ Microbiol. 2000;66:4555–8.
8. Ahmed N, Dobrindt U, Hacker J, Hasnain SE. Genomic fluidity and pathogenic bacteria: applications in diagnostics, epidemiology and intervention. Nat Rev Microbiol. 2008;6:387–94.
9. Hacker J, Kaper JB. Pathogenicity islands and the evolution of microbes. Annu Rev Microbiol. 2000;54:641–79.
10. Dobrindt U, Hochhut B, Hentschel U, Hacker J. Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol. 2004;2:414–24.
11. Groisman EA, Ochman H. Pathogenicity islands: bacterial evolution in quantum leaps. Cell. 1996;87:791–4.
12. Fais T, Delmas J, Barnich N, Bonnet R, Dalmasso G. Colibactin: more than a new bacterial toxin. Toxins (Basel). 2018;10:151.
13. Nougayrede JP, Homburg S, Taieb F, Boury M, Brzuszkiewicz E, Gottschalk G, Buchrieser C, Hacker J, Dobrindt U, Oswald E. *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. Science. 2006;313:848–51.
14. Cuevas-Ramos G, Petit CR, Marcq I, Boury M, Oswald E, Nougayrede JP. *Escherichia coli* induces DNA damage in vivo and triggers genomic instability in mammalian cells. Proc Natl Acad Sci U S A. 2010;107:11537–42.
15. McCarthy AJ, Martin P, Cloup E, Stabler RA, Oswald E, Taylor PW. The Genotoxin Colibactin is a determinant of virulence in *Escherichia coli* K1 experimental neonatal systemic infection. Infect Immun. 2015;83:3704–11.
16. Arthur JC, Perez-Chanona E, Muhlbauer M, Tomkovich S, Uronis JM, Fan TJ, Campbell BJ, Abujamel T, Dogan B, Rogers AB, et al. Intestinal inflammation targets cancer-inducing activity of the microbiota. Science. 2012;338:120–3.
17. Cougnoux A, Dalmasso G, Martinez R, Buc E, Delmas J, Gibold L, Sauvanet P, Darcha C, Dechelotte P, Bonnet M, et al. Bacterial genotoxin colibactin promotes colon tumor growth by inducing a senescence-associated secretory phenotype. Gut. 2014;63:1932–42.
18. Nouri R, Hasani A, Masnadi Shirazi K, Alivand MR, Sepehri B, Sotoudeh S, Hemmati F, Fattahzadeh A, Abdinia B, Ahangarzadeh Rezaee M. Mucosa-Associated *Escherichia coli* in Colorectal Cancer Patients and Control Subjects: Variations in the Prevalence and Attributing Features. Can J Infect Dis Med Microbiol. 2021;2021:2131787.
19. Wernke KM, Xue M, Tirla A, Kim CS, Crawford JM, Herzon SB. Structure and bioactivity of colibactin. Bioorg Med Chem Lett. 2020;30:127280.
20. Iftekhar A, Berger H, Bouznad N, Heuberger J, Boccellato F, Dobrindt U, Hermeking H, Sigal M, Meyer TF. Genomic aberrations after short-term exposure to colibactin-producing *E. Coli* transform primary colon epithelial cells. Nat Commun. 2021;12:1003.
21. Dejea CM, Fathi P, Craig JA-O, Boleij AA-O, Taddese RA-O, Geis AA-O, et al. Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria. Science. 2018;359:592–7.
22. Dohlman AB, Arguijo Mendoza D, Ding S, Gao M, Dressman H, Iliev ID, Lipkin SM, Shen X. The cancer microbiome atlas: a pancancer comparative analysis to distinguish tissue-resident microbiota from contaminants. Cell Host Microbe. 2021;29:281–e298285.
23. Arthur JC, Perez-Chanona E, Mühlbauer M, Tomkovich S, Uronis JM, Fan TJ, et al. Intestinal inflammation targets cancer-inducing activity of the microbiota. Science. 338:120–3.
24. Cougnoux A, Dalmasso G, Martinez R, Buc E, Delmas J, Gibold L, et al. Bacterial genotoxin colibactin promotes colon tumor growth by inducing a senescence-associated secretory phenotype. Gut. 2014.63:1932–42.
25. Dziubańska-Kusibab PJ, Berger HA-O, Battistini F, Bouwman BA-O, Iftekhar A, Katainen R, et al. Colibactin DNA-damage signature indicates mutational impact in colorectal cancer. Nat Med. 2020;26:1063–69.
26. Pleguezuelos-Manzano C, Puschhof J, Rosendahl Huber A, van Hoeck AA-O, Wood HA-O, Nomburg J, et al. Mutational signature in colorectal cancer caused by genotoxic pks(+) *E. coli*. Nature. 2020;580:269–273.
27. Massip C, Branchu P, Bossuet-Greif N, Chagneau CV, Gaillard D, Martin P, Boury M, Secher T, Dubois D, Nougayrede JP, Oswald E. Deciphering the interplay between the genotoxic and probiotic activities of *Escherichia coli* Nissle 1917. PLoS Pathog. 2019;15:e1008029.
28. Putze J, Hennequin C, Nougayrede JP, Zhang W, Homburg S, Karch H, Bringer MA, Fayolle C, Carniel E, Rabsch W, et al. Genetic structure and distribution of the colibactin genomic island among members of the family Enterobacteriaceae. Infect Immun. 2009;77:4696–703.
29. Buc E, Dubois D, Sauvanet P, Raisch J, Delmas J, Darfeuille-Michaud A, Pezet D, Bonnet R. High prevalence of mucosa-associated *E. Coli* producing cyclomodulin and genotoxin in colon cancer. PLoS ONE. 2013;8:e56964.
30. Dejea CM, Fathi P, Craig JM, Boleij A, Taddese R, Geis AL, Wu X, DeStefano Shields CE, Hechenbleikner EM, Huso DL, et al. Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria. Science. 2018;359:592–7.
31. Marcq I, Martin P, Payros D, Cuevas-Ramos G, Boury M, Watrin C, Nougayrede JP, Olier M, Oswald E. The genotoxin colibactin exacerbates lymphopenia and decreases survival rate in mice infected with septicemic *Escherichia coli*. J Infect Dis. 2014;210:285–94.
32. Secher T, Payros D, Brehin C, Boury M, Watrin C, Gillet M, Bernard-Cadenat I, Menard S, Theodorou V, Saoudi A, et al. Oral tolerance failure upon neonatal gut colonization with *Escherichia coli* producing the genotoxin colibactin. Infect Immun. 2015;83:2420–9.
33. Lu MC, Chen YT, Chiang MK, Wang YC, Hsiao PY, Huang YJ, Lin CT, Cheng CC, Liang CL, Lai YC. Colibactin contributes to the hypervirulence of pks(+) K1 CC23 *Klebsiella pneumoniae* in mouse meningitis infections. Front Cell Infect Microbiol. 2017;7:103.
34. Martin HM, Campbell BJ, Hart CA, Mpofu C, Nayar M, Singh R, Englyst H, Williams HF, Rhodes JM. Enhanced *Escherichia coli* adherence and invasion in Crohn's disease and colon cancer. Gastroenterology. 2004;127:80–93.
35. Liu HXB, Zheng J, Zhong H, Yu Y, Peng D, Sun M. Build a bioinformatics analysis platform and apply it to routine analysis of microbial genomics and comparative genomics. Protocol exchange, 2022.
36. Beghain J, Bridier-Nahmias A, Le Nagard H, Denamur E, Clermont O. ClermonTyping: an easy-to-use and accurate in silico method for Escherichia Genus strain phylotyping. Microb Genom. 2018;4:000192.
37. Jolley KA, Maiden MC. BIGSdb: scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. 2010;11:595.
38. Bessonov K, Laing C, Robertson J, Yong I, Ziebell K, Gannon VPJ, et al. ECTyper: in silico *Escherichia coli* serotype and species prediction from raw and assembled whole-genome sequence data. Microb Genom. 2021;7:000728.
39. Nascimento M, Sousa A, Ramirez M, Francisco AP, Carrico JA, Vaz C. PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. Bioinformatics. 2017;33:128–9.
40. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.
41. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31:3691–3.
42. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE. 2010;5:e9490.
43. Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V, Philippon A, Allesoe RL, Rebelo AR, Florensa AF, et al. ResFinder 4.0 for predictions of phenotypes from genotypes. J Antimicrob Chemother. 2020;75:3491–500.
44. Chen L, Zheng D, Liu B, Yang J, Jin Q. VFDB 2016: hierarchical and refined dataset for big data analysis–10 years on. Nucleic Acids Res. 2016;44:D694–697.
45. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. Bioinformatics. 2011;27:1009–10.

46. Chagneau CV, Garcie C, Bossuet-Greif N, Tronnet S, Brachmann AO, Piel J, Nougayrede JP, Martin P, Oswald E. The Polyamine Spermidine Modulates the Production of the Bacterial Genotoxin Colibactin. mSphere. 2019;4.
47. Sarshar M, Scribano D, Marazzato M, Ambrosi C, Aprea MR, Aleandri M, Pronio A, Longhi C, Nicoletti M, Zagaglia C, et al. Genetic diversity, phylogroup distribution and virulence gene profile of pks positive *Escherichia coli* colonizing human intestinal polyps. Microb Pathog. 2017;112:274–8.
48. 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;46:3906–11.
49. Dubois D, Delmas J, Cady A, Robin F, Sivignon A, Oswald E, Bonnet R. Cyclomodulins in urosepsis strains of *Escherichia coli*. J Clin Microbiol. 2010;48:2122–9.
50. Auvray F, Perrat A, Arimizu Y, Chagneau CV, Bossuet-Greif N, Massip C, et al. Insights into the acquisition of the pks island and production of colibactin in the *Escherichia coli* population. Microb Genom. 2021;7:000579.
51. Suresh A, Ranjan A, Jadhav S, Hussain A, Shaik S, Alam M, Baddam R, Wieler LH, Ahmed N. Molecular Genetic and Functional Analysis of pks-Harboring, Extra-intestinal Pathogenic *Escherichia coli* from India. Front Microbiol. 2018;9:2631.
52. Suresh A, Shaik S, Baddam R, Ranjan A, Qumar S, Jadhav S, Semmler T, Ghazi IA, Wieler LH, Ahmed N. Evolutionary Dynamics Based on Comparative Genomics of Pathogenic *Escherichia coli* Lineages Harboring Polyketide Synthase (pks) Island. mBio. 2021;12.
53. Zhang H, Shi Y, Lin C, He C, Wang S, Li Q, Sun Y, Li M. Overcoming cancer risk in inflammatory bowel disease: new insights into preventive strategies and pathogenesis mechanisms including interactions of immune cells, cancer signaling pathways, and gut microbiota. Front Immunol. 2024;14:1338918.
54. Li J, Ji Y, Chen N, Dai L, Deng H. Colitis-associated carcinogenesis: crosstalk between tumors, immune cells and gut microbiota. Cell Biosci. 2023;13:194.
55. Zhan Y, Cheng X, Mei P, Wu J, Ou Y, Cui Y. Risk and incidence of colorectal stricture progressing to colorectal neoplasia in patients with inflammatory bowel disease: a systematic review and meta-analysis. Eur J Gastroenterol Hepatol. 2023;35:1075–87.
56. Albuquerque A, Cappello C, Stirrup O, Selinger CP. Anal high-risk human papillomavirus infection, squamous intraepithelial lesions, and Anal Cancer in patients with inflammatory bowel disease: a systematic review and Meta-analysis. J Crohns Colitis. 2023;17:1228–34.
57. Sato Y, Tsujinaka S, Miura T, Kitamura Y, Suzuki H, Shibata C. Inflammatory bowel Disease and Colorectal Cancer: Epidemiology, etiology, Surveillance, and management. Cancers (Basel). 2023;15:4154.
58. Tomkovich S, Yang Y, Winglee K, Gauthier J, Mühlbauer M, Sun X, Mohamadzadeh M, Liu X, Martin P, Wang GP, et al. Locoregional effects of Microbiota in a preclinical model of Colon carcinogenesis. Cancer Res. 2017;77:2620–32.
59. Arthur JC, Gharaibeh RZ, Mühlbauer M, Perez-Chanona E, Uronis JM, McCafferty J, Fodor AA, Jobin C. Microbial genomic analysis reveals the essential role of inflammation in bacteria-induced colorectal cancer. Nat Commun. 2014;5:4724.
60. Tang-Fichaux M, Branchu P, Nougayrède JP, Oswald E. Tackling the threat of Cancer due to pathobionts Producing Colibactin: is mesalamine the magic bullet? Toxins (Basel). 2021;13:897.

Publisher's note