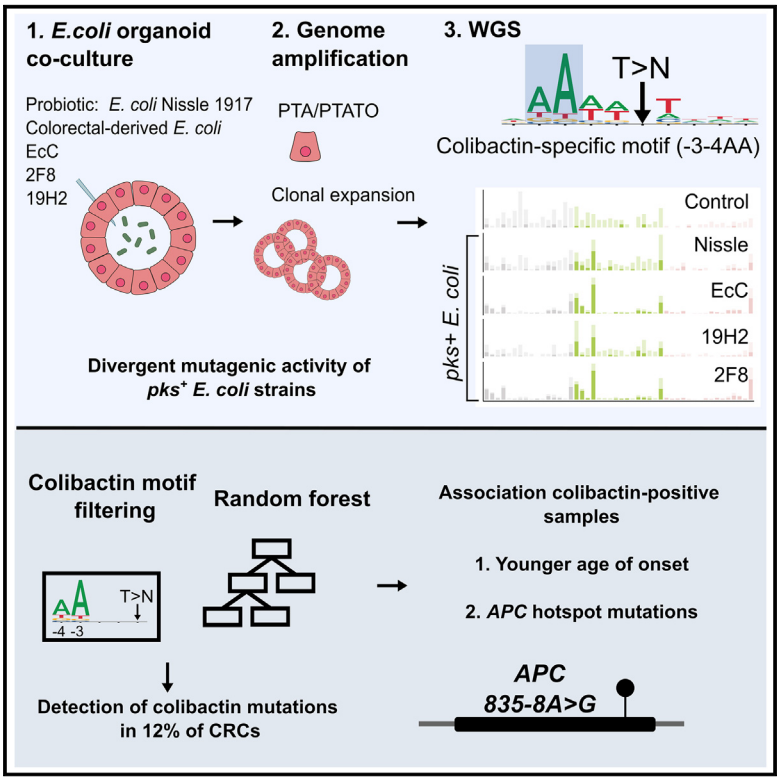
Cancer Cell

Report

Improved detection of colibactin-induced mutations by genotoxic *E. coli* in organoids and colorectal cancer

Graphical abstract



Highlights

- Detection of specific mutations induced by *pks*$^{+}$ *E. coli* strains, including Nissle 1917
- Mutation classifier indicates 12% of CRC display colibactin mutagenesis
- Colibactin-associated CRC cases have *APC* mutations at colibactin motifs
- Colibactin-CRC cases have a younger age of onset in multiple cohorts

Authors

Axel Rosendahl Huber, Cayetano Pleguezuelos-Manzano, Jens Puschhof, Joske Ubels, ..., Hans Clevers, Ruben van Boxtel

Correspondence

jens.puschhof@dkfz-heidelberg.de (J.P.), h.clevers@hubrecht.eu (H.C.), r.vanboxtel@prinsesmaximacentrum.nl (R.v.B.)

In brief

In this study, Rosendahl Huber et al. show the mutagenic properties of *pks*$^{+}$ *E. coli* strains, including probiotic *E. coli* Nissle 1917, using the extended target sequence context of colibactin and with a machine-learning model. These approaches allow for better distinguishing of colibactin-associated colorectal cancer cases, which are younger and are enriched for *APC* mutations matching the colibactin motif.

Rosendahl Huber et al., 2024, Cancer Cell *42*, 487–496
March 11, 2024
https://doi.org/10.1016/j.ccell.2024.02.009

CellPress

Cancer Cell

CellPress
OPEN ACCESS

Report

Improved detection of colibactin-induced mutations by genotoxic *E. coli* in organoids and colorectal cancer

Axel Rosendahl Huber,[1,3,4,11] Cayetano Pleguezuelos-Manzano,[2,3,11] Jens Puschhof,[2,3,5,11,*] Joske Ubels,[1,3,11] Charelle Boot,[2,3] Aurelia Saftien,[2,3,5] Mark Verheul,[1,3] Laurianne T. Trabut,[1,3] Niels Groenen,[1,3] Markus van Roosmalen,[1,3] Kyanna S. Ouyang,[5] Henry Wood,[6] Phil Quirke,[6] Gerrit Meijer,[7] Edwin Cuppen,[8,9] Hans Clevers,[2,3,10,*] and Ruben van Boxtel[1,3,12,*]

[1]Princess Máxima Center for Pediatric Oncology, Heidelberglaan 25, 3584 CS Utrecht, the Netherlands
[2]Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences (KNAW) and UMC Utrecht, 3584 CT Utrecht, the Netherlands
[3]Oncode Institute, Utrecht, the Netherlands
[4]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Carrer de Baldiri Reixac, 10, 08028 Barcelona, Spain
[5]Microbiome and Cancer Division, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
[6]Pathology and Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK
[7]Department of Pathology, Netherlands Cancer Institute, Amsterdam, the Netherlands
[8]University Medical Center Utrecht, Utrecht, the Netherlands
[9]Hartwig Medical Foundation, Amsterdam, the Netherlands
[10]Roche Pharmaceutical Research and Early Development, 4058 Basel, Switzerland
[11]These authors contributed equally
[12]Lead contact
*Correspondence: jens.puschhof@dkfz-heidelberg.de (J.P.), h.clevers@hubrecht.eu (H.C.), r.vanboxtel@prinsesmaximacentrum.nl (R.v.B.)
https://doi.org/10.1016/j.ccell.2024.02.009

SUMMARY

Co-culture of intestinal organoids with a colibactin-producing *pks*$^+$ *E. coli* strain (EcC) revealed mutational signatures also found in colorectal cancer (CRC). *E. coli* Nissle 1917 (EcN) remains a commonly used probiotic, despite harboring the *pks* operon and inducing double strand DNA breaks. We determine the mutagenicity of EcN and three CRC-derived *pks*$^+$ *E. coli* strains with an analytical framework based on sequence characteristic of colibactin-induced mutations. All strains, including EcN, display varying levels of mutagenic activity. Furthermore, a machine learning approach attributing individual mutations to colibactin reveals that patients with colibactin-induced mutations are diagnosed at a younger age and that colibactin can induce a specific *APC* mutation. These approaches allow the sensitive detection of colibactin-induced mutations in ~12% of CRC genomes and even in whole exome sequencing data, representing a crucial step toward pinpointing the mutagenic activity of distinct *pks*$^+$ *E. coli* strains.

INTRODUCTION

E. coli strains associated with increased colorectal cancer (CRC) risk harbor the polyketide synthase (*pks*) operon.[1–5] This operon is responsible for the production of the genotoxin colibactin. Recent studies demonstrate that colibactin can alkylate adenines bivalently and cause DNA cross-links.[6,7] Indeed, *pks*$^+$ strains induce DNA-double strand breaks (DSB) in cell lines.[2] Furthermore, the co-culture of *pks*$^+$ *E. coli* with intestinal organoids and subsequent whole genome sequencing (WGS) revealed its ability to cause single base substitutions (SBS) and short insertions-deletions (ID) in the form of mutational signatures SBS88 and ID18, respectively.[8] SBS88- and ID18-related mutations are characterized by T > N substitutions and T deletions in adenine- and thymine-rich genomic regions,[8] in line with other reports indicating on colibactin-induced DSBs.[9] Simultaneous presence of SBS88 and ID18 could be detected in tumor genomes, of which the majority were CRC cases, pointing to colibactin as a source of mutations in CRC genomes.[10]

Several *E. coli* strains that belong to specific B2 phylogroup lineages harbor the *pks* operon,[11] but it is not clear if they have an equal capability to induce mutations in the epithelium.[12] *E. coli* Nissle 1917 (EcN) is a well-studied probiotic strain, commonly used to treat inflammatory bowel disease (IBD).[13] Notably, EcN harbors the *pks* operon in its genome.[2] Current evidence shows that EcN has diminished ability to cause DSBs compared to other *pks*$^+$ strains.[14] Additionally, a recent study using the *HPRT* gene assay indicates a mutagenic effect of EcN in CHO cells.[15] However, no evidence for genome-wide EcN-induced mutations in primary human cells exists to date, and

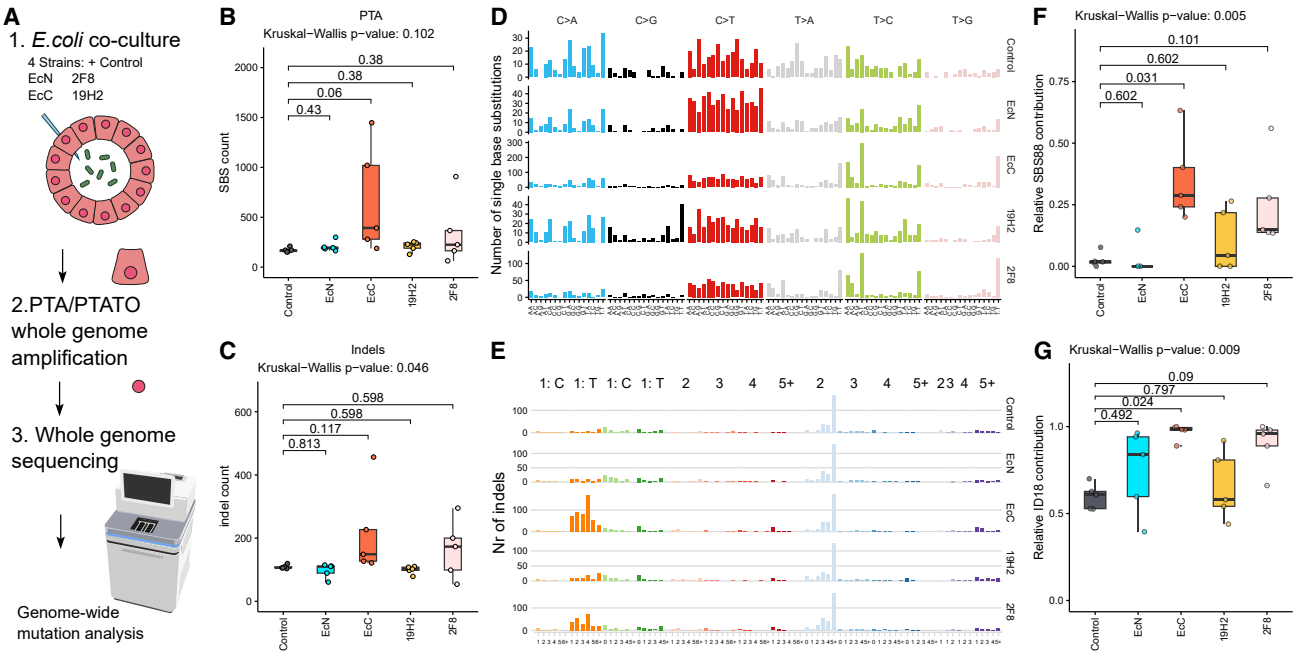


Figure 1. Mutational profiles and signature contributions of intestinal organoids exposed to *pks*$^+$ *E. coli* strains amplified by PTA

(A) Co-culture of *E. coli* strains with intestinal organoids and PTA genome amplification.

(B) Number of SBSs in EcN-, EcC-, 19H2-, 2F8-co-cultured or control organoids. N = 5 for all conditions. Box: upper and lower quartiles, center line: mean. Whiskers: largest or smallest value no more than 1.5 times the interquartile box range. p-values in B, and all other plots: Dunn's post-hoc test with FDR correction.

(C) Number of indels detected in EcN-, EcC-, 19H2-, 2F8-co-cultured, or control organoids. Boxplots according to B. N = 5 for all conditions.

(D) SBS spectra in organoids co-cultured with *E. coli* strains or control.

(E) 83-channel indel spectra in co-cultured organoids.

(F) SBS88 contribution in organoids co-cultured with *E. coli* strains or control. Signatures considered: SBS1, SBS5, SBS18 (*in vitro*), and SBS88 (colibactin). N = 5 for all conditions.

(G) ID18 contribution in organoids co-cultured with *E. coli* strains or control. Signatures considered: ID1, ID2 (*in vitro* culture), and ID18 (colibactin). N = 5 for all conditions. Also see Figure S1 and Table S1.

its relative mutagenicity to other *pks*$^+$ *E. coli* strains is unknown. To address this, we determined the mutational consequences of a panel of *pks*$^+$ *E. coli* strains, consisting of EcN and 3 CRC-derived strains, using the previously established human organoid co-culture system followed by WGS.[8,16] Here, we develop 2 computational approaches, (1) relying on the colibactin DNA target motif and (2) a random forest model, to improve the detection of individual colibactin-induced mutations.

RESULTS

pks$^+$ *E. coli* co-culture screening reveals heterogenous mutagenic activity by different strains

First, we established an intestinal organoid co-culture panel comprising EcN and 3 additional CRC-derived *pks*$^+$ *E. coli* strains, CFF16-2F8 (2F8), CFF159-19H2 (19H2) (both from[17]), and the previously tested strain EcC[18] (STAR Methods). All strains showed comparable growth dynamics in co-culture, although EcN displayed slightly reduced expansion potential (Figure S1A). EcN caused DNA damage in organoids exposed for 24h, measured by the presence of nuclear γH2AX foci, a DSB marker (Figures S1B and S1C). While EcN did not induce the same level of DSBs as EcC, the DNA damage level was considerably increased over both negative controls, which were injected with dye or EcCΔ*clbQ*, an EcC *pks* mutant strain unable to produce colibactin (FDR-adjusted p values Wilcoxon test, EcN: 0.004; EcC: 0.004, EcCΔ*clbQ:* 0.0313, dye: 0.25) Figures S1B and S1C).[8,18]

To characterize the mutagenic effects of the *pks*$^+$ *E. coli* strains on co-cultured organoids, we performed single-cell WGS by primary template amplification (PTA) using the PTA analysis toolbox (PTATO, STAR Methods)[19,20] (Figure 1A; Table S1). SBS and indels numbers were similar across conditions in our experiments (Figures 1B and 1C). The SBS and ID mutational signature profiles of EcC- and 2F8-exposed organoids were similar to SBS88 and ID18, while EcN and 19H2 showed limited similarity (Figures 1D and 1E), evaluated by cosine similarity (Figure S2H). This is partially in line with the mutational signature refitting results (Including colibactin-induced SBS88 and ID18 and *in vitro* signatures SBS1, SBS5, SBS18, ID1, and ID2) where only organoids exposed to EcC have a significant contribution of SBS88 (p value 0.031; Dunn's test with FDR correction) (Figure 1F) and ID18 (p value 0.024; Dunn's test with FDR correction) (Figure 1G). However, traces of the most characteristic SBS88 peaks are observable for both EcN-, 19H2-, and especially in 2F8-exposed organoids, suggesting that mutational signature refitting is not sensitive enough in samples with low signal-to-noise ratio. We repeated the experiment using clonal expansion of organoids exposed to dye, EcC or EcN with comparable results (Figures S2D–S2L; Table S1).

CellPress
OPEN ACCESS

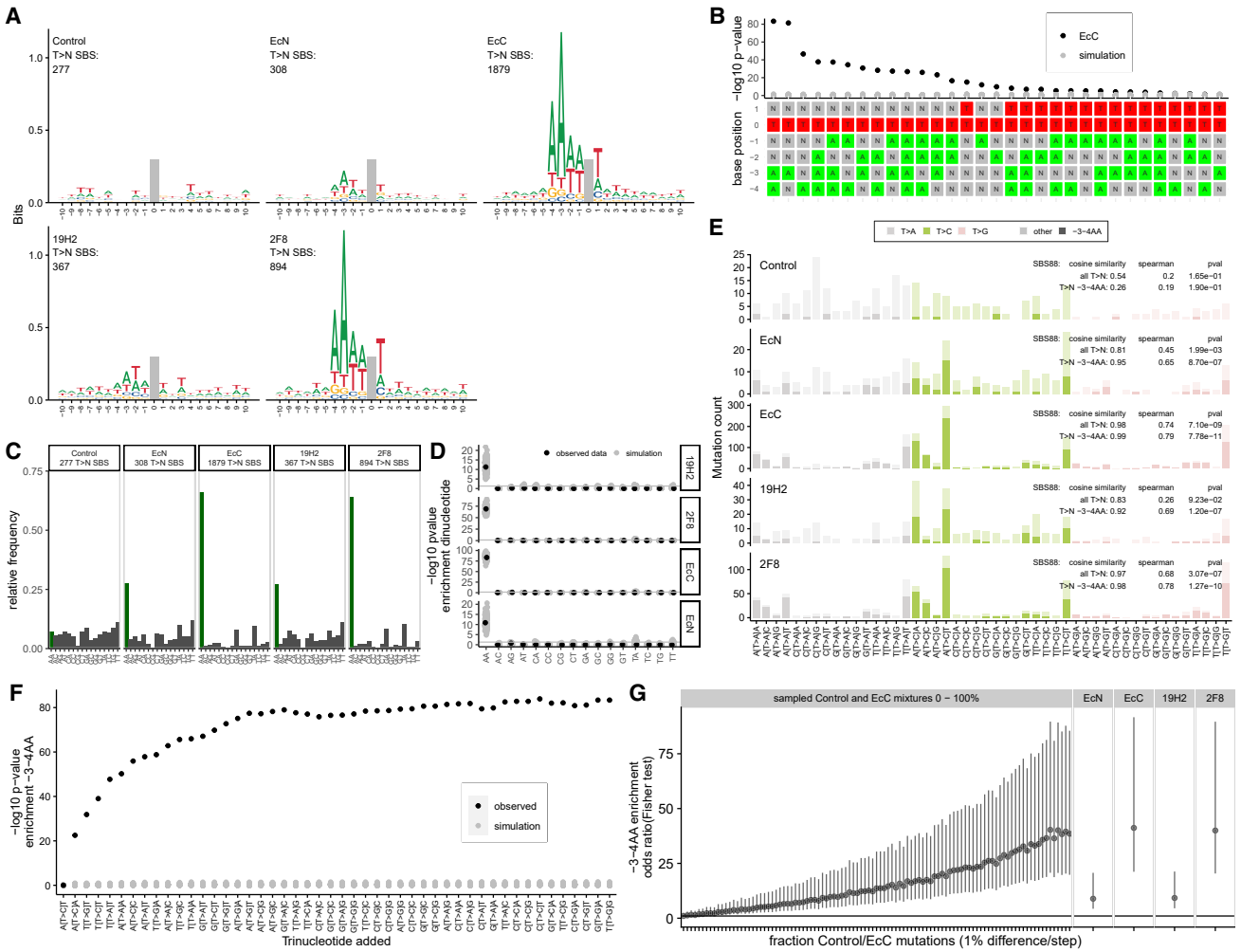


Figure 2. A colibactin-specific -3-4AA mutational motif is enriched in organoids co-cultured with four *pks*$^{+}$ *E. coli* strains, including EcN

(A) Sequence logo indicating the enrichment of bases flanking T > N mutations (gray) using information content.

(B) p values from one-tailed Fisher' exact test comparing the enrichment of motifs between colibactin-exposed organoids to control. Lower figure panel: DNA motifs tested.

(C) Relative levels of dinucleotide occurrence at −3 and −4 position from T > N mutations in organoids exposed to the different *E. coli* strains or control.

(D) p values for enrichment of dinucleotide occurrence at −3 and −4 position from T > N mutations relative to mutations in control organoids.

(E) T > N trinucleotide SBS mutations by −3 and −4 upstream bases. Dark: SBS mutations with -3-4AA. Light: all other mutations. Cosine similarity, Spearman correlation and FDR-corrected p value indicated.

(F) Optimization of trinucleotide motifs (STAR Methods): -log10 transformed p values from one-tailed Fisher's exact tests. Sequentially accumulating mutations in most prevalent SBS88 trinucleotides.

(G) Estimation of the relative mutagenicity of *pks*$^{+}$ strains compared to EcC/control sampled mixtures. Dots: odds ratio's obtained from a two-tailed Fisher's exact test. Vertical lines: 95% confidence interval. x axis left: stepwise increase of EcC-derived mutations. Right: Enrichment scores of *E. coli* strains. Also see Figures S2 and S3.

Motif filtering improves detection of colibactin-induced mutations

Since the presence of other mutational processes affects the detection of a given signature in mutational datasets,[21] we used the extended DNA contexts of colibactin-induced mutations[8] to optimize the detection. These contexts appeared in organoids exposed to each strain, and not in control (Figure 2A). The presence of two adenines 3 and 4 bases upstream of the T > N mutation (-3-4AA) was the most significantly enriched motif when comparing mutations from EcC and control (p value $4.40*10^{-84}$, one-sided Fisher's exact test) (Figure 2B). All exposed organoid genomes (to EcN, 19H2, and 2F8) presented a significant enrichment of colibactin-induced mutations with adenines at the -3-4AA positions (Figures 2C and 2D, p value EcN = $1.28*10^{-11}$, 2F8 = $7.50*10^{-70}$, 19H2 = $5.10*10^{-12}$; one-sided Fisher's exact test). Cosine similarity and Spearman correlation indicated similarity between T > N trinucleotide profiles of all *pks*$^{+}$-exposed organoids and SBS88. This similarity was more pronounced when considering only mutations with -3-4AA colibactin motif, including those to EcN, 2F8, and 19H2 (Figure 2E).

We further optimized the number T > N trinucleotides used to best distinguish colibactin mutations (STAR Methods). Using mutations occurring at the 17 most frequent SBS88 trinucleotides

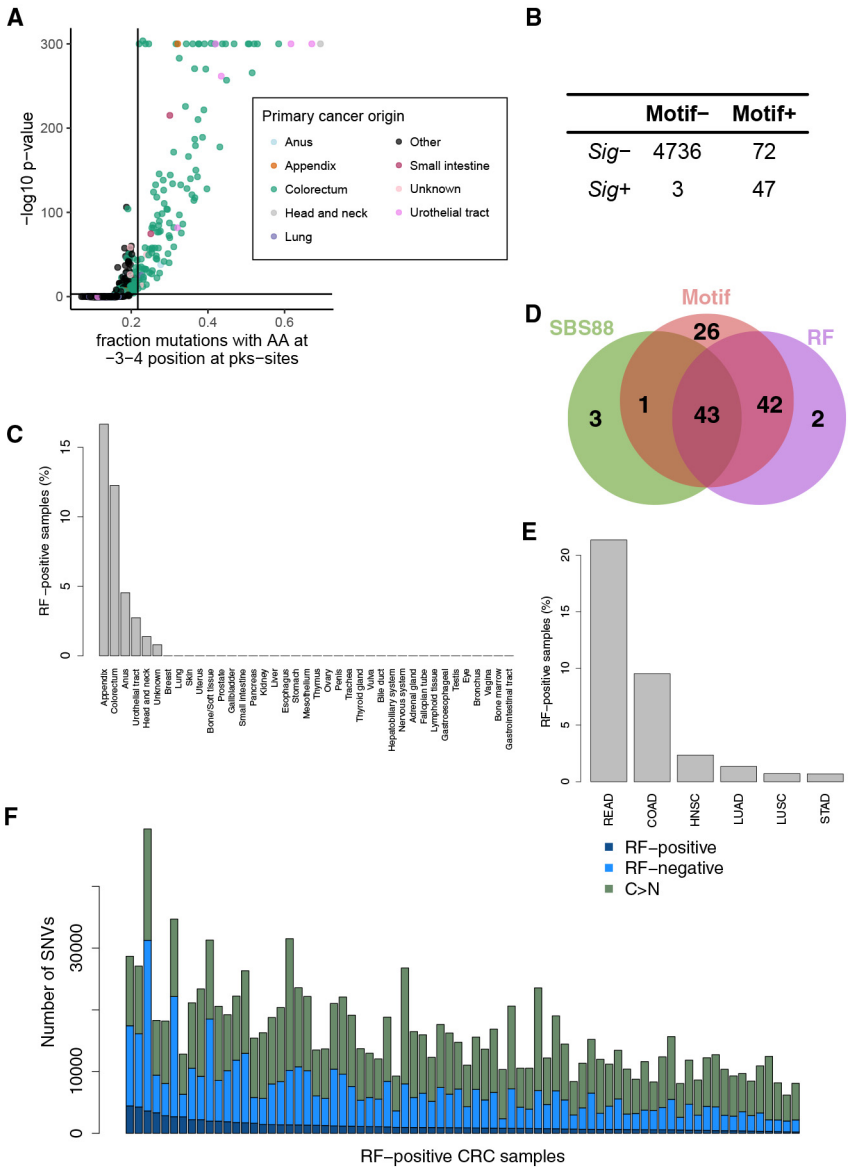


Figure 3. Motif-based and RF-based classification improves the detection of colibactin mutations WGS cancer data

(A) -$\log_{10}$ transformed p values (one-sided Fisher's exact test with FDR correction) versus the fraction of -3-4AA at colibactin T > N mutations for the HMF cohort.

(B) Summary of HMF WGS sample classification using motif-based and mutational signature refitting-based approaches.

(C) Percentage of samples showing colibactin-induced damage per tumor type in the HMF cohort, as classified by the RF model.

(D) Venn diagram showing the overlap between colibactin-postive samples using signature refitting, -3-4AA motif counting and the RF model.

(E) Percentage of samples showing colibactin-induced damage per tumor type in the TCGA cohort.

(F) Number of mutations in all CRC samples in the HMF cohort attributed to colibactin. Also see Figures S4–S6.

resulted in the most significant enrichment of -3-4AA presence against the control (Figure 2F, p value = $3.67*10^{-78}$ for -3-4AA enrichment in EcC-treated organoids compared to control, $2.03*10^{-13}$ for EcN compared to control, one-sided Fisher's exact test). Finally, by generating a sampled range of mutations from control and EcC-exposed organoids, we estimated the -3-4AA motif fraction of all strains relative to EcC (STAR Methods). In our organoid co-culture system, EcN had an estimated 32.9% (95% confidence interval between 21.2% and 62.9%) -3-4AA motif fraction relative to the EcC strain (Figure 2G). Additionally, 19H2 and 2F8 induced 33.7% (95% confidence interval between 21.7% and 64.4%) and 112,0% (95% confidence interval between 62.1% and 238.3%) of the -3-4AA motif fraction of EcC, respectively (Figure 2G). This variability in motif enrichment across strains was also demonstrated by the relative fraction of T > N substitutions with -3-4AA in exposed organoid cells, and enrichments remain stable in resampling-based analyses (Figures S3H and S3I). Additionally, all motif-based enrichments were similar to those obtained using data from clonally expanded organoids exposed to EcC and EcN (Figures S2A–S2G, S2J, and S2K).

Genetic differences between the pks island of genotoxic *E. coli* strains

To investigate if this divergent mutagenicity could be linked to sequence differences in the *pks* island, we compared the *pks* island genetic sequences of each strain. In line with a recent report on *pks* island diversity,[11] we found only few variants, of which most were single base changes (Figures S3A and S3B; Table S2). While some coding changes could influence colibactin production and secretion, most occurred in the self-protection gene *clbS* of the *pks* operon. In EcC most of these mutations occurred with an allele frequency of roughly 0.5 and the coverage of *clbS* was increased compared to neighboring regions, suggesting an allele duplication. (Figure S3C). Nevertheless, the small overall differences suggested other causes than *pks* island mutations as the source of mutagenic heterogeneity.

Detection of -3-4AA colibactin-mutations in cancer sequencing datasets

To test if this analytical framework could improve the detection of colibactin mutagenesis, we studied a WGS cohort consisting of more than 4,800 metastatic cancers (Hartwig Medical Foundation dataset; HMF).[22] 119 out of 4,858 samples (2.4%) displayed a significant (p value < 0.001, Fisher's exact test, one-sided) enrichment for the colibactin motif with a -3-4AA fraction higher than 0.22. We set this cutoff (Figure S4A; ROC curve, Youden index, optimal cutoff: 0.16, STAR Methods) to exclude potential false positive samples from tissues with implausible colibactin exposure, such as brain and bone tumors (Figure 3A, STAR Methods). The cohort

contained 656 CRC samples, of which 105 were classified as colibactin motif positive (16%) (Figures 3A and 3B). In addition, colibactin mutagenesis was detected in 2 out of 22 rectal (9.1%), 2 out of 66 small intestine (3%), 5 out of 191 urothelial tract (2.6%), 1 out of 73 head and neck (1.3%), and 1 out of 622 lung (0.16%) samples from the HMF cohort (Figure 3A). Next, we compared the motif classification method to signature refitting of SBS88 and ID18 (Figures S4B–S4D). The motif-based method allowed detection of samples with lower levels of SBS88 and ID18 mutational signatures or with high contribution of other mutational processes (Figures S4C and S4D), highlighting the advantage of using the -3-4AA motif to detect colibactin mutagenesis.

The motif-based analysis revealed a cluster of four -3-4AA colibactin motif-positive samples characterized by a high mutational load (Figure S5A) of which three harbored *POLE* hotspot mutations (*POLE*^{mut}) (Figure S5A). *POLE* encodes the catalytic subunit of DNA polymerase epsilon and hotspot driver mutations are known to result in a hypermutator phenotype.[23,24] POLE^{mut} samples are associated with mutational signatures SBS10a, SBS10b, and SBS28, the latter of which is marked by T>G mutations at T[T>G]T.[25] *POLE*^{mut}-associated T > N mutations displayed -3-4AA enrichment (Figure S5B),[9] only at T[T>G]T mutations (Figures S5C and S5D) and were more similar to SBS28 than SBS88 (cosine similarity, S5B). Thus, *POLE*^{mut} SBS28-enriched samples can be classified as false positives because of the presence of -3-4AA enrichment in T[T>G]T substitutions. Further, assessment of the specific detection of colibactin-induced mutations was demonstrated by Pearson correlation between the number of –3-4AA mutations and contributions of COSMIC mutational signatures. SBS88 was the only signature showing a clear correlation (R^2 0.88, $p < 2.2*10^{-16}$) (Figures S5E and S5F).

A random forest model for colibactin-linked mutation detection

To further investigate the specific mutations caused by colibactin, we employed a random forest (RF) model that can predict the probability that a mutation was caused by colibactin. We trained a model on both WGS data of EcC exposed organoids and CRC patients (Figure S6A, STAR Methods). These models place particular importance on the −3 and −4 position, in concordance with the motif analysis (Figure S6B). For the final probability, we multiply the posterior probability of both models. When classifying the CRC samples included in the HMF dataset, we observed a near perfect correlation between the relative contribution of SBS88 and the fraction of colibactin-induced mutations above the 10% threshold (Pearson's correlation = 0.92, p value < $2*10^{-16}$, Figure S6C), and a far lower correlation below that threshold, albeit still significant (Pearson's correlation = 0.44, p value < $2*10^{-14}$). Any sample with more than 10% contribution is considered positive. This RF prediction correlates with the -3-4AA enrichment found using the motif-based method (Figure S6D). The fraction of colibactin-induced mutations did not correlate meaningfully with any other mutational signature than SBS88 (Figures S5E and S5F).

We re-classified all samples in the HMF dataset with the RF model and found 12.3% of CRC samples to be positive (Figure 3C). In total 27 out of 635 samples were called as positive by the motif method but negative by the RF method (Figure 3D). Given the design of RF method to eliminate false positives, this classification could be more accurate.

Large WES datasets may yield additional information on prevalence and timing of colibactin-induced mutagenicity. However, for signature analysis, WGS data are preferred.[26] We assessed the performance of both the -3-4AA motif and the RF model on WES data (down-sampled from WGS; WGS classification considered as reference). Only considering mutations in exonic regions, the RF score showed a near perfect correlation between WES and WGS (Pearson correlation = 0.99, p value = $2.2*10^{-16}$ Figure S8E), while the -3-4AA counting showed more spread (Figure S6F). We thus used the RF model to classify a large WES cohort consisting of 2825 cancer WES genomes of The Cancer Genome Atlas (TCGA) (Figure 3E). Using an adapted WES threshold (Figure S6E), we showed that in total 12.5% of CRC samples are positive, in line with the estimations in the WGS HMF cohort (Figure 3C). We also observed a clear enrichment for positive cases in the rectal cancer samples, which was in line with earlier findings.[9]

CRC driver mutation analysis

Despite the detection of colibactin-induced mutagenesis by the RF, even in positive samples most mutations (93%, SD = 3.4%) have been caused by other processes (Figure 3F). To test whether colibactin-induced mutagenesis can contribute to oncogenesis, we classified mutations in known CRC driver genes. We selected genes with mutations in more than 5% of the samples in the IntOGen database[27] and in at least 5 CRC cases in the HMF dataset. This resulted in 10 CRC driver genes: *APC*, *TP53*, *KRAS*, *BRAF*, *PIK3CA*, *SMAD4*, *FBXW7*, *TCF7L2*, *FAT4*, and *ATM* (Figure 4A). We found that in randomly selected genes the mean difference in probability that mutations are caused by colibactin between the colibactin-positive and negative classes was 0.116 (SD = 0.21). The CRC genes *APC*, *SMAD4*, *BRAF*, *FBXW7*, *ATM*, and *TCF7L2* showed a larger difference in probability, ranging from 0.157–0.213 (Figure 4A). Of these, *APC* and *BRAF* mutations were significantly more likely to be induced by colibactin in colibactin-positive samples after correction for multiple testing (p value = 0.011 and 0.041, Student t test, respectively).

TP53 was significantly less likely to harbor colibactin-induced mutations compared to random genes in colibactin-positive CRC samples (p *value* = 0.0029, Student t test). When we classified all positions within *TP53*, we found that only 378 positions (3.9% of total positions) had a posterior probability to be mutated by colibactin above 0.5, whereas *APC* contains 5778 (8.5%) such positions. The probability distribution of *APC* mutations displayed an enrichment in mutations with high probability (Figure 4B), indicating specific colibactin-induced damage. The randomly selected genes showed the background distribution of colibactin damage, with a long tail in the positive class. This was absent in the probability distribution of *TP53*, showing the depletion of colibactin-induced damage.

Interestingly, we found a high probability mutation hotspot at c.835-8A-G in *APC*. This hotspot was strongly enriched in colibactin-positive CRC with 7.7% of the patients harboring this specific mutation versus 2.2% of negative patients (p value = 0.01, one-sided Fisher's exact test). The hotspot is in intron 8 of *APC* and has a predicted pathogenicity score by FATHMM[28] of 0.92 and an RF probability of 0.7592, and leads to a premature stop codon.[29] It has been reported in patients with both familial

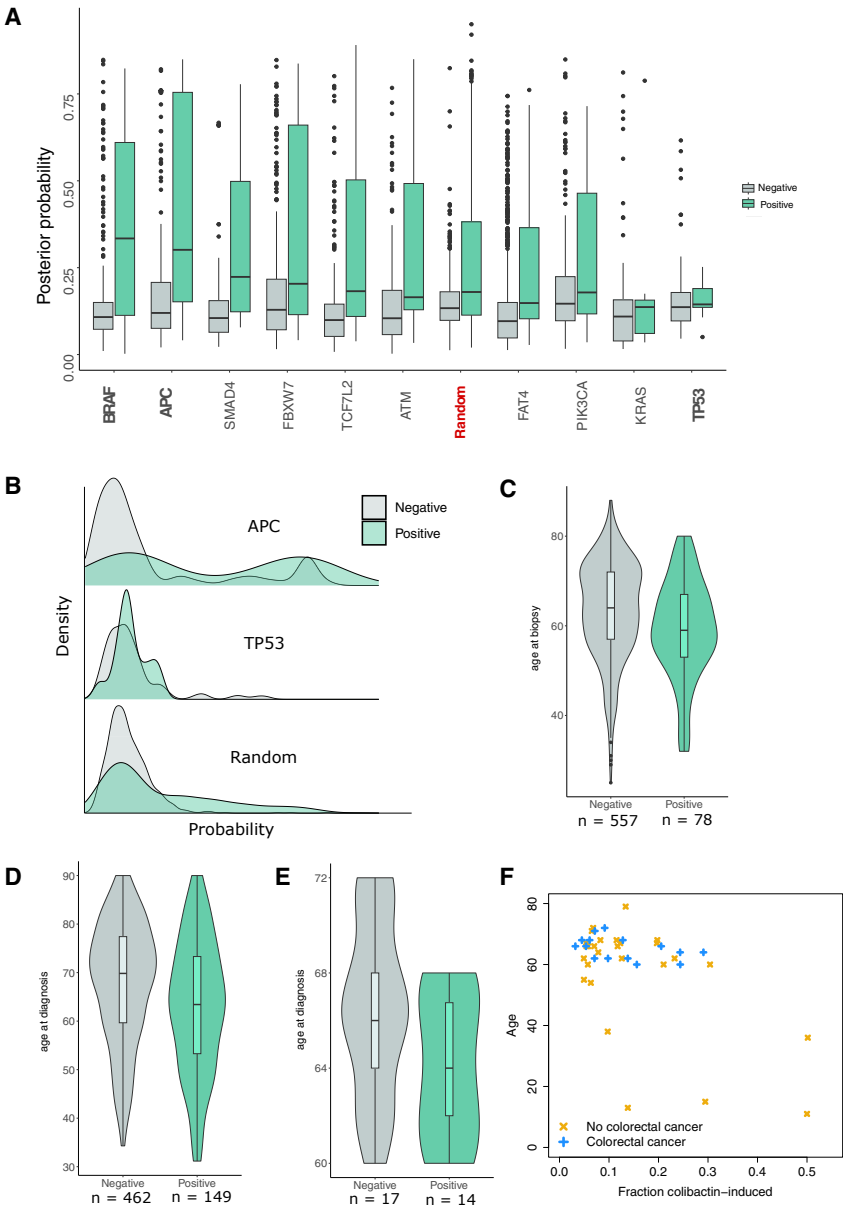


Figure 4. Colibactin-motif enables reliable detection of colibactin-induced mutagenesis in WES cohorts

(A) Posterior probability in the *pks*$^{+}$ and *pks*$^{-}$ class per driver gene and randomly sampled genes, sorted by difference between the classes. For all boxplots—whiskers indicate largest or smallest value no more than 1.5 times the interquartile range of the box.

(B) Posterior probability distribution of all SNVs in the RF-negative and positive class for APC, TP53, and the randomly sampled genes.

(C) Age of biopsy of metastasis for the HMF cohort for the positive and negative samples.

(D) Age at diagnosis in the TCGA cohort.

(E) Age at diagnosis for the screening cohort.

(F) Fraction of colibactin-induced mutations classified by the RF model in the screening cohort and age at biopsy, with no CRC being diagnosed in yellow points and CRC being diagnosed in the blue points.

adenomatous polyposis and unexplained colorectal polyposis.[30] This suggests a role for this specific mutation in the development of CRC.

Colibactin-linked mutations correlate with earlier CRC onset

APC mutations in CRC are predominantly explained by the aging signature SBS1.[31] However, the enrichment of mutations with high colibactin probability in *APC* could imply that colibactin-induced mutagenesis might accelerate the development of CRC in colibactin-positive individuals. For the HMF dataset, at tumor metastasis sampling colibactin-positive CRC patients were significantly younger than colibactin-negative patients (mean age: 58.48 versus 63.51 years, respectively Wilcoxon test, p value = 0.004) (Figure 4C). In the TCGA cohort, which consists of primary cancers, positive patients had a mean age of 62.96 at diagnosis vs. 67.96 in the negative set (Wilcoxon test, p value = $7*10^{-5}$) (Figure 4D). We also investigated a CRC screening cohort,[32] where healthy colon crypts were sequenced from patients undergoing a colonoscopy. Here we found that in the patient group where CRC was diagnosed, the mean age at diagnosis was lower for the colibactin-positive compared to negative patients albeit not significant (63.43 vs. 67.38 years, respectively, p value = 0.1, Wilcoxon test, Figure 4E). There is no correlation between the fraction of colibactin-induced damage, age of the patient, and whether CRC was diagnosed (Figure 4F). Taken together, these results suggested that while colibactin can prime cells for transformation early during life, additional hits that are caused by other mutational processes are necessary for tumorigenesis.

DISCUSSION

Both the motif-based and random forest-based classification allow to distinguish a larger group of tumors with enrichment of colibactin-linked mutations compared to mutational signature refitting. The colibactin-positive samples detected were mostly CRC, amounting to more than 12% of CRC cases. All other positive samples originated from organs harboring a microbiota, like the urinary tract, head and neck, lung, rectum, or small intestine. The absence of tumors from organs without a microbiota is indicative of preserved specificity. In addition, in healthy colorectal crypts, 21% of patients were reported using signature analysis to contain SBS88 and ID18-positive crypts.[32,33] The RF model enables interrogation of WES cohorts with a much lower false positive rate compared to signature refitting. WES signature refitting resulted in 30 samples being falsely classified as colibactin linked. However, the motif-based approach and RF model enable reliable detection of true

positive samples. This opens the door to systematically interrogating WES datasets for colibactin and potentially, other mutational patterns.

This study adds to evidence on EcN's DNA-damaging and mutagenic properties in relationship to its probiotic role,[15,34] yet the variance in DNA damage among colibactin producers[11,14] remains unexplained. We explored the *pks* island sequences of all strains used in this study and were able to detect a small number of genomic variants across strains. Beyond these, differences in production levels of rate limiting components of the *pks* enzymatic machinery, differences in how the toxin is exported and reaches the eukaryotic nuclei, as well as strain differences in metabolism of iron, spermidine, glucose, or inulin, which have been proposed to affect colibactin production ability,[35–39] could explain the differential mutagenic capacity. Finally, the relatively lower expansion speed of EcN compared to the cancer-derived strains in our experimental setup (Figure S1A) may potentially lead to an underestimation of its mutagenic capacity *in vivo*. Overall, the lack of correlation between intra-organoid expansion and mutagenicity across the whole strain panel suggests further factors influencing relative genotoxicity of strains. The human gut with a complete microbiota, mature mucus layer, and immune system, inter-individual differences in DNA repair efficiency and the duration of the exposure could further influence the mutagenic potential of *pks*$^+$ bacteria, including EcN. Whether cell-intrinsic or -extrinsic, the factors regulating colibactin production could be of clinical interest to target and reduce the mutagenic ability of *pks*$^+$ bacteria. Given that healthy colon cells accumulate only ~40 SBS mutations each year,[31,32] prolonged exposure of the human gut to even lowly mutagenic *pks*$^+$ strains could result in a markedly increased mutation load.

While earlier studies report hotspot mutations resulting in truncated *APC*,[8,14,30] *in vivo* evidence of colibactin-induced mutagenesis leading to transformation is lacking. Comparison of EcN with other *pks*$^+$ bacterial strains in such *in vivo* studies will help to elucidate the relative mutagenicity and specific risk caused by this probiotic strain. As EcN is used as a probiotic in conditions of varying severity and even in young patient groups, a careful assessment of its potential long-term mutagenicity in relation to clinical benefits is warranted for each of these use cases. Assessment of EcN-linked mutations in animal models and patients treated with EcN is required to determine the safety of this commonly prescribed probiotic. The framework presented in this manuscript is expected to translate well to *in vivo* datasets and could thereby contribute to future clinical assessment of EcN mutagenicity.

Limitations of the study

The mutational features of the frameworks used are not exclusively present in colibactin-induced mutations. A low background level of mutations within a -3-4AA motif or classified by the RF is present in all mutation catalogs, including our control organoid dataset that has been completely devoid of any colibactin exposure. Therefore, a minimal threshold of colibactin-induced mutations is needed for classification to *pks*$^+$ *E. coli*. In our study, we did not observe any specific differences in the mutational profiles induced by the different *pks*$^+$ *E. coli* strains. Thus, strain-specific classification, or even determining the influence of any probiotic treatment by assessing mutation characteristics is not possible. Although the RF model predicts that a particular driver mutation was likely to have been caused by colibactin-induced mutagenesis, this study does not allow us to casually link *pks*$^+$ *E. coli* exposure to the induction of cancer. Dedicated epidemiological studies, coupled with WGS/WES to determine past mutagenic activity of *pks*$^+$ *E. coli,* may help addressing whether exposure to *pks*$^+$ *E. coli*, including Nissle, increases the risks of cancer onset.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Organoid culture
  - *pks*$^+$ *E. coli* strains and co-culture with organoids
  - DNA damage quantification
  - DNA isolation and sequencing
  - Mapping and variant calling
  - Variant filtering
  - Mutational signature analysis
  - Extended context selection and enrichment testing
  - Estimation of relative mutagenicity of *E. coli strains* with respect to EcC
  - Monte-Carlo re-sampling analyses
  - *Pks* island sequence analysis
  - Assessment of extended context motifs enrichment in a cohort of metastatic cancer samples
  - Simulation of whole-exome (WES) data from the HMF cohort
  - TCGA analysis
  - Random forest model
  - Threshold for *pks* status of sample
  - Investigating driver genes
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.ccell.2024.02.009.

ACKNOWLEDGMENTS

CRUK grant OPTIMISTICC (C10674/A27140), the Netherlands Organ-on-Chip Initiative (024.003.001) from the Netherlands Organisation for Scientific Research (NWO) funded by the Ministry of Education, Culture and Science of the government of the Netherlands, and by a Stand Up to Cancer International Translational Cancer Research Grant, a program of the Entertainment Industry Foundation administered by the AACR (SU2C-3.1416) (J.P., C.P.-M., C.B., A.S., and H.C.). In addition, this work has been supported by a VIDI grant

from the Netherlands Organisation for Scientific Research (NWO) (no. 016.Vidi.171.023) (A.R.H. and R.v.B.), the Oncode Institute (partly financed by the Dutch Cancer Society) (A.R.H., J.P., C.P.-M., C.B., A.S., H.C., and R.v.B.) and the New York Stem Cell Foundation. R.v.B. is a New York Stem Cell Foundation—Robertson Investigator. Research reported in this publication was supported by Oncode Accelerator, a Dutch National Growth Fund project under grant number NGFOP2201. This work has been funded by an ERC consolidator grant from the European Research Council (ERC) no. 864499 to R.v.B. (J.U.).

The results published here are in whole or part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov. This publication and the underlying study have been made possible partly on the basis of the data that Hartwig Medical Foundation and the Center of Personalised Cancer Treatment (CPCT) have made available to the study.

We thank Guillaume Dalmasso and Richard Bonnet for derivation and sharing of the strains CCR (EcC), CFF16-2F8 (2F8), and CFF159-19H2 (19H2); as well as Sridhar Mani for sharing strain *Escherichia coli Nissle* 1917 (EcN).

AUTHOR CONTRIBUTIONS

Conceptualization: A.R.H., C.P.-M, J.P., J.U., H.C., and R.v.B.; software: J.U., A.R.H., and M.v.R.; validation: H.W.; formal analysis: J.U., A.R.H, H.W., and K.S.O.; resources: P.Q., E.C., and G.M.; investigation: C.P.-M, J.P., C.B., A.S., M.V., L.T.T, and N.G.; writing – original draft: A.R.H., C.P.-M, J.P., J.U., H.C., and R.v.B; writing – review & editing: all authors.

DECLARATION OF INTERESTS

H.C. is a full-time member of the executive board of F. Hoffmann-La Roche Ltd. as head of Pharma, Research and Early Development (pRED) in Basel, Switzerland. Additionally, H.C. is inventor on multiple organoid patents, licensed by the KNAW to the Foundation HUB in Utrecht.

Received: September 21, 2022
Revised: January 26, 2024
Accepted: February 14, 2024
Published: March 11, 2024

REFERENCES

1. Arthur, J.C., Perez-Chanona, E., Mühlbauer, M., Tomkovich, S., Uronis, J.M., Fan, T.-J., Campbell, B.J., Abujamel, T., Dogan, B., Rogers, A.B., et al. (2012). Intestinal inflammation targets cancer-inducing activity of the microbiota. Science (New York, N.Y.) *338*, 120–123. https://doi.org/10.1126/science.1224820.
2. Nougayrède, J.P., Homburg, S., Taieb, F., Boury, M., Brzuszkiewicz, E., Gottschalk, G., Buchrieser, C., Hacker, J., Dobrindt, U., and Oswald, E. (2006). Escherichia coli Induces DNA Double-Strand Breaks in Eukaryotic Cells. Science *313*, 848–851. https://doi.org/10.1126/science.1127059.
3. Wirbel, J., Pyl, P.T., Kartal, E., Zych, K., Kashani, A., Milanese, A., Fleck, J.S., Voigt, A.Y., Palleja, A., Ponnudurai, R., et al. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat. Med. *25*, 679–689. https://doi.org/10.1038/s41591-019-0406-6.
4. Yachida, S., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., Watanabe, H., Masuda, K., Nishimoto, Y., Kubo, M., et al. (2019). Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat. Med. *25*, 968–976. https://doi.org/10.1038/s41591-019-0458-7.
5. Pleguezuelos-Manzano, C., Puschhof, J., and Clevers, H. (2022). Gut Microbiota in Colorectal Cancer: Associations, Mechanisms, and Clinical Approaches. Annu. Rev. Cancer Biol. *6*, 65–84. https://doi.org/10.1146/annurev-cancerbio-070120-095211.
6. Xue, M., Kim, C.S., Healy, A.R., Wernke, K.M., Wang, Z., Frischling, M.C., Shine, E.E., Wang, W., Herzon, S.B., and Crawford, J.M. (2019). Structure Elucidation of Colibactin and its DNA Cross-Links. Science (New York, N.Y.) *365*, aax2685. https://doi.org/10.1126/science.aax2685.
7. Wilson, M.R., Jiang, Y., Villalta, P.W., Stornetta, A., Boudreau, P.D., Carrá, A., Brennan, C.A., Chun, E., Ngo, L., Samson, L.D., et al. (2019). The Human Gut Bacterial Genotoxin Colibactin Alkylates DNA. Science (New York, N.Y.) *363*, aar7785. https://doi.org/10.1126/science.aar7785.
8. Pleguezuelos-Manzano, C., Puschhof, J., Rosendahl Huber, A., Van Hoeck, A., Wood, H.M., Nomburg, J., Gurjao, C., Manders, F., Dalmasso, G., Stege, P.B., et al. (2020). Mutational signature in colorectal cancer caused by genotoxic pks+ E. coli. Nature *580*, 269–273. https://doi.org/10.1038/s41586-020-2080-8.
9. Dziubańska-Kusibab, P.J., Berger, H., Battistini, F., Bouwman, B.A.M., Iftekhar, A., Katainen, R., Cajuso, T., Crosetto, N., Orozco, M., Aaltonen, L.A., and Meyer, T.F. (2020). Colibactin DNA-damage signature indicates mutational impact in colorectal cancer. Nat. Med. *26*, 1063–1069. https://doi.org/10.1038/s41591-020-0908-2.
10. Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N., et al. (2020). The repertoire of mutational signatures in human cancer. Nature *578*, 94–101. https://doi.org/10.1038/s41586-020-1943-3.
11. Auvray, F., Perrat, A., Arimizu, Y., Chagneau, C.V., Bossuet-Greif, N., Massip, C., Brugère, H., Nougayrède, J.P., Hayashi, T., Branchu, P., et al. (2021). Insights into the acquisition of the pks island and production of colibactin in the Escherichia coli population. Microb. Genom. *7*, 000579. https://doi.org/10.1099/mgen.0.000579.
12. Bossuet-Greif, N., Vignard, J., Taieb, F., Mirey, G., Dubois, D., Petit, C., Oswald, E., and Nougayrède, J.P. (2018). The Colibactin Genotoxin Generates DNA Interstrand Cross-Links in Infected Cells. mBio *9*, e02393-17. https://doi.org/10.1128/mBio.02393-17.
13. Schultz, M. (2008). Clinical use of E. coli Nissle 1917 in inflammatory bowel disease. Inflamm. Bowel Dis. *14*, 1012–1018. https://doi.org/10.1002/ibd.20377.
14. Iftekhar, A., Berger, H., Bouznad, N., Heuberger, J., Boccellato, F., Dobrindt, U., Hermeking, H., Sigal, M., and Meyer, T.F. (2021). Genomic aberrations after short-term exposure to colibactin-producing E. coli transform primary colon epithelial cells. Nat. Commun. *12*, 1003. https://doi.org/10.1038/s41467-021-21162-y.
15. Nougayrède, J.P., Chagneau, C.V., Motta, J.P., Bossuet-Greif, N., Belloy, M., Taieb, F., Gratadoux, J.J., Thomas, M., Langella, P., and Oswald, E. (2021). A Toxic Friend: Genotoxic and Mutagenic Activity of the Probiotic Strain Escherichia coli Nissle 1917. mSphere *6*, e0062421. https://doi.org/10.1128/mSphere.00624-21.
16. Puschhof, J., Pleguezuelos-Manzano, C., Martinez-Silgado, A., Akkerman, N., Saftien, A., Boot, C., de Waal, A., Beumer, J., Dutta, D., Heo, I., and Clevers, H. (2021). Intestinal organoid cocultures with microbes. Nat. Protoc. *16*, 4633–4649. https://doi.org/10.1038/s41596-021-00589-z.
17. Buc, E., Dubois, D., Sauvanet, P., Raisch, J., Delmas, J., Darfeuille-Michaud, A., Pezet, D., and Bonnet, R. (2013). High Prevalence of Mucosa-Associated E. coli Producing Cyclomodulin and Genotoxin in Colon Cancer. PLoS One *8*, e56964. https://doi.org/10.1371/journal.pone.0056964.
18. Cougnoux, A., Dalmasso, G., Martinez, R., Buc, E., Delmas, J., Gibold, L., Sauvanet, P., Darcha, C., Déchelotte, P., Bonnet, M., et al. (2014). Bacterial genotoxin colibactin promotes colon tumour growth by inducing a senescence-associated secretory phenotype. Gut *63*, 1932–1942. https://doi.org/10.1136/gutjnl-2013-305257.
19. Gonzalez-Pena, V., Natarajan, S., Xia, Y., Klein, D., Carter, R., Pang, Y., Shaner, B., Annu, K., Putnam, D., Chen, W., et al. (2021). Accurate genomic variant detection in single cells with primary template-directed amplification. Proc. Natl. Acad. Sci. USA *118*, e2024176118. https://doi.org/10.1073/pnas.2024176118.

20. Middelkamp, S., Manders, F., Peci, F., Roosmalen, M.J.v., González, D.M., Bertrums, E.J.M., Werf, I.v.d., Derks, L.L.M., Groenen, N.M., Verheul, M., et al. (2023). Comprehensive Single-Cell Genome Analysis at Nucleotide Resolution Using the PTA Analysis Toolbox. Preprint at bioRxiv. https://doi.org/10.1101/2023.02.15.528636.

21. Kucab, J.E., Zou, X., Morganella, S., Joel, M., Nanda, A.S., Nagy, E., Gomez, C., Degasperi, A., Harris, R., Jackson, S.P., et al. (2019). A Compendium of Mutational Signatures of Environmental Agents. Cell *177*, 821–836.e16. https://doi.org/10.1016/j.cell.2019.03.001.

22. Priestley, P., Baber, J., Lolkema, M.P., Steeghs, N., de Bruijn, E., Shale, C., Duyvesteyn, K., Haidari, S., van Hoeck, A., Onstenk, W., et al. (2019). Pan-cancer whole-genome analyses of metastatic solid tumours. Nature *575*, 210–216. https://doi.org/10.1038/s41586-019-1689-y.

23. Campbell, B.B., Light, N., Fabrizio, D., Zatzman, M., Fuligni, F., de Borja, R., Davidson, S., Edwards, M., Elvin, J.A., Hodel, K.P., et al. (2017). Comprehensive Analysis of Hypermutation in Human Cancer. Cell *171*, 1042–1056.e10. https://doi.org/10.1016/j.cell.2017.09.048.

24. Rayner, E., van Gool, I.C., Palles, C., Kearsey, S.E., Bosse, T., Tomlinson, I., and Church, D.N. (2016). A panoply of errors: polymerase proofreading domain mutations in cancer. Nat. Rev. Cancer *16*, 71–81. https://doi.org/10.1038/nrc.2015.12.

25. Hodel, K.P., Sun, M.J.S., Ungerleider, N., Park, V.S., Williams, L.G., Bauer, D.L., Immethun, V.E., Wang, J., Suo, Z., Lu, H., et al. (2020). POLE Mutation Spectra Are Shaped by the Mutant Allele Identity, Its Abundance, and Mismatch Repair Status. Mol. Cell *78*, 1166–1177.e6. https://doi.org/10.1016/j.molcel.2020.05.012.

26. Koh, G., Degasperi, A., Zou, X., Momen, S., and Nik-Zainal, S. (2021). Mutational signatures: emerging concepts, caveats and clinical applications. Nat. Rev. Cancer *21*, 619–637. https://doi.org/10.1038/s41568-021-00377-7.

27. Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M.P., Jene-Sanz, A., Santos, A., and Lopez-Bigas, N. (2013). IntOGen-mutations identifies cancer drivers across tumor types. Nat. Methods *10*, 1081–1082. https://doi.org/10.1038/nmeth.2642.

28. Rogers, M.F., Shihab, H.A., Mort, M., Cooper, D.N., Gaunt, T.R., and Campbell, C. (2018). FATHMM-XF:Accurate prediction of pathogenic point mutations via extended features. Bioinformatics. Published online. https://doi.org/10.1093/bioinformatics/btx536.

29. Fostira, F., Thodi, G., Sandaltzopoulos, R., Fountzilas, G., and Yannoukakos, D. (2010). Mutational spectrum of APC and genotype-phenotype correlations in Greek FAP patients. BMC Cancer *10*, 389. https://doi.org/10.1186/1471-2407-10-389.

30. Terlouw, D., Suerink, M., Boot, A., van Wezel, T., Nielsen, M., and Morreau, H. (2020). Recurrent APC Splice Variant c.835-8A>G in Patients With Unexplained Colorectal Polyposis Fulfilling the Colibactin Mutational Signature. Gastroenterology *159*, 1612–1614.e5. https://doi.org/10.1053/j.gastro.2020.06.055.

31. Blokzijl, F., de Ligt, J., Jager, M., Sasselli, V., Roerink, S., Sasaki, N., Huch, M., Boymans, S., Kuijk, E., Prins, P., et al. (2016). Tissue-specific mutation accumulation in human adult stem cells during life. Nature *538*, 260–264. https://doi.org/10.1038/nature19768.

32. Lee-Six, H., Olafsson, S., Ellis, P., Osborne, R.J., Sanders, M.A., Moore, L., Georgakopoulos, N., Torrente, F., Noorani, A., Goddard, M., et al. (2019). The landscape of somatic mutation in normal colorectal epithelial cells. Nature *574*, 532–537. https://doi.org/10.1038/s41586-019-1672-7.

33. Cagan, A., Baez-Ortega, A., Brzozowska, N., Abascal, F., Coorens, T.H.H., Sanders, M.A., Lawson, A.R.J., Harvey, L.M.R., Bhosle, S., Jones, D., et al. (2022). Somatic mutation rates scale with lifespan across mammals. Nature *604*, 517–524. https://doi.org/10.1038/s41586-022-04618-z.

34. Massip, C., Branchu, P., Bossuet-Greif, N., Chagneau, C.V., Gaillard, D., Martin, P., Boury, M., Sécher, T., Dubois, D., Nougayrède, J.P., and Oswald, E. (2019). Deciphering the interplay between the genotoxic and probiotic activities of Escherichia coli Nissle 1917. PLoS Pathog. *15*, e1008029. https://doi.org/10.1371/journal.ppat.1008029.

35. Oliero, M., Calvé, A., Fragoso, G., Cuisiniere, T., Hajjar, R., Dobrindt, U., and Santos, M.M. (2021). Oligosaccharides increase the genotoxic effect of colibactin produced by pks+ Escherichia coli strains. BMC Cancer *21*, 172. https://doi.org/10.1186/s12885-021-07876-8.

36. Wallenstein, A., Rehm, N., Brinkmann, M., Selle, M., Bossuet-Greif, N., Sauer, D., Bunk, B., Spröer, C., Wami, H.T., Homburg, S., et al. (2020). ClbR Is the Key Transcriptional Activator of Colibactin Gene Expression in Escherichia coli. mSphere *5*, e00591-20. https://doi.org/10.1128/mSphere.00591-20.

37. Tronnet, S., Garcie, C., Rehm, N., Dobrindt, U., Oswald, E., and Martin, P. (2016). Iron Homeostasis Regulates the Genotoxicity of Escherichia coli That Produces Colibactin. Infect. Immun. *84*, 3358–3368. https://doi.org/10.1128/iai.00659-16.

38. Chagneau, C.V., Garcie, C., Bossuet-Greif, N., Tronnet, S., Brachmann, A.O., Piel, J., Nougayrède, J.P., Martin, P., and Oswald, E. (2019). The Polyamine Spermidine Modulates the Production of the Bacterial Genotoxin Colibactin. mSphere *4*, e00414-19. https://doi.org/10.1128/mSphere.00414-19.

39. Dougherty, M.W., and Jobin, C. (2021). Shining a Light on Colibactin Biology. Toxins *13*, 346. https://doi.org/10.3390/toxins13050346.

40. Buc, E., Dubois, D., Sauvanet, P., Raisch, J., Delmas, J., Darfeuille-Michaud, A., Pezet, D., and Bonnet, R. (2013). High prevalence of mucosa-associated E. coli producing cyclomodulin and genotoxin in colon cancer. PLoS One *8*, e56964. https://doi.org/10.1371/journal.pone.0056964.

41. Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* (New York: Springer-Verlag). https://ggplot2.tidyverse.org.

42. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinf. *12*, 77. https://doi.org/10.1186/1471-2105-12-77.

43. Manders, F., Brandsma, A.M., de Kanter, J., Verheul, M., Oka, R., van Roosmalen, M.J., van der Roest, B., van Hoeck, A., Cuppen, E., and van Boxtel, R. (2022). MutationalPatterns: the one stop shop for the analysis of mutational processes. BMC Genom. *23*, 134. https://doi.org/10.1186/s12864-022-08357-3.

44. Sato, T., Stange, D.E., Ferrante, M., Vries, R.G.J., Van Es, J.H., Van Den Brink, S., Van Houdt, W.J., Pronk, A., Van Gorp, J., Siersema, P.D., and Clevers, H. (2011). Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. Gastroenterology *141*, 1762–1772. https://doi.org/10.1053/j.gastro.2011.07.050.

45. Bartfeld, S., Bayram, T., van de Wetering, M., Huch, M., Begthel, H., Kujala, P., Vries, R., Peters, P.J., and Clevers, H. (2015). In vitro expansion of human gastric epithelial stem cells and their responses to bacterial infection. Gastroenterology *148*, 126–136.e6. https://doi.org/10.1053/j.gastro.2014.09.042.

46. Osorio, F.G., Rosendahl Huber, A., Oka, R., Verheul, M., Patel, S.H., Hasaart, K., de la Fonteijne, L., Varela, I., Camargo, F.D., and van Boxtel, R. (2018). Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. Cell Rep. *25*, 2308–2316.e4. https://doi.org/10.1016/j.celrep.2018.11.014.

47. Alexandrov, L.B., Jones, P.H., Wedge, D.C., Sale, J.E., Campbell, P.J., Nik-Zainal, S., and Stratton, M.R. (2015). Clock-like mutational processes in human somatic cells. Nat. Genet. *47*, 1402–1407. https://doi.org/10.1038/ng.3441.

48. Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of gastric adenocarcinoma. Nature *513*, 202–209. https://doi.org/10.1038/nature13480.

49. Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330–337. https://doi.org/10.1038/nature11252.

50. Cancer Genome Atlas Research Network (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature *511*, 543–550. https://doi.org/10.1038/nature13385.

51. Cancer Genome Atlas Research Network (2012). Comprehensive genomic characterization of squamous cell lung cancers. Nature *489*, 519–525. https://doi.org/10.1038/nature11404.

52. Cancer Genome Atlas Network (2015). Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature *517*, 576–582. https://doi.org/10.1038/nature14129.

53. Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C., and Lichtenstein, L. (2019). Calling Somatic SNVs and Indels with Mutect2. Preprint at bioRxiv. https://doi.org/10.1101/861054.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
anti-γH2AX; clone JBW301	Millipore	#05-636; RRID:AB_309864
goat anti-mouse AF-647	Thermo Fisher	#A-21235; RRID:AB_2535804
Bacterial and virus strains		
Escherichia coli strain CCR	Guillaume Dalmasso and Richard Bonnet; Cougnoux et al.[18]	EcC
Escherichia coli strain CFF16-2F8	Guillaume Dalmasso and Richard Bonnet; Buc et al.[40]	2F8
Escherichia coli strain CFF159-19H2	Guillaume Dalmasso and Richard Bonnet; Buc et al. 5[40]	19H2
Escherichia coli Nissle 1917	Sridhar Mani	EcN
Escherichia coli strain CCR*ΔclbQ*	Guillaume Dalmasso and Richard Bonnet; Cougnoux et al.[18]	EcC*ΔclbQ*
Chemicals, peptides, and recombinant proteins		
Cultrex Basement Membrane Extract (BME), Growth Factor Reduced, Type 2	R&D Systems, Bio-Techne	3533-001-02
Advanced DMEM/F12	Thermo Fisher scientific	12634-010
B-27 Supplement	Thermo Fisher scientific	17504044
GlutaMAX	Thermo Fisher scientific	35050061
HEPES	Thermo Fisher scientific	15630080
Penicillin-Streptomycin	Thermo Fisher scientific	15140122
Wnt NGS	U-Protein Express	Custom order
Noggin conditioned medium	U-Protein Express	Custom order
R-spondin conditioned medium	U-Protein Express	Custom order
N-Acetyl-L-cysteine	Sigma-Aldrich	A9165
Nicotinamide	Sigma-Aldrich	N0636
Human EGF	Peprotech	AF-100-15
A83-01	Tocris	2939
Prostaglandin E2	Tocris	2296
A83-01	Tocris	2939
SB 202190	Sigma-Aldrich	S7076
Y-27632 dihydrochloride	Abmole	M1817
Primocin	Invivogen	ant-pm-2
DAPI	Thermo Fisher scientific	D1306
Formaldehyde solution 4%	Sigma-Aldrich	1.00496
Gentamicin	Sigma-Aldrich	G1397
Fast Green FCF	Sigma-Aldrich	F7252
TrypIE	Thermo Fisher Scientific	12605010
Critical commercial assays		
DNeasy Blood & Tissue kit	Qiagen	Cat#69504
TruSeq Nano kit	Illumina	Cat#20015964
ResolveDNA Whole Genome Amplification Kit	BioSkryb Genomics	100545
Deposited data		
WGS of *pks*$^{+}$ *E. coli* exposed organoids	This paper, deposited in EGA	EGA:EGAD00001005416EGA: EGAD00001008687, EGA:EGAD50000000304
Experimental models: Cell lines		
Human intestinal organoid line ASC-5a	Utrecht Medical Center; Blokzijl et al.[31]	ASC-5a

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
GraphPad PRISM 9	GraphPad	NA
Las X	Leica	NA
Adobe illustrator	Adobe inc.	NA
Fiji	NIH, Fiji developers	https://imagej.net/Fiji
NF-IAP v1.2	University Medical Center Utrecht	https://github.com/UMCUGenetics/NF-IAP
PTATO	Middelkamp et al.[20]	https://github.com/ToolsVanBox/PTATO
Somatic Mutations Rechecker and Filtering	This paper	https://github.com/ToolsVanBox/SMuRF
Ggplot 3.4.4	ggplot2[41] (https://doi.org/10.1007/978-0-387-98141-3)	https://ggplot2.tidyverse.org/
ggpubr v0.6.0		https://rpkgs.datanovia.com/ggpubr/
pROC v.1.15.5	Robin et al.[42]	https://xrobin.github.io/pROC/
randomForest	https://www.stat.berkeley.edu/~breiman/RandomForests/	https://cran.r-project.org/web/packages/randomForest
MutationalPatterns v3.8.1	Manders et al.[43]	https://bioconductor.org/packages/release/bioc/html/MutationalPatterns.html
Code to analyze the data generated in this study and create the figures. Repository also includes processed mutation data files from WGS organoid sequencing	This paper	https://github.com/ProjectsVanBox/colibactin_detection
Other		
SP8 confocal microscope	Leica	NA
Novaseq 6000	Illumina	NA
FemtoJet 4i microinjector	Eppendorf	cat. no. 5252 000.013
Thin-wall glass capillaries	World Precision Instruments	cat. no. TW100F-4
Ultrospec 10 Cell Density Meter	Biochrom	cat. no. 80-2116-30

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Ruben van Boxtel (R.vanBoxtel@prinsesmaximacentrum.nl).

Materials availability

The biological reagents used in this study are available from the lead contact upon request.

Restrictions to sharing human organoid lines apply due to ethical regulations.

Data and code availability

Raw sequencing reads are deposited at the European Genome-Phenome Archive (EGA), under the dataset accession numbers (EGA: EGAD00001005416, EGAD50000000304, EGAD00001008687). The whole genome sequencing data from the screening cohort is available under EGA: EGAD00001004192.

Filtered.vcf files containing somatic mutations acquired during culture and R scripts used to perform all analyses can be retrieved from:

https://github.com/ProjectsVanBox/colibactin_detection.

This study made use of somatic SBS and indel mutations obtained from the whole-genome sequencing from the Hartwig Medical Foundation (HMF). More information how to obtain access to the HMF data can be retrieved from: https://www.hartwigmedicalfoundation.nl/en/data/data-access-request/.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Clonal human intestinal organoid line ASC-5a was established previously.[31] Ethical approval was obtained from the ethics committee of the University Medical Center Utrecht. Written informed consent was obtained from the tissue donor. All experiments and analyses were performed in compliance with the applicable ethical regulations.

The CRC-derived strain EcC and isogenic EcCΔ*clbQ* were isolated and described in Cougnoux et al., 2014.[15] CRC-derived strains 2F8 (CFF16-2F8) and 19H2 (CFF159-19H2) were isolated and described in Buc et al., 2013.[17] EcN was shared by the laboratory of Sridhar Mani.

METHOD DETAILS

Organoid culture

Organoid experiments were performed as previously described[44] based on the protocols described in Pleguezuelos-Manzano et al., 2020.[8,16] The clonal wild-type human intestinal organoid line ASC-5a (described in Blokzijl et al., 2016[31]) was cultured in 10 μl domes of Cultrex Pathclear Reduced Growth Factor Basement Membrane Extract (BME) (3533-001, Amsbio) submerged in a growth medium consisting of Advanced DMEM/F12 (Gibco), 1× B27, 1× glutamax, 10 mmol/l HEPES, 100 U/ml penicillin-streptomycin (all Thermo Fisher), 1.25 mM N-acetylcysteine, 10 μM nicotinamide, 10 μM p38 inhibitor SB202190 (all Sigma-Aldrich) and the following growth factors: 0.5 nM Wnt surrogate-Fc fusion protein, 2% noggin conditioned medium (both U-Protein Express), 20% Rspo1 conditioned medium (in-house), 50 ng/ml EGF (Peprotech), 0.5 μM A83-01, and 1 μM PGE2 (both Tocris). To derive clonal lines, organoids were dissociated to single cells using TrypLE express (Gibco) and subjected to fluorescence-activated cell sorting (FACS). After sorting, cells were seeded at a density of 50 cells per μl in BME. The Rho Kinase inhibitor Y-27632 (10 μM; Abmole, M1817) was added for the first week of growth. Upon reaching a size of >100 μm diameter, organoids were picked and transferred to separate wells of a 48 well plate per organoid. The organoid line identity was regularly confirmed using SNP testing and WGS. Mycoplasma tests were consistently negative throughout the experiments.

pks^{+} *E. coli* strains and co-culture with organoids

The *pks*$^{+}$ *E. coli* bacterial cultures were performed according to previously described protocols.[8,16] Bacteria were cultured in Advanced DMEM (Gibco) supplemented with glutamax and HEPES to an optical density (OD) of 0.4. Luminal microinjection into human intestinal organoids was performed as previously described.[16,45] Bacteria were injected at a multiplicity of infection of 1 together with 0.05% (w/v) FastGreen dye (Sigma) to visualize injected organoids. 5 μg/ml of the non-permeant antibiotic gentamicin was added to the medium right after injection to prevent overgrowth of bacteria outside the organoid lumen. Bacterial growth was determined by harvesting whole wells or picking single-organoid co-cultures, organoid dissociation with 0.5% saponin for 10 min and re-plating of serial dilutions on LB plates. Colony forming units (CFUs) were counted after a 16 h culture at 37°C. For long-term co-cultures, the bacteria were killed with 1× Primocin (InvivoGen) after 3 days (for clonal WGS) or 4 days (for PTA WGS), after which organoids were kept in culture to recover for 4 days before being passaged. Upon reaching a cystic organoid phenotype again (typically after 2–3 weeks), the injection cycle was repeated. This procedure was repeated 3 (for clonal WGS) or 2 times (for PTA WGS) to enable accumulation of mutations and ensure an even exposure of most cells.

DNA damage quantification

Organoids co-cultured with EcN, EcC or EcCΔ*clbQ* (as described in Cougnoux et al., 2014[18]) were collected in cell recovery solution (Corning) and incubated at 4°C for 30min under gentle rocking in order to remove attached BME from the organoids. The samples were fixed in 4% formalin for 16 h at 4°C. Organoids were permeabilized with 0.5% Triton-X (Sigma), 2% donkey serum (BioRad) in PBS for 30min at 4°C and blocked with 0.1% Tween-20 (Sigma) and 2% donkey serum in PBS for 15min at room temperature. Subsequently, the samples were incubated with primary mouse anti-γH2AX antibody (Millipore; clone JBW301; 1:1,000 dilution) for 16h at 4°C. Then, organoids were washed four times with PBS and incubated with secondary goat anti-mouse AF-647 antibody (Thermo Fisher, catalogue number A-21235, 1:500 dilution) for 3h at room temperature under the exclusion of light and washed again with PBS. The samples were imaged on an SP8 confocal microscope (Leica). Fluorescent microscopic images of γH2AX foci were quantified by classifying each nucleus as having either no foci or one or more foci. The fraction of nuclei containing foci divided by the sum of all nuclei is displayed as one datapoint per organoid. Statistical significance was evaluated using Prism GraphPad software version 8.4.3 (686). Wilcoxon test was performed to obtain p-values. FDR correction was applied using the R function p.adjust() with the parameter method set to "BH".

DNA isolation and sequencing

Genomic DNA was isolated from organoid pellets using the Qiagen DNeasy Blood & Tissue kit. DNA was eluted in 50 μL Low EDTA (10 mM Tris base, 0.1 mM EDTA). DNA sequencing libraries were made with a TruSeq Nano kit (Illumina) from 50 ng of genomic DNA using manufacturers' instructions. For the samples sequenced by PTA, PTA was performed using the ResolveDNA Whole Genome Amplification Kit (BioSkryb Genomics) according to the manufacturer's protocol. Instead of 10 minutes cell lysis on ice as indicated in this protocol, lysis was performed at room temperature for 20 minutes on a shaker. These libraries were sequenced at a depth of 15x or 30x using a Novaseq 6000 at the Hartwig Medical foundation (www.hartwigsequencingservices.nl).

Mapping and variant calling

Aligned sequencing data from previously sequenced organoids co-cultured with CCR, CCRΔ*clbQ*, Δ*clbQ*, CCR:Δ*clbQ*:*clbQ* and injection dye were included in the analysis (Table S1), and all analyses were performed starting from the FASTQ raw sequencing data.[8] Clones were sequenced at 30x base coverage using an Illumina Novaseq 6000, except for the clones exposed for a single injection

round. These clonal lines, and the parental clonal line were sequenced at 30x using an Illumina Hiseq X10 sequencing machine. Sequencing reads from all samples were mapped to the human reference GRCh38 genome using the Burrows-Wheeler Aligner v0.7.17 "BWA-MEM -c 100 -M". Duplicate sequencing reads were marked using Sambamba MarkDup v0.6.8. A full description and source code for the NF-IAP version 1.2 pipeline can be retrieved from: https://github.com/UMCUGenetics/NF-IAP.

Variants in the mapped data were called using GATK Haplotypecaller version 4.1.3.0 using default settings. Variants were filtered using GATK 4.1.3.0 using the following filter settings for SBS: –filter-expression 'QD < 2.0' –filter-expression 'MQ < 40.0' –filter-expression 'FS > 60.0' –filter-expression 'HaplotypeScore > 13.0' –filter-expression 'MQRankSum < -12.5' –filter-expression 'ReadPosRankSum < -8.0' –filter-expression 'MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1)' –filter-expression 'DP < 5' –filter-expression 'QUAL < 30' –filter-expression 'QUAL >= 30.0 && QUAL < 50.0' –filter-expression 'SOR > 4.0' –filter-name 'SNP_LowQualityDepth' –filter-name 'SNP_MappingQuality' –filter-name 'SNP_StrandBias' –filter-name 'SNP_HaplotypeScoreHigh' –filter-name 'SNP_MQRankSumLow' –filter-name 'SNP_ReadPosRankSumLow' –filter-name 'SNP_HardToValidate' –filter-name 'SNP_LowCoverage' –filter-name 'SNP_VeryLowQual' –filter-name 'SNP_LowQual' –filter-name 'SNP_SOR' -cluster 3 -window 10". The following settings were used to filter all other variants: filter_criteria = "–filter-expression 'QD < 2.0' –filter-expression 'ReadPosRankSum < -20.0' –filter-expression 'FS > 200.0' –filter-name 'INDEL_LowQualityDepth' –filter-name 'INDEL_ReadPosRankSumLow' –filter-name 'INDEL_StrandBias'".

For PTA samples, artefacts were filtered out using the PTATO pipeline.[20] This pipeline uses a random-forest based filtering approach to remove recurrent artifacts from the sequencing data.

Variant filtering

To filter out mutations induced during sequencing, clonal expansion or library preparation, we filtered genomic variants using an in-house filtering pipeline, SMuRF v2.1.1 (https://github.com/ToolsVanBox/SMuRF. Briefly, the variant allele frequency (VAF) was calculated for each variant by pileup of all bases mapped at the mutation position. Variant data derived from organoid clones sequenced at 30x depth were filtered for the following criteria: VAF $\geq$ 0.3, base coverage $\geq$ 10 and an MQ quality $\geq$ 60. For organoid clones sequenced at 15x depth two deviations from the filter settings were introduced: VAF $\geq$ 0.15, base coverage $\geq$ 5. To select only mutations occurring during *in-vitro* culture, variants present in the clonal parental organoid line were removed. Recurrent mapping or sequencing artifacts were removed by filtering against a blacklist containing variants present in healthy bone marrow mesenchymal stromal cells.[46]

Mutational signature analysis

The resulting filtered variants were analyzed using the R package MutationalPatterns v3.8.1[43] to read vcf files, annotate mutations and generate 96-trinucleotide and indel plots (https://bioconductor.org/packages/release/bioc/html/MutationalPatterns.html). In brief, mutations were categorized in 96-trinucleotide categories for SBS mutations and 83 categories for indel mutations. To compare profiles against COSMIC mutational signatures, version 3.1, the cosine similarity measure was used.[10] Mutational signatures for the HMF cohort were extracted in the same manner as in Pleguezuelos-Manzano et al., 2020.[8] For refitting of EcN clones to SBS and ID mutational profiles, we used the colibactin-induced (SBS88 and ID18) mutational signatures, COSMIC version v3.2, (https://cancer.sanger.ac.uk/signatures/). In the re-fitting, we included aging clock-like (SBS1 and 5 for SBSs and ID1 and 2 for IDs)[47] and cell-culture induced signatures (SBS18) active in organoids during cell culture.[31]

Extended context selection and enrichment testing

For all analyses, only unique T>N SBS mutations occurring in each exposed culture condition were considered. For PTA-expanded organoid cells, all conditions (Control, EcC, EcN, 2F8, 19H2, all conditions n = 5), all data was compared to control. For the clonally expanded organoids, a negative population consisting of WGS from 6 samples exposed to dye and 6 samples exposed to EcCΔ*clbQ* strain no longer capable of producing colibactin were used. This was compared to 9 samples exposed to EcC and three samples exposed to EcN. To determine the presence of extended context patterns of *pks* profiles, we selected all bases present in the 10-bp extended context of *pks* mutations occurring at a frequency of 45% or higher.[1] We tested the enrichment of all possible motifs using a one-tailed Fisher's exact test. and compared the enrichment of mutations in the motif against all other mutations present in control. The most significant position, AA at -3 and -4 was selected for further analysis. To compare enrichments, enrichment for all dinucleotide occurrences was tested using a one-tailed Fishers' exact test, using 'greater' as an alternative hypothesis. To select the optimal trinucleotide (Figure 2F), we tested AA enrichment within selected *pks* trinucleotides, selecting only mutations with the most frequent trinucleotide in the SBS88 signature. We tested against the control mutations using a one-tailed Fisher's exact test, and stepwise added the next most occurring trinucleotide for all T>N mutations. The most specific *pks* trinucleotide combination was determined as the combination of trinucleotides which exhibited the lowest *p*-value using a one-tailed Fishers' test, testing for enrichment for both PTA and Clonal expansion conditions. The lowest *p*-values (Figures 2F and S3F) were obtained when selecting the 17 trinucleotides with the highest contribution to SBS88: A[T>C]T, A[T>C]A, T[T>G]T, T[T>C]T, T[T>A]T, A[T>A]A, A[T>C]C, A[T>A]T, T[T>G]A, A[T>A]C, T[T>G]C, T[T>A]A, G[T>A]T, G[T>G]T, G[T>C]T, A[T>G]A, A[T>G]T.

Estimation of relative mutagenicity of *E. coli strains* with respect to EcC

We generated synthetic mixtures of control and EcC mutations by sampling mutations from all unique mutations present in control (control dye and EcCΔ*clbQ*-exposed) and EcC-exposed organoids. For each concentration, increasing in steps of 1% EcC-content,

we generated mixtures containing 0 to 100% EcC mutations. The same total number of mutations) as present in EcC-exposed organoids was sampled for each replicate with replacement. The '-3-4AA' fraction of mutations was determined by taking the extended context of mutations at positions -3 and -4 at the previously defined 17 *pks* motif trinucleotides. Enrichment for these motifs was tested against all control mutations using a one-sided Fisher's exact test. To determine the relative mutagenicity of EcN, 2F8 and 19H2, we modeled a linear model across the enrichment-odds-ratios obtained from the sampled mixtures, and predicted relative mutagenicity rates using the lower, estimated and higher confidence intervals for all three strains.

Monte-Carlo re-sampling analyses

For each condition for both PTA and Clonally expanded organoids, we re-sampled mutations (with replacement), and sampled the same number of mutations as were present in that condition. For each of the resampled conditions, the relative enrichment of mutations with –3-4AA motif, and the cosine similarity of T>N mutations to the T>N fraction of SBS88 were compared. Additionally, we calculated for each of the resampled conditions the enrichment relative to the observed control population using a one-sided Fishers exact test. Simulated data are indicated as grey dots in Figures 2 and S3.

Pks island sequence analysis

To assess the *pks* island sequences of all 4 strains used in this study for differences which may explain the divergent mutagenic activity, we performed whole genome sequencing on DNA derived from liquid cultures of EcN, EcC, 19H2 and 2F8 using Illumina Paired End Sequencing. We assessed the quality of our whole genome sequencing reads with fastqc. These reads were assembled into *de novo* assemblies with SPAdes. We then assessed the quality of these *de novo* assemblies with QUAST. The genome of the *pks*$^+$ IHE3034 strain was downloaded from NCBI. We extracted the sequence of the *pks* pathogenicity island using the beginning of *clbS* (position 2193827) and the end of *clbA* (position 2244594) as the start and end positions of the whole island, respectively. (Thus, our reference *pks* sequence includes regions between the *clb* genes.) We then used MUMmer to align the *de novo* assemblies of our four samples to this reference *pks* sequence and to call single nucleotide polymorphisms (SNPs). Finally, we predicted the effect of these SNPs on the amino acid sequence based on codon changes, allowing us to identify the missense mutations.

Assessment of extended context motifs enrichment in a cohort of metastatic cancer samples

Ethical approval was obtained from the ethics committee of the HMF. Written informed consent was obtained from all patients. All experiments and analyses were performed in compliance with the applicable ethical regulations.

Three cancer genomes containing < 100 somatic SBS mutations were removed from all subsequent analyses from the HMF data. Trinucleotide counts for all mutations in the HMF dataset were determined as in Pleguezuelos-Manzano et al.[8] The occurrence of -3-4AA was determined for all 17 *pks* trinucleotides across the cohort. Enrichment of mutations with 'AA' at the -3 and -4 position was determined compared against AA and other dinucleotide presence in all other samples of the HMF cohort using a one-sided Fisher's exact test with fdr correction. Colibactin motif enrichment per sample was defined as having a $p < 0.001$ (Fisher's exact test, one-sided) and a -3-4AA fraction higher than 0.22. Nervous system and bone/soft tissue samples were considered to be unlikely to be exposed to *pks*$^+$ bacteria prior to carcinogenesis and used as a negative population to set thresholds. For analyses of *POLE*-hypermutated samples in the HMF cohort, somatic mutations were checked to contain any of the 21 mutations in the *POLE* hotspot mutations.[24]

Simulation of whole-exome (WES) data from the HMF cohort

To simulate WES data, exonic sites were considered when reported as exonic region in Ensembl v75 (GCRh37) and coordinates were converted to GCRh38 genome using UCSC liftOver. 36 cancer genomes containing no exonic SBS or no exonic indel mutations were removed from the dataset. Mutational signature re-fitting and calculation of -3-4AA fraction were performed similarly as for the whole-genome mutations in the HMF-cohort. Receiver-operator curves (ROCs) and Area under the curve (AUCs) were determined using the R-package 'pROC', using the WGS motif classification as true positive and negative values.

TCGA analysis

Whole exome mutation calls from TCGA cohorts STAD (stomach adenocarcinoma),[48] COAD (colon adenocarcinoma),[49] READ (rectal adenocarcinoma),[49] LUAD (lung adenocarcinoma),[50] LUSC (lung squamous cell carcinoma)[51] and HNSC (head and neck squamous cell carcinoma)[52] were downloaded from the National Cancer Institute GDC data portal https://portal.gdc.cancer.gov/.

MuTect2[53] output was chosen from the various called mutation files available as the filters used in the TCGA MuTect2 pipeline most closely matched those from the HMF cohort, namely ''alt_allele_in_normal'' (Evidence seen in the normal sample), ''bPcr'' (variant allele shows a bias towards one PCR template), ''bSeq'', (Variant allele shows a bias towards one sequencing strand), ''clustered_events'' (Clustered events observed in the tumor), ''germline_risk'', (Evidence indicates this site is germline, not somatic), ''homologous_mapping_event'', (More than three events were observed in the tumor), ''multi_event_alt_allele_in_normal'', (Multiple events observed in tumor and normal), ''oxog'' (Failed dToxoG), ''panel_of_normals'' (Seen in at least 2 samples in the panel of normal), ''str_contraction'' (Site filtered due to contraction of short tandem repeat region), ''t_lod_fstar'' (Tumor does not meet likelihood threshold), ''triallelic_site'' (Site filtered because more than two alt alleles pass tumor LOD). Additionally, any SNV calls less than 2bp from another call were removed.The same mutational analyses were then performed as described for the HMF WES data.

Random forest model

We first trained a random forest model with the R package randomForest[53] on the organoid data, using only the T>N mutations since the mutational signature is dominated by these. All T>N mutations from organoids injected with colibactin-producing EcC are labeled as positive, all T>N mutation from the knockout strain as negative. As features we include the 10 bases up- and downstream from the mutation, the replication timing, transcriptional strand bias, distance to closest simple repeat and distance to closest gene body. We train 1000 trees with an mtry of 3 to prevent overtraining.

When we test this model on patient data from patients with a clear SBS88 signature and associated blood (expected to be *pks*-negative), we find many false positives. We therefore train a second model on WGS from patients with or without a clear signature. As training data, we use all mutations with a posterior probability > 0.8 from the positive samples (as per the SBS88 signature) as positive set and all mutations with a posterior probability > 0.5 from the negative samples as negative set. Thus, our second model is specifically trained on recognizing the false positives from the first model as negative. The final posterior probability for new mutations is calculated by multiplying the posterior probability of both models.

Threshold for *pks* status of sample

The threshold to call a sample as *pks*-positive is set at 10% of all T > N mutations being colibactin-induced. This threshold ensures all blood samples are negative and all patient samples with clear SBS88 contributions are called positive. Since we have no ground truth, a true optimization of the threshold cannot be performed. We confirm this threshold by comparing the fraction *pks*-positive to the relative contribution of SBS88 and find no correlation below 10%. The probability scores in WES data are on average higher, such that the equivalent of the 0.1 threshold we used in WGS data is 0.1448 in WES data.

Investigating driver genes

We select all driver genes occurring in more than 5% of the samples in IntOGen and which occur in at least 5 samples in our dataset. We then compute the mean probability of all mutations in these genes separately in the *pks*-positive and *pks*-negative patients, with no selection on consequence of these mutations. We then sample random non-driver genes that are mutated in less than 5% of the samples in both classes. We compute the mean difference in probability between the mutations in driver genes and non-driver genes. By doing this separately in the *pks*-positive and *pks*-negative class we ensure that any difference is not just driven by bias in gene length or motif occurrence in a particular gene. We repeat this sampling 1000 times to ensure a robust comparison. We then compute a p-value between the difference in the negative and the positive class with Student's t-test.

QUANTIFICATION AND STATISTICAL ANALYSIS

Significance of γH2AX damage was tested with FDR-adjusted *p*-values Wilcoxon test. N indicates number of organoids. Mean and SEM are shown. Differences in indel and SBS mutation load in exposed organoids were analyzed using the Kruskal-Wallis test, using Dunn's post-hoc correction using base R (version 4.2.2) and the "ggpubr" package (version 0.6.0) using FDR correction. Monte-Carlo subsampling methods were implemented manually in R by subsampling from unique mutations for each exposure category, using the "fisher.test" function from the base R package, using a one-tailed distribution. Similarities between 96-trinucleotide profiles were measured using either cosine similarity, making use of the "cos_sim" function from the package "MutationalPatterns" (version 3.8.1). Spearman correlations were calculated using the base R function "cor" using the 'method = "spearman"' option. *P*-values resulting from the spearman correlation test were corrected using FDR correction. Pearson correlations were calculated with the "cor" function using 'method = "pearson"'. Differences in RF probability between driver genes were assessed with a Student's T-test, using the stats R function "t.test". P-values resulting from this test were Bonferroni corrected for multiple testing. Significance of the enrichment of the APC hotspot mutation was assessed with a one-sided Fisher's exact test, using the function "fisher.test". The difference in age at diagnosis between patients was assessed with a Wilcoxon test using the "wilcox.test" function.