scientific reports

Check for updates

OPEN

Genomic analysis of *Enterobacteriaceae* from colorectal cancer patients at a tertiary hospital in Ghana: a case-control study

Sarah V. Bachelle[1], Saikou Y. Bah[2], Richmond T. Addo[3], Antoinette A. A. Bediako-Bowan[4,5], Beverly Egyir[6], Sandra E. Tsatsu[4,5], Bartholomew Dzudzor[1]✉ & Vincent Amarh[1]✉

Colorectal cancer (CRC) is a severe gastrointestinal cancer and a leading cause of cancer-related deaths in Ghana. The potential role of gut *Enterobacteriaceae* in the increasing incidence of CRC in Ghana is yet to be thoroughly investigated. In this study, *Enterobacteriaceae* from CRC patients and healthy control participants were analyzed by whole genome sequencing to identify genomic features that are associated with CRC. Socio-demographic data showed a significant association between age and alcohol consumption and CRC. *Escherichia coli* was the most abundant *Enterobacteriaceae* isolated from the study participants and they were predominantly intestinal commensals. *Escherichia coli* isolates belonging to phylogroup D encoded the highest number of virulence genes. The *agn43* and *int* genes were widespread in *Escherichia coli* isolates from the CRC patients. Multilocus sequence types of potentially pathogenic *Escherichia coli* from the CRC patients also encoded genes involved in aggregation, adherence and biofilm formation. The *ampC2* and *ampH* antimicrobial resistance genes were also widespread in the genome of the *Escherichia coli* isolates. This study highlights the virulence tendencies of *Escherichia coli* from CRC patients and their ability to transfer virulence determinants to other *Enterobacteriaceae* residing in the gut.

Keywords Genomics, Colorectal cancer, *Enterobacteriaceae*, Virulence factors, Antimicrobial resistance

Colorectal cancer (CRC) is a leading cause of cancer-related morbidity and mortality, globally[1,2]. In Ghana, CRC is one of the prevalent gastrointestinal malignancies affecting the population[3]. The incidence of CRC is expected to rise in the next decade due to increasing Western habits in the daily lifestyle of the population in Ghana. Lifestyle factors that contribute to development of CRC include smoking, high intake of alcohol and processed meat, less physical activity, obesity and low intake of fruits and vegetables[4,5]. Family history and germline mutations in human DNA mismatch repair genes (*MLH1*, *MSH2*, MSH6, *PMS1* and *PMS2*) are genetic risk factors for CRC[2,6–10].

The gut microbiome have been reported to be associated with onset and progression of CRC via perturbation of host immune system and secretion of metabolites capable of triggering oncogenesis[11–16]. The *pks*-positive *Enterobacteriaceae* are notable gut bacteria that have been implicated in development of CRC due to their ability to synthesize colibactin, which is a genotoxin encoded by the *pks*-genomic island. Colibactin causes alkylation and crosslinks in DNA of host eukaryotic cells, leading to formation of DNA double-strand breaks, cell cycle arrest at G2 phase and intestinal tumorigenesis[12,17–19]. The *pks*-genomic island has been identified in *Escherichia coli*, *Klebsiella pneumoniae*, *Enterobacter aerogenes*, and *Citrobacter koseri*[20]. In *Escherichia coli*, the *pks*-genomic island is mainly found in isolates of the B2 phylogenetic group, and to a lesser extent in isolates belonging to phylogenetic groups A, B1 and D[21,22]. Gut bacteria also secrete metabolites such as cytotoxic necrotizing factors (Cnf), cycle-inhibiting factor (Cif) and cytolethal distending toxins (Cdt), which modulate proliferation,

[1]Department of Medical Biochemistry, University of Ghana Medical School, Korle-Bu, Accra, Ghana. [2]School of Infection & Immunity, University of Glasgow, Glasgow, UK. [3]Central Laboratory, Korle-Bu Teaching Hospital, Korle-Bu, Accra, Ghana. [4]Department of Surgery, University of Ghana Medical School, Korle-Bu, Accra, Ghana. [5]Department of Surgery, Korle-Bu Teaching Hospital, Korle-Bu, Accra, Ghana. [6]Bacteriology Department, Noguchi Memorial Institute for Medical Research, Accra, Ghana. ✉email: bdzudzor@ug.edu.gh; vamarh@ug.edu.gh

differentiation and apoptosis in host eukaryotic cells. Moreover, the CdtB protein causes DNA damage in the host and may predispose infected cells to oncogenesis[23].

Beyond secretion of cell cycle-modulating toxins, gut bacteria encode a myriad of virulence factors that have the potential to perturb vital cellular events and signaling pathways, ultimately contributing to initiation and progression of colorectal tumorigenesis[24]. For instance, *Escherichia coli* can attach intimately to intestinal mucosa using an adhesin protein intimin (encoded by *eae* gene) and cause downregulation of DNA mismatch proteins in colorectal cells leading to tumourigenesis[25]. Virulence genes encoding fimbriae in *Klebsiella pneumoniae* (*mrkCDF*) and *Escherichia coli* (*fimE*) have also been reported to be significantly associated with CRC[26]. *Escherichia coli* isolates exhibiting profound cytoadherence and expressing the *cnf1* gene were associated with CRC, which might indicate a potential role of adhesins in bacteria-mediated oncogenesis[27]. Genes encoding siderophores were also predominant in gut bacteria from early-stage CRC samples, possibly facilitating iron acquisition and bacterial survival in the host tissues[26,28,29]. Collectively, these observations from previous studies highlight an important role of bacterial virulence genes in instigating and promoting CRC.

Even though several studies have investigated CRC-associated microbiota[30–34], very limited data exist in the context of the Ghanaian population. Since the composition and diversity of gut bacteria varies across different geographical locations[35], it is essential to investigate gut microbial characteristics in CRC patients in the Ghanaian population in this era of increasing prevalence of CRC. In the present study, *Enterobacteriaceae* isolated from CRC patients and healthy control participants at a tertiary hospital in Ghana were analyzed via whole genome sequencing to identify novel and already known genomic features uniquely associated with gut bacteria from the cancer patients.

Results

Demographic information of the study participants

Demographic information was obtained from 58 CRC patients and 50 control participants (Table 1). The minimum ages recorded for the CRC patients and control participants were 16 years and 19 years, respectively, whiles the oldest participants were 88 years for the CRC patients, and 93 years for the control participants. The mean ages were 52.6 ± 15.0 years for the CRC patients and 48.9 ± 19.8 years for the control participants. None of the 108 participants reported a family history of CRC. Forty-two out of the 58 CRC participants had a habit of alcohol consumption. Only six CRC participants indicated they had a habit of smoking. Alcohol consumption was associated with increased likelihood of developing CRC among the participants [OR 2.625 (1.180–5.838), $p = 0.018$], but smoking habit was not significantly associated with the disease [OR 0.863 (0.260–2.868), $p = 0.810$]. Participants that were at least 30 years old also had an increased likelihood of developing CRC indicating that both young and older adults are prone to the disease [OR 4.706 (1.124–19.705), $p = 0.034$ for participants from 30 to 49 years; OR 4.493 (1.110-18.191), $p = 0.035$ for participants ≥ 50 years old].

Enterobacteriaceae isolated from the study participants

Stool samples were obtained from 47 CRC patients and 50 healthy control participants (Fig. 1). All stool samples from the study participants were inoculated onto MacConkey agar, which led to isolation of *Enterobacteriaceae* from 44 out of the 47 CRC patients (94%) whereas these isolates were obtained from 36 out of the 50 healthy control participants (72%). These observations indicate that *Enterobacteriaceae* were more readily detected in stool samples from CRC patients compared to the control participants ($p = 0.007$). Two stool samples from CRC patients generated two *Enterobacteriaceae* each, whilst one isolate was obtained from each of the 36 control

Demography	CRC Frequency (%)	Control Frequency (%)	*P*-value	OR (95% CI)
Gender				
Female	19 (32.8)	22 (44)	0.231	0.620 (0.284–1.356)
Male	39 (67.2)	28 (56)		
Alcohol consumption				
Yes	42 (72.4)	25 (50)	0.018	2.625 (1.180–5.838)
No	16 (27.6)	25 (50)		
Smoking				
Yes	6 (10.5)	6 (12)	0.810	0.863 (0.260–2.868)
No	51 (89.5)	44 (88)		
Family history				
Yes	0 (0)	0 (0)	-	-
No	58 (100)	50 (100)		
Age				
Less than 30 years	3 (5.2)	10 (20)		1 (Ref)
30–49 years	24 (41.4)	17 (34)	0.034	4.706 (1.124–19.705)
50 years and above	31 (53.4)	23 (46)	0.035	4.493 (1.110-18.191)

Table 1. Demographic information of the study participants. Total number of CRC patients (cases) = 58. Total number of control participants = 50.

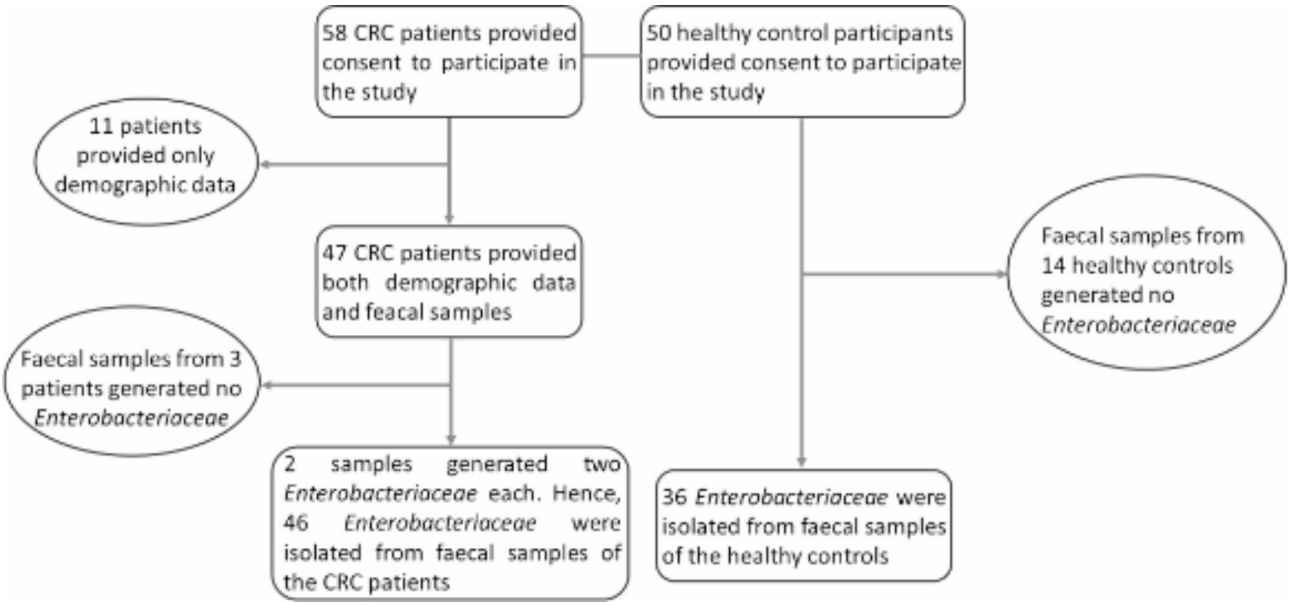

Fig. 1. Schematic summary for recruitment of study participants and collection of socio-demographic data and stool samples for isolation of *Enterobacteriaceae*.

participants (Fig. 1). Hence, a total of 46 *Enterobacteriaceae* were isolated from the CRC patients and 36 isolates were obtained from the control participants.

Escherichia coli was the predominant *Enterobacteriaceae* isolated from stool samples of the CRC patients (76.1%) and the control participants (80.6%; Fig. 2A). The other *Enterobacteriaceae* isolated from the study participants were *Klebsiella pneumoniae*, *Proteus mirabilis*, *Escherichia fergusonii*, *Enterobacter cloacae*, *Enterobacter hormaechei*, *Providencia stuartii*, *Alcaligenes faecalis*, *Raoultella ornithinolytica* and *Morganella morganii* (Fig. 2A). The *Enterobacteriaceae* from the CRC patients were more diverse compared to the isolates from the healthy control participants, highlighting a potential effect of CRC on dysbiosis.

Genomic DNA sequencing and quality control statistics of assembled genomes

The genomic DNA of each *Enterobacteriaceae* was used for DNA library preparation, multiplexing and paired-end sequencing on the Illumina NextSeq 2000 system. Sequence reads were assembled using SPAdes, and the quality of the assembled genomes was assessed using Quast. Draft assembled genomes were annotated using Prokka (1.14.6) and core genomes were analyzed using Roary. The median contig length of the draft assembled genomes was 61 bp (range is 11–162 bp), and the N50 median was 221,614 bp (72,883–2,457,026 bp). The median GC content was 50.73 (38.7–57.35), with a total median genome length of 4,776,786 bp (3,872,290–5,673,707 bp), which is within the range for *E. coli* genome size, and there was no evidence of contamination.

Cumulative virulence genes encoded by *Escherichia coli* phylogroups

Nineteen (54.3%) of the 35 *Escherichia coli* isolates from the CRC patients were categorized as phylogroup A, and nine (25.7%) other isolates as phylogroup B1, demonstrating the majority of the *Escherichia coli* isolates from the CRC patients were most likely intestinal commensal strains (Fig. 2B). One (2.9%) *Escherichia coli* isolate from the CRC patients was categorized as phylogroup B2 and four (11.4%) other isolates were categorized as phylogroup D. The *Escherichia coli* isolates from the control participants were also predominantly in phylogroups A and B1. The B2 phylogroup of *Escherichia coli*, notable for their hypervirulence traits and possession of the *pks*-genomic island, was not significantly associated with CRC [OR 0.18 (0.02–1.75), $p = 0.140$].

The *Escherichia coli* isolates from the control participants encoded relatively higher number of virulence genes compared to the isolates from the CRC patients, even though no significant association was observed between number of virulence genes per isolate and CRC (p-value = 0.101; Fig. 3A). Moreover, the non-commensal isolates (phylogroups B2, D and F) from the CRC patients encoded significantly higher number of virulence genes compared to the CRC commensal (phylogroups A and B1) isolates, suggesting a hypervirulent potential of the non-commensal isolates ($p < 0.0001$, Fig. 3B,D). The non-commensal isolates from the control participants also encoded significantly higher number of virulence genes compared to the commensal isolates of the control ($p = 0.037$, Fig. 3C,D). Since the isolate of phylogroup B2 from the CRC patient encoded relatively higher number of virulence genes compared to the corresponding isolates from the control participants, we infer that the phylogroup B2 isolate from CRC patient might represent a potential hypervirulent strain compared to the four isolates of phylogroup B2 from the control participants.

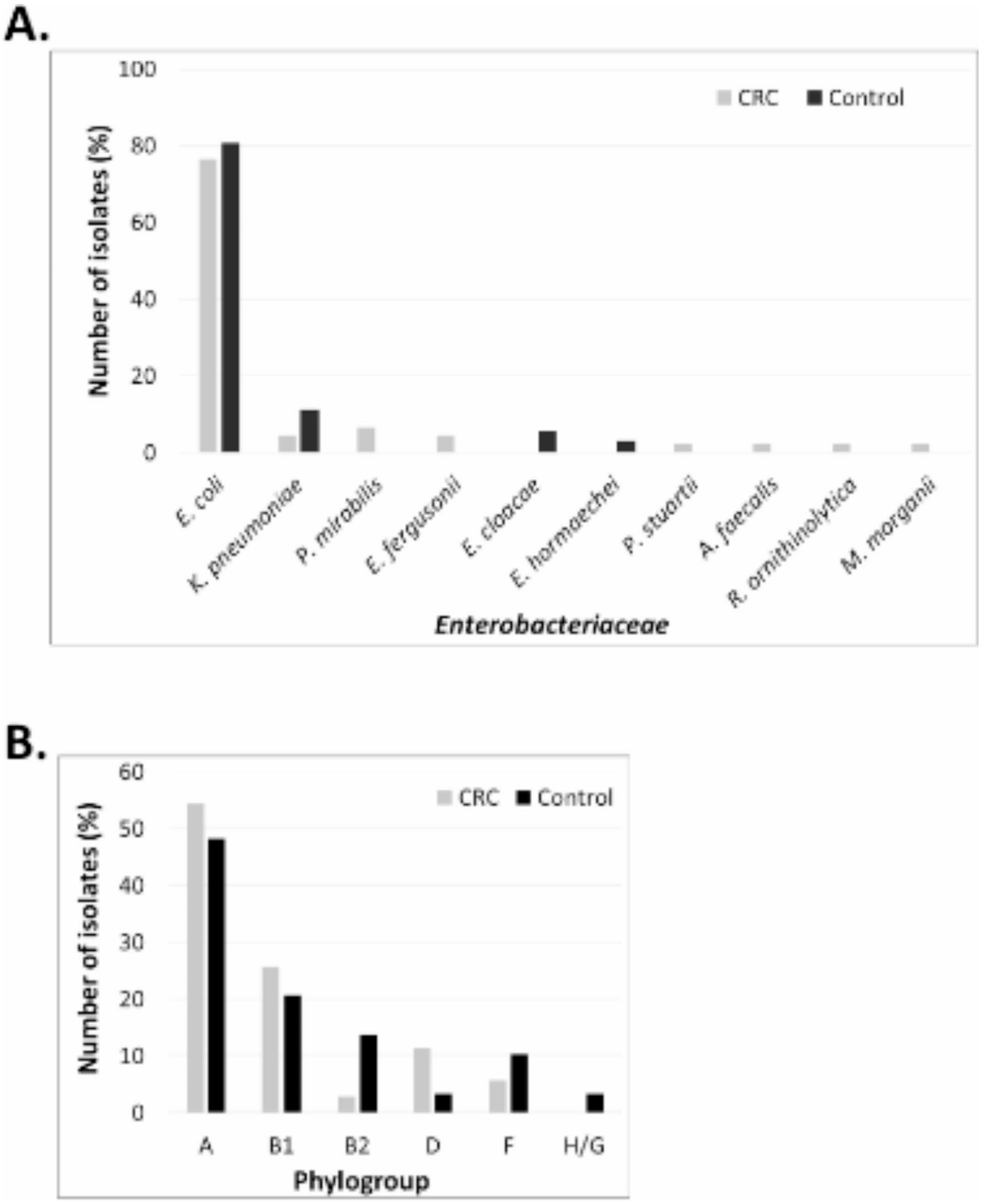

Fig. 2. *Enterobacteriaceae* isolated from the study participants. (**A**) Identity of *Enterobacteriaceae* obtained from the study participants. (**B**) Phylogroups of the *Escherichia coli* isolates determined via the EzClermont web app[36].

Analysis of toxin-encoding genes in the genome of the *Escherichia coli* isolates

The *pks*-genomic island (*clbA-clbS*), which has been reported to be overrepresented in *Enterobacteriaceae* from CRC patients, was not detected in any of the sixty-four *Escherichia coli* isolates from the study participants (Fig. 4A). Genes encoding cytolethal distending toxin (*cdtA-cdtC*), cytotoxic necrotizing factor (*cnf1-cnf3*) and cycle inhibiting factors (*cif*) were also not detected in the genome of the *Enterobacteriaceae* from the CRC patients

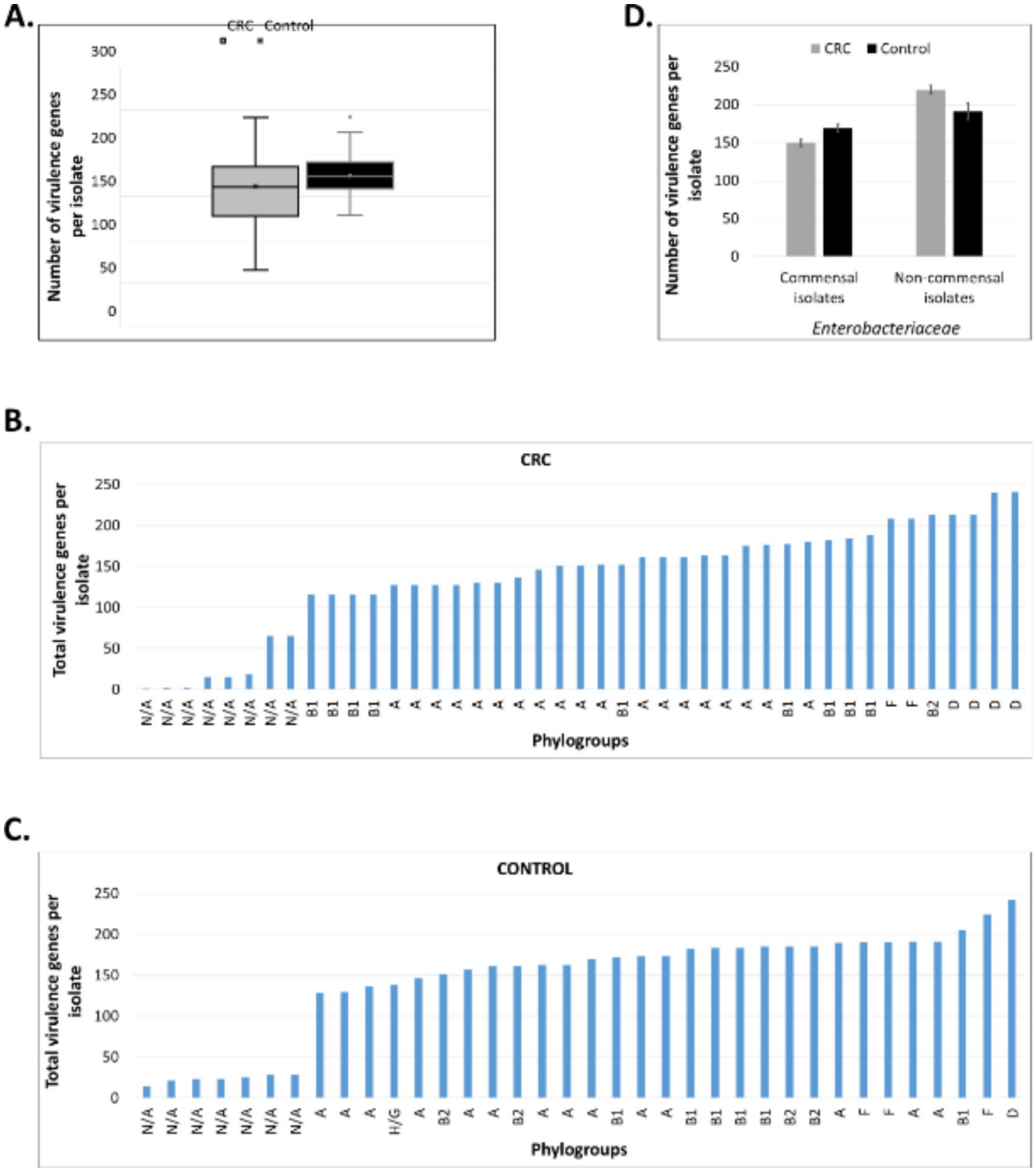

Fig. 3. Cumulative virulence genes encoded by the genome of the *Enterobacteriaceae* isolates. (**A**) Comparison of the total number of virulence genes encoded by the genome of the *Escherichia coli* isolates from the CRC patients and the control participants. (**B**) Cumulative virulence genes encoded by the phylogroups of *Escherichia coli* from the CRC patients. (**C**) Cumulative virulence genes encoded by the phylogroups of *Escherichia coli* from the control participants. (**D**) Number of virulence genes encoded by the commensal and non-commensal *Escherichia coli* isolates. Error bars represent the standard error of the mean. N/A denotes the non-*Escherichia coli* isolates, which cannot be classified into phylogroups. The seven non-*Escherichia coli* isolates from the control participants are *K. pneumoniae*, *E. cloacae*, *K. pneumoniae*, *K. pneumoniae*, *E. hormaechei*, *K. pneumoniae* and *E. cloacae*. The eight non-*Escherichia coli* isolates from the CRC patients are *M. morganii*, *P. mirabilis*, *P. mirabilis*, *K. pneumoniae*, *K. pneumoniae*, *R. ornithinolytica*, *E. fergusonii*, and *E. fergusonii*.

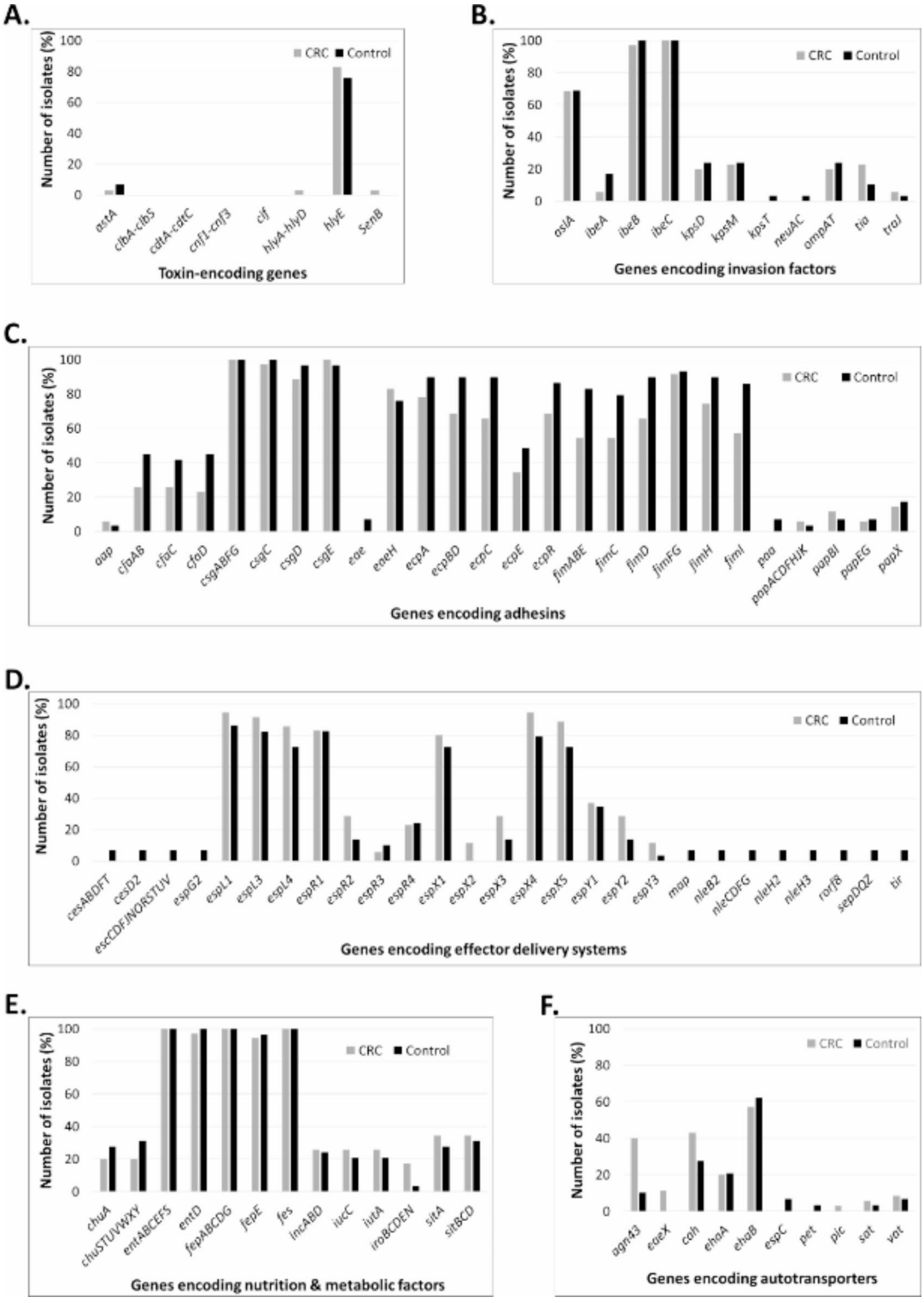

Fig. 4. Categories of virulence genes encoded by the genome of the *Enterobacteriaceae* isolates. Genes encoding toxins, invasion factors, adhesins, effector delivery systems, nutrition and metabolic factors and autotransporters are shown in (**A**)–(**F**).

and the control participants. Enterotoxin-encoding genes (*astA* and *senB*), and the *hlyA-hlyD* genes responsible for the synthesis and secretion of haemolysin, were detected in the genome of only one *Escherichia coli* isolate from the CRC patients. In contrast, the *hlyE* gene encoding the pore-forming haemolysin E (α-haemolysin) toxin was detected in the genome of most of the *Escherichia coli* isolates from the CRC patients, even though there was no statistically significant association between the *hlyE* gene and CRC (OR 1.538 (0.453–5.226); p-value = 0.49).

Hence, *Escherichia coli* genes capable of causing genotoxic effects or cell cycle perturbations in host eukaryotic cells might have no role in colorectal tumorigenesis of the patients recruited for this study. Alternatively, the genotoxic *Escherichia coli* could previously have been present in the gut microbiota of the CRC patients, but were cleared and undetected in the present study.

Analysis of genes encoding invasion and adhesion factors

The *aslA*, *ibeB* and *ibeC* genes, encoding notable invasion factors, were detected in at least 69% of the *Escherichia coli* isolates from the CRC patients and the control participants (Fig. 4B). In contrast, the *ibeA*, *kpsDM*, *ompAT* and *tia* genes were less prevalent in the isolates from the study participants. Other invasion factor genes such *kpsT*, *neuAC* and *traJ* were rare in the genomes of the *Escherichia coli* isolates. None of these invasion factor genes encoded by the *Escherichia coli* genome was significantly associated with the CRC patients (Supplementary Table S1).

Genes encoding the curli fimbriae (*csg*) were detected in almost all the *Escherichia coli* isolates, indicating this class of adhesins might be relevant for adhesion and colonization phenotypes of these isolates in host eukaryotic cells during infection (Fig. 4C). *Escherichia* common pili (*ecp*) and type I fimbriae (*fim*) genes were detected in several of the isolates, even though they were modestly over-represented in the isolates from the control participants compared to the CRC patients. P fimbriae (*pap*) genes were rare in the genome of the isolates from the study participants despite the similar morphological identity of P fimbriae and type I fimbriae. The *eae* gene, encoding intimin, was not detected in the genome of the *Escherichia coli* isolates from the CRC patients but was identified in a few (2 out of 29) isolates from the control participants. Since most of the *Escherichia coli* isolates from the CRC patients and control participants encode several adhesins, it can be inferred that these isolates can use their attachment to host cells as a vital survival mechanism during infection. Nevertheless, none of these adhesins were significantly associated with the *Escherichia coli* isolates from the CRC patients in comparison to the control participants (Supplementary Table S1).

Analysis of genes encoding effector proteins, nutrition/metabolic factors and autotransporters

Generally, bacterial strains (especially enteropathogenic isolates) are capable of directly delivering effector proteins into host cells via the type III secretion system, leading to perturbation of cellular processes in the host, and ultimately disease conditions. Genes encoding notable bacterial effector proteins, such as *tir*, *map*, *espG2*, *nleB2*, *nleC*, *nleD*, *nleG*, *nleF* and *nleH2* were absent in the genome of the *Escherichia coli* isolates from the CRC patients, but were detected in 2 out of the 29 isolates from the control participants (Fig. 4D). Less notable effector proteins including *espL1*, *espL3*, *espL4*, *espR1*, *espX1*, *espX4* and *espX5* were detected in several *Escherichia coli* isolates from the study participants, with modest over-representation in the isolates from the CRC patients in comparison to the control participants. None of the genes encoding effector proteins was significantly associated with CRC (Supplementary Table S1).

Bacterial genes that facilitate bioavailability or acquisition of iron in host cells are crucial for bacterial survival, especially when the bacteria reside in regions of the host where iron exist in an insoluble state or is complexed with iron-binding proteins of the host. The *ent* operon, which encodes enterobactin and is well known for its role in iron acquisition in a host, was detected in the genome of almost all the *Escherichia coli* isolates from the study participants (Fig. 4E). Ferric-enterobactin uptake (*fepABCDEG*) and utilization (*fes*) genes were also detected in almost all the *Escherichia coli* isolates, indicating these genes are vital for facilitating iron bioavailability and acquisition in the *Escherichia coli* isolates from the study participants. Nevertheless, these genes were not significantly associated with the *Escherichia coli* isolates from the CRC patients in comparison to the control participants (Supplementary Table S1).

Autotransporters (type V secretion system) are functionally relevant for bacterial survival and pathogenesis. The *espC*, *pet*, *pic*, *sat* and *vat* autotransporter genes, belonging to the chymotrypsin-like serine protease family, were rare in the genome of the study participants (Fig. 4F). *ehaB* was the most widespread autotransporter gene in the genome of the *Escherichia coli* isolates from both the CRC patients and the control participants. *agn43* and *cah*, both encoding self-associating autotransporters associated with biofilm formation, were modestly over-represented in the genome of the isolates from the CRC patients. Moreover, the *eaeX* gene encoding an intimin-like protein, was only detected in the isolates from the CRC patients (4 out of 35). None of the genes encoding autotransporters was significantly associated with CRC (Supplementary Table S1).

Analysis of *Escherichia coli* virulence genes encoded by the non-*Escherichia coli* isolates

Virulome analysis was conducted for 15 out of the 18 non-*Escherichia coli* isolates from the study participants using the *Escherichia coli* virulence genes database. This analysis was vital for determining potential transmission of *Escherichia coli* virulence genes to the other strains of *Enterobacteriaceae* isolated from the study participants. Virulence genes that were detected in the genome of at least one of the non-*Escherichia coli* isolates are indicated in Fig. 5. The *pks*-genomic island was not detected in the genome of the 15 non-*Escherichia coli* strains used for the analysis. Most virulence genes that were commonly detected in the genome of the *Escherichia coli* isolates from this study (*hlyE*, *aslA*, *ibeBC*, *csg*, *fimDFGH*, *espL1/3/4*, *espR1*, *espX1/4/5*, *ent*, *fep*, *fes*, *ehaB*) were also encoded by the genome of at least one of the two *Escherichia fergusonii* strains. The only exceptions were *eaeH*, *ecp* and *fimABCE* adhesin genes, which were absent in the genome of the *Escherichia fergusonii* strains. The *eaeH* and *fimABCE* genes were also not detected in the genome of any of the non-*Escherichia coli* strains from the study participants while *ecp* was detected in the genome of the *Klebsiella pneumoniae* strains.

The *ompA*, *entAB* and *fepC* genes were identified in all the *Klebsiella pneumoniae* strains, in addition to the *ecp* genes. The *cah* gene, which was over-represented in *Escherichia coli* isolates from the CRC patients, was identified in *Proteus mirabilis* isolates from the CRC patients and one isolate of *Klebsiella pneumoniae* from the

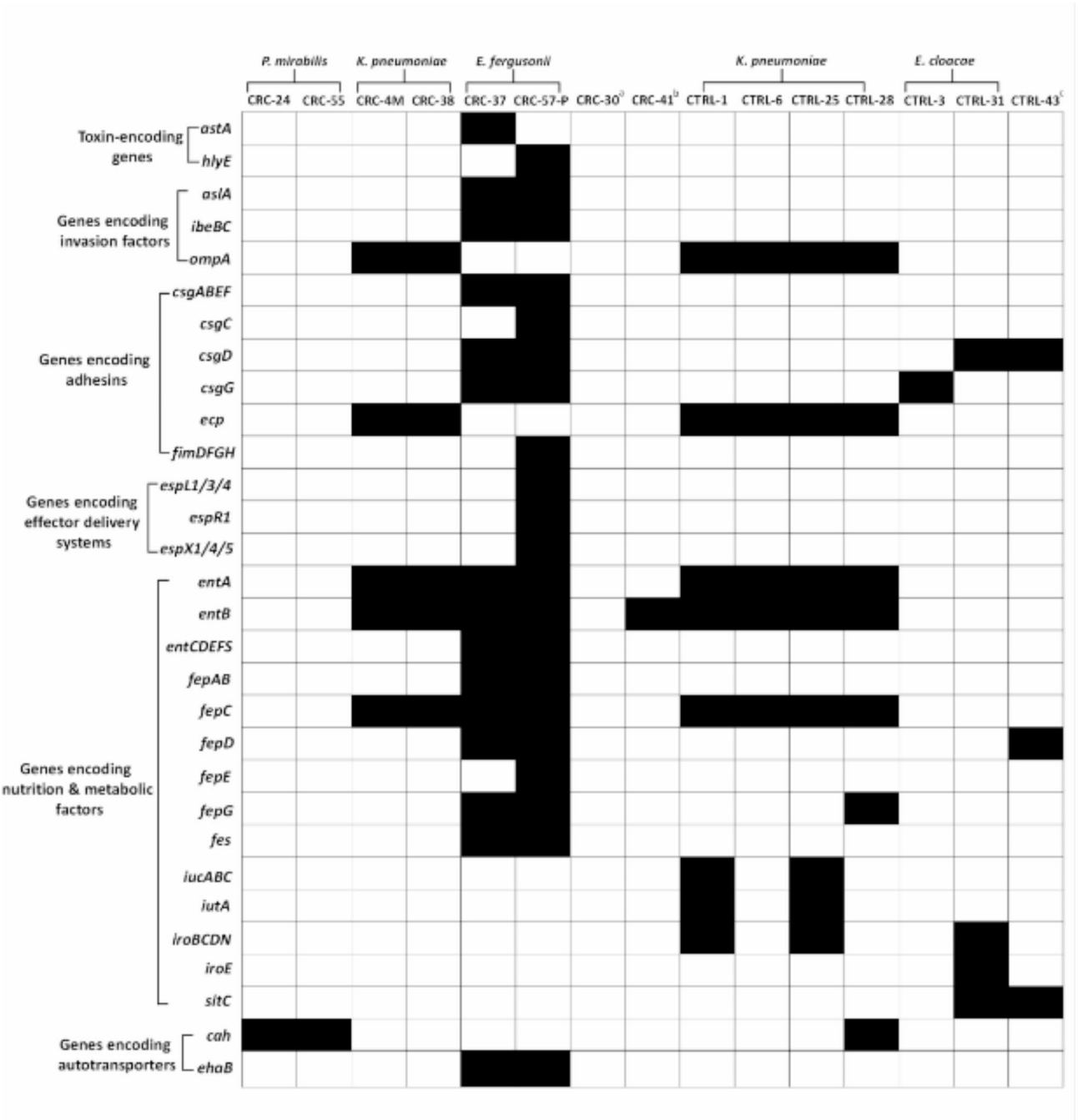

Fig. 5. Detection of *Escherichia coli* virulence genes in the genome of the other *Enterobacteriaceae* isolated from the study participants. Black shaded boxes represent presence of the indicated virulence gene while unshaded boxes represent absence of the gene. a, b, and c denote *Morganella morganii*, *R. ornithinolytica* and *E. hormaechei*, respectively.

control participants. The *agn43* gene, which was widespread in *Escherichia coli* strains from the CRC patients, was not detected in the genome of the non-*Escherichia coli* strains. None of the virulence genes analyzed in this study were identified in the *Morganella morganii* strain while *Escherichia fergusonii* encoded most of the virulence genes, depicting a close relatedness of the latter to *Escherichia coli* (Figs. 3B,C and 5).

Genes unique to putative hypervirulent *Escherichia coli* from CRC patients

The seven *Escherichia coli* isolates from CRC patients, encoding the highest number of virulence genes (> 200 virulence genes) were referred to as putative hypervirulent strains (Fig. 3B). The sequence types of these seven isolates, in descending order of the total virulence genes encoded, were ST1380 (two isolates of phylogroup D), ST69 (two isolates of phylogroup D), ST636 (one isolate of phylogroup B2) and ST3901 (two isolates of phylogroup F). Genomic analysis of these seven isolates showed that the transcriptional activator gene *aggR*, the

Genes	Sequence type	Function
aggR	ST1380 (Phylogroup D)	Transcription regulator of AAF/I
aaiA-X		Aai chromosomal type VI secretion system
hdaA-D		Aggregative adherence genes
eaeX		Invasin/intimin homologue
gspC-M		Type II secretion system
hlyA-D	ST636 (Phylogroup B2)	Alpha-haemolysin
mcbA		Biofilm formation
tcpC		Inhibits Toll-like receptor signaling
eaeX	ST69 (Phylogroup D)	intimin homologue involved in aggregation and adherence
gspC-M		Type II secretion system
gspC-M	ST3901 (Phylogroup F)	Type II secretion system

Table 2. Genes unique to putative hypervirulent *Escherichia coli* from CRC patients.

Phylogroups	CRC	Control
A	2 (19)	1 (14)
B1	7 (9)	1 (6)
B2	1 (1)	0 (4)
D	4 (4)	1 (1)
F	0 (2)	0 (3)
H/G	0 (0)	0 (1)
Total	14 (35)	3 (29)

Table 3. Distribution of *agn43* gene in the *Escherichia coli* phylogroups.

Aai chromosomal type VI secretion system (*aaiA-X*) and the *hdaA-D* adhesion genes were unique to ST1380, even though *aaiT* and *aaiW* were detected in the genome of a few other isolates (Table 2). The *hlyA-D*, *mcbA* and *tcpC* genes were identified only in the genome of isolates belonging to ST636. The *eaeX* gene was detected in both ST1380 and ST69 only, while the *gspC-M* genes were identified in ST1380, ST69 and ST3901.

Analysis of *agn43* gene and *Escherichia coli* phylogroups

The *agn43* gene has previously been reported to be more prevalent in *Escherichia coli* phylogroup E isolates[37]. Since the *agn43* gene was modestly over-represented in *Escherichia coli* isolates from the CRC patients, despite a $q > 0.05$, we investigated the distribution of *agn43* in the *Escherichia coli* phylogroups. Three out of the thirty-three phylogroup A isolates encoded the *agn43* gene, indicating the gene was significantly under-represented in Phylogroup A compared to the other phylogroups ($p = 0.002$; Table 3). Seven out of the nine phylogroup B1 isolates from the CRC patients encoded *agn43* whilst the gene was detected in only one of the six isolates of phylogroup B1 from the control participants. All the five phylogroup D isolates encoded the *agn43* gene, highlighting the prevalence and significant association of the gene with the phylogroup D isolates compared to the other *Escherichia coli* phylogroups ($p < 0.001$).

Analysis of phylogeny and antimicrobial resistance genes

Phylogenetic analysis of the *Enterobacteriaceae* from the study participants was conducted to identify clustering patterns of the isolates from the CRC patients and the control participants. The analysis demonstrated *Escherichia coli* strains were clustered according to their phylogroups and multilocus sequence sub-types, but not the source of isolation (CRC patients or control participants; Fig. 6). Furthermore, the phylogenetic analysis confirmed the close relatedness of *Escherichia fergusonii* to the *Escherichia coli* strains and a distant relationship between the *Escherichia coli* isolates and *Morganella morganii*, as was initially deduced from the virulome analysis.

Further analysis of the genomic data revealed the *int* gene, which is capable of facilitating acquisition of exogenous DNA into bacterial chromosomes, was modestly over-represented in the genome of *Enterobacteriaceae* from the CRC patients ($q = 0.1128$; Fig. 6 and Supplementary Table S1). This observation prompted the investigation of antimicrobial resistance genes encoded by the genome of the bacterial isolates from the study participants. β-lactamase resistance genes such as *E. coli ampC2* and *ampH* genes were encoded by the genome of most of the *Enterobacteriaceae* from the study participants. Penicillin binding proteins-mediated mechanisms for resistance to β-lactams was also predominant in the *Enterobacteriaceae* from both the CRC patients and control participants. In contrast, the *CTX-M-55* and *OXA-10* β-lactamase resistance genes were encoded by two *Escherichia coli* isolates from the CRC patients, indicating they were less prevalent in the genome of the *Enterobacteriaceae* from the study participants. Resistance genes affecting aminoglycosides (*strA* and *strB*), macrolide (*mphA*), sulphonamides (*sul1* and *sul2*), tetracycline (*tetA* and *tetR*) and trimethoprim (*dfrA14*) were also encoded by several *Enterobacteriaceae* from the study participants. Nonetheless, only the *sul2* gene was

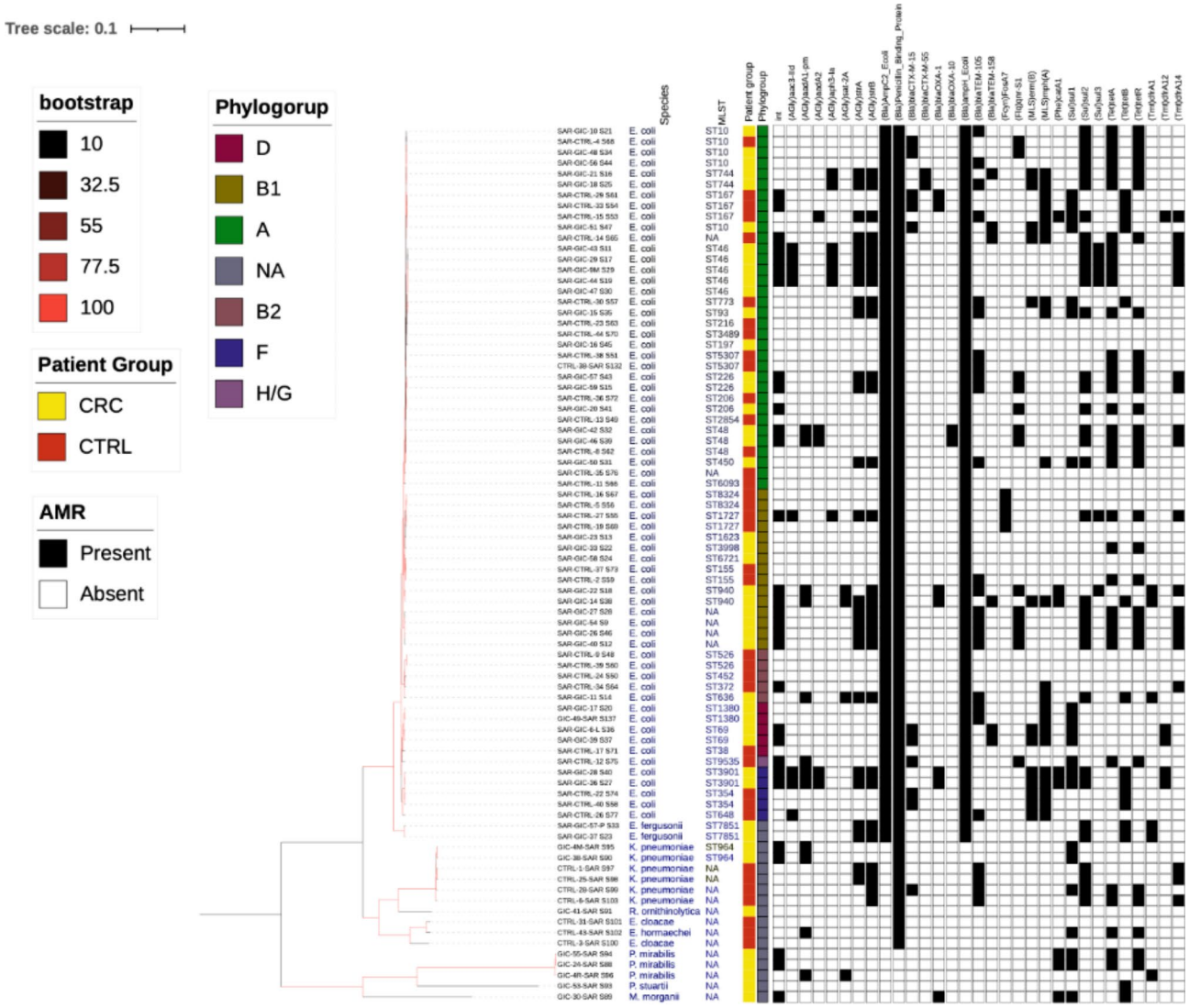

Fig. 6. Analysis of the phylogeny and antimicrobial resistance genes in the genome of the *Enterobacteriaceae* isolates. The maximum likelihood phylogeny was constructed using RAxML and the bootstrap are displayed ranging from black (less confident) to red (high confident) score. The antibiotic class are shown in parenthesis; AGly is aminoglycosides, Bla is β-lactam antibiotics, Fcyn is fosfomycin, Flq is fluoroquinolone, MLS is macrolide, lincosamide, and streptogramin, Phe is phenicol, Sul is sulphonamide, Tet is tetracycline and Tmt is trimethoprim. int denotes the *int* gene. NA denotes isolates in which the phylogroup or sequence type is not available.

significantly associated with *Enterobacteriaceae* from the CRC patients in comparison to the control participants [OR 4.79 (1.79–12.81), q = 0.045; Supplementary Table S2].

Discussion

In the present study, whole genome sequencing was used for analysis of *Enterobacteriaceae* isolated from CRC patients and healthy control participants at a tertiary hospital in Ghana. Socio-demographic data from study participants showed alcohol consumption and age (≥30 years), but not gender and smoking status, were significantly associated with the CRC patients. We had previously reported that smoking habit was not significantly associated with lower gastrointestinal cancers, which includes colorectal cancers, in the Ghanaian population[3]. However, the earlier study found no significant association between alcohol consumption and lower gastrointestinal cancers, as was detected in the present study, and this might be attributed to the small sample size used for both studies. Another intriguing observation from this present study was the lower age limit (30 years) of the Ghanaian population that was significantly associated with CRC. In the United States, the incidence rate of CRC increased by 80–100% with each 5-year group until the age of 50 years, whilst the incidence rate increased by 20–30% among the older adults (≥55 years) in the year 2015–2019[38]. These observations highlight the need to actively conduct routine genetic test for CRC in both young and older adults, especially in the Ghanaian population.

Enterobacteriaceae, such as *Escherichia coli*, *Klebsiella pneumonia*, *Enterobacter spp* and *Morganella morganii* have been associated with CRC in previous reports[39–42]. The present study identified higher microbial diversity in the samples from the CRC patients compared to the control group. In addition, detection of *Enterobacteriaceae* from the samples of the CRC patients was significantly higher than the control group. These observations might indicate that there is perturbation in gut microbial content in the CRC patients and hence, necessitated thorough evaluation of the virulence tendencies of the isolates to confirm whether the dysbiosis favoured growth of harmful microbiota while causing loss of beneficial *Enterobacteriaceae*.

Escherichia coli was the predominant *Enterobacteriaceae* isolated from stool samples of the participants, in the present study, and they were mostly commensal strains[43,44]. These commensal strains usually lack virulence determinants implicated in disease conditions caused by intestinal and extraintestinal pathogenic *Escherichia coli*[45]. Our data indicate that the B2 phylogroup of *Escherichia coli* was under-represented in the study participants, and this could be due to the isolation of *Enterobacteriaceae* from fecal samples instead of the site of malignant tumors at the colon[46]. *Escherichia coli* from the CRC patients, belonging to phylogroups B2, D and F, also encoded more virulence factors compared to isolates of phylogroups A and B1 (commensal strains). Moreover, the isolates of phylogroups B2, D and F from the CRC patients had different sequence typing in comparison to the isolates of the same phylogroups from the control participants. These observations highlight the uniqueness of diverse isolates belonging to the same phylogroups but obtained from different sources, which might be indicative of differences in virulence tendencies during infection. Confounding factors caused by variations in microbial content between bowel movement were not taken into account in the present study, and might represent a potential limitation of the study design.

None of the bacterial cyclomodulin-encoding genes was detected in the genome of the *Enterobacteriaceae* isolated from the CRC patients and control participants. Notably, the *pks*-genomic island previously reported to be associated with CRC tumorigenesis was not detected in the genome of the *Enterobacteriaceae* from the participants in the present study[47]. Nevertheless, there is a possibility that *Enterobacteriaceae* encoding the cyclomodulin genes might have previously inhabited the gut of the CRC patients and increased the risk of CRC tumorigenesis in these patients. Extensive analysis of virulence genes available in the virulence factor database revealed the virulome profile of the *Enterobacteriaceae* from the CRC patients and control participants. Genes encoding curli fimbriae, type I fimbriae, *Escherichia* common pili, enterobactin and proteins utilized in ferric iron transport were encoded by most of the *Escherichia coli* isolates, as previously reported by a genomic study of *Escherichia coli* obtained from human, animal and food sources[48]. These *Escherichia coli* virulence genes were also identified in the genomes of *Escherichia fergusonii*, and to a lesser extent *Klebsiella pneumoniae*, highlighting the potential role of horizontal gene transfer in boosting bacterial virulence in the gut of the host. This present study also showed that the *agn43* autotransporter gene, which is associated with biofilm formation, autoaggregation and attachment of bacteria to host cells was modestly over-represented in *Escherichia coli* isolates from the CRC patients[49–51]. Biofilm formation and attachment to host cells are mechanisms of pathogenic bacteria that have been reported to be associated with CRC[52,53]. Hence, it is possible that the *agn43*-encoding *Enterobacteriaceae* isolated in this study could have facilitated CRC tumorigenesis in the patients via these bacterial pathogenesis-related mechanisms.

Analysis of genes that are unique to the *Escherichia coli* isolates from the CRC patients encoding the highest total number of virulence genes revealed additional virulence factors and mechanisms that might be associated with CRC tumorigenesis in the patients enrolled for this study. The type II secretion system and virulence factors involved in aggregation and adhesion were identified in the isolates of phylogroups D and F. In contrast, alpha haemolysin toxin, biofilm formation and inhibition of innate immune system in host cells were the cellular processes that are affected by the virulence genes identified in the B2 phylogroup isolate. These observations highlight potential cellular process in host cells that can be targeted by pathogenic *Escherichia coli* in order to enhance the risk of tumorigenesis in host eukaryotic cells during exposure to these pathogens.

The *int* gene allows mobility of genetic elements between organisms via horizontal gene transfer, preferably transduction, and can facilitate the spread of antimicrobial resistance genes[54,55]. Virulome analysis in the present study showed the *int* gene was modestly over-represented in *Escherichia coli* isolates from the CRC patients. Even though the *int* gene is only one of the several key mediators of horizontal gene transfer in bacteria, this observation prompted the analysis of antimicrobial resistance genes encoded by the *Enterobacteriaceae* from the study participants. Antimicrobial resistance genes for β-lactam antibiotics were widespread in the genome of the *Enterobacteriaceae* from the study participants, as previously reported by a research conducted at the same study site[56]. Other antibiotic resistance genes we detected in the genome of the *Enterobacteriaceae* have also been reported in bacterial isolates from human inhabitants, livestock and water sources in Ghana[57–59]. Nonetheless, antimicrobial susceptibility assays are needed to confirm the expression of the antimicrobial resistance genes identified in the present study, and the extent of antibiotic resistance conferred on the *Enterobacteriaceae*.

Methods

Study design and participants

This genomic study utilized *Enterobacteriaceae* isolated from stool samples of CRC patients and healthy control participants at the Korle Bu Teaching Hospital (KBTH). The study participants were randomly recruited at the study site from February 2022 – January 2023. All the participants consented for using their socio-demographic data and stool samples in this study (Fig. 1). The CRC patients were either on admission at the Colorectal unit or outpatients at the Oncology and Surgical Clinics of the KBTH. The healthy participants were individuals who were not diagnosed with CRC or other forms of ailments at the time of sample collection. Individuals on antibiotic or cancer chemotherapy were excluded from the study.

Collection of socio-demographic data and stool samples

Paper-based questionnaires were administered to the study participants to obtain their socio-demographic data such as age, gender, family history of CRC, alcohol consumption and smoking status. Stool samples were also collected into 10 mL Cary-Blair transport media (Thermo Fisher Scientific) according to the manufacturer's instructions. Socio-demographic data and stool samples were collected only once from participants that consented to participate in the study. All socio-demographic data from study participants were handled anonymously and confidentially. Anonymity was ensured by the use of codes generated from the respondents' initials. These data were stored and analyzed electronically using Microsoft Office (Excel) version 2021.

Isolation and identification of *Enterobacteriaceae* from study participants

All the stool samples provided by the study participants were used for microbial analysis. Stool samples were inoculated onto MacConkey agar (BBL™, BD) for isolation of *Enterobacteriaceae*. Isolates of *Enterobacteriaceae* were initially identified using a standard biochemical test protocol. The identities of the isolates were confirmed via MALDI-TOF mass spectrometry, as previously reported[59].

Whole genome sequencing and analysis

Extraction of genomic DNA from the *Enterobacteriaceae* isolates was conducted using QIAamp® DNA mini kit (QIAGEN Inc. GmbH, Holden, Germany), according to the manufacturer's protocol. Genomic DNA was used for library preparation, followed by multiplexing and paired-end sequencing on the Illumina NextSeq 2000 system (Illumina, San Diego, CA, USA). Quality control was performed using FastQC (v0.12.1) and quality trimming was done using trimmomatic (v0.39) to remove low quality calls and sequencing adapters[60]. *De novo* assembly of the reads were performed using SPAdes (v3.15.5) with the following Kmers (21, 33, 55 and 77) and quality of the draft assemblies were checked using Quast (5.2.0)[61,62]. Draft genome assemblies were annotated using Prokka (1.14.6) and core genome were analyzed using Roary[63,64]. Multi-locus sequence type was determined using get_sequence_type from mlst_check of the Sanger Institute Pathogen group (https://github.com/sanger-pathogens/mlst_check). ABRicate (v1.0.1; https://github.com/tseemann/abricate) was used to identify AMR genes and *E. coli* virulence genes in the Virulence Factor Database (VFDB)[65]. A minimum coverage threshold of 80% and a minimum identity threshold of 80% was used for identification of AMR and virulence genes. Phylogenetic tree was generated from the core aligned single nucleotide polymorphisms (SNPs) using RAxML and visualised and annotated with iTOL (v8.2.12)[66,67].

The fastQ files of all the sequence data were submitted to the National Center for Biotechnology Information (NCBI) and assigned accession numbers under Bioproject ID PRJNA1126843 (ID 1126843-BioProject-NCBI (nih.gov)).

Determination of the phylogroups of the *Escherichia coli* isolates

Fasta files from the whole genome sequence analysis were used for the determination of the *Escherichia coli* phylogroups via the EzClermont web app[36].

Statistical analysis

Odds ratio was used to determine the association of parameters from socio-demographic data (age, gender, family history of CRC, alcohol consumption and smoking status), and CRC at 95% confidence interval. Association between virulence/antimicrobial resistance genes and CRC was also determined by odds ratio at 95% confidence interval, and the p-values were corrected using the Benjamini–Hochberg method. Odds ratio was also used for determining the associations between *Escherichia coli* phylogroups and *agn43* gene, B2 phylogroup and CRC, and *Escherichia coli* commensal status and number of virulence genes per isolate. The Mann-Whitney U Test was used for evaluating statistical significance between cumulative number of virulence genes per isolate and CRC. Fisher exact test was used to determine the association between detection of *Enterobacteriaceae* on MacConkey agar and CRC.

Data availability

Sequence data that support the findings of this study have been deposited in the National Center for Biotechnology Information (NCBI) Genbank with the primary accession code PRJNA1126843(ID 1126843-BioProject-NCBI (nih.gov)).

Received: 12 July 2024; Accepted: 25 September 2024
Published online: 05 October 2024

References

1. Wong, M. C. S. et al. Differences in incidence and mortality trends of colorectal cancer worldwide based on sex, age, and anatomic location. *Clin. Gastroenterol. Hepatol.* **19** (5), 955–966e61 (2021).
2. Sawicki, T. et al. A review of colorectal cancer in terms of epidemiology, risk factors, development, symptoms and diagnosis. *Cancers (Basel).* **13** (9), 2025 (2021).
3. Boi-Dsane, N. A. A. et al. Cross-sectional study for investigation of the association between modifiable risk factors and gastrointestinal cancers at a tertiary hospital in Ghana. *Cancer Control.* **30**, 10732748231155702 (2023).
4. Arnold, M. et al. Global patterns and trends in colorectal cancer incidence and mortality. *Gut.* **66** (4), 683–691 (2017).
5. Hossain, M. S. et al. Colorectal cancer: a review of carcinogenesis, global epidemiology, current challenges, risk factors, preventive and treatment strategies. *Cancers (Basel).* **14** (7), 1732 (2022).
6. Barnetson, R. A. et al. Identification and survival of carriers of mutations in DNA mismatch-repair genes in colon cancer. *N Engl. J. Med.* **354** (26), 2751–2763 (2006).

7. Hampel, H. et al. Feasibility of screening for Lynch syndrome among patients with colorectal cancer. *J. Clin. Oncol.* **26** (35), 5783–5788 (2008).
8. Hampel, H. et al. Screening for the Lynch syndrome (hereditary nonpolyposis colorectal cancer). *N. Engl. J. Med.* **352** (18), 1851–1860 (2005).
9. Mork, M. E. et al. High prevalence of hereditary cancer syndromes in adolescents and young adults with colorectal cancer. *J. Clin. Oncol.* **33** (31), 3544–3549 (2015).
10. Yurgelun, M. B. et al. Identification of a variety of mutations in cancer predisposition genes in patients with suspected Lynch syndrome. *Gastroenterology* **149**(3), (2015). 604 – 13.e20.
11. Kim, J. & Lee, H. K. Potential role of the gut microbiome in colorectal cancer progression. *Front. Immunol.* **12**, 807648 (2022).
12. Arthur, J. C. et al. Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science*. **338** (6103), 120–123 (2012).
13. Cipe, G., Idiz, U. O., Firat, D. & Bektasoglu, H. Relationship between intestinal microbiota and colorectal cancer. *World J. Gastrointest. Oncol.* **7** (10), 233–240 (2015).
14. Kostic, A. D. et al. Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell. Host Microbe*. **14** (2), 207–215 (2013).
15. Rubinstein, M. R. et al. Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. *Cell. Host Microbe*. **14** (2), 195–206 (2013).
16. Zackular, J. P. et al. The gut microbiome modulates colon tumorigenesis. *mBio*. **4** (6), e00692–e00613 (2013).
17. Nougayrède, J. P. et al. *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. *Science*. **313** (5788), 848–851 (2006).
18. Cougnoux, A. et al. Bacterial genotoxin colibactin promotes colon tumour growth by inducing a senescence-associated secretory phenotype. *Gut*. **63** (12), 1932–1942 (2014).
19. Cuevas-Ramos, G. et al. Escherichia coli induces DNA damage in vivo and triggers genomic instability in mammalian cells. *Proc. Natl. Acad. Sci. USA*. **107** (25), 11537–11542 (2010).
20. Putze, J. et al. Genetic structure and distribution of the colibactin genomic island among members of the family *Enterobacteriaceae*. *Infect. Immun.* **77** (11), 4696–4703 (2009).
21. Sarshar, M. et al. Genetic diversity, phylogroup distribution and virulence gene profile of *pks* positive *Escherichia coli* colonizing human intestinal polyps. *Microb. Pathog.* **112**, 274–278 (2017).
22. Auvray, F. et al. Insights into the acquisition of the *pks* island and production of colibactin in the *Escherichia coli* population. *Microb. Genom.* **7** (5), 000579 (2021).
23. Oswald, E., Nougayrède, J. P., Taieb, F. & Sugai, M. Bacterial toxins that modulate host cell-cycle progression. *Curr. Opin. Microbiol.* **8** (1), 83–91 (2005).
24. Sun, J. Impact of bacterial infection and intestinal microbiome on colorectal cancer development. *Chin. Med. J. (Engl).* **135** (4), 400–408 (2022).
25. Maddocks, O. D., Short, A. J., Donnenberg, M. S., Bader, S. & Harrison, D. J. Attaching and effacing *Escherichia coli* downregulate DNA mismatch repair protein *in vitro* and are associated with colorectal adenocarcinomas in humans. *PLoS One* **4**(5), e5517.
26. Mathlouthi, N. E. H., Kriaa, A., Keskes, L. A., Rhimi, M. & Gdoura, R. Virulence factors in colorectal cancer metagenomes and association of microbial siderophores with advanced stages. *Microorganisms*. **10** (12), 2365 (2022).
27. Martin, H. M. et al. Enhanced *Escherichia coli* adherence and invasion in Crohn's disease and colon cancer. *Gastroenterology*. **127** (1), 80–93 (2004).
28. Chardalias, L., Papaconstantinou, I., Gklavas, A., Politou, M. & Theodosopoulos, T. Iron deficiency anemia in colorectal cancer patients: is preoperative intravenous iron infusion indicated? A narrative review of the literature. *Cancer Diagn. Progn.* **3** (2), 163–168 (2023).
29. Ellermann, M. & Arthur, J. C. Siderophore-mediated iron acquisition and modulation of host-bacterial interactions. *Free Radic Biol. Med.* **105**, 68–78 (2017).
30. Sobhani, I. et al. Microbial dysbiosis in colorectal cancer (CRC) patients. *PloS One*. **6** (1), e16393 (2011).
31. Zeller, G. et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10** (11), 766 (2014).
32. Dejea, C. M. et al. Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria. *Sci. (New York N Y)*. **359** (6375), 592–597 (2018).
33. Thomas, A. M. et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **25** (4), 667–678 (2019).
34. Purcell, R. V., Visnovska, M., Biggs, P. J., Schmeier, S. & Frizelle, F. A. Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. *Sci. Rep.* **7** (1), 11590 (2017).
35. Gupta, V. K., Paul, S., Dutta, C. & Geography Ethnicity or subsistence-specific variations in Human Microbiome Composition and Diversity. *Front. Microbiol.* **8**, 1162 (2017).
36. Waters, N. R., Abram, F., Brennan, F., Holmes, A. & Pritchard, L. Easy phylotyping of *Escherichia coli via* the EzClermont web app and command-line tool. *Access. Microbiol.* **2** (9), acmi000143 (2020).
37. Ageorges, V. et al. Genome-wide analysis of Antigen 43 (Ag43) variants: New insights in their diversity, distribution and prevalence in Bacteria. *Int. J. Mol. Sci.* **24** (6), 5500 (2023).
38. Siegel, R. L., Wagle, N. S., Cercek, A., Smith, R. A. & Jemal, A. Colorectal cancer statistics, 2023. *CA Cancer J. Clin.* **73** (3), 233–254 (2023).
39. Cao, Y. et al. Commensal microbiota from patients with inflammatory bowel disease produce genotoxic metabolites. *Science*. **378** (6618), eabm3233 (2022).
40. Yurdakul, D., Yazgan-Karataş, A. & Şahin, F. Enterobacter strains might promote colon cancer. *Curr. Microbiol.* **71** (3), 403–411 (2015).
41. Strakova, N., Korena, K. & Karpiskova, R. *Klebsiella pneumoniae* producing bacterial toxin colibactin as a risk of colorectal cancer development—A systematic review. *Toxicon*. **197**, 126–135 (2021).
42. Wassenaar, T. M. E. Coli and colorectal cancer: a complex relationship that deserves a critical mindset. *Crit. Rev. Microbiol.* **44** (5), 619–632 (2018).
43. Picard, B. et al. The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection. *Infect. Immun.* **67** (2), 546–553 (1999).
44. Duriez, P. et al. Commensal *Escherichia coli* isolates are phylogenetically distributed among geographically distinct human populations. *Microbiol. (Reading)*. **147** (Pt 6), 1671–1676 (2001).
45. Russo, T. A. & Johnson, J. R. Proposal for a new inclusive designation for extraintestinal pathogenic isolates of *Escherichia coli*: ExPEC. *J. Infect. Dis.* **181** (5), 1753–1754 (2000).
46. Buc, E. et al. High prevalence of mucosa-associated *E. Coli* producing cyclomodulin and genotoxin in colon cancer. *PloS One* **8**(2), e56964 (2013).
47. Taieb, F., Petit, C., Nougayrède, J. P. & Oswald, E. The enterobacterial genotoxins: Cytolethal distending toxin and colibactin. *EcoSal Plus*. **7** (1). https://doi.org/10.1128/ecosalplus.ESP-0008-2016 (2016).
48. Aguirre-Sánchez, J. R. et al. Phylogenetic group and virulence profile classification in *Escherichia coli* from distinct isolation sources in Mexico. *Infect. Genet. Evol.* **106**, 105380 (2022).
49. Danese, P. N., Pratt, L. A., Dove, S. L. & Kolter, R. The outer membrane protein, antigen 43, mediates cell-to-cell interactions within *Escherichia coli* biofilms. *Mol. Microbiol.* **37** (2), 424–432 (2000).

50. Kjaergaard, K., Schembri, M. A., Hasman, H. & Klemm, P. Antigen 43 from *Escherichia coli* induces inter- and intraspecies cell aggregation and changes in colony morphology of *Pseudomonas fluorescens*. *J. Bacteriol.* **182** (17), 4789–4796 (2000).
51. Klemm, P., Hjerrild, L., Gjermansen, M. & Schembri, M. A. Structure-function analysis of the self-recognizing Antigen 43 autotransporter protein from *Escherichia coli*. *Mol. Microbiol.* **51** (1), 283–296 (2004).
52. Dahmus, J. D., Kotler, D. L., Kastenberg, D. M. & Kistler, C. A. The gut microbiome and colorectal cancer: a review of bacterial pathogenesis. *J. Gastrointest. Oncol.* **9** (4), 769–777 (2018).
53. Mirzaei, R. et al. Bacterial biofilm in colorectal cancer: what is the real mechanism of action? *Microb. Pathog.* **142**, 104052 (2020).
54. Court, D. L., Oppenheim, A. B. & Adhya, S. L. A new look at bacteriophage lambda genetic networks. *J. Bacteriol.* **189** (2), 298–304 (2007).
55. Burmeister, A. R. Horizontal gene transfer. *Evol. Med. Public. Health*. **2015** (1), 193–194 (2015).
56. Obeng-Nkrumah, N., Twum-Danso, K., Krogfelt, K. A. & Newman, M. J. High levels of extended-spectrum beta-lactamases in a major teaching hospital in Ghana: the need for regular monitoring and evaluation of antibiotic resistance. *Am. J. Trop. Med. Hyg.* **89** (5), 960–964 (2013).
57. Ohene Larbi, R., Ofori, L. A., Sylverken, A. A., Ayim-Akonor, M. & Obiri-Danso, K. Antimicrobial resistance of *Escherichia coli* from broilers, pigs, and cattle in the Greater Kumasi Metropolis, Ghana. *Int. J. Microbiol.* 5158185 (2021).
58. Egyir, B. et al. Antimicrobial resistance and genomic analysis of staphylococci isolated from livestock and farm attendants in Northern Ghana. *BMC Microbiol.* **22** (1), 180 (2022).
59. Amuasi, G. R. et al. Enterococcus species: insights into antimicrobial resistance and whole-genome features of isolates recovered from livestock and raw meat in Ghana. *Front. Microbiol.* **14**, 1254896 (2023).
60. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. **30** (15), 2114–2120 (2014).
61. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19** (5), 455–477 (2012).
62. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. **29** (8), 1072–1075 (2013).
63. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. **30** (14), 2068–2069 (2014).
64. Page, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. **31** (22), 3691–3693 (2015).
65. Liu, B., Zheng, D., Jin, Q., Chen, L. & Yang, J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.* **47** (D1), D687–D692 (2019).
66. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. **30** (9), 1312–1313 (2014).
67. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47** (W1), W256–W259 (2019).

Acknowledgements

Patient recruitment and initial sample processing were supported by the BANGA-Africa Seed Research Grant from the University of Ghana. Bacterial identification and whole genome sequencing were funded by the SeqAfrica project, which is supported by the UK Department of Health and Social Care's Fleming Fund through UK aid. The views expressed in this publication are those of the authors and do not necessarily reflect those of the UK Department of Health and Social Care or its management agent, Mott MacDonald. The authors are grateful to Christian Owusu-Nyantakyi and Felicia Amoa Owusu at the Bacteriology Department of the Noguchi Memorial Institute For Medical Research for their vital roles in bacterial identification, whole genome sequencing and initial bioinformatic analyses.

Author contributions

S.V.B.: Methodology, formal analysis, investigation, data curation. S.Y.B.: Formal analysis, writing (review and editing). R.T.A.: investigation. A.A.A.B.B.: Writing (review and editing), supervision. B.E.: Methodology, formal analysis, writing (review and editing). S.E.T.: Validation. B.D.: Conceptualization, writing (review and editing), supervision. V.A.: Conceptualization, methodology, formal analysis, data curation, writing (original draft preparation, review and editing), supervision, funding.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

Ethical approval for this study was obtained from the Institutional Review Board of the College of Health Sciences at the University of Ghana (CHS-Et/M1-P4.2/2021–2022). All methods used in this study were in accordance with the regulations approved by the above-mentioned Ethics Board. Prior to collection of socio-demographic data and stool samples, written informed consent was obtained from the CRC patients and healthy control participants for their anonymized information to be used for the purposes of this study.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-74299-3.

Correspondence and requests for materials should be addressed to B.D. or V.A.