Microenvironment of Adjacent Non-Neoplastic Regions Determines Prognostic Outcomes in Locally Advanced Colorectal Cancer after Surgical Resection: A Multi-Center & Multi-Omics study

Woong-Yang Park (woongyang@gmail.com)

Yeong Hak Bang
Samsung Medical Center

Ji Hye Choi
Samsung Medical Center

Kyunghie Park
Samsung Medical Center

https://orcid.org/0000-0003-0511-4958

Jinyeong Lim
Samsung Medical Center, Sungkyunkwan University School of Medicine

https://orcid.org/0000-0001-9358-9552

Boram Lee
Samsung Medical Center

https://orcid.org/0000-0002-7784-5971

Kyung Yeon Han
Samsung Medical Center

Dae Hee Pyo
Samsung Medical Center

Yong Beom Cho
Samsung Medical Center

Tae-You Kim
Seoul National University Hospital

Kyu Joo Park
Department of Surgery, Seoul National University Hospital, Seoul National University College of Medicine

Seung-Bum Ryoo
Seoul National University Hospital

https://orcid.org/0000-0001-6407-0444

Sung-Bum Kang
Department of Surgery, Seoul National University Bundang Hospital, Seoul National University College of Medicine

Chang Sik Yu
Department of Surgery, Division of Colon and Rectal Surgery, University of Ulsan College of Medicine and Asan Medical Center

Jaeim Lee  
Uijeongbu St. Mary's Hospital

Kil-yong Lee  
Uijeongbu St. Mary's Hospital

Kyu-Tae Kim  
Ajou University

Jin-Young Lee  
Yonsei University  https://orcid.org/0000-0002-5366-7488

Young-Joon Kim  
Yonsei University  https://orcid.org/0000-0001-5061-587X

Article

Keywords: Colorectal cancer, normal adjacent to tumor, microbiota, bacterial dysbiosis

Posted Date: November 14th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3375190/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License

Additional Declarations: There is NO Competing Interest.
Abstract

We aimed to predict postoperative prognosis of stage II/III colorectal cancer (CRC). Tumor-like normal microenvironment (tNME) and healthy normal microenvironment (hNME) groups were created using RNA sequencing (RNA-seq) data derived from 273 tumors and 273 paired normal tissues from margins of resected specimens. With a median follow-up of 58.2 months, the tNME group showed poor 5-year recurrence-free survival (54.7 vs. 73.0%, HR = 1.94, \(P = 0.002\)) and 5-year overall survival (78.2 vs. 83.0%, HR = 1.76, \(P = 0.033\)). Only the tNME group showed similar microbiome diversity between tumor and normal tissues. scRNA-seq revealed a positive co-occurrence of Th17, \(GZMK^{\text{high}}\)CD8+T cells, and \(IL1B^{\text{high}}\) neutrophils in the tumor and normal tissues of tNME. A high proportion of \(EMP1^{\text{high}}\) epithelial cells was observed in the tNME group, with upregulated epithelial-mesenchymal transition and leukocyte signaling pathways. In conclusion, the microenvironmental status of normal tissues offers a promising biomarker for stage II/III CRC.

Introduction

Colorectal cancer (CRC) is highly heterogeneous at both genomic and transcriptomic levels. Genomic biomarkers, namely microsatellite instability (MSI) and extended RAS and BRAF mutational status, are routinely used for prognostication and treatment prediction in clinical practice in metastatic settings\(^1\)–\(^4\). However, their use is not recommended in adjuvant settings, except for MSI, because of lack of value in predicting treatment benefits\(^3\)–\(^6\). Overall, 30–40% of surgically resected CRC cases relapse, and considering the poor prognosis of recurrent CRC, there is an unmet need for additive predictive biomarkers for recurrence\(^7\).

Previous studies have focused on tumor mutational profiles or microenvironments as potential biomarkers. However, heterogeneities in the molecular characteristics and microbiota within tumors present a substantial impediment in facilitating comparative analyses of the microenvironment across heterogeneous patients\(^8\)–\(^11\). To bypass intratumoral heterogeneity obstacles, we focused on the microenvironment of non-neoplastic region adjacent to tumor tissue (herein referred to as “normal”) derived from surgical tumor specimens. Disrupting protective mucus layer with heightened interactions between dysbiotic microbial networks is a crucial step in CRC carcinogenesis\(^12\),\(^13\) and evaluating the normal mucosal status may serve as an indicator of local tumor recurrence. Furthermore, normal tissues may be correlated with metachronous recurrence. Because bacteria-infected cancer cells have been reported to be related to colon cancer metastasis\(^8\) and bacterial biofilms, aggregations of the microbial community that contact unshielded epithelial cells and invade the tumor have been observed not only in tumors but also in normal tissues, even far from the tumor\(^13\)–\(^15\).

Increasing evidence suggests that normal tissues adjacent to tumors have molecular features intermediate between those of tumors and normal tissues from healthy controls, showing potential as a hallmark of tumorigenesis or tumor progression\(^16\)–\(^20\). However, most studies lack interpretation or
validation of clinical utility. Hence, we aimed to evaluate the prognostic role of the normal tissue-based microenvironment classification in stage II/III CRC after surgery and analyzed 16S rRNA sequencing and single-cell RNA sequencing (scRNA-seq) of tumor and normal tissues for biological interpretation.

Results

We investigated chemotherapy-naive colorectal tumor and paired normal tissues obtained from the furthest margins of resected specimens from 273 patients with stage II/III CRC, who underwent R0 surgical resection at the Samsung Medical Center (SMC), Seoul National University Hospital (SNUH), Bundang Seoul National University Hospital (SNUBH), Asan Medical Center (AMC), and Uijeongbu St. Mary’s Hospital (USMH), Republic of Korea, between July 2009 and July 2019.

We hypothesized that the more similar the normal microenvironment was to the tumor microenvironment (TME), the more likely it was to recur. As a proof-of-concept, we classified subgroups using RNA-seq data derived from tumor and paired normal tissues from surgical specimens of stage II/III CRCs and investigated the clinical outcomes of these subgroups. To quantify this similarity, we first analyzed the differentially expressed genes (DEGs) between tumor and normal tissues in 111 patients at the SMC and divided them into groups according to the mean score of the single-sample gene set enrichment analysis (ssGSEA$^{21}$) of the top 20 tumor DEGs set (Supplementary Table 1) in the normal tissue (herein referred to as “tumor signature”) (Extended Data Fig. 1a,b). The groups were defined as follows: patient with tumor-like normal microenvironment (tNME), with high level of tumor signature in normal tissue; patient with healthy normal microenvironment (hNME), with low tumor signature in normal tissue. Subsequently, the impact of this classification was evaluated in 273 cohorts, including the SMC and another multicenter cohort. Additionally, we externally validated The Cancer Genomic Atlas (TCGA) cohort consisting of 39 patients (Fig. 1a).

In the SMC cohort, 52 tumors and 13 normal tissues derived from 47 patients were available for scRNA-seq, whereas 44 tumor and paired normal tissues derived from 44 patients were available for 16S rRNA gene sequencing (Fig. 1a–b). Additionally, we performed hematoxylin and eosin (H&E) staining and targeted RNAscope-in situ hybridization (RNAscope-ISH) imaging using a eubacterial probe (RNAscope™ Probe- EB-16S-rRNA) to visually confirm the spatial distribution of the microbiota in the tumor and normal tissues.

Baseline characteristics of study cohort

Baseline characteristics of the study cohort are presented in Table 1. Among the total cohort of 273 patients, approximately two-thirds had left-sided CRC (N = 182, 66.7%), and most patients were diagnosed with stage III CRC (N = 204, 74.7%). Most patients exhibited a microsatellite stable (MSS) phenotype (N = 220, 80.6%), and a significant proportion of patients underwent adjuvant treatment (N = 207, 75.8%). A total of 103 (37.7%) patients were classified as tNME; no significant differences were observed in the baseline characteristics, including age, sex, TNM stage, tumor sidedness, proportion of patients who
underwent adjuvant chemotherapy, consensus molecular subtypes (CMS), and TME subtypes\textsuperscript{22} between the subgroups (Fig. 1c–e, Supplementary Table 2). We could measure the distance between tumor and paired normal tissues in 44 surgical specimens conducted 16S rRNA gene sequencing, and no significant differences were observed in tumor signature levels according to the distance between tumor and normal tissues, TNM stage, microsatellite instability (MSI) status, medical center, or tumor sidedness (Extended Data Fig. 1c–e). In the SMC cohort, 47 patients were available for scRNA-seq, and 17 (36.2\%) were classified as tNME. Approximately, 12 (25.5\%) patients with left-sided CRC and 19 (40.4\%) microsatellite instability-high (MSI-H) patients were included (Fig. 1e).
Table 1
Baseline patient characteristics

<table>
<thead>
<tr>
<th></th>
<th>Total Cohort (N = 273)</th>
<th>SMC Cohort (N = 111)</th>
<th>Multicenter Cohort (N = 162)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age (y) (mean±SD)</strong></td>
<td>62.3 ± 12.7</td>
<td>63.8 ± 12.6</td>
<td>61.3 ± 12.6</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>171 (62.6%)</td>
<td>66 (59.5%)</td>
<td>105 (64.8%)</td>
</tr>
<tr>
<td>Female</td>
<td>102 (37.4%)</td>
<td>45 (40.5%)</td>
<td>57 (35.2%)</td>
</tr>
<tr>
<td><strong>Tumor site side</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right sided</td>
<td>91 (33.3%)</td>
<td>51 (45.9%)</td>
<td>40 (24.7%)</td>
</tr>
<tr>
<td>Left sided</td>
<td>182 (66.7%)</td>
<td>60 (54.1%)</td>
<td>122 (75.3%)</td>
</tr>
<tr>
<td><strong>TNM stage</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage II</td>
<td>69 (25.3%)</td>
<td>28 (25.2%)</td>
<td>41 (25.3%)</td>
</tr>
<tr>
<td>Stage III</td>
<td>204 (74.7%)</td>
<td>83 (74.8%)</td>
<td>121 (74.7%)</td>
</tr>
<tr>
<td><strong>MSI status</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSI-H</td>
<td>39 (14.3%)</td>
<td>24 (21.6%)</td>
<td>15 (9.3%)</td>
</tr>
<tr>
<td>MSI-L</td>
<td>10 (3.7%)</td>
<td>1 (0.9%)</td>
<td>9 (5.6%)</td>
</tr>
<tr>
<td>MSS</td>
<td>220 (80.6%)</td>
<td>84 (75.7%)</td>
<td>136 (84.0%)</td>
</tr>
<tr>
<td>NA</td>
<td>4 (1.5%)</td>
<td>2 (1.8%)</td>
<td>2 (1.2%)</td>
</tr>
<tr>
<td><strong>CMS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMS 1</td>
<td>35 (12.8%)</td>
<td>15 (13.5%)</td>
<td>20 (12.3%)</td>
</tr>
<tr>
<td>CMS 2</td>
<td>67 (24.5%)</td>
<td>34 (30.6%)</td>
<td>33 (20.4%)</td>
</tr>
<tr>
<td>CMS 3</td>
<td>57 (20.9%)</td>
<td>19 (17.1%)</td>
<td>38 (23.5%)</td>
</tr>
<tr>
<td>CMS 4</td>
<td>89 (32.6%)</td>
<td>34 (30.6%)</td>
<td>55 (34.0%)</td>
</tr>
<tr>
<td>NA</td>
<td>25 (9.2%)</td>
<td>9 (8.1%)</td>
<td>16 (9.9%)</td>
</tr>
</tbody>
</table>

Data are presented as n (%), unless otherwise indicated.

Abbreviations: MSI, microsatellite instability; MSI-H, microsatellite instability high; MSI-L, microsatellite instability low; MSS, microsatellite stable; CMS, consensus molecular subtype; tNME, tumor-like normal microenvironment; hNME, healthy normal microenvironment; NA, not available for analysis.
### Normal tissue-based classification

<table>
<thead>
<tr>
<th></th>
<th>Total Cohort (N = 273)</th>
<th>SMC Cohort (N = 111)</th>
<th>Multicenter Cohort (N = 162)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>tNME</strong></td>
<td>103 (37.7%)</td>
<td>38 (34.2%)</td>
<td>65 (40.1%)</td>
</tr>
<tr>
<td><strong>hNME</strong></td>
<td>170 (62.3%)</td>
<td>73 (65.8%)</td>
<td>97 (59.9%)</td>
</tr>
</tbody>
</table>

### Adjuvant chemotherapy

<table>
<thead>
<tr>
<th></th>
<th>Total Cohort (N = 273)</th>
<th>SMC Cohort (N = 111)</th>
<th>Multicenter Cohort (N = 162)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>207 (75.8%)</td>
<td>69 (62.2%)</td>
<td>138 (85.2%)</td>
</tr>
<tr>
<td>No</td>
<td>66 (24.2%)</td>
<td>42 (37.8%)</td>
<td>24 (14.8%)</td>
</tr>
</tbody>
</table>

Data are presented as n (%), unless otherwise indicated.

Abbreviations: MSI, microsatellite instability; MSI-H, microsatellite instability high; MSI-L, microsatellite instability low; MSS, microsatellite stable; CMS, consensus molecular subtype; tNME, tumor-like normal microenvironment; hNME, healthy normal microenvironment; NA, not available for analysis.

## Recurrence and survival outcomes of the total cohort (N = 273) according to the classification

With a median follow-up of 58.2 months, the tNME group showed significantly poorer 5-year recurrence-free survival (RFS; 54.7 vs. 73.0%, HR = 1.94 [95% CI, 1.26–2.97], \(P = 0.002\)) (Fig. 2a) and 5-year overall survival (OS; 78.2 vs. 83.0%, HR = 1.76 [95% CI, 1.05–2.97], \(P = 0.033\)) (Fig. 2b). The most frequent site of recurrence was the liver (40.5%), followed by the lungs (34.5%); there was no significant difference in the location of recurrence between tNME and hNME (Supplementary Table 3). In the subgroup analysis, tNME showed poor prognosis regardless of MSI status, whereas the difference in prognosis was significant in patients stratified for left-sided CRC (HR = 2.40 [95% CI, 1.46–3.95], \(P < 0.001\)) (Fig. 2c, Extended Data Fig. 2). Multivariate analysis revealed that tNME was an independent negative prognostic factor for RFS (vs. hNME, HR = 2.07 [95% CI, 1.34–3.18], \(P = 0.001\)), and OS (vs. hNME, HR = 1.87 [95% CI 1.11–3.15], \(P = 0.019\)) (Supplementary Table 4). In addition, tNME showed a trend toward poorer 5-year OS compared to that of hNME (34.5 vs. 68.8%, HR = 3.76 [95% CI, 0.81–17.53], \(P = 0.071\)) in the TCGA validation cohorts (Fig. 2d).

### Pathway analysis

Gene set enrichment analysis (GSEA) with gene ontology (GO) showed tNME tumors presenting a relatively decreased maintenance pathway of the epithelium, whereas enrichment of leukocyte chemotaxis signals and antimicrobial responses were observed (Fig. 3a). Leukocyte pathway enrichment was also observed in normal tissues of tNME, and interestingly, the vitamin and flavonoid metabolic pathways, which were previously reported to be correlated with the modulation of colon mucosal barrier...
permeability, protection of the mucus layer, regulation of the intestinal immune system, fight against oxidative stress, and positively shaped microbiota\textsuperscript{23,24} were significantly decreased in normal tissues of tNME (Fig. 3b).

**Metagenomic analysis of tumor and normal tissues according to subgroups**

We performed 16S rRNA gene sequencing of 44 tumors and 44 paired normal tissues from 44 patients (Extended Data Fig. 3a). As hypothesized that the microbiota of normal tissues of tNME was similar to that of the tumors, we compared the tumor and normal microbiota of tNME and hNME. Although not statistically significant, the normal tissues of hNME showed higher alpha diversity (Berger-Parker index) than that of the tumors of hNME (Wilcoxon signed-rank test, $P = 0.077$), whereas similar alpha diversity was observed between the tumor and normal tissues of tNME (Fig. 3c). Principal coordinate analysis (PCoA) with beta diversity (Bray-Curtis dissimilarity) of the bacterial community showed differences between the tumor and normal tissues in the hNME group (Permutational multivariate analysis of variance [PERMANOVA], $F = 1.907$, $P = 0.020$). However, there was no significant difference in the tNME group (PERMANOVA, $F = 0.919$, $P = 0.510$) (Fig. 3d). *Prevotella, Bacteroides, Treponema, Fusobacterium, Leptotrichia, Campylobacter*, and *Selenomonas* were the most dominant genera in the tumors of both tNME and hNME, whereas the proportions of these genera differed between the normal tissues of tNME and hNME (Fig. 3e). In addition, we compared the proportional abundance of genera between tumor and normal tissues, and notably, significantly lower proportions of carcinogenic or enterotoxigenic genera (*Fusobacterium*\textsuperscript{25–27}, *Bacteroides*\textsuperscript{28–30}, *Leptotrichia*\textsuperscript{31}, *Campylobacter*\textsuperscript{32}, *Selenomonas*\textsuperscript{31,33}, *Hungatella*\textsuperscript{34}, *Lachnoanaerobaculum*\textsuperscript{35}, *Peptacetobacter*\textsuperscript{36}) were uniquely present in normal tissues of hNME than in tumors of hNME, whereas no significant differences between tumor and normal tissues were observed in tNME group. In contrast, more *Parabacteroides* genera, which inhibit TLR4 suppression and have an antitumor effect,\textsuperscript{37,38} were found in normal tissues than in tumors in the hNME group (Fig. 3f, Extended Data Fig. 3b).

**Single-cell RNA seq analysis of tumor and normal tissues, derived from 12 patients**

We analyzed 15 tumors and 13 normal samples derived from 12 patients for which both tumor and paired normal scRNA-seq data were available. Among them, four patients were classified as tNME, and eight were hNME. A total of 32,221 tumor and 22,984 normal cells were analyzed (Extended Data Fig. 4a–d). We identified 11 different myeloid clusters, and macrophage subsets featured the canonical markers $C1QA$, $C1QB$, $CD68$\textsuperscript{39}, were subsequently divided based on DEGs ($C1QC$, $SPP1$, $MKI67$). Monocytes were characterized by the expression of *S100A8* and *S100A9*\textsuperscript{39}. In addition, *S100A12*, $GOS2$, and $CSF3R$ were used to identify two distinct neutrophil subgroups\textsuperscript{40}. Among these neutrophil subgroups, one was distinguished by the expression of $IL1B^{\text{high}}$. Other myeloid cell subtypes were defined based on following canonical markers: mast ($KIT$, $TPSAB1$, and $CPA3$)\textsuperscript{41} and dendritic (DC) cell markers.
(conventional type 1 DC [cDC1]: CLEC9A, XCR1, and CADM1; conventional type 2 DC [cDC2]: CLEC10A, CD1C, and CD1E; plasmacytoid DC [pDC]: LILRA4, GZMB, IL3RA; lymphoid DC: LAMP3, CCR7, and FSCN1) (Fig. 4a).

Unsupervised clustering analysis of T and NK cells identified 11 subtypes: GZMK\textsuperscript{high}CD8\textsuperscript{+} T cells, characterized by a high expression of GZMK along with GZMA, PRF1, GZMB, GZMH, IFNG, and NKG7, but without ENTPD1 expression. The exhausted T (Tex) cluster was characterized by the highest expression of exhaustion-related markers, PDCD1 (PD-1), TIGIT, and LAG3 and resident-tissue memory CD8\textsuperscript{+} T (Trm) cells were identified using the canonical markers ITGAE (CD103) and CD69. CD4\textsuperscript{+} T cells were defined based on the following canonical markers (follicular helper T cells [Tfh]: MAF, CXCL13; Th17: CCR6, KLRB1, ROA, and IL17A; Regulatory T cells [Treg]: FOXP3, IL2RA, and IKZF2). In addition, the central memory CD4\textsuperscript{+} (Tcm) cells were featured based on the expression of CD69, IL7R, and GPR183; T cells were characterized based on TRGC2 and TRDC expression; Proliferating T cells identified with expression of MKI67, PCNA, and STMN1; Naïve-like T cells expressed TCF7, CCR7, and SELL and NK cells expressed FCGR3A, KLRD1, and KLRF1, respectively (Fig. 4b).

We conducted Spearman's correlation analysis of the proportions of immune cells (T and myeloid cells) between tumor and normal tissues to identify clusters that comprised a similar microenvironment in tNME. The proportions of Th17 cells, GZMK\textsuperscript{high}CD8\textsuperscript{+} T cells, and IL1B\textsuperscript{high} neutrophils were correlated between tumor and normal cells in tNME (Fig. 4c). Interestingly, Th17 cells are known to recruit neutrophils and highly cytotoxic T cells, and a recent study reported that neutrophils interact with CD8\textsuperscript{+} positive T cells, inducing the production of high levels of GZMK, which in turn decreases E-cadherin expression in the intestinal epithelium and promotes tumor progression.

Ten stromal or endothelial cells were identified using fibroblast (COL3A1, DCN, and THY1) and endothelial cell (EC) markers (ENG and PECAM1). Subsequently, fibroblast subsets were characterized by each DEG (ADAMDEC1, SFRP2, and FAP) and ECs were identified with canonical markers (telocyte: SOX6, F3; pericyte: NDUFA4L2, RGS5, and NOTCH3; tip-like EC: RGCC, RAMP3; stalk-like EC: ACKR1, SELP; lymphatic EC: LYVE1, PROX1; smooth muscle cell: SYNPO2, CRYAB, CNN1, and DES; glial cell: SOX10, S100B, and PLP1; Fig. 4d). Figure 4e shows the scRNA-seq analysis of epithelial cells. In addition to colonocytes (CA1, CA2, and KRT20), goblet (TFF3, MUC2, and FCGBP), enterochromaffin (NEUROD1, FEV, and CHGB), and Paneth (DEFA5, DEFA6) cells, epithelial cells were characterized based on the expression of EMP1, LGR5, and MKI67, and additional PLCG2, OLFM4, and BEST4 epithelial cell clusters were identified based on DEGs. Differences between tumor and normal tissues were prominent in the UMAP embedding fibroblasts and endothelial, and epithelial cells. In particular, cancer associated fibroblast (CAF), EMP1\textsuperscript{high}, LGR5\textsuperscript{+}, and proliferating epithelial cell clusters were distinctly observed in the tumor tissues (Fig. 4d, e and Extended Data Fig. 4c–d).
As most patients experienced metachronous rather than local recurrence, we further investigated cancer cells and the TME using available scRNA-seq data from 52 tumor samples derived from 47 patients. Among these, 17 (36.2%) were classified as tNME. Figure 5 presented the scRNA-seq analysis of tumor tissues, and the total of 126,628 tumor cells were available for analysis (Extended Data Fig. 4e–f). The cell clusters were identified using canonical markers described above. In addition, we further divided the CAF and T cells using other canonical markers (inflammatory CAF [iCAF]: CFD, CXCL8, and CXCL1; myofibroblast-like CAF [myCAF]: MMP11, HOPX; mucosal-associated invariant T [MAIT]: SLC4A10, TRAV1-2; stress response T [Tstr]: NR4A1, BAG3, HSPA1A; CD4+ cytotoxic T [CTL]: GZMA, GNLY, PRF1, GZMB, and GZMK). The tNME comprised a higher proportion of inflammatory CAF (iCAF) and GZMK<sup>high</sup>CD8<sup>+</sup>T cells, which may indicate an inflammatory TME (Fig. 5a–c). Figure 5d shows the scRNA-seq analysis of epithelial cells, in which the clusters were divided based on EMP1, LGR5, MKI67 expression, and an additional PLCG2<sup>high</sup> epithelial cell cluster. Notably, a significantly higher proportion of EMP<sup>high</sup> epithelial cell cluster was observed in the tNME group. This cluster was correlated with poor prognosis, along with iCAF and myCAF inferred by the SCISSOR method, a computational method that deduced cell-type transcriptomic signatures<sup>53</sup>, based on the survival data of the 111 SMC cohort (Fig. 5e). The EMP<sup>high</sup> epithelial cell cluster displayed high expression of Epithelial-Mesenchymal Transition (EMT), inflammatory response, and angiogenesis pathway gene signatures in the Molecular Signatures Database (MSigDB) (Fig. 5f). Additionally, pathway enrichment analysis demonstrated that this cluster showed upregulated leukocyte transendothelial migration and bacterial invasion of epithelial cell pathways in the KEGG database<sup>54</sup> (Fig. 5g).

We further conducted a CellChat analysis<sup>55</sup> to investigate the interactions between epithelial cell clusters and inflammatory immune clusters. Cell-to-cell interaction analysis revealed a significantly upregulated interaction between EMP<sup>high</sup> epithelial cells and GZM<sup>high</sup>CD8<sup>+</sup>T cells, IL1B<sup>high</sup> neutrophils, C1QC<sup>high</sup> macrophages and SPP1<sup>high</sup> macrophages (Fig. 5h). The most prominent pathway in tNME is the midkine (MK) pathway, which promotes EMT, cancer invasion, and metastasis through a combination of mitogenic, pro-inflammatory, and angiogenic functions<sup>56</sup>. Furthermore, the vascular endothelial growth factor (VEGF)<sup>57,58</sup> and secreted phosphoprotein 1 (SPP1) pathways<sup>59,60</sup>, which are key signaling pathways that mediate angiogenesis and promote EMT, were consistently elevated in tumors of tNME. In addition, TIGIT-PVR and TIGIT-NECTIN2 interactions between EMP<sup>high</sup> epithelial and Tex cells were increased in tumors of tNME, which are related to inhibitory signals in TCR expression or T cell activations<sup>61</sup> (Extended Data Fig. 5).

**The tNME group showed more proportion of bacteria-infected cancer cells**

Recent study reported the bacteria-infected CRC cells cause the cancer metastasis<sup>8</sup>. In addition, tNME showed a higher proportion of EMP<sup>high</sup> epithelial cells, which are related to bacterial invasion and leukocyte-related pathways. Therefore, we used CSI-microbes<sup>62</sup>, a PathSeq<sup>63</sup>-based computational
method, to identify intracellular microbes. Notably, a higher proportion of bacteria-infected cancer cells was detected in the tNME group. In addition, intracellular microbial taxa differed between subgroups, with a greater proportion of *Fusobacterium*, *Treponema* genera-infected cancer cells in the tNME (Fig. 5i).

**Differences in the spatial distribution of microbiome and histological characteristics between tNME and hNME**

To evaluate the spatial distribution of the microbiota, we conducted RNAscope-ISH on representative tumor and paired normal samples of tNME and hNME. Both the samples were subjected to 16S rRNA gene sequencing. Patient#1 was classified as tNME, and 16S rRNA gene sequencing revealed *Treponema* and *Prevotella* genera-dominant microbiota in the tumor and paired normal tissue. Bacterial colonies (brown) were observed in both the tumor and normal mucosal areas (Fig. 6a–c). Patient#2 was classified as hNME. As shown in Fig. 6d, bacterial aggregation was observed in the tumor, whereas it was scarcely present in the normal tissue. The 16S rRNA sequencing of the tumor revealed a *Treponema* genus-dominant microbiota, similar to that of patient#1, whereas 16S rRNA sequencing of normal tissues showed different genera, such as *Erysipelatoclostridium* and *Phocaeicola* (Fig. 6.d–g). Meanwhile, a relatively large number of lymphocytes surrounding the disordered crypts were observed in the H&E slides of another tNMEs (Fig. 6b, Extended Data Fig. 5a), whereas most of the normal tissues of hNMEs showed relatively ordered crypts (Fig. 6e, Extended Data Fig. 5b).

**Putative mechanism**

The tNME group exhibited features of bacterial dysbiosis, showing similar microbiota between the tumor and normal tissues, with an activated bacterial humoral response and decreased mucosal maintenance. tNME presented a positive co-occurrence of Th17, *GZMK*\(^{\text{high}}\)CD8+T cells, and *IL1B*\(^{\text{high}}\) neutrophils in the tumor and normal tissues. The tumors of tNME showed a higher proportion of *EMP1*\(^{\text{high}}\) epithelial cells, which were related to bacterial invasion, leukocyte signaling, inflammatory and interferon signaling, and the EMT pathway (Fig. 6g). This epithelial cell cluster has characteristics similar to those of recently reported bacteria-infected cancer cells, which invade the surrounding environment and recruit myeloid cells to the bacterial regions. This cluster also exhibited upregulation of EMT-related genes and downregulation of cell cycle-related signaling pathways. Figure 6h illustrates this putative mechanism.

Microbiota dysbiosis may induce Th17 cells, and Th17 cells promote neutrophil retention, increasing crosstalk with CD8 + T cells, and inducing the production of high levels of *GZMK*, which in turn decreases E-cadherin in the intestinal epithelium and promote tumor progression. A weak intestinal barrier may provoke more bacteria-infected cancer cells, promoting cancer cell progression via EMT and recruiting *IL1B*\(^{\text{high}}\) neutrophils and *GZMK*\(^{\text{high}}\)CD8+T cells, resulting in a vicious cycle. Organized microniches with immune and epithelial cells may induce metachronous recurrence after surgical resection.

**Discussion**
We investigated the prognostic role of paired normal tissue-based subgroups in stage II/III CRC. The tNME group showed poor prognosis and was associated with a disrupted intestinal barrier and microbiota dysbiosis. Microbiota dysbiosis may induce Th17 cells, $IL1B^{\text{high}}$ neutrophils, and $GZMK^{\text{high}}CD8^+T$ cells, which attack the intestinal epithelial integrity and promote tumor progression. In addition, tumors of tNME showed more infected cancer cells and a higher proportion of $EMP1^{\text{high}}$ epithelial cells, which strongly interacted with $IL1B^{\text{high}}$ neutrophils and $GZMK^{\text{high}}CD8^+T$ cells. Our study underscores that the normal tissue adjacent to a tumor could be a novel prognostic or predictive biomarker for patients with localized CRC after surgical resection.

A previous study reported the presence of bacterial biofilms that invaded the colonic mucosa. In addition, biofilm-covered CRC is known to exhibit bacterial invasion into the tumor. Given that tNME features resemble the microbiota and inflammatory microenvironment of tumor and normal tissues, a higher proportion of bacteria-infected cancer cells, and colonization of bacteria in both the mucosal areas of tumor and normal tissues, we speculated that tNME might be related to bacterial biofilm formation. Considering that bacterial biofilms are less frequently observed in the left-sided CRCs than in the right-sided CRCs, and the prognostic difference between tNME and hNME is prominent in left-sided CRCs, the bacterial biofilm in the left-sided CRCs may be crucial for prognostic outcomes after surgical resection.

In addition to these prognostic differences after surgical resection, recent US nationwide cohort-based prospective studies reported that microbiota dysbiosis was uniquely related to the incidence of left-sided CRC, which was associated with a Western dietary pattern, including processed red meat and a sulfur microbial diet. In contrast, prudent diets rich in whole grains and dietary fiber are associated with a lower incidence for patients detected with *Fusobacterium nucleatum* in CRC tumor tissue. Based on these findings, we can infer that western diet-induced microbiota dysbiosis may be crucially related to the incidence and prognosis of left-sided CRC. In this context, our pathway analysis highlighted that normal tissues of tNME showed decreased flavonoid or vitamin pathway activity, which is known to affect the colonic mucosa and positively shape the microbiota. A recent randomized controlled trial also reported that a high flavonoid intake showed clinical benefits in patients who underwent surgical resection for CRC. Although we could not determine the causality, it seems certain that bacterial dysbiosis was negatively correlated with a prudent diet, and our data could be the rationale for personalized nutritional support.

Our results highlight poor clinical outcomes of tNME, and this prognostication was consistent regardless of the MSI status. These results provide evidence for the potential application of different management strategies. A recent human-like mouse model-based study investigated the mechanisms underlying metachronous recurrence after surgical resection of CRC. This study reported that the $EMP1^{\text{high}}$ epithelial cell cluster was enriched in liver micrometastases with high T-cell infiltration and became progressively immune-excluded during outgrowth. These results suggest a potential role for perioperative immunotherapy in CRC. As tNME features a high proportion of $EMP1^{\text{high}}$ epithelial cell cluster, its prognosis can be prolonged by adjuvant immunotherapy after surgery. Given that the clinical utility of
adjuvant immunotherapy for MSI-H or POLE exonuclease domain mutations (NCT02912559 and NCT03827044) is currently under investigation, benefit to the tNME subgroups requires further evaluation. Furthermore, even in the absence of immunotherapy, a clinical trial of an intensified treatment, including modified FOLFIRINOX (oxaliplatin, L-leucovorin, and irinotecan on day 1, and continuous intravenous 5-FU for 46 h), is ongoing (NCT02967289), making subgroup analysis based on our classification worth considering.

Our study had several limitations. First, although our study was based on multiple tertiary referral institutions and validated using TCGA cohort, intratumoral or intranormal mucosal microbial heterogeneity could affect the results. Second, there were discrepancies between the study and the scRNA-seq cohorts, which included a higher proportion of right-sided and MSI-H cancers. To overcome this discrepancy, we integrated the transcriptome data from the study cohort with scRNA-seq data using the SCISSOR method. Third, because our scRNA-seq data were not generated to check for bacteria-infected cells, there is a possibility of underestimating bacteria-infected cancers. Fourth, as the objective of this study was to compare normal tissues, a possibility of being affected by the distance from the tumor to the normal tissue exists. Although no correlation between distance and tumor signature was observed in normal tissues, a detailed protocol is required to identify agnostic biomarkers.

In conclusion, classification based on the normal microenvironment may be a promising biomarker for stage II/III CRC. Microbiota dysbiosis and a weak intestinal barrier due to colon inflammation are associated with poor prognosis. Based on our classification, prudent dietary habits or more advanced treatment strategies may be required to improve survival outcomes.

**Online Methods**

**Survival analysis**

Recurrence-free survival (RFS) was calculated from the date of surgery to the date of the first tumor recurrence or death from any cause, whichever occurred first. Overall survival (OS) was defined as the time from the date of surgery to the date of death from any cause or last follow-up. The Kaplan–Meier method was used to estimate survival outcomes, and the log-rank test was used to compare survival outcomes among subgroups. Univariate and multivariate analyses of RFS and OS were performed using Cox proportional hazards models. Variables with a potential relationship \( (P < 0.1) \) in the univariate analysis were included in the multivariate analysis. A \( P \)-value of < 0.05 was considered statistically significant. All statistical analyses were performed using the R software (version 4.0.5; R Foundation for Statistical Computing, Vienna, Austria).

**RNA sequencing**

**RNA extraction and sequencing**
Tumor and matched normal fresh-frozen tissues obtained from patients with CRC were dissected at 20–40 mg and homogenized 3–4 times for 15 s at a frequency of 30 Hz using a Tissue Lyser II (Qiagen, Hilden, Germany). RNAs were extracted using the Rneasy Mini Kit (Qiagen, Hilden, Germany) according to the standard tissue RNA extraction protocol. Total RNA concentration was calculated using Quant-IT RiboGreen (R11490; Invitrogen, Waltham, MA, USA). To assess the integrity of the total RNA, samples were evaluated using a TapeStation RNA screentape (5067–5576, Agilent, Santa Clara, CA, USA). Only high-quality RNA preparations (RNA integrity number [RIN] > 7.0) were used for the RNA library construction. The total RNAs were subjected to rRNA depletion using Ribo-Zero Gold rRNA Removal Kit (MRZG12324, Illumina, San Diego, CA, USA) and then added with the 2 µL of 100-fold diluted ERCC Mix2 solution of ERCC RNA Spike-In Mix (4456740, Ambion, Austin, TX, USA). An RNA-sequencing library was prepared using a TruSeq RNA Sample Prep Kit (Illumina, San Diego, CA, USA). The libraries were then subjected to an Illumina HiSeq2000 platform (Illumina, Inc., San Diego, CA, USA), and paired-end (2×100 bp) sequencing was performed by Macrogen, Inc.

RNA pre-processing

The human GRCh38 reference genome and gene annotation GTF file (GENCODE version 27) were downloaded from GENCODE (https://www.gencodegenes.org/human/), and the ERCC sequences FASTA and GTF annotations were downloaded from Thermo Fisher Scientific (https://assets.thermofisher.com/TFS-Assets/LSG/manuals/ERCC92.zip). Genome indexing and alignment were performed using STAR (v2.5.3a) and gene expression levels were quantified using RSEM (v1.3.0) \(^67\).

Tumor signature gene selection and classification of the group

Differentially expressed genes (DEGs) were identified using edgeR software (version 3.38.4) \(^68\). Count matrices and genes were generated using the DGEList and filterByExpr functions. Principal component analysis (PCA) was performed to investigate batch effects. Trimmed mean of M-values (TMM) normalization was performed using calcNormFactors. Finally, the negative binomial dispersion parameters were estimated using estimateGLMCommonDisp, estimateGLMTrendedDisp, and estimateGLMTagwiseDisp. We obtained DEGs by running glmQLFit and glmQLFTest and filtered them using logFC > 1, FDR < 0.05, and logCPM > 3 thresholds. Among the DEGs, we selected the top 20 enriched genes in tumors as tumor signature genes.

To quantify the tumor signature, we performed ssGSEA \(^21\) using the tumor signature genes for each sample. Counts per million (CPM) were generated using edgeR \(^68\) with log = T. ssGSEA was performed using the gsva function in the gsva package with the options \texttt{mx.diff = F, kcdf=“Poisson”, method=“ssgsea”, ssgsea.norm = F}.

Groups were divided based on the mean tumor signature score (ssGSEA score = -6764) of 111 paired normal samples derived from patients at the SMC. The tNME was defined as those who scored higher
than the mean tumor signature score, while the hNME was defined as those who scored lower than the mean tumor signature score. We applied the same tumor signature score criteria to another multicenter and TCGA cohorts.

**Differentially expressed genes and related pathway analysis**

Gene Set Enrichment Analysis (GSEA\(^69\)) was performed using gseGO in the clusterProfiler package (version 4.4.4). gseGO was performed using the options `keyType=“SYMBOL”, pvalueCutoff = 0.05, OrgDb = org.Hs.eg.db` and `pAdjustMethod=“fdr”`.

**CMS prediction and TME classification**

The CMS was predicted using the CMScaller package (version 0.99.2)\(^70\). TME classification was performed using the molecular functional portrait (MFP)\(^22\).

**DNA isolation and 16S rRNA gene sequencing**

A total of 44 matched normal and tumor samples obtained from the same individuals with CRC who underwent resection surgery were used for 16S rRNA analysis. V3–V4 amplicon sequencing data for 16S rRNAs were obtained using the Illumina MiSeq Reagent Kit v3 (2 × 300 bp, Illumina, USA). PCR primers, forward (CCTACGGGNGGCWGCAG) and reverse (GACTACHVGGGTATCTAATCC), were designed based on the hypervariable regions (V3–V4) of the 16S rRNAs. PCR was conducted using 2× KAPA HiFi HotStart ReadyMix (Roche) under the following conditions: 95°C solution chain for 3 min, 25 cycles of 95°C for 30 s, 55°C for 30 s, and 72°C for 45 s, followed by a 72°C extension for 5 min. Sequencing libraries were then constructed using a TruSeq® DNA PCR-Free Sample Preparation Kit (Illumina, USA) and TruSeq® Nextera XT index primer (Illumina, USA), and 2× KAPA HiFi HotStart ReadyMix (Roche) using the PCR products after purification. Subsequently, paired-end reads were generated by sequencing on a MiSeq platform after determining the quality of the library using a Tapestation 4200 platform (Agilent Technologies) and a Qubit Fluorometer (Thermo Fisher Scientific, WA, USA).

**16S rRNA data analysis**

To enhance the sensitivity of estimating the abundance of microbiota in each donor, we first sought to mitigate the sequencing bias. We removed contaminated human reads and adapter sequences from the 16S rRNA reads using the Trimmomatic software (version 0.36)\(^71\). Kraken2 (version 2.12)\(^72\) was used to detect microbial reads and assign taxonomic classifications using the default settings. We then computed the abundance of taxa at the genus level using Braken (version 2.6.2)\(^73\) with the default settings. At the genus level, the relative abundance was calculated as the read count of a specific genus divided by the total number of read genera in each sample. Microbiota \(\alpha\)-diversity of individual bacterial families was calculated Berger-parker with alpha_diversity.py from KrakenTools\(^74\) and analyzed between tumor and paired normal tissue of tNME and hNME using a Wilcoxon signed-rank test. Microbiota \(\beta\)-diversity was calculated on the Bray–Curtis's dissimilarity and analyzed by permutational multivariate analysis of variance (PERMANOVA) using the vegan\(^75\) package version 2.6.
Single-cell RNA sequencing

Single-cell preparation for 10x sequencing

For scRNA-seq, tissue dissociation was performed using a Tumor Dissociation Kit (Miltenyi Biotech, Germany) according to the manufacturer's instructions. Briefly, tissues were cut into pieces of 2–4 mm in size and transferred to a C tube containing the enzyme mix (enzymes H, R, and A in RPMI 1640 medium). GentleMACS programs (h_tumor_01, h_tumor_02, and h_tumor_02) were run in a MACSmix tube rotator (Miltenyi) with two 30-min incubations periods at 37°C between each run. The digested samples were filtered through a 70 µm strainer and washed with RPMI 1640 medium. Each cell suspension purified using a Ficoll-Paque PLUS (GE Healthcare, Chicago, IL, USA) was processed with 10x Chromium Single Cell 3’ Reagent Kits v3 (10x Genomics) according to the manufacturer's protocol.

scRNA-seq and data processing

Single-cell RNA-seq libraries using the 10x Single Cell 3’ v2 Reagent Kit were prepared according to the manufacturer's protocol (10x Genomics, Pleasanton, CA, USA). Sequencing libraries were sequenced on an Illumina HiSeq 4000 platform, targeting 100,000 reads per cell according to the manufacturer's instructions (Illumina). Both reads were aligned to the GRCh38 human genome reference sequence and quantified using the CellRanger software (version 7.0.1). SoupX and DoubletFinder were used to remove the ambient RNA and doublets. Gene expression was analyzed using the Seurat software (version 4.0.5). In total, 149,612 cells were considered, based on the following criteria: >200 detected genes, < 7,000 detected genes, and < 25% mitochondrial content. The data were normalized using a log-normalized function with a scale factor of 10,000. Variable features were identified using the FindVariableFeatures function, returning 3,000 features. Subsequently, a PCA was performed based on a processed expression matrix containing highly variable genes. Next, we applied the Harmony batch correction package to each sample ID to adjust for potential batch-derived effects across samples. Uniform manifold approximation and projection (UMAP) was used to visualize the cells in two-dimensional space, followed by the FindNeighbors and FindClusters functions of Seurat. Major cell types were annotated by comparing canonical marker genes and DEGs for each cluster using FindAllMarkers with the Wilcoxon rank-sum test.

Quantifying immune phenotype similarity between tumor and normal

To evaluate the similarity in the immune microenvironment between tumors and normal tissues, we conducted patient-level Spearman correlation analysis of the proportion of immune clusters (T and myeloid cells) and hierarchical clustering analysis using pheatmap (version 1.0.12).

SCISSOR method
We used SCISSOR\textsuperscript{53} to associate phenotypic data from the bulk RNA-seq experiments with single-cell data. SCISSOR\textsuperscript{53} was performed on the primary tumor cells of patients according to the SCISSOR tutorial, using recurrence-free survival (Cox regression) as the dependent variable. We computed the fractions of scissor-positive (positively associated with poor survival) and scissor-negative (negatively associated with survival) cells.

**Epithelial cluster characteristic**

We identified each epithelial cluster trait using Seurat's AddModuleScore, which calculates the average expression level of each cell cluster using selected MSigDB Hallmark gene sets. A heatmap was generated using the heatmap function with \textit{scale="row"}. Additionally, to identify the EMP\textsuperscript{high} epithelial cluster characteristics, we identified DEGs using the FindAllMarkers function in the Seurat R package. The DEG lists were filtered based on the following criteria: expression in at least 20% of the cluster cells, average expression log2-fold change > 0.5, and q-value < 0.05. We next conducted Enrichr\textsuperscript{81} (version 3.2) with "KEGG_2021_Human" database\textsuperscript{54} for Over-representation analysis.

**CellChat analysis**

CellChat\textsuperscript{55} was used to assess cell-cell communication via interaction network analysis. The tNME and hNME data were processed separately, and each Seurat object was used as an input for CellChat following the standard protocol (https://github.Com/sqjin/CellChat). Population size was considered in computeCommunProb function with the option \textit{population.size = T} and minimum count of communicating cell is 10. Cell-cell communication networks were calculated using the getMaxWeight function, and circle plots were generated using netVisual_aggregate.

**CSI-microbes**

CSI-microbes\textsuperscript{62} was used to infer cell type-specific intracellular microbial taxa from single-cell RNA sequencing. CSI-microbes was run on the scRNA-seq of each patient according to the CSI-microbes tutorial. The scRNA-seq reads were mapped non-uniquely to microbial genomes using GATK PathSeq\textsuperscript{63} after filtering the reads aligned to the host genome and spike-in transcripts. The presence of a particular microbial taxon in each cell was considered one if \( \geq 3 \) UMIs were assigned to the relevant genome(s) using PathSeq\textsuperscript{63}.

**RNAscope-in situ hybridization**

RNAscope-in situ hybridization for mRNA expression was performed on tumor and paired normal tissues using an RNAscope 2.5 LS Reagent Brown Kit (catalog number 322370; Advanced Cell Diagnostics, Newark, CA, USA) according to the manufacturer's instructions. A total of 3-mm thick sections were cut from formalin-fixed, paraffin-embedded tissue samples. Staining was performed using a Leica Bond Rx
Autostainer (Leica, Wetzlar, Germany). All probes used for RNA in situ hybridization were obtained from Advanced Cell Diagnostics. RNAscope™ Probe- EB-16S-rRNA was applied.

Declarations

Competing of interest statement: The authors declare no conflicts of interest.

Acknowledgements: This study was supported by the Bio & Medical Technology Development Program of the National Research Foundation, funded by the Ministry of Science and ICT (grant no. NRF2017M3A9A7050803). We would like to express our sincere gratitude to all patients and their families for their contributions to this study.


Consent for publication: All data were reported in an anonymized, aggregated manner.

Patient data: Records were collected from electronic medical records and reviewed. This study was conducted in accordance with the Declaration of Helsinki and was approved by the Institutional Review Board (IRB) of Samsung Medical Center (IRB number:2017-01-131), Asan Medical Center (IRB number:2017-1350), Seoul National University Hospital (IRB number:2103-121-1206), Bundang Seoul National University Hospital (IRB number: B-1709-423-306), Uijeongbu St. Mary’s Hospital Institutional Review Board (IRB number: XC17TNDI0068), Yonsei University (IRB number:7001988-201910-BR-727-02).

References

25. Wu J, Li Q, Fu X: Fusobacterium nucleatum Contributes to the Carcinogenesis of Colorectal Cancer by Inducing Inflammation and Suppressing Host Immunity. Transla Oncol 12:846-851, 2019


60. Kazakova E, Rakina M, Sudarskikh T, et al: Angiogenesis regulators S100A4, SPARC and SPP1 correlate with macrophage infiltration and are prognostic biomarkers in colon and rectal cancers. Front Oncol 13, 2023

77. Young MD, Behjati S: SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. Gigascience 9, 2020


Figures
Figure 1

(a) Study schematics. DEG, differentially expressed gene; scRNA-seq, single-cell RNA sequencing; ISH, in situ hybridization; H&E, hematoxylin and eosin stain

(b) Summary of the cohorts. SMC, Samsung Medical Center; SNUH, Seoul National University Hospital; SNUBH, Bundang Seoul National University Hospital; AMC, Asan Medical Center, USMH, Uijeongbu St. Mary's Hospital. Alluvial plot presenting various

---

Study outline

(a) Study schematics. DEG, differentially expressed gene; scRNA-seq, single-cell RNA sequencing; ISH, in situ hybridization; H&E, hematoxylin and eosin stain

(b) Summary of the cohorts. SMC, Samsung Medical Center; SNUH, Seoul National University Hospital; SNUBH, Bundang Seoul National University Hospital; AMC, Asan Medical Center, USMH, Uijeongbu St. Mary's Hospital. Alluvial plot presenting various
consensus molecular subtype (CMS) (c) and tumor microenvironment (TME) subtypes (d) according to the normal tissue-based classification (tNME and hNME). IE, Immune-enriched, non-fibrotic; IE/F, Immune-enriched, fibrotic; F, Fibrotic; D, Desert (e) Clinical metadata of the study samples.

**Figure 2**

Survival outcomes according to the normal tissue-based classification
(a) Recurrence-free and (b) Overall survival of the study cohort. (c) Forest plot of HR (dot) and 95% CI (arrow) for Recurrence-free survival according to tNME and hNME. (d) Overall survival in the TCGA cohort.

**Figure 3**

Pathway and 16S rRNA metagenomics analyses
Gene set enrichment analysis (GSEA) for tumor (a) and normal (b) tissues of tNME. P-value was adjusted by Benjamini-Hochberg (c) Alpha diversity calculated using Berger-parker and Wilcoxon signed-rank test. (d) Principal coordinate analysis (PCoA) plot representing beta-diversity (Bray-Curtis dissimilarity) of bacterial communities at the genus level from tumor and normal tissues of tNME and hNME, and PERMANOVA analyses. (e) Proportion bar presenting relative abundance of bacterial genera according to tumors and normal tissues of tNME and hNME. The relative abundance suggests the total read count of a specific genus per read counts from total genera. (f) Wilcoxon signed-rank test of abundance of bacterial genus between tumor and normal tissues of tNME and hNME. The dot indicates the read count of a specific genus divided by the total genera read counts.

Figure 4

Single-cell analysis of tumor and paired normal tissues derived from 12 patients

(a–b and d–e) UMAP denote myeloid (a), T (b), stromal and endothelial (d), and epithelial (e) cells. Proportion bars on the right side indicate the comparison of the proportion of cell clusters. The comparison was conducted separately (tumor of tNME vs. tumor of hNME; normal of tNME vs. normal of hNME; Wilcoxon rank-sum test, *P < 0.05). The proportion is colored by cell clusters. The bottom figure
indicates the dot plot of canonical markers. (c) Spearman correlation analysis for co-occurrence of cell cluster frequencies between tumor and normal tissues, derived from tNME and hNME. Positive and negative co-occurrences patterns are in red and blue, respectively.

**Figure 5.**

**Single-cell analysis of tumor tissues derived from 47 patients**
(a–d) The top figure denotes UMAP embedding of myeloid (a), stromal and endothelial (b), T (c), and epithelial (d) cells of tumor tissues. Proportion bars on the right side indicate the comparison of the proportion of cell clusters between tNME and hNME (Wilcoxon rank-sum test, *P < 0.05). The proportion is colored by cell clusters. (a–b) The bottom figure indicates the dot plot of canonical markers. (c–d) The bottom figure indicates the normalized expression of selected marker genes. (e) SCISSOR analysis relating prognostic information from bulk RNA-seq data from 111 SMC patients with scRNA-seq data. The red color bar represents the single-cell fraction related to poor prognosis, and the blue color bar denotes good prognosis. (f) Heatmap illustrating expression of MSigDB Hallmark gene set across epithelial clusters. (g) Enrichment pathway analysis of EMP1^{high} epithelial cells compared to other epithelial cells based on the KEGG pathway database. P-value was adjusted by Benjamini-Hochberg (h) Heatmap of interaction score of ligand and receptor between selected fibroblasts, myeloid cells, epithelial cells, and CD8+T cells. (i) Pie charts indicate the composition of the intracellular bacterial genus in the tumor samples from tNME and hNME. Top 10 most dominant genera in each group were shown in pie chart, and other genera were represented by ‘Others’. Proportion bar on the right upper side presents the comparison of the infected cancer cell proportion inferred by CSI-microbes between tNME and hNME.
Figure 6

Representative cases with histologic pattern and spatial distribution of eubacteria

(a–f) Representative cases of tNME (a–c) and hNME (d–f). (a,d) RNAscope-in situ hybridization (RNAscope-ISH) images indicate the spatial distribution of eubacteria across the tumor and normal tissue from the surgical specimen (brown color indicates eubacterial probe) with higher magnification (x4) of
the bacterial colonization area. (b,e) H&E-stained slides with higher magnification (x8) of the lesion that inflammatory cells infiltration. (c,f) Pie charts show the composition of the bacterial genera derived from the 16S rRNA gene sequencing. (g) Graphical model presenting normal tissue-based classification. (h) Putative biological mechanism underlying vicious cycle in the tNME.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- NATSuppleTablever1.2.docx
- Supple1.tif
- Supple2.tif
- Supple3.tif
- Supple4.tif
- Supple5.tif
- Supple6.tif