Article

Mutational signature in colorectal cancer caused by genotoxic *pks*$^+$ *E. coli*

https://doi.org/10.1038/s41586-020-2080-8

Received: 14 September 2019

Accepted: 17 February 2020

Published online: 27 February 2020

Check for updates

Cayetano Pleguezuelos-Manzano[1,2,21], Jens Puschhof[1,2,21], Axel Rosendahl Huber[2,3,21], Arne van Hoeck[2,4], Henry M. Wood[5], Jason Nomburg[6,7,8], Carino Gurjao[7,8], Freek Manders[2,3], Guillaume Dalmasso[9], Paul B. Stege[10], Fernanda L. Paganelli[10], Maarten H. Geurts[1,2], Joep Beumer[1], Tomohiro Mizutani[1,2], Yi Miao[11,12,13], Reinier van der Linden[1], Stefan van der Elst[1], Genomics England Research Consortium*, K. Christopher Garcia[11,12,13], Janetta Top[10], Rob J. L. Willems[10], Marios Giannakis[7,8], Richard Bonnet[9,14], Phil Quirke[5], Matthew Meyerson[7,8,15,16], Edwin Cuppen[2,4,17,18], Ruben van Boxtel[2,3]✉ & Hans Clevers[1,2,3]✉

Various species of the intestinal microbiota have been associated with the development of colorectal cancer[1,2], but it has not been demonstrated that bacteria have a direct role in the occurrence of oncogenic mutations. *Escherichia coli* can carry the pathogenicity island *pks*, which encodes a set of enzymes that synthesize colibactin[3]. This compound is believed to alkylate DNA on adenine residues[4,5] and induces double-strand breaks in cultured cells[3]. Here we expose human intestinal organoids to genotoxic *pks*$^+$ *E. coli* by repeated luminal injection over five months. Whole-genome sequencing of clonal organoids before and after this exposure revealed a distinct mutational signature that was absent from organoids injected with isogenic *pks*-mutant bacteria. The same mutational signature was detected in a subset of 5,876 human cancer genomes from two independent cohorts, predominantly in colorectal cancer. Our study describes a distinct mutational signature in colorectal cancer and implies that the underlying mutational process results directly from past exposure to bacteria carrying the colibactin-producing *pks* pathogenicity island.

The intestinal microbiome has long been suggested to be involved in colorectal cancer (CRC) tumorigenesis[1,2]. Various bacterial species have been reported to be enriched in stool and biopsy samples from patients with CRC[6–9], including genotoxic strains of *E. coli*[3,6,10,11]. The genome of these genotoxic *E. coli* harbours a 50-kb hybrid polyketide-nonribosomal peptide synthase operon (*pks*, also referred to as *clb*) that is responsible for the production of the genotoxin colibactin. *pks*$^+$ *E. coli* are present in a substantial fraction of individuals (about 20% of healthy individuals, about 40% of patients with inflammatory bowel disease, and about 60% of patients with familial adenomatous polyposis or CRC)[6,10,11]. *pks*$^+$ *E. coli* induce—among other things—interstrand crosslinks (ICLs) and double-strand breaks (DSBs) in epithelial cell lines[3,10–12] and in gnotobiotic mouse models of CRC, in which they can also contribute to tumorigenesis[6,11]. Recently, two studies have reported that colibactin–adenine adducts are formed in mammalian cells exposed to *pks*$^+$ *E. coli*[4,5]. Whereas the chemistry of the interaction between colibactin and DNA is thus well-established, the outcome of this process in terms of recognizable mutations remains to be determined. Recent advances in sequencing technologies and the application of novel mathematical approaches allow somatic mutational patterns to be classified. More than 50 mutational signatures have been defined in using a mutational signature analysis that includes the bases immediately 5′ and 3′ to a single-base substitution (SBS), and a number of different contexts that characterize insertions and deletions (indels)[13,14]. For some of these mutational signatures, the underlying causes (for example, tobacco smoke, UV light, specific genetic DNA repair defects) are known[13,15,16]. However, for many the underlying aetiology remains unclear. Human intestinal organoids, which are established from primary crypt stem cells[17], have been useful for identifying the underlying causes of mutational signatures[18]: After being exposed to a specific mutational agent in culture, the organoids can be subcloned and analysed by whole-genome sequencing (WGS) to identify the resulting mutational signature[16,19,20].

To define the mutagenic characteristics of *pks*$^+$ *E. coli*, we developed a co-culture protocol in which a *pks*$^+$ *E. coli* strain (originally derived from a CRC biopsy[21]) was microinjected into the lumen of clonal human

[1]Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences (KNAW) and UMC Utrecht, Utrecht, The Netherlands. [2]Oncode Institute, Utrecht, The Netherlands. [3]The Princess Máxima Center for Pediatric Oncology, Utrecht, The Netherlands. [4]Center for Molecular Medicine, University Medical Centre Utrecht, Utrecht, The Netherlands. [5]Pathology and Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK. [6]Graduate Program in Virology, Division of Medical Sciences, Harvard Medical School, Boston, MA, USA. [7]Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA, USA. [8]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [9]University Clermont Auvergne, Inserm U1071, INRA USC2018, M2iSH, Clermont-Ferrand, France. [10]Department of Medical Microbiology, University Medical Center Utrecht, Utrecht, The Netherlands. [11]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA, USA. [12]Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, CA, USA. [13]Department of Structural Biology, Stanford University School of Medicine, Stanford, CA, USA. [14]Department of Bacteriology, University Hospital of Clermont-Ferrand, Clermont-Ferrand, France. [15]Department of Genetics, Harvard Medical School, Boston, MA, USA. [16]Department of Medicine, Harvard Medical School, Boston, MA, USA. [17]Hartwig Medical Foundation, Amsterdam, The Netherlands. [18]CPCT Consortium, Rotterdam, The Netherlands. [21]These authors contributed equally: Cayetano Pleguezuelos-Manzano, Jens Puschhof, Axel Rosendahl Huber.
*A list of authors and their affiliations appears at the end of the paper. ✉e-mail: R.vanBoxtel@prinsesmaximacentrum.nl; h.clevers@hubrecht.eu

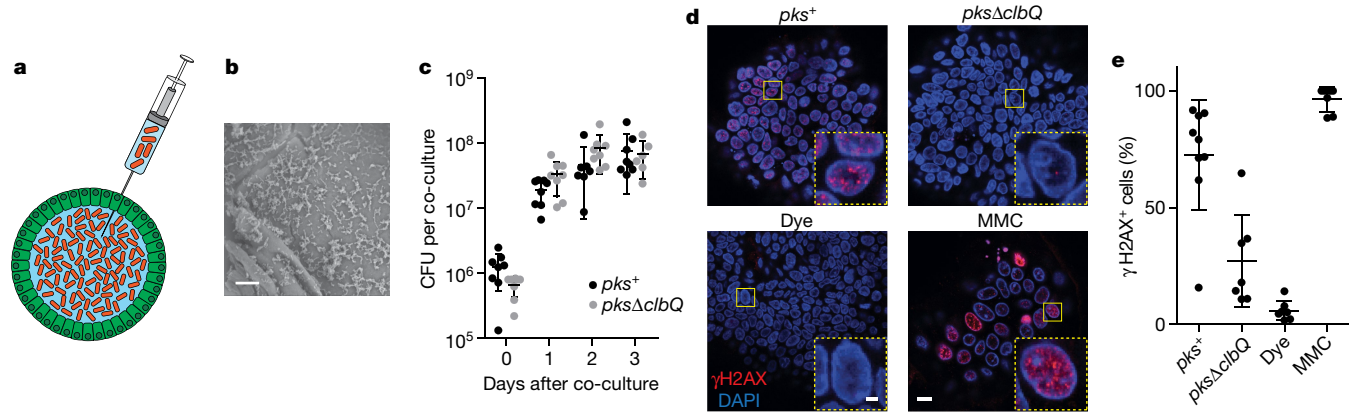


Fig. 1 | Co-culture of healthy human intestinal organoids with genotoxic *E. coli* induces DNA damage. a, Schematic representation of microinjection of genotoxic *E. coli* into the lumen of human intestinal organoids. **b**, Scanning electron microscopy image illustrating direct contact between organoid apical side and *pks*$^+$ *E. coli* after 24 h co-culture. Scale bar, 10 µm. **c**, Mean ± s.d. bacterial load of *pks*$^+$ or *pks*Δ*clbQ* at 0, 1, 2 and 3 days after co-culture establishment ($n = 8$ co-cultures per condition and time point, except *pks*$^+$ day 2 ($n = 7$) and *pks*Δ*clbQ* day 3 ($n = 6$)). CFU, colony-forming units. **d**, Representative images of DNA damage induction after 1 day of co-culture, measured by γH2AX immunofluorescence. One organoid is shown per image with one nucleus in the inset (expansion of boxed area). Scale bars, 10 µm (main image); 2 µm (inset). MMC (mitomycin C), positive control for double-strand break induction. **e**, Quantification of data from **d**: mean ± s.d. percentage of nuclei positive for γH2AX foci in organoids injected with *pks*$^+$ *E. coli* ($n = 9$ organoids), *pks*Δ*clbQ E. coli* ($n = 7$ organoids), dye ($n = 7$ organoids) or mitomycin C (MMC) ($n = 7$ organoids) after 1 day of co-culture.

intestinal organoids[22] (Fig. 1a, b). An isogenic *clbQ* knockout strain that cannot produce active colibactin[21,23] served as a negative control. Both bacterial strains were viable for at least three days in co-culture and followed similar growth dynamics (Fig. 1c). DSBs and ICLs, visualized by γH2AX and FANCD2 immunofluorescence, respectively, were induced specifically in epithelial cells exposed to *pks*$^+$ *E. coli* (Fig. 1d, e,

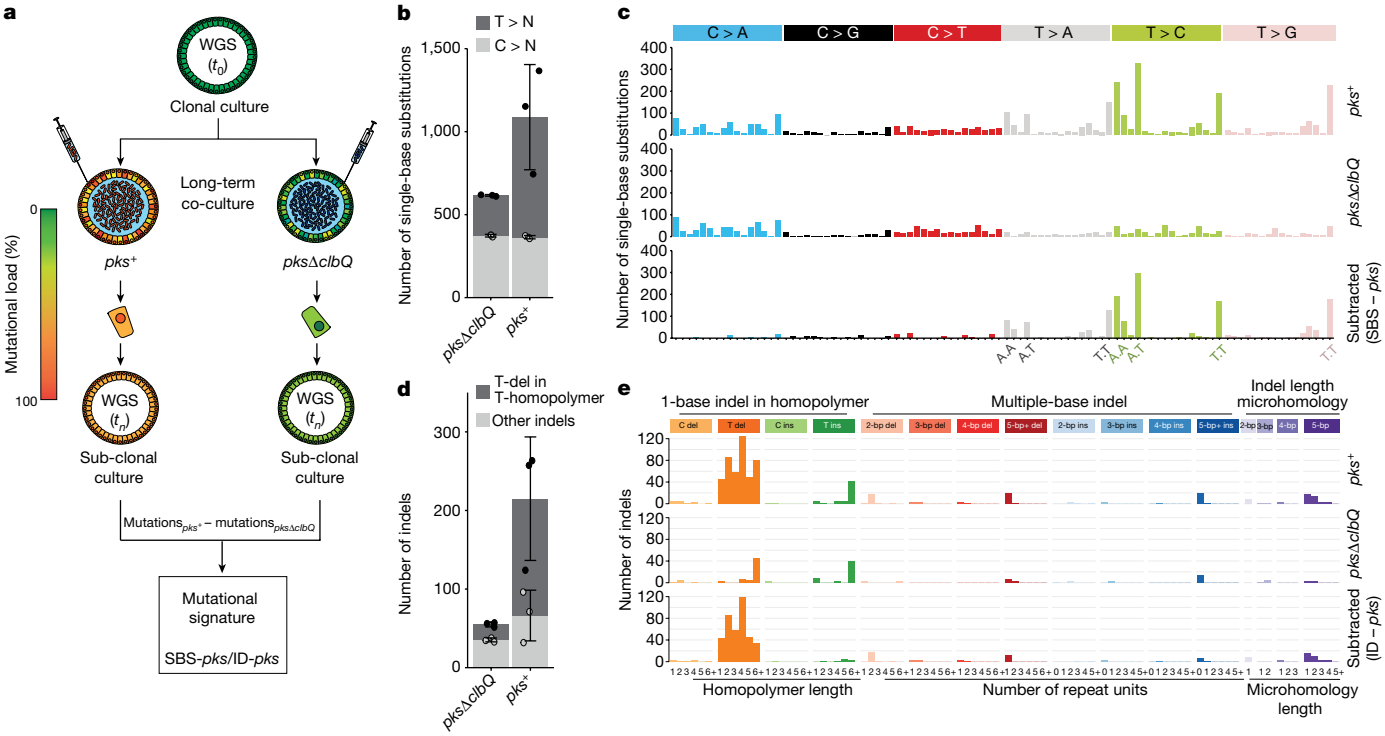


Fig. 2 | Long-term co-culture with *pks*$^+$ *E. coli* induces SBS-*pks* and ID-*pks* mutational signatures. a, Schematic representation of the experimental setup. **b**, Bar segment height indicates the mean ± s.d. number of SBSs that accumulated in organoids co-cultured with either *pks*$^+$ or *pks*Δ*clbQ E. coli* ($n = 3$ clones). Dot position above the bottom of the corresponding bar segment (T > N, black; C > N, grey) indicates the number of mutations for each clone. **c**, SBS 96-trinucleotide mutational spectra in organoids exposed to either *pks*$^+$ (top) or *pks*Δ*clbQ* (middle) *E. coli*. The bottom panel depicts the SBS-*pks* signature, which was defined by subtracting SBS mutations under the *pks*Δ*clbQ* condition from those under the *pks*$^+$ condition. SBSs are indicated above the plot. Most mutated trinucleotide sequences are highlighted below the bottom axis as '5′ base.3′ base', with the dot indicating the position of the substituted nucleotide. **d**, Bar segment height indicates the mean ± s.d. number of indels that accumulated in organoids co-cultured with either *pks*$^+$ or *pks*Δ*clbQ E. coli* ($n = 3$ clones). Dot position above the bottom of the corresponding bar segment (T deletion in T-homopolymer, black; other indels, grey) indicates the number of mutations for each clone. **e**, Indel mutational spectra observed in organoids exposed to either *pks*$^+$ (top) or *pks*Δ*clbQ* (middle) *E. coli*. The bottom panel depicts the ID-*pks* signature, which was defined by subtracting indel mutations under the *pks*Δ*clbQ* condition from those under the *pks*$^+$ condition.

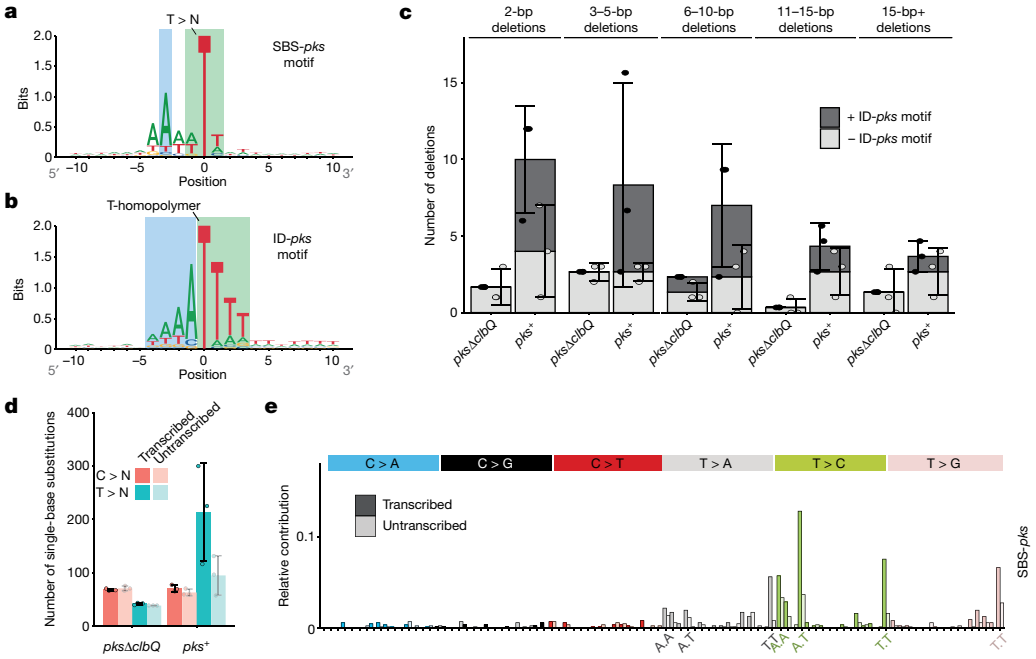


Fig. 3 | Consensus motifs and extended features of SBS-*pks* and ID-*pks* mutational signatures. a, Two-bit representation of the extended sequence context of T > N mutations observed in organoids exposed to *pks*$^+$ *E. coli*. Green, highlighted T > N trinucleotide sequence; blue, highlighted A-enriched position characteristic of the SBS-*pks* mutations. **b**, Two-bit representation of the extended sequence context of single T deletions in T homopolymers observed in organoids exposed to *pks*$^+$ *E. coli*. Green, highlighted T homopolymer with deleted T; blue, highlighted characteristic poly-A stretch. **c**, Bar segment height indicates the mean ± s.d. occurrence of deletions comprising more than 1 bp in organoids exposed to *pks*$^+$ or *pks*Δ*clbQ E. coli* ($n = 3$ clones). Dot position above the bottom of the corresponding bar segment (matching ID-*pks* motif, black; lacking ID-*pks* motif, grey) indicates the number of mutations for each clone. **d**, Transcriptional strand bias of T > N and C > N mutations in organoids exposed to *pks*$^+$ *E. coli* or *pks*Δ*clbQ E. coli*. Pink, C > N; blue, T > N; dark colour, transcribed strand; bright colour, untranscribed strand; mean ± s.d. number of events ($n = 3$ clones). **e**, Transcriptional strand bias of the 96-trinucleotide SBS-*pks* mutational signature. Colour, transcribed strand; white, untranscribed strand.

Extended Data Fig. 1a), confirming that *pks*$^+$ *E. coli* induced DNA damage in our model. There was no substantial difference in viability between organoids exposed to *pks*$^+$ *E. coli* and those exposed to *pks*Δ*clbQ E. coli*, although there was a modest decrease for both when compared to the organoids injected with dye only (Extended Data Fig. 1b, c). We then performed repeated injections (with *pks*$^+$ *E. coli*, *pks*Δ*clbQ E. coli* or dye only) into single cell-derived organoids, in order to achieve long-term exposure over a period of five months. Subsequently, we established subclonal organoids from individual cells extracted from the exposed organoids. For each condition, three subclones were subjected to WGS (Fig. 2a). We also subjected the original clonal cultures to WGS to subtract somatic mutations that were already present before co-culture. Organoids exposed to *pks*$^+$ *E. coli* presented increased numbers of SBSs compared to those exposed to *pks*Δ*clbQ E. coli*, with a bias towards T > N substitutions (Fig. 2b). These T > N substitutions occurred preferentially in ATA, ATT and TTT (with the middle base mutated). From this, we defined a *pks*-specific single-base substitution signature (SBS-*pks*; Fig. 2c). This mutational signature was not observed in organoids exposed to *pks*Δ*clbQ E. coli* or dye only (Fig. 2b, c, Extended Data Fig. 2a–c), proving that it is a direct consequence of exposure to *pks*$^+$ *E. coli*. Furthermore, exposure to *pks*$^+$ *E. coli* induced a characteristic small indel signature (ID-*pks*), which was characterized by single T deletions at T homopolymers (Fig. 2d, e, Extended Data Fig. 2d–f). SBS-*pks* and ID-*pks* were replicated in an independent human intestinal organoid line (Extended Data Fig. 3a–d; SBS cosine similarity, 0.77; ID cosine similarity, 0.93) and with a *clbQ*-knockout *E. coli* strain recomplemented with the *clbQ* locus (*pks*Δ*clbQ:clbQ*) (Extended Data Fig. 3e–h; SBS cosine similarity, 0.95; ID cosine similarity, 0.95).

Next, we investigated whether the SBS-*pks* and ID-*pks* mutations were characterized by other recurrent patterns. First, the assessed DNA stretch was extended beyond the nucleotide triplet. This uncovered the preferred presence of an adenine residue 3 bp upstream of the mutated SBS-*pks* T > N site (Fig. 3a). Similarly, mutations that contributed to the ID-*pks* signature in poly-T stretches showed enrichment of adenines immediately upstream of the affected poly-T stretch (Fig. 3b). Notably, the lengths of the adenine stretch and the T-homopolymer were inversely correlated, consistently resulting in a combined length of five or more A/T nucleotides (Extended Data Fig. 4a). While SBS-*pks* and ID-*pks* are the predominant mutational outcomes of colibactin exposure, we also observed longer deletions at sites containing the ID-*pks* motif in organoids treated with *pks*$^+$ *E. coli* (Fig. 3c). Additionally, the SBS-*pks* signature exhibited a striking transcriptional strand bias (Fig. 3d, e). We speculate that these observations reflect preferential repair of alkylated adenosines on the transcribed strand by transcription-coupled nucleotide excision repair. These features clearly distinguish the *pks* signature from published signatures of alkylating agents or other factors[19].

We then investigated whether the experimentally deduced SBS-*pks* and ID-*pks* signatures occur in human tumours, by interrogating WGS data from a Dutch collection of 3,668 solid cancer metastases[24]. The mutations acquired by a cancer cell at its primary site will be preserved even in metastases, so that these provide a view of the entire mutational history of a tumour. We first performed non-negative matrix factorization (NMF) on genome-wide mutation data obtained from 496 CRC metastases in this collection. Encouragingly, this unbiased approach identified an SBS signature that highly resembled SBS-*pks* (cosine similarity, 0.95; Extended Data Fig. 5a, b). We then determined the contributions of SBS-*pks* and ID-*pks* to the mutations of each sample in the cohort. This analysis revealed that the two *pks* signatures were strongly enriched in CRC-derived metastases when compared to all other cancer types (Fisher's exact test, $P < 0.0001$; Fig. 4a, b, Extended Data Table 1). With a cut-off contribution value of 0.05, 7.5% of CRC samples were

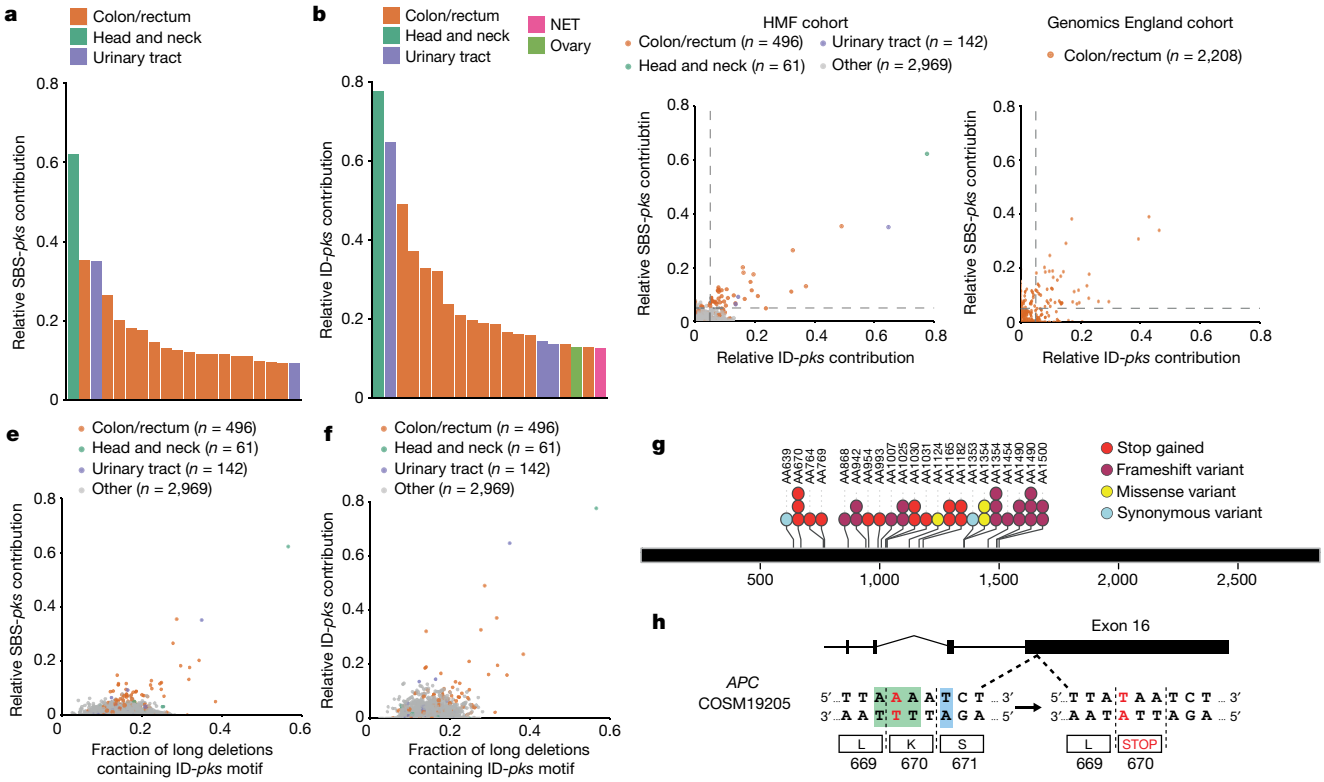


Fig. 4 | SBS-*pks* and ID-*pks* mutational signatures are present in a subset of CRC samples from two independent cohorts. a, Top 20 out of 3,668 metastases from the HMF cohort, ranked by the fraction of SBSs attributed to SBS-*pks*. CRC metastases (orange) are enriched. **b**, Top 20 out of 3,668 metastases from the HMF cohort. Samples are ranked by the fraction of indels attributed to ID-*pks*. CRC metastases (in orange) are also enriched here. NET, neuroendocrine tumour. **c**, Scatterplot of the fraction of SBSs and indels attributed to SBS-*pks* and ID-*pks* in 3,668 metastases fro the HMF cohort. Each dot represents one metastasis. Samples high for both SBS-*pks* and ID-*pks* (more than 5% contribution, dashed lines) are enriched in CRC (orange). SBS-*pks* and ID-*pks* are correlated ($R^2 = 0.46$; only CRC, $R^2 = 0.7$). **d**, Scatterplot of SBS-*pks* and ID-*pks* contributions in 2,208 CRC tumour samples, predominantly of primary origin, from the Genomics England cohort. SBS-*pks* and ID-*pks* are correlated ($R^2 = 0.35$). Each dot represents one primary tumour sample. Dashed lines delimit samples with high SBS-*pks* or ID-*pks* contribution (more than 5%). **e**, Scatterplot of SBS-*pks* and deletions longer than 1 bp with ID-*pks* pattern in the HMF cohort. **f**, Scatterplot of ID-*pks* and deletions longer than 1 bp with ID-*pks* pattern in the HMF cohort. **a**–**f**, Colours indicate tissue of origin. **g**, Exonic *APC* driver mutations found in the IntOGen collection matching the colibactin target SBS-*pks* or ID-*pks* motifs. **h**, Schematic representation of a driver mutation in *APC* causing a premature stop codon matching the SBS-*pks* motif, found in the IntOGen collection and in two independent patients from the HMF cohort with high SBS-*pks* and high ID-*pks*.

enriched for SBS-*pks*, 8.8% for ID-*pks* and 6.25% for both SBS-*pks* and ID-*pks* (Fig. 4c, Extended Data Table 1). As expected, the SBS-*pks* and ID-*pks* signatures were positively correlated in this metastasis data set ($R^2 = 0.46$ (all samples); $R^2 = 0.70$ (CRC-only); Fig. 4c), in line with their co-occurrence in our in vitro data set. The longer deletions at ID-*pks* sites were also found to co-occur with SBS-*pks* and ID-*pks* (Fig. 4e, f). In addition, we evaluated the levels of the SBS-*pks* or ID-*pks* mutational signatures in an independent cohort, generated in the framework of the Genomics England 100,000 Genomes Project. This data set comprises WGS data from 2,208 CRC tumours, predominantly of primary origin. SBS-*pks* and ID-*pks* were enriched in 5.0% and 4.4% of patients, respectively, while 44 samples (2.0%) were high in both SBS-*pks* and ID-*pks* (Fig. 4d). The relative contribution of both *pks* signatures correlated with an R^2 of 0.35 (Fig. 4d).

Finally, we also investigated to what extent the *pks* signatures can cause oncogenic mutations. To this end, we investigated the most common driver mutations found in seven cohorts of patients with CRC[25] for hits matching the extended SBS-*pks* or ID-*pks* target motifs (Fig. 3a, b). This analysis revealed that 112 out of 4,712 CRC driver mutations (2.4%) matched the colibactin target motif (Supplementary Table 1). *APC*, the most commonly mutated gene in CRC, contained the highest number of mutations that matched the SBS-*pks* or ID-*pks* target sites, with 52 out of 983 driver mutations (5.3%) matching the motifs (Fig. 4g). We then explored the mutations in the 31 SBS/ID-*pks* high CRC metastases from the HMF cohort for putative driver mutations that matched the extended motif. In total, this approach detected 209 changes in protein-coding sequences (Supplementary Table 2). Remarkably, an identical *APC* driver mutation matching the SBS-*pks* motif was found in two independent donors (Fig. 4h).

A recent publication[26] identified mutational signatures occurring in healthy human colon crypts. The authors of that study note the co-occurrence of two mutational signatures in subsets of crypts from some of the subjects. These signatures were termed SBS-A and ID-A. The authors derived hierarchical lineages of the sequenced crypts, which allowed them to conclude that the unknown mutagenic agent was active only during early childhood. Notably, SBS-A and ID-A closely match SBS-*pks* and ID-*pks*, respectively. Our data imply that *pks*$^+$ *E. coli* is the mutagenic agent that causes the SBS-A and ID-A signatures observed in healthy crypts. We assessed whether the SBS-*pks* mutational signature contributed early to the mutational load of metastatic samples from the Dutch cohort by evaluating their levels separately in clonal (pre-metastasis) or non-clonal (post-metastasis) mutations. The accumulation of SBS-*pks* and ID-*pks* at the primary tumour site or even earlier was substantiated by the abundant presence of SBS-*pks* in clonal mutations in the cohort (Extended Data Fig. 5c). In addition to CRCs, one head and neck-derived tumour and three urinary tract-derived tumours from this cohort also displayed a clear SBS-*pks* and ID-*pks* signature (Fig. 4c). Both tissues have been described as sites of *E. coli* infection[27–29]. This rare

occurrence of the *pks* signatures in non-CRC tumours was substantiated by a preprint report[30] of signatures that closely resembled SBS-*pks* and ID-*pks* in a patient with oral squamous cell carcinoma.

The distinct motifs at sites of colibactin-induced mutations may serve as a starting point for deeper investigations into the underlying processes. There is increasing evidence that colibactin forms interstrand crosslinks between two adenosines[4,5,12], and our data suggest that there is a distance of 3–4 bases between these adenosines. These crosslinks formed by a bulky DNA adduct could be resolved in different ways, including induction of DSBs, nucleotide excision repair or translesion synthesis, which in turn could result in various mutational outcomes. While our study identifies single-base substitutions and deletions as a mutational consequence, the underlying mechanisms will need to be elucidated in more detailed DNA-repair studies.

In summary, we find that prolonged exposure of wild-type human organoids to genotoxic *E. coli* allows the extraction of a unique SBS and indel signature. As organoids do not model immune or inflammation effects or other microenvironmental factors, this provides evidence that colibactin directly causes mutations in host epithelial cells. The adenine-enriched target motif is consistent with the proposed mode of action of colibactin's 'double-warhead' in attacking closely spaced adenine residues[4,5,12]. The pronounced sequence specificity reported here may inspire more detailed investigations into the interaction of colibactin with specific DNA contexts. As stated above, Stratton and colleagues[26] are likely to have described SBS-*pks* and ID-*pks* mutational signatures of the same aetiology in primary human colon crypts. This agrees with the notion that *pks*$^+$ *E. coli*-induced mutagenesis occurs in the healthy colon of individuals that harbour genotoxic *E. coli* strains[31] and that such individuals may be at an increased risk of developing CRC. The small number of *pks* signature-positive cases of urogenital and head-and-neck cancer suggests that *pks*$^+$ bacteria act beyond the colon. Notably, the presence of the *pks* island in another strain of *E. coli*, Nissle 1917, is closely linked to its probiotic effect[32]. This strain has been investigated for decades in relation to various disease indications[33]. Our data suggest that *E. coli* Nissle 1917 may induce the characteristic SBS/ID-*pks* mutational patterns described here. Future research should clarify whether this is the case in vitro, and in patients treated with *pks*$^+$ bacterial strains. This study implies that detection and removal of *pks*$^+$ *E. coli*, as well as re-evaluation of probiotic strains harbouring the *pks* island, could decrease the risk of cancer in a large group of individuals.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2080-8.

1. Allen, J. & Sears, C. L. Impact of the gut microbiome on the genome and epigenome of colon epithelial cells: contributions to colorectal cancer development. *Genome Med.* **11**, 11 (2019).
2. Gagnaire, A., Nadel, B., Raoult, D., Neefjes, J. & Gorvel, J.-P. Collateral damage: insights into bacterial mechanisms that predispose host cells to cancer. *Nat. Rev. Microbiol.* **15**, 109–128 (2017).
3. Nougayrède, J.-P. et al. *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. *Science* **313**, 848–851 (2006).
4. Wilson, M. R. et al. The human gut bacterial genotoxin colibactin alkylates DNA. *Science* **363**, eaar7785 (2019).
5. Xue, M. et al. Structure elucidation of colibactin and its DNA cross-links. *Science* **365**, eaax2685 (2019).
6. Dejea, C. M. et al. Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria. *Science* **359**, 592–597 (2018).
7. Bullman, S. et al. Analysis of *Fusobacterium* persistence and antibiotic response in colorectal cancer. *Science* **358**, 1443–1448 (2017).
8. Kostic, A. D. et al. *Fusobacterium nucleatum* potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe* **14**, 207–215 (2013).
9. Wirbel, J. et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679–689 (2019).
10. Buc, E. et al. High prevalence of mucosa-associated *E. coli* producing cyclomodulin and genotoxin in colon cancer. *PLoS One* **8**, e56964 (2013).
11. Arthur, J. C. et al. Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* **338**, 120–123 (2012).
12. Bossuet-Greif, N. et al. The colibactin genotoxin generates DNA interstrand cross-links in infected cells. *mBio* **9**, e02393-17 (2018).
13. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
14. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
15. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
16. Drost, J. et al. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science* **358**, 234–238 (2017).
17. Sato, T. et al. Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. *Gastroenterology* **141**, 1762–1772 (2011).
18. Tuveson, D. & Clevers, H. Cancer modeling meets human organoid technology. *Science* **364**, 952–955 (2019).
19. Kucab, J. E. et al. A compendium of mutational signatures of environmental agents. *Cell* **177**, 821–836.e16 (2019).
20. Jager, M. et al. Measuring mutation accumulation in single human adult stem cells by whole-genome sequencing of organoid cultures. *Nat. Protocols* **13**, 59–78 (2018).
21. Cougnoux, A. et al. Bacterial genotoxin colibactin promotes colon tumour growth by inducing a senescence-associated secretory phenotype. *Gut* **63**, 1932–1942 (2014).
22. Bartfeld, S. et al. In vitro expansion of human gastric epithelial stem cells and their responses to bacterial infection. *Gastroenterology* **148**, 126–136.e6 (2015).
23. Li, Z.-R. et al. Divergent biosynthesis yields a cytotoxic aminomalonate-containing precolibactin. *Nat. Chem. Biol.* **12**, 773–775 (2016).
24. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
25. Gonzalez-Perez, A. et al. IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **10**, 1081–1082 (2013).
26. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
27. McLellan, L. K. & Hunstad, D. A. Urinary tract infection: pathogenesis and outlook. *Trends Mol. Med.* **22**, 946–957 (2016).
28. Zawadzki, P. J. et al. Identification of infectious microbiota from oral cavity environment of various population group patients as a preventive approach to human health risk factors. *Ann. Agric. Environ. Med.* **23**, 566–569 (2016).
29. Banerjee, S. et al. Microbial signatures associated with oropharyngeal and oral squamous cell carcinomas. *Sci. Rep.* **7**, 4036 (2017).
30. Boot, A. et al. Identification of novel mutational signatures in Asian oral squamous cell carcinomas associated with bacterial infections Preprint at https://doi.org/10.1101/368753 (2019).
31. Payros, D. et al. Maternally acquired genotoxic *Escherichia coli* alters offspring's intestinal homeostasis. *Gut Microbes* **5**, 313–325 (2014).
32. Olier, M. et al. Genotoxicity of *Escherichia coli* Nissle 1917 strain cannot be dissociated from its probiotic activity. *Gut Microbes* **3**, 501–509 (2012).
33. Jacobi, C. A. & Malfertheiner, P. *Escherichia coli* Nissle 1917 (Mutaflor): new insights into an old probiotic bacterium. *Dig. Dis.* **29**, 600–607 (2011).

Genomics England Research Consortium

J. C. Ambrose[19], P. Arumugam[19], E. L. Baple[19], M. Bleda[19], F. Boardman-Pretty[19,20], J. M. Boissiere[19], C. R. Boustred[19], H. Brittain[19], M. J. Caulfield[19,20], G. C. Chan[19], C. E. H. Craig[19], L. C. Daugherty[19], A. de Burca[19], A. Devereau[19], G. Elgar[19,20], R. E. Foulger[19], T. Fowler[19], P. Furió-Tarí[19], J. M. Hackett[19], D. Halai[19], A. Hamblin[19], S. Henderson[19,20], J. E. Holman[19], T. J. P. Hubbard[19], K. Ibáñez[19,20], R. Jackson[19], L. J. Jones[19,20], D. Kasperaviciute[19,20], M. Kayikci[19], L. Lahnstein[19], L. Lawson[19], S. E. A. Leigh[19], I. U. S. Leong[19], F. J. Lopez[19], F. Maleady-Crowe[19], J. Mason[19], E. M. McDonagh[19,20], L. Moutsianas[19,20], M. Mueller[19,20], N. Murugaesu[19], A. C. Need[19,20], C. A. Odhams[19], C. Patch[19,20], D. Perez-Gil[19], D. Polychronopoulos[19], J. Pullinger[19], T. Rahim[19], A. Rendon[19], P. Riesgo-Ferreiro[19], T. Rogers[19], M. Ryten[19], K. Savage[19], K. Sawant[19], R. H. Scott[19], A. Siddiq[19], A. Sieghart[19], D. Smedley[19,20], K. R. Smith[19,20], A. Sosinsky[19,20], W. Spooner[19], H. E. Stevens[19], A. Stuckey[19], R. Sultana[19], E. R. A. Thomas[19,20], S. R. Thompson[19], C. Tregidgo[19], A. Tucci[19,20], E. Walsh[19], S. A. Watters[19], M. J. Welland[19], E. Williams[19], K. Witkowska[19,20], S. M. Wood[19,20] & M. Zarowiecki[19]

[19]Genomics England, London, UK. [20]William Harvey Research Institute, Queen Mary University of London, London, UK.

Methods

Human material and organoid cultures

Ethical approval was obtained from the ethics committees of the University Medical Center Utrecht, Hartwig Medical Foundation and Genomics England. Written informed consent was obtained from patients. All experiments and analyses were performed in compliance with relevant ethical regulations.

Organoid culture

Clonal organoid lines were derived and cultured as described previously[16,17]. In brief, wild-type human intestinal organoids (clonal lines ASC-5a and ASC-6a, previously described[34]) were cultured in domes of Cultrex Pathclear Reduced Growth Factor Basement Membrane Extract (BME) (3533-001, Amsbio) covered by medium containing 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). For derivation of clonal lines, cells were sorted by fluorescence-activated cell sorting (FACS) and grown at a density of 50 cells per µl in BME. The ROCK 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 one well per organoid. All organoid lines were regularly tested to rule out mycoplasma infection and authenticated using SNP profiling. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Organoid bacteria co-culture

The genotoxic *pks*$^{+}$ *E. coli* strain was previously isolated from a patient with CRC and isogenic *pks*Δ*clbQ* knock-out and *pks*Δ*clbQ:clbQ* recomplemented strains were generated from this strain[21]. Bacteria were initially cultured in Advanced DMEM (Gibco) supplemented with glutamax and HEPES to an optical density (OD) of 0.4. They were then microinjected into the lumens of organoids as previously described[22,35]. Bacteria were injected at a multiplicity of infection of 1 together with 0.05% (w/v) FastGreen dye (Sigma) to allow tracking of injected organoids. At this point, 5 µg/ml of the non-permeant antibiotic gentamicin was added to the medium to prevent overgrowth of bacteria outside the organoid lumen. Cell viability was assessed as follows: organoids were harvested after 1, 3 or 5 days (bacteria were removed by primocin treatment at day 3) of co-culture in cold DMEM (Gibco) and incubated in TrypLE Express (Gibco) at 37 °C for 5 min with repeated mechanical shearing. Single cells were resuspended in DMEM with added DAPI, incubated on ice for at least 15 min and assessed for viability on a BD FACS Canto. Cells positive for DAPI were considered dead, while cells maintaining DAPI exclusion were counted as viable. Bacterial growth kinetics were assessed by harvesting, organoid dissociation with 0.5% saponin for 10 min and re-plating of serial dilutions on LB plates. CFUs were quantified after overnight culture at 37 °C. *E. coli* were killed with 1× Primocin (InvivoGen) after 3 days of co-culture, after which organoids were left to recover for 4 days before being passaged. When the organoids reached a cystic stage again (typically after 2–3 weeks), the injection cycle was repeated. This procedure was repeated five times (three times for ASC clone 6-a and the *clbQ* recomplementation experiment in ASC clone 5-a) to reduce injection heterogeneity and ensure accumulation of enough mutations for reliable signature detection.

Whole-mount organoid immunofluorescence, DNA damage quantification and scanning electron microscopy

Organoids co-cultured with *pks*$^{+}$ or *pks*Δ*clbQ E. coli*[21] were collected in cell recovery solution (Corning) and incubated at 4 °C for 30 min with regular shaking in order to free them from BME. For FANCD2 staining, organoids were pre-permeabilized with 0.2% Triton-X (Sigma) for 10 min at room temperature. Then, organoids were fixed in 4% formalin overnight at 4 °C. Subsequently, organoids were permeabilized with 0.5% Triton-X (Sigma), 2% donkey serum (BioRad) in PBS for 30 min at 4 °C and blocked with 0.1% Tween-20 (Sigma) and 2% donkey serum in PBS for 15 min at room temperature. Organoids were incubated with mouse anti-γH2AX (Millipore; clone JBW301; 1:1,000 dilution) or rabbit anti-FANCD2 (affinity purified as described[36]; 1 mg/ml) primary antibody overnight at 4 °C. Then, organoids were washed four times with PBS and incubated with either secondary goat anti-mouse AF-647 (Thermo Fisher, catalogue number A-21235, 1:500 dilution) or goat anti-rabbit AF-488 (Life Technologies, catalogue number A21206, 1:500 dilution) antibodies, respectively, for 3 h at room temperature in the dark and washed again with PBS. Organoids were imaged using an SP8 confocal microscope (Leica). Fluorescent microscopic images of γH2AX foci were quantified as follows: nuclei were classified as containing either no foci or one or more foci. The fraction of nuclei containing foci over all nuclei is displayed as one datapoint per organoid. Organoids co-cultured with bacteria for 24 h were harvested as described above and processed for scanning electron microscopy as previously described[35].

WGS and read alignment

For WGS, clonal and subclonal cultures were generated for each condition. DNA was isolated from these clonal cultures using the DNeasy Blood and Tissue Kit (Qiagen) according to the manufacturer's instructions. Illumina DNA libraries were prepared using 50 ng of genomic DNA isolated from the (sub-)clonal cultures isolated using a TruSeq DNA Nano kit. The parental ASC 5a clone was sequenced on a HiSeq XTEN instrument at 30× base coverage. All other samples were sequenced using an Illumina Novaseq 6000 with 30× base coverage. Reads were mapped against the human reference genome version GRCh37 using Burrows–Wheeler Aligner[37] (BWA) version v0.7.5 with settings bwa mem -c 100 -M. Sequences were marked for duplicates using Sambamba (v0.4.732) and realigned using GATK IndelRealigner (GATK version 3.4-46). The full description and source code of the pipeline is available at https://github.com/UMCUGenetics/IAP.

Mutation calling and filtration

Mutations were called using GATK Haplotypecaller (GATK version 3.4-46) and GATK Queue to produce a multi-sample Vcf file[20]. The quality of the variants was evaluated using GATK VariantFiltration v3.4-46 using the following settings: -snpFilterName SNP_LowQualityDepth -snpFilterExpression "QD < 2.0" -snpFilterName SNP_MappingQuality -snpFilterExpression "MQ < 40.0" -snpFilterName SNP_StrandBias -snpFilterExpression "FS > 60.0" -snpFilterName SNP_HaplotypeScoreHigh -snpFilterExpression "HaplotypeScore > 13.0" -snpFilterName SNP_MQRankSumLow -snpFilterExpression "MQRankSum < -12.5" -snpFilterName SNP_ReadPosRankSumLow -snpFilterExpression "ReadPosRankSum < -8.0" -snpFilterName SNP_HardToValidate -snpFilterExpression "MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1)" -snpFilterName SNP_LowCoverage -snpFilterExpression "DP < 5" -snpFilterName SNP_VeryLowQual -snpFilterExpression "QUAL < 30" -snpFilterName SNP_LowQual -snpFilterExpression "QUAL >= 30.0 && QUAL < 50.0" -snpFilterName SNP_SOR -snpFilterExpression "SOR > 4.0" -cluster 3 -window 10 -indelType INDEL -indelType MIXED -indelFilterName INDEL_LowQualityDepth -indelFilterExpression "QD < 2.0" -indelFilterName INDEL_StrandBias -indelFilterExpression "FS > 200.0" -indelFilterName INDEL_ReadPosRankSumLow -indelFilterExpression "ReadPosRankSum < -20.0" -indelFilterName INDEL_HardToValidate -indelFilterExpression "MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1)" -indelFilterName INDEL_LowCoverage -indelFilterExpression "DP < 5" -indelFilterName INDEL_VeryLowQual -indelFilterExpression "QUAL < 30.0" -indelFilterName INDEL_LowQual

-indelFilterExpression "QUAL >= 30.0 && QUAL < 50.0" -indelFilterName INDEL_SOR -indelFilterExpression "SOR > 10.0.

Somatic SBS and indel filtering

To obtain high-confidence catalogues of mutations induced during culture, we applied extensive filtering steps as previously described[20]. First, only variants obtained by GATK VariantFiltration with a GATK phred-scaled quality score of ≥100 for SBSs and ≥250 for indels were selected. Subsequently, we considered only variants with at least 20× read coverage in control and sample. We additionally filtered base substitutions with a GATK genotype score (GQ) lower than 99 or 10 in WGS(t_n) or WGS(t_0), respectively (Fig. 2a). Indels were filtered when GQ scores were higher than 60 WGS(t_n) or 10 in WGS(t_0). All variants were filtered against the Single Nucleotide Polymorphism Database v137.b3730, from which SNPs present in the COSMICv76 database were excluded. To exclude recurrent sequencing artefacts, we excluded all variants that were variable in at least three individuals in a panel of bulk-sequenced mesenchymal stromal cells[38]. Next, all variants present at the start of co-culture (WGS(t_0) in Fig. 2a) were filtered from those detected in the clonal *pks*$^+$ *E. coli*, *pks*Δ*clbQ E. coli* co-cultures (WGS(t_n) in Fig. 2a) or dye culture. Indels were selected only when no called variants in WGS(t_0) were present within 100 bp of the indel and if not shared in WGS(t_0). In addition, both indels and SNVs were filtered for the additional parameters: mapping quality (MQ) of at least 60 and a variant allele frequency (VAF) of 0.3 or higher to exclude variants obtained during the clonal step. Finally, all multi-allelic variants were removed. Scripts used for filtering SBSs (SNVFIv1.2) and indels (INDELFIv1.5) can be found at https://github.com/ToolsVanBox/.

Mutational profile analysis

To extract mutational signatures from the high-quality mutational catalogues after filtering, we used the R package MutationalPatterns to obtain 96-trinucleotide SBS and indel subcategory counts for each clonally cultured sample[39] (Extended Data Fig. 1a, d). To identify the additional mutational effects induced by *pks*$^+$ *E. coli* (SBS and ID), we pooled mutation numbers for each culture condition (*pks*Δ*clbQ* and *pks*$^+$), and subtracted the mutational counts of *pks*Δ*clbQ* from *pks*$^+$ (Fig. 2c, e, Extended Data Fig. 2b, d). For the clones exposed to *pks*Δ*clbQ:clbQ E. coli*, we subtracted relative levels of the *pks*Δ*clbQ* mutations in the same organoid line. This enabled us to correct for the background of mutations induced by *pks*Δ*clbQ E. coli* and the injection dye. To determine the transcriptional strand bias of mutations induced during *pks*$^+$ *E. coli* exposure, we selected all SBSs within gene bodies and checked whether the mutated C or T was located on the transcribed or non-transcribed strand. We defined the transcribed area of the genome as all protein-coding genes based on Ensembl v75 (GCRh37)[40] and included introns and untranslated regions. The extended sequence context around mutation sites was analysed and displayed using an in-house script ('4_extended_sequence_context.R'). Two-bit sequence motifs were generated using the R package ggseqlogo. Cosine similarities between indel and SBS profiles were calculated using the function 'cos_sim_matrix' from the MutationalPatterns package.

Analysis of clonal mutations in the SBS/ID-*pks*-high CRC tumours

From the 31 SBS/ID-*pks*-high CRC tumours, clonal and subclonal SBSs were defined to contain a purity- or ploidy-adjusted allele-fraction (PURPLE_AF) of <0.4 or >0.2, respectively[41]. Signature re-fitting on both fractions was performed with the same signatures as described above for the initial re-fitting of the HMF cohort.

Analysis of >1-bp deletions matching *pks*-motif

For each >1-bp T-deletion observed in organoid clones or the HMF cohort, the sequence of the deleted bases and 5-bp flanking regions was retrieved using the R function getSeq from the package BSgenome. Retrieved sequences were examined for the presence of a 5-bp motif matching the *pks*-motifs identified (Extended Data Fig. 4a): AAAAT, AAATT, AATTT or ATTTT. Sequences containing one or more matches with the motifs were marked as positive for containing the motif.

NMF extraction of signatures from the HMF CRC cohort

To identify SBS-*pks* in an unbiased manner, signature extraction was performed on all 496 samples from colorectal primary tumours present in the HMF metastatic cancer database[24]. All variants containing the 'PASS' flag were used for analysis. Signature extraction was performed using non-negative matrix factorization (NMF), using the R package MutationalPatterns, function 'extract_signatures' with the following settings: rank = 17, nrun = 200. The cosine similarity of the extracted signature matching SBS-*pks* was re-fitted to the COSMIC SigProfiler signatures and SBS-*pks* was determined as described above to determine similarity (Extended Data Fig. 5a, b).

Signature re-fitting on HMF cohort

Mutation catalogues containing somatic variants processed as described[24] were obtained from the HMF. All variants containing the 'PASS' flag in the HMF data set were selected. Single-base trinucleotide and indel subcategory counts were extracted using the R package MutationalPatterns and in house-written R scripts, respectively. To determine the contribution of SBS-*pks* and ID-*pks* to these mutational catalogues, we re-fitted the COSMIC SigProfiler mutational SBS and ID signatures v3 (https://cancer.sanger.ac.uk/cosmic/signatures/), in combination with SBS-*pks* and ID-*pks*, to the mutational catalogues using the MutationalPatterns function 'fit_to_signatures'. Signatures marked as possible sequencing artefacts were excluded from the re-fitting. Cut-off values for high SBS-*pks* and ID-*pks* levels were manually set at 5% each. The numbers of SBS/ID-*pks*-positive samples were compared between CRC and other cancer types by Fisher's exact test (two-tailed).

Mutation calling and filtration (Genomics England cohort)

As part of the Genomics England 100,000 Genomes Project (main programme version 7)[42] standard pipeline, 2,208 CRC genomes were sequenced on the Illumina HiSeq X platform. Reads were aligned to the human genome (GRCh38) using the Illumina iSAAC aligner 03.16.02.1[43]. Mutations were called using Strelka and filtered in accordance with the HMF data set[24].

Before examining somatic mutations for the *pks* mutational signature, mutation calls were first subjected to additional filtering steps similar to those previously described[24]. All calls present in the matched normal sample were removed. The calls were split into high and low confidence genomic regions according to lists available at ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv3.3.1/GRCh38/. Somatic mutation calls in high-confidence regions were passed with a somatic score (QSI or QSS) of 10, while calls in low confidence regions were passed with a score of 20. A pool of 200 normal samples was constructed, and any calls present in three or more normal samples were removed. Any groups of single-nucleotide variants within 2 bp were considered to be miscalled multiple nucleotide variants and were removed. Finally, all calls had to pass the Strelka 'PASS' filter. Mutational signatures were then analysed as described above for the HMF cohort.

Detection of *pks*-signature mutations in protein-coding regions

Mutations were extracted from the 31 SBS/ID-*pks*-high CRC samples. Exonic regions were defined as all autosomal exonic regions reported in Ensembl v75 (GCRh37)[40]. All extracted CRC mutations were filtered for localization in exonic regions using the Bioconductor packages GenomicRanges[44] and BSgenome. In a second filtering step, the sequence context of mutations was required to match the following criteria. For SBS-*pks*: T > N mutation, A or T directly upstream and downstream, A three bases upstream. For ID-*pks*: single T deletion, A directly

upstream, a stretch of an A homopolymer followed by a T polymer with combined length of at least five nucleotides, but no stretch exceeding ten nucleotides in length. Mutations passing both filter steps were further filtered for presence of a predicted 'high' or 'moderate' score in the transcript with the highest impact score according to the reported SnpEff annotation.

To assess the mutagenic effect of *pks*, we obtained all mutations from the 50 highest mutated genes in CRC from IntOGen[25], release 2019.11.12. Mutations were filtered to match the *pks* motif according to the sequence criteria stated above apart from the predicted impact score. Mutations in *APC* were plotted using the R package rtrackViewer, using only exonic mutations.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Whole-genome sequence data have been deposited in the European Genome–Phenome Archive (https://ega-archive.org); accession number EGAS00001003934. The data used from the Hartwig Medical Foundation and Genomics England databases consist of patient-level somatic variant data (annotated variant call data) and are considered privacy sensitive and available through access-controlled mechanisms. Patient-level somatic variant and clinical data were obtained from the Hartwig Medical Foundation under data request number DR-084. Somatic variant and clinical data are freely available for academic use from the Hartwig Medical Foundation through standardized procedures. Privacy and publication policies, including co-authorship policies, can be retrieved from: https://www.hartwigmedicalfoundation.nl/en/data-policy/. Data request forms can be downloaded from https://www.hartwigmedicalfoundation.nl/en/applying-for-data/. To gain access to the data, this data request form should be emailed to info@hartwigmedicalfoundation.nl, upon which it will be evaluated within six weeks by the HMF Scientific Council and an independent Data Access Board. When access is granted, the requested data become available through a download link provided by HMF. Somatic variant data from the Genomics England data set were analysed within the Genomics England Research Environment secure data portal, under Research Registry project code RR87, and exported from the Research Environment following data transfer request 1000000003652 on 3 December 2019. The Genomics England data set can be accessed by joining the community of academic and clinical scientist via the Genomics England Clinical Interpretation Partnership (GeCIP), https://www.genomicsengland.co.uk/about-gecip/. To join a GeCIP domain, the following steps have to be taken: 1. Your institution has to sign the GeCIP Participation Agreement, which outlines the key principles that members of each institution must adhere to, including our Intellectual Property and Publication Policy. 2. Submit your application using the relevant form found at the bottom of the page (https://www.genomicsengland.co.uk/join-a-gecip-domain/). 3. The domain lead will review your application, and your institution will verify your identity for Genomics England and communicate confirmation directly to Genomics England. 4. Your user account will be created. 5. You will be sent an email containing a link to complete Information Governance training and sign the GeCIP rules (https://www.genomicsengland.co.uk/wp-content/uploads/2019/07/GeCIP-Rules_29-08-2018.pdf). Completing the training and signing the GeCIP Rules are requirements for you to access the data. After you have completed the training and signed the rules, you will need to wait for your access to the Research Environment to be granted. 6. This will generally take up to one working day. You will then receive an email letting you know your account has been given access to the environment, and instructions for logging in (for more detail, see: https://www.genomicsengland.co.uk/join-a-gecip-domain/). Details of the data access agreement can be retrieved from https://figshare.com/articles/GenomicEnglandProtocol_pdf/4530893/5. All requests will be evaluated by the Genomics England Access Review Committee taking into consideration patient data protection, compliance with legal and regulatory requirements, resource availability and facilitation of high-quality research. All analysis of the data must take place within the Genomics England Research Environment secure data portal, https://www.genomicsengland.co.uk/understanding-genomics/data/ and exported following approval of a data transfer request. Regarding co-authorship, all publications using data generated as part of the Genomics England 100,000 Genomes Project must include the Genomics England Research Consortium as co-authors. The full publication policy is available at https://www.genomicsengland.co.uk/about-gecip/publications/. All other data supporting the findings of this study are available from the corresponding author upon request.

Code availability

All analysis scripts are available at https://github.com/ToolsVanBox/GenotoxicEcoli.

34. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
35. Heo, I. et al. Modelling *Cryptosporidium* infection in human small intestinal and lung organoids. *Nat. Microbiol.* **3**, 814–823 (2018).
36. Pace, P. et al. FANCE: the link between Fanconi anaemia complex assembly and activity. *EMBO J.* **21**, 3414–3423 (2002).
37. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
38. Osorio, F. G. et al. Somatic mutations reveal lineage relationships and age-related mutagenesis in human hematopoiesis. *Cell Rep.* **25**, 2308–2316.e4 (2018).
39. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018).
40. Cunningham, F. et al. Ensembl 2015. *Nucleic Acids Res.* **43**, D662–D669 (2015).
41. Cameron, D. L. et al. GRIDSS, PURPLE, LINX: unscrambling the tumor genome via integrated analysis of structural variation and copy number. Preprint at https://doi.org/10.1101/781013 (2019).
42. Genomics England *The National Genomics Research and Healthcare Knowledgebase* https://www.genomicsengland.co.uk/the-national-genomics-research-and-healthcare-knowledgebase/ (2017).
43. Raczy, C. et al. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics* **29**, 2041–2043 (2013).
44. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLOS Comput. Biol.* **9**, e1003118 (2013).

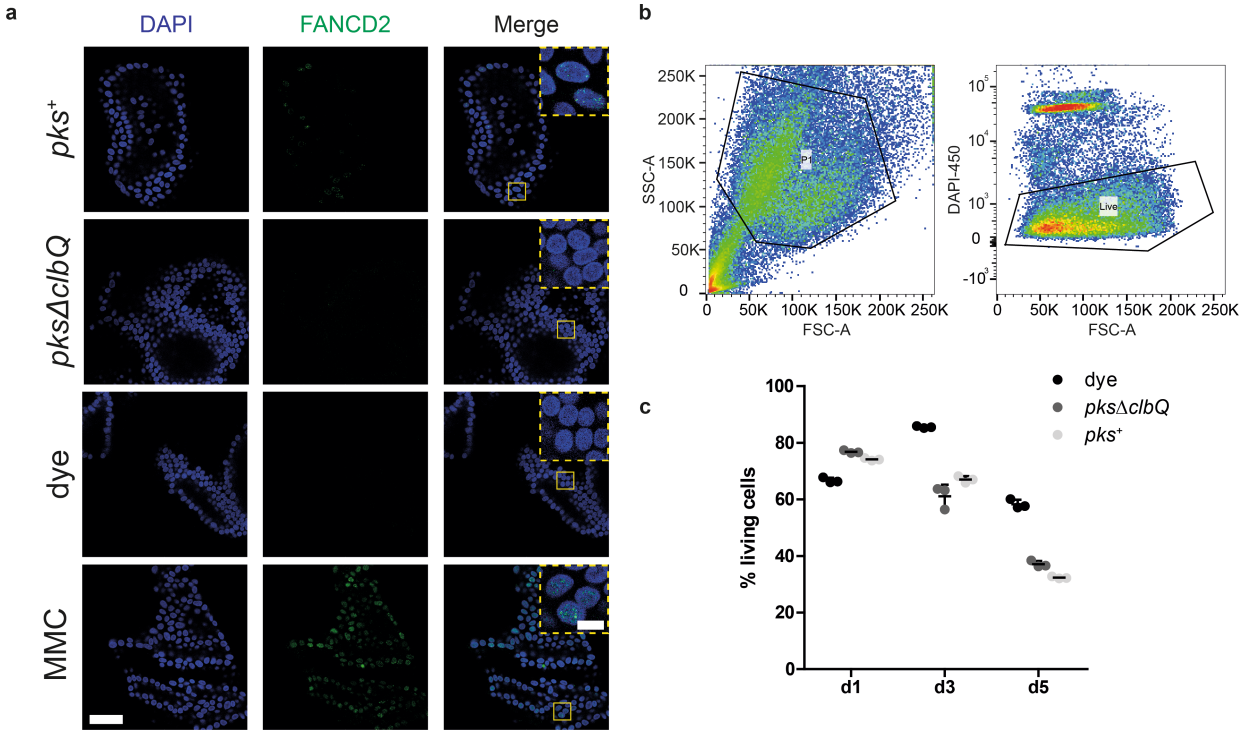
Acknowledgements We thank J. H. J. Hoeijmakers, P. Knipscheer and J. I. Garaycoechea for discussions on DNA damage, and P. Robinson, K. Vervier, T. Lawley, and M. Stratton for explorative analysis and discussions. This publication and the underlying study have been made possible partly on the basis of the data that Hartwig Medical Foundation and the Center of Personalised Cancer Treatment (CPCT) have made available to the study. This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support. This work was supported by CRUK grant OPTIMISTICC (C10674/A27140), the Gravitation projects CancerGenomiCs.nl and 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 (C.P.-M., J.P.), the Oncode Institute (partly financed by the Dutch Cancer Society), the European Research Council under ERC Advanced Grant Agreement no. 67013 (J.P., T.M., H.C.), a VIDI grant from the NWO (no. 016.Vidi.171.023) to R.v.B. that supports A.R.H. and NWO building blocks of life project: Cell dynamics within lung and intestinal organoids (737.016.009) (M.H.G.). With financial support from ITMO Cancer AVIESAN (Alliance Nationale pour les Sciences de la Vie et de la Santé, National Alliance for Life Sciences & Health) within the framework of the Cancer Plan (HTE201601) (G.D., R.B.) as well as Howard Hughes Medical Institute, Mathers Foundation, and NIH-1R01DK115728-01A1 (Y.M., K.C.G.).

Author contributions C.P.-M., J.P., A.R.H. and H.C. conceived the study; C.P.-M., J.P., A.R.H., R.v.B. and H.C. wrote the manuscript; A.R.H., H.M.W., F.M. and R.v.B. performed signature analysis; A.R.H., A.v.H., H.M.W., J.N., C.G., P.Q., M.G., M.M. and E.C. provided access to and analysed patient WGS data; G.D. and R.B. isolated bacterial strains and generated knockouts; C.P.-M., J.P., T.M., R.v.d.L., M.H.G. and S.v.E. established and performed organoid cloning experiments; C.P.-M., J.P. and J.B. performed organoid co-culture experiments; P.B.S., F.L.P., J.T.

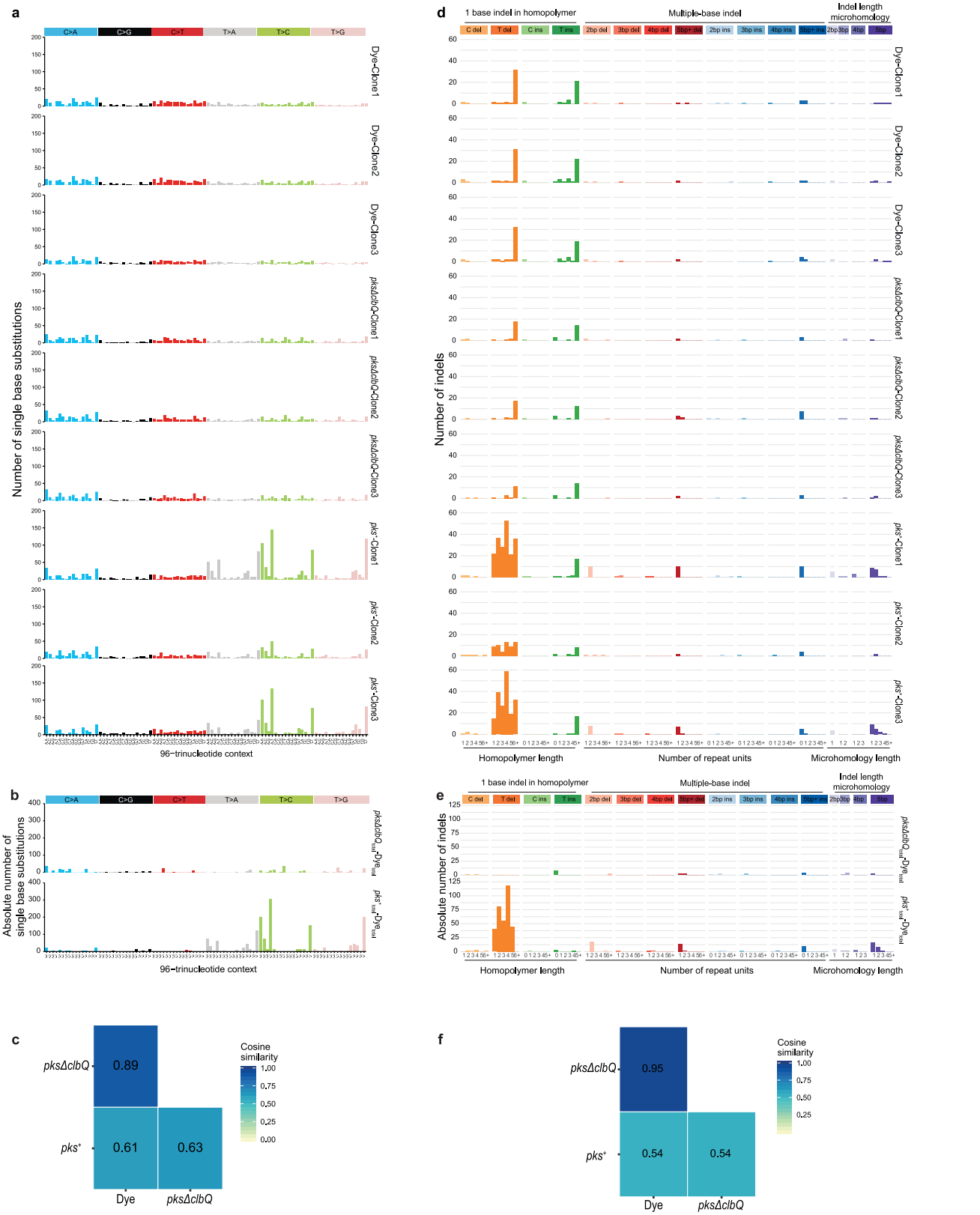
and R.J.L.W. performed bacteria validation and assays. Y.M. and K.C.G. provided and advised on the use of the Wnt surrogate-Fc fusion reagent.

Competing interests H.C. is inventor on several patents related to organoid technology; his full disclosure is given at https://www.uu.nl/staff/JCClevers/. M.M. is scientific advisory board chair and a consultant for OrigiMed, receives research support from Bayer, Janssen, and Ono, and receives royalty payments from Labcorp. H.C and K.C.G are co-founders of Surrozen.

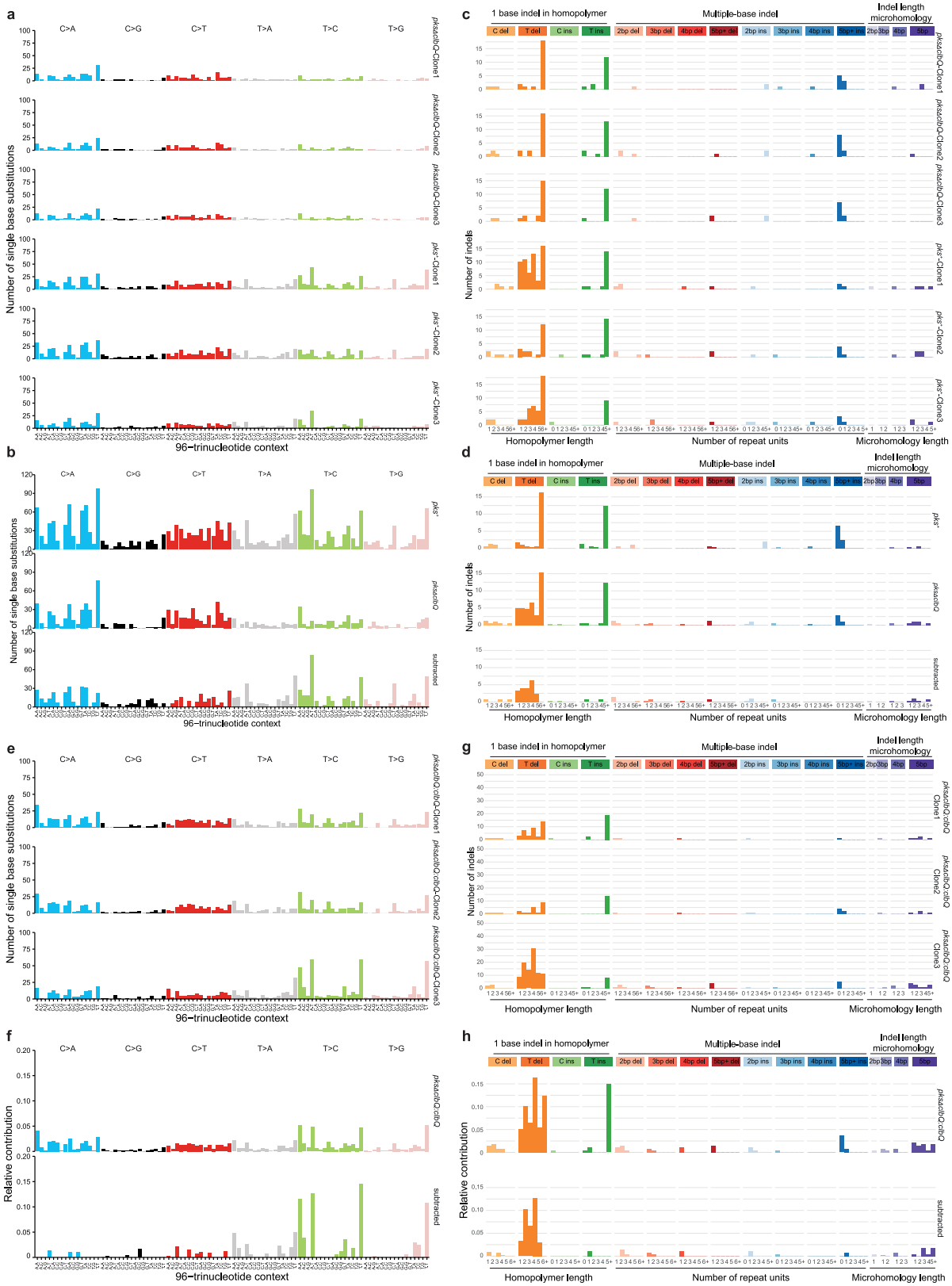
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2080-8.
Correspondence and requests for materials should be addressed to R.v.B. or H.C.
Peer review information *Nature* thanks Bogdan Fedeles, Christian Jobin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



Extended Data Fig. 1 | Co-culture with genotoxic *pks*$^+$ *E. coli* induces DNA interstrand crosslinks in healthy human intestinal organoids. **a**, Representative images (out of $n = 5$ organoids per group) of DNA interstrand crosslink formation after 1 day of co-culture, measured by FANCD2 immunofluorescence (green). Nuclei were stained with DAPI (blue). Yellow boxes represent inset area. Scale bars, 50 µm (main image); 10 µm (inset). Experiment was repeated independently twice with similar results. **b**, Gating strategy to select epithelial cells (left) and to quantify viable cells (right). **c**, Mean ± s.d. viability of intestinal organoid cells after 1, 3 or 5 days of co-culture ($n = 3$ technical replicates) (bacteria eliminated after 3 days of co-culture). Points are independent replicates.



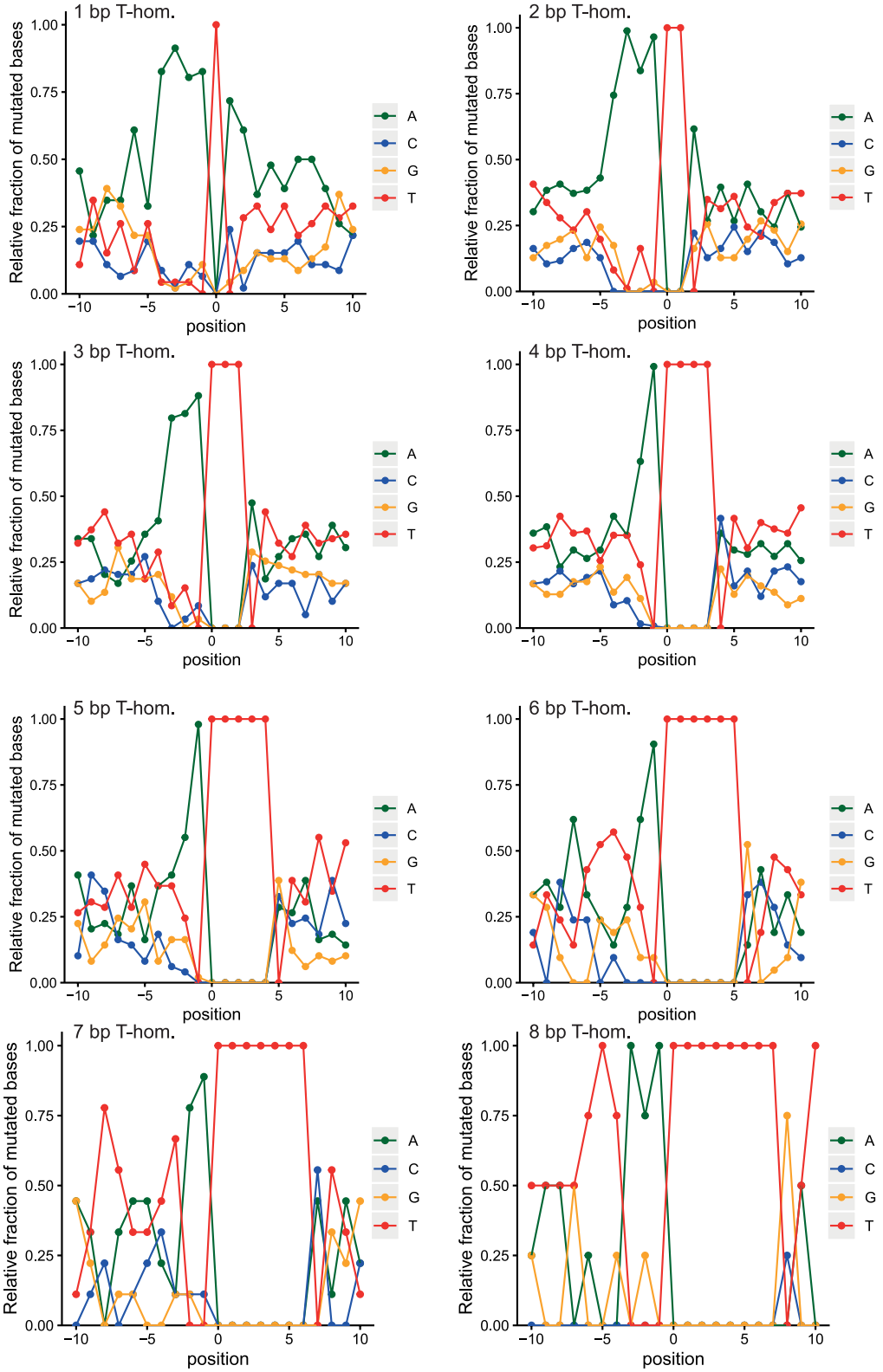
Extended Data Fig. 2 | Genotoxic *pks*$^+$ *E. coli* induce SBS-*pks* and ID-*pks* mutational signatures after long-term co-culture with wild-type intestinal organoids. a, Ninety-six-trinucleotide mutational spectra of SBSs in each of the three individual clones sequenced per condition. Top three, dye; middle three, *pks*Δ*clbQ E. coli*; bottom three, *pks*$^+$ *E. coli*. **b**, Total 96-trinucleotide mutational spectra of organoids injected with *pks*$^+$ *E. coli* or *pks*Δ*clbQ E. coli* from which SBSs in dye-injected organoids have been subtracted. **c**, Heatmap depicting cosine similarity between 96-trinucleotide mutational profiles of organoids injected with dye, *pks*$^+$ *E. coli* or *pks*Δ*clbQ E. coli*. **d**, Indel mutational spectra plots from each of the three individual clones sequenced per condition. Top three, dye; middle three, *pks*Δ*clbQ E. coli*; bottom three, *pks*$^+$ *E. coli*. **e**, Total indel mutational spectra of organoids injected with *pks*$^+$ *E. coli* and *pks*Δ*clbQ E. coli* from which indels in dye-injected organoids have been subtracted. **f**, Heatmap depicting cosine similarity between indel mutational profiles of organoids injected with dye, *pks*$^+$ *E. coli* or *pks*Δ*clbQ E. coli*.



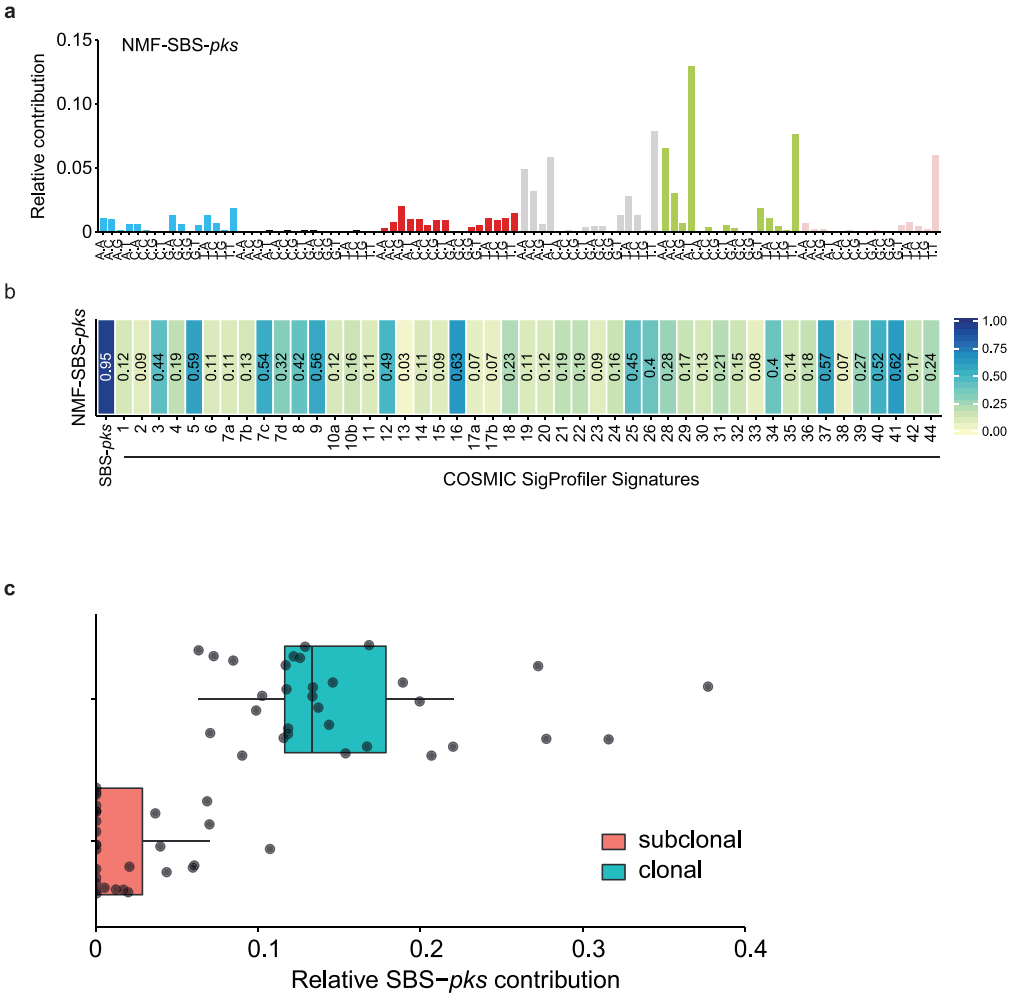
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Genotoxic *pks*$^+$ *E. coli* and isogenic strain reconstituted with *pks*Δ *clbQ:clbQ* induce SBS-*pks* and ID-*pks* mutational signatures after co-culture. a, Ninety-six-trinucleotide mutational spectra of SBSs in three individual clones from the independent human healthy intestinal organoid line ASC-6a co-cultured for three rounds with *pks*$^+$ or *pks*Δ*clbQ E. coli*. **b**, Top, total 96-trinucleotide mutational spectra from the three clones co-cultured with from *pks*$^+$ or *pks*Δ*clbQ E. coli* shown in **a**. Bottom, resulting 96-trinucleotide mutational spectrum from ASC-6a organoids co-cultured with *pks*$^+$ *E. coli* after the subtraction of background mutations from three parallel *pks*Δ*clbQ E. coli* co-cultures (cosine similarity to SBS-*pks* = 0.77). **c**, Indel mutational spectra from the three independent ASC-6a clones co-cultured for three rounds with *pks*$^+$ or *pks*Δ*clbQ E. coli*. **d**, Top, total indel mutational spectra from the three clones co-cultured with *pks*$^+$ or *pks*Δ*clbQ E. coli* shown in **c**. Bottom, resulting indel mutational spectrum from the independent ASC-6a organoids co-cultured with *pks*$^+$ *E. coli* after the subtraction of background mutations from three parallel *pks*Δ*clbQ E. coli* co-cultures (cosine similarity to ID-*pks* = 0.93). **e**, Ninety-six-trinucleotide mutational spectra from three individual clones of the ASC-5a line co-cultured for three rounds with the isogenic recomplemented *E. coli* strain *pks*Δ*clbQ:clbQ*. **f**, Top, total 96-trinucleotide mutational spectrum from the three clones co-cultured with *pks*Δ*clbQ:clbQ E. coli* shown in **e**. Bottom, resulting mutational spectrum after subtracting *pks*Δ*clbQ* background (cosine similarity to SBS-*pks* = 0.95). **g**, Indel mutational spectra from three individual clones of the ASC-5a line co-cultured for three rounds with the isogenic recomplemented *E. coli* strain *pks*Δ*clbQ:clbQ*. **h**, Top, total indel mutational spectrum from the three clones co-cultured with *pks*Δ*clbQ:clbQ E. coli* shown in **g**. Bottom, resulting mutational spectrum after subtracting *pks*Δ*clbQ* background (cosine similarity to ID-*pks* = 0.95).

a



Extended Data Fig. 4 | Detailed sequence context for ID-*pks* and longer deletions by length. a, Ten-base up- and downstream profile shows an upstream homopolymer of adenosines that favours induction of T deletions. The length of the adenosine stretch decreases with increasing T homopolymer length (1–8, top left to bottom right).



Extended Data Fig. 5 | Signature extraction and clonal contribution of SBS-*pks* in CRC metastases. a, De novo NMF-SBS-*pks* signature extracted by NMF on all 496 CRC metastases in the HMF data set. **b**, Cosine similarity scores between the de novo extracted SBS signature in **a** and COSMIC SigProfiler signatures, including our experimentally defined SBS-*pks* signature (left). **c**, Relative contribution of SBS-*pks* to clonal (corrected variant allele frequency >0.4, blue) and subclonal fractions (corrected variant allele frequency <0.2, red) of mutations in the 31 SBS/ID-*pks* high CRC metastases from the HMF cohort. Box, upper and lower quartiles; centre line, mean; whiskers, largest value no more than 1.5 times the interquartile range extending from the box; points, individual CRC metastases.

Extended Data Table 1 | SBS-*pks* and ID-*pks* levels across tissue types

Primary Tumor Location	Total number	SBS-*pks* > 0.05	ID-*pks* > 0.05	SBS-*pks* > 0.05 & ID-*pks* > 0.05
CRC	496	37 (7.5%)	44 (8.8%)	31 (6.25%)
Head & Neck	61	1 (1.6%)	1 (1.6%)	1 (1.6%)
Urinary Tract	142	3 (2.1%)	6 (4.2%)	3 (2.1%)
Other	2969	12 (0.4%)	134 (4.5%)	1 (0.03%)

Sample numbers are displayed by primary tumour type per row. Numbers of samples with more than 5% contribution of ID-*pks*, SBS-*pks* or both are shown; the percentage of positive samples per tissue is indicated in parentheses.

nature research

Corresponding author(s): Hans Clevers

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Leica LAS X (v1.1), BD FACSDiva Software (v8.0.1)
Data analysis	ImageJ (Fiji, Version 1.51n), Microsoft Excel 2016, R Studio (v1.2.1335), R (v3.6.0), Leica LAS X (v1.1), GraphPad PRISM (v5.0), FlowJo (v10.1r1), BWA (v0.7.5), Sambamba (v0.4.732), GATK (v3.4-46), SNVFI (v1.2), INDELFI (v1.5), Illumina iSAAC aligner (v03.16.02.1), the code of the read alignment pipeline is available on https://github.com/UMCUGenetics/IAP, customized code is made available on https://github.com/ToolsVanBox/GenotoxicEcoli

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Whole-genome sequence data have been deposited in the European Genome-phenome Archive (EGA; https://ega-archive.org); accession number: EGAS00001003934. Metastasis WGS data can be requested at https://www.hartwigmedicalfoundation.nl/en/. Primary CRC WGS data are available from Genomics England Ltd, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Genomics England Ltd.
All other data supporting the findings of this study are available from the corresponding author on reasonable request.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences

Study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample-size calculation was performed. Results obtained were highly significant and consistent and did not require larger experimental groups. The sequencing of 3 clones per condition is standard practice and allows faithful identification of novel patterns in mutational signature studies (Drost et al. 2017, Kucab et al. 2019). Cosine similarities to patient-derived signatures confirm this approach and group sizes chosen. The HMF metastasis cohort and Genomics England 100,000 Genomes Project primary CRC cohort are among the largest WGS datasets available for CRC and allowed for unbiased extraction of the SBS/ID-pks signatures.
Data exclusions	No datapoints were excluded from the presented datasets.
Replication	The number of times each experiment has been repeated with similar results is stated in each figure legend or the methods section. The key finding of the manuscript, i.e. the SBS-pks and ID-pks mutational signatures, have been reproduced in the line of an independent donor and with a ClbQ recomplemented strain. They were further confirmed by detection in 2 large, independent cohorts of cancer patients. All attempts of replication were successful.
Randomization	No randomization was performed. The organoid lines were subjected to clonal outgrowth and therefore genetically identical at the start of bacterial exposure. Group allocation was performed by injecting different wells with identical genotype with different bacteria. For this reason, no randomization had to be performed.
Blinding	Quantification of nuclei positive for yH2AX was performed in a blinded fashion, with C.P.M. and J.P. taking sets of images and blinded counting by the other.

Materials & experimental systems

Policy information about availability of materials

n/a	Involved in the study
☒	☐ Unique materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Research animals
☐	☒ Human research participants

Antibodies

Antibodies used	1) mouse anti-yH2AX Millipore; clone JBW301. Catalog number: 05-636-I; monoclonal, 1:1000 dilution 2) rabbit anti-FANCD2 (kind gift by J. Garaycoechea, affinity purified antibody - see reference 34, final concentration: 1 mg/ml) 3) goat anti-mouse AF-647 (Invitrogen); Catalog number: A-21235; polyclonal, 1:500 dilution 4) goat anti-rabbit AF-488 (Life Technologies, catalog number A21206, polyclonal, 1:500 dilution

Validation	The used antibodies are validated for the purposes of DNA double strand break/interstrand crosslink detection and secondary detection of a primary mouse/rabbit antibody by the supplier.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	The organoid lines in use (STEM0072/ASC donor 5-a, STEM0076/ASC donor 6-a) was derived by intestinal endoscopic biopsy as indicated in reference 32. The patient's informed consent was obtained and the study was approved by the ethics committee of the University Medical Center Utrecht.
Authentication	Organoid lines were authenticated based on their WGS results and compared against previous studies such as reference 15.
Mycoplasma contamination	All organoid lines used in this study were regularly assessed for mycoplasma presence and scored negatively without exception.
Commonly misidentified lines (See ICLAC register)	There are no reports of the organoid line in use being misidentified.

Human research participants

Policy information about studies involving human research participants

Population characteristics	The organoid lines were derived as previously described (reference 32), ethical approval was obtained from the ethics committees of the University Medical Center Utrecht. WGS data from 3668 patients with metastatic cancer was analyzed with approval from the ethics board of the Hartwig Medical Foundation. WGS data from 2208 patients with predominantly primary colorectal cancer was analyzed with approval from Genomics England Ltd. Written informed consent was obtained from patients.

Method-specific reporting

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ Magnetic resonance imaging

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Organoids were dissociated using TrypLE to single cells and incubated with DMEM containing DAPI for at least 15 minutes.
Instrument	Cells were sorted with a BD FACSCanto flow sorter.
Software	Data was collected using BD FACSDiva and analyzed using FlowJo (version v10)
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	First gate: SSC-A vs. FSC-A (select epithelial cells in P1) Second gate: FSC-A vs DAPI (select for live, DAPI negative cells)

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.