

## Metabolite quantification of faecal extracts from colorectal cancer patients and healthy controls

Gwénaëlle Le Gall<sup>1,\*</sup>, Kiran Guttula<sup>2,\*</sup>, Lee Kellingray<sup>1,\*</sup>, Adrian J. Tett<sup>3</sup>, Rogier ten Hoopen<sup>2</sup>, Kate E. Kemsley<sup>1</sup>, George M. Savva<sup>1</sup>, Ashraf Ibrahim<sup>2,\*</sup> and Arjan Narbad<sup>1,\*</sup>

<sup>1</sup>Quadram Institute Bioscience, Norwich Research Park, Norwich, UK

<sup>2</sup>Division of Molecular Histopathology, Department of Pathology, University of Cambridge, Cambridge, UK

<sup>3</sup>Centre for Integrative Biology, University of Trento, Trento, Italy

\*These authors contributed equally to this work

**Correspondence to:** Gwénaëlle Le Gall, **email:** gwenaelle.legalln@gmail.com

**Keywords:** NMR; colorectal cancer; markers; metabolite; metabolomics

**Received:** February 07, 2018

**Accepted:** August 10, 2018

**Published:** September 07, 2018

**Copyright:** Le Gall et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License 3.0 (CC BY 3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

### ABSTRACT

**Colorectal cancer (CRC), a primary cause of morbidity and mortality worldwide is expected to rise in the coming years. A better understanding of the metabolic changes taking place during the disease progression is needed for effective improvements of screening strategies and treatments. In the present study, Nuclear Magnetic Resonance (NMR) metabolomics was used to quantify the absolute concentrations of metabolites in faecal extracts from two cohorts of CRC patients and healthy controls. The quantification of over 80 compounds revealed that patients with CRC had increased faecal concentrations of branched chain fatty acids (BCFA), isovalerate and isobutyrate plus valerate and phenylacetate but diminished concentrations of amino acids, sugars, methanol and bile acids (deoxycholate, lithodeoxycholate and cholate). These results suggest that alterations in microbial activity and composition could have triggered an increase in utilisation of host intestinal slough cells and mucins and led to an increase in BCFA, valerate and phenylacetate. Concurrently, a general reduction in the microbial metabolic function may have led to reduced levels of other components (amino acids, sugars and bile acids) normally produced under healthy conditions. This study provides a thorough listing of the most abundant compounds found in human faecal waters and presents a template for absolute quantification of metabolites. The production of BCFA and phenylacetate in colonic carcinogenesis warrants further investigations.**

### INTRODUCTION

The global burden of colorectal cancer (CRC), which accounted for about 1.4 million new cases and almost 700 000 deaths in 2012, is expected to rise by 60% by 2030 [1]. Although the incidence is decreasing in countries with high human development index (HDI) where around two-thirds of cases and deaths worldwide are occurring, rapid increases in incidence and mortality

are now seen in many medium HDI countries [1, 2]. The risk of colorectal cancer increases with age, high consumption of red or processed meat or alcohol, low intake of fruit and vegetables, smoking, high body mass index and low physical activity [2]. Early detection is key to a favourable 5-year survival rate [3]. Screening programs in many countries are based on a combination of non-invasive clinical markers (faecal occult blood and faecal immunochemical tests) and endoscopic techniques

(flexible sigmoidoscopy, colonoscopy and computed tomographic colonography) but due to the variable sensitivity of existing non-invasive tests, new non-invasive procedures are still urgently needed. DNA-based stool [4] and serum-based tests such as carcinoembryonic antigen and carbohydrate antigen 19-9 [5, 6] hold promise but low sensitivity remains an issue. The widespread use of colonoscopy has resulted in a significant decrease in the mortality of colorectal cancer due its high rate sensitivity [7] but is invasive, costly and associated with bleeding and perforation.

Metabolomic analysis consists of measuring and comparing the levels of metabolites across samples to discover potential biomarkers. This non-invasive approach has been applied to tumour and adjacent tissue [8–9], blood plasma [10–12], urine [13, 9] and faecal extracts [14–20] to search for markers of early diagnosis and for staging of CRC. The commonality in the biopsy studies was an increase in amino acids and lactate in tumour tissue [8–9]. Changes in plasma partly mirrored those findings since amino acid levels differed [10–12] and lactate levels were found to be higher in CRC samples [10]. Additionally, intermediates of purines, pyrimidines and the tricarboxylic acid (TCA) cycle were altered [11–12]. Variation in amino acids and TCA cycle pathways was also observed in urine [13]. The studies on faecal metabolomics reported changes in short chain fatty acid (SCFA), amino acid, and lipid metabolism. Butyrate levels were depleted in many of the studies [14, 16, 19–20] but with one exception [18]. Similarly, conflicting results emerged on acetate with three studies reporting an increase in concentration [16, 18, 20] while two others stated a decrease [14, 19]. A more consensual trend of elevated amino acid levels was found for most studies [14, 16, 20] except for two studies that reported a diminution of the level of glutamine [18–19]. Results from metabolomics studies typically do not use absolute quantitation since the primary aim with this approach is to provide a rapid screening for group comparisons. The novelty in the present study is two-fold: firstly, the provision of a full list of quantified faecal metabolites in a healthy human and secondly, detecting the metabolic differences between healthy individuals and two independent cohorts of CRC patients. Absolute quantification of faecal SCFA and amino acids has previously been published [21–22] but reports on the amounts of organic acids, sugars, nucleosides, and other molecules present in human faeces are scarce. Establishing the composition and the expected quantity of compounds in faecal extracts would help to clarify the role of faecal metabolites in the development of CRC and other gastrointestinal tract diseases. In this study, we analysed the faecal microbiome of one cohort and the metabolomes of two cohorts of CRC patients and healthy controls and have identified quantifiable differences in the composition and function of the gut microbiomes of CRC patients.

## RESULTS

### Patient demographics and study design

The first set of samples from 20 CRC patients and 20 healthy controls was analysed in 2012. The second cohort consisting of 30 CRC patients and 30 healthy individuals were analysed two years later. Hence 50 age and sex-matched pairs of stool samples were used for <sup>1</sup>H NMR profiling of faecal metabolites. One outlier was excluded from the second set due to poor spectral quality. Metagenomics analysis was also applied to the faecal microbiome of the first set ( $n = 40$ ). Patients and tumour characteristics are outlined in Table 1.

### Faecal metabolite quantification

The faecal <sup>1</sup>H NMR spectra were dominated by signals arising from the three main SCFA namely, acetate, propionate, and butyrate and characterised by low levels of many other metabolites (Figure 1A). Over 80 compounds were identified with 2-dimensional NMR experiments, the literature data [18, 23] and the human metabolome database and quantified in an absolute manner (Supplementary Table 1). Compounds included energy related metabolites such as fatty, organic, and amino acids, sugars, osmolytes, amines, alcohols, phenolic compounds, nucleobases, nucleosides, nucleotides, vitamin B3 and bacterial degradation products. Findings prior (Supplementary Table 2) and after (Supplementary Table 1) combining the two sets of data did not differ substantially.

### Biomarkers of CRC

Visual inspection indicated that a subset of cancer profiles was characterised by high levels of isovalerate, isobutyrate and phenylacetate (Figure 1A and 1B). Principal component analysis (PCA) of the samples from both sets showed overlapping but some separation between cancer and healthy groups (Supplementary Figure 1). However, a large number of principal components were needed to account for a significant proportion of variation in the dataset.

Conversely, with PLS-DA, cross-validation within each set suggested that using two components was optimal for prediction. When two components models were estimated in each set and applied the predictive power was similar in each case (Figure 2A). Using 11-fold cross-validation within the combined set led to an C-statistic of 0.80 (Figure 2B). A sensitivity of 80% was achieved at a specificity of 70%, while specificity of 80% could be achieved with sensitivity of 67%. With a threshold of 50% used to identify cases from controls, the PLS-DA model classified 74 (75%) of 99 cases correctly. This suggests that while it is not possible to completely identify CRC

**Table 1: Patient demographics and tumour characteristics**

Patients		set 1		set 2	
		Colorectal cancer	Healthy	Colorectal cancer	Healthy
		N = 20	N = 20	N = 30	N = 29
Age, years	Mean	67	67	66	66
	Range	61–72	60–74	60–74	60–74
Sex	Women	8	8	9	9
	Men	12	12	21	21
Tumour site	caecum	2		3	
	ascending	2		2	
	transversal	3		0	
	descending	1		2	
	sigmoid	4		17	
	rectum	6		6	
Cancer size, mm	Mean	35		24	
	Range	12–70		15–40	
Dukes's stage	A	3		3	
	B	7		2	
	C	5		11	

from control patients using faecal metabolites a reasonable degree of discrimination is possible, even within a cohort that has screened positive using FOBt.

In univariate analysis, twenty metabolites had FDR-adjusted *p*-value lower than 0.05 when values were compared between cancer patients and controls. Fifteen had significantly lower levels among cancer patients, five had significantly higher levels (Table 2). Distributions of each metabolite concentration in which a significant difference was observed (at FDR < 0.05) are shown in Supplementary Figures 2 (raw concentrations) and 3 (Box-Cox transformed).

While not all identified metabolites show statistically significant differences in both sets of patients, ratios of effects are largely consistent across sets. The distribution of *p*-values comparing the differences in effects across sets was uniform, suggesting that there was no difference between sets with respect to associations between metabolites and cancer status (Supplementary Table 2). Some of the markers were unique to a set (succinate for set 1, *p*-cresol for set 2) but the majority showed a consistent trend of increased or decreased levels in both sets, although for each circumstance the metabolite had a non-significant *p* value in one of the sets.

Table 2 shows the ratios of concentrations of metabolites that are statistically significantly higher or lower between groups (at FDR corrected *p*-value < 0.05).

Cancer patients are characterised by higher concentrations of iso short-chain fatty acids (valerate, isobutyrate, isovalerate), phenylacetate and, a sugar-phosphate whose signals arise at 5.61 ppm, and lower concentrations of methanol, amino acids (glutamine, ornithine, isoleucine, taurine, and *b*-alanine), sugars (glucose, galactose and xylose), and bile acids (deoxycholate, lithodeoxycholate and cholate). There is some evidence for differences in many other metabolites with individual *p* values of less than 0.05 although their statistical significance may arise through the large number of hypotheses being tested. (Supplementary Table 1). Correlations between selected metabolites are shown in Supplementary Figure 4. The highest levels of isoacids, phenylacetate and phenylpropionate were all associated with the same five cancer patients.

### Gut microbiota composition of set 1

The gut microbiota composition of the first cohort (20 CRC patients versus 20 age and sex matched healthy controls) was assessed using 454 pyrosequencing and the QIIME pipeline. This analysis produced 167 572 sequence reads, with an average of  $4189 \pm 1476$  reads per sample, which clustered into 5762 operational taxonomic units at 97% identity. PCoA plots were generated to investigate whether the microbiota of patients with CRC were more

similar to one another than those obtained from healthy controls. The unweighted analysis, whose results are based on which taxa are shared between samples, suggested a difference in the composition of the microbiota between the two groups (Figure 3A). However, these differences were less clear following the weighted analysis, which also takes the relative abundances of taxa into consideration (Figure 3B). At the family level it was observed that, on average, CRC patients had a larger relative abundance of Ruminococcaceae ( $32.65\% \pm 8.72\%$  vs  $20.35\% \pm 13.34\%$  ( $P = 0.001$ )) and a lower proportion of Lachnospiraceae ( $30.34\% \pm 11.49\%$  vs  $42.57\% \pm 18.33\%$  ( $P = 0.016$ )). Of the seven bacterial families that were significantly different between CRC patients and healthy controls, five are members of the Clostridiales order (unclassified Clostridiales ( $P = 0.027$ ), Christensenellaceae ( $P = 0.002$ ), Mogibacteriaceae ( $P = 0.017$ ), and Lachnospiraceae & Ruminococcaceae), as well as Porphyromonadaceae (Bacteroidales) and an unclassified family member of the order RF39 ( $P = 0.006$ ) (Table 3). Interestingly, although only present as a small proportion of the microbiota, the Archaeal family Methanobacteriaceae were found in 50% (10/20) of CRC patients compared to 10% (2/20) of healthy controls ( $P = 0.009$ ). Bar charts depicting the taxonomic composition of each faecal sample at the genera level did not suggest a community profile that may be a signature of CRC (data not shown). However, some statistically significant differences in the relative proportions of certain genera were observed between the two groups, with 14 of the 17 taxa identified found at a higher proportion in the CRC samples (Table 3). Further statistical analysis of the microbiota composition of the subset of CRC patients, which were identified as outliers through metabolomic analyses, produced a list of taxa that were present at significantly different proportions to the remaining CRC cohort (Supplementary Table 3).

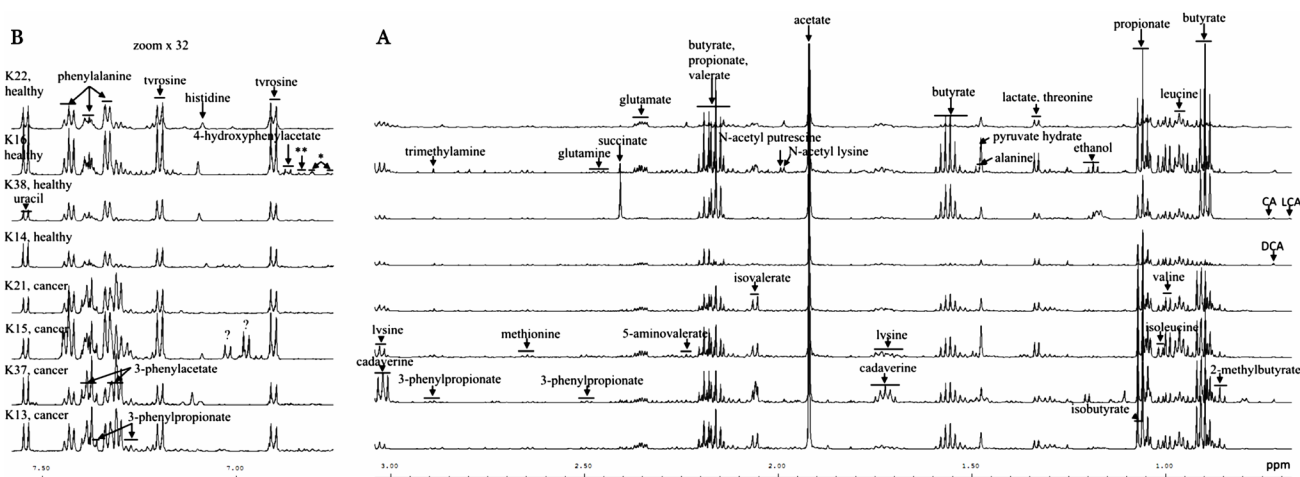
A canonical correlation analysis on the metabolite data and the matching microbiota data I set 1 (Supplementary Figure 5) showed there was a weak but significant link between the two microbiota and NMR data sets but no separation between the two groups of samples (cancer vs controls). No good correlation was found in the heap map correlating individual metabolite with individual microbial trait (not shown).

## DISCUSSION

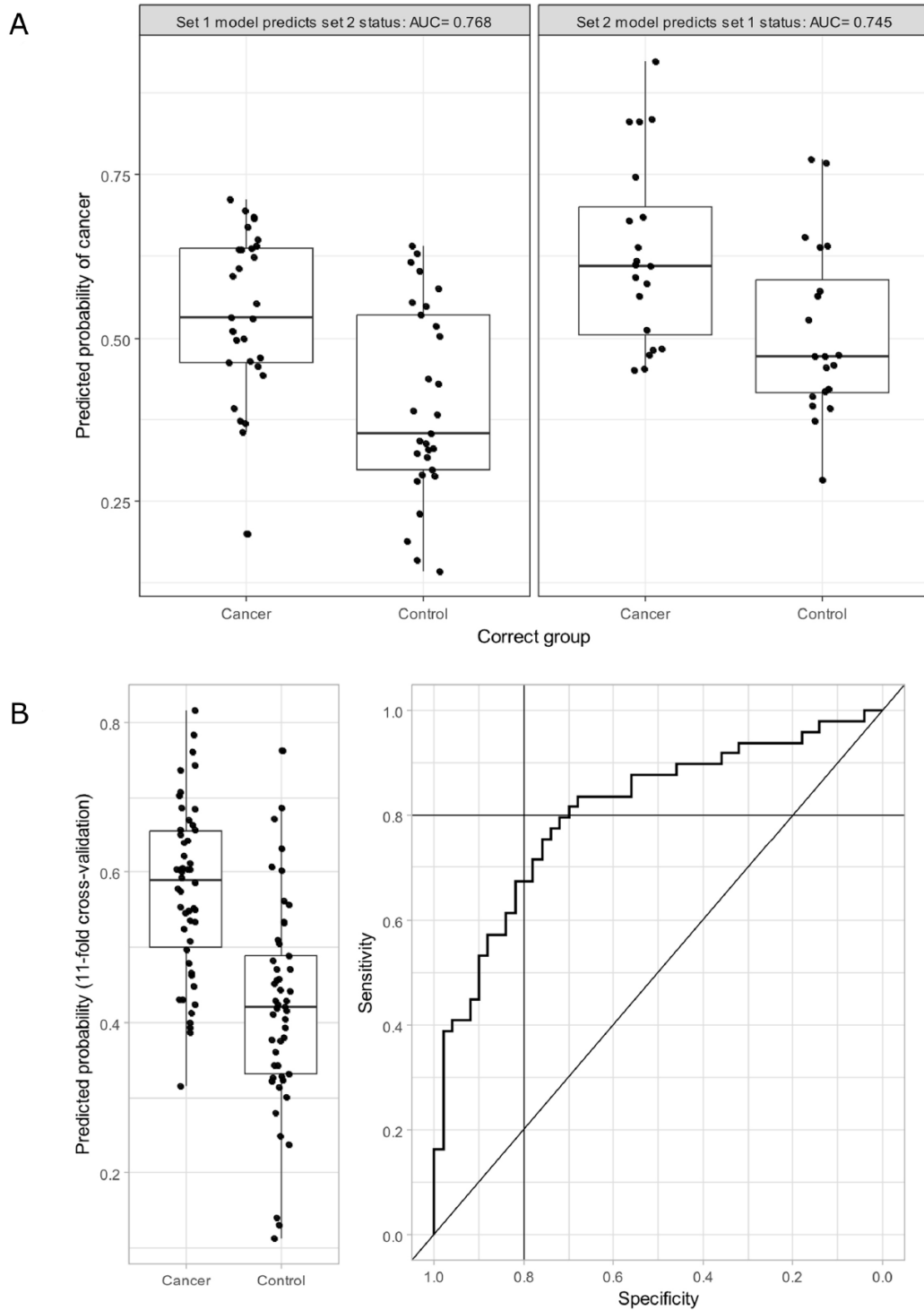
In the present study, we have presented a comprehensive list of faecal metabolites expressed in concentration units among 50 CRC patients and 49 controls, recruited through a national screening programme who had screened positive using FOBt.

NMR quantification, an approach widely applied to urine and blood samples [24, 25] has not yet been applied to faecal water extracts. We have used our expertise in metabolite identification of faecal waters by  $^1\text{H}$  NMR [26] to prepare a thorough list and used a specialised software to quantify the metabolites detectable in faecal extracts.

Previous faecal metabolomics studies on CRC have qualitatively detected metabolites such as short chain fatty acids and amino acids [14–20], but to our knowledge this is the first time that faecal absolute concentrations in healthy and cancer patients have been determined. The concentrations of SCFA, amino acids, lactate, phenol, p-cresol, and bile acids are consistent with those reported in the literature [21–22, 27, 28]. The absolute concentration of the other compounds has not been reported before. Dietary polysaccharides/fibre that reach the large intestine are broken down by bacteria into end-products such as SCFA, mainly acetate, propionate, and butyrate, lactate, ethanol, methane, hydrogen, and  $\text{CO}_2$  [29]. Colonic health is associated with a diet rich in non-



**Figure 1: Typical 600 MHz  $^1\text{H}$  NMR spectra of aqueous faecal extracts from 4 CRC patients and age and sex matched controls.** High and mid (A) and low (B) field regions of the  $^1\text{H}$  NMR spectra. Key: \*, 3-hydroxyphenylpropionate; \*\* p-cresol.



**Figure 2:** (A) The predicted probability of cancer estimated by PLS-DA predictive power using each set and validated by applying to the other. AUC = area under the receiver operator characteristic (ROC) curve, and reflects the probability that a randomly selected CRC patient has a higher predicted probability of cancer than a randomly selected control. (B) Left panel shows the predicted probability of cancer estimated by scaled PLS-DA models using Box-Cox transformed metabolite concentrations, stratified by cancer status. 11-fold cross-validation was used, hence each predicted probability is estimated independently of the true cancer status of the patient. Right hand panel shows the ROC curve estimated using the same data (AUC = 0.8) with solid lines indicating sensitivity and specificity of 80%.

**Table 2: The ratio of metabolite concentrations between CRC patients and controls**

Metabolite	Mean concentration ratio (CRC/control)	<i>t</i> -statistic	<i>p</i> -value ( <i>t</i> -test)	<i>p</i> -value (FDR adjusted)
Lower among cancer patients				
Cholate	0.13	-5.06	0.00000	0.0002
Taurine	0.39	-4.59	0.00001	0.001
Glutamine	0.67	-4.24	0.00005	0.002
β-Alanine	0.35	-4.13	0.00008	0.002
Glucose	0.38	-3.91	0.00018	0.003
Lithodeoxycholate	0.01	-4.02	0.00015	0.003
Xylose	0.38	-3.31	0.00134	0.012
Deoxycholate	0.44	-3.28	0.00145	0.012
Ornithine	0.59	-3.33	0.00124	0.012
Glycerol	0.68	-3.20	0.00189	0.015
Guanosine	0.21	-3.16	0.00248	0.018
Isoleucine	0.70	-3.01	0.00339	0.021
Methanol	0.62	-2.77	0.00679	0.040
Galactose	0.61	-2.66	0.00913	0.048
4-Aminohippurate	0.39	-2.61	0.01058	0.050
Higher among cancer patients				
Isovalerate	1.75	3.59	0.00052	0.007
Hexose-phosphate*	2.02	3.37	0.00107	0.012
Phenylacetate	1.73	3.01	0.00335	0.021
Isobutyrate	1.54	2.74	0.00741	0.041
Valerate	1.42	2.62	0.01029	0.050

*T*-statistics and *p*-values are calculated using Box-Cox transformed concentrations. Adjusted *p*-values are calculated using the false discovery rate method of Benjamini and Hochberg to correct for the large number of hypotheses being tested.

\*signal at 5.61 ppm, most likely glucose- or galactose-1-phosphate.

digestible starch and is reflected by fairly high levels of butyrate, acetate, and propionate [30–31]; however, the presence/absence of other elements from the metabolite composition also contribute to the maintenance of a healthy gut [29–31]. Host derived glycans notably mucins, dietary amino acids and proteins are other major nutrient sources for gut bacteria [29, 32]. In this study, findings from two independent sets of CRC patients and healthy controls consistently showed an elevation of isovalerate, isobutyrate, valerate and phenylacetate levels in CRC and a diminution in the concentrations of amino acids, sugars and methanol.

Two previous studies have reported an increase of isovalerate, isobutyrate and valerate levels in CRC faecal extracts [16, 20], one study reported a decrease [33] and another one no differences [34]. Straight-chain SCFA (butyrate, propionate, acetate and valerate) are products of saccharides and amino acids while branched SCFA (BCFA), isobutyrate, isovalerate and 2-methylbutyrate are specifically attributed to the degradation of branched

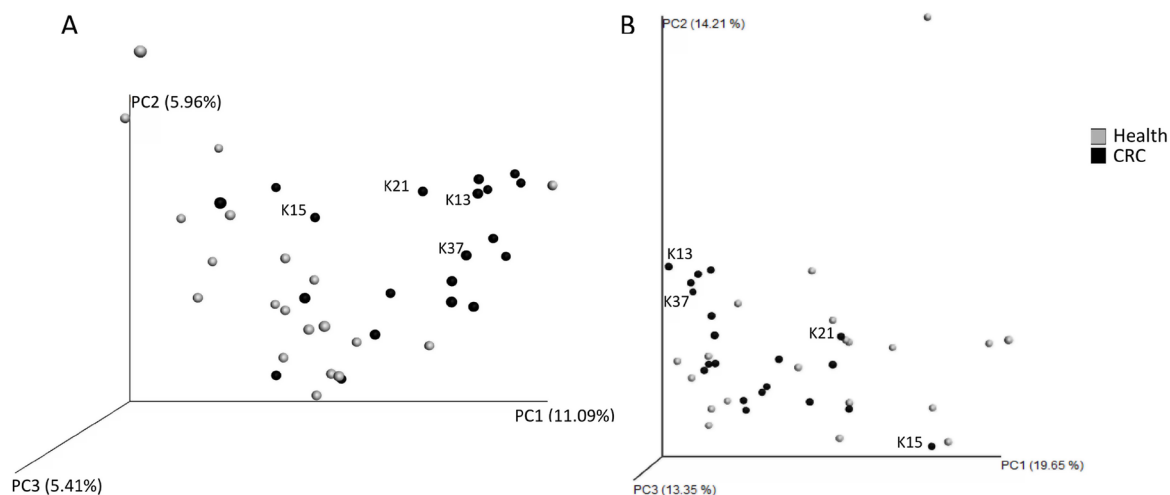
amino acids (BAA, valine, leucine and isoleucine) [35]. In the past, BCFA and valerate have been associated with an augmented risk of developing CRC principally because protein fermentation has been tied in with a high protein intake [29]. However, contrary to other microbial and chemical products such as hydrogen sulphide, *p*-cresol, phenol, haem iron, *N*-nitroso compounds, polycyclic aromatic hydrocarbons, and heterocyclic amines which are reported to be detrimental [29, 31, 36], the isoacids although indicative of putrescible fermentation, are not associated with cell toxicity [35]. The World Health Organization recently classified red meat and processed meat as carcinogens however no mechanisms or causal link have yet been established [36]. Additionally, a recent publication showed an increase in faecal BCFA after a high protein intake but no toxicity [37]. Moreover, the presence of a small amount of branched SCFA and valerate (2–4 mmol/kg) is of normal occurrence in healthy adult individuals [38] and neonates [39]. Increased transit time in the GI tracts has also been associated

with a high concentration of putrefactive products [27]; however contrary to BCFA, the levels of putrescine, N-acetylputrescine, cadaverine and p-cresol were not increased suggesting no enhanced protein putrefaction in our study. Interestingly, Andrieux and colleagues [40] proposed an alternative source of BCFA production. They reported an increase in isobutyrate and isovalerate levels with age and attributed the differences to a change in bacterial mucin degradation. Concomitantly, a high degree of correlation exists between the levels of isobutyrate and isovalerate regardless of the host species [21, 41] which suggests the presence of a universal substrate. In line with those findings our data showed strong correlations between isobutyrate and isovalerate, isobutyrate and valerate and isovalerate and valerate (Supplementary Figure 4). There are two main sources of endogenous microbial substrate: intestinal sloughed cells and host glycan mucins [40, 41]. Production of BCFA from sloughed cells is a plausible event since intestinal tissue is particularly rich in BAA [42]. It is also tempting to speculate that changes in microbial mucolytic activity could occur in CRC since an increase in abundance of mucinophilic bacteria (*Fusobacterium nucleatum* and *Akkermansia spp*) has recently been reported [31, 43]. A causal link has not yet been determined [43] but opportunistic mucin degradation is a theory that has been proposed [44]. Proline and BAA are abundant constituents of human mucin [45], thus microbial mucolytic activity followed by the fermentation of proline, valine and leucine could contribute to an increase in valerate [46] and BCFA respectively. Phenylacetate which originates from phenylalanine is another degradation metabolite associated with a high intake of protein [29, 47] but as with BCFA no toxicity to epithelial cells has yet been reported [48].

Similar to BAA, phenylalanine is an abundant component of gut tissue [42]. This infers the possibility that a portion of phenylacetate may originate from the bacterial degradation of slough cells. The detection of a smaller amount of amino acids, sugars, secondary bile acids and other bacterial products (alcohols, polyols, amines) in faecal extracts is an expected occurrence [14, 18–19, 47].

Previous studies reported higher levels of amino acids in the extracts from the CRC patients [14, 16, 19–20] but we found an inverse trend of lower abundance of amino acids, sugars, secondary bile acids and other compounds (Table 2). This general metabolic decrease may mirror a lower “normal” bacterial activity due to the aetiology of the disease which could result in a lower concentration of metabolites in stool.

The beta-diversity analysis of the microbiota of CRC patients and healthy controls indicated that although it seemed that the faecal samples clustered by health status in the unweighted analysis, this separation of samples was less distinct when relative abundance of taxa was taken into consideration (Figure 3). This may indicate that CRC patients share a common pool of taxa, but at varying abundances. Further investigation indicated that at the family level, CRC patients harboured a greater proportion of Ruminococcaceae and a lower relative abundance of Lachnospiraceae compared to healthy controls, both of which are butyrate-producing members of the Clostridiales order. Multiple members of the Clostridiales, including *Oscillospira*, were present at higher proportions in the microbiota of CRC patients compared to healthy controls, as has been identified previously [49]. A more thorough investigation into the Clostridiales and their metabolic products may shed light on how important bacteria within this order are in colorectal cancer. The methanogenic



**Figure 3: Beta-diversity analysis of faecal microbiota of healthy controls (grey) and colorectal cancer patients (black).** The data-points associated with the subset of CRC patients (K13, K15, K21, & K37) identified from the metabolomic analyses are labelled. (A) unweighted beta-diversity analysis and (B) weighted beta-diversity analysis were performed using the Unifrac metric in QIIME 1.9.1, and visualised as 3D principal coordinates analysis plots using Emperor.

**Table 3: Statistically significant taxa that differ between healthy and colorectal cancer patients**

Microbial taxa	Healthy (%)	CRC (%)	P value
o_Clostridiales	2.14 ± 2.96	4.66 ± 4.14	0.027
o_RF39	0.20 ± 0.55	1.95 ± 2.26	0.006
o_Clostridiales; f_Ruminococcaceae	1.29 ± 1.62	3.76 ± 2.47	0.001
f_Christensenellaceae	0.12 ± 0.31	0.81 ± 0.75	0.002
f_Mogibacteriaceae	0.13 ± 0.14	0.37 ± 0.26	0.002
f_Coriobacteriaceae	0.20 ± 0.19	0.58 ± 0.49	0.001
f_Erysipelotrichaceae;g_Clostridium	0.45 ± 1.07	0.05 ± 0.15	0.044
f_Ruminococcaceae;g_Ruminococcus	5.98 ± 6.70	11.40 ± 5.23	0.015
g_Methanobrevibacter	<0.01 ± 0.03	0.10 ± 0.15	0.01
g_Parabacteroides	1.47 ± 1.77	0.62 ± 0.52	0.014
g_Clostridium	0.37 ± 0.53	0.05 ± 0.07	0.012
g_Peptostreptococcus	<0.01 ± 0.01	0.05 ± 0.08	0.029
g_Anaerofilum	<0.01 ± 0	0.03 ± 0.05	0.011
g_Oscillospira	1.23 ± 1.07	2.64 ± 1.89	0.006
g_Sporobacter	0 ± 0	<0.01 ± 0.01	0.045
g_Eubacterium	1.58 ± 3.01	4.21 ± 6.21	0.049
g_cc_115	<0.01 ± 0.01	0.07 ± 0.16	0.03

Preceding letter indicates taxonomic level: o = order; f = family; g = genus. Values shown are mean ± SD.

archaea *Methanobrevibacter* was found to be present in 50% (10/20) of the CRC microbiota, compared to 10% (2/20) of the healthy controls. This taxon has previously been linked to CRC and has been considered a putative causal agent of various cancers [49], however further studies are required to elucidate the importance of this taxon in cancer, and whether gut Eukaryotes, such as fungi and protists, may play a role. A subset of 5 patients were associated to the highest levels of iso short chain fatty acids and phenylacetate. Interestingly, the bacterial taxa that were present at significantly different proportions in the proposed CRC subset (Supplementary Table 3), compared to the remaining CRC patients, have been associated with the human intestinal mucosa, inflammation, and/or an increased risk of CRC [50–51]. It was not possible to conclude further as to why those values were consistently high for those patients. Linking the microbiota and metabolomics data was attempted and although there was a weak but significant link between the two microbiota and NMR data sets, no further separation between the two groups of samples was detected (Supplementary Figure 5). Nor was a good correlation found in the heap map correlating individual metabolite with individual microbial trait (not shown).

Our sample was derived from a population-based screening programme and had each already screened

positive for further investigation using FOBt and were scheduled for colonoscopy, hence our findings suggest further risk stratification using faecal metabolites might be possible within this cohort. The groups were age and sex matched, and so this cannot explain any observed differences, but we could not control for differences in lifestyle including diet which may account for our findings.

Problems of data dispersal are often encountered with chemical data originating from human samples. The <sup>1</sup>H NMR spectra were characterized by heterogeneous profiles within each group (Figure 1) which showed large intra-group variability and some large outliers for specific metabolites. Nevertheless, strong evidence for differences between controls and cases for several metabolites was seen. In our cross-validated PLS-DA analysis faecal metabolites were able to discriminate between CRC and control patients with a C-statistic of 0.8, with sensitivity of 74% and specificity of 76% (classification accuracy of 75% in our sample) when a threshold of probability > 0.5 was used as a cut-off. This is less powerful than existing faecal tests, but our sample had already screened positive so we do not know what the discriminatory power would be in an unselected population. A 2014 review of faecal immunochemical tests (FIT) suggested a combined sensitivity of 79% and specificity of 94%, while faecal



occult blood tests [52]. Nevertheless, our findings, if repeated in larger cohorts, suggest that faecal metabolite profiles might augment existing markers to produce a more reliable non-invasive tests and that further investigation in this area is needed.

In summary, our findings clearly demonstrate that there are significant alterations in the metabolite composition of faecal extracts from patients with CRC compared to controls. As has been reported previously, we confirmed an increase in levels of isovalerate, isobutyrate and valerate that could originate from an increase in intestinal slough cells utilisation or an increase in mucolytic activity from a subset of microbes. We also reported a decrease in levels of amino acids, sugars and various microbial products (amines, alcohols and secondary bile acids) that could be attributed to a possible generalised reduction in the metabolic activity of gut bacteria. The mechanisms underlying the observed changes are still unidentified and require further investigation.

## MATERIALS AND METHODS

### Clinical characteristics of patients

Patients referred to Addenbrookes Hospital in Cambridge, UK were enrolled after having received information about the study and given their written informed consent. Stool specimens were collected as part of a study involving the National Health Service Bowel cancer screening programme (Cambridge 2 LREC reference: 08/H0308/13). The stool samples included in the study were collected between 2009 and 2013.

All patients involved in the study had a positive Faecal occult blood test (FOBT) and were invited for a colonoscopy within the National Health Service Bowel cancer screening programme (NHSBCSP). The stool samples were collected prior to the patients starting the bowel preparation for colonoscopy and stored at  $-80^{\circ}\text{C}$  immediately on arrival. Patients who had biopsy proven CRC were classified as cancers in this study. Those classified as healthy had no evidence of CRC on colonoscopy. Out of these, two sets of stool samples of patients who had colorectal cancer (CRC) and age and sex matched normal controls were analysed. The first set of samples from 40 individuals (20 CRC patients and 20 healthy controls) was analysed in 2012. A second larger set of samples ( $n = 60$ , 30 CRC and 30 healthy patients) was then selected and analysed in 2014. Patient demographics are shown in Table 1.

### Sample preparation

To get an adequate representation of the sample, a total of 50 g was collected from 5 different portions of the whole frozen stool block and homogenised. Samples were aliquoted in duplicates (50 mg  $\pm$  1 mg). The first aliquot was

lyophilised to measure the water content which ranged from 62–89% of fresh weight for set 1 and 51–93% for set 2. To obtain the normalised metabolomics data, each metabolite value was multiplied by the ratio obtained by dividing the sample water content and the maximum water content found in the set it belongs to. The ratio factors ranged between 0.69 and 1 for set 1 and 0.59 and 1 for set 2. In each set, the values of the sample with the maximum water content were left unchanged (they were multiplied by 1) and the values of the other samples were multiplied by a factor comprised between 0.59 and 1 to compensate for their lower water content. The data were thus normalised to dry weight. The second aliquot was thawed at room temperature and prepared for  $^1\text{H}$  NMR spectroscopy by mixing the faecal aliquot with 600  $\mu\text{L}$  NMR buffer (0.26 g  $\text{NaH}_2\text{PO}_4$  and 1.41 g  $\text{K}_2\text{HPO}_4$ ) made up in 100%  $\text{D}_2\text{O}$  (100 ml), containing 0.1%  $\text{NaN}_3$  (100 mg), and 1 mM sodium 3-(Trimethylsilyl)propionate-*d*4, (TSP) (17 mg) as a chemical shift reference. The sample was mixed, centrifuged and 500  $\mu\text{L}$  was transferred into a 5-mm NMR tube for spectral acquisition. The  $^1\text{H}$  NMR spectra were recorded at 600 MHz on a Bruker Avance spectrometer (Bruker BioSpin GmbH, Rheinstetten, Germany) running Topspin 3.2 software and fitted with a TCI probe. Each  $^1\text{H}$  NMR spectrum was acquired with 2816 scans, a spectral width of 12300 Hz and an acquisition time of 2.7 s and delay time of 3 s. The “noesygppr1d” presaturation sequence was used to suppress the residual water signal with a low-power selective irradiation at the water frequency during the recycle delay. Spectra were transformed with a 0.3-Hz line broadening, manually phased, baseline corrected, and referenced by setting the TSP methyl signal to 0 ppm. Spectra were prepared for statistical analysis using the Bruker AMIX software v3.9. The “underground removal tool” of AMIX was applied to all spectra (filter width = 20 Hz) to remove the broad irregular envelope that extends from  $\sim 0.7$  to 4.5 ppm. Metabolites were identified using information found in the literature [18, 23] or on the web (Human Metabolome Database, <http://www.hmdb.ca/>) and by use of the 2D-NMR methods, COSY, HSQC, and HMBC. The metabolites were quantified using the NMR Suite v7.6. Profiler (Chenomx, Inc., Edmonton, Canada).

### Statistical analysis

Statistical analysis was conducted for both sets of patients concurrently. Data were described as mean, standard deviation, median and quartiles for each group (CRC and control) and the relationship between metabolites as the ratio of means between each group.

### Univariate analysis

The distribution of each metabolite concentration across participants was heavily skewed, the presence of large numbers of tied values made standard non-parametric

tests impractical and large outliers in some groups made permutation tests invalid. Hence a separate Box-Cox transformation was applied to each metabolite, adding the minimum non-zero value for each metabolite to each value to enable zero values to be transformed. Visual inspection showed that the Box-Cox transformation removed skew for most metabolites well and stabilised variances across groups including outliers. Hence *t*-tests were then applied to transformed values in order to calculate the statistical significance of differences between metabolite concentrations in each group. In cases where there were large numbers of zeros the transformation was not able to remove skew, but a sensitivity analysis was conducted using Fishers exact tests comparing the proportion of non-zero metabolite values between cancer and control groups; *p*-values from Fishers exact tests in these cases were close to *p*-values from *t*-tests of transformed data supporting the validity of the *t*-tests. *P*-values from *t*-tests are reported along with adjusted *p*-values corrected for multiple testing using the procedure of Benjamini and Hochberg.

### Multivariate analysis

Scaled principal components analysis of Box-Cox transformed concentrations was used to estimate the relationships between metabolites, and to show the relationship between the first two principal components and cancer status.

To test whether metabolite concentrations predict cancer status, scaled partial least squares discriminant analysis (PLS-DA) of transformed values was then conducted. First, cross-validation was used within each set to select the number of components to use. PLS regression models using the optimal number of components were then estimated in each set and then validated in the other graphically and using C-statistics. A final model was then validated with data from both sets combined using 11-fold cross-validation, whereby the combined dataset was randomly split into 11 sets of 9 observations each, with the predictions for each group based on models estimated in the other 10. For this validation the Box-Cox transformations were recalculated within each group before model estimation. All metabolite analysis was conducted using R statistical software version 3.5.0. Sensitivity and specificity at different thresholds and C-statistics were calculated using the pROC package. Transformations and PLS-DA model estimation were conducted using the caret package.

### 16S rRNA gene sequencing analysis

The gut microbiota composition of the first cohort (20 CRC versus 20 age and sex matched healthy patients) was assessed using 454 pyrosequencing. The FastDNA SPIN Kit for Soil (MP Biomedicals, UK) was

used following the manufacturer's instructions, with an additional bead-beating step, to extract the microbial DNA from the collected faecal samples. The quality and yield of the DNA was assessed using gel electrophoresis, and the NanoDrop ND-1000 UV/vis spectrophotometer (NanoDrop Technologies, Inc., USA), respectively. The DNA was sent to the Animal Health and Veterinary Laboratories Agency (UK), where the V4 and V5 regions of the 16S rRNA genes were amplified using the U515F (5'-GTGYCAGCMGCCGCGGTA) and U927R (5'-CCCYCAATTCMTTTRAGT) primers, prior to the amplicons being subjected to 454 pyrosequencing [53]. Analysis of the sequencing reads was performed using Quantitative Insights Into Microbial Ecology (QIIME) 1.9.1 software and RDP classifier (version 2.10) 16S rRNA gene sequence database [54]. All sequences were filtered to meet the following criteria: read length between 200 and 1,000 bp; maximum of 6 ambiguous bases; minimum average quality score of 25 within a 50 bp window; and exact match to primer sequences. ChimeraSlayer was used to filter the trimmed reads for chimeric sequences, RDP classifier enabled microbial taxonomy assignment with a confidence value threshold of 50%, and the trimmed reads clustered into operational taxonomic units at 97% identity level.

### ACKNOWLEDGMENTS AND FUNDING

We thank the patients who participated in the study. This work was supported by the BBSRC BBSRC Institute Strategic Programme Gut Health and Food Safety BB/J004529/1 and a Clinician Scientist fellowship from Cancer Research UK (AI).

### CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

### REFERENCES

1. Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Freddie Bray F. Global patterns and trends in colorectal cancer incidence and mortality. *Gut*. 2017; 66:683–691.
2. American cancer society cancer facts and figures. American Cancer Society. 2016.
3. Brenner H, Kloor M, Pox CP. Colorectal cancer. *Lancet*. 2014; 383:1490–1502.
4. Bailey JR, Aggarwal A, Imperiale TF. Colorectal cancer screening: Stool DNA and other non-invasive modalities. *Gut Liver*. 2016; 10:204–211.
5. Duffy MJ. Carcinoembryonic antigen as a marker for colorectal cancer: Is It Clinically Useful. *Clin Chem*. 2001; 47:624–630.

6. Vukobrat-Bijedic Z, Husic-Selimovic A, Sofic A, Bijedic N, Bjelogrljic I, Gogov B, Mehmedovic A. Cancer antigens (CEA and CA 19-9) as markers of advanced stage of colorectal carcinoma. *Med Arh.* 2013; 67:397–401.
7. Hadjipetrou A, Anyfantakis D, Galanakis CG, Kastanakis M, Kastanakis S. Colorectal cancer, screening, and primary care: a mini literature review. *World J Gastroenterol.* 2017; 23:6049–6058.
8. Qiu Y, Cai G, Zhou B, Li D, Zhao A, Xie G, Li H, Cai S, Xie D, Huang C, Ge W, Zhou Z, Xu LX, et al. A distinct metabolic signature of human colorectal cancer with prognostic potential. *Clin Cancer Res.* 2014; 20:2136–2146.
9. Jiménez B, Mirnezami R, Kinross J, Cloarec O, Keun HC, Holmes E, Goldin RD, Ziprin P, Darzi A, Nicholson JK. 1H HR-MAS NMR spectroscopy of tumor-induced local metabolic “field-effects” enables colorectal cancer staging and prognostication. *J Proteome Res.* 2013; 12:959–68.
10. Qiu Y, Cai G, Su M, Chen T, Zheng X, Xu Y, Ni Y, Zhao A, Xu LX, Cai S, Jia W. Serum Metabolite profiling of human colorectal cancer using GC-TOFMS and UPLC-QTOFMS. *J Proteome Res.* 2009; 8:4844–4850.
11. Tan B, Qiu Y, Zou X, Chen T, Xie G, Cheng Y, Dong T, Zhao L, Feng B, Hu X, Xu LX, Zhao A, Zhang M, et al. Metabonomics identifies serum metabolite markers of colorectal cancer. *J Proteome Res.* 2013; 12:3000–3009.
12. Zhu J, Djukovic D, Deng L, Gu H, Himmati F, Chiorean EG, Raftery D. Colorectal cancer detection using targeted serum metabolic profiling. *J Proteome Res.* 2014; 13:4120–4130.
13. Cheng Y, Xie G, Chen T, Qiu Y, Zou X, Zheng M, Tan B, Feng B, Dong T, He P, Zhao L, Zhao A, Xu LX, et al. Distinct urinary metabolic profile of human colorectal cancer. *J Proteome Res.* 2012; 11:1354–1363.
14. Monleon D, Morales JM, Barrasa A, López JA, Vázquez C, Celda B. Metabolite profiling of fecal water extracts from human colorectal cancer. *NMR Biomed.* 2009; 22:342–348.
15. Bezabeh T, Somorjai R, Dolenko B, Bryskina N, Levin B, Bernstein CN, Jeyarajah E, Steinhart AH, Rubin DT, Smith IC. Detecting colorectal cancer by 1H magnetic resonance spectroscopy of fecal extracts. *NMR Biomed.* 2009; 22:593–600.
16. Weir TL, Manter DK, Sheflin AM, Barnett BA, Heuberger AL, Ryan EP. Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults. *PLoS One.* 2013 8:e70803.
17. Phua LC, Chue XP, Koh PK, Cheah PY, Ho HK, Chan EC. Non-invasive fecal metabonomic detection of colorectal cancer. *Cancer Biol Ther.* 2014; 15:389–397.
18. Amiot A, Dona AC, Wijeyesekera A, Tournigand C, Baumgaertner I, Lebaleur Y, Sobhani I, Holmes E. 1H NMR spectroscopy of fecal extracts enables detection of advanced colorectal neoplasia. *J Proteome Res.* 2015; 14:3871–3781.
19. Lin Y, Ma C, Liu C, Wang Z, Yang J, Liu X, Shen Z, Wu R. NMR-based fecal metabolomics fingerprinting as predictors of earlier diagnosis in patients with colorectal cancer. *Oncotarget.* 2016; 7:29454–464. <https://doi.org/10.18632/oncotarget.8762>.
20. Wang X, Wang J, Rao B, Deng L. Gut flora profiling and fecal metabolite composition of colorectal cancer patients and healthy individuals. *Exp Ther Med.* 2017; 13:2848–2854.
21. Siigur U, Norin KE, Allgood G, Schlagheck T, Midtvedt T. Concentrations and correlations of faecal short-chain fatty acids and faecal water content in man. *Microb Ecol Health Dis.* 1994; 7:287–294.
22. Gardana C, Del Bo' C, Simonetti P. Validation and application of an ultrahigh-performance liquid chromatographic-Orbitrap mass spectrometric method for the simultaneous detection and quantification of volatile and non-volatile organic acids in human faecal samples. *J Pharm Biomed Anal.* 2017; 141:46–51.
23. Gratton J, Phetcharaburanin J, Mullish BH, Williams HR, Thursz M, Nicholson JK, Holmes E, Marchesi JR, Li JV. An optimized sample handling strategy for metabolic profiling of human feces. *Anal Chem.* 2016; 88:4661–4668.
24. Bouatra S, Aziat F, Mandal R, Guo AC, Wilson MR, Knox C, Bjorndahl TC, Krishnamurthy R, Saleem F, Liu P, Dame ZT, Poelzer J, Huynh J, et al. The human urine metabolome. *PLoS One.* 2013; 8:e73076.
25. Psychogios N, Hau DD, Peng J, Guo AC, Mandal R, Bouatra S, Sinelnikov I, Krishnamurthy R, Eisner R, Gautam B, Young N, Xia J, Knox C, et al. The human serum metabolome. *PLoS One.* 2011; 6:e16957.
26. Le Gall G, Noor SO, Ridgway K, Scovell L, Jamieson C, Johnson IT, Colquhoun IJ, Kemsley EK, Narbad A. Metabolomics of fecal extracts detects altered metabolic activity of gut microbiota in ulcerative colitis and irritable bowel syndrome. *J Proteome Res.* 2011; 10:4208–4218.
27. Ikeda N, Saito Y, Shimizu J, Ochi A, Mizutani J, Watabe J. Variations in concentrations of bacterial metabolites, enzyme activities, moisture, pH and bacterial composition between and within individuals in faeces of seven healthy adults. *J Appl Bacteriol.* 1994; 77:185–194.
28. Tanida N, Hikasa Y, Shimoyama T, Setchell KDR. Comparison of faecal bile acid profiles between patients with adenomatous polyps of the large bowel and healthy subjects in Japan. *Gut.* 1984; 25:824–32.
29. Louis P, Hold GL, Flint HJ. The gut microbiota, bacterial metabolites and colorectal cancer. *Nat Rev Microbiol.* 2014; 12:661–672.
30. De Filippis F, Pellegrini N, Vannini L, Jeffery IB, La Stora A, Laghi L, Serrazanetti DI, Di Cagno R, Ferrocino I, Lazzi C, Turroni S, Cocolin L, Brigidi P, et al. High-level adherence to a Mediterranean diet beneficially impacts the gut microbiota and associated metabolome. *Gut.* 2016; 65:1812–1821.
31. O’Keefe SJD. Diet, microorganisms and their metabolites, and colon cancer. *Nat Rev Gastroenterol Hepatol.* 2016; 13:691–706.

32. Tailford LE, Crost EH, Kavanaugh D, Juge N. Mucin glycan foraging in the human gut microbiome. *Front Genet.* 2015; 6:81.
33. Ohigashi S, Sudo K, Kobayashi D, Takahashi O, Takahashi T, Asahara T, Nomoto K, Onodera H. Changes of the intestinal microbiota, short chain fatty acids, and fecal pH in patients with colorectal cancer. *Dig Dis Sci.* 2013; 58:1717–1726.
34. Chen HM, Yu YN, Wang JL, Lin YW, Kong X, Yang CQ, Yang L, Liu ZJ, Yuan YZ, Liu F, Wu JX, Zhong L, Fang DC, et al. Decreased dietary fiber intake and structural alteration of gut microbiota in patients with advanced colorectal adenoma. *Am J Clin Nutr.* 2013; 97:1044–1052.
35. Blachier F, Mariotti F, Huneau JF, Tomé D. Effects of amino acid-derived luminal metabolites on the colonic epithelium and physiopathological consequences. *Amino Acids.* 2007; 33:547–562.
36. Johnson IT. The cancer risk related to meat and meat products. *Br Med Bull.* 2017; 121:73–81.
37. Windey K, De Preter V, Louat T, Schuit F, Herman J, Vansant G, Verbeke K. Modulation of protein fermentation does not affect fecal water toxicity: A randomized cross-over study in healthy subjects. *PLoS One.* 2012; 7:e52387
38. Ou J, DeLany JP, Zhang M, Sharma S, O'Keefe SJ. Association between low colonic short-chain fatty acids and high bile acids in high colon cancer risk populations. *Nutr Cancer.* 2012; 64:34–40.
39. Pourcyrus M, Nolan VG, Goodwin A, Davis SL, Buddington RK. Fecal short-chain fatty acids of very-low-birth-weight preterm infants fed expressed breast milk or formula. *J Pediatr Gastroenterol Nutr.* 2014; 59:725–731.
40. Andrieux C, Membré JM, Cayuela C, Antoine JM. Metabolic characteristics of the faecal microflora in humans from three age groups. *Scand J Gastroenterol.* 2002; 37:792–798.
41. Cardona ME, Collinder E, Stern S, Tjellström B, Norin E, Midtvedt T. Correlation between faecal iso-butyric and iso-valeric acids in different species. *Microb Ecol Health Dis.* 2005; 17:177–182.
42. Wang Y, Holmes E, Comelli EM, Fotopoulos G, Dorta G, Tang H, Rantalainen MJ, Lindon JC, Corthésy-Theulaz IE, Fay LB, Kochhar S, Nicholson JK. Topographical variation in metabolic signatures of human gastrointestinal biopsies revealed by high-resolution magic-angle spinning 1H NMR spectroscopy. *J Proteome Res.* 2007; 6:3944–3951.
43. Amitay EL, Werner S, Vital M, Pieper DH, Höfler D, Gierse IJ, Butt J, Balavarca Y, Cuk K, Brenner H. Fusobacterium and colorectal cancer: causal factor or passenger? Results from a large colorectal cancer screening study. *Carcinogenesis.* 2017; 38:781–788.
44. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, Amiot A, Böhm J, Brunetti F, Habermann N, Hercog R, Koch M, Luciani A, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol.* 2014; 10:766.
45. Wesley A, Mantle M, Man D, Qureshi R, Forstner G, Forstner J. Neutral and acidic species of human intestinal mucin. Evidence for different core peptides. *J Biol Chem.* 1985; 260:7955–7959.
46. Allison MJ, Bryant MP. Biosynthesis of branched-chain amino acids from branched-chain fatty acids by rumen bacteria. *Arch Biochem Biophys.* 1963; 101:269–277.
47. Beaumont M, Portune KJ, Steuer N, Lan A, Cerrudo V, Audebert M, Dumont F, Mancano G, Khodorova N, Andriamihaja M, Airinei G, Tomé D, Benamouzig R, et al. Quantity and source of dietary protein influence metabolite production by gut microbiota and rectal mucosa gene expression: a randomized, parallel, double-blind trial in overweight humans. *Am J Clin Nutr.* 2017; 106:1005–1019.
48. Kim E, Coelho D, Blachier F. Review of the association between meat consumption and risk of colorectal cancer. *Nutr Res.* 2013; 33:983–994.
49. Hibberd AA, Lyra A, Ouwehand AC, Rolny P, Lindegren H, Cedgård L, Wettergren Y. Intestinal microbiota is altered in patients with colon cancer and modified by probiotic intervention. *BMJ Open Gastroenterol.* 2017; 4:e000145.
50. Bajaj JS, Hylemon PB, Ridlon JM, Heuman DM, Daita K, White MB, Monteith P, Noble NA, Sikaroodi M, Gillevet PM. Colonic mucosal microbiome differs from stool microbiome in cirrhosis and hepatic encephalopathy and is linked to cognition and inflammation. *Am J Physiol Gastrointest Liver Physiol.* 2012; 303:G675–685.
51. Zackular JP, Baxter NT, Iverson KD, Sadler WD, Petrosino JF, Chen GY, Schloss PD. The gut microbiome modulates colon tumorigenesis. *MBio.* 2013; 4:e00692–13.
52. Lee JK, Liles EG, Bent S, Levin TR, Corley DA. Accuracy of Fecal Immunochemical Tests for Colorectal Cancer: Systematic Review and Meta-analysis. *Ann Intern Med.* 2014; 160:171–205.
53. Ellis RJ, Bruce KD, Jenkins C, Stothard JR, Ajarova L, Mugisha L, Viney ME. Comparison of the Distal Gut Microbiota from People and Animals in Africa. *PLoS One.* 2013; 8:e54783.
54. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, et al. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods.* 2010; 7:335–336.