

Tumor microenvironment interplay amid microbial community, host gene expression and pathological features elucidates cancer heterogeneity and prognosis risk

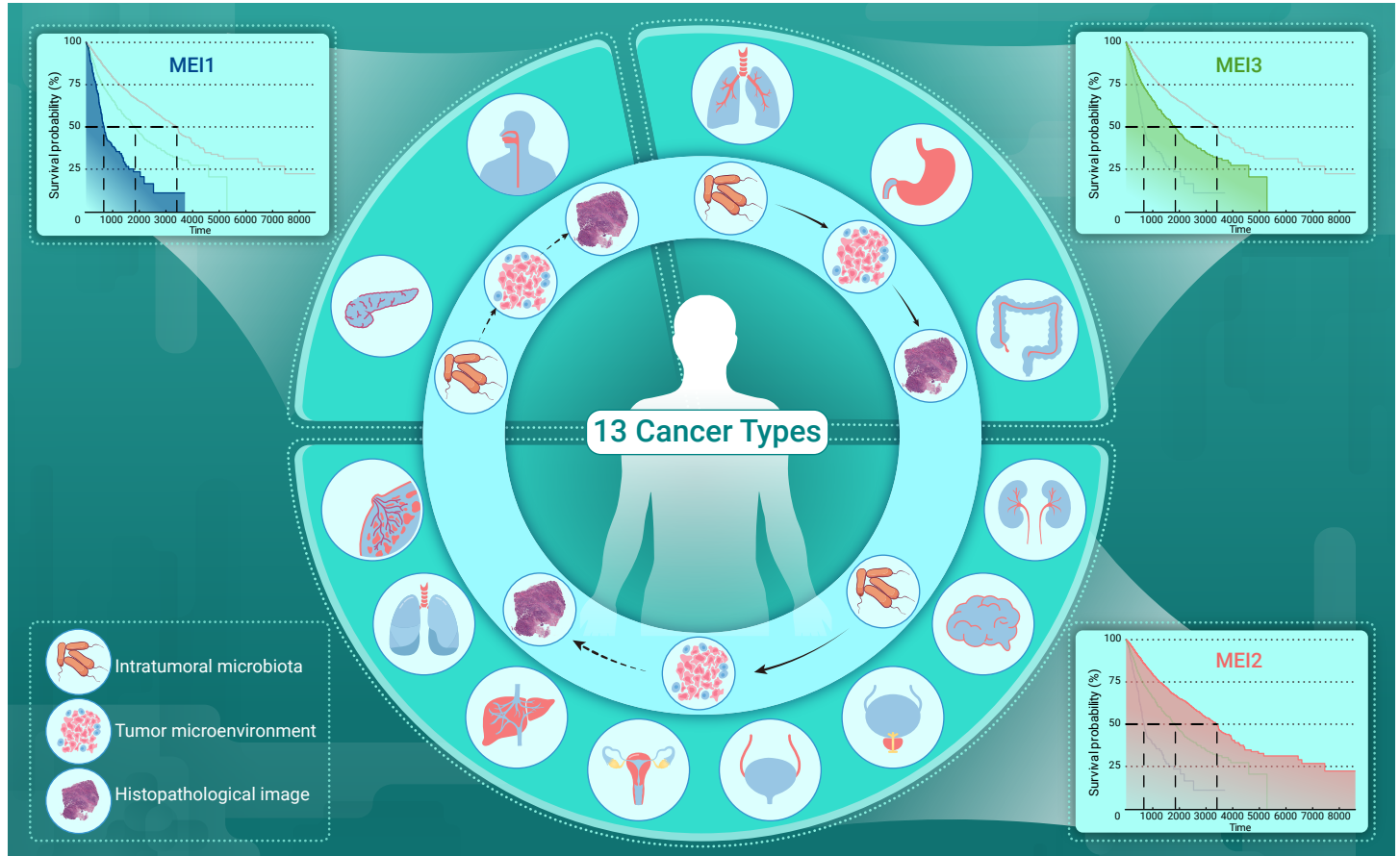
Lei Ji,^{1,2,3} Haohong Zhang,² Geng Tian,^{1,3} Shuxue Xi,^{1,3} Yuwen Chu,^{1,3,4} Yumeng Zhang,^{1,3,5} Jinyang Liu,^{1,3} Kang Ning,^{2,*} and Jialiang Yang^{1,3,*}

*Correspondence: ningkang@hust.edu.cn (K.N.); yangjl@geneis.cn (J.Y.)

Received: July 19, 2023; Accepted: September 12, 2023; Published Online: September 13, 2023; <https://doi.org/10.59717/j.xinn-life.2023.100028>

© 2023 The Author(s). This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- There is a link between intratumoral microbiota, host, and histopathology images in cancer patients.
- Patterns of association between intratumoral microbiota, host, and histopathology images exhibit tumor heterogeneity.
- Associations between intratumoral microbiota, host, and histopathology images underlie the prognostic risk of cancers.



Tumor microenvironment interplay amid microbial community, host gene expression and pathological features elucidates cancer heterogeneity and prognosis risk

Lei Ji,^{1,2,3} Haohong Zhang,² Geng Tian,^{1,3} Shuxue Xi,^{1,3,4} Yuwen Chu,^{1,3,4} Yumeng Zhang,^{1,3,5} Jinyang Liu,^{1,3} Kang Ning,^{2,*} and Jialiang Yang^{1,3,*}

¹Geneis Beijing Co., Ltd., Beijing 100102, China

²Key Laboratory of Molecular Biophysics of the Ministry of Education, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

³Qingdao Geneis Institute of Big Data Mining and Precision Medicine, Qingdao 266000, China

⁴School of Electrical & Information Engineering, Anhui University of Technology, Maanshan 243002, China

⁵School of Mathematics and Statistics, Hainan Normal University, Haikou 571158, China

*Correspondence: ningkang@hust.edu.cn (K.N.); yangjl@geneis.cn (J.Y.)

Received: July 19, 2023; Accepted: September 12, 2023; Published Online: September 13, 2023; <https://doi.org/10.59717/j.xinn-life.2023.100028>

© 2023 The Author(s). This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Citation: Ji L., Zhang H., Tian G., et al., (2023). Tumor microenvironment interplay amid microbial community, host gene expression and pathological features elucidates cancer heterogeneity and prognosis risk. *The Innovation Life* 1(2), 100028.

Microbial community, host gene expression, and pathological image in the tumor microenvironment (TME) are interrelated, and they can provide valuable insights into tumor heterogeneity. However, the complex and important interconnection patterns among these multimodal omics in pan-cancer remain far from being elucidated. Here, we developed a framework for decoding the association among the microbial community, host gene expression, and pathological image in the TME for a broad spectrum of cancers in multiple independent cohorts. Our results indicated three types of microbe-expression-image triplet (MEI-triplet) interplay patterns in human cancers. MEI1, including esophageal cancer (ESCA) and pancreatic adenocarcinoma (PAAD), was characterized by no significant intratumoral microbe-host interaction. MEI2, which included eight cancers (liver hepatocellular carcinoma [LIHC] is one of these), was characterized by a strong association of tumor microbiota and host gene expression, but changes in the TME were not reflected in pathological images. Meanwhile, MEI3, including colorectal cancer (CRC), stomach adenocarcinoma (STAD), and lung squamous cell carcinoma (LUSC), was characterized by strong associations among tumor microbiota, host gene expression, and pathological images. Further experiments with validation datasets for several representative cancer types also supported these patterns. Out of the three MEI patterns, patients with MEI1 had the worst prognosis risk, followed by those with MEI3 and MEI2. Our findings provide new insights into tumor heterogeneity, which could assist with providing more precise diagnosis and prognosis assessment in clinics.

INTRODUCTION

The microbial community, host gene expression, and pathological image in the tumor-microenvironment (TME) have all been found to be associated with the onset and progress of cancer.¹⁻³ Numerous studies performed over the past decade have elucidated tumor heterogeneity using multi-modal omics technology, including at the microbial, transcriptional, and phenotypic levels, which are inter-related.⁴⁻⁷ Studies have observed correlations between microbes and host gene expression in the TME of several types of cancers.^{8,9} The direct alteration of mutagenesis, the modulation of carcinogenic pathways, and the regulation of the immune system provide explanations for the association between local microbes and cancer.¹⁰⁻¹² Additionally, several studies have applied computational image analysis of tissue samples to classify stroma, immune, and tumor epithelial cells.^{13,14} However, at the pan-cancer level, the association among the microbial community, host gene expression, and pathological images remains unknown.

In this work, we took advantage of the comprehensiveness of The Cancer Genome Atlas (TCGA) database, which covers large-scale tumor samples of numerous cancer types, and multi-omics data, including numerous genes, to perform a comprehensive pan-cancer level study of the microbe-expression-image interplay in human tumors. We classified each cancer based on the clustering of color and morphological features of the hematoxylin and eosin (H&E)-stained slides extracted in a fully automated process. Next, we compared the differences and biological associations between subtypes in

the tumor microbiome, TME, and histological features. Finally, the results of these analyses were compared across 13 cancer types to reveal the tumor heterogeneity from the perspective of the interrelationship between the microbial community, host gene expression, and pathological images.

MATERIALS AND METHODS

Study population and data collection

A total of 6,095 patients with cancer from TCGA database, covering 13 cancer types, were included in this study. The microbial abundance data of the tumor tissues for various cancer types were downloaded from Rob Knight's article.¹⁵ Rob Knight's team developed a Kraken TCGA microbial-detection pipeline that used a ultrafast Kraken algorithm to map sequence readings not aligning to the human reference genome to known bacterial, viral, and archaea microbial genomes.¹⁶ The cancer microbiome data used in this study and the metadata were available at ftp://ftp.microbio.me/pub/cancer_microbiome_analysis/. The gene expression quantification by RNA-Seq, as well as the histopathological slides for each cancer patient were downloaded from <https://portal.gdc.cancer.gov/>. Data were analyzed with the R (version 3.4.0) and R Bioconductor packages.

Validation dataset and analysis

Tissue samples from 32 Chinese patients with liver cancer were collected to independently verify tumor microbiome and host interactions. The raw RNA-Seq sequencing data (PRJNA576155) were downloaded from the Sequence Read Archive (SRA). For taxonomic assignment, Kraken2 was used to obtain the microbial abundance data with the standard Kraken2 database. To verify the results of LIHC in TCGA, species matching the results of Poore et al was used for subsequent analysis.¹⁵ For gene expression quantification, HISAT2 was used for RNA-Seq alignment,¹⁷ with hg19 as the reference genome. SAM-tools was used for data conversion and sorting of the alignment results.¹⁸ Finally, the gene expression was quantified by StringTie,¹⁹ with hg19 as the reference transcriptome, and the Fragments Per Kilobase Million (FPKM) values were extracted as the expression levels. In addition, the AC-ICAM cohort,²⁰ comprising 246 patients with colon adenocarcinoma (COAD), was used to validate the tissue microbiome-host interaction patterns in COAD. Clinical metadata and gene expression data were obtained from cBioPortal. The tissue microbiome abundances were downloaded from FigShare.

To independently verify the difference in the TME between the classifications based on image feature clustering, the pathological images and matched gene expression data of 104 patients with COAD were downloaded from <https://www.cancerimagingarchive.net/>.

Feature extraction of pathological images

For each whole slide image (WSI), our pipeline started with an automatic partition of the area of tissues. WSI was read into memory at a downsampled resolution (e.g., 32x), converted from RGB to HSV color space. A binary mask of the tissue area (foreground) was calculated based on the threshold of the saturated channel of the image after median blurring to smooth the edges, followed by additional morphological closure to fill in small gaps and

holes. Subsequently, the detected approximate contour of the foreground object was filtered and stored for downstream processing based on the region threshold, while making the segmentation mask of each slide available for optional visual inspection (see the article by Lu et al. for more details.²¹). Next, the color features on four components R, G, B, and grayscale for each WSI were calculated, including the mean, variance, and skewness, to describe the color characteristics. Additionally, Hu's invariant moments on the four components R, G, B, and grayscale of each image were extracted by cv2.HuMoments provided by OpenCV.²² Hu introduced a set of seven orthogonal image moments that are insensitive to translation, scaling, and rotation (see the article by Hu for more details on the characteristics of Hu moments.²²).

Clustering analysis of pathological image features

All of the pathological images were clustered based on the pathological image features using partitioning around medoids (PAM) clustering with the "cluster" package in R. We calculated the silhouette width when the clustering number was 1 to 10. The silhouette width ranges from -1 to +1, where higher values indicate that a point matches well with its own cluster and poorly with its neighboring clusters. If most of the points have high silhouette width (high average silhouette width), the current clustering method is appropriate, whereas if many points have low or negative values (low average silhouette width), the current clustering method is considered unsuitable. The optimal PAM clustering number based on pathological image features was the clustering number when the highest average silhouette width calculated with package "fpc" in R was obtained.

Procrustes analysis

The correlation between tumor microbes and host gene expression was verified by Procrustes analysis. Jaccard dissimilarities among samples were calculated using the R function "vegdist" in the "vegan" package based on the onco-microbial community at the genus level and host gene expression, respectively. The R function "procrustes" in the "vegan" package was used to conduct the Procrustes analysis. The correlation, significance, and Procrustes sum of squares were calculated by the R function "protest" in the "vegan" package, with 999 permutations. Procrustes correlations between onco-microbial community and host were visualized using the "ggplot2" package.

Mantel test

The Mantel test was used to examine the correlation between the tumor microbial community and host gene expression with the "vegan" package in R. Jaccard dissimilarities were calculated using the R function "vegdist" based on the tumor microbial community and host gene expression, respectively. Then, the Spearman correlation coefficient and p-value were calculated by the R function "Mantel", with 999 permutations.

Microbial diversity analysis

The richness and Shannon indices were used to characterize the α -diversity of the tumor microbial community. The Shannon index was calculated using the R function "diversity" in the "vegan" package. Wilcoxon rank-sum test was used to test the differences in α -diversity between groups, while β -diversity was represented by measurements of the Bray-Curtis distance. Permutational multivariate analysis of variance (PERMANOVA) was used to examine differences in beta-diversity with the "vegan" package in R.

TME characterized by stromal score and immune score

The stromal score and immune score of all the samples in TCGA were available from <https://bioinformatics.mdanderson.org/estimate/>. ESTIMATE is a method that uses gene expression characteristics to infer the ratio of stromal cells and immune cells in a tumor sample.²³ In addition, the abundance of eight immune cell types, namely, CD3+ T cells, CD8+ T cells, cytotoxic lymphocytes, natural killer (NK) cells, B lymphocytes, cells originating from monocytes (monocytic lineage), myeloid dendritic cells, neutrophils, and two stromal cells (endothelial cells and fibroblasts), were calculated using the R function "MCPcounter.estimate" in the "MCPcounter" package.²⁴

Linear discriminant analysis Effect Size (LEfSe) analysis, correlation analysis, and survival analysis

LEfSe analysis was performed to identify the significantly different bacteria in terms of abundance between groups. The threshold value of the linear discriminant analysis (LDA) score was 2 when identifying the differential bacteria of the two groups obtained by clustering the image features of a single cancer. Correlation analysis was performed for the differential microbes and stromal scores using the R function "corr.test" in the "psych" package, with $p < 0.05$ as the significance threshold. Kaplan-Meier (KM) analysis was performed to plot the survival curves with the "survminer" and "survival" packages in R, and the significance was determined by log-rank test.

RESULTS

Tumor microbiome profiles of 13 cancer types

Our cohort consisted of 6,095 tumor samples covering 13 types of cancer (Figure 1A). A matching H&E-stained slide sample, tissue microbiome sample, and host transcriptome sample were available for each patient. *Pseudomonas* was the genus with the highest relative abundance in 10 tumor tissues, while *Mycobacterium* was the dominant bacterium of colorectal cancer (CRC), stomach adenocarcinoma (STAD), and uterine corpus endometrial carcinoma (UCEC) (Figure 1B). Notably, *Xanthomonas* was overabundant in liver tumor tissue. Additionally, we found that STAD and esophageal carcinoma (ESCA) tumors had the highest microbial richness and the lowest Shannon index (Figure 1C). Tumors with high microbial richness included UCEC and breast cancer (BRCA). Consistent with previous findings,²⁵ we observed significant differences in the tissue microbial communities of the 13 cancer types (Figure 1D; $R^2 = 21\%$, $p = 0.001$).

Two patterns of tissue microbiome-host interactions in the 13 cancer types

Our results revealed two patterns of tissue microbiome-host interactions in the 13 cancer types (Figure 1E). Procrustes analysis showed a significant correlation between microbiome and host expression in 11 of the 13 cancer types ($p < 0.05$). For ESCA and pancreatic adenocarcinoma (PAAD), we found no significant microbe-host interplay pattern ($p > 0.05$). The Mantel test was used to verify the robustness of the results. Consistent with the results of the Procrustes analysis, ESCA and PAAD showed weak correlation between tissue microbe and host (Figure 1F). Consequently, from the perspective of the microbe-host interaction, the 13 cancer types could be divided into two categories, representing strong and weak microbe-host association patterns.

Next, two non-western cohorts were collected to validate the tissue microbiome-host interaction patterns. First, the AC-ICAM cohort, consisting of 246 Qatari patients with COAD, was used to validate the microbe-host interaction model of COAD. Procrustes analysis showed a significant correlation between tissue microbes and host gene expression (Figure S1A). Another independent cohort of 32 Chinese patients with liver cancer also showed the same results with TCGA (Figure S1B). Consequently, two tissue microbiome-host interaction patterns presented by TCGA with independent cohorts demonstrated generalization ability.

Stratification of cancer types based on the interaction among the microbial community, host gene expression, and pathological image in the TME

Next, we aimed to explore whether tissue microbe-host interactions manifest themselves in histology slides. Here, a fully automatic process, involving section standardization, tissue contour recognition, and feature extraction, was employed to extract color and texture features from the histology slides (Figure 2). Additionally, we calculated the stromal and immune score, which quantified the TME by the ESTIMATE algorithm based on the mRNA expression data.²³ First, for each cancer, we performed the PAM clustering for pathological image samples. The optimal clustering number k was obtained based on the silhouette width. Then, we compared the differences in tumor microbes and the TME between the k groups.

Using this methodology, these 13 cancers were further categorized into the following three subtypes: MEI1, MEI2, and MEI3. MEI1, including PAAD and ESCA, was characterized by no significant microbe-host interaction; MEI2,

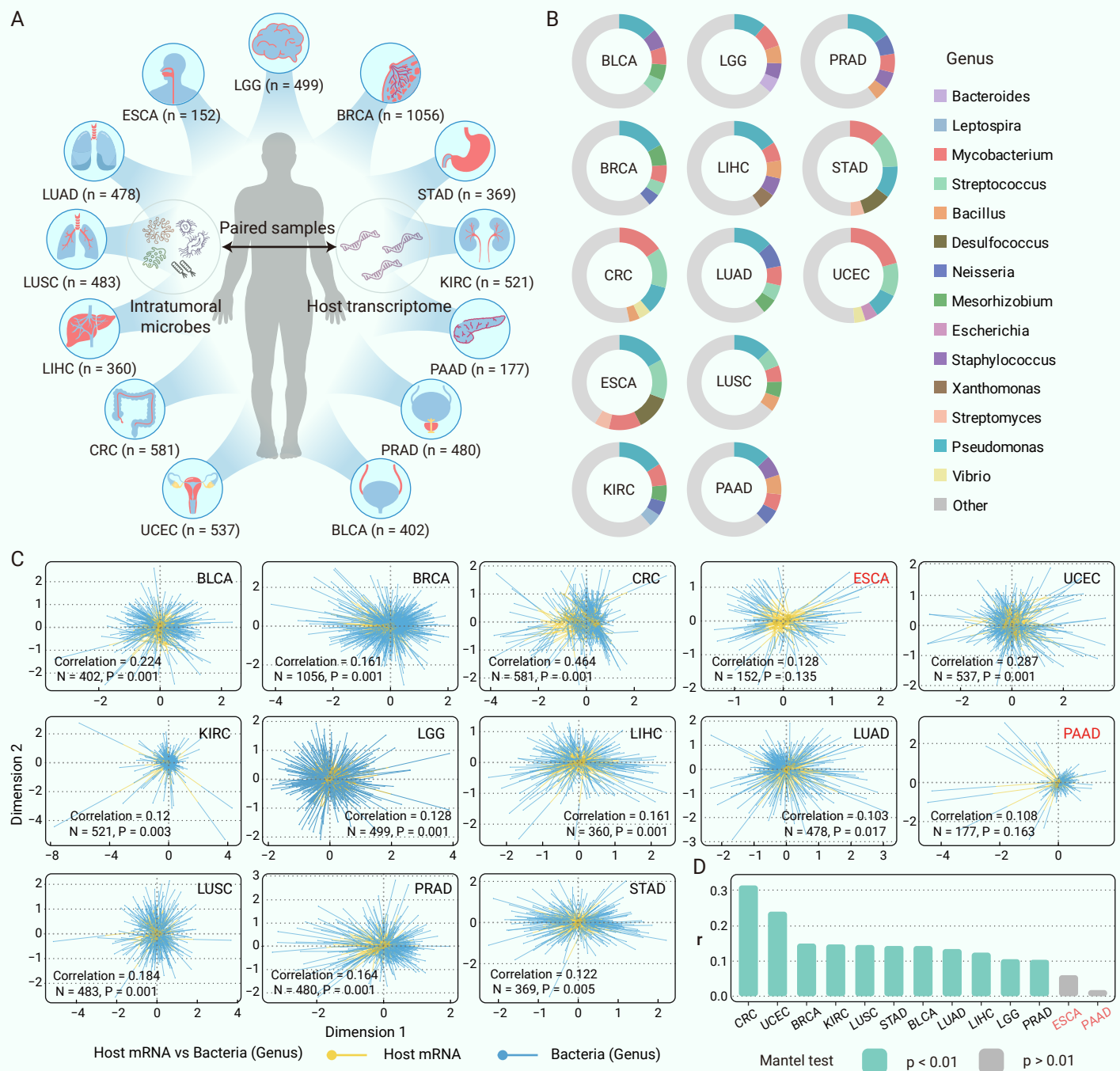


Figure 1. Tumor microbial profiles of the 13 cancer types and their host interactions (A) Matched tumor microbiota and host transcriptome sample sizes for the 13 cancer types. (B) Pie chart showing the top five bacteria at the genus level in relative abundance for each cancer. (C) Procrustes analysis of the association between the microbial community and that of the host gene expression in each cancer. The correlation coefficient and p-values were calculated by the "protest" function, where n represents the sample size. (D) Histogram showing the correlation coefficients of the tumor microbiota and host gene expression calculated by the Mantel test for 13 cancer types. The red column indicates a p-value < 0.01, and the gray column indicates a p-value > 0.01.

including BRCA, lung adenocarcinoma (LUAD), liver hepatocellular carcinoma (LIHC), kidney renal clear cell carcinoma (KIRC), prostate adenocarcinoma (PRAD), UCEC, bladder cancer (BLCA), and lower-grade glioma (LGG), was characterized by the fact that the tumor microbes significantly affected the TME, but the changes in the TME were not reflected in the histology slides; and MEI3, including STAD, CRC, and lung squamous cell carcinoma (LUSC), was characterized by the fact that the tumor microbes significantly influenced the TME, which was further reflected in the histology slides.

Association patterns in STAD

STAD was the first cancer found to belong to the MEI3 pattern. PAM clustering divided patients with STAD into two clusters (STAD-T1 and STAD-T2)

(Figures 3A–C). STAD-T1 and STAD-T2 showed significant differences in microbial community in tumor (Figure 3D, $p = 0.001$). The microbial richness of STAD-T2 was significantly higher than that of STAD-T1 (Figure 3E, $p < 0.001$). Moreover, the abundance of Bacteroidetes in STAD-T2 was significantly higher than that in STAD-T1 (Figure 3F, $p < 0.05$), while that of Firmicutes was the opposite ($p < 0.001$). In addition, we found that the Firmicutes/Bacteroidetes (F/B) ratio of STAD-T1 was significantly lower than that of STAD-T2 ($p < 0.01$). Next, we identified the differentially enriched bacteria between STAD-T1 and STAD-T2 (Figure 3G). The relative abundance of *Bordetella* and *Staphylococcus* was increased in STAD-T1, while the genus *Streptococcus* was more abundant in STAD-T2.

At the TME level, no significant difference was found between STAD-T1

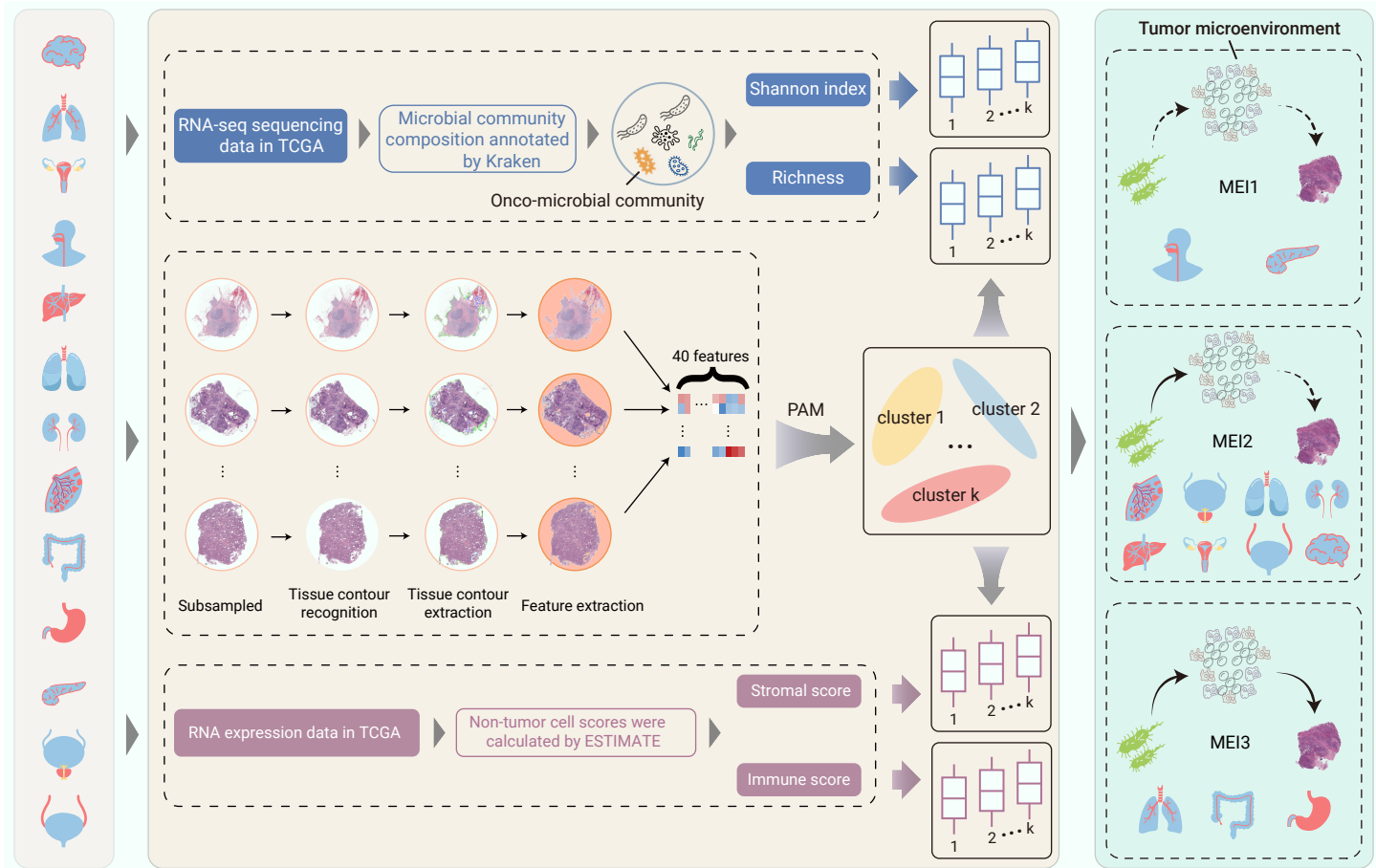


Figure 2. Schematic overview of the workflow employed to generate the microbe-expression-image triplet (MEI-triplet) interplay pattern classification platform Microbial abundance data were obtained by Kraken's annotation of TCGA RNA-seq sequencing data. The tumor microenvironment characterized by the stromal score and immune score, was calculated using the ESTIMATE algorithm based on TCGA mRNA expression data. The tissue regions of pathological images were identified and the features were extracted. Then, PAM clustering was conducted based on the extracted features. The 13 cancers were subdivided into three MEI patterns by comparing the characteristics of the TME and the microbial characteristics among the clusters obtained by PAM clustering (please see the Results for further details).

and STAD-T2 in terms of the immune score (Figure S2, $p = 0.41$), but STAD-T1 was significantly higher than STAD-T2 in terms of the stromal score (Figure 3H, $p < 0.01$). The abundance of endothelial cells, a type of stromal cell, was significantly higher in STAD-T1 than that in STAD-T2 ($p < 0.05$). Considering the interactions between host cells and tumor microbes,^{26, 27} Spearman correlation analysis was performed for stromal score and differential bacteria (Figure 3J). The genera *Bordetella*, *Staphylococcus*, and *Acidovorax* were significantly positively correlated with stromal score ($p < 0.01$), while these three genera were enriched in STAD-T1. Additionally, *Pseudothermotoga* and *Denitrovibrio* enriched in STAD-T2 had a significantly negative correlation with stromal score ($p < 0.001$). Thus, the finding that the stromal score of STAD-T1 was significantly higher than that of STAD-T2 confirmed the role of tumor microbes on non-cancerous cells in the TME.

Association patterns in LUSC

Similarly, patients with LUSC, also belonging to the MEI3 pattern, were categorized into two clusters (Figures 4A–C). The tumor microbiome was significantly different between the two clusters (Figure 4D, $p = 0.001$), while the microbial richness of LUSC-T1 was significantly higher than that of LUSC-T2 (Figure 4E, $p < 0.0001$). A total of 30 significantly different bacteria were identified, among which, 14 were significantly enriched in LUSC-T1 and 16 were overabundant in LUSC-T2 (Figure 4F). The genera *Pseudomonas*, *Shewanella*, and *Bordetella* were overabundant in LUSC-T1, while *Neisseria*, *Mesorhizobium*, and *Escherichia* were enriched in LUSC-T2. Moreover, we found that patients in LUSC-T1 had significantly poorer survival than those in LUSC-T2 (Figure 4G, $p = 0.031$).

At the TME level, the stromal score of LUSC-T1 was significantly higher than that of LUSC-T2 (Figure 4H, $p < 0.05$). The relative abundances of

Bordetella, *Pseudomonas*, and *Acidovorax*, all of which were highly enriched in LUSC-T2, were significantly positively correlated with stromal score (Figure 4I, $p < 0.01$). Additionally, bacteria with a significantly negative correlation with stromal score were overabundant in LUSC-T2, such as the genera *Sanguibacteroides* and *Waddlia* ($p < 0.001$). Consequently, for LUSC, variation in the TME affected by tumor microbes was further reflected in the pathological images and was related to patient prognosis.

Association patterns in CRC

As with STAD and LUSC, PAM clustering divided patients with CRC (the final cancer belonging to MEI3) into two clusters (Figures 5A–C). The tumor microbial community was significantly different between the two clusters (Figure 5D, $p < 0.001$). Specifically, the microbial richness of CRC-T1 was significantly higher than that of CRC-T2 (Figure 5E, $p < 0.001$), while the Shannon index was higher in CRC-T2 (Figure 5F, $p < 0.0001$). A total of 117 differential bacteria at the genus level were identified (Figure 5G, Supplementary Table S1), among which, 43 were enriched in CRC-T1 and 74 were overabundant in CRC-T2. Additionally, the stromal score of CRC-T2 was significantly higher than that of CRC-T1 (Figure 5H, $p < 0.05$). Specifically, the abundances of endothelial cells ($p < 0.01$) and fibroblasts ($p < 0.05$) were significantly higher in CRC-T2 than in CRC-T1 (Figure 5I). Spearman correlation analysis showed that the bacteria that were significantly positively associated with stromal score were all enriched in CRC-T2 (Figure 5G, $p < 0.01$). Finally, there was no statistical significance (Figure 5J, $p = 0.42$), but the survival of patients in CRC-T2 was worse than that of patients in CRC-T1.

Independent validation of MEI3 association pattern with matching images and host expression samples for COAD

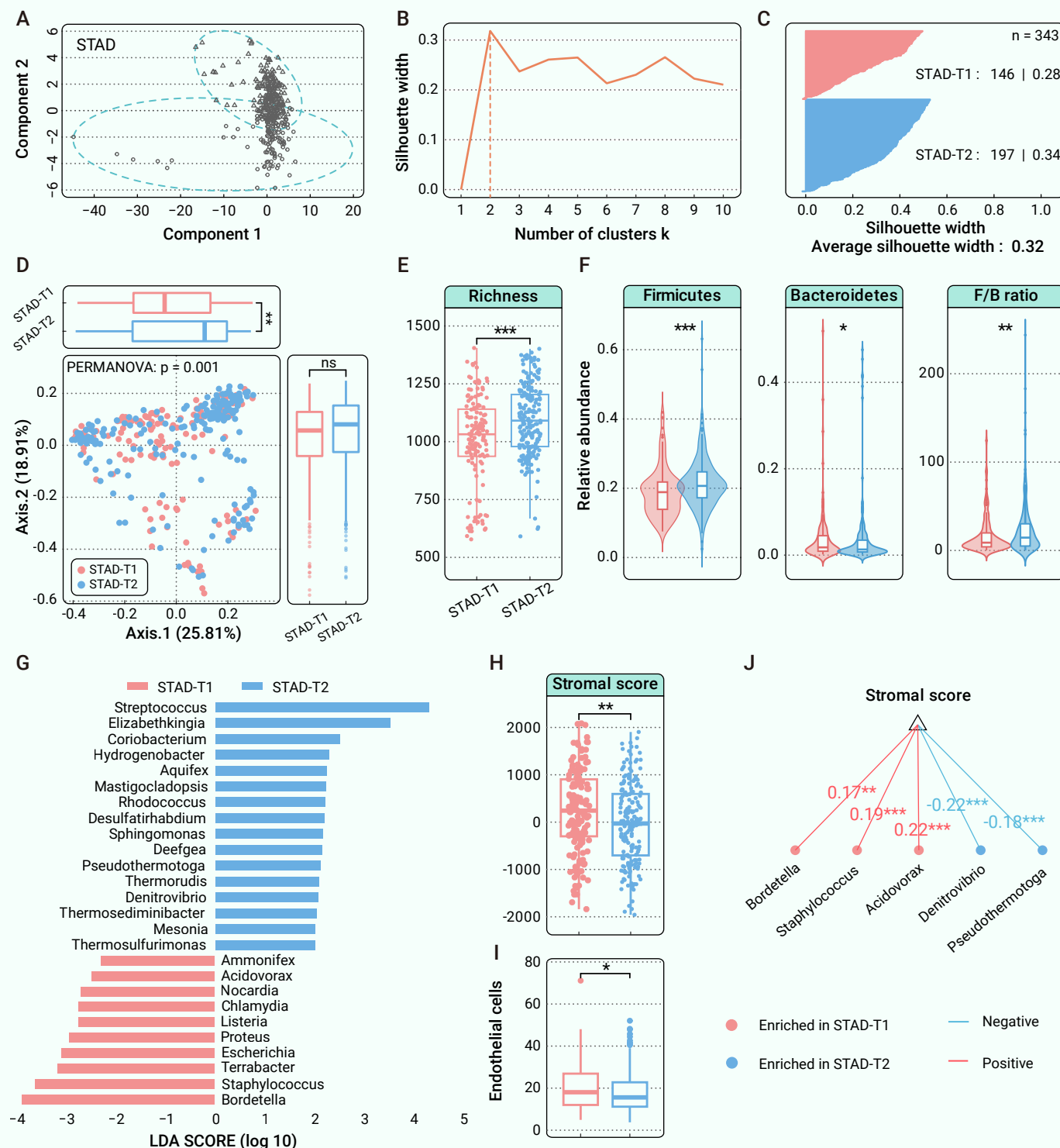


Figure 3. Differences and associations of STAD subtypes at the level of tumor microbes and tumor microenvironment (TME) based on pathological image features (A) Two subtypes of STAD obtained by PAM clustering based on image features. (B) Silhouette width with clustering numbers of 1 to 10. (C) The silhouette width of each sample in the two STAD subtypes and the average silhouette width of the two subtypes. (D) PCoA showing the difference in the tumor microbiota between STAD-T1 and STAD-T2. The p-value was generated by a PERMANOVA test. (E) Boxplot showing the difference in tumor microbial richness between the two subtypes. ***p-value < 0.001. (F) Boxplots showing the differences in the relative abundance of Firmicutes, Bacteroidetes, and the Firmicutes/Bacteroidetes ratio between the two subtypes. (G) Significantly different bacteria identified by LEfSe between the two clusters; an LDA score of 2 was the threshold. (H) Boxplot showing the differences in stromal score between the two subtypes. (I) Boxplot showing the differences in the abundance of endothelial cells between the two subtypes. (J) Spearman correlation between significantly different bacteria in terms of abundance and stromal score. **p-value < 0.01. The number represents the correlation coefficient, and the red and blue lines indicate positive and negative correlations, respectively.

To independently verify the MEI3 association pattern, we obtained 104 patients with COAD with matching images and host expression samples. Similarly, PAM clustering based on image features divided the patients into two groups (Figures S3A–C). Then, we compared the stromal score of these two groups of patients and found a significant difference (Figure S3D, $p <$

0.05). Taken together, our findings were not inherent in TCGA dataset, but were reproducible across-the-board results.

Prognostic potentials of MEI patterns

To determine whether MEI patterns were associated with differences in

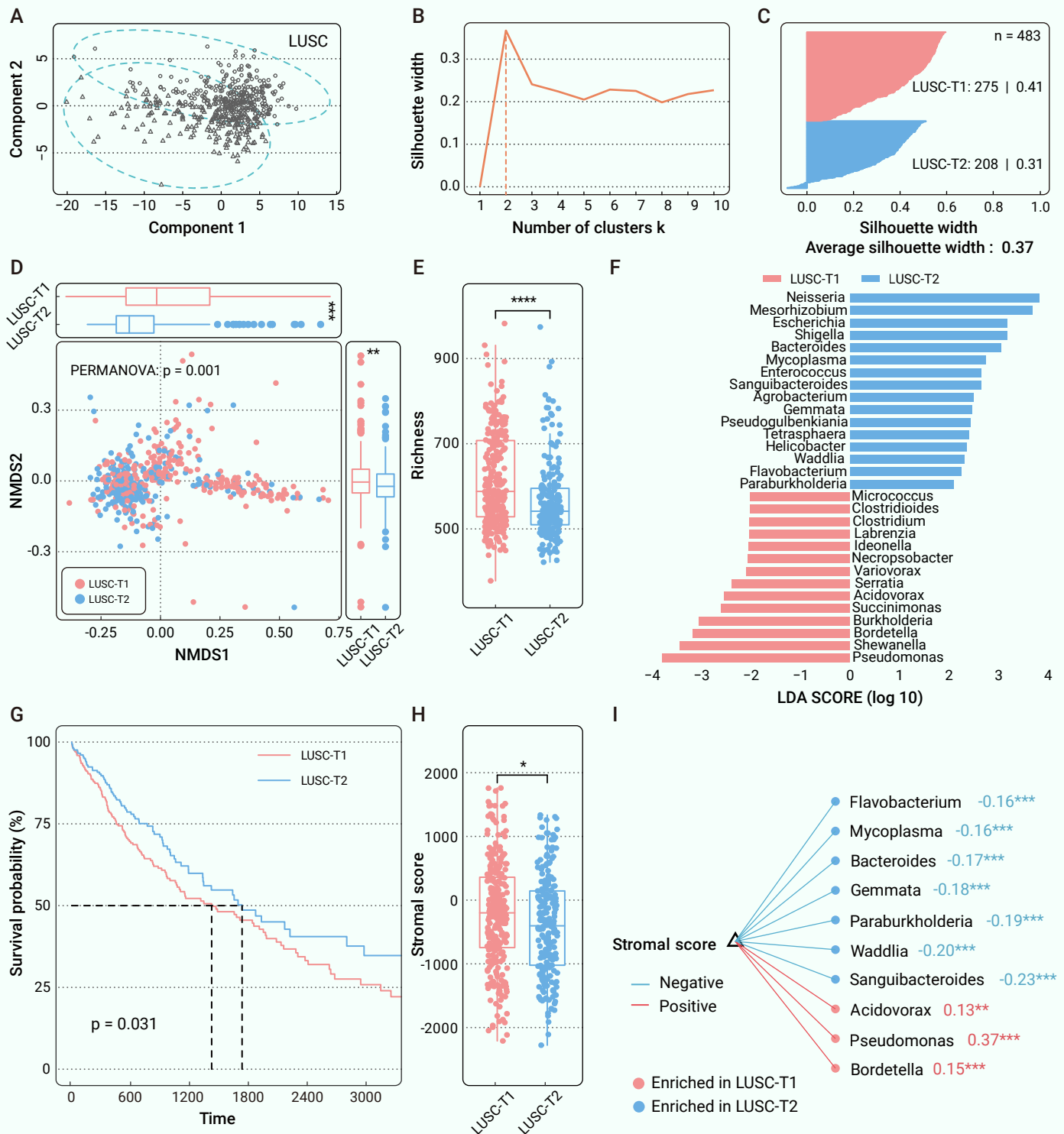


Figure 4. Differences in, and associations of LUSC subtypes at the tumor microbe and TME levels based on pathological image features (A) Two subtypes of LUSC obtained by PAM clustering based on image features. (B) Silhouette width with clustering numbers of 1 to 10. (C) The silhouette width of each sample in the two LUSC subtypes and the average silhouette width of the two subtypes. (D) PCoA showing the difference in tumor microbiota between LUSC-T1 and LUSC-T2. The p-value was generated by a PERMANOVA test. (E) Boxplot showing the difference in tumor microbial richness between LUSC-T1 and LUSC-T2. **** p -value < 0.0001. (F) Significantly different bacteria identified by LefSe between the two clusters; an LDA score of 2 was the threshold. (G) Overall survival of patients in LUSC-T1 and LUSC-T2. The p-value was generated by a log-rank test. (H) Boxplot showing the differences in stromal score between the two subtypes. (I) Spearman correlation between significantly different bacteria in terms of abundance and stromal score; the red and blue lines indicate positive and negative correlations, respectively.

prognosis, we next examined the overall survival of patients with each of these 13 cancer types. First, we plotted survival curves, which showed significant differences in patient survival (Figure 6A, $p < 0.0001$). Specifically, compared to other cancers, patients with PAAD and ESCA had the worst survival, with a median survival of 607 and 801 days, respectively (Figure 6B & Table S2). We then plotted survival curves for the three MEI patterns, which

confirmed the significant differences in survival for the three patterns of cancer (Figure 6C, $p < 0.0001$). MEI1 had the worst prognosis, followed by MEI3, while MEI2 had the best prognosis, with a median survival of 666, 1,849, and 3,409 days, respectively (Figure 6D). These associations with prognosis indicated that the tumor MEI classification had clinical relevance.

We next explored the non-tumor components of the MEIs in more detail.

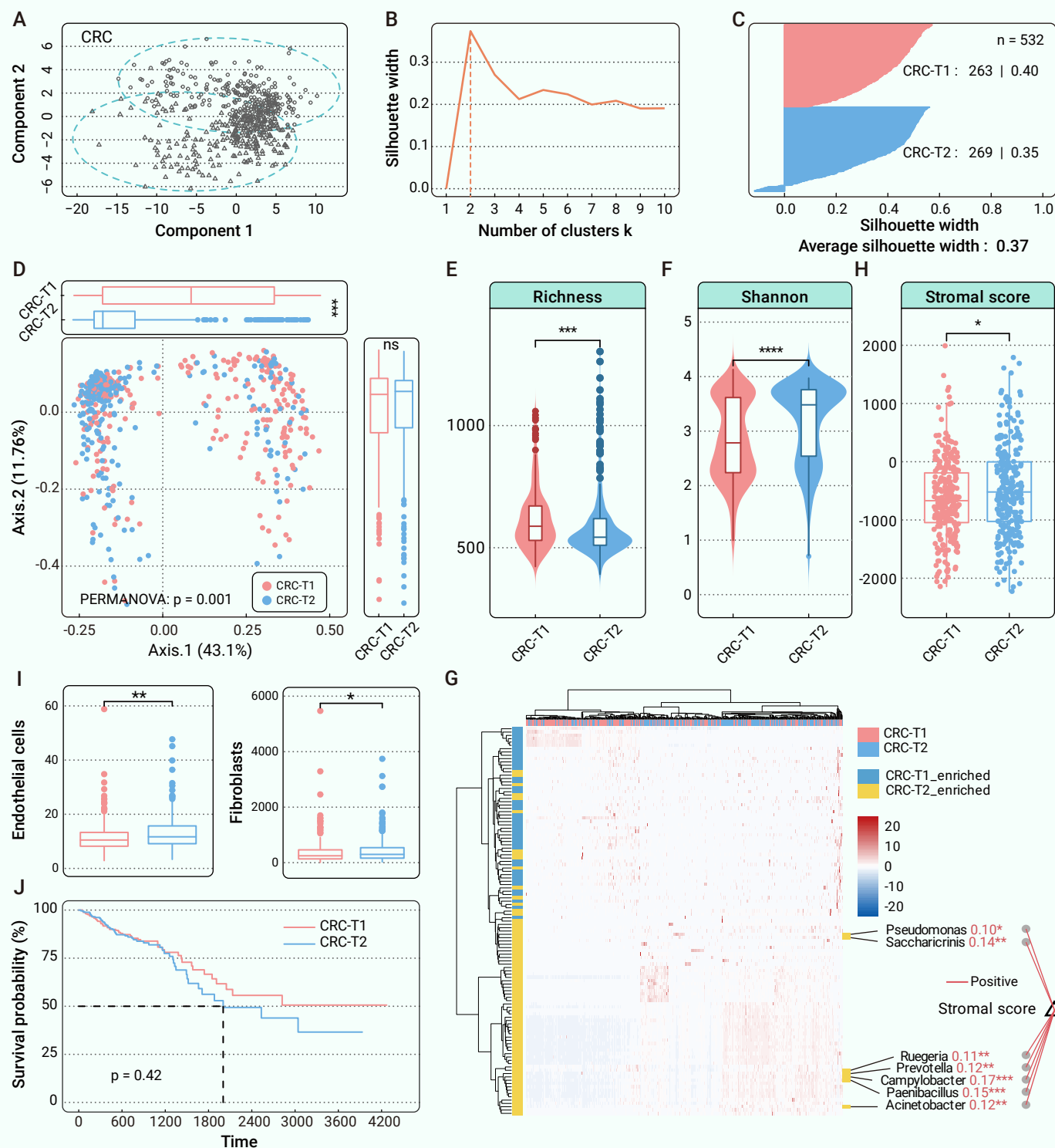


Figure 5. Differences in, and associations of CRC subtypes at the level of the tumor microbes and TME based on pathological image features (A) Two subtypes of CRC obtained by PAM clustering based on image features. (B) Silhouette width with clustering numbers of 1 to 10. (C) The silhouette width of each sample in the two CRC subtypes and the average silhouette width of the two subtypes. (D) PCoA plot showing the difference in tumor microbiota between CRC-T1 and CRC-T2. The p-value was generated by a PERMANOVA test. Boxplot showing the differences in (E) tumor microbial richness and (F) Shannon index between CRC-T1 and CRC-T2. (G) Heatmap of genera with significant differences between the two clusters identified by LEfSe; the red lines indicate a significant positive correlation between these bacteria and stromal scores. *p-value < 0.05, **p-value < 0.01, ***p-value < 0.001. (H) Boxplot showing the differences in stromal score between CRC-T1 and CRC-T2. (I) Boxplot showing the differences in the abundances of endothelial cells and fibroblasts between CRC-T1 and CRC-T2. (J) Overall survival of patients in CRC-T1 and CRC-T2. The p-value was generated by a log-rank test.

We estimated the stromal fraction to be 1 minus tumor purity. To estimate how much of the stromal fraction was due to immune cell infiltration, we plotted the immune score versus the stromal fraction (Figure 6E). In all of the MEIs, the stromal fraction was defined largely by immune cell infiltration. Additionally, MEI2 had the highest median tumor purity (Figure 6F), while MEI1 had the highest stromal score (Figure 6G).

DISCUSSION

Our investigation into the associations between tumor microbes, host gene expression, and pathological images of patients with pan-cancer in multiple independent cohorts revealed a microbe-expression-image-based classification system for pan-cancer that allowed the categorization of tumors into one of three distinct MEI patterns, each with marked differences in microbe-

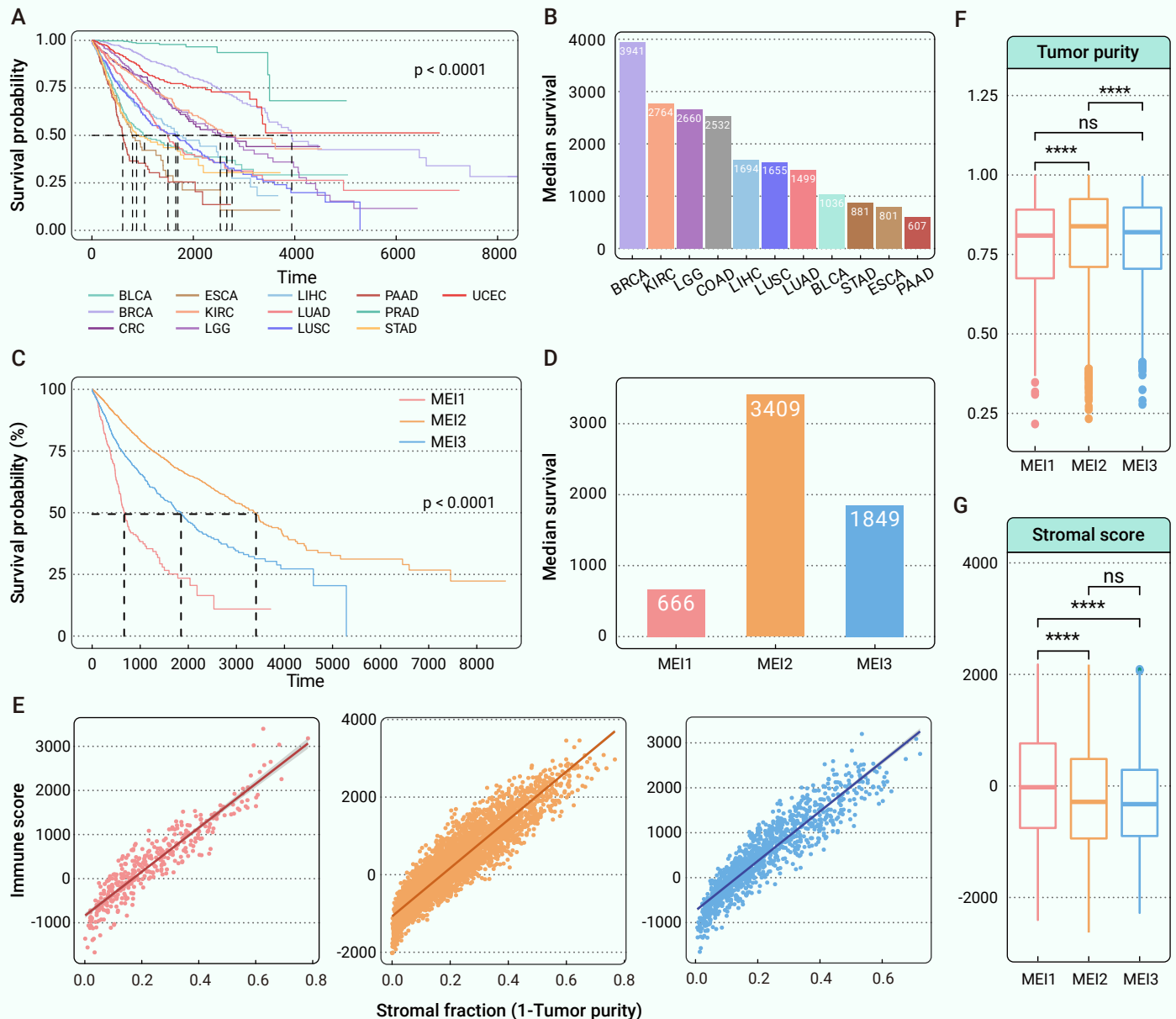


Figure 6. Prognostic potential of the MEI-triplet interplay pattern (A) Overall survival of patients with 13 cancer types. (B) Bar plot showing the median survival of the 13 cancer types. (C) Overall survival of patients with the three MEI-triplet interplay patterns. The p-value was generated by a log-rank test. (D) Bar plot showing the median survival of the three MEI patterns. (E) Linear regression analysis showing the correlation between immune score and stromal fraction among the three MEI patterns. Stromal fraction = 1 - tumor purity. Boxplot showing the differences in the (F) tumor purity and (G) stromal score among the three MEI patterns. Wilcoxon test, ****p value < 0.0001, ns: No significant difference.

expression-image interplay and outcome associations. Moreover, these three patterns could offer actionable knowledge for clinical decision-making and pave the way for precision medicine.

Our results emphasized the importance of understanding the associations among the microbial community, host gene expression, and pathological image in the TME for a deep understanding of inter-tumor heterogeneity, and highlighted the potential of quantifying tumor MEI as a prognostic marker for multiple human cancers. Hoadley et al. found that only one-third of tumors exhibited homogeneity, while the remaining two-thirds exhibited varying degrees of heterogeneity.²⁸ It is worth noting that while both cancers in MEI1 belonged to the digestive system, a mixture of the digestive, respiratory, and reproductive system was present in MEI2, while the digestive and respiratory systems were also included in MEI3. Our results suggested that tumors originating from different systems could have many commonalities at the multimodal association level, whereas those originating from the same system could have distinct association maps. We found associations between MEI patterns and survival in 13 cancers, and the tumor MEI could be further

applied to the classification of other solid tumors to verify its prognostic association. Despite in-depth research and new treatment strategies, most cancers remain incurable, largely due to the high heterogeneity of the tumor.²⁹ Despite recent advances in the development of sophisticated techniques to assess tumor heterogeneity, the clinical application of tumor heterogeneity remains limited. Our study improved the understanding of the mechanisms underlying tumor heterogeneity, although the underlying prognostic mechanisms require further investigation. Future development of effective treatment strategies for heterogeneous tumors will be required, although this will not be possible without improved preclinical models that truly reproduce the heterogeneity of human disease, both between individuals and between tumors.

Our work elucidated tumor heterogeneity at multiple levels from the microbial community, host gene expression, and pathological images, all of which underlie differences in cancer risk in humans. Multimodal data fusion is probably the only way to advance precision oncology.³⁰ The limited view of disease heterogeneity from single modality data may not offer sufficient

information to stratify patients and capture all of the events that occur in response to treatments.³¹ Our work provided new insights into inter- and intra-tumor heterogeneity from a multimodal perspective of tumor microbes, host genes, and pathological images. Phenotypes derived from medical images can be used as substitutes for molecular phenotypes.^{32, 33} Furthermore, by analyzing the biological links between the three modalities data, we believe that image phenotypes can define tumor tissue microbial phenotypes and TME phenotypes for a broad spectrum of cancers.

In conclusion, our findings provided a framework for tumor subtyping across multiple types of cancers based on the association among the microbial community, host gene expression, and pathological image in the TME. Our findings indicated that MEI classification demarcated human cancers with specific tumor microbe-expression-image interplay patterns and prognoses. Our data provided biologically and clinically relevant information on the association of the microbial community, host gene expression, and pathological images in human tumors, which could assist with providing more precise diagnosis and prognosis assessment in the clinic.

REFERENCES

- Cullin, N., Azevedo Antunes, C., Straussman, R., et al. (2021). Microbiome and cancer. *Cancer Cell* **39**:1317-1341. DOI: 10.1016/j.ccell.2021.08.006.
- Donehower, L.A., Soussi, T., Korkut, A., et al. (2019). Integrated analysis of TP53 gene and pathway alterations in the cancer genome atlas. *Cell Rep.* **28**: 1370-1384.e5. DOI: 10.1016/j.celrep.2019.07.001.
- Tsai, P.C., Lee, T.H., Kuo, K.C., S., et al. (2023). Histopathology images predict multi-omics aberrations and prognoses in colorectal cancer patients. *Nat. Commun.* **14**: 2102. DOI: 10.1038/s41467-023-37179-4.
- Sepich-Poore G.D., Zitvogel, L., Straussman, R., et al. (2021). The microbiome and human cancer. *Science* **371**: eabc4552. DOI: 10.1126/science.abc4552.
- Lee, Y., Bogdanoff, D., Wang, Y., et al. (2021). XYSeq: Spatially resolved single-cell RNA sequencing reveals expression heterogeneity in the tumor microenvironment. *Sci. Adv.* **7**: eabg4755. DOI: 10.1126/sciadv.abg4755.
- Dentro, S.C., Leshchiner, I., Haase, K., et al. (2021). Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* **184**: 2239-2254.e39. DOI: 10.1016/j.cell.2021.03.009.
- Zuo, C., Zhang, Y., Cao, C., et al. (2022). Elucidating tumor heterogeneity from spatially resolved transcriptomics data by multi-view graph collaborative learning. *Nat. Commun.* **13**: 5962. DOI: 10.1038/s41467-022-33619-9.
- Ma, J., Huang, L., Hu, D., et al. (2021). The role of the tumor microbe microenvironment in the tumor immune microenvironment: Bystander, activator, or inhibitor. *J. Exp. Clin. Cancer Res.* **40**: 327. DOI: 10.1186/s13046-021-02128-w.
- Wong-Rolle, A., Wei, H.K., Zhao, C., Jin, C. (2021). Unexpected guests in the tumor microenvironment: Microbiome in cancer. *Protein Cell.* **12**: 426-435. DOI: 10.1007/s13238-020-00813-8.
- Ramirez-Labrada, A.G., Isla, D., Artal, A., et al. (2020). The influence of lung microbiota on lung carcinogenesis, immunity, and immunotherapy. *Trends Cancer.* **6**: 86-97. DOI: 10.1016/j.trecan.2019.12.007.
- Dzutsev, A., Badger, J.H., Perez-Chanona, E., et al. (2017). Microbes and Cancer. *Annu. Rev. Immunol.* **35**: 199-228. DOI: 10.1146/annurev-immunol-051116-052133.
- Garrett, W.S. (2015). Cancer and the microbiota. *Science* **348**: 80-86. DOI: 10.1126/science.aaa4972.
- van Pelt, G.W., Sandberg, T.P., Morreau, H., et al. (2018). The tumour-stroma ratio in colon cancer: The biological role and its prognostic impact. *Histopathology* **73**: 197-206. DOI: 10.1111/his.13489.
- Foersch, S., Glasner, C., Woerl, A.C., et al. (2023). Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer. *Nat. Med.* **29**: 430-439. DOI: 10.1038/s41591-022-02134-1.
- Poore, G.D., Kopylova, E., Zhu, Q., et al. (2020). Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579**: 567-574. DOI: 10.1038/s41586-020-2095-1.
- Wood, D.E. and Salzberg, S.L. (2014) Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**: R46. DOI: 10.1186/gb-2014-15-3-r46.
- Kim, D., Paggi, J.M., Park, C., et al. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**: 907-915. DOI: 10.1038/s41587-019-0201-4.
- Danecek, P., Bonfield, J.K., Liddle, J., et al. (2021) Twelve years of SAMtools and BCFtools. *Gigascience* **10**: giab008. DOI: 10.1093/gigascience/giab008.
- Pertea, M., Pertea, G.M., Antonescu, C.M., et al. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**: 290-295. DOI: 10.1038/nbt.3122.
- Roelands, J., Kuppen, P.J.K., Ahmed, E.I., et al. (2023). An integrated tumor, immune and microbiome atlas of colon cancer. *Nat. Med.* **29**: 1273-1286. DOI: 10.1038/s41591-023-02324-5.
- Lu, M.Y., Williamson, D.F.K., Chen, T.Y., et al. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**: 555-570. DOI: 10.1038/s41551-020-00682-w.
- Hu, M.-K. (1962). Visual pattern recognition by moment invariants. **8**: 179-187. DOI: 10.1109/TIT.1962.1057692.
- Yoshihara, K., Shahmoradgol, M., Martinez, E., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**: 2612. DOI: 10.1038/ncomms3612.
- Becht, E., Giraldo, N.A., Lacroix, L., et al. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17**: 218. DOI: 10.1186/s13059-016-1070-5.
- Nejman, D., Livyatan, I., Fuks, G., et al. (2020). The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science* **368**: 973-980. DOI: 10.1126/science.aay9189.
- Ghaddar, B., Biswas, A., Harris, C., et al. (2022). Tumor microbiome links cellular programs and immunity in pancreatic cancer. *Cancer Cell* **40**: 1240-1253 e5. DOI: 10.1016/j.ccell.2022.09.009.
- Llorens-Rico, V., Simcock, J.A., Huys, G.R.B., et al. (2022). Single-cell approaches in human microbiome research. *Cell* **185**: 2725-2738. DOI: 10.1016/j.cell.2022.06.040.
- Hoadley, K.A., Yau, C., Hinoue, T., et al. (2018). Cell-of-origin patterns dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**: 291-304 e296.
- Li, Z., Seehawer, M., and Polyak, K. (2022). Untangling the web of intratumour heterogeneity. *Nat. Cell Biol.* **24**: 1192-1201. DOI: 10.1038/s41556-022-00969-x.
- Steyaert, S., Pizurica, M., Nagaraj, D., et al. (2023). Multimodal data fusion for cancer biomarker discovery with deep learning. *Nat. Mach. Intell.* **5**: 351-362. DOI: 10.1038/s42256-023-00633-5.
- Grossman, J.E., Vasudevan, D., Joyce, C.E., et al. (2021). Is PD-L1 a consistent biomarker for anti-PD-1 therapy. The model of balstilimab in a virally-driven tumor. *Oncogene* **40**: 1393-1395. DOI: 10.1038/s41388-020-01611-6.
- Liu, Y., Kim, J., Balagurunathan, Y., et al. (2016). Radiomic features are associated with EGFR mutation status in lung adenocarcinomas. *Clin. Lung Cancer* **17**: 441-448.e6. DOI: 10.1016/j.clcc.2016.02.001.
- Gevaert, O., Echeagaray, S., Khuong, A., et al. (2017). Predictive radiogenomics modeling of EGFR mutation status in lung cancer. *Sci. Rep.* **7**: 41674. DOI: 10.1038/srep41674.

FUNDING AND ACKNOWLEDGMENTS

The results presented here are part based upon data generated by TCGA Research Network: <https://www.cancer.gov/tcga>. We thank LetPub for its linguistic assistance during the preparation of this manuscript.

AUTHOR CONTRIBUTIONS

L.J., K.N., and J.Y. conceived the research idea. H.Z., K.N., and J.Y. designed the study. L.J. conducted the main analysis. L.J., H.Z., and G.T. wrote the manuscript. S.X., Y.C., Y.Z., and J.L. reviewed and edited the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

ETHICAL STATEMENT AND PATIENT CONSENT

Not applicable.

DATA AND CODE AVAILABILITY

The onco-microbial community abundance data and metadata are available at ftp://ftp.microbio.me/pub/cancer_microbiome_analysis/. The host gene expression and the histopathological slides are available at <https://portal.gdc.cancer.gov/>.

SUPPLEMENTAL INFORMATION

It can be found online at <https://doi.org/10.59717/j.xinn-life.2023.100028>

LEAD CONTACT WEBSITE

<https://www.researchgate.net/profile/Jialiang-Yang>