scientific reports

Check for updates

OPEN

Gut microbiota composition in colorectal cancer patients is genetically regulated

Francesca Colombo[1,6], Oscar Illescas[2,6], Sara Noci[2], Francesca Minnai[1], Giulia Pintarelli[2,5], Angela Pettinicchio[2], Alberto Vannelli[3], Luca Sorrentino[4], Luigi Battaglia[4], Maurizio Cosimelli[4], Tommaso A. Dragani[2]✉ & Manuela Gariboldi[2]

The risk of colorectal cancer (CRC) depends on environmental and genetic factors. Among environmental factors, an imbalance in the gut microbiota can increase CRC risk. Also, microbiota is influenced by host genetics. However, it is not known if germline variants influence CRC development by modulating microbiota composition. We investigated germline variants associated with the abundance of bacterial populations in the normal (non-involved) colorectal mucosa of 93 CRC patients and evaluated their possible role in disease. Using a multivariable linear regression, we assessed the association between germline variants identified by genome wide genotyping and bacteria abundances determined by 16S rRNA gene sequencing. We identified 37 germline variants associated with the abundance of the genera *Bacteroides*, *Ruminococcus*, *Akkermansia*, *Faecalibacterium* and *Gemmiger* and with alpha diversity. These variants are correlated with the expression of 58 genes involved in inflammatory responses, cell adhesion, apoptosis and barrier integrity. Genes and bacteria appear to be involved in the same processes. In fact, expression of the pro-inflammatory genes *GAL*, *GSDMD* and *LY6H* was correlated with the abundance of *Bacteroides*, which has pro-inflammatory properties; abundance of the anti-inflammatory genus *Faecalibacterium* correlated with expression of KAZN, with barrier-enhancing functions. Both the microbiota composition and local inflammation are regulated, at least partially, by the same germline variants. These variants may regulate the microenvironment in which bacteria grow and predispose to the development of cancer. Identification of these variants is the first step to identifying higher-risk individuals and proposing tailored preventive treatments that increase beneficial bacterial populations.

Colorectal cancer (CRC) is the fourth most diagnosed tumor and the second leading cause of cancer-related deaths in the world[1]. Its incidence is increasing, especially in developing countries that are undergoing modifications in lifestyle[2]. The vast majority of CRCs are considered to be sporadic[3]. Numerous studies indicated that sporadic CRC is the result of a complex interplay of genetic variants and environmental factors (reviewed in)[4]. Genome-wide association studies have provided evidence for 53 unique CRC susceptibility loci across ethnicities[5]. Among environmental factors, the gut microbiota has emerged as important for some cancers, including CRC, where an imbalance in its composition can contribute to the development of the disease[6]. Interacting closely with host epithelial cells, the microbiota can influence colorectal carcinogenesis via a variety of mechanisms, including microbe-derived factors.

Human gut microbiota composition and host metabolic functions are intimately related and mutually regulated[7]. Studies that investigated whether the microbiota composition is influenced by host genetics highlighted some degree of heritability[8,9]. Genome-wide association studies on host genetic factors associated with the microbiome composition have led to the identification of several microbial quantitative trait loci (mbQTLs). These mbQTLs are often associated with host genes that participate in nutrition-related metabolic pathways and in immune traits such as barrier integrity or inflammation[8,10]. Moreover, a study of inflammatory bowel disease (IBD) patients suggested that genetic variants associated with microbiota dysregulation can also affect

[1]Institute of Biomedical Technologies, National Research Council (ITB-CNR), Segrate, MI, Italy. [2]Genetic Epidemiology and Pharmacogenomics Unit, Department of Research, Fondazione IRCCS Istituto Nazionale dei Tumori, Via G.A. Amadeo 42, 20133 Milan, Italy. [3]Department of Surgery, Valduce Hospital, Como, Italy. [4]Colorectal Surgery Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy. [5]Present address: Azienda Ospedaliera Universitaria Integrata, UOC Laboratorio Analisi, Verona, Italy. [6]These authors contributed equally: Francesca Colombo and Oscar Illescas. ✉email: tommaso.dragani@istitutotumori.mi.it

Characteristic	Value
Age at surgery, median (range), years	64 (38–86)
Sex, n (%)	
Male	56 (60)
Female	37 (40)
Smoking habit, n (%)	
Ever smoker	38 (41)
Never smoker	47 (50)
Not available	8 (9)
Tumor site, n (%)	
Colon	38 (41)
Rectum	55 (59)
Pathological stage, n (%)	
I	23 (25)
II	23 (25)
III	33 (35)
IV	11 (12)
Not available	3 (3)
Neo-adjuvant therapy, n (%)	
Yes	14 (15)
No	79 (85)

Table 1. Clinical characteristics of 93 patients with colorectal cancer.

the immune system and induce inflammation[11]. These observations are of great importance in colorectal cancer, where both inflammation and certain bacterial populations play a role in tumorigenesis[12,13]. However, no studies associating bacterial populations with genomic variants in CRC patients have been done to date.

We characterized the microbiota in normal colonic mucosa from CRC patients, to identify bacterial populations regulated by host germline variants. Since the vast majority of genetic variants mapped outside coding regions, we looked for possible regulatory effects (i.e. affecting gene expression or splicing) of the identified mbQTLs, with the aim of understanding their functional role.

Results

Genomic DNA was obtained from resected colorectal mucosa from 95 patients with CRC and subjected to microbiota profiling and SNP genotyping. Genotyping failed in one case and another sample was excluded due to a low call rate. Therefore, mbQTL analyses were done for 93 patients (Table 1). The patients had a median age of 64 years, 60% were men, and 50% had never smoked. The tumor affected the rectum in 59% of cases. There was a broad distribution of pathological stage, with a mode of III (35% of cases). Finally, most patients (85%) had not received neo-adjuvant treatment. Among the 14 neoadjuvant-treated patients, one received only chemotherapy and one received only radiotherapy. All the other 12 patients were treated both with chemo- and radiotherapy.

Microbiota profiling of non-involved colorectal mucosa. Bacteria associated with patients' non-involved colorectal mucosa were identified by 16S rRNA gene sequencing, and species diversity was expressed with alpha and beta metrics. The mean Shannon index was higher in V1-V2-V3 16S rRNA gene sequencing data than in V4-V5-V6 data (mean, 6.2 vs. 5.6; $P=0.001$) (Supplementary Table 1). In contrast, the mean number of observed OTUs and mean Chao1 estimator were significantly lower in V1-V2-V3. Using Bray–Curtis dissimilarity to estimate beta diversity, we detected significant differences between V1-V2-V3 and V4-V5-V6 data with ANOSIM ($P=0.001$) and ADONIS ($P=0.001$). Considering the differences between V1-V2-V3 and V4-V5-V6 datasets in both alpha and beta diversities, reflecting the differential ability of different 16S variable regions in resolving specific taxonomic groups, we decided to individually analyze the V1-V2-V3 and V4-V5-V6 datasets in this study.

Alpha diversity comparisons between patients grouped according to the clinical characteristics age at surgery, sex or smoking habit showed no differences using either the V1-V2-V3 or V4-V5-V6 dataset. Differently, when patients were grouped according to tumor site, the median Chao1 estimator in the V4-V5-V6 dataset was lower for patients with tumors in the colon than in the rectum (183 vs. 194, respectively, $P=0.049$; Fig. 1A). Similarly, the median number of observed OTUs was lower in patients with colonic tumors than with rectal tumors (181 vs. 193, respectively, $P=0.049$; Fig. 1B).

A total of 469 OTUs at genus level was found in both V1-V2-V3 and V4-V5-V6. After data filtering, there were 13 and 12 OTUs, respectively, for analysis. Eight OTUs (i.e. *Bacteroides*, *Blautia*, *Coprococcus*, *Dorea*, *Faecalibacterium*, *Pseudomonas, Roseburia*, and *Ruminococcus* genera) were identified in both datasets, for 17 unique OTUs (Table 2). In multivariate linear regression, seven unique OTUs independently associated with patients' characteristics, namely age at surgery (three OTUs), smoking habit (three OTUs), and tumor site (two

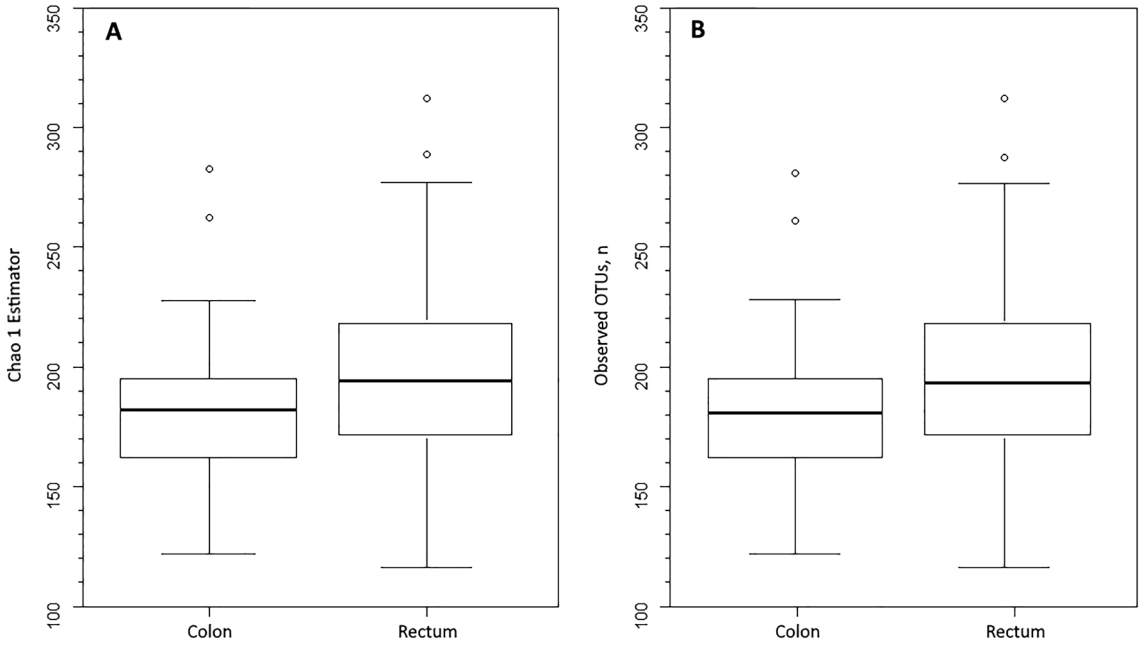

Figure 1. Microbiota diversity in resected colorectal mucosa from patients with colorectal cancer. Data grouped by tumor site, according to the V4-V5-V6 dataset. (**A**) Chao1 estimator, $P = 0.049$, Kruskal–Wallis rank-sum test with continuity correction. (**B**) Number of observed operational taxonomic units (OTUs), $P = 0.049$. The box marks the interquartile range, and the central horizontal line marks the median. Outliers (extreme values, > 1.5 times the interquartile range) are shown as circles.

OTU	V1-V2-V3 Dataset[a]	V4-V5-V6 Dataset	Clinical characteristic	Beta (*P*)[b]
Akkermansia	N	Y	None	–
Alistipes	Y	N	None	–
Bacteroides	Y	Y	None	–
Blautia	Y	Y	None	–
Collinsella	N	Y	Smoking	1.1 (0.045)[2]
Coprococcus	Y	Y	None	–
Dorea	Y	Y	Age at surgery	− 0.046 (0.023)[2]
Escherichia	Y	N	Age at surgery	0.079 (0.00074)[1]
Faecalibacterium	Y	Y	None	–
Gemmiger	N	Y	Tumor site[c]	− 0.97 (0.033)[2]
Parabacteroides	Y	N	Smoking	0.89 (0.029)[1]
Prevotella	Y	N	None	–
Pseudomonas	Y	Y	Tumor site[c]	0.93 (0.030)[2]
Robinsoniella	N	Y	None	–
Roseburia	Y	Y	Smoking	0.95 (0.035)[1]
			Age at surgery	− 0.050 (0.0063)[1]
			Age at surgery	− 0.045 (0.014)[2]
Ruminococcus	Y	Y	None	–
Stenotrophomonas	Y	N	None	–

Table 2. Association of operational taxonomic units (OTUs) and clinical characteristics. OTUs observed in > 25% of samples and with mean relative abundance > 1%, and associations between the clr-transformed relative abundance and patient characteristics, by multivariate linear regression. [a]Sequence data from the six hypervariable regions were aggregated into two datasets (V1-V2-V3 and V4-V5-V6) and analyzed separately. [b]Beta and *P*-value from [1]V1-V2-V3 or [2]V4-V5-V6 datasets. [c]Rectum/colon.

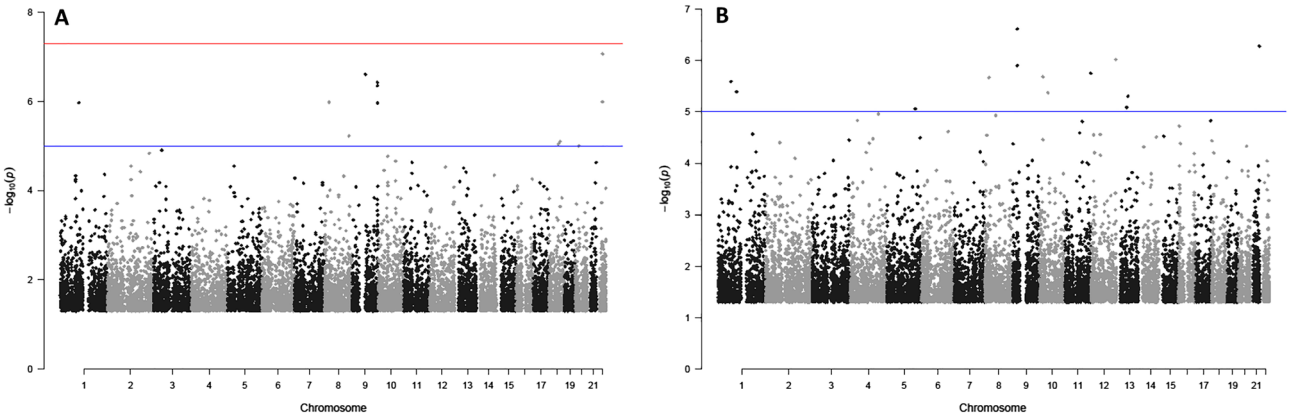

Figure 2. Manhattan plots of Shannon index-specific mbQTLs. (**A**) V1-V2-V3 dataset. (**B**) V4-V5-V6 dataset. Red horizontal line, threshold for genome-wide significance ($P = 5.0 \times 10^{-8}$). Blue horizontal line, FDR = 0.10.

OTUs). The relative abundance of *Roseburia* associated with age and smoking habit, with lower abundance in older patients (beta, -0.050 and -0.045 for V1-V2-V3 and V4-V5-V6 datasets, respectively) and greater abundant in ever smokers (beta = 0.95). *Colinsella* and *Parabacteroides* OTUs were also more abundant in ever smokers than never smokers (beta, 1.1 and 0.89, respectively). The abundance of *Dorea* decreased with age, whereas that of *Escherichia* increased. Finally, *Pseudomonas* was more abundant and *Gemmiger* less abundant in patients who had a rectal than colonic tumor.

We also observed that many OTU abundances were significantly correlated among each other (Pearson's $|r| > 0.20$ and $P < 0.05$; Supplementary Fig. 1). In the V1-V2-V3 dataset, the number of correlations per OTU ranged from three for *Blautia* to 11 for *Faecalibacterium*, *Alistipes*, and *Dorea*. In the V4-V5-V6 dataset, the number of correlations per OTU ranged from three for *Blautia* to 10 for *Roseburia*, *Ruminococcus*, and *Faecalibacterium*. In both datasets, more correlations were positive than negative.

Association of germline variants with microbiota-related quantitative traits. Genome-wide genotyping on Axiom Precision Medicine Research Arrays provided data on 856,427 variants. Of these, 17,919 were removed since they had a call rate < 98% and 536,891 variants with a MAF < 5% were also removed. Therefore, data on 301,884 germline variants were available for analysis.

Multivariable linear regression was used to combine genotype data and microbiota-related quantitative traits to identify mbQTLs. Analysis using the Shannon index identified 12 loci comprising 18 variants (FDR < 0.1) on seven unique chromosomes. Of the 18 variants, 10 (on four chromosomes) derived from V1-V2-V3 sequencing data (Fig. 2a; Supplementary Table 2); these variants all had an FDR < 0.05, and two of them (rs78578513 and rs5994535 on 22q12.3) almost reached the genome-wide significance threshold of nominal $P = 5.0 \times 10^{-8}$. The remaining eight variants derived from V4-V5-V6 sequencing (Fig. 2b; Supplementary Table 2); these variants were distributed on seven chromosomes and had an FDR < 0.1. For all these mbQTLs, there was a negative correlation between the number of minor alleles in the genotype and the Shannon index, indicating a reduction in microbial diversity with an increasing number of minor alleles. No mbQTLs were found when the Chao1 estimator and number of observed OTUs were analyzed.

When multivariable linear regression was run using the clr-transformed relative abundance of OTUs as the microbiota-related quantitative trait, we identified 37 unique mbQTLs in 23 loci on 11 chromosomes (FDR < 0.10) (Supplementary Table 2). *Bacteroides* had the highest number of unique mbQTLs (n = 28), including four (rs7815797, rs72691617, rs7817574, rs79744701) at 8q24.3 that were found in both V1-V2-V3 and V4-V5-V6 datasets (Fig. 3A, B). A single mbQTL was associated with *Ruminococcus* abundance in the V1-V2-V3 dataset (Fig. 3C). Another eight mbQTLs in seven loci were identified as being associated with *Akkermansia*, *Faecalibacterium* or *Gemmiger* abundance (Fig. 3D–F). The most significant association overall ($P = 1.23 \times 10^{-9}$) was observed between *Akkermansia* and rs4527077 on chromosome 8q24.22. This variant, together with its neighbor rs4736470 and with rs12472313 on chromosome 2, were the only three polymorphisms that correlated directly with microbial abundance, meaning that an increasing number of minor alleles of these single nucleotide polymorphisms (SNPs) associated with more abundant *Akkermansia*. All the other mbQTL variants correlated inversely with microbial abundance. Among the mbQTLs found, only two have so far been reported to be associated with a specific trait, according to the NHGRI-EBI Catalog of Human Genome-Wide Association Studies. In particular, variants rs79744701 (G > A) and rs7817574 (T > C) were both found to associate with serum levels of apolipoprotein A1 and HDL cholesterol[14,15].

Regulatory effects of mbQTLs in non-involved colorectal mucosa. Most identified mbQTLs mapped outside coding regions. To determine if the mbQTLs have genetic regulatory effects, we interrogated GTEx and Ensembl VEP databases and identified 58 genes whose expression or splicing (in any tissue) had been reported to be associated with the variants we identified as colorectal mbQTLs. With our Shannon index-specific mbQTLs, we identified 28 genes whose expression (n = 24) or splicing (n = 9) varied according to genotype (Supplementary Table 3). In particular, two genes were upregulated and two downregulated in colorectal mucosa.

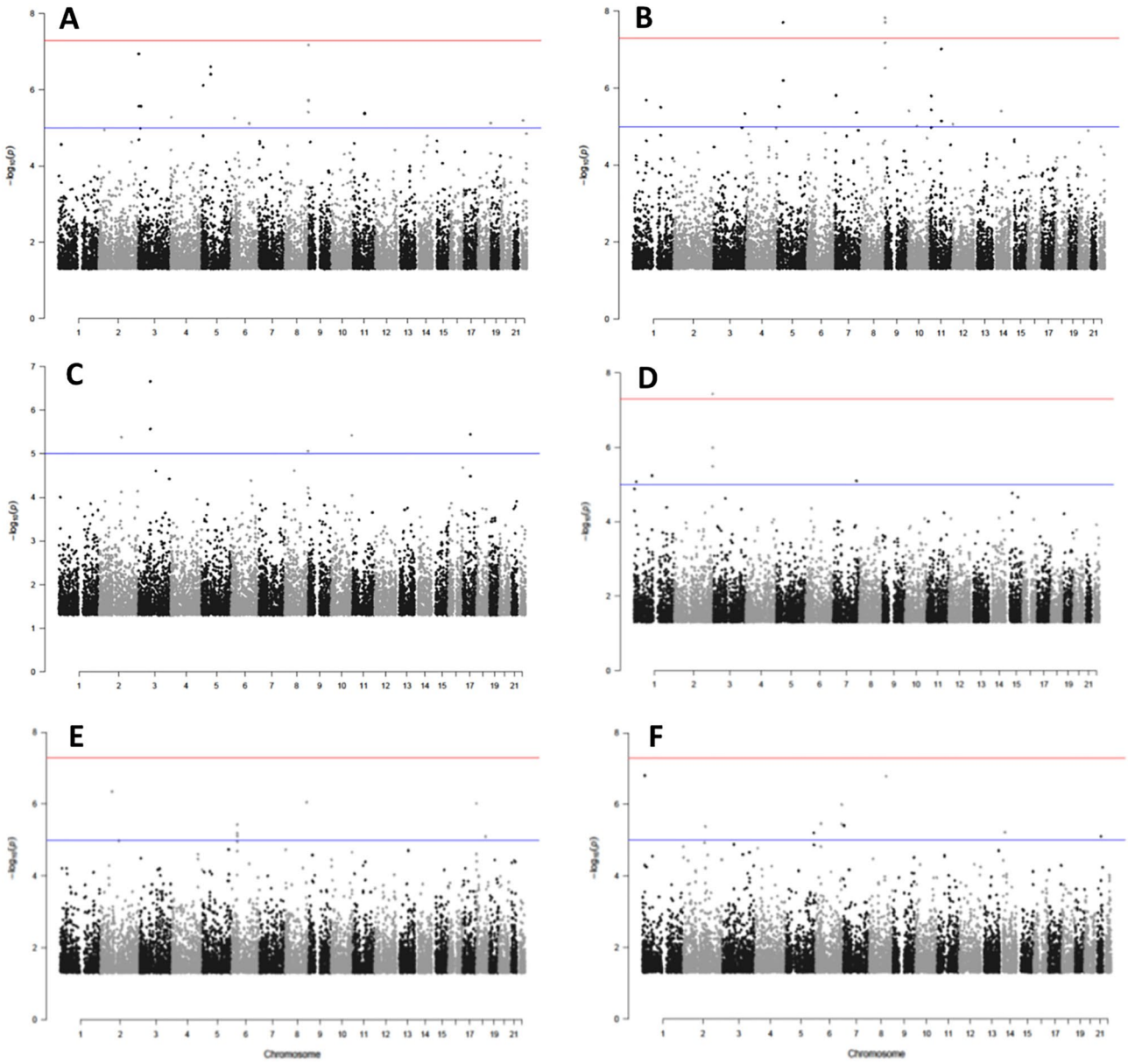

Figure 3. Manhattan plots of OTU-specific mbQTLs. (**A**) *Bacteroides* V1-V2-V3, (**B**) *Bacteroides* V4-V5-V6, (**C**) *Ruminococcus* V4-V5-V6, (**D**) *Akkermansia*, V4-V5-V6 (**E**) *Faecalibacterium* and (**F**) *Gemmiger* V4-V5-V6. Red horizontal line, threshold for genome-wide significance ($P = 5.0 \times 10^{-8}$). Blue horizontal line, FDR = 0.10.

With our OTU-specific mbQTLs, we identified 30 genes regulated by genotype (Table 3). We found 25 genes linked to *Bacteroides*-specific mbQTLs, including one upregulated and two downregulated in colorectal mucosa. We also found three genes with *Akkermansia*-specific and one each with *Faecalibacterium*- and *Gemmiger*-specific mbQTLs. Of the 58 identified genes, 29 are known to be differentially expressed in colorectal adenocarcinoma, with respect to normal tissue, according to The Cancer Genome Atlas database (Supplementary Table 4). Of these, 10 were associated with the Shannon index and 19 associated with *Bacteroides, Akkermansia* or *Gemmiger* abundance in this study.

Ontology annotations in the DAVID Bioinformatic Database indicate that the products of 23 of the 58 genes localize to the membrane, 10 are extracellular, and 17 are glycosylated (Supplementary Table 5). Also, four genes (*GAL*, *GSDMD*, *IL1RAP* and *ITGB2*) are associated with the inflammatory response, three with cell adhesion (*ATP1B1, LOXL2* and *ITGB2*) and three with apoptosis (*THOC1, TP63* and *ITGB2*) (Supplementary Table 6). Four variants associated with the Shannon index and seven with *Bacteroides,* as well as one variant for *Akkermansia*, *Faecalibacterium* and *Gemmiger* each, map in predicted promoter or enhancer regions or in CTCF (CCCTC-binding factor) binding sites (Supplementary Table 7). The single variant associated with *Ruminococcus* has no reported effects on the expression of any gene nor is found in a regulatory region. Finally, we observed the pathogenic species *Bacteroides fragilis* in 36 patients, with an abundance greater than 1% in 13 of them, according to data from 16S rRNA gene V1-V2-V3 regions (Supplementary Table 8).

OTU	SNPs	Gene	Locus	Expression change[a]	Splicing change
Bacteroides	rs111574827, rs3790893	ADGRL2	1p31.1	–	IR
	rs12563507	ATP1B1	1q24.2	↑	–
	rs12563507	NME7	1q24.2	↑	AS
	rs12563507	CCDC181	1q24.2	–	AS
	rs711582	LRRN1	3p26.1	↑*	IR
	rs34222640	ITPR1	3p26.1	–	IR
	rs78482149	TP63	3q28	–	IR
	rs78482149, rs56201661	IL1RAP	3q28	↑	–
	rs28368180, rs35361432, rs116871956	SLC38A9	5q11.2	↓*	AS, IR
	rs75284212	RP11-274B21.2	7q32.1	↓	–
	rs72691617, rs7817574, rs79744701	LY6H	8q24.3	↓	–
	rs7815797, rs72691617, rs7817574,	GPIHBP1	8q24.3	↓*	–
	rs79744701	ZFP41	8q24.3	↓	–
	rs79744701	GLI4	8q24.3	↑	–
	rs79744701	MINCR	8q24.3	↑	–
	rs79744701	TOP1MT	8q24.3	↑	AS
	rs7815797, rs72691617, rs7817574	ZC3H3	8q24.3	↓	–
	rs7815797, rs72691617, rs7817574	GSDMD	8q24.3	↓	–
	rs7815797, rs72691617, rs7817574	RP11-661A12.5	8q24.3	↓	–
	rs3833782	RPL27A	11p15.4	–	IR
	rs3833782, rs1104774	STK33	11p15.4	↓	AS
	rs3833782, rs1104774	DENND2B	11p15.4	–	IR
	rs74830761	GAL	11q13.3	↓	–
	rs4901170	RP11-280K24.4	14q22.1	–	AS
	rs4901170	GNG2	14q22.1	–	IR
Akkermansia	rs598300	ROCK1P1	18p11.32	↑	–
	rs598300	USP14	18p11.32	↓	–
	rs598300	THOC1	18p11.32	–	IR
Faecalibacterium	rs7526230	KAZN	1p36.21	–	IR
Gemmiger	rs73129818	IQCA1	2q37.3	↓↑	–

Table 3. Genetic regulatory effects of operational taxonomic unit (OTU)-specific mbQTLs. [a]Upregulation (↑) or downregulation (↓) in the colon (*) or other tissues, associated with an increasing frequency of minor alleles (GTEx database; FDR < 0.05). ↓↑ Divergent results in two tissues; –, no change reported. AS, alternative splicing-derived isoforms (GTEx database); IR, intron retention (Ensembl VEP).

Discussion

This study investigated associations between the microbiota of non-involved colorectal mucosa in CRC patients and both patients' clinical characteristics and genetic variants. We found that changes in abundance of the genera *Dorea*, *Roseburia* and *Escherichia* were associated with age at surgery, *Gemmiger* with tumor site, and *Collinsella* and *Parabacteroides* with smoking habit. Moreover, changes in the abundance of *Bacteroides*, *Akkemansia*, *Faecalibacterium* and *Gemmiger* correlated with germline variants already known to be associated with the expression of 30 genes, while germline variants associated with the Shannon index were correlated with 28 genes. These genes participate in mechanisms of great relevance in cancer development, namely the inflammatory response, apoptosis, cell adhesion and epithelial barrier function, and 29 are known to be dysregulated in CRC. Interestingly, similar mechanisms were associated with *Bacteroides, Akkermansia* or *Gemmiger* abundance.

The risk and incidence of CRC and inflammatory diseases are higher in older patients[1]. Accordingly, we found that two genera with anti-inflammatory properties, namely the short-chain fatty acid producers *Roseburia* and *Dorea*[16,17], were negatively associated with age, while the potentially pro-inflammatory *Escherichia* was positively associated[18,19]. A smoking habit was positively associated with the abundance of *Collinsella*, *Parabacteroides* and *Roseburia* genera. *Collinsella* has been reported to increase mucosal barrier permeability and induce the expression of pro-inflammatory IL-17 and NFkB1 in patients with rheumatoid arthritis[20]. Comparing patients with rectal vs. colonic tumors, we found that the opportunistic pathogen *Pseudomonas*[21], was more abundant in the rectum. Instead, *Gemmiger* was more abundant in the colon; the abundance of this genus has been shown to be reduced in patients with inflammatory diseases[22]. These observations are consistent with findings that the gut bacterial composition varies along the gastrointestinal tract[23].

This study found 18 mbQTLs associated with bacterial diversity (Shannon index) and 37 mbQTLs associated with the abundance of *Bacteroides*, *Akkemansia*, *Faecalibacterium*, *Gemmiger* and *Ruminococcus.* Genes associated with mbQTLs were already known to influence the microenvironment by regulating host metabolism, immunity

and cell and tissue barrier integrity, potentially favouring the growth of specific bacteria[8,10]. Of all the mbQTLs identified here, 34 were already known to be expression or splicing QTLs (here, referred to as "mbesQTLs") for 58 genes. The products of 23 of these genes localize to the cell membrane and may alter gut barrier–microbiota interactions. Eleven of these genes are known to participate in immune mechanisms, including the innate response, inflammation and barrier integrity[24]. Additionally, six genes are known to be involved in apoptosis and transcription regulation, two pathways commonly dysregulated in cancer.

Our study suggests that a pro-inflammatory environment reduces diversity of the normal colonic mucosa-associated microflora, as already reported for IBD[25]. Indeed, among the 28 genes with Shannon index-associated mbesQTLs, *BPIFC* (bactericidal/permeability-increasing fold-containing family C protein), *PIK3IP1* (phosphoinositide-3-kinase-interacting protein 1), *ITGB2* (integrin subunit beta 2), and *TRPM3* (transient receptor potential cation channel subfamily M member 3) are involved in inflammation and immune mechanisms[26–30]. In particular, *BPIFC* belongs to a family of genes expressing antibacterial peptides that are released by neutrophils and bind lipopolysaccharides[31]. PIK3IP1 is a transmembrane protein that inhibits T cell responses to tumoral cells and intracellular bacteria[32,33]. The presence of these mbesQTLs corresponds to a lower Shannon index, lower expression of the anti-inflammatory *PIK3IP1* gene and higher expression of the pro-inflammatory *BPIFC* gene. Genes involved in immunity were also identified with the mbesQTLs associated with *Bacteroides*, *Akkermansia* and *Faecalibacterium* genera. *Bacteroides* comprises pathogenic species with pro-inflammatory and invasive properties; they are able to colonize epithelial cells and their abundance is increased in CRC[6,34,35]. *Bacteroides fragilis* has a known role in colon carcinogenesis[36,37].

We observed germline variants in 5 genes predisposing to a more efficient barrier and lower inflammation that would hinder *Bacteroides* expansion. *GAL* (galanin and GMAP prepropeptide) encodes for two proteins: galanin and galanin message-associated peptide (GMAP). Galanin regulates intestinal motility and induces chloride ion secretion in pathogen-derived inflammation-associated diarrhea[38], while GMAP expression increases in response to lipopolysaccharide (an inducer of inflammation)[39]. Gasdermin D *(GSDMD)* is the main catalyst of pyroptosis, a cellular programmed necrosis induced by the presence of intracellular lipopolysaccharide[40]. IL-1 receptor accessory protein (*IL1RAP*) is essential for IL-1 signaling[41]. It is associated with IL-1 receptor (IL-1R1), which participates in CRC-associated inflammation and tumor progression[42]. *LY6H* (lymphocyte antigen 6 family member H) is a pro-inflammatory regulator that inhibits the nicotinic acetylcholine receptor[43], which is a negative regulator of innate immunity antimicrobial peptides (AMPs)[44]. AMPs are potent pro-inflammatory molecules with a broad antibacterial spectrum, central to the innate epithelial immune response, and their expression can be regulated by the microbiota itself[45,46]. The product of *ATP1B1* (ATPase Na + /K + transporting subunit beta 1) upregulates the expression of tight-junction proteins and increases the epithelial barrier function in the lung[47]. The abundance of *Bacteroides* was lower in the presence of all its associated mbesQTLs. The expression of the pro-inflammatory genes *GAL, GSDMD* and *LY6H* was also reduced. *ATP1B1* and *IL1RAP* were instead more expressed.

Akkermansia muciniphila, the most abundant *Akkermansia* species in the human gut, is a mucin degrader considered beneficial for its anti-inflammatory and gut barrier function-enhancing properties[48–50]. However, it is enriched in CRC[51]. We observed an increase in *Akkermansia* abundance and *USP14* (ubiquitin-specific protease 14) expression in the presence of the mbesQTL. *USP14* is often overexpressed in cancer[52], and its product activates NF-κB and ERK1/2 in response to microbial infection[53]. Interestingly, NF-κB and ERK1/2 induce mucin expression[54], which may explain the observed increase in *Akkermansia* abundance.

Faecalibacterium prausnitzii, the *Faecalibacterium* species most represented in the human gut, has anti-inflammatory properties and is reduced in various intestinal disorders[55,56]. Kazrin (*KAZN*), which acts as a cytolinker by functionally connecting adherens junctions and desmosomes[57], was identified with the mbesQTL associated with *Faecalibacterium* and presents a splicing variation with unknown function.

A search of the 58 genes within The Cancer Genome Atlas revealed that 10 of the 28 genes identified with Shannon index-associated mbesQTLs and 19 of the 29 mbesQTLs associated with *Bacteroides, Akkermansia* and *Gemmiger* are dysregulated in CRC. Among the genes associated with mbesQTLs and known to be differently expressed in colorectal adenocarcinoma tissue according to The Cancer Genome Atlas database, *USP14, LOXL2* (lysyl oxidase-like 2) and *TP63* (tumor protein 63) link the microbiota composition with cancer development. *LOXL2* is involved in extracellular matrix (ECM) remodeling[58], and is predicted to have intron retention with unknown function in presence of the Shannon index-associated mbesQTL. Changes in ECM structure or composition can promote inflammation due to bacterial invasion that can lead to tumor development[59]. *TP63* expresses a tumor suppressor from the p53 family and an important regulator of cell differentiation, proliferation and survival[60]. It also enhances CXCR4 expression[61], which can be activated by various pathogens such as *Porphyromonas gingivalis* and *Chlamydia pneumoniae* to modulate the immune response and facilitate infection[62,63]. We found an intron retention with unknown function in *TP63* associated with *Bacteroides* mbQTLs.

We observed the presence of the *B. fragilis* in 36 patients. While other *Bacteroides* species are commensals normally present in healthy microbiota, *B. fragilis* is a pathogen. It is the most common anaerobe isolated from extraintestinal infections, and it has pro-inflammatory activities and contributes to colon carcinogenesis[36]. However, this OTU was not associated with clinical data in this study.

In conclusion, we identified mbesQTLs associated with genes involved in the regulation of the inflammatory environment and with the abundance of bacterial populations with immune-modulatory properties. We observed a correlation between pro-inflammatory genes and the abundance of the pro-inflammatory *Bacteroides*, while an inverse correlation was observed for barrier-reinforcing genes. mbesQTLs also showed a correlation between the mucin-degrading *Akkermansia* and expression of genes that regulate mucin expression. mbesQTL-associated genes may regulate the microenvironment in which bacteria grow and predispose to the development of cancer, some of these genes are also known to participate in carcinogenesis. Identifying variants that promote the growth of disease-associated bacterial populations may help find higher-risk individuals who could benefit

Primer ID	Sequence (5' – 3')	Refs.
V1-V2-V3 regions		
27F	**tcgtcggcagcgtcagatgtgtataagagacag**agagtttgatcntggctcag	[69]
519R	**gtctcgtgggctcggagatgtgtataagagacag**gwnttacngcggckgctg	[70]
V4-V5-V6 regions		
533F	**tcgtcggcagcgtcagatgtgtataagagacag**gtgccagcagccgcggtaa	[71]
1100R	**gtctcgtgggctcggagatgtgtataagagacag**agggttgcgctcgttg	[70]

Table 4. Universal primers used for 16S rRNA amplification. Nextera adaptor sequences in bold.

from preventive treatment such as antibiotics that inhibit cancer-associated bacteria and supplementation with commensal bacteria (probiotics) or metabolites generated by microbiota (postbiotics). Further studies with larger sample sizes and control populations are needed to clarify whether genetic variants associated with the regulation of both intestinal microbiota composition and gene expression have a role in CRC development. To include colonic mucosa from normal subjects, samples collected by less invasive procedures such as endoscopic colonic lavage or luminal brushing, both of which are good tissue surrogates for studying microbiota[64].

Methods

Patient series and biological material. The study included 95 patients with CRC who underwent surgery at the Colorectal Surgery Unit of Fondazione IRCCS Istituto Nazionale dei Tumori between 2009 and 2010 in the context of a larger study on CRC genetics[65]. The protocol for recruiting patients and collecting tissue specimens and clinical data had been approved by the Committee for Ethics of the Fondazione IRCCS Istituto Nazionale dei Tumori (authorization number: INT 08/08). All patients who donated samples provided written informed consent for the use of their materials for research purposes. Experimental procedures were performed in accordance with relevant institutional guidelines and regulations of the European Commission.

Patients with CRC at any pathological stage, determined following the American Joint Committee on Cancer (AJCC) TNM system[66], were eligible for inclusion in the study. To avoid potential consequences of inflammation on gut microbiota, patients were excluded from this study if signs of acute or chronic colorectal inflammation or stenosis were observed during surgery. Furthermore, patients were excluded if they had received antibiotics in the 6 months before surgery or underwent a colostomy or ileostomy before surgery. Clinical data were collected regarding age at surgery, sex, self-reported smoking habit (scored as ever or never smoker of tobacco-containing products), tumor site (colon or rectum), pathological stage and whether the patients received neo-adjuvant therapy before surgery (yes/no).

As per standard operating procedures, patients underwent bowel preparation with 2 sachets of sodium picosulfate (4 h apart in the evening before surgery) and received cefazolin (2 g intravenously) and metronidazole (500 mg intravenously) 40 min before incision. At the end of surgery, biopsies were collected from non-involved mucosa as far as possible from the tumor in the resected colorectal tissue. The samples were placed in RNAlater Stabilization Solution (Sigma-Aldrich) and then used for DNA extraction with the DNeasy Blood & Tissue Kit (Qiagen). Genomic DNA was quantified using the NanoDrop 2000c UV–Vis spectrophotometer (Thermo Fisher Scientific), diluted in water and stored at – 20 °C until use in microbiota profiling and SNP genotyping.

Microbiota profiling. Mucosal-associated bacteria are thought to play a critical role in interactions with the host immune system[67]. Therefore, genomic DNA samples from non-involved colorectal mucosa of the 95 patients were used for microbiota profiling by sequencing the bacterial 16S rRNA gene. We decided to sequence both V1-V3 and V4-V6 regions to better estimate the microbial taxonomic diversity, knowing that each set of regions would give better resolution on different taxa[68]. To amplify the first three hypervariable regions (V1, V2 and V3) of the 16S rRNA gene, we used AmpliTaq Gold DNA Polymerase (Thermo Fisher Scientific) and the 27F and 519R universal primers (Table 4)[69,70]. For the V4, V5 and V6 hypervariable regions, we ran PCR with the 533F and 1100R universal primers[70,71]. The primers sequences were modified to include Illumina Nextera index PCR adaptor sequences (Illumina).

PCR products were purified and sent to Eurofins Genomics (Edersberg, Germany) for index PCR, pooling and normalization of amplicons, and sequencing on an Illumina MiSeq sequencer with the 2 × 300 bp paired-end read module. Eurofins Genomics also did the read processing (according to primer sequences using proprietary scripts), taxonomy assignation and copy number correction. The company provided the data in files with fasta and biom formats for further analyses in our institutes.

Paired-end read merging was performed using FLASH software v. 2.2.00 (http://ccb.jhu.edu/software/FLASH/)[72]. Chimeric reads were identified and removed with UCHIME, implemented with VSEARCH software v. 2.7.1 (https://github.com/torognes/vsearch)[73,74]. Reads were sorted into operational taxonomic units (OTUs) through the minimum entropy decomposition method[75,76]. Taxonomic assignment was performed with BLAST in QIIME software v. 1.9.1 (http://qiime.org/)[77,78], using the NCBI_nt database as reference and with a minimum identity of 70% across at least 80% of the OTU representative sequence. OTU abundances were normalized considering their lineage-specific gene copy number with CopyRighter software v. 0.46 (https://github.com/fangly/AmpliCopyrighter)[79]. Microbial profiling was carried out separately for V1-V2-V3 and V4-V5-V6 amplicons.

QIIME software was used to analyze bacterial species diversity. Alpha diversity was assessed with the Shannon index, total number of observed OTUs, and the Chao1 estimator[80]. Beta diversity metric was assessed with the Bray–Curtis index of dissimilarity. Analyses of microbiota abundance were performed at the genus and species levels. To analyze OTUs at the genus level, we first removed OTUs missing a genus-level classification. To minimize the number of tests, we only considered taxa present in at least 25% of samples and with mean relative abundance greater than 1% among all samples. To remove compositional constraints typical of microbiota data, zero values were substituted with a pseudo-value less than the lesser relative abundance of the dataset, and relative abundances were transformed using centered log ratio (clr) transformation with the clr function of the Compositions R package[81].

SNP genotyping. Genomic DNA (100 ng per sample) was subjected to genome-wide genotyping using Axiom Precision Medicine Research Arrays (Thermo Fisher Scientific) on an Affymetrix Gene Titan platform at Eurofins Genomics Europe Genotyping (Galten, DK). Raw data provided by Eurofins were analyzed in our institutes using the Axiom Analysis Suite (Thermo Fisher Scientific). Genotype data were subjected to quality control using PLINK software v1.90b6.16 (https://www.cog-genomics.org/plink/1.9/)[82]. Per-sample quality check discarded samples with ≥ 2% missing genotypes. Genotyped variants with a call rate < 98% or a minor allele frequency (MAF) < 5% were excluded from analysis.

Statistical analyses. Alpha diversity metrics were compared between samples grouped by hypervariable regions V1-V2-V3 and V4-V5-V6 using a non-parametric *t*-test, adjusted for multiple testing by calculating the false discovery rate (FDR) using the Benjamini–Hochberg method. Differences in the Bray–Curtis index of dissimilarity were tested for significance with a permutation test with pseudo-F ratios (ADONIS function) and an analysis of similarities (ANOSIM function). These analyses were done in QIIME software v. 1.9.1, and the results were considered statistically significant when $P < 0.01$.

To determine if alpha diversity indexes differed between subgroups of patients defined by clinical characteristics (age, sex, smoking habit and tumor site), the non-parametric Kruskal–Wallis rank-sum test with continuity correction was used. To study associations between the clr-transformed relative abundance of OTUs and patient characteristics, we used multivariate linear regression. Pearson's correlation in relative abundance between OTUs was assessed using the rcorr function of the Hmisc package of R, and correlograms were drawn using R package corrplot. The significance threshold was set at $P < 0.05$, and correlation was defined as $|r| > 0.02$.

To identify mbQTLs, we used multivariable linear regression analysis to assess associations between germline polymorphisms and microbiota-related traits, including alpha and beta diversity metrics and clr-transformed relative abundance of OTUs. Sex, age, pathological stage, tumor site, and smoking habit were entered as covariates. The distance to define two loci on the same chromosome as independent loci was set to > 1 Mb. We adjusted for multiple testing using the Benjamini–Hochberg procedure to obtain the FDR[83], and a significance threshold was set at FDR < 0.10. We also consider the commonly accepted genome-wide significance threshold, set at the nominal $P < 5.0 \times 10^{-8}$[84]. These analyses were done using PLINK software. Manhattan plots were drawn using the manhattan function of the qqman package in R.

To investigate possible regulatory effects of the identified mbQTLs, we consulted the Genotype-Tissue Expression (GTEx) project database v7[85], and the Ensembl Variant Effect Predictor (VEP)[86], on 19 April 2021. In GTEx, we looked for matches between our mbQTLs and any expression QTL or splicing QTL in any tissue. Genes associated with these expression or splicing QTLs and genes predicted by VEP to retain an intron on their products in the presence of the minor allele of the identified mbQTLs were considered for analysis. Ontology annotation of these genes was obtained from DAVID Bioinformatic Database v6.8[24], and information on differential expression was obtained from The Cancer Genome Atlas[87].

Data availability

The genotyping data that support the findings of this study have been deposited in the European Genome-Phenome Archive with the accession code EGASXXXXXXX. The 16S rRNA gene sequencing data that support the findings of this study are available from https://icedrive.net/s/7RhCk77Fw5jAFw87AzNRixBGa8WA.

Received: 23 March 2022; Accepted: 21 June 2022
Published online: 06 July 2022

References

1. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
2. Arnold, M. *et al.* Global patterns and trends in colorectal cancer incidence and mortality. *Gut* **66**, 683–691 (2017).
3. Keum, N. & Giovannucci, E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nat. Rev. Gastroenterol. Hepatol.* **16**, 713–732 (2019).
4. Yang, T. *et al.* Gene–environment interactions and colorectal cancer risk: an umbrella review of systematic reviews and meta-analyses of observational studies. *Int. J. Cancer* **145**, 2315–2329 (2019).
5. Law, P. J. *et al.* Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nat. Commun.* **10**, 1–15 (2019).
6. Gagnière, J. *et al.* Gut microbiota imbalance and colorectal cancer. *World J. Gastroenterol.* **22**, 501 (2016).
7. Tremaroli, V. & Bäckhed, F. Functional interactions between the gut microbiota and host metabolism. *Nature* **489**, 242–249 (2012).
8. Kurilshikov, A. *et al.* Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nat. Genet.* **53**, 156–165 (2021).
9. Grieneisen, L. *et al.* Gut microbiome heritability is nearly universal but environmentally contingent. *Science* **373**, 181–186 (2021).
10. Davenport, E. R. Elucidating the role of the host genome in shaping microbiome composition. *Gut Microbes* **7**, 178–184 (2016).
11. Hu, S. *et al.* Whole exome sequencing analyses reveal gene–microbiota interactions in the context of IBD. *Gut* **70**, 285–296 (2021).

12. Schmitt, M. & Greten, F. R. The inflammatory pathogenesis of colorectal cancer. *Nat. Rev. Immunol.* **10**, 653–667 (2021).
13. Wong, S. H. & Yu, J. Gut microbiota in colorectal cancer: mechanisms of action and clinical applications. *Nat. Rev. Gastroenterol. Hepatol.* **16**, 690–704 (2019).
14. Richardson, T. G. *et al.* Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: a multivariable Mendelian randomisation analysis. *PLoS Med.* **17**, e1003062 (2020).
15. Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet.* **53**, 185–194 (2021).
16. Louis, P. & Flint, H. J. Formation of propionate and butyrate by the human colonic microbiota. *Environ. Microbiol.* **19**, 29–41 (2017).
17. Bang, S.-J. *et al.* The influence of in vitro pectin fermentation on the human fecal microbiome. *AMB Express* **8**, 1–9 (2018).
18. Rueter, C. & Bielaszewska, M. Secretion and delivery of intestinal pathogenic Escherichia coli virulence factors via outer membrane vesicles. *Front. Cell Infect. Microbiol.* **10**, 91 (2020).
19. Kittana, H. *et al.* Commensal Escherichia coli strains can promote intestinal inflammation via differential interleukin-6 production. *Front. Immunol.* **9**, 2318 (2018).
20. Chen, J. *et al.* An expansion of rare lineage intestinal microbes characterizes rheumatoid arthritis. *Genome Med.* **8**, 1–14 (2016).
21. Azam, M. W. & Khan, A. U. Updates on the pathogenicity status of *Pseudomonas aeruginosa*. *Drug Discov. Today* **24**, 350–359 (2019).
22. Forbes, J. D. *et al.* A comparative study of the gut microbiota in immune-mediated inflammatory diseases—does a common dysbiosis exist?. *Microbiome* **6**, 1–15 (2018).
23. Vuik, F. *et al.* Composition of the mucosa-associated microbiota along the entire gastrointestinal tract of human individuals. *United Eur. Gastroenterol. J.* **7**, 897–907 (2019).
24. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
25. Ott, S. *et al.* Reduction in diversity of the colonic mucosa associated bacterial microflora in patients with active inflammatory bowel disease. *Gut* **53**, 685–693 (2004).
26. Barber, D. F., Faure, M. & Long, E. O. LFA-1 contributes an early signal for NK cell cytotoxicity. *J. Immunol.* **173**, 3653–3659 (2004).
27. Ley, K., Laudanna, C., Cybulsky, M. I. & Nourshargh, S. Getting to the site of inflammation: the leukocyte adhesion cascade updated. *Nat. Rev. Immunol.* **7**, 678–689 (2007).
28. Bai, M. *et al.* CD177 modulates human neutrophil migration through activation-mediated integrin and chemoreceptor regulation. *Blood* **130**, 2092–2100 (2017).
29. Mulier, M. *et al.* Upregulation of TRPM3 in nociceptors innervating inflamed tissue. *Elife* **9**, e61103 (2020).
30. Boonen, B. *et al.* Differential effects of lipopolysaccharide on mouse sensory TRP channels. *Cell Calcium* **73**, 72–81 (2018).
31. Wiesner, J. & Vilcinskas, A. Antimicrobial peptides: the ancient arm of the human immune system. *Virulence* **1**, 440–464 (2010).
32. Chen, Y. *et al.* Pik3ip1 is a negative immune regulator that inhibits antitumor T-cell immunity. *Clin. Cancer Res.* **25**, 6180–6194 (2019).
33. Uche, U. U. *et al.* PIK3IP1/TrIP restricts activation of T cells through inhibition of PI3K/Akt. *J. Exp. Med.* **215**, 3165–3179 (2018).
34. Wexler, H. M. Bacteroides: the good, the bad, and the nitty-gritty. *Clin. Microbiol. Rev.* **20**, 593–621 (2007).
35. Nakano, V. *et al.* Adherence and invasion of *Bacteroidales* isolated from the human intestinal tract. *Clin. Microbiol. Infect.* **14**, 955–963 (2008).
36. Cheng, W. T., Kantilal, H. K. & Davamani, F. The mechanism of *Bacteroides fragilis* toxin contributes to colon cancer formation. *Malays. J. Med. Sci.* **27**, 9 (2020).
37. Liu, Q.-Q. *et al.* Enterotoxigenic *Bacteroides fragilis* induces the stemness in colorectal cancer via upregulating histone demethylase JMJD2B. *Gut Microbes* **12**, 1788900 (2020).
38. Benya, R. V., Matkowskyj, K. A., Danilkovich, A. & Hecht, G. Galanin causes Cl– secretion in the human colon: potential significance of inflammation-associated NF-κB activation on galanin-1 receptor expression and function. *Ann. N. Y. Acad. Sci.* **863**, 64–77 (1998).
39. Rauch, I., Lundström, L., Hell, M., Sperl, W. & Kofler, B. Galanin message-associated peptide suppresses growth and the budded-to-hyphal-form transition of *Candida albicans*. *Antimicrob. Agents Chemother.* **51**, 4167–4170 (2007).
40. Shi, J., Gao, W. & Shao, F. Pyroptosis: gasdermin-mediated programmed necrotic cell death. *Trends Biochem. Sci.* **42**, 245–254 (2017).
41. Greenfeder, S. A. *et al.* Molecular cloning and characterization of a second subunit of the interleukin 1 receptor complex*. *J. Biol. Chem.* **270**, 13757–13765 (1995).
42. Dmitrieva-Posocco, O. *et al.* Cell-type-specific responses to interleukin-1 control microbial invasion and tumor-elicited inflammation in colorectal cancer. *Immunity* **50**(166–180), e167 (2019).
43. Moriwaki, Y. *et al.* Endogenous neurotoxin-like protein Ly6H inhibits alpha7 nicotinic acetylcholine receptor currents at the plasma membrane. *Sci. Rep.* **10**, 1–11 (2020).
44. Kishibe, M., Griffin, T. M. & Radek, K. A. Keratinocyte nicotinic acetylcholine receptor activation modulates early TLR2-mediated wound healing responses. *Int. Immunopharmacol.* **29**, 63–70 (2015).
45. Kobatake, E. & Kabuki, T. S-layer protein of *Lactobacillus helveticus* SBT2171 promotes human β-defensin 2 expression via TLR2–JNK signaling. *Front. Microbiol.* **10**, 2414 (2019).
46. Álvarez, Á. H., Velázquez, M. M. & de Oca, E. P. M. Human β-defensin 1 update: Potential clinical applications of the restless warrior. *Int. J. Biochem. Cell Biol.* **104**, 133–137 (2018).
47. Bai, H. *et al.* The Na+, K+-ATPase β1 subunit regulates epithelial tight junctions via MRCKα. *JCI Insight.* **6**, e134881 (2021).
48. Zhai, R. *et al.* Strain-specific anti-inflammatory properties of two *Akkermansia muciniphila* strains on chronic colitis in mice. *Front. Cell Infect. Microbiol.* **9**, 239 (2019).
49. Ottman, N. *et al.* Pili-like proteins of *Akkermansia muciniphila* modulate host immune responses and gut barrier function. *PLoS ONE* **12**, e0173004 (2017).
50. Chelakkot, C. *et al.* Akkermansia muciniphila-derived extracellular vesicles influence gut permeability through the regulation of tight junctions. *Exp. Mol. Med.* **50**, e450–e450 (2018).
51. Borges-Canha, M., Portela-Cidade, J. P., Dinis-Ribeiro, M., Leite-Moreira, A. F. & Pimentel-Nunes, P. Role of colonic microbiota in colorectal carcinogenesis: a systematic review. *Rev. Esp. Enferm. Dig.* **107**, 659–671 (2015).
52. Zhang, B., Li, M., Huang, P., Guan, X. Y. & Zhu, Y. H. Overexpression of ubiquitin specific peptidase 14 predicts unfavorable prognosis in esophageal squamous cell carcinoma. *Thorac. Cancer* **8**, 344–349 (2017).
53. Liu, N. *et al.* Ubiquitin-specific protease 14 regulates LPS-induced inflammation by increasing ERK1/2 phosphorylation and NF-κB activation. *Mol. Cell Biochem.* **431**, 87–96 (2017).
54. Liu, Y. *et al.* The role of MUC2 mucin in intestinal homeostasis and the impact of dietary components on MUC2 expression. *Int. J. Biol. Macromol.* https://doi.org/10.1016/j.ijbiomac.2020.07.191 (2020).
55. Sokol, H. *et al. Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 16731–16736 (2008).
56. Lopez-Siles, M., Duncan, S. H., Garcia-Gil, L. J. & Martinez-Medina, M. *Faecalibacterium prausnitzii*: from microbiology to diagnostics and prognostics. *ISME J.* **11**, 841–852 (2017).

57. Groot, K. R., Sevilla, L. M., Nishi, K., DiColandrea, T. & Watt, F. M. Kazrin, a novel periplakin-interacting protein associated with desmosomes and the keratinocyte plasma membrane. *J. Cell Biol.* **166**, 653–659 (2004).
58. Rodriguez, H. M. *et al.* Modulation of lysyl oxidase-like 2 enzymatic activity by an allosteric antibody inhibitor. *J. Biol. Chem.* **285**, 20964–20974 (2010).
59. Mohan, V., Das, A. & Sagi, I. Emerging roles of ECM remodeling processes in cancer. *Semin. Cancer Biol.* **62**, 192–200 (2020).
60. Gonfloni, S., Caputo, V. & Iannizzotto, V. P63 in health and cancer. *Int. J. Dev. Biol.* **59**, 87–93 (2015).
61. DeCastro, A. J., Cherukuri, P., Balboni, A. & DiRenzo, J. ΔNP63α transcriptionally activates chemokine receptor 4 (CXCR4) expression to regulate breast cancer stem cell activity and chemotaxis. *Mol. Cancer Ther.* **14**, 225–235 (2015).
62. Hajishengallis, G., Wang, M., Liang, S., Triantafilou, M. & Triantafilou, K. Pathogen induction of CXCR4/TLR2 cross-talk impairs host defense function. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 13532–13537 (2008).
63. Miao, G. *et al.* TLR2/CXCR4 coassociation facilitates *Chlamydia pneumoniae* infection-induced atherosclerosis. *Am. J. Physiol. Heart Circ. Physiol.* **318**, H1420–H1435 (2020).
64. Tang, Q. *et al.* Current sampling methods for gut microbiota: a call for more precise devices. *Front. Cell Infect. Microbiol.* **10**, 151 (2020).
65. Noci, S. *et al.* A subset of genetic susceptibility variants for colorectal cancer also has prognostic value. *Pharmacogenom. J.* **16**, 173–179 (2016).
66. Amin, M. B. *et al.* The eighth edition AJCC cancer staging manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA Cancer J. Clin.* **67**, 93–99 (2017).
67. Miyauchi, E. *et al.* Analysis of colonic mucosa-associated microbiota using endoscopically collected lavage. *Sci. Rep.* **12**, 1–8 (2022).
68. Pinna, N. K., Dutta, A., Monzoorul Haque, M. & Mande, S. S. Can targeting non-contiguous V-regions with paired-end sequencing improve 16S rRNA-based taxonomic resolution of microbiomes?: an in silico evaluation. *Front. Genet.* **10**, 653 (2019).
69. Lane, D. J. 16S/23S rRNA sequencing. In *Nucleic Acid Techniques in Bacterial Systematics* (eds Stackebrandt, E. & Goodfellow, M.) 115–175 (John Wiley & Son, 1991).
70. Turner, S., Pryer, K. M., Miao, V. P. & Palmer, J. D. Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. *J. Eukaryot. Microbiol.* **46**, 327–338 (1999).
71. Weisburg, W. G., Barns, S. M., Pelletier, D. A. & Lane, D. J. 16S ribosomal DNA amplification for phylogenetic study. *J. Bacteriol.* **173**, 697–703 (1991).
72. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
73. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194–2200 (2011).
74. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
75. Eren, A. M. *et al.* Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* **4**, 1111–1119 (2013).
76. Eren, A. M. *et al.* Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J.* **9**, 968–979 (2015).
77. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
78. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods.* **7**, 335–336 (2010).
79. Angly, F. E. *et al.* CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome* **2**, 1–13 (2014).
80. Colwell, R. Biodiversity: concepts, patterns, and measurement. In *The Princeton Guide to Ecology* (eds Levin, S. *et al.*) 257–263 (Princeton University Press, 2009).
81. Aitchison, J. The statistical analysis of compositional data. *J. R. Statist. Soc. B* **44**, 139–160 (1982).
82. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
83. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57**, 289–300 (1995).
84. Xu, C. *et al.* Estimating genome-wide significance for whole-genome sequencing studies. *Genet. Epidemiol.* **38**, 281–290 (2014).
85. Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
86. McLaren, W. *et al.* The ensembl variant effect predictor. *Genome Biol.* **17**, 1–14 (2016).
87. Li, M., Sun, Q. & Wang, X. Transcriptional landscape of human cancers. *Oncotarget* **8**, 34534 (2017).

Acknowledgements

The authors acknowledge the contribution of Valerie Matarese, PhD, who provided scientific editing, and the COST action TRANSCOLONCAN (CA17118). Alberto Vannelli—formerly of the Colorectal Surgery Unit, Department of Surgery, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy.

Author contributions

T.A.D., F.C., O.I. and M.G. conceived the study. A.V., L.B., L.S., and M.C. provided biological samples from colorectal cancer patients. S.N., G.P., and A.P. prepared DNA samples and amplified hypervariable regions of 16S rRNA gene. O.I. analyzed microbiota data. F.C. and F.M. were involved in genotyping data management and analysis. F.C., O.I., M.G. and T.A.D. were involved in experimental design and manuscript preparation. All authors participated in critical revision of the article and approved the final manuscript.

Funding

This work was supported in part by The European H2020 research project Oncobiome (Grant Number 825410) and by institutional funds obtained through an Italian law that allows taxpayers to allocate 0.5% of their income tax contribution to a research institution of their choice. OI was recipient of a postdoctoral fellowship from the National Council of Science and Technology of Mexico (CONACYT) during the study and currently holds a Fondazione Umberto Veronesi scholarship grant 2021. The funding organizations had no role in design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-15230-6.

Correspondence and requests for materials should be addressed to T.A.D.

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

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