

Table S1 Details of quality-score based filtering of the input sequencing reads using the QIIME2-DADA2 algorithm

Sample-id	Input	Filtered	% of input passed filter	Denosed	Merged	% of input merged	Non-chimeric	% of input non-chimeric
221-100	80286	59136	73.66	58556	56264	70.08	55860	69.58
221-101	88160	66844	75.82	66333	64360	73	62743	71.17
221-102	93058	69105	74.26	68590	66476	71.44	65741	70.65
221-103	89703	68455	76.31	67553	63809	71.13	63014	70.25
221-104	59400	44958	75.69	44185	41900	70.54	41624	70.07
221-105	94989	71068	74.82	70203	66587	70.1	65211	68.65
221-106	65142	50770	77.94	50321	48751	74.84	46577	71.5
221-107	40474	31077	76.78	30687	29449	72.76	29154	72.03
221-108	26617	21420	80.47	20939	17934	67.38	14412	54.15
221-109	102069	78168	76.58	77507	74872	73.35	72888	71.41
221-110	127464	95837	75.19	94711	91373	71.69	89452	70.18
221-111	81604	59106	72.43	58564	56084	68.73	56024	68.65
221-112	82951	65849	79.38	65124	62516	75.36	61721	74.41
221-113	111257	77847	69.97	77378	73035	65.65	70983	63.8
221-114	83759	62849	75.04	62423	59705	71.28	57503	68.65
221-115	48545	33139	68.26	32750	31050	63.96	31002	63.86
221-79	147535	108266	73.38	106740	101214	68.6	99155	67.21
221-80	71022	51056	71.89	50411	48508	68.3	48107	67.74
221-81	139977	82930	59.25	81612	72080	51.49	65837	47.03
221-82	152177	113246	74.42	111886	107572	70.69	106329	69.87
221-83	80978	65730	81.17	65365	61740	76.24	61365	75.78
221-84	128180	99043	77.27	98164	95137	74.22	93080	72.62
221-85	118558	86956	73.34	85953	83638	70.55	83356	70.31
221-86	256847	194346	75.67	192665	179599	69.92	177151	68.97
221-87	218101	164309	75.34	162494	155499	71.3	151978	69.68
221-88	100298	76134	75.91	72208	60250	60.07	51202	51.05
221-89	147559	113430	76.87	111505	100120	67.85	82576	55.96
221-90	332554	237411	71.39	234488	218460	65.69	198576	59.71
221-91	60731	45225	74.47	44503	42811	70.49	41349	68.09
221-92	81810	64543	78.89	63991	62274	76.12	61769	75.5
221-93	87123	63182	72.52	61277	54542	62.6	50938	58.47
221-94	113726	80326	70.63	79168	75810	66.66	72102	63.4
221-96	75753	57818	76.32	57275	53904	71.16	51942	68.57
221-97	149305	107617	72.08	106831	103282	69.18	100340	67.2
221-98	384311	251555	65.46	250730	228567	59.47	222054	57.78
221-99	67416	50611	75.07	50055	47780	70.87	46749	69.34

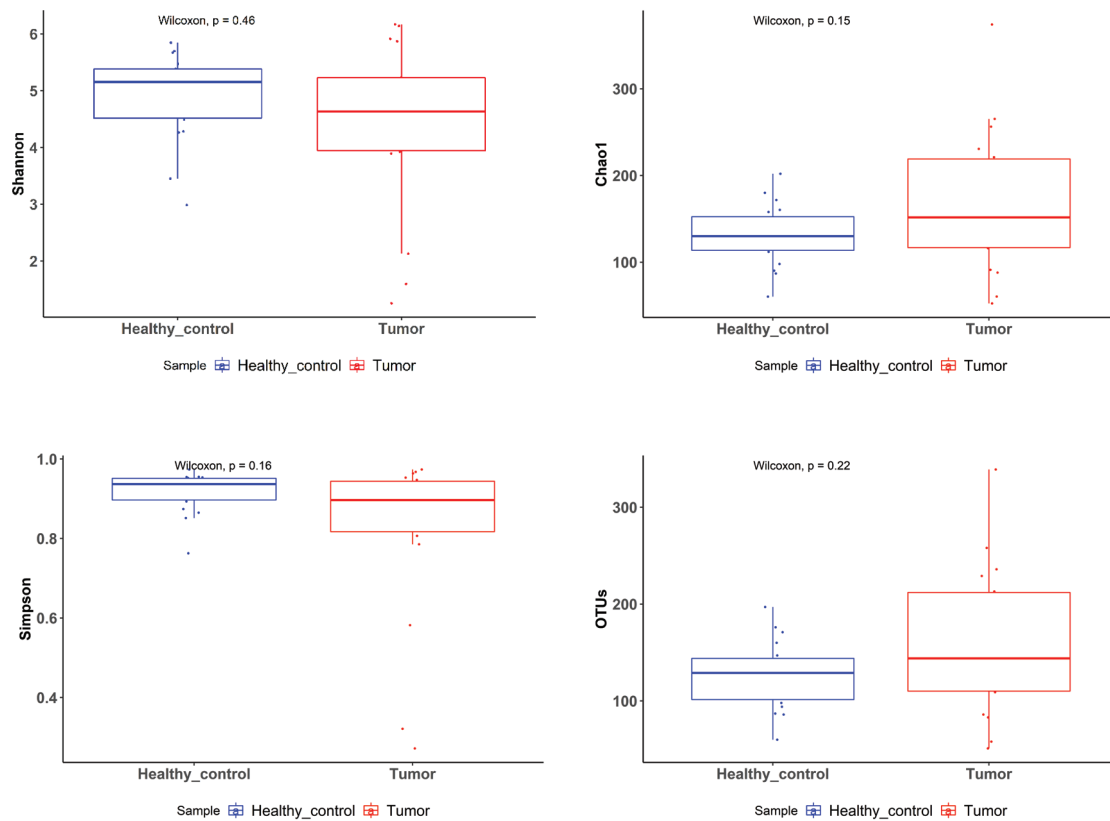


Figure S1 Reflects the alpha diversity measurements for CRC patients *vs.* healthy controls, measured by observed species, Chao1 richness estimator, Simpson diversity, and Shannon diversity and plotted for patients with CRC (red) and healthy controls (blue). The line inside the box represents the median; the whiskers represent the lowest and highest values within the 1.5 interquartile range (IQR). Outliers as well as individual sample values are shown as dots. CRC, colorectal cancer.

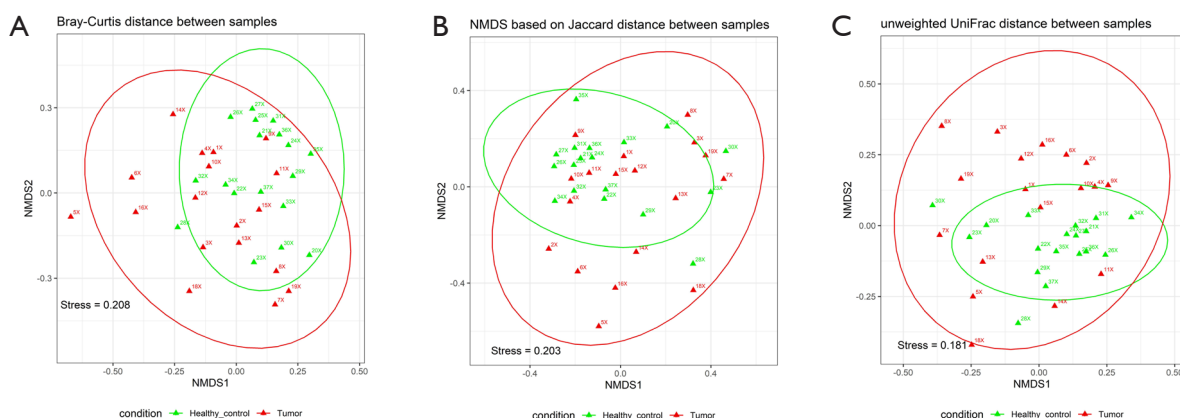


Figure S2 Beta diversity analysis: Non-metric Multidimensional Scaling (NMDS) ordination plots showing beta diversity comparisons between healthy control and tumor (CRC) samples. Ellipses represent the 95% confidence interval for healthy control and tumor groups, each dot in the ellipse represents a sample from that group. Statistics were calculated using PERMANOVA with 999 permutations. Beta diversity based on, (A) *Bray-Curtis* distance ($P=0.004$), (B) *Jaccard* distance ($P=0.002$), and (C) *unweighted UniFrac* distance ($P=0.0006$).

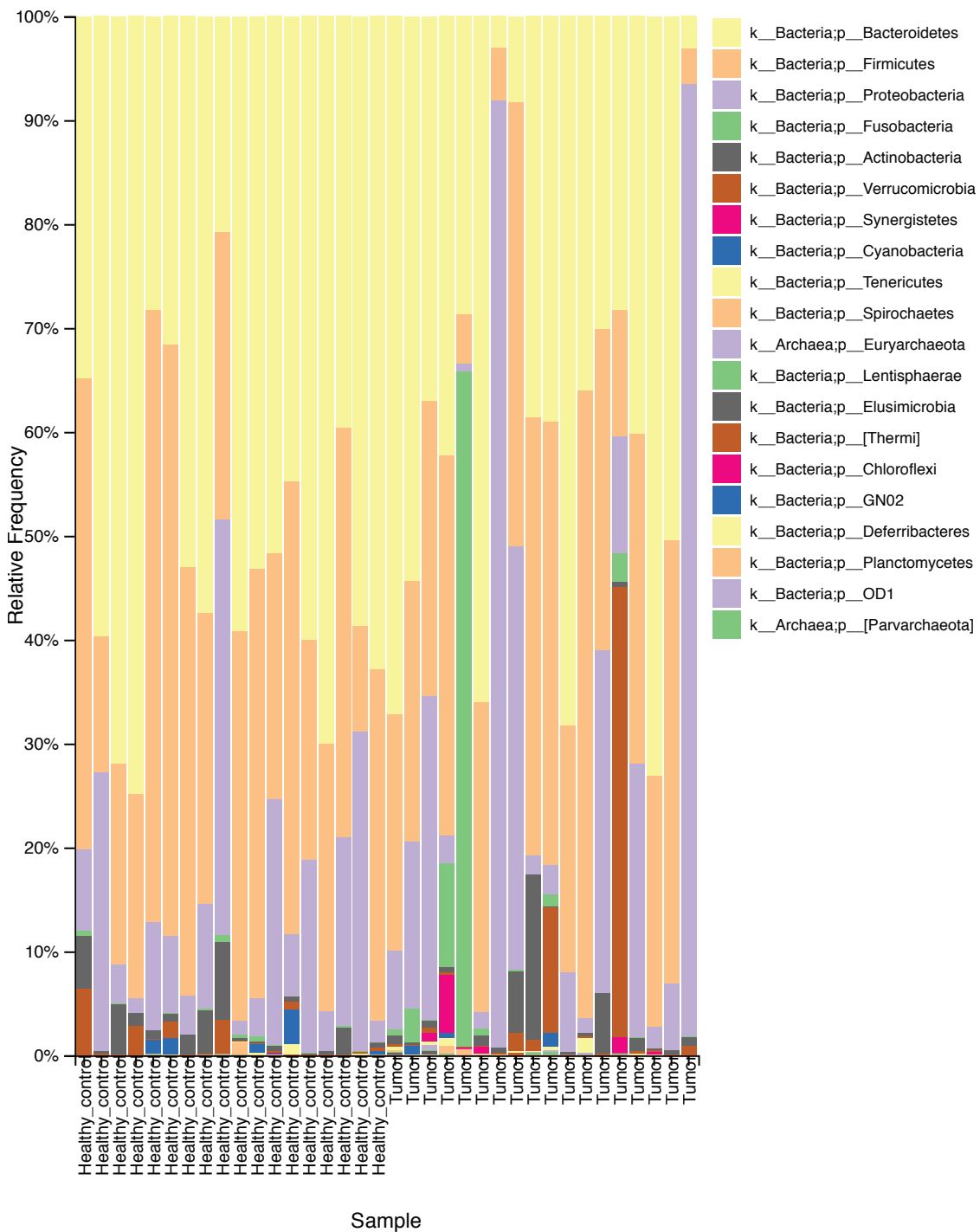


Figure S3 Relative frequencies of microbial taxa in CRC and healthy control samples at the phylum level.

Cladogram

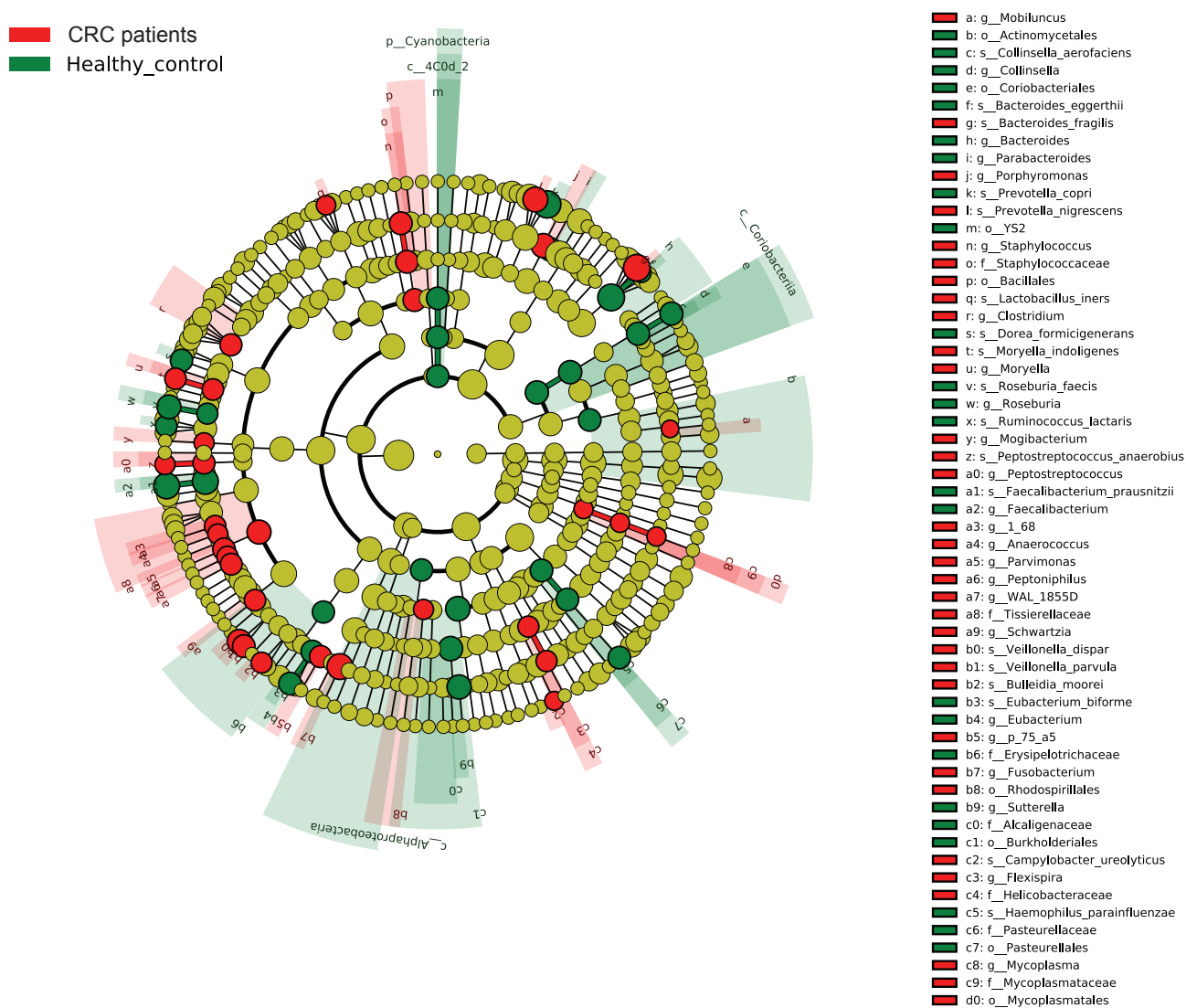


Figure S4 CRC patients and healthy controls differ in gut microbial composition. Cladogram for differentially distributed taxa ($P < 0.05$, $LDA > 2$) based on taxonomic abundance at the species level. The circles in the cladogram represent the phylogenetic tree from phylum to the species level. Each colored circle/dot on each phylogenetic level represents the taxa, and the diameter of the dot is proportional to the relative abundance of the taxa. Taxa with no significant differences are represented by the color yellow, whereas the other significant different taxa colored according to the group. CRC, colorectal cancer; LDA, Linear Discriminant Analysis.

Cladogram

- CRC patients over 40 years of age
- CRC patients under 40 years of age

- a: s_*Bacteroides_ovatus*
- b: s_*Prevotella_copri*
- c: s_*Lactobacillus_salivarius*
- d: s_*Clostridium_symbiosum*
- e: s_*Clostridium_amosum*
- f: g_*Herbaspirillum*
- g: f_*Oxalobacteraceae*
- h: g_*Enhydrobacter*
- i: f_*Xanthomonadaceae*
- j: o_*Xanthomonadales*

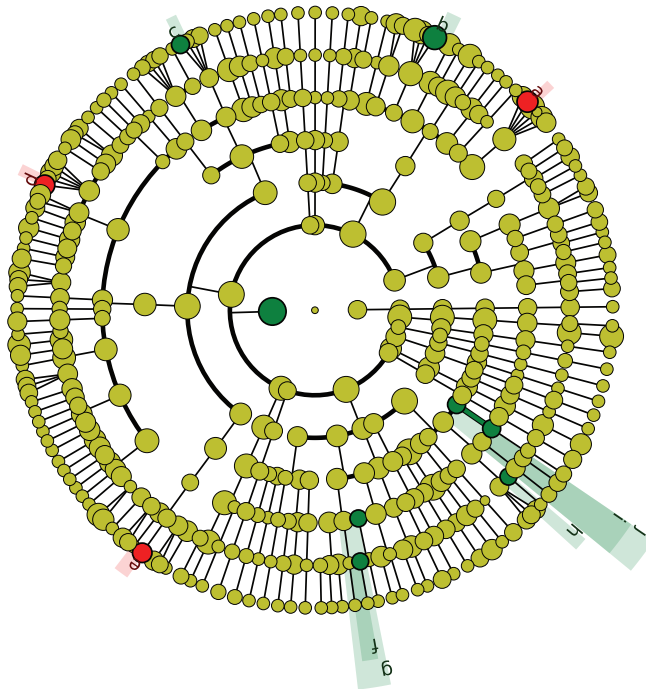


Figure S5 CRC patients under 40 years of age and CRC patients over 40 years of age differ in gut microbial composition. Cladogram for differentially distributed taxa ($P < 0.05$, LDA > 2) based on taxonomic abundance at the species level. The circles in the cladogram represent the phylogenetic tree from phylum to the species level. Each colored circle/dot on each phylogenetic level represents the taxa, and the diameter of the dot is proportional to the relative abundance of the taxa. Taxa with no significant differences are represented by the color yellow, whereas the other significant different taxa colored according to the group. CRC, colorectal cancer; LDA, Linear Discriminant Analysis.