Supplementary Figures

Supplementary Fig. 1 The analysis pipeline of taxonomic identification and quantification of tissue-resident microbes. Related to main Figure 1.

Supplementary Fig. 2 Assessments of microbial composition across different anatomic locations of colon and rectum. Related to main Figure 1.

Supplementary Fig. 3 Associations between somatic alterations and tumor-enriched taxa. Related to main Figure 2.

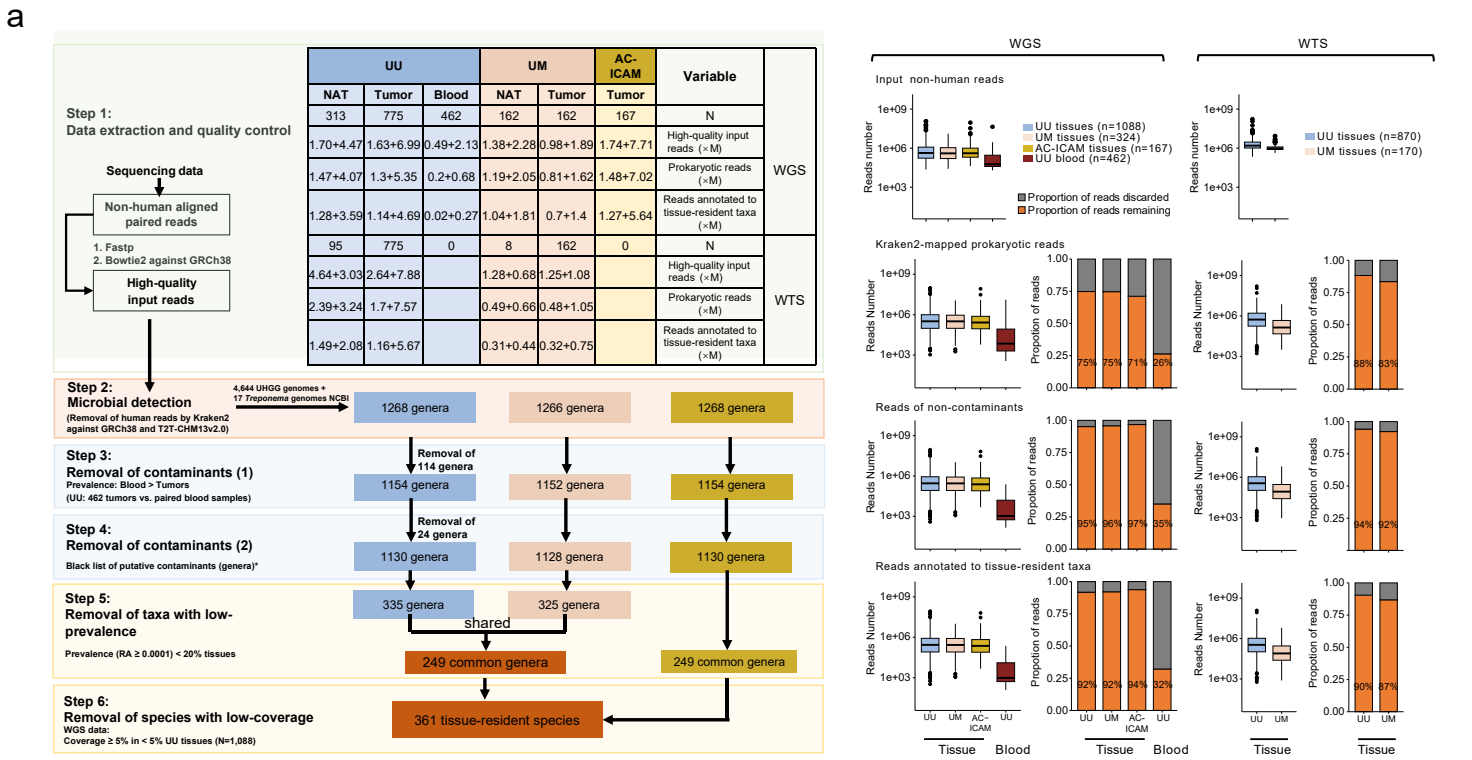
Supplementary Fig. 4
Prevalence of polyketide synthase (PKS) genes and relationship with CRC-associated mutational signature. Related to main Figure 3.

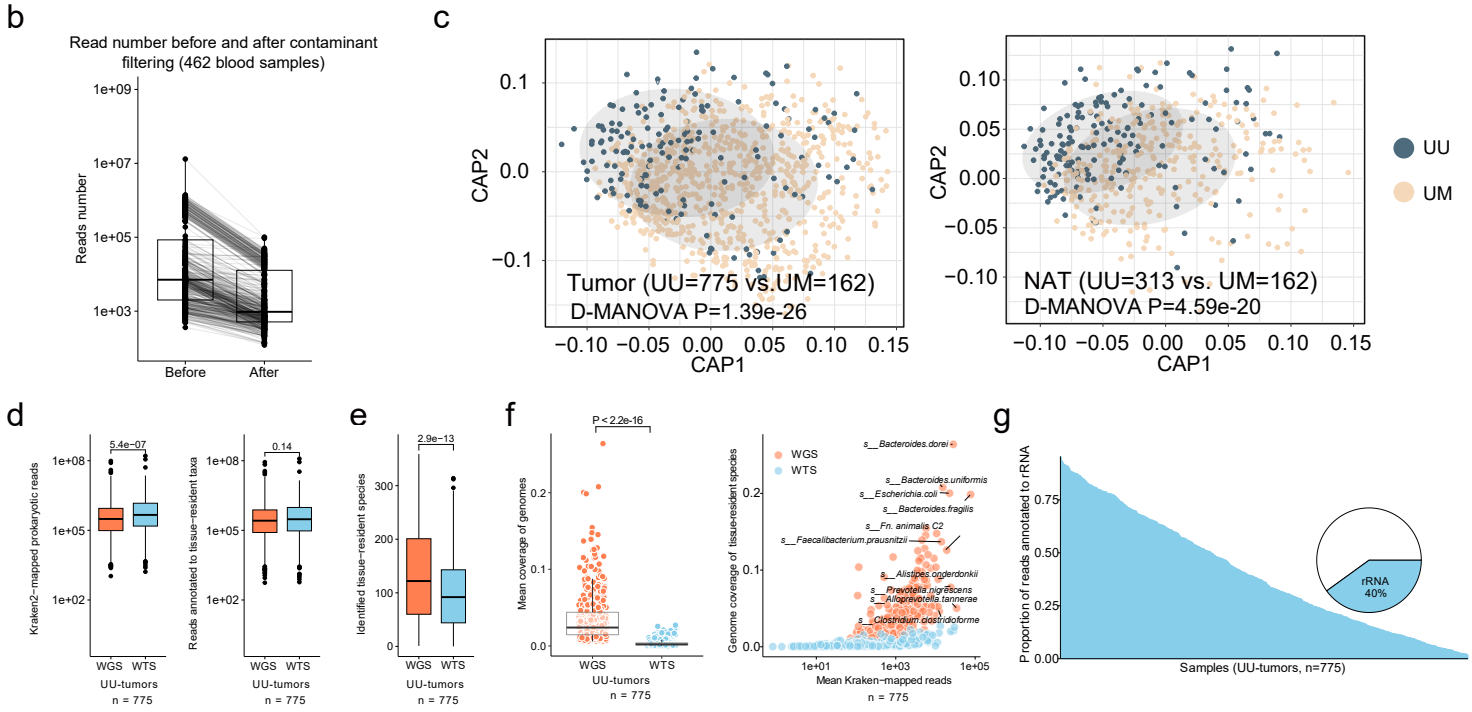
Supplementary Fig. 5
CMS-dependent association patterns between prognostic taxa and host expressions levels of CRC related pathways and immune cell populations. Related to main Figure 4.

Supplementary Fig. 6
Assessment of the prognostic effects of host features and microbiota-derived risk score on survival in stage I-III CRC patients. Related to main Figure 5.

Supplementary Fig. 7 Validation of the prognostic performance of microbiota-derived risk score based on the 16S amplicon data of tumors and paired NATs in the AC-ICAM cohort. Related to main Figure 5.

Supplementary Fig. 8 The differential expression patterns of pro-inflammation genes between tumors from UU, UM and AC-ICAM with high and low MRS-T values. Related to main Figure 5.


a
Step 1:
Data extraction and quality control
Sequencing data
Non-human aligned paired reads
1. Fastp
2. Bowtie2 against GRCh38
High-quality input reads
UU
UM
AC-ICAM
Variable
NAT
Tumor
Blood
NAT
Tumor
Tumor
313 775 462 162 162 167 N
1.70+4.47 1.63+6.99 0.49+2.13 1.38+2.28 0.98+1.89 1.74+7.71 High-quality input reads (×M)
1.47+4.07 1.3+5.35 0.2+0.68 1.19+2.05 0.81+1.62 1.48+7.02 Prokaryotic reads (×M)
1.28+3.59 1.14+4.69 0.02+0.27 1.04+1.81 0.7+1.4 1.27+5.64 Reads annotated to tissue-resident taxa (×M)
WGS
95 775 0 8 162 0 N
4.64+3.03 2.64+7.88 1.28+0.68 1.25+1.08 High-quality input reads (×M)
2.39+3.24 1.7+7.57 0.49+0.66 0.48+1.05 Prokaryotic reads (×M)
1.49+2.08 1.16+5.67 0.31+0.44 0.32+0.75 Reads annotated to tissue-resident taxa (×M)
WTS
Step 2:
Microbial detection
(Removal of human reads by Kraken2 against GRCh38 and T2T-CHM13v2.0)
4,644 UHGG genomes + 17 Treponema genomes NCBI
1268 genera
1266 genera
1268 genera
Step 3:
Removal of contaminants (1)
Prevalence: Blood > Tumors
(UU: 462 tumors vs. paired blood samples)
Removal of 114 genera
1154 genera
1152 genera
1154 genera
Step 4:
Removal of contaminants (2)
Black list of putative contaminants (genera)*
Removal of 24 genera
1130 genera
1128 genera
1130 genera
Step 5:
Removal of taxa with low-prevalence
Prevalence (RA ≥ 0.0001) < 20% tissues
335 genera
325 genera
shared
249 common genera
249 common genera
Step 6:
Removal of species with low-coverage
WGS data:
Coverage ≥ 5% in < 5% UU tissues (N=1,088)
361 tissue-resident species
WGS
WTS
Input non-human reads
Reads number
UU tissues (n=1088)
UM tissues (n=324)
AC-ICAM tissues (n=167)
UU blood (n=462)
Proportion of reads discarded
Proportion of reads remaining
UU tissues (n=870)
UM tissues (n=170)
Kraken2-mapped prokaryotic reads
Reads Number
Propotion of reads
75% 75% 71% 26%
88% 83%
Reads of non-contaminants
95% 96% 97% 35%
94% 92%
Reads annotated to tissue-resident taxa
92% 92% 94% 32%
90% 87%
UU UM AC-ICAM UU
Tissue
Blood
UU UM
Tissue


b
Read number before and after contaminant filtering (462 blood samples)
Reads number
Before
After
c
CAP2
CAP1
Tumor (UU=775 vs.UM=162)
D-MANOVA P=1.39e-26
NAT (UU=313 vs. UM=162)
D-MANOVA P=4.59e-20
UU
UM

d
Kraken2−mapped prokaryotic reads
5.4e−07
Reads annotated to tissue−resident taxa
0.14
WGS WTS
UU-tumors
n = 775
e
Identified tissue−resident species
2.9e−13
f
Mean coverage of genomes
P < 2.2e-16
Genome coverage of tissue-resident species
Mean Kraken−mapped reads
s__Bacteroides.dorei
s__Bacteroides.uniformis
s__Escherichia.coli
s__Bacteroides.fragilis
s__Fn.animalis.C2
s__Faecalibacterium.prausnitzii
s__Alistipes.onderdonkii
s__Prevotella.nigrescens
s__Alloprevotella.tannerae
s__Clostridium.clostridioforme
g
Proportion of reads annotated to rRNA
rRNA 40%
Samples (UU-tumors, n=775)

Supplementary Fig. 1

The analysis pipeline of taxonomic identification and quantification of tissue-resident microbes.

a. Analytical pipeline for taxonomic identification and quantification of tissue-resident microbes. A two-step host read removal strategy was applied to enhance data quality and minimize human read contamination (Bowtie2 against GRCh38, and Kraken2 against GRCh38 and T2T-CHM13v2.0). A high-quality reference database (4, 644 species from the UHGG, and 17 *Treponema* species from NCBI) were used for microbial detection. Microbial contaminants were filtered by comparing WGS data from 462 paired UU blood (positive controls) and tumor samples. Box plots display input reads, while bar plots show proportions of discarded reads in each step. Please see **Methods** section for details on quality control and non-human reads extraction, microbial detection, removal of putative microbial contaminants and rare taxa with low prevalence and genome coverage.

b. Distribution of mapped prokaryotic reads before and after contaminant filtering in UU blood samples (n=462).

c. Distance-based redundancy analysis (db-RDA) shows a clear separation of tissue-resident microbiota for tumors and NATs between the UU and UM WGS cohorts. D-MANOVA P-values were presented.

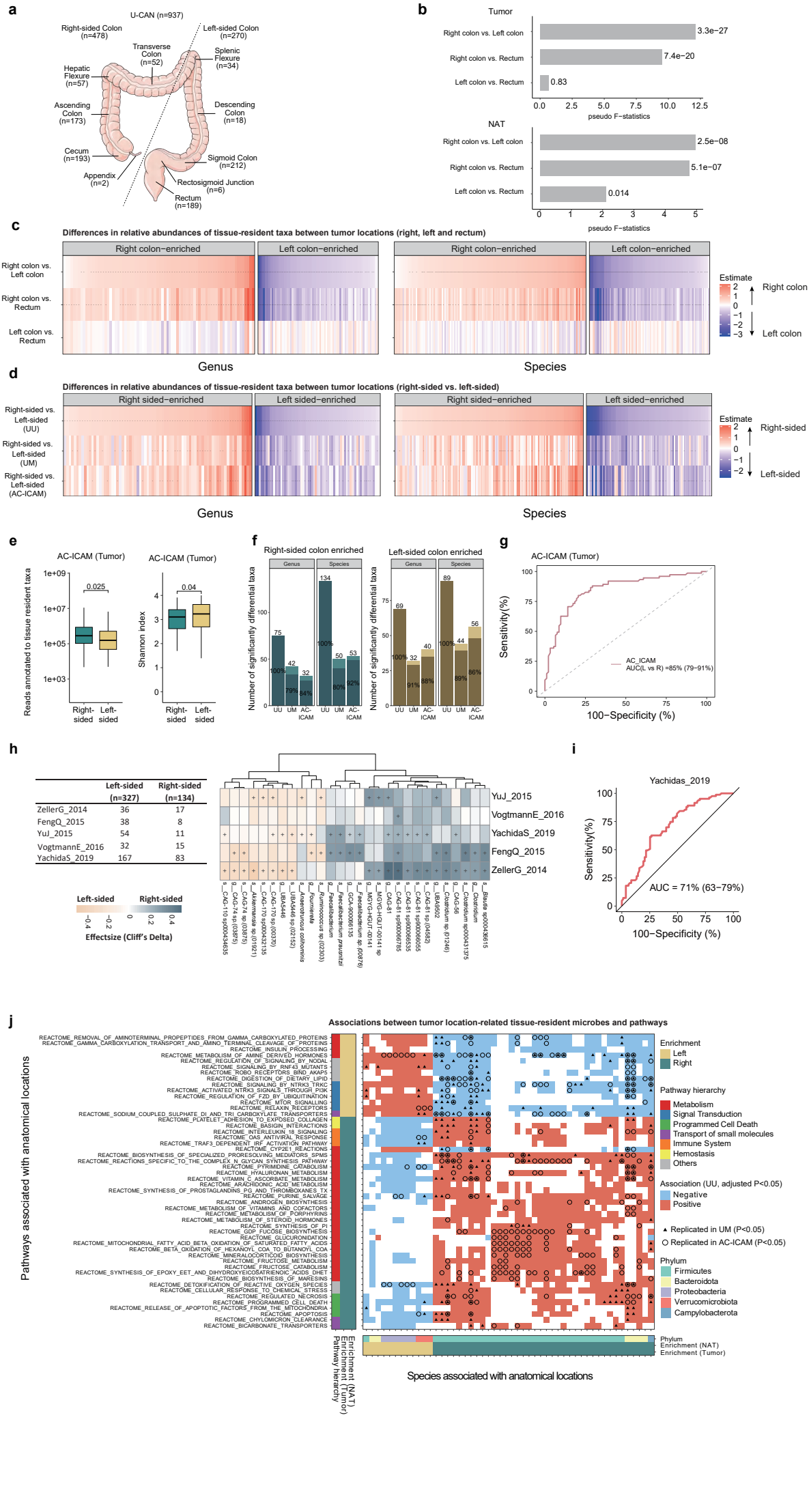
d. Number of prokaryotic reads (left), and read counts annotated to tissue-resident microbes (right) between paired WGS and WTS samples from 775 UU tumors.

e. Number of detected tissue-resident species per sample between paired WGS and WTS data from 775 UU tumors.

f. Left: Average genome coverage of tissue-resident species between paired WGS and WTS samples. Right: Average genome coverage and mapped reads number per species between paired WGS and WTS data from 775 UU tumors. The name of the top 10 species by Kraken2-mapped reads across WGS samples are shown.

g. Proportion of prokaryotic rRNA reads to the total number of prokaryotic reads across WTS data. The average percentage of rRNA reads is shown in the pie chart.

For **d**, **e**, and **f**, P-values were calculated using two-sided Wilcoxon signed-rank test. For the boxplots, boxes represent the interquartile ranges (IQRs) between the first and third quartiles, the center line represents the median, and the whiskers extend 1.5 times the IQR from the top and bottom of the box.


a
U-CAN (n=937)
Right-sided Colon (n=478)
Left-sided Colon (n=270)
Transverse Colon (n=52)
Splenic Flexure (n=34)
Hepatic Flexure (n=57)
Ascending Colon (n=173)
Descending Colon (n=18)
Cecum (n=193)
Sigmoid Colon (n=212)
Appendix (n=2)
Rectosigmoid Junction (n=6)
Rectum (n=189)
b
Tumor
Right colon vs. Left colon
Right colon vs. Rectum
Left colon vs. Rectum
3.3e−27
7.4e−20
0.83
pseudo F−statistics
NAT
2.5e−08
5.1e−07
0.014
c
Differences in relative abundances of tissue-resident taxa between tumor locations (right, left and rectum)
Right colon−enriched
Left colon−enriched
Estimate
Right colon
Left colon
Genus
Species
d
Differences in relative abundances of tissue-resident taxa between tumor locations (right-sided vs. left-sided)
Right sided−enriched
Left sided−enriched
Right-sided vs. Left-sided (UU)
Right-sided vs. Left-sided (UM)
Right-sided vs. Left-sided (AC-ICAM)
Right-sided
Left-sided
e
AC-ICAM (Tumor)
Reads anndated to tissue resident taxa
Shannon index
0.025
0.04
Right-sided
Left-sided
f
Right-sided colon enriched
Left-sided colon enriched
Genus
Species
Number of significantly differential taxa
UU
UM
AC-ICAM
g
AC-ICAM (Tumor)
Sensitivity(%)
100−Specificity (%)
AC_ICAM
AUC(L vs R) =85% (79−91%)
h
Left-sided (n=327)
Right-sided (n=134)
ZellerG_2014 36 17
FengQ_2015 38 8
YuJ_2015 54 11
VogtmannE_2016 32 15
YachidaS_2019 167 83
Left-sided
Right-sided
Effectsize (Cliff's Delta)
YuJ_2015
VogtmannE_2016
YachidaS_2019
FengQ_2015
ZellerG_2014
i
Yachidas_2019
Sensitivity(%)
AUC = 71% (63−79%)
100−Specificity (%)
j
Associations between tumor location-related tissue-resident microbes and pathways
Pathways associated with anatomical locations
REACTOME_REMOVAL_OF_AMINOTERMINAL_PROPEPTIDES_FROM_GAMMA_CARBOXYLATED_PROTEINS
REACTOME_GAMMA_CARBOXYLATION_TRANSPORT_AND_AMINO_TERMINAL_CLEAVAGE_OF_PROTEINS
REACTOME_INSULIN_PROCESSING
REACTOME_METABOLISM_OF_AMINE_DERIVED_HORMONES
REACTOME_REGULATION_OF_SIGNALING_BY_NODAL
REACTOME_SIGNALING_BY_RNF43_MUTANTS
REACTOME_ROBO_RECEPTORS_BIND_AKAP5
REACTOME_DIGESTION_OF_DIETARY_LIPID
REACTOME_SIGNALING_BY_NTRK3_TRKC
REACTOME_ACTIVATED_NTRK3_SIGNALS_THROUGH_PI3K
REACTOME_REGULATION_OF_FZD_BY_UBIQUITINATION
REACTOME_MTOR_SIGNALLING
REACTOME_RELAXIN_RECEPTORS
REACTOME_SODIUM_COUPLED_SULPHATE_DI_AND_TRI_CARBOXYLATE_TRANSPORTERS
REACTOME_PCATELET_ADHESION_TO_EXPOSED_COLLAGEN
REACTOME_BASIGIN_INTERACTIONS
REACTOME_INTERLEUKIN_18_SIGNALING
REACTOME_OAS_ANTIVIRAL_RESPONSE
REACTOME_TRAF3_DEPENDENT_IRF_ACTIVATION_PATHWAY
REACTOME_CYP2E1_REACTIONS
REACTOME_BIOSYNTHESIS_OF_SPECIALIZED_PRORESOLVING_MEDIATORS_SPMS
REACTOME_REACTIONS_SPECIFIC_TO_THE_COMPLEX_N_GLYCAN_SYNTHESIS_PATHWAY
REACTOME_PYRIMIDINE_CATABOLISM
REACTOME_HYALURONAN_METABOLISM
REACTOME_VITAMIN_C_ASCORBATE_METABOLISM
REACTOME_ARACHIDONIC_ACID_METABOLISM
REACTOME_SYNTHESIS_OF_PROSTAGLANDINS_PG_AND_THROMBOXANES_TX
REACTOME_PURINE_SALVAGE
REACTOME_ANDROGEN_BIOSYNTHESIS
REACTOME_METABOLISM_OF_VITAMINS_AND_COFACTORS
REACTOME_METABOLISM_OF_PORPHYRINS
REACTOME_METABOLISM_OF_STEROID_HORMONES
REACTOME_SYNTHESIS_OF_PI
REACTOME_GDP_FUCOSE_BIOSYNTHESIS
REACTOME_GLUCURONIDATION
REACTOME_MITOCHONDRIAL_FATTY_ACID_BETA_OXIDATION_OF_SATURATED_FATTY_ACIDS
REACTOME_BETA_OXIDATION_OF_HEXANOYL_COA_TO_BUTANOYL_COA
REACTOME_MINERALOCORTICOID_BIOSYNTHESIS
REACTOME_FRUCTOSE_METABOLISM
REACTOME_FRUCTOSE_CATABOLISM
REACTOME_SYNTHESIS_OF_EPOXY_EET_AND_DIHYDROXYEICOSATRIENOIC_ACIDS_DHET
REACTOME_BIOSYNTHESIS_OF_MARESINS
REACTOME_DETOXIFICATION_OF_REACTIVE_OXYGEN_SPECIES
REACTOME_CELLULAR_RESPONSE_TO_CHEMICAL_STRESS
REACTOME_REGULATED_NECROSIS
REACTOME_PROGRAMMED_CELL_DEATH
REACTOME_RELEASE_OF_APOPTOTIC_FACTORS_FROM_THE_MITOCHONDRIA
REACTOME_APOPTOSIS
REACTOME_CHYLOMICRON_CLEARANCE
REACTOME_BICARBONATE_TRANSPORTERS
Enrichment (NAT)
Enrichment (Tumor)
Pathway hierarchy
Enrichment
Left
Right
Pathway hierarchy
Metabolism
Signal Transduction
Programmed Cell Death
Transport of small molecules
Immune System
Hemostasis
Others
Association (UU, adjusted P<0.05)
Negative
Positive
Replicated in UM (P<0.05)
Replicated in AC-ICAM (P<0.05)
Phylum
Firmicutes
Bacteroidota
Proteobacteria
Verrucomicrobiota
Campylobacterota
Phylum
Enrichment (NAT)
Enrichment (Tumor)
Species associated with anatomical locations

Supplementary Fig. 2 Assessment of tissue microbial composition across different anatomical locations of colon and rectum.

a. Schematic diagram of the ten sampling sites of U-CAN patients with primary CRC. Samples were classified right-sided colorectal cancers or left-sided colorectal cancers based on the tumor location at diagnosis. Right-sided colon: caecum, ascending, hepatic flexure and transverse colon. Left-sided colon: splenic flexure, descending colon, sigmoid and rectosigmoid junction, and rectum (distal 15 cm from the anal verge). Illustration elements were adapted from Servier Medical Art (https://smart.servier.com), licensed under CC BY 3.0.

b. Differences in overall microbial composition between anatomical locations in tumors (n=775) and NATs (n=313) from the UU cohort, assessed by PERMANOVA. BH-adjusted D-MANOVA P-values were presented (two-sided).

c. Differences in relative abundance of tissue-resident taxa between tumor locations in UU tumors, assessed by MaAsLin2. *, BH-adjusted P< 0.05 (two-sided test). Each column represents a single taxon. Red indicates right colon-enriched taxa, and blue indicates left and rectum-enriched taxa.

d. Heatmap showing location-based enrichment patterns of tissue-resident genera and species in the tumors from UU, UM and AC-ICAM cohorts, assessed by MaAsLin2. *, BH-adjusted P-value< 0.05, indicated by "*".

e. Differences in number of microbial reads (left) and genus-level microbial Shannon index (right) between right- and left-sided tumors in AC-ICAM cohort (n=167), assessed by MaAsLin2.

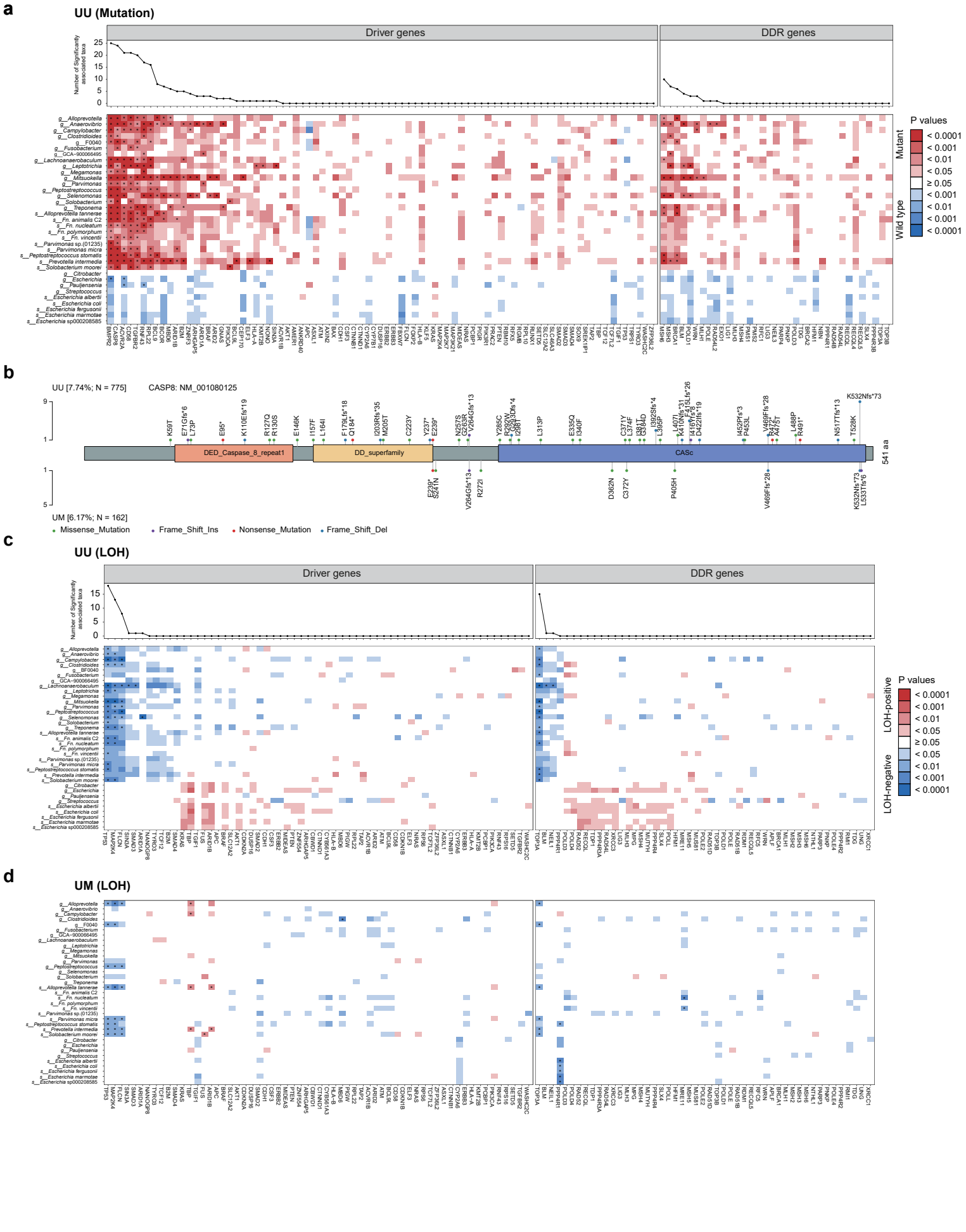
f. Number of significantly differential abundant taxa between right- and left-sided tumors in the UU (BH-adjusted P<0.05), UM (P<0.05) and AC-ICAM (P<0.05) cohorts (MaAsLin2, two-sided test).

g. Performance of the random forest model in distinguishing between right- (n=92) and left-sided (n=75) tumors in the AC-ICAM cohort. A two-sided 95% confidence interval for AUC is shown.

h. Validation of location-specific taxa in fecal metagenomic samples of CRC patients (n=461) from five publicly available studies. Abundance differences was assessed using the two-sided Wilcoxon rank-sum test (P<0.05). 28 taxa meeting the threshold in at least two cohorts are shown.

i. Performance of a separate, five-fold cross-validation-based RF model for distinguishing right- (n=83) from left-sided (n=167) samples in the Yachidas_2019 dataset. Other fecal datasets have been excluded because of small sample sizes with annotated tumor locations (n≤20 per group).

j. Discovery and validation of associations between tumor location-related tissue-resident microbes and pathways assessed by partial Spearman's correlation (two-sided). Red represents positive, and blue represents negative associations (BH-adjusted P< 0.05 for UU and P < 0.05 for UM and AC-ICAM cohorts).


a
UU (Mutation)
Driver genes
DDR genes
Number of Significantly associated taxa
P values
Mutant
Wild type
< 0.0001
< 0.001
< 0.01
< 0.05
≥ 0.05
b
UU [7.74%; N = 775]
CASP8: NM_001080125
DED_Caspase_8_repeat1
DD_superfamily
CASc
541 aa
UM [6.17%; N = 162]
Missense_Mutation
Frame_Shift_Ins
Nonsense_Mutation
Frame_Shift_Del
c
UU (LOH)
Driver genes
DDR genes
LOH-positive
LOH-negative
d
UM (LOH)

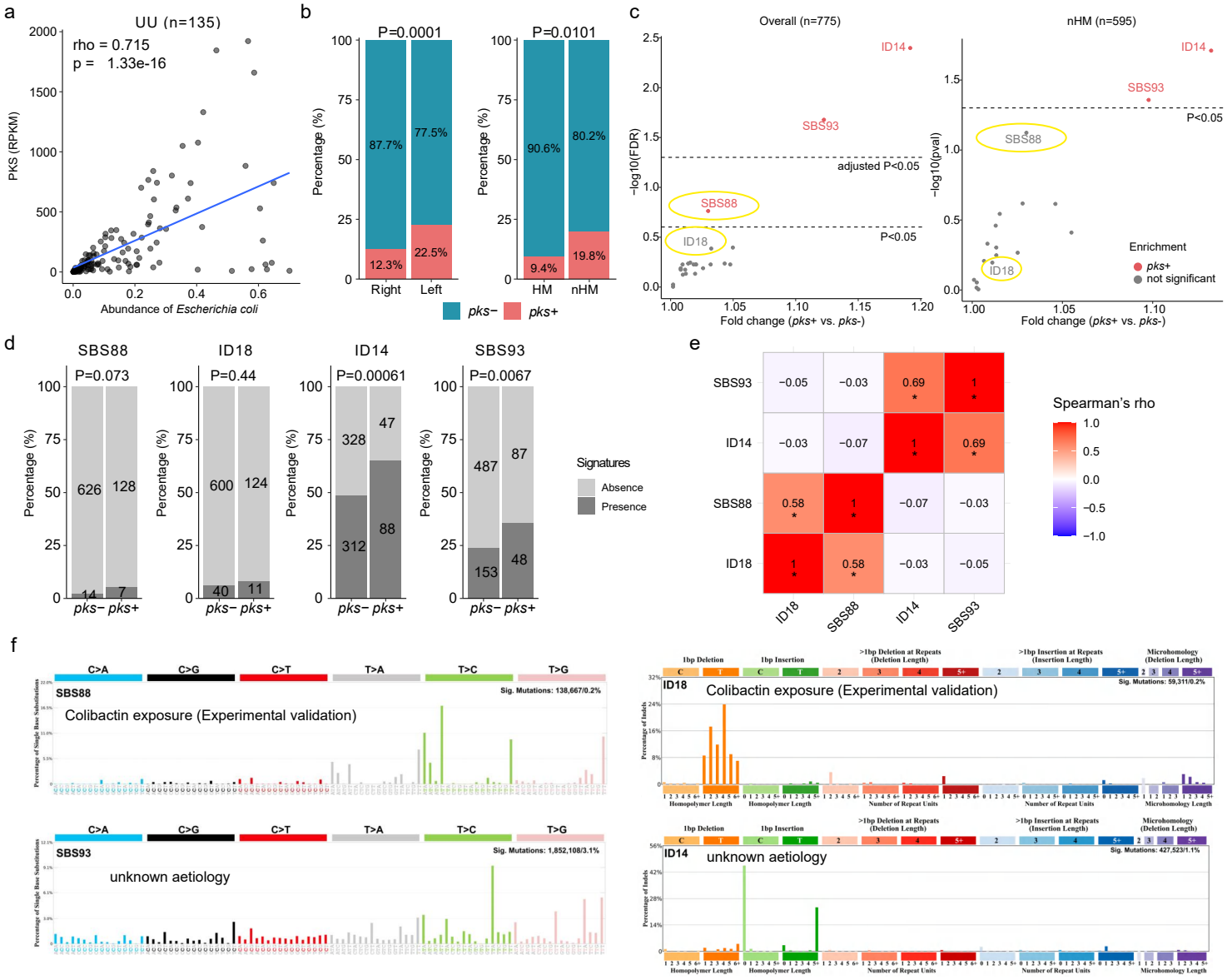
Supplementary Fig. 3 Associations between somatic mutations and tumor-enriched taxa

a. The relationships between gene mutations and 35 tumor-enriched taxa in UU (n=775) cohort. Heatmap shows relationships between taxon abundances and mutant or wild-type (WT) samples for individual drivers (n=83) or DNA damage response and repair (DDR, n=35) genes, assessed by MaAsLin2 (*, BH-adjusted P< 0.05, two-sided test). The top curve shows the total number of taxa that are significantly different between mutant and WT samples by gene. Red represents enrichment in mutant samples, and blue represents enrichment in WT samples.

b. Lollipop plot produced with Maftools displaying *CASP8* somatic coding mutations in the UU and UM cohorts. The length of the lollipop represents the number of mutations detected at each codon position. Lollipops are colored according to mutation type: missense mutations in green, nonsense mutations in red, and frameshift insert/deletion mutations in purple and blue, respectively. The colored boxes represent specific functional domains of CASP8: DED: death-effector-domain, DD: death domain, CAS: catalytic domain.

c. The discovery of relationships between loss of heterozygosity (LOH) events and tumor-enriched microbes in UU cohort (n=775). Heatmap shows associations between taxon abundances and LOH-positive or LOH-negative samples for individual drivers (n=62) or DDR (n=49) genes, assessed by MaAsLin2 (*, BH-adjusted P< 0.05, two-sided test). The top curves show the total number of taxa that are significantly different between LOH-positive and LOH-negative samples by gene. Red represents enrichment in LOH-positive samples, and blue represents enrichment in LOH-negative samples.

d. The validation of relationships between LOH events and tumor-enriched microbes in UM cohort (n=162). A cutoff of P<0.05 was used.


a
UU (n=135)
rho = 0.715
p = 1.33e-16
PKS (RPKM)
Abundance of Escherichia coli
b
P=0.0001
P=0.0101
Percentage (%)
87.7%
77.5%
12.3%
22.5%
90.6%
80.2%
9.4%
19.8%
Right
Left
HM
nHM
pks−
pks+
c
Overall (n=775)
nHM (n=595)
ID14
SBS93
SBS88
ID18
adjusted P<0.05
P<0.05
−log10(FDR)
−log10(pval)
Fold change (pks+ vs. pks-)
Enrichment
pks+
not significant
d
SBS88
ID18
ID14
SBS93
P=0.073
P=0.44
P=0.00061
P=0.0067
626
128
14
7
600
124
40
11
328
47
312
88
487
87
153
48
Signatures
Absence
Presence
e
SBS93
ID14
SBS88
ID18
−0.05
−0.03
0.69
1
−0.03
−0.07
1
0.69
0.58
1
−0.07
−0.03
1
0.58
−0.03
−0.05
Spearman's rho
f
C>A
C>G
C>T
T>A
T>C
T>G
SBS88
Sig. Mutations: 138,667/0.2%
Colibactin exposure (Experimental validation)
SBS93
Sig. Mutations: 1,852,108/3.1%
unknown aetiology
ID18
Sig. Mutations: 59,311/0.2%
Colibactin exposure (Experimental validation)
ID14
Sig. Mutations: 427,523/1.1%
unknown aetiology
1bp Deletion
1bp Insertion
>1bp Deletion at Repeats (Deletion Length)
>1bp Insertion at Repeats (Insertion Length)
Microhomology (Deletion Length)
Homopolymer Length
Number of Repeat Units
Microhomology Length

Supplementary Fig. 4 Prevalence of polyketide synthase (*pks*) gene island and its relationship with associated mutational signatures

a. Spearman's rank correlation (two-sided) between the relative abundances of *E. coli* and pks gene island (RPKM) across 135 pks-positive (*pks*+) samples from UU cohort.

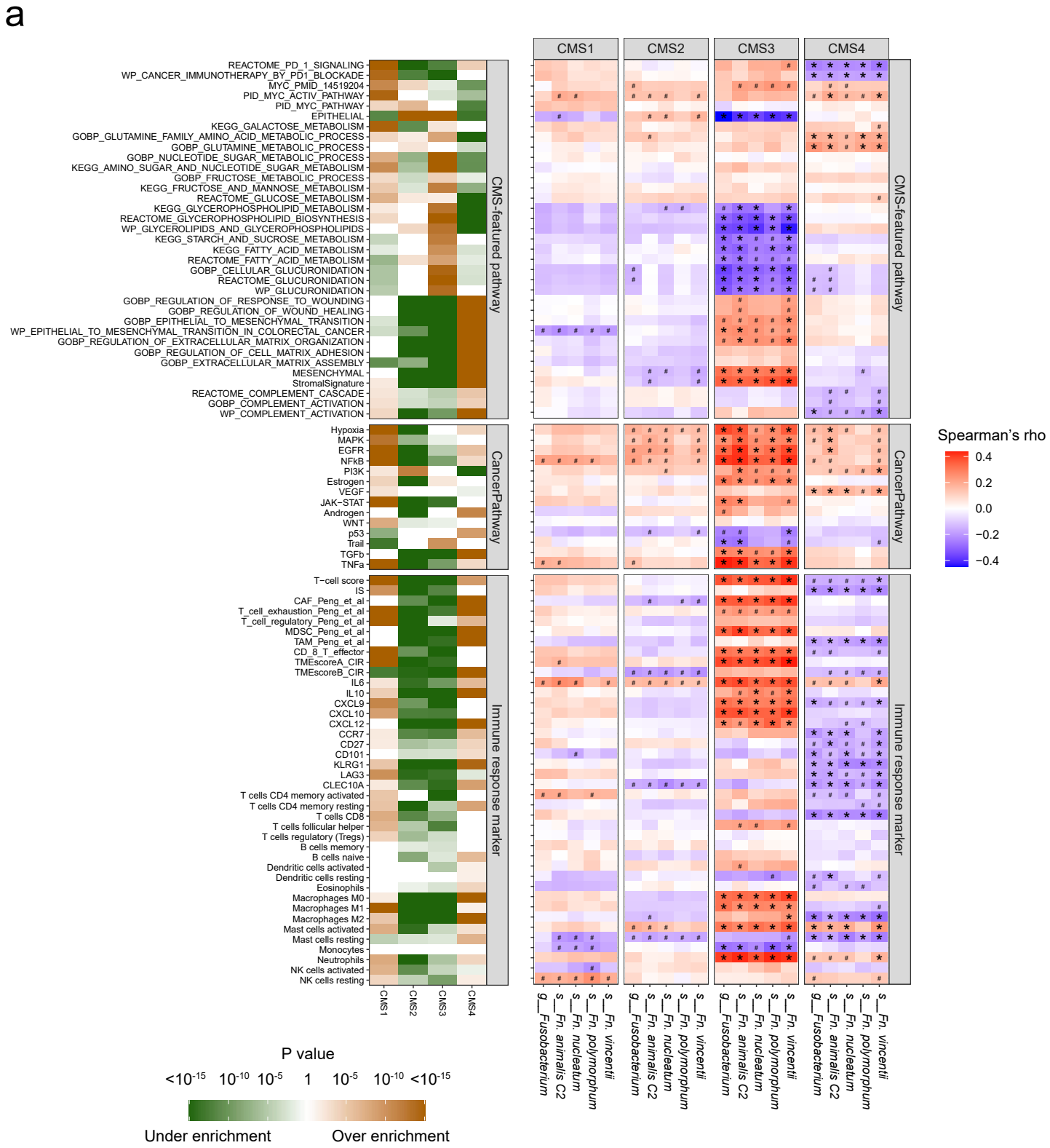
b. Percentages of pks-positive (*pks*+) and pks-negative (*pks*-) tumors in right- and left-sided, and hypermutated (HM) and non-hypermutated (nHM) tumors from the UU cohort. P values were calculated by two-sided Chi-squared test.

c. Mutational signatures enriched in *pks*+ tumors compared to *pks*- tumors in the UU cohort (n=775) and individuals with nHM tumors (n=595), assessed by two-sided partial Spearman's rank correlation, with adjustment for age, sex, tumor location, and prokaryotic reads count. SBS, single base substitution. DBS, doublet base substitution. ID, small insertions and deletions.

d. Percentages of samples carrying mutational signature SBS88, SBS93, ID18 or ID14 in tumors with *pks*+ and *pks*- in the UU cohort. P-values were calculated with two-sided Fisher test.

e. Spearman's rank correlation (two-sided) between the mutational signature activities of the four selected signatures. *, BH-adjusted $P<0.05$.

f. Profiles of the four *pks*+ associated mutational signatures. Proposed aetiology of each signature was taken from the COSMIC database.


a
CMS1
CMS2
CMS3
CMS4
CMS-featured pathway
CancerPathway
Immune response marker
REACTOME_PD_1_SIGNALING
WP_CANCER_IMMUNOTHERAPY_BY_PD1_BLOCKADE
MYC_PMID_14519204
PID_MYC_ACTIV_PATHWAY
PID_MYC_PATHWAY
EPITHELIAL
KEGG_GALACTOSE_METABOLISM
GOBP_GLUTAMINE_FAMILY_AMINO_ACID_METABOLIC_PROCESS
GOBP_GLUTAMINE_METABOLIC_PROCESS
GOBP_NUCLEOTIDE_SUGAR_METABOLIC_PROCESS
KEGG_AMINO_SUGAR_AND_NUCLEOTIDE_SUGAR_METABOLISM
GOBP_FRUCTOSE_METABOLIC_PROCESS
KEGG_FRUCTOSE_AND_MANNOSE_METABOLISM
REACTOME_GLUCOSE_METABOLISM
KEGG_GLYCEROPHOSPHOLIPID_METABOLISM
REACTOME_GLYCEROPHOSPHOLIPID_BIOSYNTHESIS
WP_GLYCEROLIPIDS_AND_GLYCEROPHOSPHOLIPIDS
KEGG_STARCH_AND_SUCROSE_METABOLISM
KEGG_FATTY_ACID_METABOLISM
REACTOME_FATTY_ACID_METABOLISM
GOBP_CELLULAR_GLUCURONIDATION
REACTOME_GLUCURONIDATION
WP_GLUCURONIDATION
GOBP_REGULATION_OF_RESPONSE_TO_WOUNDING
GOBP_REGULATION_OF_WOUND_HEALING
GOBP_EPITHELIAL_TO_MESENCHYMAL_TRANSITION
WP_EPITHELIAL_TO_MESENCHYMAL_TRANSITION_IN_COLORECTAL_CANCER
GOBP_REGULATION_OF_EXTRACELLULAR_MATRIX_ORGANIZATION
GOBP_REGULATION_OF_CELL_MATRIX_ADHESION
GOBP_EXTRACELLULAR_MATRIX_ASSEMBLY
MESENCHYMAL
StromalSignature
REACTOME_COMPLEMENT_CASCADE
GOBP_COMPLEMENT_ACTIVATION
WP_COMPLEMENT_ACTIVATION
Hypoxia
MAPK
EGFR
NFkB
PI3K
Estrogen
VEGF
JAK−STAT
Androgen
WNT
p53
Trail
TGFb
TNFa
T−cell score
IS
CAF_Peng_et_al
T_cell_exhaustion_Peng_et_al
T_cell_regulatory_Peng_et_al
MDSC_Peng_et_al
TAM_Peng_et_al
CD_8_T_effector
TMEscoreA_CIR
TMEscoreB_CIR
IL6
IL10
CXCL9
CXCL10
CXCL12
CCR7
CD27
CD101
KLRG1
LAG3
CLEC10A
T cells CD4 memory activated
T cells CD4 memory resting
T cells CD8
T cells follicular helper
T cells regulatory (Tregs)
B cells memory
B cells naive
Dendritic cells activated
Dendritic cells resting
Eosinophils
Macrophages M0
Macrophages M1
Macrophages M2
Mast cells activated
Mast cells resting
Monocytes
Neutrophils
NK cells activated
NK cells resting
g__Fusobacterium
s__Fn. animalis C2
s__Fn. nucleatum
s__Fn. polymorphum
s__Fn. vincentii
Spearman's rho
0.4
0.2
0.0
−0.2
−0.4
P value
<10-15 10-10 10-5 1 10-5 10-10 <10-15
Under enrichment
Over enrichment
b
SBS93
1.00
0.75
0.50
0.25
0.00
0.017
0.0002
1.2e−14
CMS1
CMS2
CMS3
CMS4
ID14
1.0
0.5
0.0
0.76
3.1e−05
2.22e−16

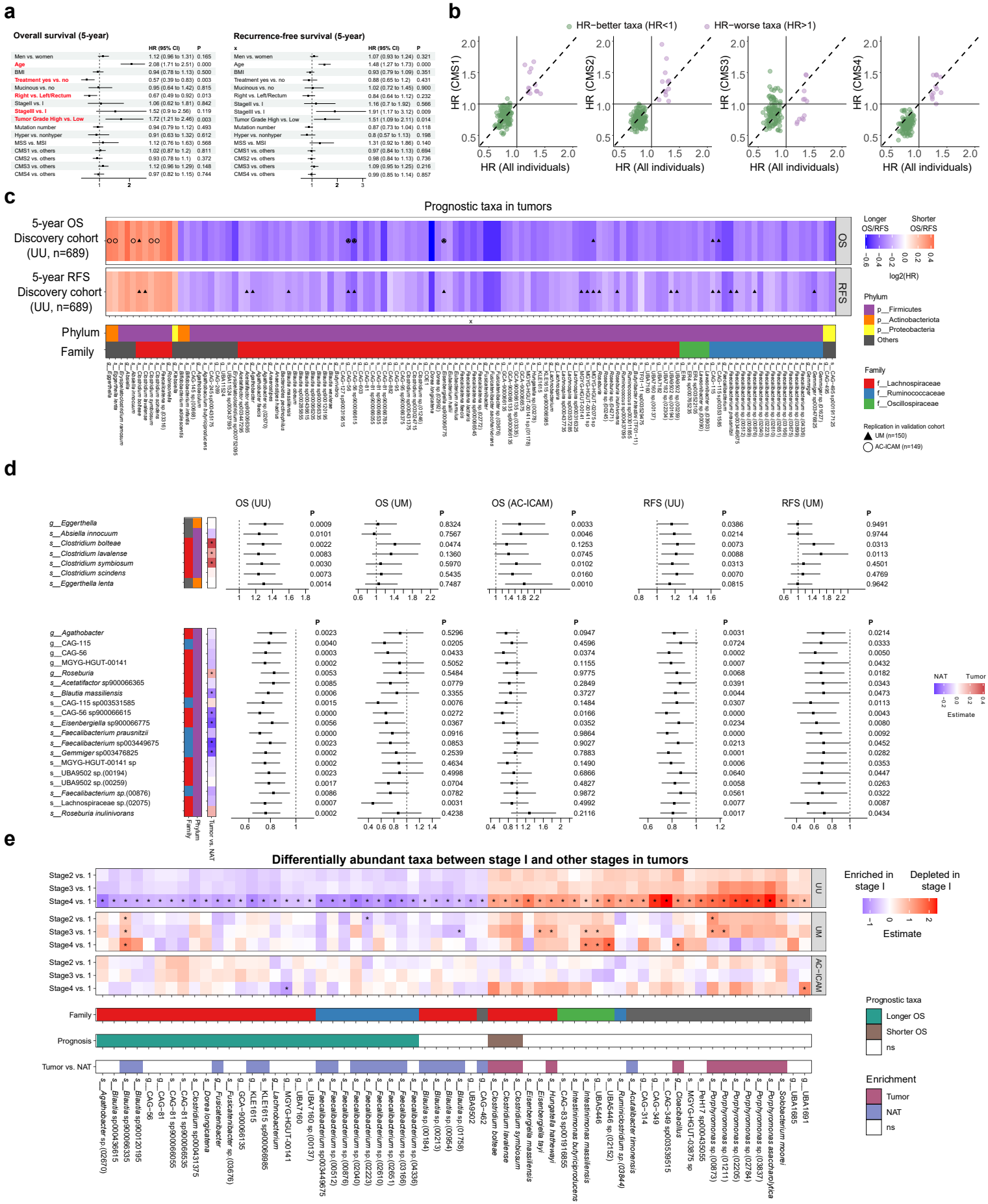
Supplementary Fig. 5

CMS-dependent associations between prognostic taxa and tumor transcriptional profiles of CRC related pathways and immune response markers

a. Heatmap (left panel) shows the differences in expression levels of 14 CRC-related pathways, CMS featured pathways and immune response markers comparing each CMS with the others. P-values were calculated with two-sided t-test.

Heatmap (right panel) shows the associations between the relative abundances of *Fusobacterium* spp. and the expression levels of expression levels of all selected CRC-related pathways, genes and immune response markers within CMS subgroups, assessed by two-sided partial Spearman's correlation with adjustment for age, sex, tumor location, and prokaryotic reads count. #, $P < 0.05$. *, BH-adjusted $P<0.05$.

b. Distribution of two *pks*+ associated mutational signatures across CMS subtypes. P values between CMS2 and other subtypes were generated by two-sided Wilcoxon rank-sum test. SBS, single base substitution. ID, small insertions and deletions.


a
Overall survival (5-year)
HR (95% CI) P
Men vs. women 1.12 (0.96 to 1.31) 0.165
Age 2.08 (1.71 to 2.51) 0.000
BMI 0.94 (0.78 to 1.13) 0.500
Treatment yes vs. no 0.57 (0.39 to 0.83) 0.003
Mucinous vs. no 0.95 (0.64 to 1.42) 0.815
Right vs. Left/Rectum 0.67 (0.49 to 0.92) 0.013
StageII vs. I 1.06 (0.62 to 1.81) 0.842
StageIII vs. I 1.52 (0.9 to 2.56) 0.119
Tumor Grade High vs. Low 1.72 (1.21 to 2.46) 0.003
Mutation number 0.94 (0.79 to 1.12) 0.493
Hyper vs. nonhyper 0.91 (0.63 to 1.32) 0.612
MSS vs. MSI 1.12 (0.76 to 1.63) 0.568
CMS1 vs. others 1.02 (0.87 to 1.2) 0.811
CMS2 vs. others 0.93 (0.78 to 1.1) 0.372
CMS3 vs. others 1.12 (0.96 to 1.29) 0.148
CMS4 vs. others 0.97 (0.82 to 1.15) 0.744
Recurrence-free survival (5-year)
x HR (95% CI) P
Men vs. women 1.07 (0.93 to 1.24) 0.321
Age 1.48 (1.27 to 1.73) 0.000
BMI 0.93 (0.79 to 1.09) 0.351
Treatment yes vs. no 0.88 (0.65 to 1.2) 0.431
Mucinous vs. no 1.02 (0.72 to 1.45) 0.900
Right vs. Left/Rectum 0.84 (0.64 to 1.12) 0.232
StageII vs. I 1.16 (0.7 to 1.92) 0.566
StageIII vs. I 1.91 (1.17 to 3.12) 0.009
Tumor Grade High vs. Low 1.51 (1.09 to 2.11) 0.014
Mutation number 0.87 (0.73 to 1.04) 0.118
Hyper vs. nonhyper 0.8 (0.57 to 1.13) 0.198
MSS vs. MSI 1.31 (0.92 to 1.86) 0.140
CMS1 vs. others 0.97 (0.84 to 1.13) 0.694
CMS2 vs. others 0.98 (0.84 to 1.13) 0.736
CMS3 vs. others 1.09 (0.95 to 1.25) 0.216
CMS4 vs. others 0.99 (0.85 to 1.14) 0.857
b
HR−better taxa (HR<1)
HR−worse taxa (HR>1)
HR (CMS1)
HR (CMS2)
HR (CMS3)
HR (CMS4)
HR (All individuals)
c
Prognostic taxa in tumors
5-year OS Discovery cohort (UU, n=689)
5-year RFS Discovery cohort (UU, n=689)
OS
RFS
Phylum
Family
Longer OS/RFS
Shorter OS/RFS
log2(HR)
p__Firmicutes
p__Actinobacteriota
p__Proteobacteria
Others
f__Lachnospiraceae
f__Ruminococcaceae
f__Oscillospiraceae
Replication in validation cohort
UM (n=150)
AC-ICAM (n=149)
d
OS (UU)
OS (UM)
OS (AC-ICAM)
RFS (UU)
RFS (UM)
g__Eggerthella
s__Absiella innocuum
s__Clostridium boltleae
s__Clostridium lavalense
s__Clostridium symbiosum
s__Clostridium scindens
s__Eggerthella lenta
g__Agathobacter
g__CAG-115
g__CAG-56
g__MGYG-HGUT-00141
g__Roseburia
s__Acetatifactor sp900066365
s__Blautia massiliensis
s__CAG-115 sp003531585
s__CAG-56 sp900066615
s__Eisenbergiella sp900066775
s__Faecalibacterium prausnitzii
s__Faecalibacterium sp003449675
s__Gemmiger sp003476825
s__MGYG-HGUT-00141 sp
s__UBA9502 sp.(00194)
s__UBA9502 sp.(00259)
s__Faecalibacterium sp.(00876)
s__Lachnospiraceae sp.(02075)
s__Roseburia inulinivorans
Phylum
Family
Tumor vs. NAT
NAT
Tumor
Estimate
e
Differentially abundant taxa between stage I and other stages in tumors
Stage2 vs. 1
Stage3 vs. 1
Stage4 vs. 1
UU
UM
AC-ICAM
Enriched in stage I
Depleted in stage I
Estimate
Family
Prognosis
Tumor vs. NAT
Prognostic taxa
Longer OS
Shorter OS
ns
Enrichment
Tumor
NAT
ns
s__Agathobacter sp.(02670)
s__Blautia sp900066335
s__Blautia sp900066615
s__Blautia sp900120195
s__CAG-56
g__CAG-81
s__CAG-81 sp900066055
s__CAG-81 sp900066535
s__Clostridium sp900431375
s__Dorea longicatena
g__Fusicatenibacter
s__Fusicatenibacter sp.(03676)
g__GCA-900066135
g__KLE1615
s__KLE1615 sp900066985
g__Lachnoclostridium
g__MGYG-HGUT-00141
g__UBA7160
g__UBA7160 sp.(00137)
s__Faecalibacterium sp003449675
s__Faecalibacterium sp.(00512)
s__Faecalibacterium sp.(00876)
s__Faecalibacterium sp.(02040)
s__Faecalibacterium sp.(02223)
s__Faecalibacterium sp.(02610)
s__Faecalibacterium sp.(02651)
s__Faecalibacterium sp.(03166)
s__Faecalibacterium sp.(04336)
s__Blautia sp.(00184)
s__Blautia sp.(00213)
s__Blautia sp.(00954)
s__Blautia sp.(01758)
g__UBA9502
g__CAG-462
s__Clostridium boltleae
s__Clostridium lavalense
s__Clostridium symbiosum
s__Eisenbergiella tayi
s__Eisenbergiella massiliensis
s__Hungatella hathewayi
s__CAG-83 sp001916855
s__Intestinimonas butyriciproducens
s__Intestinimonas massiliensis
g__UBA5446
s__UBA5446 sp.(02152)
s__Ruminiclostridium sp.(03844)
s__Acutalibacter timonensis
g__CAG-314
g__CAG-349
s__CAG-349 sp003539515
g__Cloacibacillus
s__MGYG-HGUT-03875 sp
s__PeH17 sp000435055
s__Porphyromonas sp.(00873)
s__Porphyromonas sp.(01211)
s__Porphyromonas sp.(02205)
s__Porphyromonas sp.(02784)
s__Porphyromonas sp.(03637)
s__Porphyromonas asaccharolytica
s__Porphyromonas somerae
g__UBA1685
g__UBA1691

Supplementary Fig. 6

Assessments of the prognostic effects of host factors and microbiota-derived risk score on survival in stage I-III CRC tumors.

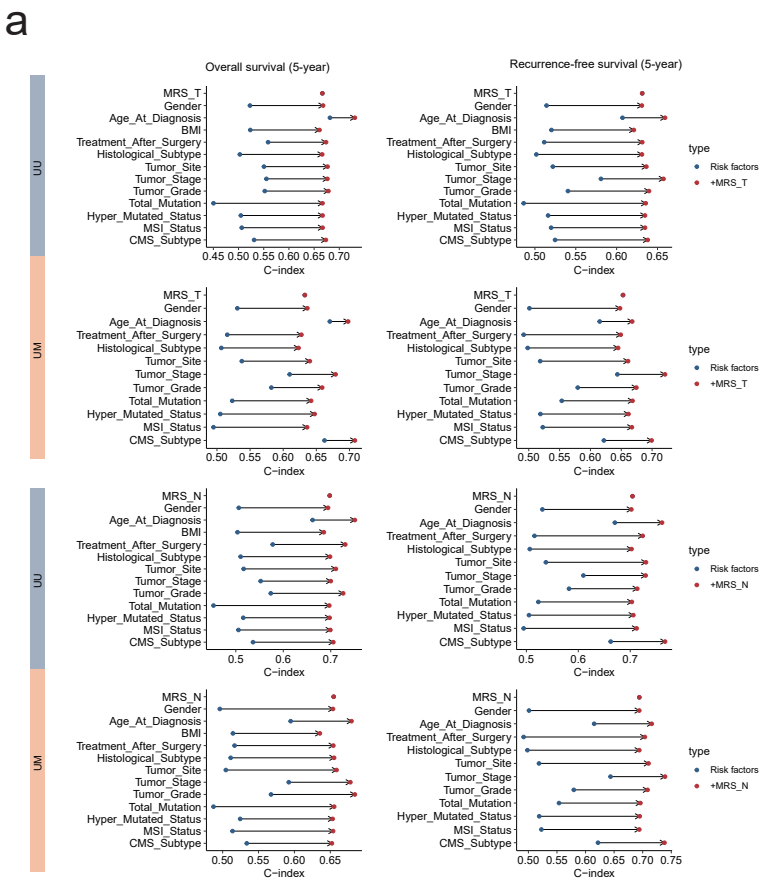
a. Forest plots display the hazard ratio (HR) values and two-sided P-values of host clinicopathological features for overall survival (OS; left) and recurrence free survival (RFS; right) in stage I-III UU tumors (n=689). Univariate survival analysis using Cox proportional hazards model. The factors (highlighted in red) significantly associated with OS or RFS were defined as covariates for microbiome-based prognosis analyses.

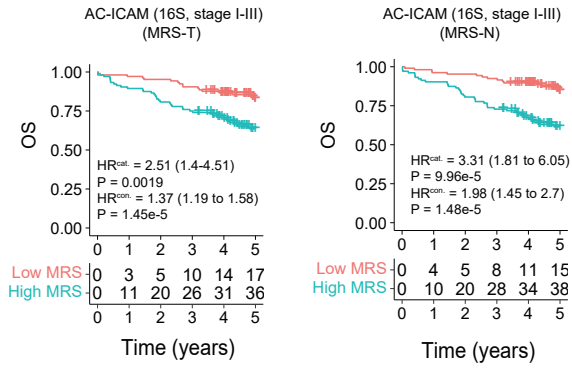
b. Scatter plots show the adjusted HR values of prognostic taxa for OS by CMS subtypes in the UU cohort. Adjusted HRs were calculated using multivariate Cox proportional hazard regression models (two-sided) after adjustment for age, tumor location, tumor stage, grade, postoperative treatments and Kraken2-mapped prokaryotic read count. Green and purple represent the taxa exhibiting significant associations with longer and shorter OS in UU tumors (n=689).

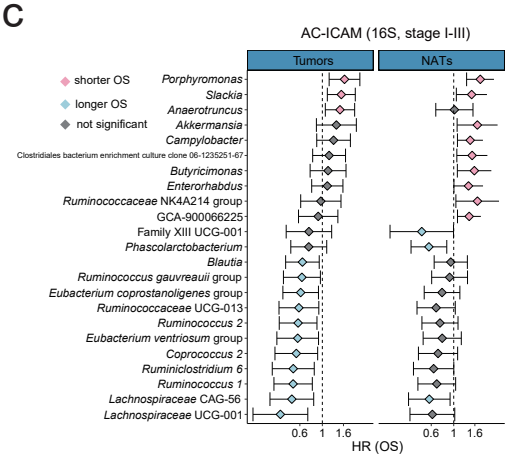
c. The validation of tumor-derived prognostic taxa identified in the UU cohort, validated in the UM (triangles) and AC-ICAM (circles) cohorts. HR and P-values were calculated using univariate Cox proportional hazards model (two-sided).

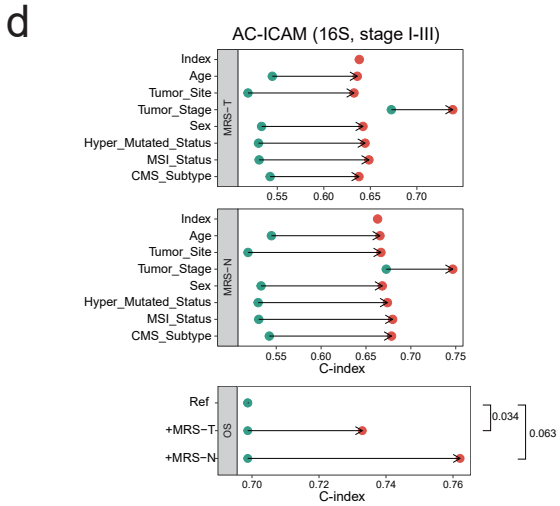
d. Associations (red frames in **c**) between selected prognostic taxa and survival outcomes (OS and RFS) across the three cohorts, estimated by multivariate Cox proportional hazard regression models (two-sided).

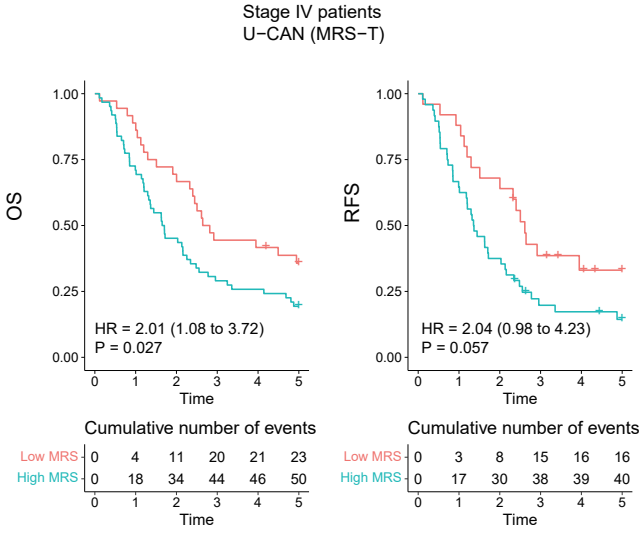
e. Heatmaps show differentially abundant taxa between stage I and other stages in tumors from the UU (BH-adjusted $P<0.05$), UM (two-sided $P<0.05$) and AC-ICAM (two-sided $P<0.05$) cohorts, assessed by MaAsLin2, with adjustment for age, sex, tumor location and prokaryotic read count. Blue and red indicate taxa significantly enriched and depleted in stage I tumors, respectively. Color bars, featuring of green, and brown, represent taxa associated with longer and shorter survival, respectively.


a
Overall survival (5-year)
Recurrence-free survival (5-year)
UU
UM
MRS_T
MRS_N
Gender
Age_At_Diagnosis
BMI
Treatment_After_Surgery
Histological_Subtype
Tumor_Site
Tumor_Stage
Tumor_Grade
Total_Mutation
Hyper_Mutated_Status
MSI_Status
CMS_Subtype
C−index
type
Risk factors
+MRS_T
+MRS_N


b
AC-ICAM (16S, stage I-III)
(MRS-T)
AC-ICAM (16S, stage I-III)
(MRS-N)
OS
HR^cat. = 2.51 (1.4-4.51)
P = 0.0019
HR^con. = 1.37 (1.19 to 1.58)
P = 1.45e-5
HR^cat. = 3.31 (1.81 to 6.05)
P = 9.96e-5
HR^con. = 1.98 (1.45 to 2.7)
P = 1.48e-5
Low MRS
High MRS
Time (years)


c
AC-ICAM (16S, stage I-III)
Tumors
NATs
shorter OS
longer OS
not significant
Porphyromonas
Slackia
Anaerotruncus
Akkermansia
Campylobacter
Clostridiales bacterium enrichment culture clone 06-1235251-67
Butyricimonas
Enterorhabdus
Ruminococcaceae NK4A214 group
GCA-900066225
Family XIII UCG-001
Phascolarctobacterium
Blautia
Ruminococcus gauvreauii group
Eubacterium coprostanoligenes group
Ruminococcaceae UCG-013
Ruminococcus 2
Eubacterium ventriosum group
Coprococcus 2
Ruminiclostridium 6
Ruminococcus 1
Lachnospiraceae CAG-56
Lachnospiraceae UCG-001
HR (OS)


d
AC-ICAM (16S, stage I-III)
MRS−T
MRS−N
OS
Index
Age
Tumor_Site
Tumor_Stage
Sex
Hyper_Mutated_Status
MSI_Status
CMS_Subtype
Ref
+MRS−T
+MRS−N
0.034
0.063
C-index


e
Stage IV patients
U−CAN (MRS−T)
OS
RFS
HR = 2.01 (1.08 to 3.72)
P = 0.027
HR = 2.04 (0.98 to 4.23)
P = 0.057
Time
Cumulative number of events
Low MRS
High MRS

Supplementary Fig. 7 Conceptional validation of the prognostic effects of the 16S amplicon-based microbiota-derived risk score of tumors and paired NATs in the AC-ICAM cohort.

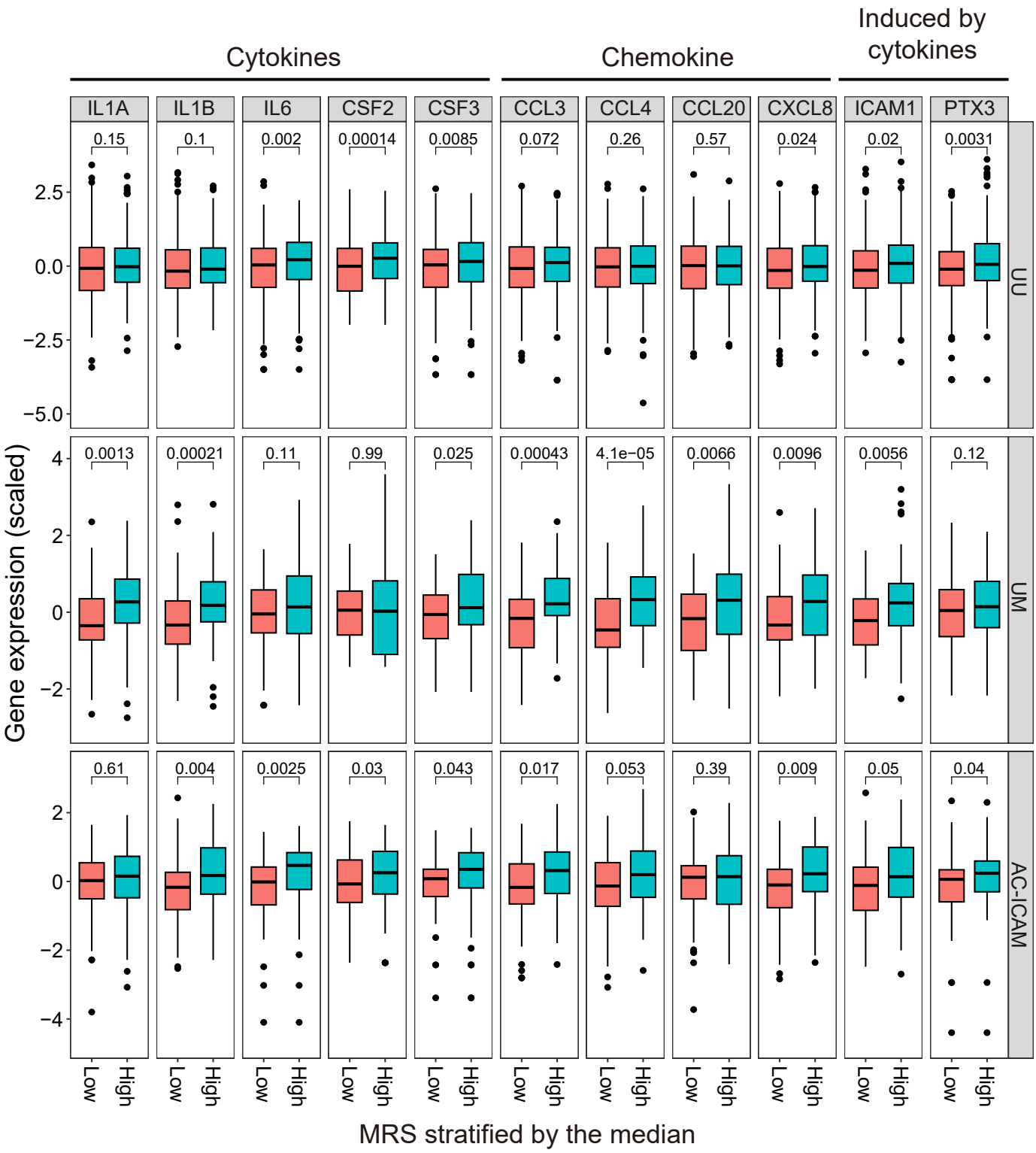
a. Prognostic performance of microbiota-derived risk score from tumors (MRS-T, **upper**) and NATs (MRS-N, **lower**) for OS and RFS. The assessment was conducted in patients with stage I-III CRC tumors in the UU cohort and validated in the UM cohort. The concordance index (C-index) of MRS alone and MRS plus each of the host clinicopathological factors were calculated.

b. Kaplan–Meier curves of the groups with low- and high MRS from tumors (MRS-T) or NATs (MRS-N) for OS in the AC-ICAM cohort (16S dataset, n=209). Adjusted HRs and two-sided P-values for both continuous and categorical MRS were presented, as determined by multivariate Cox proportional hazard models.

c. Forest plots display the adjusted HR values of 16S rRNA gene amplicon-based prognostic genera (continuous variables) associated with OS in tumors or NATs in AC-ICAM patients with stage I-III tumors (two-sided). Pink indicates taxa significantly associated with shorter OS; blue indicates taxa significantly associated with longer OS.

d. The C-index of MRS alone and MRS plus each of the host factors are calculated for 5-year OS using the 16S rRNA gene amplicon dataset from the 209 paired tumors and NATs in AC-ICAM patients with stage I-III tumors.

e. Kaplan–Meier curves of the groups with low- and high MRS-T (defined by median) for OS and RFS in 98 CRC patients with stage IV tumors in U-CAN cohort (86 patients from UU and 12 from UM). Models were adjusted for age, tumor location, grade, post-surgical treatment, and prokaryotic reads counts (two-sided).


Cytokines
Chemokine
Induced by cytokines
IL1A
IL1B
IL6
CSF2
CSF3
CCL3
CCL4
CCL20
CXCL8
ICAM1
PTX3
0.15
0.1
0.002
0.00014
0.0085
0.072
0.26
0.57
0.024
0.02
0.0031
0.0013
0.00021
0.11
0.99
0.025
0.00043
4.1e−05
0.0066
0.0096
0.0056
0.12
0.61
0.004
0.0025
0.03
0.043
0.017
0.053
0.39
0.009
0.05
0.04
UU
UM
AC-ICAM
Gene expression (scaled)
Low
High
MRS stratified by the median

Supplementary Fig. 8 Differential expressions of pro-inflammatory genes between low- and high MRS-T tumors across the three CRC cohorts.

Box plots show the differential expression patterns of 11 pro-inflammatory genes between the low- and high MRS-T tumors (by median) from the UU, UM and AC-ICAM cohorts. P-values were calculated by two-sided Wilcoxon rank-sum test.