

Microbiome confounders and quantitative profiling challenge predicted microbial targets in colorectal cancer development

Received: 18 November 2022

Accepted: 29 March 2024

Published online: 30 April 2024

 Check for updates

Raúl Y. Tito ^{1,2,11}, Sara Verbandt^{3,11}, Marta Aguirre Vazquez³, Leo Lahti ^{1,4}, Chloe Verspecht^{1,2}, Verónica Lloréns-Rico^{1,2,5}, Sara Vieira-Silva ^{1,6,7}, Janine Arts⁸, Gwen Falony^{1,2,6}, Evelien Dekker⁹, Joke Reumers ¹⁰, Sabine Tejpar ^{3,11} & Jeroen Raes ^{1,2,11} 

Despite substantial progress in cancer microbiome research, recognized confounders and advances in absolute microbiome quantification remain underused; this raises concerns regarding potential spurious associations. Here we study the fecal microbiota of 589 patients at different colorectal cancer (CRC) stages and compare observations with up to 15 published studies (4,439 patients and controls total). Using quantitative microbiome profiling based on 16S ribosomal RNA amplicon sequencing, combined with rigorous confounder control, we identified transit time, fecal calprotectin (intestinal inflammation) and body mass index as primary microbial covariates, superseding variance explained by CRC diagnostic groups. Well-established microbiome CRC targets, such as *Fusobacterium nucleatum*, did not significantly associate with CRC diagnostic groups (healthy, adenoma and carcinoma) when controlling for these covariates. In contrast, the associations of *Anaerococcus vaginalis*, *Dialister pneumosintes*, *Parvimonas micra*, *Peptostreptococcus anaerobius*, *Porphyromonas asaccharolytica* and *Prevotella intermedia* remained robust, highlighting their future target potential. Finally, control individuals (age 22–80 years, mean 57.7 years, standard deviation 11.3) meeting criteria for colonoscopy (for example, through a positive fecal immunochemical test) but without colonic lesions are enriched for the dysbiotic Bacteroides2 enterotype, emphasizing uncertainties in defining healthy controls in cancer microbiome research. Together, these results indicate the importance of quantitative microbiome profiling and covariate control for biomarker identification in CRC microbiome studies.

Colorectal cancer (CRC) incidence is steadily increasing¹, especially in people under 50 years². It is estimated that approximately 16 and approximately 14 individuals per 100,000 people in the United States and Belgium, respectively, die every year from CRC³. As medical

interventions can effectively reduce CRC progression and associated mortality, it is imperative to identify individuals at increased risk.

Colonoscopies with polypectomy of adenomas reduce up to 90% of CRC risk⁴. Early identification of individuals with polyps would

reduce the global burden of CRC. Yet, ascertainment of patients at an increased risk remains challenging, highlighting the need for population-wide screening.

Microbiota shifts have been associated with a wide array of disease phenotypes⁵. Some bacterial markers, such as *Fusobacterium*, have been reported enriched in lesions and stools of patients with CRC^{6–14} across developing and developed countries¹⁵, suggesting a potential role for microbiome-based diagnostics and/or prognostics.

Although microbiome profiles are affected by multiple variables that may confound or compound biological phenomena, covariate control is far from standard. For example, moisture content, a proxy for transit time, remains uncontrolled despite showing the biggest explanatory power for overall gut microbiota variation in multiple cohorts^{16,17}. Intestinal inflammation, measured as fecal calprotectin^{18,19} that reflects increased neutrophil shedding into the intestinal lumen²⁰, is more sensitive than fecal occult blood for identifying patients with CRC²¹, thus a potential untapped target for molecular stool CRC-screening¹⁹.

Relative microbiome profiling (RMP, taxon abundances are expressed in percentages) remains the dominant approach in microbiome research. However, given issues with compositionality²² and interpretation of relative profiles²³, the use of experimental and quantitative approaches is increasingly recommended^{23–25}. This reduces both false-positive and false-negative rates in downstream analyses, thereby lowering the risk of erroneous interpretation of microbiome associations, and allows focusing clinical programs on biologically relevant targets²⁵. Although quantitative microbiome profiling (QMP) facilitates normalized comparisons across different samples or conditions^{24,25}, so far, no QMP CRC microbiota studies were performed.

In this Article, we address these two gaps in CRC microbiota studies: (1) to quantitatively characterize the microbiota profile associated with malignant colonic transformation and (2) to identify microbiota covariates that may obscure biological phenomena behind microbiota-CRC associations. To this end, we examined the microbial profiles of 589 Belgian patients from Universitair Ziekenhuis Leuven (UZL) who warranted colonoscopies based on clinical presentations, including patients with CRC, and compared these to existing published datasets (total $n = 4,439$ patients and controls). To the best of our knowledge, this is the first large scale study of the gut microbiota across colonic cancer developmental stages that combines QMP analysis with extensive analysis of microbiota covariates to disentangle disease-associated from confounder-based signals to identify taxa specifically associated with CRC.

Results

Intestinal inflammation is higher in patients with colorectal tumors

We recruited 650 volunteers referred for colonoscopy and colonic resections at UZL between 2017 and 2018 who provided a stool sample before the colonic procedure. Most participants were from the Flemish region of Belgium. For this study, cancer developmental stages were

defined as diagnosis groups, and we classified participants into three groups according to a thorough colonoscopy and clinical assessment: (1) patients without evidence of colonic lesions (CTLs, $n = 205$), (2) patients with polyps (considering polyps as a precancerous lesion; $n < 10$ and size between 6 and 10 mm) (ADE, $n = 337$) and (3) patients with CRC ($n = 47$; 2 (4%) stage 0, 14 (30%) stage I, 13 (28%) stage II, 11 (23%) stage III, 3 (6%) stage IV and 4 (9%) of undetermined stage). We excluded patients outside these criteria, as well as those with insufficient clinical and molecular data. The final Leuven CRC Progression Microbiome (LCPM) study cohort consisted of 589 patients. The most frequent indications for colonoscopy were either a positive fecal immunochemical test (FIT) or adenoma surveillance. Other indications included familial risk, abdominal symptoms and change in bowel habits (Fig. 1a and Supplementary Table 1). The study was registered at clinicaltrials.gov (NCT02947607).

We collected an extensive set of 165 universal metadata variables (nonspecific for any of the three groups) from each participant. After curation, we excluded variables that were colinear (if Pearson $|r| > 0.8$, we kept the variable with fewer missing data) or had incomplete data collection (variables missing more than 20% of the values). The final set consisted of 95 high-quality variables (Supplementary Table 2).

To identify metadata variables associated with diagnosis groups, we applied two statistical approaches: (1) nonparametric Kruskal–Wallis (KW) test and its η^2 effect size (Supplementary Table 3) for all numerical variables and (2) chi-square (CS) tests and Cramer's V effect size (CV) (Supplementary Table 4) for categorical variables, followed by the Benjamini–Hochberg method for multiple testing correction (adjusted P). We found eight variables associated with diagnosis groups (false discovery rate $< 5\%$), namely: age, body mass index (BMI), calprotectin, reported hours of sleep, previous cancer (including CRC), dental status (complete, partial and so on), diabetes treatment and high blood pressure (Supplementary Tables 3 and 4). The CTL patients were younger ($n = 589$, KW test, $\eta^2 = 0.058$, $\chi^2 = 35.77$, adjusted $P = 2.6 \times 10^{-7}$; post hoc Dunn (pHD) tests, adjusted $P < 0.05$ for CTL versus ADE or CRC groups), had a lower BMI ($n = 553$, KW test, $\eta^2 = 0.023$, $\chi^2 = 15.73$, adjusted $P = 1.9 \times 10^{-3}$; pHD tests, adjusted $P < 0.05$ for CTL versus ADE) and reported fewer hours of sleep than participants from the other two diagnosis groups ($n = 557$, KW test, $\eta^2 = 0.019$, $\chi^2 = 13.41$, adjusted $P = 4.6 \times 10^{-3}$; pHD tests, adjusted $P < 0.05$ for CTL versus ADE; Fig. 1; see Supplementary Table 3 for full results). Moisture content, an important microbiota covariate¹⁶, was not significant across diagnosis groups ($n = 589$, KW test, $\eta^2 = -0.001$, $\chi^2 = 1.32$, adjusted $P = 7.0 \times 10^{-1}$).

The calprotectin levels were positively associated with malignant transformation. The patients with CRC showed higher intestinal inflammation, measured by fecal calprotectin^{18,26} (Fig. 1a and Supplementary Table 3). Specifically, CRC exhibited higher levels ($219.42 \mu\text{g g}^{-1}$, range 2.74–1,114.42, $n = 47$) compared to ADE ($70.24 \mu\text{g g}^{-1}$, range 1.87–487.21, $n = 337$) or CTL ($73.25 \mu\text{g g}^{-1}$, range 2.42–884.82, $n = 202$) (Fig. 1a, $N = 583$, KW test, $\eta^2 = 0.047$, $\chi^2 = 29.43$, adjusted $P = 3.0 \times 10^{-6}$; pHD tests, adjusted $P < 0.05$ for CRC versus CTL and CRC versus ADE).

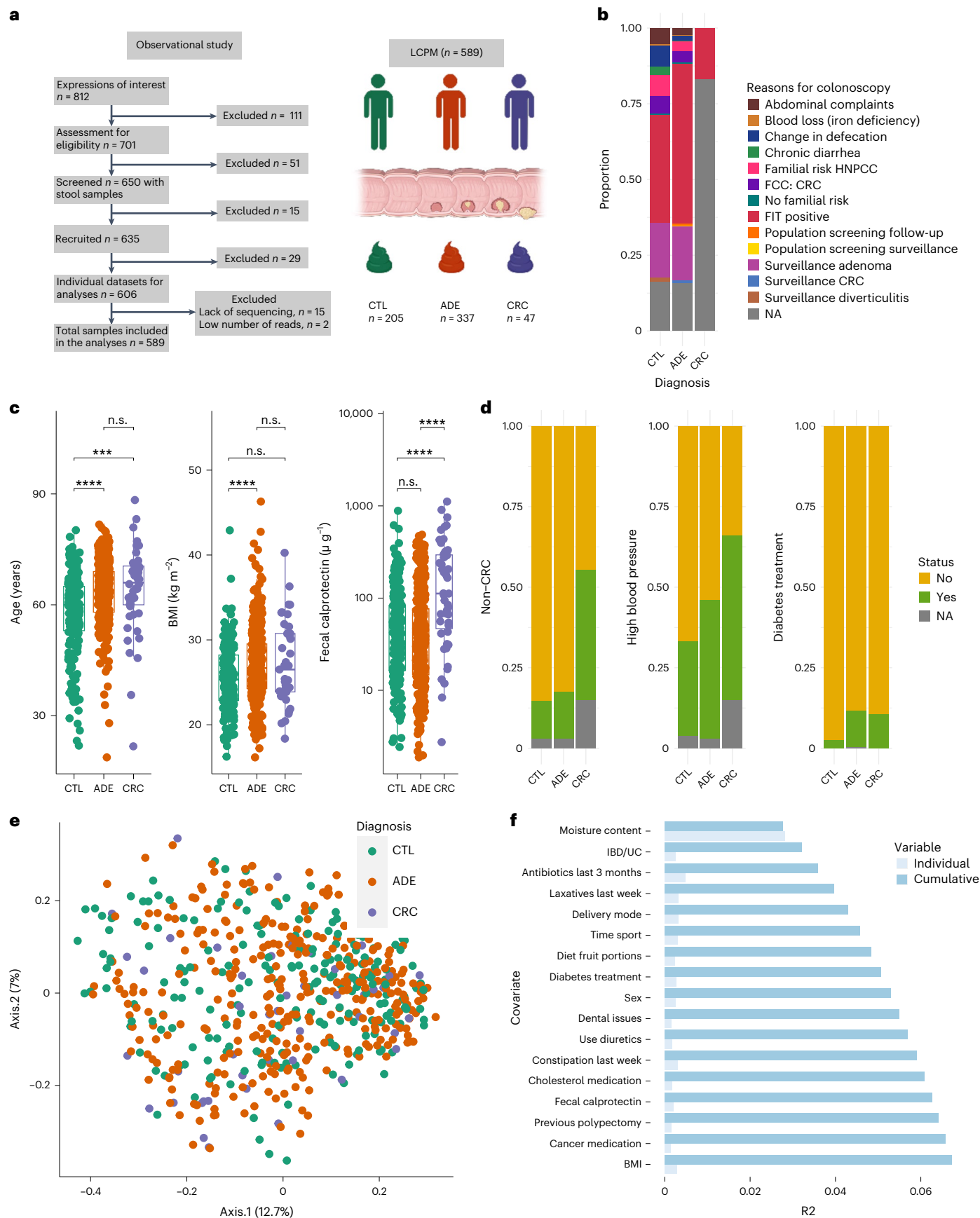
Fig. 1 | The LCPM cohort and gut microbiota covariates in CRC progression.

a, STROBE flowchart and cohort size. CTL represents patients without colonic lesions, ADE denotes patients with colonic polyps and CRC refers to patients with colorectal tumors (generated in [BioRender.com](https://www.biorender.com)). **b**, Colonoscopy referral reasons for patients of the LCPM cohort: positive FIT, adenoma surveillance, familial risk cancer (FCC), hereditary nonpolyposis CRC (HNPCC) and changes in defecation. NA, denotes the proportion of patients without information. **c**, Age, BMI and calprotectin are associated with diagnosis groups. The patients without lesions were younger ($n = 589$, two-sided KW test $\chi^2 = 35.77$, adjusted $P = 2.6 \times 10^{-7}$; pHD tests) and had lower BMI ($n = 553$, two-sided KW test $\chi^2 = 15.73$, adjusted $P = 1.9 \times 10^{-3}$; pHD tests), while patients with tumors had higher fecal calprotectin levels ($n = 583$, two-sided KW test $\chi^2 = 29.43$, adjusted $P = 3.0 \times 10^{-6}$; pHD tests, adjusted $P < 0.001$, $^{***}P < 0.01$, $^{*}P < 0.05$ and n.s., non-significant $P > 0.05$; Supplementary Table 3). The box plot center represents the median value whiskers extend from the quartiles to the last data point within 1.5 times of the interquartile range, with outliers beyond.

d, Previous non-CRC cancer, high blood pressure and diabetes treatment are associated with the distribution of diagnosis groups. The patients with CRC have a higher proportion of previous cancer (47.5% versus 15.0% and 12.1%, two-sided CS test, CV effect size of 0.24, $\chi^2 = 31.65$, d.f. of 2, adjusted $P = 1.98 \times 10^{-2}$) and high blood pressure (60.0% versus 44.3% and 30.5%, CV of 0.17, two-sided CS test, $\chi^2 = 16.55$, d.f. of 2, adjusted $P = 1.98 \times 10^{-2}$) while the CTL group has the lowest proportion of patients with diabetes treatment (2.4% versus 10.3 and 10.6, two-sided CV effect size of 0.15, CS test, $\chi^2 = 13.79$, d.f. of 2, adjusted $P = 1.98 \times 10^{-2}$). **e**, PCoA on BCD representing QMP species-level microbiota variation in the LCPM cohort ($n = 589$), PCoA1 (Axis.1) and PCoA2 (Axis.2) respectively explained 12.7% and 7% of the variance. Each dot represents one sample, colored by assigned diagnosis group. **f**, Cumulative effect sizes of significant covariates on microbiota community variation (cumulative bars; stepwise dbRDA on BCD) as compared to individual effect sizes (R^2) assuming covariate independence in the LCPM cohort ($n = 589$; Supplementary Table 5). UC, ulcerative colitis.

We also observed increased fecal calprotectin in patients reporting previous cancers (primarily breast and prostate cancer) (Wilcoxon ranksum (WR) test, $W = 11,067$, adjusted $P = 4.1 \times 10^{-3}$), consumption of

cancer medication (WR test, $W = 3,671$, adjusted $P < 0.05$), heartburn complaints (WR test, $W = 11,067$, adjusted $P = 1.0 \times 10^{-10}$) and lower dietary fiber (WR test, $W = 20,964$, adjusted $P = 3.3 \times 10^{-2}$).



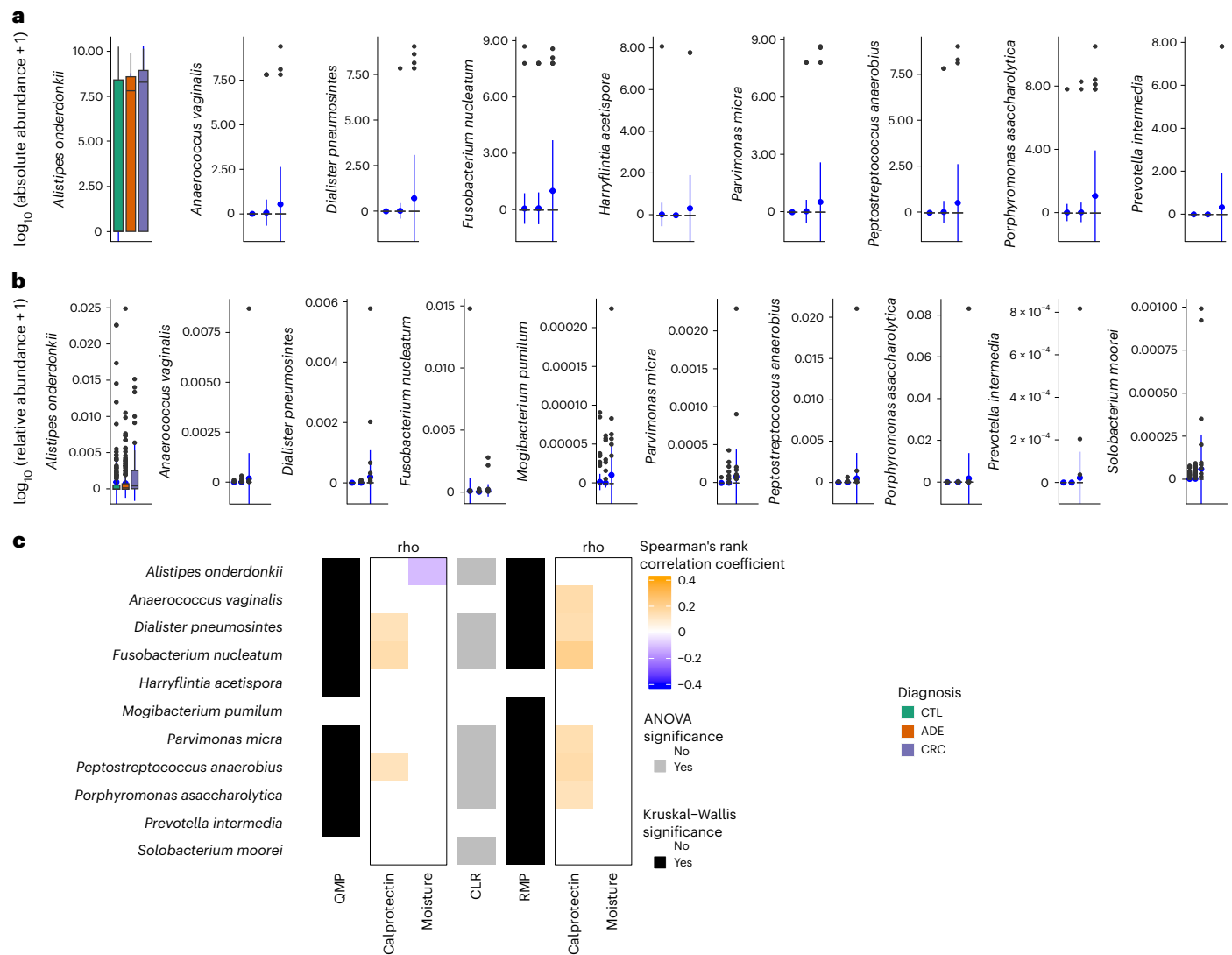


Fig. 2 | Microbial biomarkers in CRC progression. a, Nine species were identified with differential absolute abundance across diagnosis groups ($n = 589$, KW test, adjusted $P < 0.05$; Supplementary Table 7). **b**, Ten species were identified with differential relative abundance across diagnosis groups ($n = 589$, KW test, adjusted $P < 0.05$; Supplementary Table 7). The center of the box plot represents the median value of the data, and the whiskers extend from the quartiles to the last data point within 1.5 times of the interquartile range, with outliers beyond. The blue circles represent the mean. **c**, Biomarkers associations and their confounders. Species Spearman's rank correlation with calprotectin levels and moisture proportions

using QMP (first rho column panel) and RMP (second rho column panel) data. The effect size of the associations between species and calprotectin, moisture and diagnosis variables for QMP and RMP ($n = 589$, Spearman's rank correlation comparison, adjusted $P < 0.05$). Significant associations were tested using two-sided KW tests for QMP and RMP data and ANOVA for CLR data. The associations for *Herryflintia acetispora*, *Parvimonas micra* and *Prevotella intermedia* are sensitive to bias by the extreme values (absolute abundance) in the higher range. Removing these values leads to loss of significance. As rank-based approaches were used, it is not clear if this loss is due to the strength of the signal or the loss of power.

The history of chronic diseases was distinct across diagnosis groups. The patients with CRC showed higher proportions of previous non-CRC cancer (47.5% versus 15.0% and 12.1%, CS test, CV of 0.24, $\chi^2 = 31.65$, d.f. of 2, adjusted $P = 1.98 \times 10^{-2}$) and high blood pressure (60.0% versus 44.3% and 30.5%, CS test, CV of 0.17, $\chi^2 = 16.55$, d.f. of 2, adjusted $P = 1.98 \times 10^{-2}$) (Fig. 1b and Supplementary Table 4). The CTL group had the lowest diabetes treatment (2.4% versus 10.3% and 10.6%, CS test, CV of 0.15, $\chi^2 = 13.79$, d.f. of 2, adjusted $P = 1.98 \times 10^{-2}$) (Fig. 1b and Supplementary Table 4) and mostly complete dental sets (53.3% versus 35.2% and 32.5%, CS test, CV of 0.03, $\chi^2 = 30.78$, d.f. of 10, adjusted $P = 1.98 \times 10^{-2}$) (Supplementary Table 4).

Known confounders, not diagnosis groups, explain overall microbiota variation across CRC developmental stages

The influence of microbiota covariates and the quantitative amplitude of observed microbiota shifts are understudied in CRC. We combined

sequencing data with flow cytometry measurements of fecal microbial load²³ to generate QMP data from our study cohort.²³ We studied the QMP variation in the context of the 94 potential covariates mentioned above (the 95th being microbial load) using established procedures¹⁷.

A principal coordinate analysis (PCoA; Fig. 1c) on a species-level Bray–Curtis dissimilarity (BCD) matrix revealed no significant separation between diagnosis groups. Furthermore, no difference in total microbial load was found between groups ($n = 589$, KW test, $\chi^2 = 0.68$, adjusted $P = 8.2 \times 10^{-1}$). Distance-based redundancy analysis (dbRDA) revealed 24 microbiota covariates associated with microbial variation in this cohort (Fig. 1d and Supplementary Table 5). We identified 17 nonredundant covariates that jointly explained 6.7% of microbiota compositional variation (Supplementary Table 5).

Consistent with previous reports^{16,17}, moisture content exhibited the highest explanatory value (2.8%) of all covariates ($n = 589$, stepwise dbRDA, $R^2 = 2.8\%$, adjusted $P = 2 \times 10^{-3}$). Intestinal bowel disease/

ulcerative colitis (IBD/UC) status, a CRC-risk factor, possibly associated with its microbial dysbiotic community and intestinal inflammation²⁷, was the second largest covariate. IBD/UC explained 0.4% of the microbiota variation ($n = 569$, stepwise dbRDA, $R^2 = 0.4\%$, adjusted $P = 2 \times 10^{-3}$). Other top microbiota covariates included antibiotics and laxatives use (Fig. 1d). Delivery mode (cesarean or natural birth) explained 0.3% variation ($n = 533$, stepwise dbRDA, $R^2 = 0.3\%$, adjusted $P = 2 \times 10^{-3}$), although it is probably confounded by diet in this cohort (proportion of dietary vegetables; CS test, $\chi^2 = 33.09$, d.f. of 14, $P = 2.8 \times 10^{-3}$, adjusted $P < 0.05$). Intestinal inflammation (fecal calprotectin) explained 0.2% ($n = 583$, stepwise dbRDA, $R^2 = 0.2\%$, adjusted $P = 2.6 \times 10^{-2}$). In contrast with our previous study in the Flemish population (Flemish Gut Flora Project, FGFP)¹⁷, age did not explain microbiota variation ($n = 589$, univariate dbRDA, $R^2 = 0.2\%$, adjusted $P = 5.9 \times 10^{-2}$). Surprisingly, the cancer diagnosis group (CTL, ADE and CRC), as a covariate, was not associated with microbial variation ($n = 589$, univariate dbRDA, $R^2 = 0.2\%$, adjusted $P = 0.22$; Supplementary Table 5).

Fusobacterium association with CRC stages disappears when controlling for confounders or when using QMP

Microbiota signals can be specific to taxonomic groups and, thus, not reflected in broad community shifts. While a multitude of microbial associations have been reported in CRC studies using RMP^{6–8,13}, we used QMP to identify species whose absolute abundance associated with diagnosis groups. The comparisons were limited to the 138 species with a prevalence of greater than 5% in at least one of the diagnosis groups of the LCPM cohort (Supplementary Table 6). Only eight species showed significant differential abundance (absolute or relative) among diagnosis groups: *Anaerococcus vaginalis* (*Anaerococcus obeisensis*), *Alistipes onderdonkii*, *Dialister pneumosintes*, *Fusobacterium nucleatum*, *Parvimonas micra*, *Peptostreptococcus anaerobius*, *Porphyromonas asaccharolytica* and *Prevotella intermedia* (KW test, adjusted $P < 0.05$; Fig. 2a,b and Supplementary Table 7). While *Fusobacterium nucleatum* has been consistently associated with colorectal lesions across cohorts of diverse backgrounds^{13,14}, in the LCPM cohort, *Fusobacterium nucleatum* absolute abundance was positively correlated with high fecal calprotectin levels (Spearman's rank and Kendall's tau correlations, adjusted $P < 0.05$; Fig. 2c, Extended Data Fig. 1 and Supplementary Table 8) and cancer progression (diagnosis groups) (KW test, $\eta^2 = 0.010$, adjusted $P = 1.84 \times 10^{-5}$; pH test adjusted $P = 8.80 \times 10^{-1}$ for CTL versus ADE, adjusted $P = 3.84 \times 10^{-7}$ for CTL versus CRC and adjusted $P = 3.84 \times 10^{-7}$ for ADE versus CRC; Fig. 2c and Supplementary Table 7). However, after deconfounding for calprotectin only or combined BMI, moisture content and calprotectin, and neither absolute nor relative *Fusobacterium nucleatum* abundance were associated with diagnosis (generalized linear model analysis of variance (ANOVA), $n = 547$, $P > 0.05$; Extended Data Fig. 2).

Multiple established CRC microbial markers are associated with transit time, intestinal inflammation and body mass index but not with CRC stages

The association of *Fusobacterium* abundance with fecal calprotectin urged us to investigate the influence of this confounder on previously reported CRC-associated genera, adding moisture content since it is the top microbiome covariate, and BMI, which showed differences among diagnosis groups.

To this end, we compiled a list of 89 CRC species-level markers from ten published cohorts^{6,9,11,13,14,28–31} (including 1,633 samples) and 67 genera-level markers from 15 cohorts^{6–9,11–15,28–32} (representing 4,439 samples). We used this compiled list of taxa as a criterion to test whether the CRC association of these taxa in our cohort is influenced by the target covariates. To reduce the impact of distinct statistical treatments, we downloaded the microbial profiles of nine out of ten studies at species level from the curated MetagenomicData³³ resource and analyzed them using the statistical component of our pipeline.

Spearman correlation between taxa abundances and the three focus covariates revealed strong associations between microbial targets and these confounders at the species (Extended Data Fig. 3a) and genus level (Fig. 3b). Most of these associations were replicated in an independent population cohort (FGFP), suggesting these associations are robust and not specifically linked to CRC (Extended Data Fig. 3). Moisture content, the known major covariate in microbiome studies¹⁷, is unsurprisingly associated with many taxa validated in both cohorts.

As we compiled the CRC-associated taxa from non-QMP studies, we conducted analyses using both RMP and QMP to assess whether confounder associations influence quantitative association of biomarkers or targets to diagnosis groups in LCPM. We found only 8% (6 out of 89) and 10% (9 out of 89) of species previously associated with CRC using QMP and RMP replicating after confounder control. *Anaerococcus vaginalis*, *Dialister pneumosintes*, *Parvimonas micra*, *Peptostreptococcus anaerobius*, *Prevotella intermedia* and *Porphyromonas asaccharolytica*, were identified by controlled QMP and RMP. Controlled QMP excluded *Fusobacterium nucleatum* and *Alistipes onderdonkii*, suggesting previous associations of these two species may be spurious (Fig. 3a).

We identified eight species previously linked to CRC (that is, using QMP and/or RMP), including *Fusobacterium nucleatum* and *Peptostreptococcus anaerobius*, to be associated with inflammation (Fig. 3 and Supplementary Tables 8 and 9). This association was previously reported for only three out of the eight taxa above (*Escherichia*, *Fusobacterium* and *Streptococcus*)²⁴. Further validation of this association was conducted using the FGFP (Extended Data Fig. 3 and Supplementary Tables 8 and 9).

Recognizing that inflammation is a risk factor, not a requirement, for CRC progression, we further investigated markers associated with diagnosis groups in relation to inflammatory status. To this end, we focused on a subset of 340 samples, which, regardless of their CRC status, exhibited normal levels of calprotectin (fecal calprotectin under $50 \mu\text{g g}^{-1}$ (ref. 34)), indicating no evidence of local inflammation (112 CTL, 216 ADE and 12 CRC). Assessment of the 89 CRC species-level markers mentioned above confirmed that the association of three of the six replicating species (*Anaerococcus vaginalis*, *Prevotella intermedia* and *Porphyromonas asaccharolytica*) is independent of intestinal inflammation (Supplementary Table 10).

Colonoscopy patients, with or without CRC, exhibit an excess of the Bacteroides2 enterotype

To study the LCPM cohort in a population context, we enterotyped participants using Dirichlet multinomial mixtures (DMM) on a genus matrix against the background of microbial variation as observed in the FGFP samples ($n = 1,045$)¹⁷. Consistent with previous description of the Flemish population²³, we identified four community types based on selecting the optimal number of clusters using the Bayesian Information Criterion (Fig. 4a,b and Extended Data Fig. 4), 'Bacteroides1' (Bact1), 'Bacteroides2' (Bact2), 'Prevotella' (Prev) and 'Ruminococcaceae' (Rum). The enterotype distribution was different between LCPM and FGFP (CS test, $\chi^2 = 34.3$, d.f. of 3, adjusted $P = 1.7 \times 10^{-7}$), but no differences were observed among diagnosis groups within the LCPM cohort (pairwise CS tests, adjusted $P > 0.1$). Pairwise comparisons of the prevalence of the dysbiotic Bact2 enterotype in the LCPM cohort diagnosis groups revealed that compared to the FGFP population, this enterotype was enriched in all CRC diagnosis groups (test of equal or given proportions, FGFP versus CTL: $\chi^2 = 15.09$, d.f. of 1, adjusted $P = 1.1 \times 10^{-4}$; FGFP versus ADE: $\chi^2 = 18.93$, d.f. of 1, adjusted $P = 2.4 \times 10^{-5}$; and FGFP versus CRC: $\chi^2 = 4.34$, d.f. of 1, adjusted $P = 3.4 \times 10^{-2}$). Although dysbiosis and CRC development were previously linked^{13,35}, the high prevalence of this enterotype in the LCPM, even in samples from patients free of lesions, is unexpected. Consistent with previous reports^{24,25}, the Bact2 enterotype in this group exhibited all hallmarks of dysbiosis: low cell count, low richness, higher calprotectin values, reduced butyrate producers and increased proinflammatory bacteria.



Fig. 3 | BMI, intestinal inflammation and moisture correlations with microbial biomarkers and CRC. a, b. Species (a) and genera (b) previously reported in association with CRC (blue and green represent enrichment or depletion; the squares indicate reported in corresponding publications, while circles represent our reanalysis of the MetaPhlan 3.0 profiles generated from the curatedMetagenomicData³³ of these cohorts using the statistical part of our pipeline). Graphic representation of Spearman's rank correlation of pairwise analysis of fecal calprotectin, BMI, and moisture values against absolute species

abundance (QMP) and RMP from the LCPM ($N = 589$) and FGFP ($N = 1,045$) cohorts (adjusted $P < 0.05$, Supplementary Table 8). The species enriched or depleted in relation to CRC diagnosis groups were tested using QMP, CLR and RMP data before ($n = 589$, two-sided KW test and Spearman's rank correlation comparison, adjusted $P < 0.05$) and after controlling for microbiota covariates (before adjustment for BMI, calprotectin and moisture; generalized linear model ANOVA, adjusted $P < 0.05$).

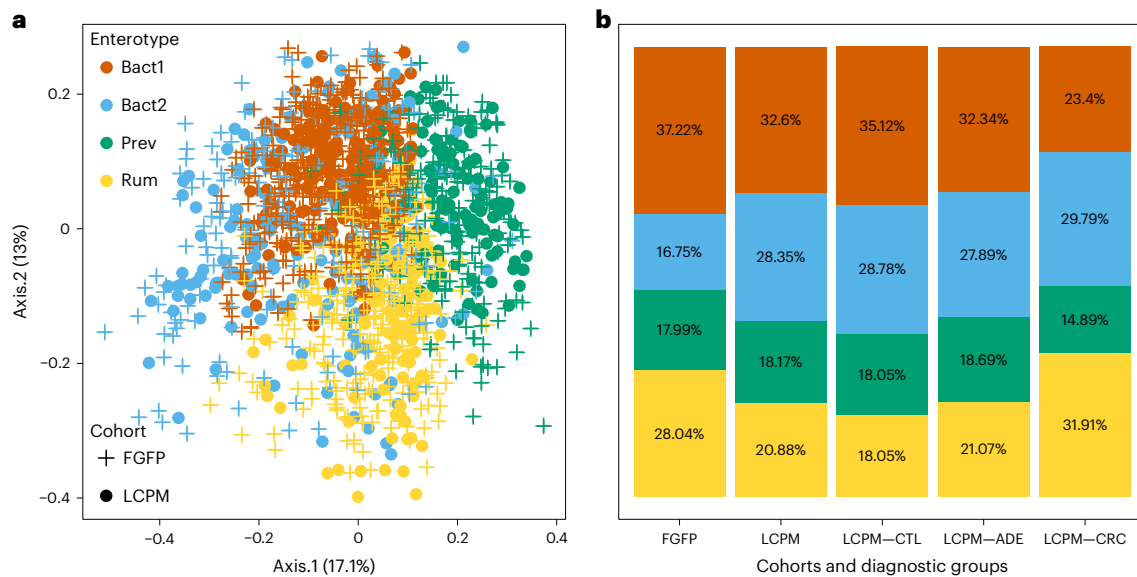


Fig. 4 | The Bact2 enterotype is enriched in patients referred for a colonoscopy (with and without colorectal lesions). **a**, PCoA of interindividual differences (BCD) in relative microbiota profiles of the LCPM cohort ($n = 589$ samples) using a cross-section of the Flemish population ($n = 1,045$ samples) as a background dataset. PCoA1 (Axis.1) and PCoA2 (Axis.2) respectively explained

13% and 17.1% of the variance of microbiota at the genus level. **b**, Enterotype distribution across the FGFP, LCPM and LCPM diagnosis groups (CTL, ADE and CRC), increased prevalence of the Bact2 enterotype in the three groups from the LCPM cohort ($n = 589$) as compared to FGFP samples ($n = 1,045$); pairwise two-sided test of equal or given proportions ($P < 0.05$).

Additional categorical variables appeared associated with the Bact2 enterotype. They included antibiotic consumption (CS test, $\chi^2 = 30.78$, d.f. of 3, adjusted $P = 2.1 \times 10^{-2}$), current treatment with anti-inflammatory medications (CS test, $\chi^2 = 30.78$, d.f. of 3, adjusted $P = 2.1 \times 10^{-2}$), diabetes treatment (CS test, $\chi^2 = 30.78$, d.f. of 3, adjusted $P = 3.3 \times 10^{-2}$), recent diarrhea (last week) (CS test, $\chi^2 = 30.78$, d.f. of 3, adjusted $P = 2.1 \times 10^{-2}$), history of gallstones (CS test, $\chi^2 = 30.78$, d.f. of 3, adjusted $P = 4.7 \times 10^{-2}$) and recent use of laxatives (last week) ($\chi^2 = 30.78$, d.f. of 3, adjusted $P = 4.2 \times 10^{-2}$) (Supplementary Table 11).

Discussion

While associations between the gut microbiota and CRC have been extensive, this is the first study using QMP and extensive metadata collection to systematically investigate microbiota covariates that potentially are masking or creating spurious associations between specific taxa and malignant transformation.

At first glance, this study yielded a gut microbial profile partially consistent with previous reports of CRC-associated taxa. Further analysis, however, suggested that many of the previously reported associations, including those of prominent biomarkers, such as *Fusobacterium nucleatum*, are confounded by microbiota covariates. A total of 17 of 94 variables explained 6.7% of the observed variation. Of those, the moisture content had highest explanatory power (2.7%), greater than eight times that of the next covariate (IBD status). The explanatory power of fecal calprotectin was lower (0.2%) but significant; age and, most importantly, diagnosis groups were not.

Some associations were complex in nature. For example, BMI, consistent with previous reports, showed an association with both microbial composition^{17,25} and cancer progression³⁶, while others, such as age, suggested to modify the BMI-association with cancer progression³⁷, were not significant in this cohort.

Inflammation is a known risk factor for CRC³⁸, but its effect size in shaping the cancer-associated microbiota is yet to be described. Fecal calprotectin is a well-documented marker of intestinal local inflammation^{39,40} and has been associated with cancer progression, probably having an effect on tumor development rather than on tumor initiation⁴¹. We observed participants with normal and elevated fecal calprotectin levels within each diagnosis group and covariate-controlled analysis of

the LCPM cohort revealed that 8 and 19 CRC-associated markers, at the species and genus levels, respectively, associated with fecal calprotectin rather than with the diagnosis group. We replicated these observations in an independent cohort of apparently healthy individuals (FGFP).

High levels of fecal calprotectin have been associated with intestinal inflammatory pathologies¹⁹. However, when removing patients with IBD from our analysis, CRC diagnosis groups remained not significant, and the significance of *Fusobacterium nucleatum*, among other six species, was unaltered after differential abundance analysis. In patients with CRC, increased levels of fecal calprotectin ($>50 \mu\text{g g}^{-1}$ stool^{18,26}) are directly associated with tumor presence, as the level decreases after tumor resection⁴². Here, fecal calprotectin was increased in CRC, consistent with previous associations between malignant transformation, local inflammation⁴³ and advanced tumor stages (T3 and T4)⁴². No difference in calprotectin levels was observed between CTL and ADE (mean 73.25 versus 70.24 $\mu\text{g g}^{-1}$), suggesting that although no lesions are visible in the colon of the CTL group, they have a detectable level of local inflammation. The potential effect of local inflammation in shaping the colonic microbiota in the context of malignant transformation, or its potential confounding effect, remains largely obscure, as most studies surveying the association between gut microbiota and CRC, including meta-analysis^{13,14}, do not control for local inflammation.

We argue that strict control of covariates is a must in any microbiota analysis assessing potential clinical associations, as for example, three of the species with repeated CRC association^{11,13,14,28–30,32}, *Escherichia coli*, *Fusobacterium nucleatum* and *Parvimonas micra*, exhibit association with local inflammation, unfortunately uncontrolled for in previous studies, that may or may not be associated with cancer progression.

Fusobacterium nucleatum is one of the species that attracts more attention as there is a substantial body of work linking it to CRC⁴⁴. In this study, *Fusobacterium* was enriched in patients with CRC. However, this apparent association disappears when the analysis is covariate controlled. Our study suggests that the association of *Fusobacterium nucleatum* to cancer may be driven by its association to intestinal inflammatory conditions; there are no differences in the abundance of *Fusobacterium nucleatum* across diagnostic groups once calprotectin is controlled for. These results suggest reassessment of the diagnostic utility of this marker. At the same time, our results do not

mean that *Fusobacterium nucleatum* is not linked to CRC; they rather suggest that the reasons behind this association might be less straightforward than originally considered. They, thus, present a cautionary tale of the importance to control for covariates as the microbiome field moves forward. Given that inflammation is a risk factor for CRC but not a requirement⁴¹, potential use of *Fusobacterium nucleatum* as a marker of CRC development could fail to identify those cases of inflammation-independent cancer progression. While not yet commercialized, there are already publications proposing the use of microbial markers, including *Fusobacterium nucleatum*, for CRC screening^{7,45}, which, in light of our results, raises concerns as uncontrolled variables may be obscuring actual biological mechanisms. We present evidence that purported CRC biomarkers, even those replicated in multiple studies, may suffer from the compounding or confounding effect of covariates, which in addition to the use of nonquantitative signals, may result in misleading conclusions on what diagnostic signals really mean—complicating the path towards potential clinical applications.

BMI, in combination or independent of inflammation, has been independently associated with changes in the gut microbiota⁴⁶, which in turn are associated with increased risk of CRC⁴⁷. Yet, microbial dysbiosis by itself does not explain the higher risk of colon cancer observed in the obese population⁴⁸, indicating that the underlying process that associates obesity and CRC is more complex and demands further investigation.

Among four described gut enterotypes, the Bact2 enterotype is defined as a dysbiotic microbial profile^{24,25}. Bact2 enrichment is observed in obesity²⁵ and in conditions such as PSC (Primary sclerosing cholangitis) and IBD²⁴, further supporting the potential disease association of this enterotype. The analysis of the LCPM cohort revealed an excess of the Bact2 enterotype across all diagnosis subgroups, regardless of BMI.

Increased Bact2 prevalence in the no-lesions group compared to FGFP is particularly striking. While patients in the CTL group have no observable lesions, they may be considered at increased risk for colorectal perturbations based on clinical referrals (blood loss in the stool, familiar risk to colonic lesion and so on) that warranted colonoscopies—something that might also be reflected by their Bact2 enterotype. Of importance, ‘healthy’ biopsies included in CRC microbiome studies are often selected using colonoscopies with a negative result as the main criterium, posing a potential problem, as no other markers of colonic health are considered to qualify these healthy individuals. The reasons for the appearance of Bact2 in the no-lesion group are multifold, but these findings suggest that such individuals, while representing a useful category for biomarker discovery, may harbor an unhealthy gut ecosystem, from a microbial point of view.

There is a plethora of variables identified as modifiers of the gut microbiota. Yet, covariate control is far from standard and notably absent from most association studies. As intestinal microbial taxa are being nominated as potential biomarkers of malignant transformation, it is imperative to explore the influence of microbiota covariates as potential confounders or compounders of observed associations. Rather than denying previous associations, our analysis emphasizes the need for covariate-controlled analysis for any microbiota study aiming to establish clinical associations, as these covariates by themselves may explain most of the stool microbiota variation, independent of CRC status.

Out of the multiple taxa previously associated with CRC, six species remain significant after strict control of covariates in this quantitative cohort. Without denying other potential biomarkers, further studies are warranted on *Anaerococcus vaginalis*, *Dialister pneumosintes*, *Parvimonas micra*, *Peptostreptococcus anaerobius*, *Prevotella intermedia* and *Porphyromonas asaccharolytica*, as their reported association to CRC^{6,7} is robust enough to remain independent of the method. Our data present a strong argument in favor of revisiting potential microbial associations with clinical phenotypes to ensure

that the purported associations are not driven by uncontrolled covariates warranting further follow up of the mechanisms underlying these associations. Refining the approaches to discover microbial biomarkers will undoubtedly impact the microbiota field, facilitating the path towards the much-coveted clinical applications.

Limitations

We aim to identify taxa associated with malignant colonic transformation. While our cohort includes a set of participants without lesions, we make no claim that these are healthy controls, as there is an apparent increased incidence of gut dysbiosis in this group. Considering that all participants in this study had a medical need for a colonoscopy, there is an implicit increased risk to CRC. Thus, the present study cannot rule out that the group without polyps is undergoing potential molecular or cellular changes that are not detectable via colonoscopy. In addition, as this is a cross-sectional study, the term cancer progression is an extrapolation of what is seen at cancer development stages (operationalized here as diagnosis groups). We cannot rule out potential particularities of our cohort that may be contributing to our observations, as most studies do not report sufficient metadata for us to compare across cohorts. It is important to consider that certain taxonomic groups may not even be represented in current databases, and specific microbial species may require longer hypervariable regions or alternative sequencing approaches to achieve accurate species-level identification. Nonetheless, the V4 region for our cohort seems to be able to resolve species taxonomy of the biomarkers previously associated with CRC, as we show for the case of *Fusobacterium*.

Furthermore, it has been proposed that the potential diagnostic value of colonic microbial profiles goes beyond bacteria, as fungal and viral species have been proposed as CRC biomarkers⁴⁹. We recognize that multidomain approaches to discover CRC biomarkers and longitudinal prospective studies to better study the dynamics of cancer progression are warranted to comprehensively inform cancer detection and treatment.

Online content

Any methods, additional references, Nature Portfolio 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/s41591-024-02963-2>.

References

1. Yang, L. et al. Changes in colorectal cancer incidence by site and age from 1973 to 2015: a SEER database analysis. *Aging Clin. Exp. Res.* **33**, 1–10 (2020).
2. Keum, N. & Giovannucci, E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nat. Rev. Gastroenterol. Hepatol.* **16**, 713–732 (2019).
3. Araghi, M. et al. Global trends in colorectal cancer mortality: projections to the year 2035. *Int. J. Cancer* <https://doi.org/10.1002/ijc.32055> (2018).
4. Rex, D. K. & Eid, E. Considerations regarding the present and future roles of colonoscopy in colorectal cancer prevention. *Clin. Gastroenterol. Hepatol.* **6**, 506–514 (2008).
5. Gupta, V. K. et al. A predictive index for health status using species-level gut microbiome profiling. *Nat. Commun.* **11**, 4635 (2020).
6. Yachida, S. et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* **25**, 968–976 (2019).
7. Young, C. et al. Microbiome analysis of more than 2,000 NHS bowel cancer screening programme samples shows the potential to improve screening accuracy. *Clin. Cancer Res.* **27**, 2246–2254 (2021).

8. Clos-Garcia, M. et al. Integrative analysis of fecal metagenomics and metabolomics in colorectal cancer. *Cancers* <https://doi.org/10.3390/cancers12051142> (2020).
9. Yu, Y. N. et al. Berberine may rescue *Fusobacterium nucleatum*-induced colorectal tumorigenesis by modulating the tumor microenvironment. *Oncotarget* **6**, 32013–32026 (2015).
10. Yu, T. C. et al. *Fusobacterium nucleatum* promotes chemoresistance to colorectal cancer by modulating autophagy. *Cell* **170**, 548–563.e16 (2017).
11. He, T., Cheng, X. & Xing, C. The gut microbial diversity of colon cancer patients and the clinical significance. *Bioengineered* **12**, 7046–7060 (2021).
12. Kasai, C. et al. Comparison of human gut microbiota in control subjects and patients with colorectal carcinoma in adenoma: terminal restriction fragment length polymorphism and next-generation sequencing analyses. *Oncol. Rep.* **35**, 325–333 (2016).
13. Thomas, A. M. et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* <https://doi.org/10.1038/s41591-019-0405-7> (2019).
14. Wirbel, J. et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* <https://doi.org/10.1038/s41591-019-0406-6> (2019).
15. Young, C. et al. The colorectal cancer-associated faecal microbiome of developing countries resembles that of developed countries. *Genome Med.* **13**, 1–13 (2021).
16. Vandeputte, D. et al. Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut* **65**, 57–62 (2016).
17. Falony, G. et al. Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).
18. Poullis, A., Foster, R., Shetty, A., Fagerhol, M. K. & Mendall, M. A. Bowel inflammation as measured by fecal calprotectin: a link between lifestyle factors and colorectal cancer risk. *Cancer Epidemiol. Biomarkers Prev.* <https://doi.org/10.1158/1055-9965.EPI-03-0160> (2004).
19. Högberg, C., Karling, P., Rutegård, J. & Lilja, M. Diagnosing colorectal cancer and inflammatory bowel disease in primary care: the usefulness of tests for faecal haemoglobin, faecal calprotectin, anaemia and iron deficiency. A prospective study. *Scand. J. Gastroenterol.* **52**, 69–75 (2017).
20. Schreuders, E. H., Grobbee, E. J., Spaander, M. C. W. & Kuipers, E. J. Advances in fecal tests for colorectal cancer screening. *Curr. Treat. Options Gastroenterol.* **14**, 152–162 (2016).
21. Røseth, A. G. et al. Faecal calprotectin: a novel test for the diagnosis of colorectal cancer? *Scand. J. Gastroenterol.* **28**, 1073–1076 (1993).
22. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiomedatasets are compositional: and this is not optional. *Front. Microbiol.* **8**, 2224 (2017).
23. Vandeputte, D. et al. Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**, 507–511 (2017).
24. Vieira-Silva, S. et al. Quantitative microbiome profiling disentangles inflammation-and bile duct obstruction-associated microbiota alterations across PSC/IBD diagnoses. *Nat. Microbiol.* **4**, 1826–1831(2019).
25. Vieira-Silva, S. et al. Statin therapy is associated with lower prevalence of gut microbiota dysbiosis. *Nature* <https://doi.org/10.1038/s41586-020-2269-x> (2020).
26. Tibble, J. A. & Bjarnason, I. Fecal calprotectin as an index of intestinal inflammation. *Drugs Today* <https://doi.org/10.1358/dot.2001.37.2.614846>(2001).
27. Quaglio, A. E. V., Grillo, T. G., De Oliveira, E. C. S., Di Stasi, L. C. & Sasaki, L. Y. Gut microbiota, inflammatory bowel disease and colorectal cancer. *World J. Gastroenterol.* **28**, 4053–4060 (2022).
28. Zeller, G. et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).
29. Feng, Q. et al. Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nat. Commun.* **6**, 6528 (2015).
30. Vogtmann, E. et al. Colorectal cancer and the human gut microbiome: reproducibility with whole-genome shotgun sequencing. *PLoS ONE* **11**, e0155362 (2016).
31. Hannigan, G. D., Duhaime, M. B., Ruffin, M. T., Koumpouras, C. C. & Schloss, P. D. Diagnostic potential and interactive dynamics of the colorectal cancer virome. *mBio* **9**, e02248-18 (2018).
32. Yu, J. et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **66**, 70–78 (2017).
33. Pasolli, E. et al. Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**, 1023–1024 (2017).
34. Bjarnason, I. The use of fecal calprotectin in inflammatory bowel disease. *Gastroenterol. Hepatol.* **13**, 53–56 (2017).
35. Dai, Z. et al. Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. *Microbiome* <https://doi.org/10.1186/s40168-018-0451-2> (2018).
36. Zheng, R. et al. Body mass index (BMI) trajectories and risk of colorectal cancer in the PLCO cohort. *Br. J. Cancer* **119**, 130–132 (2018).
37. Carr, P. R. et al. Association of BMI and major molecular pathological markers of colorectal cancer in men and women. *Am. J. Clin. Nutr.* <https://doi.org/10.1093/ajcn/nqz315> (2020).
38. Rutter, M. et al. Severity of inflammation is a risk factor for colorectal neoplasia in ulcerative colitis. *Gastroenterology* **126**, 451–459 (2004).
39. Costa, F. et al. Role of faecal calprotectin as non-invasive marker of intestinal inflammation. *Digest. Liver Dis.* **35**, 642–647 (2003).
40. Konikoff, M. R. & Denson, L. A. Role of fecal calprotectin as a biomarker of intestinal inflammation in inflammatory bowel disease. *Inflamm. Bowel Dis.* <https://doi.org/10.1097/OO054725-200606000-00013>(2006).
41. Terzić, J., Grivennikov, S., Karin, E. & Karin, M. Inflammation and colon cancer. *Gastroenterology* **138**, 2101–2114 (2010).
42. Lehmann, F. S. et al. Clinical and histopathological correlations of fecal calprotectin release in colorectal carcinoma. *World J. Gastroenterol.* <https://doi.org/10.3748/wjg.v20.i17.4994> (2014).
43. Pathirana, W. G. W., Chubb, S. P., Gillett, M. J., & Vasikaran, S. D. Faecal calprotectin. *Clin. Biochem. Rev.* <https://doi.org/10.1097/mpg.0000000000001847> (2018).
44. Bullman, S. et al. Analysis of *Fusobacterium* persistence and antibiotic response in colorectal cancer. *Science* **358**, 1443–1448 (2017).
45. Osman, M. A. et al. *Parvimonas micra*, *Peptostreptococcus stomatis*, *Fusobacterium nucleatum* and *Akkermansia muciniphila* as a four-bacteria biomarker panel of colorectal cancer. *Sci. Rep.* **11**, 1–12 (2021).
46. Turnbaugh, P. J. et al. A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484 (2009).
47. Moghaddam, A. A., Woodward, M. & Huxley, R. Obesity and risk of colorectal cancer: a meta-analysis of 31 studies with 70,000 events. *Cancer Epidemiol. Biomarkers Prev.* **16**, 2533–2547 (2007).
48. Greathouse, K. L. et al. Gut microbiome meta-analysis reveals dysbiosis is independent of body mass index in predicting risk of obesity-associated CRC. *BMJ Open Gastroenterol.* <https://doi.org/10.1136/bmjgast-2018-000247> (2019).
49. Liu, N. N. et al. Multi-kingdom microbiota analyses identify bacterial–fungal interactions and biomarkers of colorectal cancer across cohorts. *Nat. Microbiol.* **7**, 238–250 (2022).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this

article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024

¹Laboratory of Molecular Bacteriology, Department of Microbiology and Immunology, Rega Institute, Katholieke Universiteit Leuven, Leuven, Belgium.

²Center for Microbiology, Vlaams Instituut voor Biotechnologie, Leuven, Belgium. ³Digestive Oncology, Department of Oncology, Katholieke Universiteit Leuven, Leuven, Belgium.

⁴Department of Computing, University of Turku, Turku, Finland. ⁵Systems Biology of Host–Microbiome Interactions Laboratory, Principe Felipe Research Center (CIPF), Valencia, Spain.

⁶Institute of Medical Microbiology and Hygiene and Research Center for Immunotherapy, University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany.

⁷Institute of Molecular Biology, Mainz, Germany. ⁸Oncology, Janssen Pharmaceutica NV, Beerse, Belgium.

⁹Department of Gastroenterology and Hepatology, Amsterdam University Medical Centers, Amsterdam, the Netherlands.

¹⁰Therapeutics Discovery, Janssen Pharmaceutica NV, Beerse, Belgium. ¹¹These authors contributed equally: Raúl Y. Tito, Sara Verbandt, Sabine Tejpar, Jeroen Raes.

✉ e-mail: jeroen.raes@kuleuven.be

Methods

Participant recruitment

The LCPM project was an observational cross-sectional survey for which procedures were approved by the medical ethics committee of the UZL (ethical approval number S57084). Between 2017 and 2018, we recruited patients through the study nurse following a standardized procedure. Briefly, we invited patients scheduled for lower gastrointestinal endoscopy or abdominal surgery for CRC removal at the UZL were invited. After explaining the research project and if they expressed their agreement, participants signed an informed consent, and no compensation was offered. A set of stool sample collection material was provided.

Each patient completed an extensive questionnaire containing information about the date of sample collection, the consistency of the stool, diet, antibiotics usage, clinical symptoms or disease among other variables¹⁷, as well as an extensive medical and clinical questionnaire using the Websurvey service of KU Leuven.

As a validation cohort we included the FGFP¹⁷, a population-wide microbiota monitoring effort, representing one of the largest and best characterized fecal microbiota database currently available. Its extensive metadata including health and lifestyle allowed the identification of 69 factors associated with microbiota variation (microbiota covariates). The QMP transformation was conducted in parallel, with the same protocol, for both the FGFP and the LCPM cohorts.

CRC status classification

We invited patients referred for colonoscopy or colectomy to participate in the study. Those that consented were instructed to collect a stool sample at home, which was kept frozen using a sample kit provided by the research team. Upon completion of the medically necessary procedures (colonoscopy or colon resection), we stratified study participants into three diagnosis groups according to their clinical phenotype: (1) patients without evidence of lesions, (2) patients with polyps ($n < 10$ and size between 6 and 10 mm) (ADE) and (3) patients with CRC. Patients whose clinical presentation did not fit any of these three groups were excluded from the study. Once the participants were included in the corresponding groups, extensive metadata was collected from their medical records as stated in the informed consent.

Sample collection

The stool samples of patients from UZL were collected as part of the LCPM project using aliquot ready mat without any buffer or preservative (Supplementary Fig. 1). The samples were kept at $-20\text{ }^{\circ}\text{C}$ freezers at the patients' homes and brought to our laboratory on icepacks. Upon arrival, samples were stored in the Raes' Lab at $-80\text{ }^{\circ}\text{C}$ until further analysis. Each stool sample had a temperature logger to make sure that, during the storage at home or transport to the laboratory, low stable temperature was maintained.

Stool sample analyses

Microbial load measurement by flow cytometry. We determined microbial loads of stool samples of LCPM patients following published procedures²³. We performed cell counting for all other samples in triplicate. Briefly, we dissolved 0.2 g frozen ($-80\text{ }^{\circ}\text{C}$) aliquots in physiological solution to a total volume of 100 ml ($8.5\text{ g l}^{-1}\text{NaCl}$; VWR International). Subsequently, the slurry was diluted 1,000 times. The samples were filtered using a sterile syringe filter (pore size of $5\text{ }\mu\text{m}$; Sartorius Stedim Biotech). Next, we stained 1 ml of the microbial cell suspension obtained with $1\text{ }\mu\text{l}$ SYBR Green I (1:100 dilution in dimethylsulfoxide; shaded for 15 min of incubation at $37\text{ }^{\circ}\text{C}$; 10,000 concentrate, Thermo Fisher Scientific) and monitored fluorescence events using the FL1 533/530 nm and FL3 $>670\text{ nm}$ optical detectors of the C6 Accuri flow cytometer (BD Biosciences). In addition, forward and sideward scattered light was collected. The BD Accuri CFlow (v.1.0.264.21) software was used to gate and separate the microbial fluorescence events on the FL1/FL3 density plot from background events Supplementary Fig. 2.

A threshold value of 2,000 was applied on the FL1 channel. We evaluated the gated fluorescence events on the forward and sideward density plot, as to exclude remaining background events. We kept instrument and gating settings identical for all samples as described previously²⁴. Based on the exact weight of the aliquots analyzed, we converted cell counts to microbial loads per gram of fecal material.

Fecal moisture content. We determined moisture content as the percentage of mass loss after lyophilization from 0.2 g frozen aliquots of nonhomogenized fecal material ($-80\text{ }^{\circ}\text{C}$) as described previously²⁴.

Fecal calprotectin measurement. We quantified fecal calprotectin concentrations using the fCAL ELISA Kit (Buhlmann). For patients and FGFP participants, we conducted analyses on frozen fecal material ($-80\text{ }^{\circ}\text{C}$) as described previously²⁴.

Microbiota phylogenetic profiling

DNA extraction and sequencing data preprocessing. The fecal microbiota profile of the FGFP cohort was described previously¹⁷. For fecal DNA extraction and microbiota profiling of the new cohort, we followed the same protocols¹⁷.

The bacterial profiling was carried out as described previously⁵⁰. Briefly, we extracted nucleic acids from frozen fecal aliquots using the MagAttract PowerMicrobiome DNA/RNA kit (Qiagen). We modified the manufacturer's protocol by the addition of a heating step at $90\text{ }^{\circ}\text{C}$ for 10 min after vortexing and excluding the steps where DNA is removed. For bacterial and archaeal characterization, we used 16S ribosomal RNA primers 515F (5'-GTGYCAGCMGCCGCGGTAA-3') and 806R (5'-GGACTACNVGGGTWTCTAAT-3') targeting the V4 region. These primers were modified to contain a barcode sequence between each primer and the Illumina adapter sequences to produce dual-barcoded libraries from the extracted DNA (dilution 1:10) in triplicate. Deep sequencing was performed on a MiSeq platform (2×250 paired end (PE) reads, Illumina). We randomized all samples and negative controls (polymerase chain reaction (PCR) and extraction controls) taken along for sequencing. After demultiplexing with sdm as part of the LotuS pipeline (v. 1.60)⁵¹ without allowing for mismatches, we further analyzed fastq sequences per sample using DADA2 pipeline (v. 1.6)⁵². Briefly, we removed the primer sequences and the first ten nucleotides after the primer. After merging paired sequences and removing chimeras, we assigned taxonomy using formatted Silva set 'SLV_nr99_v138.1'. We performed taxonomic assignments at the domain, class, order, family, genus and species levels were performed using the 'assignTaxonomy' function from the DADA2 R library, by a naive Bayesian classifier method with a minimum bootstrap confidence of 50, using the 'silva_nr99_v138.1_wSpecies_train_set.fa.gz' training database (Extended Data Fig. 5). Deep sequencing was performed on a MiSeq platform from the DADA2 R library with the formatted Silva SSU database 'silva_species_assignment_v138.1.fa.gz' to obtain species assignments for the amplicon sequence variants (ASVs). We labeled any unassigned ASVs at any taxonomic level, with the prefix 'uc' along with the assigned taxonomic level (not species level) to avoid the lack of labels.

Before the analyses, we removed sequences annotated to the class Chloroplast, family mitochondria or unknown archaea and bacteria from eukaryotic origin. phyloseq (v. 1.36.0)⁵³ and MicroViz (v. 0.11.0)⁵⁴ libraries were used for data curation and figure generation.

RMP. For the relative microbiome matrix, we transformed ASV counts to relative abundances. In other words, we divided ASV counts by the total counts of ASV per sample. We agglomerated ASV to species level using the phyloseq (v. 1.36.0)⁵³ function 'tax_glom'.

RMP (CLR). We agglomerated ASV to the species level, and the abundance matrix was centered log-ratio (CLR)-transformed using 'codaSeq.clr' in the CoDaSeq (v. 0.99.6)⁵⁵ using the minimum proportional abundance detected for each taxon for the imputation of zeros.

Workflow Assessment. We conducted a workflow assessment using (1) a commercial mock community, ZymoBIOMICS Gut, and (2) two *Fusobacterium* species: *Fusobacterium hwasookii* (THCT14E2) and *Fusobacterium nucleatum* (DSM 20482T). The assessment followed our standard methods, involving the amplification, sequencing and analysis of the extracted DNA. This evaluation aimed to assess the performance of our full methodology, as depicted in Extended Data Fig. 6.

Quality control assessment for amplicon sequencing data (16S rRNA) using RMP. In short, we sequenced all samples in six MiSeq runs (Extended Data Fig. 7a). Per each run, we used a set of internal controls to identify: 1) Technical variation within and between runs 1) Contamination events during the DNA extraction, 2) Contamination events during the amplification and sequencing procedures and, 3) Carry-over contamination at the sequencing facility and barcode crosstalk.

We amplified all samples, including biological material (stool samples), positive controls (DNA from a stool sample previously profiled and RS: nonhuman gut bacteria strain '*Runella slithyiformis*'), negative controls (negative control of extraction (NCE) and negative control during PCR (NCP)) in triplicate using a unique barcode combination, while omitting several barcode combinations to control for primer synthesis cross contamination. We used *Runella slithyiformis* in duplicate within each sequencing library to detect barcode crosstalk during the sequencing procedure (Extended Data Fig. 7b). This genus is not detected in human gut samples; therefore, we expected no *Runella slithyiformis* reads in any of the stool samples analyzed. We determined technical variation based on the BCD of positive control samples (Extended Data Fig. 7c). Finally, we included NCEs along the whole process from extraction to bioinformatic analysis. For amplification and sequencing contamination⁵⁶, we used NCP and NCE (Extended Data Fig. 7d and Supplementary Table 12), and for carry-over contamination events, we used a different set of barcode combinations in consecutive MiSeq runs⁵⁶.

QMP. We built the QMP matrix as described previously²³. In brief, we downsized samples to even sampling depth, defined as the ratio between sampling size (16S rRNA gene copy number-corrected sequencing depth) and microbial load (the average total cell counts per gram of frozen fecal material; Supplementary Table 2). We imputed 16S rRNA genome copies (GC) numbers using RasperGade16S (v. 0.0.1)⁵⁷, a new tool that utilizes a heterogeneous pulsed evolution model for predicting 16S rRNA GC. It not only predicts the GC but also provides confidence estimates for the predictions⁵⁷. We used a minimum rarefied read count of less than 150 for QMP analyses. We converted rarefied ASV abundances into numbers of cells per gram. The QMP matrices had a final size of 589 samples for the study cohort and 1,045 samples for the FGFP validation cohort¹⁷. We agglomerate the QMP matrix at ASV level to species level using the phyloseq (v. 1.36.0)⁵³ function 'tax_glom'. We used the resulting species QMP matrix for the main analysis.

Statistical analysis

We performed all statistical analyses with R (Version 4.2.1, RStudio v.2022.12.0 + 353, 86_64-apple-darwin17.0 (64-bit)) and packages phyloseq (v. 1.36.0)⁵³, vegan (v. 2.6.2)⁵⁸, coin (v. 1.4.2)⁵⁹, effectsize (v. 0.8.3), vcd (1.4.11)⁶⁰, DirichletMultinomial (v. 1.34.0)⁶¹, pairwiseAdonis (v. 0.4.1) and microbiome (v. 1.14.0)⁶². We used nonparametric statistical tests for robust comparisons among unbalanced groups. For multiple testing, we corrected all *P* values using the Benjamini–Hochberg method (reported as adjusted *P*) as appropriate on lists ($n > 1$) of features (for example, taxa–metadata or metadata–metadata associations) and also when performing multiple pairwise group ($n > 2$) comparisons (for example, KW test with pH test).

Fecal microbiota derived features and visualization. We visualized microbiota interindividual variation by PCoA using BCD on the species QMP matrix^{24,25}. All the rest of the microbiota derived features were

calculated based on QMP. We determined the contribution of metadata variables to microbiota community variation (effect size) of each of 94 variables by dbRDA on a species-level BCD with the capscale function in the vegan package⁵⁸. We visualized absolute abundance species as log₁₀ (abundance + 1). This was the same for relative abundance.

Microbiota and physiological features associations. We excluded from analyzes any taxa unclassified at the species level or present in less than 5% of samples per each diagnosis group (Supplementary Table 6). We used Spearman correlations for rank–order correlations, between continuous variables complemented by Kendall's tau correlation, including species abundances, calprotectin values and moisture content. We used the Mann–Whitney *U*-test to test median differences of continuous variables between two different groups. For more than two groups, for example, for differential abundance analysis for QMP and RMP taxa versus diagnosis groups, we used the KW test with pH test. For differential abundance analysis among diagnosis groups and bacteria species abundances from CLR transformed data, we performed an ANOVA test.

We evaluated statistical differences in the proportions of categorical variables (enterotypes) between patient groups using pairwise CS tests. We tested for deconfounded microbiota contributions to the diagnosis groups variable by using a nested model comparison (ANOVA) of generalized linear models as follows:

$$\begin{aligned} [\text{null model}] \text{glm0} &= \text{rank}(\text{abundance}) + \text{rank}(\text{calprotectin}) \\ &+ \text{rank}(\text{moisture}) + \text{rank}(\text{BMI}) \end{aligned}$$

[alternative model] glm1 = rank(abundance) + rank(calprotectin) + rank(moisture) + rank(BMI) + diagnosis, where the diagnosis groups were recoded as 1, 2 and 3 for patients without evidence of CTLs, patients with polyps and patients with CRC, respectively. We treated this variable as a continuous variable, translating the directional increase in disease progression, from healthy to lesions, in the colonic mucosa. For the nested model comparison, we used taxa abundances (quantitative or relative) as explanatory variables, the diagnosis groups variable as response variable and BMI, fecal calprotectin and moisture as covariates. Additionally, we employed rank-transformed modeling to perform nonparametric testing on data that is not normally distributed, such as species abundances.

Previous reported CRC microbial markers. To compile a list of published CRC markers that would define taxa that should be tested against covariates in our data set, we conducted a PubMed search query using the keywords 'CRC AND microbiome AND stool AND human AND biomarkers'. We found ten studies that met our inclusion criteria, namely: (1) a sample size minimum of 60 and (2) the CRC biomarker described at the species level, with statistical significance, in the main text of the publication. We included this list of published biomarkers in our correlation analysis between taxa and the three main covariates (fecal calprotectin, BMI and moisture) within the LCPM cohort. A similar procedure was followed at the genus level, which included 15 studies found in our PubMed search.

CRC microbial markers identification. We performed differential abundance analyzes on nine different CRC shotgun datasets as part of 'curatedMetagenomicData'³³ using MetaPhlan 3.0 profiles to compare the results while controlling for potential differences arising from the classification tools and statistical methods used in each independent study. The results of the meta-analysis are presented in Extended Data Fig. 8 and Supplementary Table 13.

Enterotyping and visualization. Using the genus matrix (agglomerated and downsized to 10,000 reads), we enterotyped and calculated observed genus richness⁵³, as already reported for previous studies^{24,25}. For enterotyping (or community typing) based on the DMM approach

we used R as described previously⁶¹. We performed enterotyping on a combined genus-level abundance RMP matrix including LCPM samples compiled with 1,045 samples originating from the FGFP¹⁷. The optimal number of Dirichlet components based on the Bayesian information criterion was four. The four clusters were named 'Bact1', 'Bact2', 'Prev' and 'Rum', as described previously²³.

Reporting summary

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

Data availability

Raw amplicon sequencing data and metadata reported in this study have been deposited in European Nucleotide Archive with accession code [EGAS00001007413](https://www.ebi.ac.uk/ena/record/EGAS00001007413). FGFP 16S rRNA gene sequencing data and metadata are available at the European Genome-phenome Archive ([EGAS00001003296](https://www.ebi.ac.uk/ena/record/EGAS00001003296)). The diagnosis metadata and processed microbiome data required for the reanalysis are provided as Supplementary Tables 1 and 14, respectively. Formatted Silva set 'SLV_nr99_v138.1' files were downloaded from Zenodo via https://zenodo.org/records/4587955/files/silva_nr99_v138.1_wSpecies_train_set.fa.gz?download=1 (silva_nr99_v138.1_wSpecies_train_set.fa.gz)⁶³ and https://zenodo.org/records/4587955/files/silva_species_assignment_v138.1.fa.gz?download=1 (silva_species_assignment_v138.1.fa.gz)⁶³. The nine CRC cohort MetaPhlan 3.0 profiles were collected from curated MetagenomicData, study names: FengQ_2015, HanniganGD_2017, ThomasAM_2018a, ThomasAM_2018b, VogtmannE_2016, WirbelJ_2018, YachidaS_2019 and Yuj_2015, ZellerG_2014 (<https://doi.org/10.18129/B9.bioc.curatedMetagenomicData>). Source data are provided with this paper.

Code availability

Analysis codes are available via Github at <https://github.com/raeslab/QMP-Microbiome-CRC-confounders>.

References

- Tito, R. Y. et al. Population-level analysis of Blastocystis subtype prevalence and variation in the human gut microbiota. *Gut* <https://doi.org/10.1136/gutjnl-2018-316106> (2018).
- Hildebrand, F., Tadeo, R., Voigt, A. Y., Bork, P. & Raes, J. LotuS: an efficient and user-friendly OTU processing pipeline. *Microbiome* **2**, 30 (2014).
- Callahan, B. J. et al. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
- McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).
- Barnett, D., Arts, I. & Penders, J. microViz: an R package for microbiome data visualization and statistics. *J. Open Source Softw.* **6**, 3201 (2021).
- Gloor, G. B., Wu, J. R., Pawlowsky-Glahn, V. & Egozcue, J. J. It's all relative: analyzing microbiome data as compositions. *Ann. Epidemiol.* **26**, 322–329 (2016).
- Seitz, V. et al. A new method to prevent carry-over contaminations in two-step PCR NGS library preparations. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkv694> (2015).
- Gao, Y. & Wu, M. Accounting for 16S rRNA copy number prediction uncertainty and its implications in bacterial diversity analyses. *ISME Commun.* **3**, 59–67 (2023).
- Oksanen, F. J. et al. Vegan: Community Ecology Package. R package Version 2.4-3 <https://CRAN.R-project.org/package=vegan> (2017).
- Hothorn, T., Hornik, K., Van De Wiel, M. A. & Zeileis, A. A Lego system for conditional inference. *Am. Stat.* <https://doi.org/10.1198/000313006x118430> (2006).

- Friendly, M. & Institute, S. A. S. *Visualizing Categorical Data* (SAS Institute, 2000).
- Holmes, I., Harris, K. & Quince, C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS ONE* **7**, e30126 (2012).
- Shetty, S. A. & Lahti, L. Microbiome data science. *J. Biosci.* **44**, 1–6 (2019).
- McLaren, M. R. & Callahan, B. J. Silva 138.1 prokaryotic SSU taxonomic training data formatted for DADA2. *Zenodo* <https://doi.org/10.5281/zenodo.4587955> (2021).

Acknowledgements

We thank all study participants and the different staff members involved in the recruitment and execution of this project. We acknowledge L. Rymenans for her contribution to sample analysis. R.Y.T., S.V. and V.L.R. are funded by postdoctoral fellowships from the Research Fund–Flanders (1234321N, 12R6119N and 12V9421N, respectively). This work was funded by the Innovatie door Wetenschap en Technologie project 'CRC_μBiome: characterization of human and microbial genetic components in premalignant adenoma and colorectal cancer'. The Raes lab is supported by Vlaams Instituut voor Biotechnologie (VIB), KU Leuven and the Rega Institute for Medical Research. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

This study was conceived by J.A., S.T., J. Reumers and J. Raes. The experiments were designed by R.Y.T. and J. Raes. The data were collected and curated by S.V., M.A.V., L.L., J. Reumers, V.L.R., S.V.S., G.F. and S.T. The molecular data were generated by C.V. and R.Y.T. The statistical analyses were planned and executed by R.Y.T. and J. Raes. R.Y.T. and J. Raes drafted the manuscript. All authors revised the article and approved the final version for publication.

Competing interests

J.A. and J. Reumers are employees of Janssen Pharmaceutica NV. J. Raes and R.T. are inventors on the patent application WO2017109059A1 in the name of VIB VZW, Katholieke Universiteit Leuven, KU Leuven R&D and Universiteit Gent covering methods for detecting the presence or assessing the risk of development of inflammatory arthritis disease. J. Raes, S.V.S. and G.F. are inventors on the patent application PCT/EP2018/084920 in the name of VIB VZW, Katholieke Universiteit Leuven, KU Leuven Research and Development and Vrije Universiteit Brussel covering microbiome features associated with inflammation described in Vieira-Silva et al. *Nature Microbiology* 2019. The other authors declare no competing interests.

Additional information

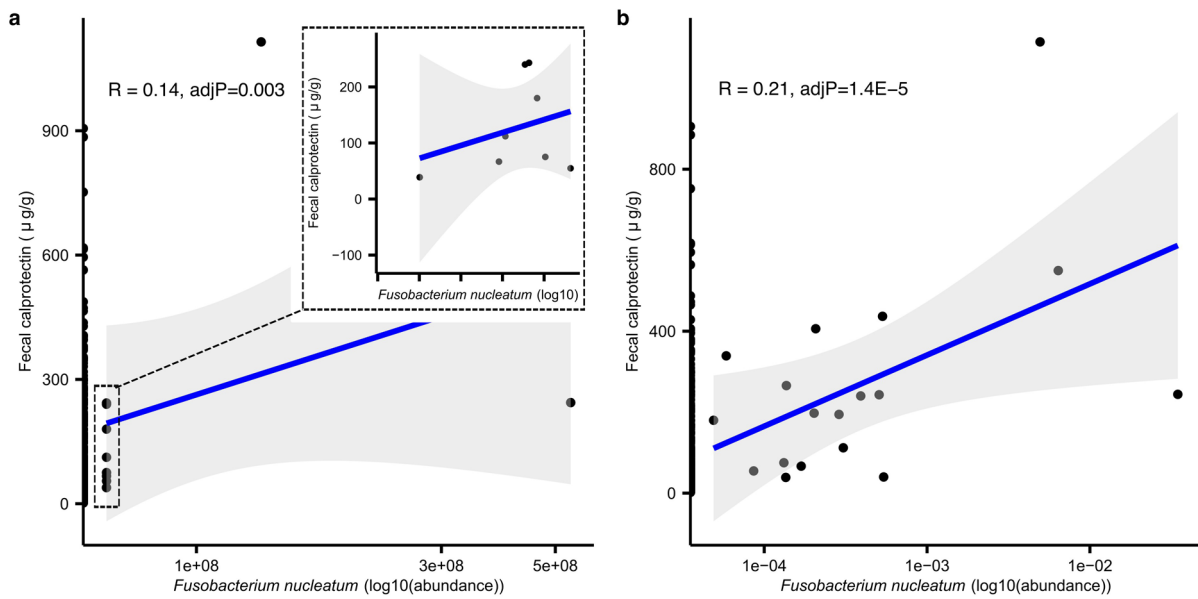
Extended data is available for this paper at <https://doi.org/10.1038/s41591-024-02963-2>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41591-024-02963-2>.

Correspondence and requests for materials should be addressed to Jeroen Raes.

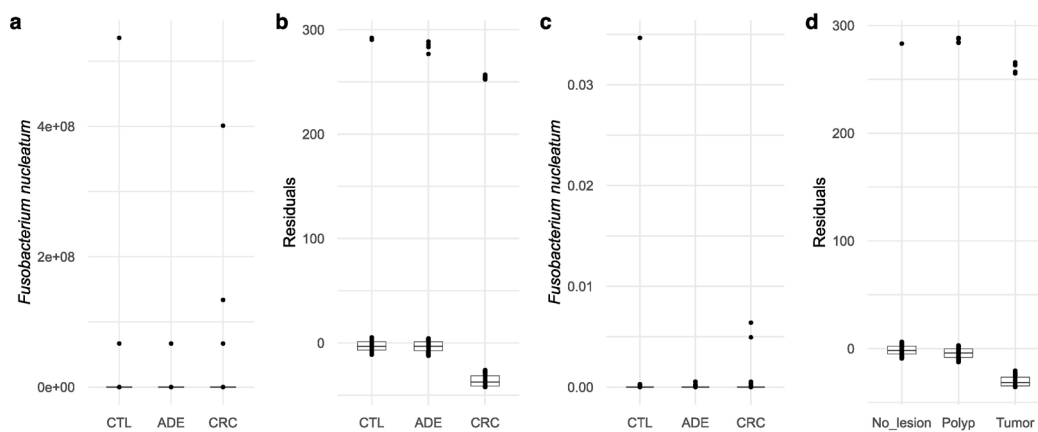
Peer review information *Nature Medicine* thanks Ruixin Zhu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling editor: Alison Farrell, in collaboration with the *Nature Medicine* team.

Reprints and permissions information is available at www.nature.com/reprints.



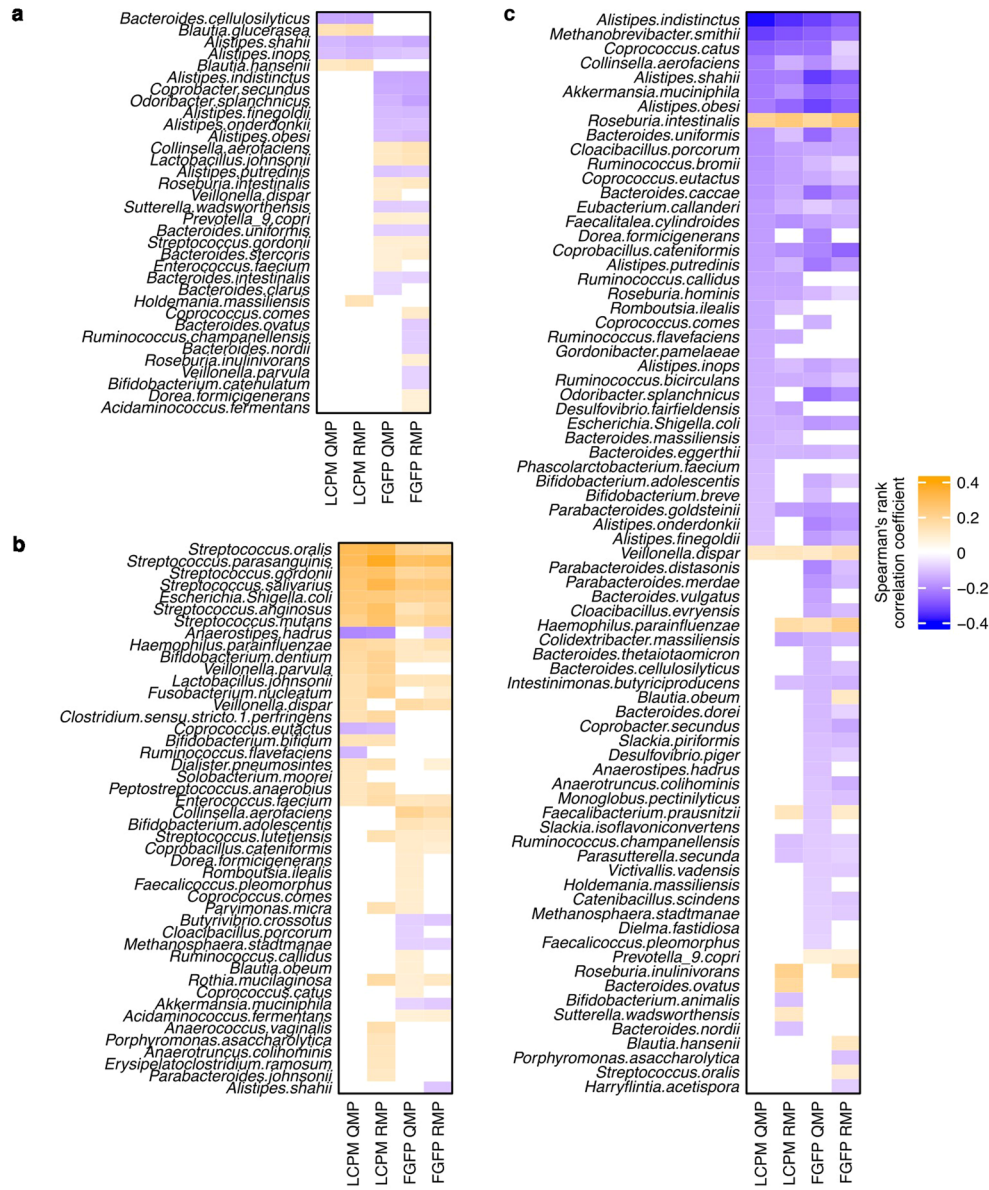
Extended Data Fig. 1 | Association of intestinal inflammation with *Fusobacterium nucleatum*. Intestinal calprotectin levels associate with *Fusobacterium nucleatum* absolute (a) and relative (b) abundance in the LCMP. Two-sided Spearman rank correlation ($\text{adj}P < 0.05$) and 'x' axes are log10 transformed just for plotting. To rule out that the observed association is driven

by a few samples with high abundance of *Fusobacterium nucleatum*, panel a has an insert of the plot removing samples with *Fusobacterium nucleatum* values above 1E8 cells per gram of stool. Best-fitting regression line in blue and 95% confidence interval shown in grey shading.



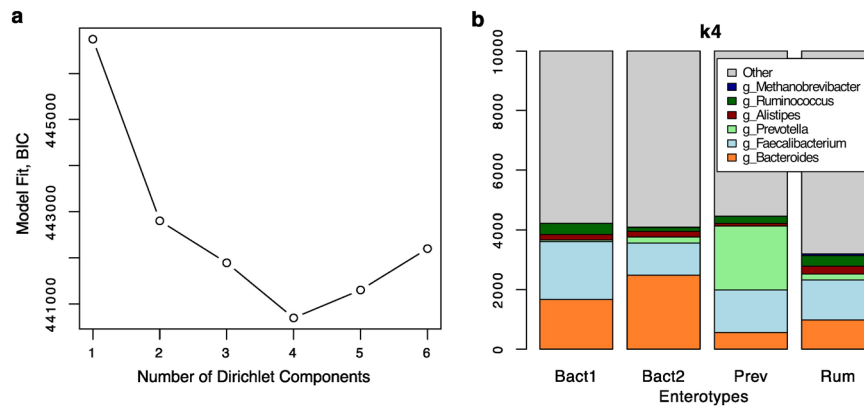
Extended Data Fig. 2 | *Fusobacterium nucleatum* abundances before and after correction for intestinal calprotectin across diagnosis groups. Absolute abundance of *Fusobacterium nucleatum* before (a) and after (b) correcting for intestinal calprotectin. Relative abundance of *Fusobacterium nucleatum* before (c) and after (d) correcting for intestinal calprotectin. The whiskers extend

from the quartiles to the last data point within 1.5× of the interquartile range, with outliers beyond. The 'y' axes for (a) are log₁₀ transformed values (absolute abundance +1). The whiskers extend from the quartiles to the last data point within 1.5× of the interquartile range, with outliers beyond.



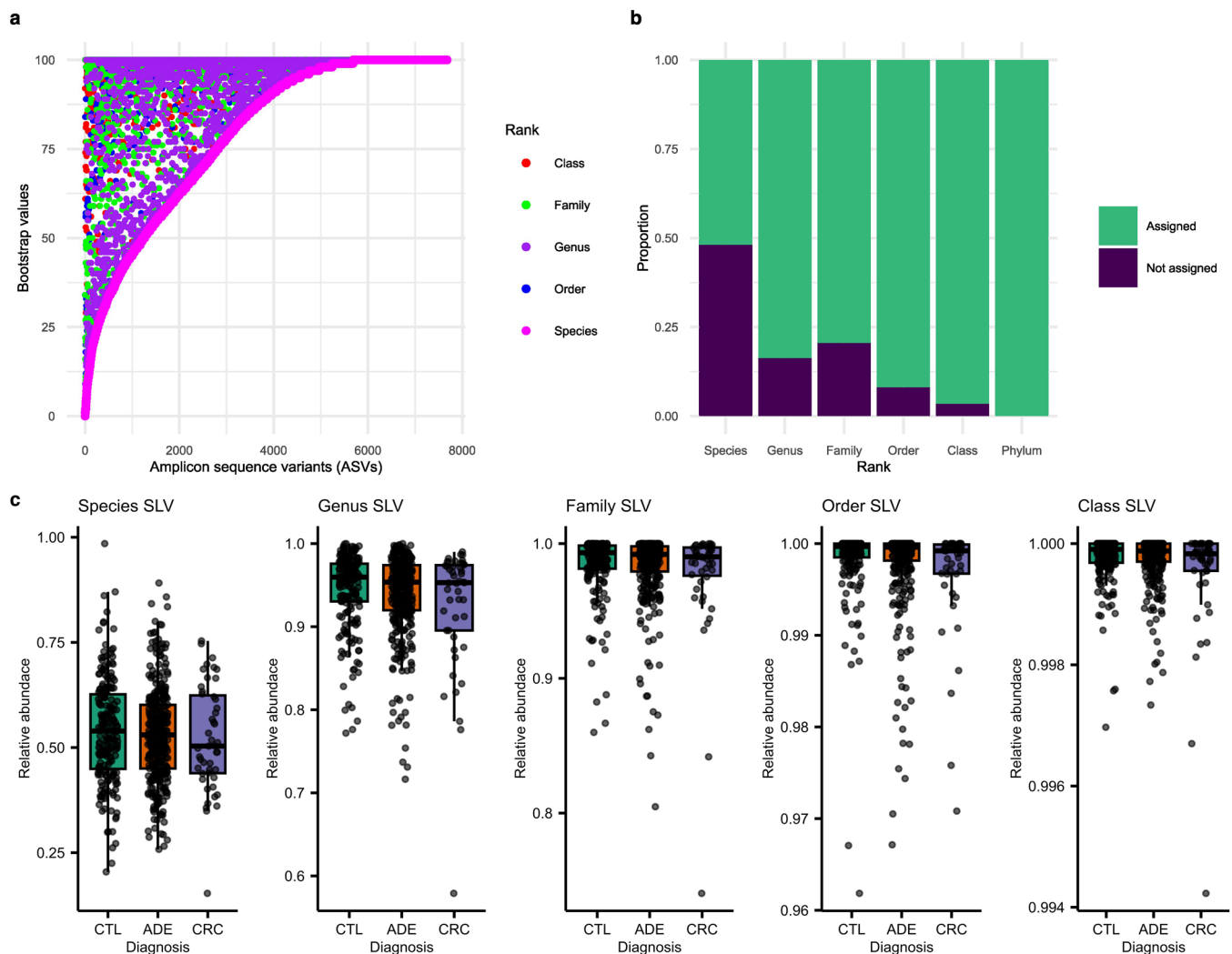
Extended Data Fig. 3 | Spearman correlation between species abundance and microbiota covariates in the LCPM and FGFP cohorts. Two-sided Spearman's rank correlation comparison between absolute species abundance (QMP) and

relative abundance (RMP) from the LCPM (N = 589 samples) and FGFP (N = 1045 samples) cohorts and a, BMI b, faecal calprotectin and c, moisture content values. Spearman correlation adjp < 0.05 (QMP and RMP, Supplementary Table 8).



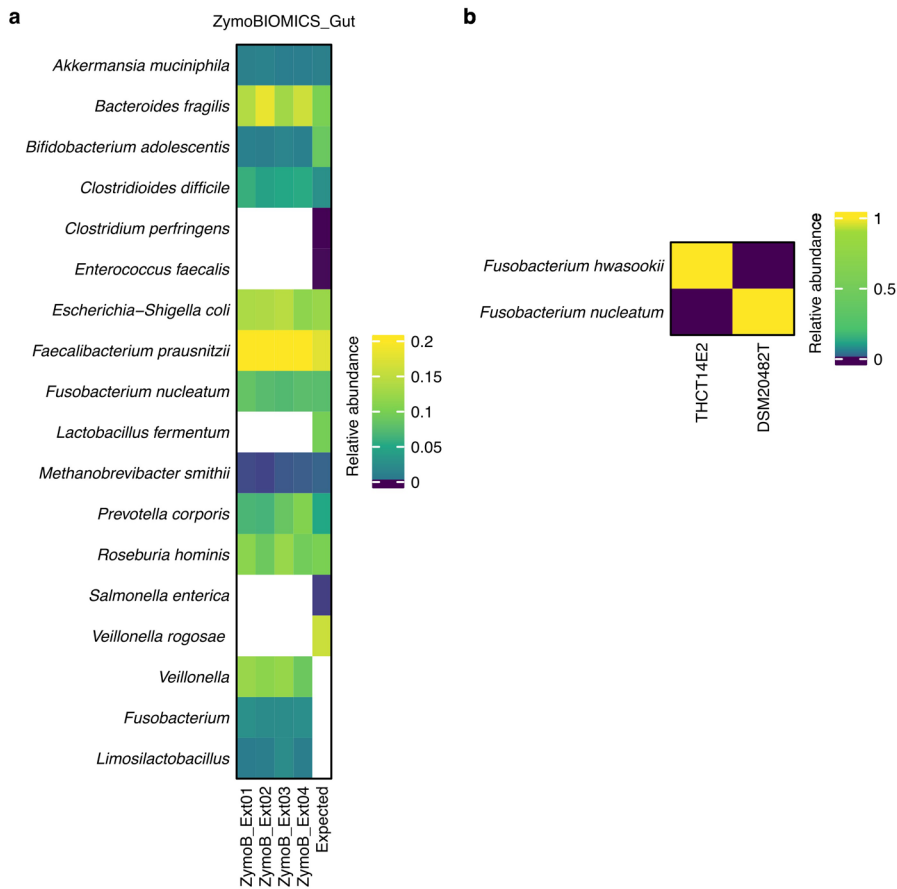
Extended Data Fig. 4 | Enterotype stratification by DMM community typing. **a**, Identification of optimal number of clusters (Dirichlet components) in the LCPM cohort (n = 589) complemented with 1045 samples from the FGFP cohort, based on the Bayesian Information Criterion (BIC). **b**, Barplot representation of

the average relative abundance of a few representative genera split into the four enterotypes identified by DMM community typing on the combined LCPM and FGFP cohorts (n = 1634).



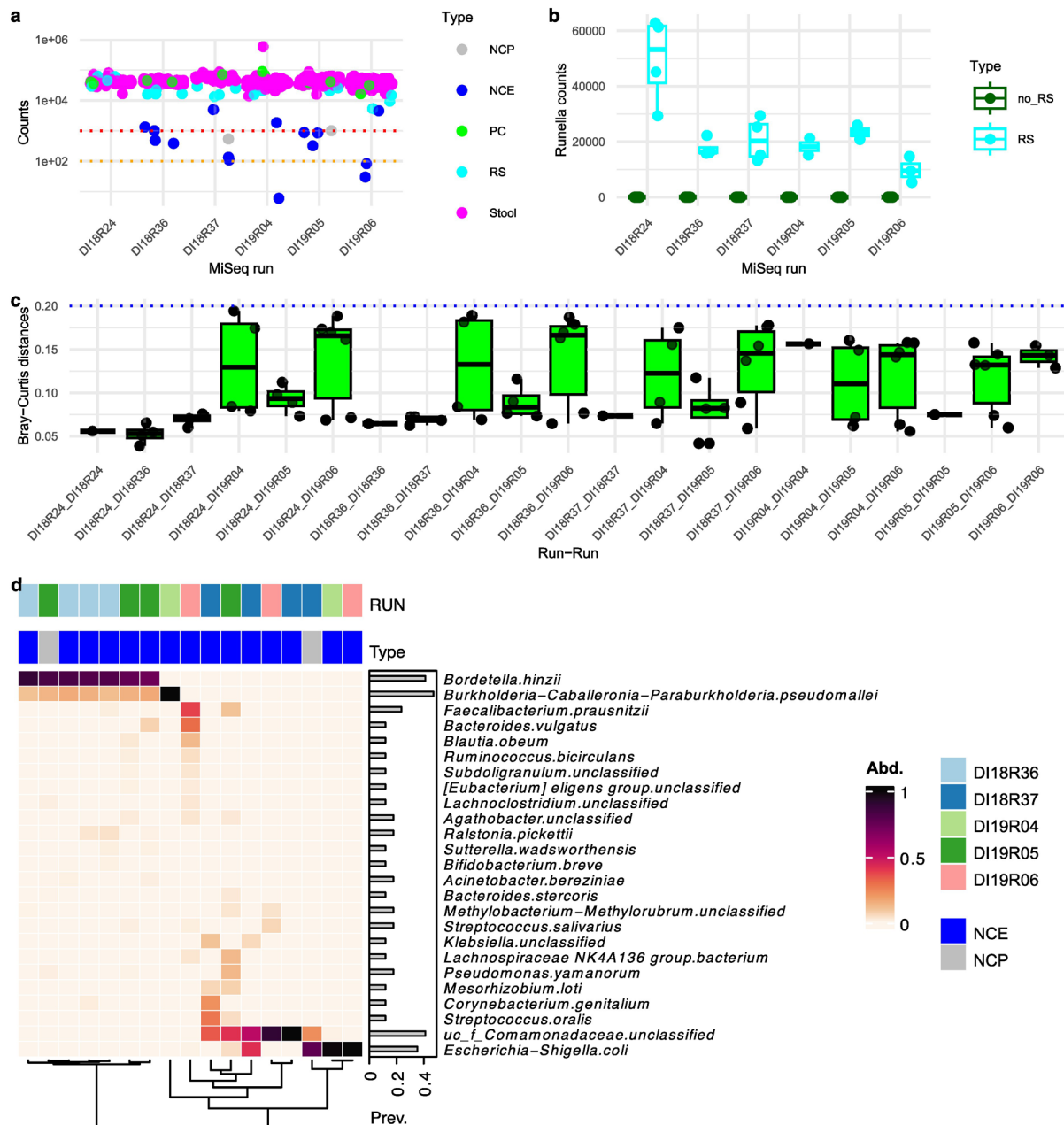
Extended Data Fig. 5 | Taxa assignment performance of the V4 amplicon marker in the LCPM. a, Bootstrap values distribution across different ranks, **b**, Proportion of ASVs assigned from species to phylum, **c**, Proportion of ASVs assigned from species to phylum to each sample. The whiskers extend from the quartiles to the last data point within 1.5× of the interquartile range, with outliers beyond. The figure below (Panel a) illustrates our taxa assignment performance, showing that more than half of the ASVs were assigned to species level with bootstrap values above 80. Panel b shows the ASV assignment proportions from

phylum (100%) to species level (50%). A comparison of proportions of ASVs assigned from each sample at different taxonomic levels revealed no significant differences in the distributions of assigned ASVs per sample across diagnosis groups, as indicated in panel c (KW test, p -values > 0.05). The center of the box plot represents the median value of the data, and the whiskers extend from the quartiles to the last data point within 1.5× of the interquartile range, with outliers beyond.



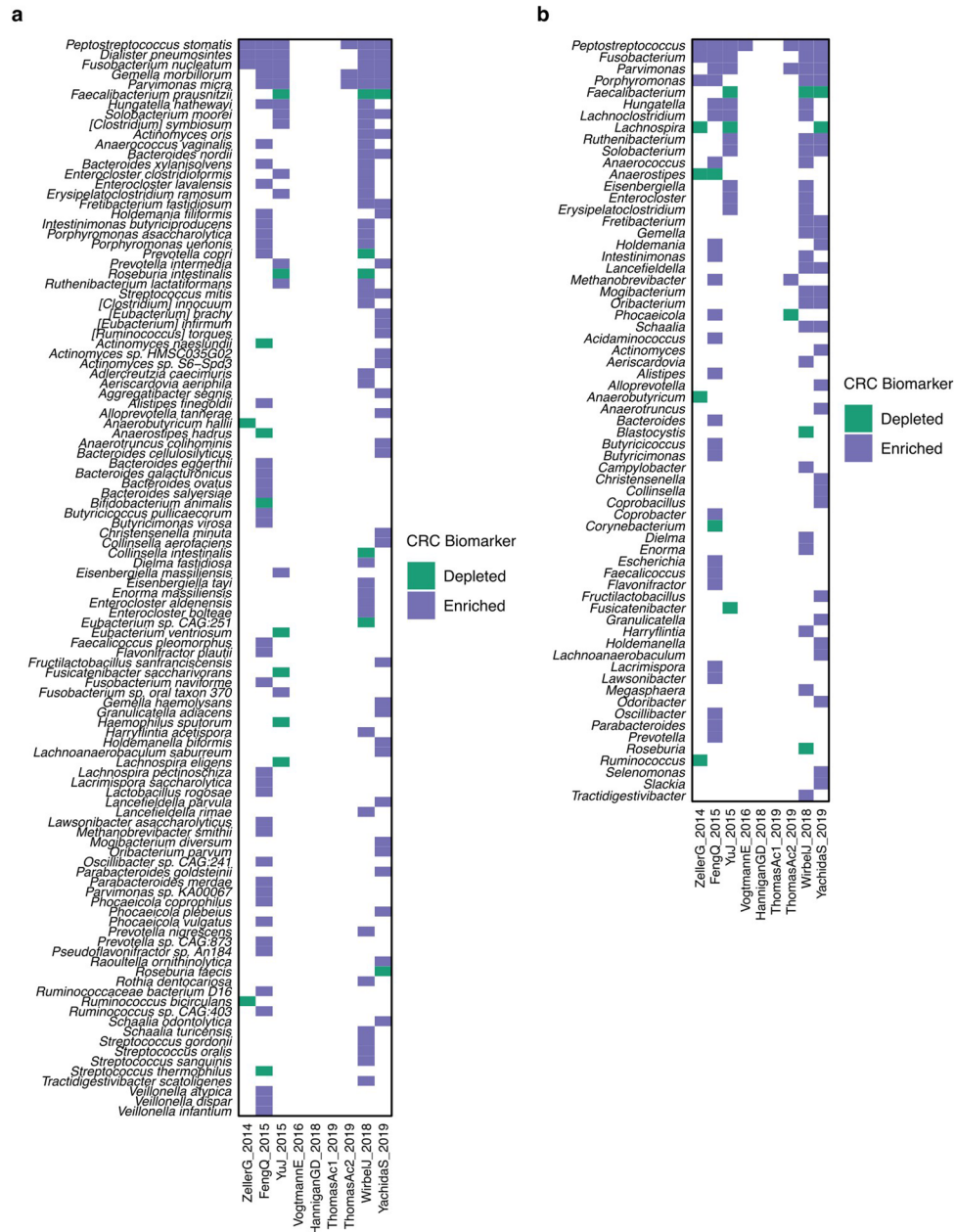
Extended Data Fig. 6 | Performance of our methodology in small communities and isolated microorganisms. a, Species composition of the ZymoBIOMICS gut controls, ten successfully identified species and

b, two *Fusobacterium* species: *Fusobacterium hwasookii* (THCT14E2) and *Fusobacterium nucleatum* (DSM 20482T) were successfully identified using our methodology.



Extended Data Fig. 7 | Quality control assessment for amplicon sequencing data (V4 16S rRNA gene). **a**, The obtained reads for each sample are shown after processing with DADA2 (red and orange dashed lines represent 10,000 and 1,000 reads, respectively; NCP: PCR negative control, NCE: DNA extraction Negative control, PC: positive control, and RS: *Runella sliathyformis* control). **b**, Sequencing controls reveal the absence of barcode crosstalk. RS sequences serve as a marker for barcode crosstalk during sequencing. The absence of RS sequences in the samples without RS (no_RS) ruled out barcode crosstalk during the sequencing or PCR setup procedures. **c**, BCD among technical replicates demonstrating reproducibility. Pairwise comparisons between PC samples within and among MiSeq runs showed values under 0.2 (depicted by the pointed

blue line). The center of the box plot represents the median value of the data, and the whiskers extend from the quartiles to the last data point within 1.5× of the interquartile range, with outliers beyond. **d**, Species composition of negative controls is presented, indicating the relative abundance and prevalence of the top 20 species. None of the species detected with differential abundance using QMP, RMP or CLR were found as background contaminants. Non-significant differences in bacteria composition were observed among DNA sequencing runs (Padj > 0.05, pairwise Adonis test). A full list of detected species is available in Supplementary Table 12. Of note, D118R24 is not shown as the negative controls (NCE and NCP) did not produce reads.



Extended Data Fig. 8 | Species and genera associated with CRC on a subset of the curated Metagenomic Data. After performing our differential abundance procedure on the MetaPhlan 3.0 profiles downloaded from the curated Metagenomic Data, 108 species (a) and 63 genera (b) were identified across the 9 metagenomics datasets.

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | n/a | Confirmed |
|-------------------------------------|--|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of all covariates tested |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A full description of the statistical parameters 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) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

- | | |
|-----------------|---|
| Data collection | [Amplicon sequencing data] After demultiplexing with sdm as part of the LotuS pipeline (v. 1.60) without allowing for mismatches, fastq raw amplicon sequencing files (2x250, Miseq illumina) were further analysed per sample using DADA2 pipeline (v. 1.6).
[Microbial load] The flow cytometry analysis was performed using a C6 Accuri flow cytometer (v.1.0.264.21, BD Biosciences). |
| Data analysis | [Amplicon sequencing data] QMP profiles were created using QMP R-script (https://github.com/raeslab/QMP-Microbiome-CRC-confounders) to rarefy the profiles to even sampling depth using the microbial load and R package: RasperGade16S (v. 0.0.1).
[Statistical analyses] Statistical analyses were performed on Rstudio with R(v.4.2.1) using the following R packages: phyloseq(v. 1.36.0), vegan (v. 2.6.2), coin(v. 1.4.2), effectsize(v. 0.8.3), vcd(1.4.11), DirichletMultinomial(v. 1.34.0), pairwiseAdonis(v. 0.4.1) and microbiome(v. 1.14.0). |

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Raw amplicon sequencing data and metadata reported in this study have been deposited in European Nucleotide Archive with accession codes EGAS00001007413. FGFP 16S rRNA gene sequencing data and metadata are available at the European Genome-phenome Archive (EGAS00001003296). The diagnosis metadata and processed microbiome data required for the reanalysis are provided as Supplementary Table 1 and Supplementary Table 14, respectively.

Formatted Silva set 'SLV_nr99_v138.1' files were downloaded from: https://zenodo.org/records/4587955/files/silva_nr99_v138.1_wSpecies_train_set.fa.gz?download=1 (silva_nr99_v138.1_wSpecies_train_set.fa.gz) and https://zenodo.org/records/4587955/files/silva_species_assignment_v138.1.fa.gz?download=1 (silva_species_assignment_v138.1.fa.gz).

The 9 colorectal cancer cohort MetaPhlan 3.0 profiles were collected from curatedMetagenomicData, study names: FengQ_2015, HanniganGD_2017, ThomasAM_2018a, ThomasAM_2018b, VogtmannE_2016, WirbelJ_2018, YachidaS_2019, Yuj_2015, ZellerG_2014 (DOI: 10.18129/B9.bioc.curatedMetagenomicData).

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender

Results only refer to sex. Gender was not recorded.

Population characteristics

A total of eight variables were found significantly associated (FDR < 5%) with diagnostic groups (namely: age, BMI, calprotectin, reported hours of sleep, previous cancer (including CRC), diabetes treatment and high blood pressure). Younger patients were more likely to exhibit no evidence of colonic lesions, whereas patients with lesions tended to be older. Additionally, patients with adenomas had a higher BMI compared to those without lesions, while patients without lesions have lower levels of intestinal calprotectin

Recruitment

Patients were recruited through the study nurse following a standardized procedure. Briefly, we invited patients scheduled for lower gastrointestinal endoscopy or abdominal surgery for CRC removal at the UZL. After explaining the research project and upon expressed interest, the informed consent was signed, and stool sample collection materials were provided. After colonic examination, patients were assigned to one of the diagnosis groups. In the case of the CLT group without colonic lesions, there might be an intrinsic increased risk of colorectal cancer due to the clinical necessity of a colonic examination.

Ethics oversight

Medical ethics committee of the UZL [Ethical approval number: S57084].

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

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

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-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. Cohort sizes were determined based on previous findings (Falony et al., Science 2016).

Data exclusions

Taxa unclassified at the species level or present in <5 % of samples per each diagnosis group were excluded from the statistical analyses.

Replication

Several correlation on species abundance and BMI (2 of 6), fecal calprotectin (17 of 29), and moisture content values (29 of 50) in the LCPM were replicated in the FGFP cohort.

Randomization

Not applicable: this was a cross-sectional study, not a randomized study. No intervention was performed on subjects, and therefore no random allocation into groups.

Blinding

Not applicable: this was a cross-sectional study, not a randomized study. Investigators were not blinded during data collection and analyses.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involvement
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involvement
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	The study protocol was registered at clinicaltrials.gov (NCT02947607)
Study protocol	The study protocol is available at https://clinicaltrials.gov/ct2/show/NCT02947607
Data collection	The final Leuven CRC Progression Microbiome (LCPM) study cohort consisted of 589 patients. The screening cohort included 650 volunteers referred for colonoscopy screening and colonic resections at UZL between 2017-2018, who provided a stool sample before the colonic procedure. Most participants were from the Flemish region of Belgium. These participants were classified within three diagnostic groups according to a thorough colonoscopy and clinical assessment: 1) patients without evidence of colonic lesions (CTL), 2) patients with polyps (n< 10 and size between 6 to 10 mm) (ADE), and 3) patients with CRC (CRC).
Outcomes	The hypotheses tested in this manuscript were not originally specified as part of the planned outcomes for the NCT02947607 study. The primary outcome of the project, characterization of "Differential host microbiome composition and abundance in healthy, adenoma, and CRC patients and its correlation to CRC risk features and host genomic and transcriptomic components," is not addressed in this manuscript. While the secondary outcome, involving the analysis of "Host microbiome composition and abundance data generated from saliva, stool, and colonic biopsies using amplicon-based 16S ribosomal RNA sequencing," is partially explored in this manuscript, particularly regarding stool microbial communities, a comprehensive examination across all specified sample types and methodologies outlined in the study protocol, were not fully developed.

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	0.2 g frozen (-80 °C) aliquots were dissolved in physiological solution to a total volume of 100 ml (8.5 g l-1 NaCl; VWR International). Subsequently, the slurry was diluted 1,000 times. Samples were filtered using a sterile syringe filter (pore size of 5 µm; Sartorius Stedim Biotech). Next, 1 ml of the microbial cell suspension obtained was stained with 1 µl SYBR Green I (1:100 dilution in dimethylsulfoxide; shaded for 15 min of incubation at 37 °C; 10,000 concentrate, Thermo Fisher Scientific).
Instrument	C6 Accuri flow cytometer (BD Biosciences)
Software	BD Accuri CFlow software v.1.0.264.21 (BD Biosciences)
Cell population abundance	not applicable. No sorting of fractions was performed.
Gating strategy	Fluorescence events were monitored using the FL1 533/30 nm and FL3 > 670 nm optical detectors. In addition, forward and

Gating strategy

sideward scattered light was collected. The BD Accuri CFlow software was used to gate and separate the microbial fluorescence events on the FL1/FL3 density plot from background events. The gated fluorescence events were evaluated on the forward and sideward density plot, as to exclude remaining background events.

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