Supplementary Tables

Supplementary Table S1. Baseline characteristics of colorectal cancer (CRC) cases and controls in cohort C1. For quantitative traits, the median, minimum and maximum are shown. FBG: fasting blood glucose; ALT/GPT: alanine transaminase/glutamate pyruvated transaminase; BMI: body mass index; DM: diabetes mellitus type 2; HDL: high density lipoprotein; TG: triglyceride; eGFR: epidermal growth factor receptor; TCHO: total cholesterol; Cr: creatinine; LDL; low density lipoprotein; TNM: tumor node metastasis staging system; Statistical tests used for identifying associations between metadata and CRC: † - Wilcoxon test, ‡ - Fisher's exact test. * - Information missing in one CRC patient sample.

Parameter	**Controls (n=54)**	**Cases (n=74)**	***P-value***	**q-value**
Age	63(50,73)	67(34,89)	0.007373932†	0.027652245
Gender (M:F)	33:21	48:26	0.7124‡	0.875363308
BMI	22.86341 (17.08744,35.08618)	23.89549 (17.28882,31.25)	0.1107815†	0.237388929
DM (%)	16 (29.6%)	29 (39.2%)	0.3488‡	0.503844136
Stage of CRC (1:2:3:4)	n.a	18:22:26:8	n.a	n.a
Distribution of detailed TNM stages (T1N0:T2N0:T3N0:T4N0:T2N1:T3N1:T3N2:T3N+:T4N2:T3N1M1:T3N3M1:T4N1M1:T4N2M1:M:M1:multiple liver met)	n.a	12:6:21:1:3:14:5:2:2:2:1:1:1:1:1	n.a	n.a

Leison location(1:2:NA)	n.a	11:54:9	n.a	n.a
Leison specific location（1:2:3:4:6:7:8:9）	n.a	3:3:3:2:6:14:5:29	n.a	n.a
Fecal sampling before or after colonoscopy* (before:after)	30:24 (56%:44%)	21:52 (29%:71%)	0.003295‡	0.016475000
Duration between colonoscopy and fecal sample collection* (days)	-1.5 (-34,106.3438)	19.89097 (-110.7083,247)	0.7586482†	0.875363308
Duration of frozen storage of fecal samples (days)	185.1076 (86.67708,2032)	149 (6.6875,1280)	0.3694857†	0.503844136
FBG	5.1 (4.3,6.9)	5.75 (4.3,13.2)	0.000131342†	0.001407555
TCHO	5 (3.2,6.7)	4.9 (2.6,8.6)	0.299775†	0.503844136
LDL	2.65 (1.4,5.2)	2.9 (0.7,5)	0.9413451†	0.989788600
HDL	1.8 (0.8,3.5)	1.3 (0.5,2.2)	0.000187674†	0.001407555
TG	1 (0.37,2.9)	1.2 (0.5,5.1)	0.01991682†	0.059750460
Cr	71.5 (43,101)	74 (41,202)	0.3257186†	0.503844136
ALT GPT	21 (9,68)	18 (10,69)	0.05068182†	0.126704550
eGFR	69.51 (50.82,115.04)	71.13 (16.81,136.52)	0.9897886†	0.989788600

Supplementary Table S2. Summary of metagenomic data from C1 and mapping to reference gene catalogue. Fourth column reports results from Wilcoxon rank-sum tests.

Parameter	**Controls**	**Cases**	***P-value***
Average raw reads	60162577	60496561	0.8082
After removing low quality reads	59423292 (98.77%)	59715967 (98.71%)	0.831
After removing human reads	59380535 ± 7378751	58112890 ± 10324458	0.419
Mapping rate	66.82%	66.27%	0.252

Supplementary Table S3. Gene number and gene alpha diversity of CRC and healthy microbiomes in cohort C1. Diversity was represented by Shannon and Simpson indices.

Parameter	**Controls**		**Cases**		**P-value**
	<u>mean</u>	<u>sd</u>	<u>mean</u>	<u>sd</u>	
Gene number	581635	165527	534440	164184	0.1127
Shannon index	11.699	0.565	11.558	0.546	0.0746
Simpson index	0.99997	2.71E-05	0.99996	3.94E-05	0.0276

Supplementary Table S4. PERMANOVA analysis of microbial gene profiles in cohort C1. The analysis was conducted to test whether clinical parameters and CRC status have significant impact on the gut microbiota with q<0.05. BMI: body mass index; DM: diabetes mellitus type 2; FBG: fasting blood glucose; HDL: high density lipoprotein; TG: triglyceride; eGFR: epidermal growth factor receptor; TNM: tumor node metastasis staging system; TCHO: total cholesterol; Cr: creatinine; LDL; low density lipoprotein. ALT/GPT: alanine transaminase/glutamate pyruvated transaminase.

Parameter	**Df**	**SumsOfSqs**	**MeanSqs**	**F.Model**	**R^2**	**Pr(>F)**	**q-value**
CRC Status	1	0.6792933	0.6792933	1.9596297	0.0153144	0.0004	0.0076
Stage of CRC	4	1.7697175	0.4424294	1.2778364	0.0398977	0.0058	0.0551
Lesion location(1:2:NA)	1	0.464298	0.464298	1.31427	0.020435	0.0536	0.2717
BMI	1	0.4600024	0.4600024	1.3200099	0.0104497	0.0572	0.2717
DM	1	0.4383585	0.4383585	1.257642	0.0098826	0.084	0.285
FBG	1	0.4319269	0.4319269	1.2300105	0.0123955	0.09	0.285
Lesion specific location（1:2:3:4:6:7:8:9:NA）	1	0.421307	0.421307	1.190278	0.018543	0.1369	0.371586
Age	1	0.3972817	0.3972817	1.1387282	0.0089566	0.1923	0.456713
HDL	1	0.3641778	0.3641778	1.0352042	0.010246	0.3578	0.722
eGFR	1	0.3585266	0.3585266	1.0231375	0.0094715	0.38	0.722
TG	1	0.3522642	0.3522642	1.001382	0.0099145	0.4329	0.747736

Duration between colonoscopy and fecal sample collection	1	0.3397823	0.3397823	0.9722612	0.0077181	0.5036	0.761608
Fecal sampling before or after colonoscopy	1	0.3378151	0.3378151	0.9665887	0.0076734	0.5211	0.761608
TNM	15	5.3000663	0.3533378	0.9890377	0.2036857	0.5781	0.766587
Cr	1	0.3281613	0.3281613	0.9330291	0.0088077	0.6052	0.766587
TCHO	1	0.3127842	0.3127842	0.8878167	0.0088	0.7198	0.854763
LDL	1	0.2994855	0.2994855	0.8502487	0.0084308	0.8146	0.863233
ALT/GPT	1	0.2976508	0.2976508	0.847193	0.007929	0.8178	0.863233
Gender	1	0.2677377	0.2677377	0.7651615	0.0060361	0.9528	0.9528

Supplementary Table S5. Principal component analysis (PCA) using microbial gene profiles from cohort C1 using 2.1 million genes. Association tests of the first five principal components (PC) with 17 parameters are reported.

	Gene level *P-value*					
Parameter	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**Statistical test for differences**
CRC Status	**0.029375**	0.7556469	0.0908164	0.2458964	**1.29E-06**	Wilcoxon rank-sum tests

Age	0.1107786	0.7187803	0.9579642	0.3323753	0.1740341	Pearson correlation test
BMI	0.1666538	0.349701	0.2799689	0.9666352	0.6664927	Pearson correlation test
Duration between colonoscopy and fecal sample collection	0.2612967	0.3677261	0.5027833	0.8471867	0.985353	Pearson correlation test
Fecal sampling before or after colonoscopy	0.3051672	0.3564576	0.6633822	0.998038	0.2695479	Wilcoxon rank-sum tests
DM	0.3304729	0.7684188	0.192732	0.4910126	**0.025695**	Wilcoxon rank-sum tests
TNM	0.3587305	0.7179382	0.4123964	0.6422653	0.2646984	Kruskal-Wallis tests
Gender	0.3762511	0.692509	0.0652127	0.6280812	0.4261203	Wilcoxon rank-sum tests
TCHO	0.3918745	0.6337139	0.8437887	0.8920492	0.8586685	Pearson correlation test
LDL	0.3996362	0.3026439	0.2289333	0.8912377	0.5012763	Pearson correlation test
eGFR	0.4185351	0.6904019	0.4945847	0.3171986	0.5644339	Pearson correlation test
Stage of CRC	0.4785966	0.4859963	0.3868685	0.6319499	0.1400903	Kruskal-Wallis tests
HDL	0.4855939	0.265294	0.9413181	0.8985499	0.1237575	Pearson correlation test
TG	0.5435062	0.7623276	0.4072886	0.1106054	0.0247417	Pearson correlation test
ALT/GPT	0.6686028	0.819014	0.5737057	0.3283116	0.6117176	Pearson correlation test
Cr	0.8059999	0.5986523	0.743814	0.7723353	0.3177772	Pearson correlation test
FBG	0.8765164	0.6637887	**0.037985**	0.8754288	0.0596181	Pearson correlation test

Supplementary Table S6. List of KEGG modules and pathways associated with CRC status at P-value<0.01 in cohort C1.

KEGG Modules						
Module ID	**Control rank mean**	**Case rank mean**	**Enrichment (0:case/1:control)**	**P-value**	**q-value**	**Definition**
M00036	48.72222222	76.01351351	0	4.36E-05	0.014810703	Leucine degradation, leucine => acetoacetate + acetyl-CoA
M00050	50.2962963	74.86486486	0	0.000141552	0.02406378	Guanine nucleotide biosynthesis, IMP => GDP/dGDP,GTP/dGTP
M00037	51.72222222	73.82432432	0	0.000655997	0.056363614	Melatonin biosynthesis, tryptophan => serotonin => melatonin
M00042	52.05555556	73.58108108	0	0.000663101	0.056363614	Catecholamine biosynthesis, tyrosine => dopamine => noradrenaline => adrenaline
M00020	53.07407407	72.83783784	0	0.002447934	0.142598694	Serine biosynthesis, glycerate-3P => serine
M00046	54.81481481	71.56756757	0	0.003088118	0.142598694	beta-Alanine biosynthesis, cytosine / uracil => beta-alanine
M00055	77.38888889	55.09459459	1	0.003404079	0.142598694	N-glycan precursor biosynthesis
M00250	76.40740741	55.81081081	1	0.003804616	0.142598694	Lipopolysaccharide transport system
M00135	53.09259259	72.82432432	0	0.004435999	0.142598694	GABA biosynthesis, eukaryotes, putrescine => GABA
M00144	52.27777778	73.41891892	0	0.004478832	0.142598694	Complex I (NADH dehydrogenase), NADH dehydrogenase I
M00267	76.2962963	55.89189189	1	0.004613487	0.142598694	PTS system, N-acetylglucosamine-specific II component

M00117	53.53703704	72.5	0	0.00506723	0.143571525	Ubiquinone biosynthesis, prokaryotes, chorismate => ubiquinone
M00319	55.03703704	71.40540541	0	0.005986981	0.156582578	Manganese/zinc/iron transport system
M00045	53.81481481	72.2972973	0	0.006759204	0.164152094	Histidine degradation, histidine => N-formiminoglutamate => glutamate
M00318	56.85185185	70.08108108	0	0.007585097	0.171928858	Iron/zinc/copper transport system
M00209	53.72222222	72.36486486	0	0.009495418	0.20177763	Osmoprotectant transport system
KEGG Pathways						
Map ID	**Control rank mean**	**Case rank mean**	**Enrichment(0:case/ 1:control)**	**P-value**	**q-value**	**Definition**
map00901	51.72222222	73.82432432	0	0.000655997	0.093807599	Indole alkaloid biosynthesis
map00965	51.72222222	73.82432432	0	0.000655997	0.093807599	Betalain biosynthesis
map00943	76.35185185	55.85135135	1	0.001077093	0.102682879	Isoflavonoid biosynthesis
map00253	76.75925926	55.55405405	1	0.002345148	0.167678104	Tetracycline biosynthesis
map00190	52.53703704	73.22972973	0	0.003379124	0.177259289	Oxidative phosphorylation
map00430	52.7962963	73.04054054	0	0.003718726	0.177259289	Taurine and hypotaurine metabolism
map00280	54.22222222	72	0	0.006119571	0.222185359	Valine, leucine and isoleucine degradation

map04724	54.68518519	71.66216216	0	0.006639056	0.222185359	Glutamatergic synapse
map00562	54.07407407	72.10810811	0	0.008048415	0.222185359	Inositol phosphate metabolism
map00061	74.27777778	57.36486486	1	0.008239217	0.222185359	Fatty acid biosynthesis
map00910	55.16666667	71.31081081	0	0.008545591	0.222185359	Nitrogen metabolism
map04940	54.53703704	71.77027027	0	0.009490601	0.226192666	Type I diabetes mellitus

Supplementary Table S7. List of KEGG orthologous groups (KOs) associated with CRC status at q-value <0.05 in cohort C1.

KO ID	Control rank mean	Case rank mean	Enrichment (0:case/1:control)	P-value	q-value	Definition
K09778	46.68519	77.5	0	2.91E-06	0.017649179	Hypothetical protein
K10670	46.44444	77.67568	0	6.94E-06	0.020377011	Glycine reductase
K09065	49.01852	75.7973	0	1.15E-05	0.020377011	N-acetylornithine carbamoyltransferase
K13772	47.46296	76.93243	0	1.34E-05	0.020377011	Rrf2 family transcriptional regulator, iron-responsive regulator
K01464	49.27778	75.60811	0	2.35E-05	0.022105986	Dihydropyrimidinase
K02656	81.26852	52.26351	1	2.51E-05	0.022105986	Type IV pilus assembly protein PilF
K08286	81.05556	52.41892	1	2.55E-05	0.022105986	Protein-serine/threonine kinase

K01096	80.68519	52.68919	1	4.04E-05	0.029007909	Phosphatidylglycerophosphatase B
K00087	49.61111	75.36486	0	4.56E-05	0.029007909	Xanthine dehydrogenase molybdenum-binding subunit
K05020	48.07407	76.48649	0	5.26E-05	0.029007909	Glycine betaine transporter
K07301	81.09259	52.39189	1	5.43E-05	0.029007909	Inner membrane protein
K01318	79.22222	53.75676	1	5.73E-05	0.029007909	Glutamyl endopeptidase
K11786	80.85185	52.56757	1	6.74E-05	0.031479285	ATP-dependent helicase STH1/SNF2
K01951	49.25926	75.62162	0	7.32E-05	0.031758681	GMP synthase (glutamine-hydrolysing)
K01459	78.51852	54.27027	1	1.31E-04	0.049313518	NA
K00132	50.64815	74.60811	0	1.38E-04	0.049313518	Acetaldehyde dehydrogenase (acetylating)
K04835	51.12963	74.25676	0	1.44E-04	0.049313518	Methylaspartate ammonia-lyase
K11337	49.16667	75.68919	0	1.51E-04	0.049313518	3-hydroxyethyl bacteriochlorophyllide a dehydrogenase
K04749	49.11111	75.72973	0	1.54E-04	0.049313518	Anti-sigma B factor antagonist

Supplementary Table S8. IMG, mOTU and MLG species associated with CRC with q-value < 0.05 in cohort C1. 86 MLG species were formed after grouping 106 MLGs with more than 100 genes using species annotation when available. MLG species identifiers starting with “Con_” are enriched in control samples, and those starting with “CRC_” are enriched in CRC samples.

28 IMG species					
	Control rank mean	**Case rank mean**	**Enrichment (0:case/1:control)**	**P-value**	**q-value**
Peptostreptococcus stomatis	37.25926	84.37838	0	5.11E-12	1.32E-08
Parvimonas micra	38.43519	83.52027	0	4.21E-11	5.43E-08
Parvimonas sp. oral taxon 393	39.81481	82.51351	0	2.79E-10	2.40E-07
Parvimonas sp. oral taxon 110	43.52778	79.80405	0	6.17E-08	3.98E-05
Gemella morbillorum	43.87037	79.55405	0	1.53E-07	7.88E-05
Fusobacterium nucleatum	45.09259	78.66216	0	3.86E-07	1.56E-04
Leptotrichia buccalis	45.60185	78.29054	0	4.44E-07	1.56E-04
Fusobacterium sp. oral taxon 370	45.02778	78.70946	0	4.83E-07	1.56E-04
Burkholderia mallei	45.19444	78.58784	0	7.93E-07	2.27E-04
Prevotella intermedia	46.47222	77.65541	0	1.92E-06	4.95E-04

Streptococcus pseudoporcinus	47.5	76.90541	0	4.03E-06	8.99E-04
Streptococcus dysgalactiae	47.06481	77.22297	0	4.18E-06	8.99E-04
Beggiatoa sp. PS	46.53704	77.60811	0	5.03E-06	9.97E-04
Malassezia globosa	46.35185	77.74324	0	8.71E-06	1.60E-03
Paracoccus denitrificans	47.48148	76.91892	0	1.18E-05	2.02E-03
Eubacterium ventriosum	80.98148	52.47297	1	1.27E-05	2.05E-03
Streptococcus constellatus	48.2037	76.39189	0	1.66E-05	2.52E-03
Filifactor alocis	49.06481	75.76351	0	3.94E-05	5.65E-03
Peptoniphilus indolicus	51.2963	74.13514	0	4.53E-05	6.14E-03
Crenothrix polyspora	48.76852	75.97973	0	5.14E-05	6.63E-03
Peptostreptococcus anaerobius	50.14815	74.97297	0	5.88E-05	7.22E-03
Streptococcus equi	50.58333	74.65541	0	6.91E-05	8.10E-03
Solobacterium moorei	47.66667	76.78378	0	8.79E-05	9.85E-03
Sulfurovum sp. SCGC AAA036-O23	52.12037	73.53378	0	1.28E-04	1.37E-02
Streptobacillus moniliformis	52.35185	73.36486	0	1.44E-04	1.49E-02

Eubacteriaceae bacterium ACC19a	51.87037	73.71622	0	1.93E-04	1.92E-02
Fusobacterium necrophorum	52.37037	73.35135	0	3.72E-04	3.55E-02
Adhaeribacter aquaticus	77.06481	55.33108	1	4.79E-04	4.41E-02
21 mOTU species					
	Control rank mean	Case rank mean	Enrichment(0:case /1:control)	P-value	q-value
Parvimonas micra	46.2963	77.78378	0	2.31E-08	7.73E-06
Peptostreptococcus stomatis	46.25	77.81757	0	2.81E-08	7.73E-06
motu_linkage_group_731	50.42593	74.77027	0	2.91E-07	5.33E-05
Gemella morbillorum	47.93519	76.58784	0	8.63E-07	1.18E-04
motu_linkage_group_407	81.13889	52.35811	1	8.51E-06	9.34E-04
motu_linkage_group_490	80.46296	52.85135	1	3.04E-05	2.78E-03
Fusobacterium nucleatum	54.62037	71.70946	0	3.56E-05	2.79E-03
Clostridium symbiosum	48.66667	76.05405	0	4.50E-05	2.99E-03
motu_linkage_group_443	79.66667	53.43243	1	4.91E-05	2.99E-03
motu_linkage_group_316	79.61111	53.47297	1	7.03E-05	3.86E-03

Eubacterium ventriosum	78.09259	54.58108	1	9.82E-05	4.90E-03
Solobacterium moorei	51.22222	74.18919	0	2.49E-04	1.14E-02
Bacteroides fragilis	51.09259	74.28378	0	3.75E-04	1.58E-02
unclassified Fusobacterium	54.22222	72	0	4.20E-04	1.59E-02
Clostridiales bacterium 1_7_47FAA	51.27778	74.14865	0	4.34E-04	1.59E-02
Clostridium ramosum	50.92593	74.40541	0	5.21E-04	1.75E-02
motu_linkage_group_611	77.2963	55.16216	1	5.50E-04	1.75E-02
Prevotella nigrescens	58.09259	69.17568	0	5.72E-04	1.75E-02
motu_linkage_group_624	51.01852	74.33784	0	1.33E-03	3.69E-02
motu_linkage_group_510	77.84259	54.76351	1	1.35E-03	3.69E-02
Clostridium bolteae	51.81481	73.75676	0	1.41E-03	3.69E-02
85 MLG species					
	Control rank mean	Case rank mean	Enrichment(0:case /1:control)	P-value	q-value
Parvimonas micra	38.40741	83.54054	0	5.56E-12	4.84E-10
Fusobacterium nucleatum	40.32407	82.14189	0	1.72E-10	7.48E-09

Solobacterium moorei	42.2037	80.77027	0	4.01E-08	1.16E-06
Clostridium symbiosum	46.31481	77.77027	0	2.67E-06	5.80E-05
Con 10180	82.03704	51.7027	1	6.06E-06	1.05E-04
CRC 2881	51.25926	74.16216	0	7.57E-06	1.10E-04
CRC 2794	51.03704	74.32432	0	1.04E-05	1.30E-04
Coprococcus sp. ART55/1	80.85185	52.56757	1	2.09E-05	2.05E-04
Clostridium hathewayi	46.77778	77.43243	0	2.12E-05	2.05E-04
Clostridiales bacterium 1_7_47FAA	48.16667	76.41892	0	2.49E-05	2.17E-04
CRC 4136	50.99074	74.35811	0	2.97E-05	2.32E-04
butyrate-producing bacterium SS3/4	80.57407	52.77027	1	3.19E-05	2.32E-04
Haemophilus parainfluenzae	80.49074	52.83108	1	4.18E-05	2.69E-04
Con 154	80.35185	52.93243	1	4.45E-05	2.69E-04
Clostridium clostridioforme	50.2037	74.93243	0	4.64E-05	2.69E-04
Bacteroides fragilis	49.09259	75.74324	0	5.56E-05	3.02E-04
Con 1979	79.94444	53.22973	1	6.03E-05	3.09E-04

Eubacterium ventriosum	78.62963	54.18919	1	6.88E-05	3.33E-04
Con 7958	75.27778	56.63514	1	7.40E-05	3.33E-04
Con 5770	79.39815	53.62838	1	7.66E-05	3.33E-04
Clostridium sp. HGF2	48.27778	76.33784	0	8.28E-05	3.43E-04
CRC 6481	52.09259	73.55405	0	9.87E-05	3.90E-04
Cloacibacillus evryensis	52.73148	73.08784	0	1.13E-04	4.23E-04
Con 1987	79.42593	53.60811	1	1.17E-04	4.23E-04
Con 4595	77.21296	55.22297	1	1.38E-04	4.81E-04
Con 1617	76.12963	56.01351	1	1.50E-04	5.03E-04
Con 1371	78.46296	54.31081	1	2.05E-04	6.60E-04
Lachnospiraceae bacterium 5_1_57FAA	49.96296	75.10811	0	2.49E-04	7.73E-04
Eubacterium biforme	74.68519	57.06757	1	3.00E-04	8.70E-04
Faecalibacterium prausnitzii	78.25926	54.45946	1	3.00E-04	8.70E-04
Con 4699	78.78704	54.07432	1	3.13E-04	8.79E-04
Desulfovibrio sp. 6_1_46AFAA	53.33333	72.64865	0	3.70E-04	9.87E-04

Con 1529	75.05556	56.7973	1	3.74E-04	9.87E-04
Ruminococcus torques	76.92593	55.43243	1	5.28E-04	1.35E-03
Coprobacillus sp. 3_3_56FAA	50.53704	74.68919	0	6.01E-04	1.46E-03
Streptococcus equinus	54.52778	71.77703	0	6.02E-04	1.46E-03
Synergistes sp. 3_1_syn1	54.37963	71.88514	0	6.89E-04	1.62E-03
Lachnospiraceae bacterium 8_1_57FAA	51.88889	73.7027	0	7.91E-04	1.81E-03
Klebsiella pneumoniae	74.7037	57.05405	1	8.33E-04	1.86E-03
Eubacterium eligens	79.53704	53.52703	1	9.07E-04	1.97E-03
Clostridium bolteae	51.39815	74.06081	0	9.27E-04	1.97E-03
Con 1513	76.59259	55.67568	1	1.02E-03	2.11E-03
Clostridium citroniae	51.71296	73.83108	0	1.08E-03	2.19E-03
Fusobacterium varium	54.57407	71.74324	0	1.15E-03	2.28E-03
Bacteroides clarus	75.55556	56.43243	1	1.29E-03	2.50E-03
Ruminococcus obeum	77.53704	54.98649	1	1.34E-03	2.54E-03
Con 2606	77.5	55.01351	1	1.42E-03	2.59E-03

Lachnospiraceae bacterium 3_1_46FAA	52.53704	73.22973	0	1.44E-03	2.59E-03
CRC 2867	52.31481	73.39189	0	1.46E-03	2.59E-03
Con 6037	77.5463	54.97973	1	1.56E-03	2.71E-03
Clostridium sp. L2-50	76.37963	55.83108	1	1.61E-03	2.75E-03
Con 1867	76.38889	55.82432	1	2.13E-03	3.57E-03
Roseburia intestinalis	76.99074	55.38514	1	2.20E-03	3.58E-03
Subdoligranulum sp. 4_3_54A2FAA	51.56481	73.93919	0	2.24E-03	3.58E-03
Con 1197	75.42593	56.52703	1	2.26E-03	3.58E-03
CRC 4069	53.7963	72.31081	0	2.56E-03	3.96E-03
Con 8757	77.17593	55.25	1	2.60E-03	3.96E-03
Con 5752	73.65741	57.81757	1	2.71E-03	4.07E-03
Con 4295	74.98148	56.85135	1	2.95E-03	4.34E-03
Eubacterium rectale	75.90741	56.17568	1	3.21E-03	4.60E-03
Con 2494	74.35185	57.31081	1	3.22E-03	4.60E-03
Con 7367	76.23148	55.93919	1	3.63E-03	5.09E-03

Con 4829	76.7963	55.52703	1	3.88E-03	5.35E-03
Con 356	75.94444	56.14865	1	3.95E-03	5.37E-03
Dorea formicigenerans	52.98148	72.90541	0	4.36E-03	5.84E-03
Con 10559	76.59259	55.67568	1	4.52E-03	5.91E-03
Con 563	72.7037	58.51351	1	4.55E-03	5.91E-03
Con 4909	75.72222	56.31081	1	4.79E-03	6.12E-03
Con 6128	76.22222	55.94595	1	4.86E-03	6.13E-03
Con 2503	74.14815	57.45946	1	6.02E-03	7.46E-03
CRC 3579	54.05556	72.12162	0	6.09E-03	7.46E-03
Con 2703	74.55556	57.16216	1	7.67E-03	9.15E-03
Con 6068	75.74074	56.2973	1	7.67E-03	9.15E-03
Con 1604	71.92593	59.08108	1	8.96E-03	1.05E-02
Con 5615	76.07407	56.05405	1	9.70E-03	1.12E-02
Lachnospiraceae bacterium 3_1_57FAA_CT1	54.07407	72.10811	0	1.04E-02	1.19E-02
Con 569	73.41667	57.99324	1	1.30E-02	1.46E-02

Con 631	70.01852	60.47297	1	1.31E-02	1.46E-02
Con 1241	76.27778	55.90541	1	1.46E-02	1.61E-02
Alistipes indistinctus	54.50926	71.79054	0	1.59E-02	1.72E-02
Con 8420	72.64815	58.55405	1	2.32E-02	2.48E-02
Burkholderiales bacterium 1_1_47	72.37963	58.75	1	2.34E-02	2.48E-02
Con 7993	73.74074	57.75676	1	3.01E-02	3.16E-02
Con 425	73.19444	58.15541	1	3.87E-02	4.01E-02
Con 561	70.5	60.12162	1	4.81E-02	4.92E-02

Supplementary Table S9. PERMANOVA analysis of variation in three CRC-enriched species measured by three different methods in cohort C1. CRC- and colonoscopy-related factors explain the variation in these three species.

Parameter	**Df**	**mOTU species**					**IMG species**					**MLG species**				
		SumsOf Sqs	**MeanSqs**	**F. Model**	**R^2**	**Pr(>F)**	**SumsOf Sqs**	**MeanSqs**	**F.Model**	**R^2**	**Pr(>F)**	**SumsOf Sqs**	**MeanSqs**	**F.Model**	**R^2**	**Pr(>F)**
CRC Status	1	5.85E-05	5.85E-05	5.1835238	0.0395135	**0.0076**	2.42E-04	2.42E-04	4.2189512	0.0323989	**0.0127**	7.02E-03	7.02E-03	5.9492807	0.0450876	**0.0072**
Duration between colonoscopy and fecal	1	4.05E-05	4.05E-05	3.5159771	0.0273583	**0.0523**	1.57E-04	1.57E-04	2.6787139	0.0209801	**0.0777**	4.25E-03	4.25E-03	3.5265637	0.0274384	**0.0569**

sample collection																
Fecal sampling before or after colonoscopy	1	3.21E-05	3.21E-05	2.7722393	0.0216967	**0.0799**	1.12E-04	1.12E-04	1.8992995	0.0149670	0.163	3.54E-03	3.54E-03	2.9217093	0.0228398	**0.0799**
Stage of CRC	4	8.38E-05	2.09E-05	1.8432688	0.0565537	0.1262	4.44E-04	1.11E-04	1.9437773	0.0594540	0.1157	1.27E-02	3.17E-03	2.7293564	0.0815236	**0.0354**
Lesion location	1	3.02E-05	3.02E-05	1.5272855	0.0236688	0.1846	1.28E-04	1.28E-04	1.2152307	0.0189243	0.1988	2.27E-03	2.27E-03	1.0493068	0.0163828	0.3215
LDL	1	2.03E-05	2.03E-05	1.4217908	0.0140186	0.2414	2.52E-05	2.52E-05	0.3436566	0.0034248	0.5793	6.77E-04	6.77E-04	0.4524804	0.0045044	0.5249
eGFR	1	5.78E-06	5.78E-06	0.4256440	0.0039622	0.5138	4.77E-06	4.77E-06	0.0692402	0.0006467	0.8438	3.31E-04	3.31E-04	0.2318740	0.0021624	0.6453
TCHO	1	1.24E-05	1.24E-05	0.8618039	0.0085444	0.3454	7.84E-06	7.84E-06	0.1067080	0.0010659	0.7915	2.81E-04	2.81E-04	0.1872153	0.0018687	0.6821
Lesion specific location	1	4.15E-06	4.15E-06	0.2052181	0.0032469	0.6648	1.41E-06	1.41E-06	0.0131386	0.0002085	0.9754	8.14E-05	8.14E-05	0.0370280	0.0005874	0.9353
HDL	1	3.24E-07	3.24E-07	0.0222985	0.0002229	0.9401	4.69E-06	4.69E-06	0.0638119	0.0006377	0.8687	3.50E-05	3.50E-05	0.0232691	0.0002326	0.955
Age	1	1.75E-07	1.75E-07	0.0148715	0.0001180	0.9652	3.05E-06	3.05E-06	0.0515304	0.0004088	0.8841	3.47E-05	3.47E-05	0.0280829	0.0002228	0.9507
FBG	1	4.03E-06	4.03E-06	0.2850014	0.0028997	0.5725	1.73E-05	1.73E-05	0.2322323	0.0023641	0.6205	1.70E-03	1.70E-03	1.1175736	0.0112752	0.2544
BMI	1	1.41E-06	1.41E-06	0.1195008	0.0009551	0.749	1.07E-05	1.07E-05	0.1801544	0.0014392	0.6958	8.11E-05	8.11E-05	0.0651803	0.0005212	0.8618
Cr	1	2.32E-06	2.32E-06	0.1668589	0.0015866	0.6698	3.16E-06	3.16E-06	0.0449746	0.0004281	0.8759	1.61E-04	1.61E-04	0.1103230	0.0010496	0.7615
ALT/GPT	1	8.01E-07	8.01E-07	0.0625344	0.0005896	0.8156	6.22E-06	6.22E-06	0.0929296	0.0008759	0.7813	5.69E-04	5.69E-04	0.4106836	0.0038594	0.4907
TNM	15	5.83E-05	3.89E-06	0.1815751	0.0448528	0.9841	3.68E-04	2.46E-05	0.2193220	0.0536766	0.9134	1.15E-02	7.68E-04	0.3435946	0.0816089	0.8323

TG	1	3.80E-07	3.80E-07	0.0261886	0.0002618	0.9144	6.05E-07	6.05E-07	0.0082320	0.0000823	0.9827	1.39E-04	1.39E-04	0.0922060	0.0009212	0.7912
Gender	1	1.07E-06	1.07E-06	0.0908585	0.0007206	0.8475	9.10E-06	9.10E-06	0.1537437	0.0012187	0.8233	1.65E-04	1.65E-04	0.1336220	0.0010594	0.7801
DM	1	5.19E-07	5.19E-07	0.0441774	0.0003505	0.9158	4.74E-06	4.74E-06	0.0800697	0.0006351	0.8975	2.34E-04	2.34E-04	0.1895356	0.0015020	0.7209

Supplementary Table S10. List of 13 genera associated with CRC status in cohort C1.

	Control rank mean	**Case rank mean**	**Enrichment(0:case/1:control)**	**P-value**	**q-value**
Parvimonas	38.55556	83.43243	0	3.97E-11	3.86E-08
Peptostreptococcus	40.55556	81.97297	0	5.49E-10	2.67E-07
Fusobacterium	45.51852	78.35135	0	6.90E-07	2.24E-04
Beggiatoa	45.89815	78.07432	0	1.78E-06	4.34E-04
Malassezia	46.35185	77.74324	0	8.71E-06	1.70E-03
Paracoccus	47.66667	76.78378	0	1.10E-05	1.79E-03
Leptotrichia	48.15741	76.42568	0	3.40E-05	4.74E-03
Filifactor	49.06481	75.76351	0	3.94E-05	4.80E-03
Crenothrix	48.76852	75.97973	0	5.14E-05	5.57E-03

Solobacterium	47.66667	76.78378	0	8.79E-05	8.56E-03
Sulfurovum	49.48148	75.45946	0	1.14E-04	9.64E-03
Eubacterium	80.07407	53.13514	1	1.19E-04	9.64E-03
Streptobacillus	52.35185	73.36486	0	1.44E-04	1.08E-02
Adhaeribacter	77.06481	55.33108	1	4.79E-04	3.33E-02
Moniliophthora	49.91667	75.14189	0	6.39E-04	4.15E-02

Supplementary Table S11. List of phyla significantly associating with CRC status in cohort C1.

Phylum	**Control rank mean**	**Case rank mean**	**Enrichment (0:case/1:control)**	**P-value**	**q-value**
Fusobacteria	44.68519	78.95946	0	0.00000014	**0.000005**
Firmicutes	73.44444	57.97297	1	0.02924627	0.259876
Cloacimonetes	69.25926	61.02703	1	0.03419421	0.259876

Supplementary Table S12. IMG, mOTU and MLG species markers. IMG, mOTU and MLG species markers identified using random forest method among species associated with CRC (**Supplementary Table 8**). Marker species are listed by their importance reported by the method. MLG species identifiers starting with “Con_” are enriched in control samples, and those starting with “CRC_” are enriched in CRC samples.

17 IMG species markers					
	Control rank mean	**Case rank mean**	**Enrichment (0:case/1:control)**	**P-value**	**q-value**
Peptostreptococcus stomatis	37.25926	84.37838	0	5.11E-12	1.32E-08
Parvimonas micra	38.43519	83.52027	0	4.21E-11	5.43E-08
Parvimonas sp. oral taxon 393	39.81481	82.51351	0	2.79E-10	2.40E-07
Parvimonas sp. oral taxon 110	43.52778	79.80405	0	6.17E-08	3.98E-05
Gemella morbillorum	43.87037	79.55405	0	1.53E-07	7.88E-05
Fusobacterium nucleatum	45.09259	78.66216	0	3.86E-07	1.56E-04
Leptotrichia buccalis	45.60185	78.29054	0	4.44E-07	1.56E-04
Fusobacterium sp. oral taxon 370	45.02778	78.70946	0	4.83E-07	1.56E-04
Burkholderia mallei	45.19444	78.58784	0	7.93E-07	2.27E-04
Prevotella intermedia	46.47222	77.65541	0	1.92E-06	4.95E-04
Streptococcus dysgalactiae	47.06481	77.22297	0	4.18E-06	8.99E-04
Beggiatoa sp. PS	46.53704	77.60811	0	5.03E-06	9.97E-04

Malassezia globosa	46.35185	77.74324	0	8.71E-06	1.60E-03
Paracoccus denitrificans	47.48148	76.91892	0	1.18E-05	2.02E-03
Eubacterium ventriosum	80.98148	52.47297	1	1.27E-05	2.05E-03
Filifactor alocis	49.06481	75.76351	0	3.94E-05	5.65E-03
Solobacterium moorei	47.66667	76.78378	0	8.79E-05	9.85E-03
7 mOTU species markers					
	Control rank mean	**Case rank mean**	**Enrichment(0:case/1:control)**	**P-value**	**q-value**
Gemella morbillorum	47.93518519	76.58783784	0	8.63E-07	1.18E-04
Parvimonas micra	46.2962963	77.78378378	0	2.31E-08	7.73E-06
Peptostreptococcus stomatis	46.25	77.81756757	0	2.81E-08	7.73E-06
motu_linkage_group_316	79.61111111	53.47297297	1	7.03E-05	3.86E-03
motu_linkage_group_407	81.13888889	52.35810811	1	8.51E-06	9.34E-04
motu_linkage_group_490	80.46296296	52.85135135	1	3.04E-05	2.78E-03
motu_linkage_group_624	51.01851852	74.33783784	0	1.33E-03	3.69E-02
27 MLG species markers					

	Control rank mean	Case rank mean	Enrichment(0:case/1:control)	P-value	q-value
Parvimonas micra	38.40741	83.54054	0	5.56E-12	4.84E-10
Fusobacterium nucleatum	40.32407	82.14189	0	1.72E-10	7.48E-09
Solobacterium moorei	42.2037	80.77027	0	4.01E-08	1.16E-06
Clostridium symbiosum	46.31481	77.77027	0	2.67E-06	5.80E-05
Con_10180	82.03704	51.7027	1	6.06E-06	1.05E-04
CRC_2881	51.25926	74.16216	0	7.57E-06	1.10E-04
Coprococcus sp. ART55/1	80.85185	52.56757	1	2.09E-05	2.05E-04
Clostridium hathewayi	46.77778	77.43243	0	2.12E-05	2.05E-04
Clostridiales bacterium 1_7_47FAA	48.16667	76.41892	0	2.49E-05	2.17E-04
CRC_4136	50.99074	74.35811	0	2.97E-05	2.32E-04
butyrate-producing bacterium SS3/4	80.57407	52.77027	1	3.19E-05	2.32E-04
Haemophilus parainfluenzae	80.49074	52.83108	1	4.18E-05	2.69E-04
Con_154	80.35185	52.93243	1	4.45E-05	2.69E-04
Bacteroides fragilis	49.09259	75.74324	0	5.56E-05	3.02E-04

Con_1979	79.94444	53.22973	1	6.03E-05	3.09E-04
Con_7958	75.27778	56.63514	1	7.40E-05	3.33E-04
Con_5770	79.39815	53.62838	1	7.66E-05	3.33E-04
CRC_6481	52.09259	73.55405	0	9.87E-05	3.90E-04
Con_1987	79.42593	53.60811	1	1.17E-04	4.23E-04
Con_4595	77.21296	55.22297	1	1.38E-04	4.81E-04
Eubacterium biforme	74.68519	57.06757	1	3.00E-04	8.70E-04
Desulfovibrio sp. 6_1_46AFAA	53.33333	72.64865	0	3.70E-04	9.87E-04
Clostridium citroniae	51.71296	73.83108	0	1.08E-03	2.19E-03
Fusobacterium varium	54.57407	71.74324	0	1.15E-03	2.28E-03
Roseburia intestinalis	76.99074	55.38514	1	2.20E-03	3.58E-03
Dorea formicigenerans	52.98148	72.90541	0	4.36E-03	5.84E-03
CRC_3579	54.05556	72.12162	0	6.09E-03	7.46E-03

Supplementary Table S13. 20 gene markers identified by the mRMR feature selection method in cohort C1. Detailed information regarding their enrichment, occurrence in CRC cases and controls, statistical test of association, taxonomy and identity percentage are listed.

Marker gene id	**Enrich-ment**	**Wilcoxon rank-sum test**		**Occurrence**				**Identity**	**Taxonomy (Blastn to IMG v400)**	**Description (Blastp to KEGG v59)**
				Control (n=54)		**Case (n=74)**				
		P-value	**q-value**	**N**	**Rate(%)**	**N**	**Rate(%)**			
2361423	Case	2.31E-13	4.88E-07	11	20.37037037	62	83.78378378	93.87	*Peptostreptococcus anaerobius*	transposase
3173495	Case	6.24E-13	6.58E-07	10	18.51851852	61	82.43243243	93.98	*Peptostreptococcus anaerobius*	transposase
2040133	Case	7.51E-10	4.06E-04	14	25.92592593	62	83.78378378	99.4	*Clostridium symbiosum*	cobalt/nickel transport system permease protein
1696299	Case	7.70E-10	4.06E-04	2	3.703703704	43	58.10810811	99.78	*Parvimonas micra*	DNA-directed RNA polymerase subunit beta
482585	Case	7.41E-09	1.05E-03	16	29.62962963	58	78.37837838	NA	NA	RNA-directed DNA polymerase
2211919	Control	4.98E-08	2.20E-03	49	90.74074074	47	63.51351351	80.99	*Coprobacillus* sp. 8_2_54BFAA	NA
4171064	Control	7.50E-08	2.61E-03	40	74.07407407	18	24.32432432	94.94	*Faecalibacterium prausnitzii*	cytidine deaminase
1704941	Case	7.53E-08	2.61E-03	2	3.703703704	39	52.7027027	99.13	*Fusobacterium nucleatum*	butyryl-CoA dehydrogenase

3319526	Control	1.08E-07	2.79E-03	32	59.25925926	10	13.51351351	90.01	*Faecalibacterium prausnitzii*	NA
3246804	Case	1.80E-07	3.24E-03	1	1.851851852	35	47.2972973	NA	NA	citrate-Mg2+:H+ or citrate-Ca2+:H+ symporter, CitMHS family
3976414	Control	4.42E-07	4.07E-03	30	55.55555556	9	12.16216216	87.12	*Faecalibacterium prausnitzii*	adenosylcobinamide-phosphate synthase CobD
4256106	Control	7.39E-07	4.53E-03	28	51.85185185	9	12.16216216	NA	NA	integrase/recombinase XerD
3531210	Control	1.44E-06	5.63E-03	13	24.07407407	0	0	NA	NA	GDP-L-fucose synthase
3611706	Control	1.68E-06	5.82E-03	15	27.77777778	0	0	NA	NA	anti-repressor protein
2206475	Control	1.81E-06	5.95E-03	28	51.85185185	9	12.16216216	98.59	*Eubacterium ventriosum*	beta-glucosidase
181682	Control	1.95E-06	6.09E-03	34	62.96296296	15	20.27027027	99.25	*Roseburia intestinalis*	NA
1804565	Control	2.03E-06	6.16E-03	22	40.74074074	4	5.405405405	NA	NA	branched-chain amino acid transport system ATP-binding protein
2736705	Case	5.71E-06	8.55E-03	2	3.703703704	32	43.24324324	99.68	*Clostridium hathewayi*	NA
1559769	Control	1.03E-05	1.04E-02	27	50	7	9.459459459	88.65	*Coprococcus catus*	polar amino acid transport system substrate-binding protein
370640	Control	2.64E-05	1.47E-02	14	25.92592593	0	0	99.4	*Bacteroides clarus*	NA

Supplementary Table S14. PERMANOVA analysis of variation in 20 CRC-associated gene markers in cohort C1. CRC status and stage explain the variation in these gene profiles, while fasting blood glucose (FBG) moderately explains the variation. See **Supplementary Table S4** for explanation of parameters in column 1.

Parameter	**Df**	**SumsOfSqs**	**MeanSqs**	**F.Model**	**R^2**	**Pr(>F)**	**q-value**
CRC Status	1	5.5793661	5.5793661	16.626711	0.116575	0.0001	0.00095
Stage of CRC	4	6.7812635	1.6953159	5.0761083	0.1416874	0.0001	0.00095
FBG	1	0.8119553	0.8119553	2.154786	0.0215146	0.0073	0.046233
Fecal sampling before or after colonoscopy	1	0.5473702	0.5473702	1.4588296	0.011536	0.0978	0.46455
Lesion location	1	0.500106	0.500106	1.4185104	0.0220202	0.1329	0.486163
Lesion specific location	7	2.7831853	0.3975979	1.1372468	0.1225468	0.1889	0.486163
HDL	1	0.4718905	0.4718905	1.2480119	0.0123263	0.203	0.486163
ALT/GPT	1	0.4650084	0.4650084	1.2366953	0.0115324	0.2047	0.486163
Duration between colonoscopy and fecal sample collection	1	0.4170429	0.4170429	1.1084063	0.0087893	0.3116	0.657822
Age	1	0.3976816	0.3976816	1.0557238	0.0083091	0.3669	0.676838
TCHO	1	0.3768657	0.3768657	0.9942006	0.0098441	0.4287	0.676838
DM	1	0.3653642	0.3653642	0.9692711	0.0076339	0.4617	0.676838
BMI	1	0.3660728	0.3660728	0.9708139	0.0077067	0.4631	0.676838

Cr	1	0.3412225	0.3412225	0.8963725	0.0084646	0.5617	0.719847
TNM	15	5.2686733	0.3512449	0.9797038	0.2021521	0.5683	0.719847
LDL	1	0.308397	0.308397	0.8136124	0.0080705	0.6624	0.741782
Gender	1	0.3092058	0.3092058	0.8193202	0.0064605	0.6637	0.741782
TG	1	0.291975	0.291975	0.7695216	0.0076365	0.7334	0.774144
eGFR	1	0.2043621	0.2043621	0.539403	0.0050159	0.9496	0.9496

Supplementary Table S15. CRC index estimated in cohort C1, a type 2 diabetes (T2D) cohort and an inflammatory bowel disease (IBD) cohort.

Cohort/group	**Median CRC index**	**Comparison with C1 patients**	
		***P*-value**	**q-value**
C1 patients	7.30636	NA	NA
C1 controls	-5.558923	3.91E-21	4.89E-21
T2D patients	0.2512602	1.71E-26	2.85E-26
T2D controls	-1.47849	2.00E-30	1.00E-29
IBD patients	-1.789305	6.00E-11	6.00E-11

IBD controls	-4.505388	1.27E-28	3.16E-28

Supplementary Table S16. Baseline characteristics of the Chinese cohort C2 consisting 47 CRC patients and 109 control individuals. For quantitative traits, the median, minimum and maximum are shown. FBG: fasting blood glucose; ALT/GPT: alanine transaminase/glutamate pyruvated transaminase; BMI: body mass index; DM: diabetes mellitus type 2; HDL: high density lipoprotein; TG: triglyceride; eGFR: epidermal growth factor receptor; TCHO: total cholesterol; Cr: creatinine; LDL; low density lipoprotein; TNM: tumor node metastasis staging system; Statistical tests used for identifying associations between metadata and CRC: † - Wilcoxon test, ‡ - Fisher's exact test.

Parameter	**Controls (n=109)**	**Cases (n=47)**	***P-value***	**q-value**
Age	58 (43,68)	69 (48,90)	3.146E-06†	1.363E-05
Gender (M:F)	40:69	25:22	0.07626‡	0.1824
BMI	23.02 (18.59,30.8)	20.94 (15.83,31.68)	0.7098†	0.7098
Stage of CRC (1:2:3:4)	n.a	4:24:15:4	n.a	n.a
Distribution of detailed TNM stages (T1N0:T3N0:T1N1:T3N1:T3N2:T4N1:T2N1M1:T3N1M1:T3N2M1:UT4:Mx)	n.a	4:23:1:9:4:1:1:1:1:1:1	n.a	n.a
Leison location (1:2:NA)	n.a	9:20:18	n.a	n.a
Leison specific location (2:3:4:6:7:8:9:NA)	n.a	3:3:3:2:7:4:7:18	n.a	n.a
Fecal sampling before or after colonoscopy	101:8 (93%:7%)	9:38 (19%:81%)	6.1669E-20‡	8.017E-19

(before:after)				
Duration between colonoscopy and fecal sample collection (days)	-63 (-202,92)	18 (-58,239)	4.064E-14[†]	2.642E-13
Duration of frozen storage of fecal samples (days)	374 (93,3526)	297 (30,3450)	0.2086[†]	0.3390
FBG	5 (4.5,6.3)	5.6 (4.5,7.9)	0.0842[†]	0.1824
TCHO	5.2 (3.8,5.9)	4.3 (3.6,5.3)	0.0769[†]	0.1824
LDL	2.9 (2,4.2)	2.5 (2.3,3.6)	0.6241[†]	0.6761
HDL	1.66 (1,2.03)	1.3 (0.9,2.6)	0.2822[†]	0.4076
TG	0.9 (0.7,2.08)	0.8 (0.5,1.9)	0.4680[†]	0.6084
Cr	74 (58,129)	70 (44,122)	0.5484[†]	0.6481
ALT/GPT	20 (14,68)	13 (10,36)	0.1043[†]	0.1937

Supplementary Table S17. Enrichment of two CRC-enriched and two control-enriched genes measured by qPCR in cohort C2.

Marker gene ID	**Gene description**	**Enrichment**	**Wilcoxon rank-sum test P-value**	**Wilcoxon rank-sum test stratified for colonoscopy**	**Mantel Haenszel Odds Ratio, adjusted for colonoscopy (95% CI)**	**Mantel Haenszel test P-value**

1704941	butyryl-CoA dehydrogenase	case	1.97E-09	1.52E-03	18.54 (2.62-131)	0.00509
482585	RNA-directed DNA polymerase	case	2.34E-03	4.55E-02	1.815 (0.653-5.05)	0.38
181682	gene with unknown function from Roseburia intestinalis	control	2.15E-01	3.13E-01	1.495 (0.456-4.9)	0.714
370640	gene with unknown function from Bacteroides clarus	control	3.11E-01	6.30E-01	1.647 (0.395-6.88)	0.778

Supplementary Table S18. Baseline characteristics of the Danish cohort (cohort D) consisting 16 CRC patients and 24 control individuals. For quantitative traits, the median, minimum and maximum are shown. BMI: body mass index; DM: diabetes mellitus type 2; TNM: tumor node metastasis staging system; Statistical tests used for identifying associations between metadata and CRC: † - Wilcoxon test, ‡ - Fisher's exact test.

Parameter	**Control (n=24)**	**Case (n=16)**	***P-value***	**q-value**
Age	65.5 (30, 87)	67.5 (47, 78)	0.4308219†	0.6376
Gender (M:F)	07:17	10:06	0.05309‡	0.15927
BMI	25.88 (18.94, 35.29)	25.89 (18.83, 33.20)	0.6328136†	0.6376
DM (YES:NO)	03:21	01:15	0.6376‡	0.6376
Stage of CRC (1:2:3:4)	n.a	1:9:5:1	n.a	n.a
Distribution of detailed TNM stages (T1N0M0V0:T3N0M0V0:T3N0M0V1: T3N1M0V0:T3N2M0V0:T4N0M0:	n.a	1:6:3:1:2:1:1:1	n.a	n.a

T4N2M0V1:T4NxMx)				
Cancer location (Distal:Proximal)	n.a	13:03	n.a	n.a
Cancer location (Adenocarcinom:Ascendens:Coecum:Rectum: Sigmoideum:Transversum)	n.a	1:1:1:9:3:1	n.a	n.a
Fecal sampling before or after colonoscopy (before:after)	24:0 (100%:0%)	12:4 (75%:25%)	0.0199[‡]	0.1194
Duration between colonoscopy and fecal sample collection (days)	7 (3, 89)	14 (-24, 252)	0.4466[†]	0.6376

Supplementary Table S19. Community structure differences between cohorts C1 and D. All comparisons were performed using Wilcoxon rank-sum test.

Gene count P-value					**Shannon index P-value**			
	D: Case	D: Control	C2: Case	C2: Control	D: Case	D: Control	C2: Case	C2: Control
D: Case		0.25991847	1.94E-05	0.000294527		0.772788361	5.84639E-05	4.02E-04
D: Control			7.86E-05	0.001729823			2.25586E-05	9.34E-04
C2: Case				0.212812929				0.178412749

Supplementary Table S20. Species annotation of the 1498 genes enriched in CRC patient microbiomes, both in cohort C1 and cohort D. A large fraction was annotated to *Parvimonas micra.* Annotated species with more than 10 genes are listed here.

Species	**Gene numbers (Total=1452)**
Parvimonas micra	389
Solobacterium moorei	204
Clostridium symbiosum	177
Clostridium sp. 7_3_54FAA	108
Parvimonas sp. oral taxon 110	93
Parvimonas sp. oral taxon 393	93
Fusobacterium nucleatum	64
Peptostreptococcus stomatis	23
Clostridium hathewayi	17
Clostridium citroniae	14
Akkermansia muciniphila	11
[Clostridium] *difficile*	11
Peptostreptococcus anaerobius	10

Supplementary Table S21. List of CRC-associated species predicted from Chinese cohort C1 and validated in Danish cohort D with q<0.05

IMG species validated in cohort D					
	Control rank mean	Case rank mean	Enrichment(0:case/1: control)	P-value	q-value
Parvimonas sp. oral taxon 110	14.54166667	29.4375	0	9.06E-05	0.000808962
Parvimonas sp. oral taxon 393	14.66666667	29.25	0	0.000127394	0.000808962
Parvimonas micra	14.70833333	29.1875	0	0.00015168	0.000808962
Gemella morbillorum	15.70833333	27.6875	0	0.001465743	0.005862972
Peptostreptococcus stomatis	16.16666667	27	0	0.003409134	0.010909228
Fusobacterium sp. oral taxon 370	16.58333333	26.375	0	0.010235287	0.024739601
Fusobacterium nucleatum	16.70833333	26.1875	0	0.010823576	0.024739601
Malassezia globosa	17	25.75	0	0.023703729	0.047407459
mOTU species validated in cohort D					

	Control rank mean	Case rank mean	Enrichment(0:case/1:control)	P-value	q-value
Peptostreptococcus stomatis	16.5	26.5	0	0.000139835	0.000978842
Parvimonas micra	16.70833333	26.1875	0	0.000749378	0.002622823
Gemella morbillorum	18	24.25	0	0.004603221	0.010740848
MLG species validated in cohort D					
	Control rank mean	**Case rank mean**	**Enrichment (1:Control;0:Case)**	**P-value**	**q-value**
Parvimonas micra	15.20833333	28.4375	0	9.13E-05	0.002329351
Solobacterium moorei	16.22916667	26.90625	0	0.000172545	0.002329351

Supplementary Table S22. List of four gene markers predicted from cohort C1 that show significant associations in cohort D with q<0.05.

Gene	**Cohort C1**			**Cohort D**			**Blastn on IMG v400**	**Blastp on KEGG v59**	
Marker ID	**P-value**	**q-value**	**Enrich**	**P-value**	**q-value**	**Enrich**	**Species taxonomy**	**KEGG ID**	**Gene annotation**
2361423	2.31148E-13	4.87836E-07	case	1.16E-04	0.00116	case	*Peptostreptococcus anaerobius*	K07485	transposase

3173495	6.23501E-13	6.57946E-07	case	1.85E-04	0.00123	case	*Peptostreptococcus anaerobius*	K07485	transposase
1696299	7.69646E-10	0.000406082	case	7.87E-05	0.00116	case	*Parvimonas micra*	K03043	DNA-directed RNA polymerase subunit beta
1704941	7.53342E-08	0.002606428	case	2.08E-03	0.01040	case	*Fusobacterium nucleatum*	K00248	butyryl-CoA dehydrogenase

Supplementary Table S23. PERMANOVA analysis of variation in four gene markers validated in cohort D (No. of permutations = 9999). CRC status explains the variation in these gene profiles.

phenotype	**Df**	**Sums Of Sqs**	**Mean Sqs**	**F.Model**	**R^2**	**Pr (>F)**
CRC Status	1	8.11E-11	8.11E-11	4.8910108	0.1140335	0.0001
Stage of CRC	4	1.15E-10	2.86E-11	1.6816488	0.1612064	0.1375
Duration between colonoscopy and fecal sample collection	1	2.03E-11	2.03E-11	1.1199259	0.028628	0.2265
Cancer location (Distal:Proximal)	1	5.20E-11	5.20E-11	1.2648699	0.0828615	0.2383
Cancer location(Adenocarcinom:Ascendens:Coecum:Rectum:Sigmoideum:Transversum)	5	3.12E-10	6.24E-11	1.9756046	0.4969319	0.2998
Age	1	1.48E-11	1.48E-11	0.8097989	0.0208658	0.3989
DM	1	5.61E-12	5.61E-12	0.3020817	0.0078868	0.5654

Gender	1	6.48E-12	6.48E-12	0.3495622	0.0091152	0.571
BMI	1	7.51E-12	7.51E-12	0.4060178	0.0105717	0.5869
DNA purification date	1	3.66E-12	3.66E-12	0.1966498	0.0051484	0.6696
Fecal sampling before or after colonoscopy	1	6.95E-12	6.95E-12	0.3749813	0.0097715	0.6878
TNM	7	1.57E-10	2.25E-11	0.3823119	0.2506686	0.7061

Supplementary Table S24. Enrichment of four marker genes in published Austrian and French cohorts (A and F, respectively).

Marker Gene ID	**Cohort A**			**Cohort F**			**Blastn on IMG v400**	**Blastp on KEGG v59**	
	P-value	**q-value**	**Enrich**	**P-value**	**q-value**	**Enrich**	**Species taxonomy**	**KEGG ID**	**Gene annotation**
2361423	9.465681e-06	3.786272e-05	case	1.805948e-06	7.223791e-06	case	*Peptostreptococcus anaerobius*	K07485	transposase
3173495	1.021888e-04	3.065663e-04	case	1.311802e-05	3.935405e-05	case	*Peptostreptococcus anaerobius*	K07485	transposase
1696299	3.089198e-03	3.089198e-03	case	3.471676e-03	3.471676e-03	case	*Parvimonas micra*	K03043	DNA-directed RNA polymerase subunit beta
1704941	5.007540e-04	1.001508e-03	case	9.687230e-05	1.937446e-04	case	*Fusobacterium nucleatum*	K00248	butyryl-CoA dehydrogenase

Supplementary Table S25. Comparison of enrichment of 20 marker genes in Chinese (C1), Danish (D), Austrian (A) and French (F) cohorts. Cells marked in red: *P*<0.05. Enrichment in case or control is only reported when *P*<0.2. Only cohort C1 was used to discover gene biomarkers, and these 20 genes were among the 102,514 that associated with CRC. In cohorts D, A and F, association of only these 20 genes were verified.

Gene id	**Chinese cohort C1**		**Danish cohort D**		**Austrian cohort A**		**French cohort F**	
	Case (1) Vs. Controls (0)		Case (1) Vs. Controls (0)		Carcinoma (1) Vs Controls (0)		Case (1) Vs. Controls (0)	
	p.value	Enrichment	p.value	Enrichment	p.value	Enrichment	p.value	Enrichment
181682	1.95E-06	0	0.900619951	NA	0.678813728	NA	0.007181249	0
370640	2.64E-05	0	0.495680726	NA	0.862554181	NA	0.901689843	NA
482585	7.41E-09	1	0.467868103	NA	0.114070684	1	0.09202366	1
1559769	1.03E-05	0	0.627103852	NA	0.613815329	NA	0.318983729	NA
1696299	7.70E-10	1	7.87E-05	1	0.003089198	1	0.003471676	1
1704941	7.53E-08	1	0.002080194	1	0.000500754	1	9.68723E-05	1
1804565	2.03E-06	0	0.345063544	NA	0.719304711	NA	1	NA
2040133	7.51E-10	1	0.923193148	NA	0.037408072	1	0.3620777	NA
2206475	1.81E-06	0	0.559844892	NA	0.239405355	NA	0.086939707	0
2211919	4.98E-08	0	0.343905238	NA	0.8730299	NA	0.403859093	NA
2361423	2.31E-13	1	0.000116036	1	9.46568E-06	1	1.80595E-06	1
2736705	5.71E-06	1	0.653175645	NA	0.085244448	1	0.321243655	NA
3173495	6.24E-13	1	0.00018455	1	0.000102189	1	1.3118E-05	1

3246804	1.80E-07	1	0.586270986	NA	0.834009147	NA	0.893668207	NA
3319526	1.08E-07	0	0.646619859	NA	0.847882874	NA	0.085059441	0
3531210	1.44E-06	0	0.23124459	NA	0.014329165	1	0.142060944	0
3611706	1.68E-06	0	1	NA	0.889823764	NA	0.346149329	NA
3976414	4.42E-07	0	0.539082044	NA	0.748143815	NA	0.458758072	NA
4171064	7.50E-08	0	0.705131044	NA	0.171937649	1	0.081938362	0
4256106	7.39E-07	0	0.702861448	NA	0.05048434	1	0.880361689	NA

Supplementary Table S26. Classification accuracy of the two marker genes measured by qPCR in cohort C2, stratified into early (I-II) and late (III-IV) stage cancer.

Group	Marker ID	Enrichment	Wilcox rank-sum test, P-value	Wilcoxon rank-sum test stratified for colonoscopy, P-value	Mantel Haenszel Odds Ratio adjusted for colonoscopy (95% CI)	Mantel-Haenszel test P-value
Stages I and II	1696299	case	6.51E-14	3.35E-06	21.5 (3.18-146)	1.38E-05
	1704941	case	4.15E-07	0.008654411	27.77 (1.64-469)	0.0322
	1696299 or 1704941		N.A.	N.A.	33.37 (4.49-248)	1.68E-06
Stages III and IV	1696299	case	1.51E-11	0.00027574	15.44(3.06-77.9)	0.00109
	1704941	case	4.40E-09	0.002700628	25.34(2.91-221)	0.00842
	1696299 or 1704941		N.A.	N.A.	15.77(3.52-70.6)	0.000653

Supplementary Table S27. Primer and probe sequences for qPCR measurement of five gene markers and controls.

Gene	**Sequence type**	**Nucleotide sequence**
1696299	Forward	AAGAATGGAGAGAGTTGTTAGAGAAAGAA
	Reverse	TTGTGATAATTGTGAAGAACCGAAGA
	Probe	AACTCAAGATCCAGACCTTGCTACGCCTCA
1704941	Forward	TTGTAAGTGCTGGTAAAGGGATTG
	Reverse	CATTCCTACATAACGGTCAAGAGGTA
	Probe	AGCTTCTATTGGTTCTTCTCGTCCAGTGGC
181682	Forward	CGGATTTGCAGTGGCAAGTT
	Reverse	TGATTGCAGACGCCAATGTC
	Probe	CGTGAAAAATCCGCGCATCTGGC
370640	Forward	TCCATCCGCAAGCCTTTACT
	Reverse	GCTTCCGGTGCCATTGACTA
	Probe	TTCATCATCACAGCCGACAACGCA

482585	Forward	AATGGGAATGGAGCGGATTC
	Reverse	CCTGCACCAGCTTATCGTCAA
	Probe	AAGCCTGCGGAACCACAGTTACCAGC
control	Forward	CGTCAGCTCGTGTCGTGAG
	Reverse	CGTCGTCCCCACCTTCC
	Probe	TTAAGTCCCACAACGAGCGCAACCC