# Pan-cancer analysis reveals tumor microbiome associations with host molecular aberrations

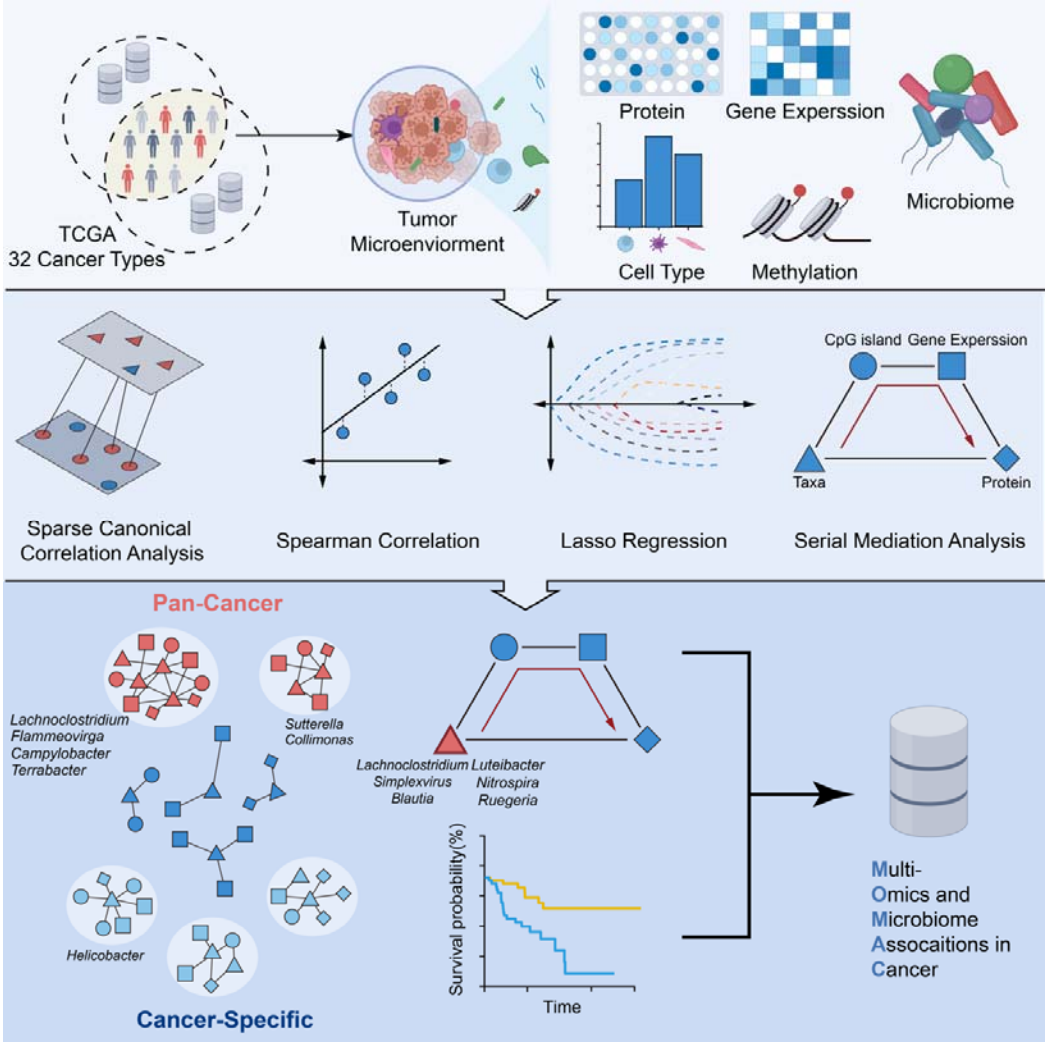
Chenchen Ma[1#], Changxing Su[1,2#], Jiaxuan Li[1], Jiuxin Qu[2], Shimin Shuai[1]*

1. Department of Human Cell Biology and Genetics, School of Medicine, Southern University of Science and Technology, Shenzhen, 518055, China
2. The Third People's Hospital of Shenzhen, Southern University of Science and Technology, National Clinical Research Center for Infectious Diseases, Shenzhen, 518115, China

# These authors contributed equally to this work.

* Correspondence: Shimin Shuai (E-mail: shuaism@sustech.edu.cn)

## Graphical abstract



**Abstract**

Host-microbiome interaction is known to play a pivotal role in the cancer ecosystem, yet the associations have not been systematically investigated at the pan-cancer and the multi-omics level. Here, we evaluated nearly 10,000 samples across 32 cancer types collected from The Cancer Genome Atlas (TCGA), to investigate the association between the tumor microbiome (taxa, n=1,630) and tumor microenvironment composition (cell types, n=20), epigenome (CpG island methylation, n=30,716), transcriptome (gene expression, n=10,216) and proteome (protein expression, n=193). We identified 836,738 candidate associations between the tumor microbiome and host molecular aberrations across multiple cancers. Besides cancer-specific associations, we also revealed recurrent pan-cancer associations between microbes (*Lachnoclostridium*, *Flammeovirga*, *Terrabacter* and *Campylobacter*) and immune cells, as well as between microbes (*Collimonas* and *Sutterella*) and fibroblasts, which were further validated by cell type estimations derived from pathological images and methylation data. We also identified several potential microbe and gene/protein expression associations mediated by DNA methylation using the sequential mediation analysis. Furthermore, our survival analysis demonstrated that tumor microbes may affect the patient's overall survival and progression-free survival. Finally, a user-friendly web portal, Multi-Omics and Microbiome Associations in Cancer (MOMAC) was constructed for users to explore potential host-microbe interactions in cancer.

**Keywords**: tumor microbiome; tumor microenvironment; host-microbe interaction; pan-cancer analysis; multi-omics

## Introduction

Recent large-scale analyses reveal that tumor-resident bacteria, archaea, viruses and fungi, collectively called the tumor microbiome [1-4], are a key component of the cancer ecosystem as "permanent resident" rather than "short-term tenant" [5]. Several functional and association studies have revealed that the tumor microbiome may play crucial roles across the entire lifespan of cancer, such as promoting oncogenesis [6] and metastasis [7], as well as modulating the cancer ecosystem and treatment response [8,9].

Treatment efficacy and cancer development may be significantly impacted by the interactions between the tumor microbiome, tumor cells, and immune cells in the cancer ecosystem [9,10], which is largely orchestrated by inflammatory cells [11], but little is known about the precise action modes of the interplay between the tumor microbiome and host cells in the cancer ecosystem. One hypothesis could be that the tumor microbiome contributes to tumor progression and development via affecting the host immune responses and cancer-related signaling pathways [4]. However, systematic analysis of associations between the tumor microbiome and host molecular profiles such as DNA methylation, RNA expression and protein levels in pan-cancer is lacking. In addition, although previous study has suggested that tumor microbial communities are largely unique to each cancer type and microbial profiles can be used to distinguish cancers [2], whether both pan-cancer and cancer-type-specific correlations between the tumor microbiome and host molecular features exist are not clear.

Recent advances have shown that the integration of multi-omics data, which include genomics, epigenomics, transcriptomics, and proteomics, can help us better understand how the host and microbe interact by correlating host molecular aberrations with the microbiota [12-14]. The Cancer Genome Atlas (TCGA), one of the largest cancer sequencing efforts, has profiled host molecular aberrations at the DNA, RNA, protein and epigenetic levels for 32 cancer types. The tumor microbiome in the entire TCGA cohort has also been systematically characterized [1,2]. Hence, it is possible and imperative to link the pan-cancer microbiome and host multi-omics data

to gain new insights into potential functional roles of the tumor microbiome in the cancer ecosystem.

Here, we comprehensively characterized host-microbe associations in 32 cancer types using multi-omics profiles from TCGA. Our analysis identified both cancer-type-specific and pan-cancer host-microbe associations. We validated and investigated several recurrent host-microbe associations with certain molecular and clinical features. Finally, we present MOMAC (Multi-Omics and Microbiome Associations in Cancer), an online web portal (https://comics.med.sustech.edu.cn/momac) for exploring host-microbe associations identified in pan-cancer.

## Results

### Identifying host-microbe associations at the pan-cancer and multi-omics level

In this study, we integrated tumor microbiome profiles consisting of 1,206 bacteria, 224 fungi, 138 viruses and 62 archaea, along with host molecular features including 10,216 gene expression, 30,716 CpG island methylation, and 193 protein expression, for nearly 10,000 samples across 32 cancer types collected from TCGA. To model the cancer ecosystem, we also integrated proportions of 20 cell types derived from the deconvolution of bulk transcriptomics data. To evaluate the associations between the tumor microbiome and host features, we employed various analytical techniques, including sparse canonical correlation analysis (sparse CCA), Spearman's rank correlation test and the randomized lasso regression analysis (**Figure 1A**).

After screening more than 4 billion microbe and host feature pairs, we identified 836,738 candidate associations (Benjamini-Hochberg adjusted p-value < 0.1 for both the Spearman's test and the lasso test, and stability > 0.6 for the randomized lasso test; **Methods**), which accounted for only 0.0198% of the total association tests. Specifically, we discovered 1,225, 365,458, 465,662, and 1,604 associations between microbes and cell types, gene expression, CpG island methylation, and protein

expression, respectively. Our results demonstrated the feasibility of identifying associations between the tumor microbiome and host features using multi-omics data in pan-cancer, and such associations can be found at different omics and cancer types.

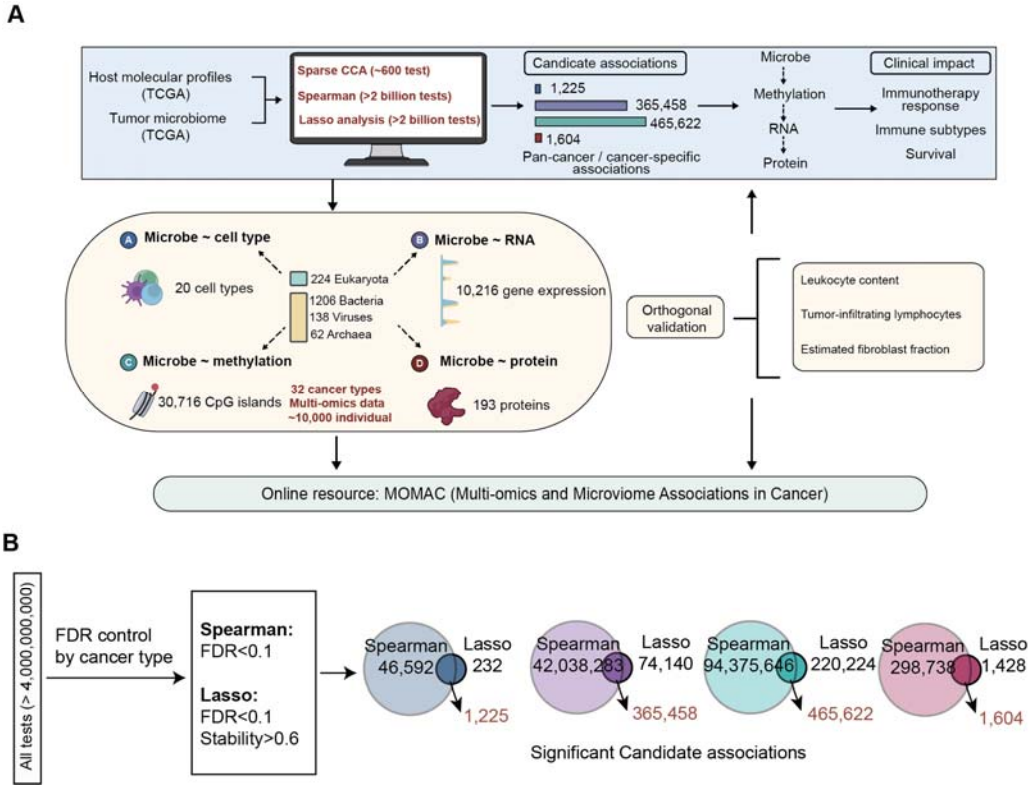



**Figure 1. Overview of the study. A)** Workflow for the identification and characterization of associations between host molecular features and the tumor microbiome at pan-cancer and multi-omics level. **B)** Selection of significant candidate associations for host cell types (blue), gene expression (purple), CpG island methylation (green) and protein expression (red). Final numbers of candidate associations are shown as the overlap of Spearman's rank correlation test and lasso regression test.

## Validating known associations on pathogenic infections

To assess whether our candidate associations captured known host-microbe interactions, we examined microbes that were causally linked to cancer. We first analyzed patients with liver hepatocellular carcinoma (LIHC) who may be infected with the Hepatitis B virus (HBV species). LIHC patients had an overabundance of *Orthohepadnavirus* (HBV genus), which is a major causative etiological factor.

Overall, 13 consistent hub genes identified by our work and recent studies [15-17] were confirmed to be associated with HBV infection (**Figure S1A, Table S1**). Our gene set overrepresentation analysis using gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) revealed that genes associated with the HBV genus were enriched in biological pathways like mitotic nuclear division, cell cycle, DNA replication, chromosome segregation and nuclear division, which were also consistent with previously identified pathways that may contribute to the pathogenesis of HBV-infected liver cancer (**Figure S1A**)[17].

Next, we analyzed *Helicobacter pylori* (within the *Helicobacter* genus) in stomach adenocarcinoma (STAD) patients, which is known to be associated with gastric carcinogenesis through p53 degradation [18]. Among 13 genes associated with the *Helicobacter* genus, nine genes were also identified in previous studies (**Figure S1B, Table S1**) [19]. These genes were involved in GO terms related to homophilic cell adhesion via plasma membrane adhesion molecules and epithelial tube morphogenesis, and KEGG pathways related to cancer (**Figure S1B**). We also discovered negative correlation ($r$ = -0.195, $p$ = 0.0025) between *Helicobacter* and *TP53* gene expression (**Table S1**). Collectively, these results indicate that our candidate associations could capture known associations between host molecules and oncogenic microbes.

**Identifying recurrent associations between tumor microbiome and host features**

To provide a more detailed insight into the distribution of candidate associations, we categorized them into cancer-type-specific and pan-cancer associations as well as gastrointestinal tract (GI) cancer and non-GI cancer associations (**Figure 2A**). Pan-cancer associations were defined as recurrent host-microbe associations if they were found in at least eight out of 32 cancer types for cell types, gene expression and CpG island methylation, and at least four cancers for protein expression. Our analysis revealed that approximately 24.98%, 54.84%, 79.93% and 86.16% of associations for cell type, RNA, methylation and protein were found only in one cancer type

(cancer-type-specific), while 36.57%, 11.13%, 0.98% and 3.05% associations were recurrent pan-cancer associations, respectively. We also noticed that seven GI cancers (colon adenocarcinoma, COAD; cholangiocarcinoma, CHOL; esophageal carcinoma, ESCA; pancreatic adenocarcinoma, PAAD; rectum adenocarcinoma, READ; STAD; LIHC) accounted for 16.77%-24.47% of all associations for different host feature types. For the rest of the manuscript, we focused on characterizing recurrent pan-cancer associations, while providing examples of cancer-type-specific associations. Notably, all potential host-microbe interactions can be explored via the MOMAC web portal.

Interestingly, for host cell type associations, we discovered a large proportion of cancer types (28/32, 87.5%), except for four cancers with limited sample size (**Figure 2B, Figure S2A-B**), exhibited strong associations between microbes and non-tumor cells in the tumor microenvironment (TME). Non-tumor cells such as immune cells and fibroblasts, are crucial components of the TME, affecting tumor immune reaction and response to immunotherapy [20,21]. Moreover, we identified two clusters of recurrent pan-cancer associations between the tumor microbiome and non-tumor cells: cluster 1 between *Lachnoclostridium*, *Flammeovirga*, *Terrabacter* and *Campylobacter* and immune cells, and cluster 2 between *Collimonas* and *Sutterella* and fibroblasts (**Figure S2C)**. Hereafter, these six microbes were referred to as pan-cancer key microbes.

Microbes within these two clusters were also associated with many other host molecular aberrations. At the gene expression level, for example, 1,275 and 795 genes were discovered to be associated with *Campylobacter* and *Lachnoclostridium*, respectively, which were among the top ten microbes having the highest number of associations with host gene expression (**Figure 2C, Table S2**). We also annotated gene pathways associated with two representative microbes from two clusters, *Lachnoclostridium and Collimonas*. *Lachnoclostridium* was discovered to be correlated with activation and adhesion of immune cells, while *Collimonas* was

associated with extracellular organizations (**Figure 2D**). Gene set enrichment analysis for the other four microbes can be found in **Figure S3A**, and the results were consistent with their non-tumor cell associations. Namely, cluster 1 microbes mainly correlated with host immune responses and cluster 2 microbes mainly correlated with extracellular organizations. Similarly, the DNA methylation status of 262 and 199 CpG islands were associated with *Lachnoclostridium* and *Terrabacter*, respectively. We also found that microbes correlated with immune cells in cluster 1 tend to be associated with a shared set of CpG islands (**Figure 2E, Table S3**). By linking CpG islands to genes, we found cluster 1 microbes were associated with pathways like actin organization and cell adhesion (**Figure S3B**). For *Lachnoclostridium*, we also identified 15 genes with both mRNA and methylation associations, and these genes were related to the immune signaling (**Figure S3C**). Interestingly, we also noticed some microbes associated with multiple CpG islands linked to the same gene. For example, *Lachnoclostridium* was found to have multiple associated CpG islands linked to *AGAP2*, *NCOR2*, *HIVEP3*, and *HLA-F*, while four CpG islands on *DOCK1* correlated with *Terrabacter* abundance (**Figure S3D**). Since only 66.9% of TCGA samples having the protein expression data, we detected merely eight recurrent associations, mainly (n = 5) for *Lachnoclostridium* (**Figure 2F, Table S4**).

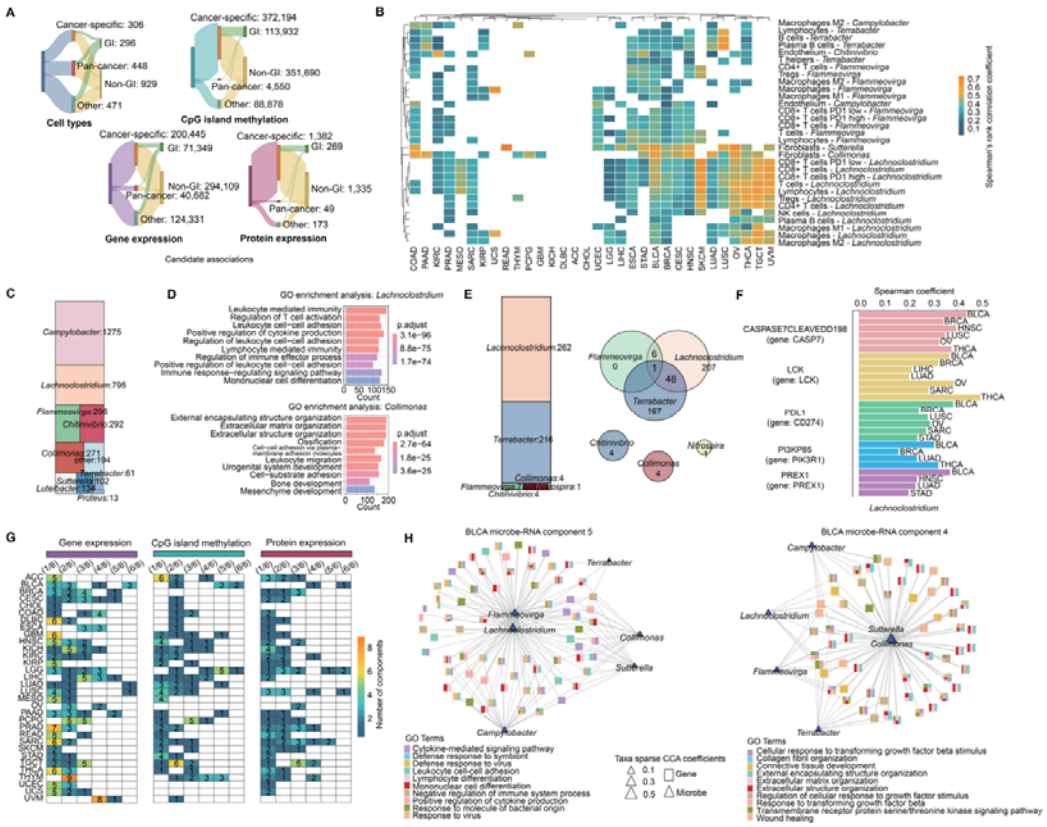


**Figure 2. Pan-cancer recurrent host-microbe associations. A)** The Sankey diagram shows candidate associations grouped by recurrence or organ system. Recurrence grouping includes cancer-specific (only found in one cancer type), pan-cancer (>=4 cancer types for protein and >=8 cancer types for the rest) or other. Organ system grouping includes gastrointestinal (GI) cancer and non-GI cancer. GI cancer includes COAD, CHOL, ESCA, LIHC, PAAD, READ and STAD. **B)** Heatmap shows pan-cancer associations between microbes and cell types. Color shows Spearman's correlation coefficients. **C)** Mosaic plot shows the top ten microbes with the highest number of associated gene expression. **D)** Gene ontology overrepresentation analysis for two representative microbes, *Lachnoclostridium* and *Collimonas*. **E)** Mosaic plot on the left shows the number of CpG islands associated with pan-cancer microbes and the Venn diagram on the right shows the overlap of associated CpG islands for pan-cancer microbes. **F)** Barplot shows pan-cancer associations between *Lachnoclostridium* and protein expression, with gene names showed in brackets. **G)** Heatmap shows the number of sparse CCA components for six pan-cancer key microbes. Columns 1 to 6 correspond to the number of pan-cancer key microbes co-occurred in the same component. **H)** Network plots show two

examples in BLCA for microbe-RNA sparse CCA components. Squared nodes in the network represent genes colored by their GO terms, and triangular nodes are microbes. The size of triangle represents the absolute value of sparse CCA coefficients for microbes. Only microbe-RNA associations with Spearman's rho > 0.3 and top 10 GO terms based on adjust *p* values are showed.

To further explore the role of pan-cancer key microbes, we employed the sparse CCA method to study the co-occurrence and host association patterns of them. In most cancer types, more than two of these six microbes could be observed in the same component (**Figure 2G, Table S5**). In several instances, we found all six microbes belonging to the same component, including associations between gene expression and microbes in bladder urothelial carcinoma (BLCA), lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), between CpG island methylation and microbes in BLCA and brain low grade glioma (LGG), and between protein expression and microbes in BLCA, LGG and LUSC. The sparse CCA analysis also allowed us to compare the relative contribution of different microbes in certain pathways. For example, in the fifth sparse CCA component of BLCA, we highlighted the pivotal role of cluster 1 microbes in comparison with cluster 2 microbes in host immune signaling pathways (**Table S6, Figure 2H**), as the absolute coefficients for *Lachnoclostridium* (-0.56), *Flammeovirga* (-0.26) and *Campylobacter* (-0.24) were higher than *Collimonas* (-0.10) and *Sutterella* (-0.11); in contrast, we noticed a higher contribution of *Collimonas* (-0.48) and *Sutterella* (-0.48) to extracellular organizations than *Lachnoclostridium* (-0.17), *Flammeovirga* (-0.17), *Campylobacter* (-0.17) and *Terrabacter* (-0.16) in the fourth sparse CCA component (**Table S6**). Similar patterns of differential contribution of cluster 1 and cluster 2 microbes in host pathways were also observed at the DNA methylation and protein expression levels using sparse CCA (**Figure S4, Table S7-8**). Altogether, these results demonstrated that recurrent host-microbe associations could be identified in pan-cancer and were mainly related to non-tumor cells and relevant signaling pathways in the TME.

**Validating recurrent associations using orthogonal data**

Recurrent host-microbe associations identified above used cell type data based on RNA-Seq, thus we wondered if they can be validated by cell type estimations based on orthogonal data types. For the purpose of validation, we selected two control microbes (*Bacteroides*, a common bacterium, and *Neisseria*, a classified pathogen) that had no associations with cell types in the initial test and whose average abundance was similar to the pan-cancer microbes. Firstly, we used leukocyte fraction estimated from DNA methylation [22], and found that cluster 1 microbes but not control microbes showed consistent associations with the leukocyte fraction across multiple cancer types (**Figure 3A**). Next, we examined the association between microbes and tumor-infiltrating lymphocytes based on Hematoxylin and Eosin (H&E) images for 13 tumor types [23]. As expected, the results were consistent with microbe and cell type associations based on RNA-Seq for *Lachnoclostridium*, *Flammeovirga* and control microbes (**Figure S5**). For example, for *Lachnoclostridium*, in 8 out of 9 tumor types showing associations with immune cells previously, we confirmed again the association between microbe and tumor-infiltrating lymphocytes, whereas this correlation was absent in three out of four cancer types lacking associations in previous analysis (**Figure 3B**). To further validate cluster 2 microbes associated with fibroblast, we defined an estimated fibroblast fraction, approximated by the non-tumor and non-leukocyte fraction of the TME (**Methods**). We again observed expected associations for cluster 2 microbes in cancer types where it had been initially discovered, but not in other cancer types nor in control microbes (**Figure 3C**). Collectively, our validation analysis indicated that recurrent pan-cancer host-microbe associations with immune cells and fibroblasts could be validated by orthogonal data.

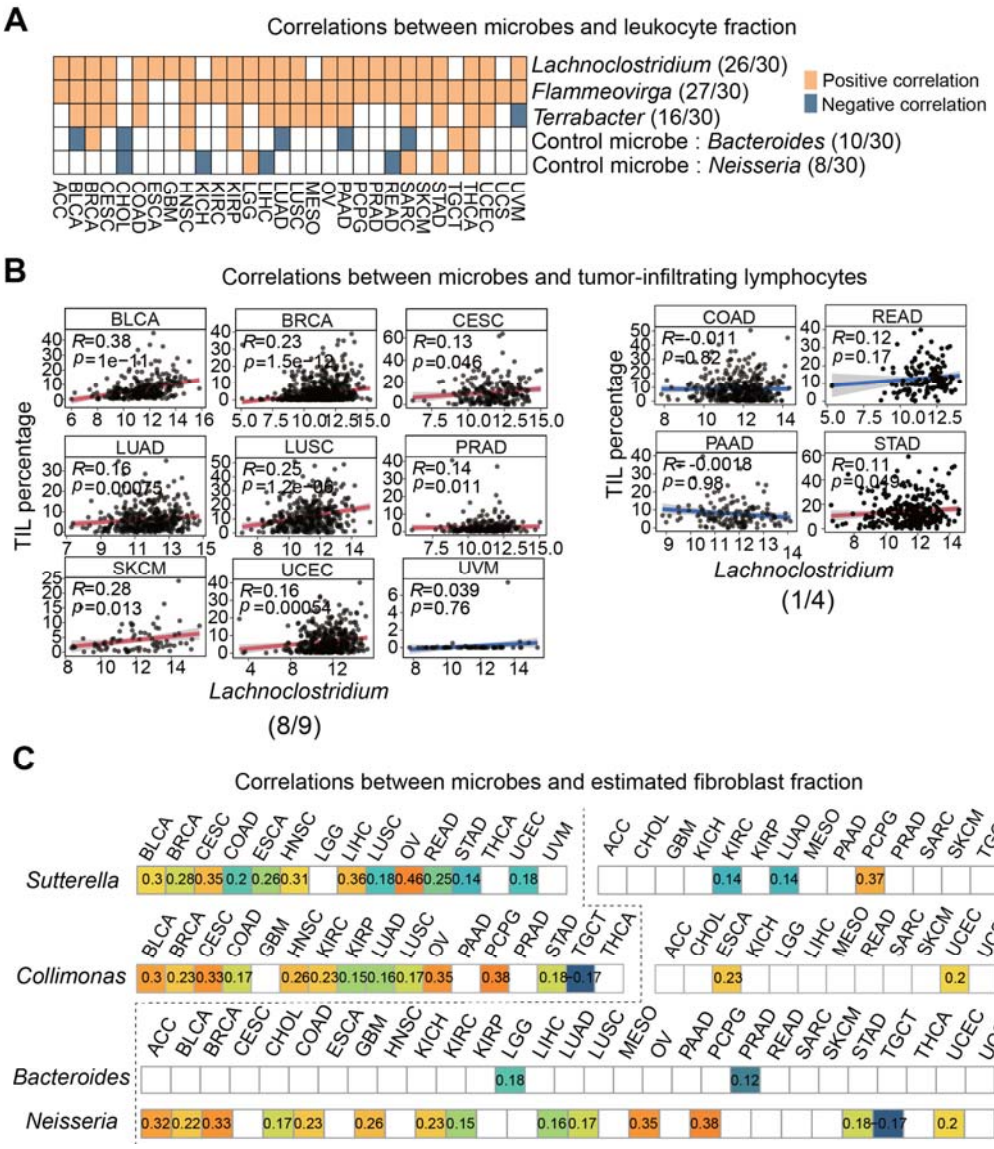



**Figure 3. Validation of associations between pan-cancer key microbes and host cell types. A-B)** Validation of associations between pan-cancer immune-related microbes (cluster 1) and leukocyte fraction from methylation (A), and tumor-infiltrating lymphocytes from pathological images (B). For B, the left shows nine cancer types with significant associations in the initial test, and the right shows four cancer types that were not; the red and blue fitted lines represent significant ($p < 0.05$) and non-significant Spearman's rank correlations in the validation analysis, respectively. **C)** Validation of associations between fibroblast-related microbes (cluster 2) and estimated fibroblast fraction. Numbers are the Spearman's rank correlation coefficients. White squares represent no correlation ($p > 0.05$).

**Identifying putative microbe-methylation-RNA-protein axis**

After analyzing host-microbe associations at the single omics level, we performed an integrative analysis to identify host-microbe associations across multiple omics, which may suggest key genes and how these genes are regulated by the tumor microbiome. As expected, we observed a high prevalence of overlaps between gene expression and proteins (**Figure 4A**). For instance, 41.67% (n = 53) microbe-protein associations in ovarian cancer (OV) were also detected at the RNA level. Furthermore, the *Lachnoclostridium* genus was found to be correlated with mRNA and protein levels of *LCK*, *CD274* and *PREX1* in multiple cancers. Interestingly, we observed varied levels (0-2,022 genes) of overlaps between microbe-RNA and microbe-methylation associations across 32 cancer types, suggesting the possibility of the tumor microbiome contributing to host gene expression via affecting CpG island methylation in certain cancers (**Figure 4B**).

In addition, we discovered 12 genes with associations across all three omics in eight cancers (**Figure 4C**). Intriguingly, although the direction of associations for microbe-RNA and microbe-protein was always consistent, the direction for microbe-methylation associations was not (**Figure 4D**), which may be related to the position of CpG islands (gene body or promoter) and the context-dependent manner of gene regulation by promoter hypermethylation [24]. For example, *Lymphocryptovirus* (Epstein–Barr virus) was enriched in EBV-infected primary STAD (**Figure S6A**), and positively correlated with T cells especially CD8+ T cells in our study (**Figure S6B**), which were consistent with previous studies[25,26]. Our results suggested that *Lymphocryptovirus* was associated with promoter hypermethylation (cg08545268, r = 0.18, $p$ = 0.00043) and increased mRNA ($r$ = 0.21, $p = 2\times10^{-5}$) and protein ($r$ = 0.18, $p$ = 0.0013) levels of *TGM2*, a known marker gene for worse prognosis in stomach cancer [25]. Additionally, the overexpression of *ERBB3* and its protein (HER3) have been reported in BLCA, which was involved in tumor progression and metastasis [27]. Our results implied that *Lachnoclostridium* might inhibit *ERBB3* expression (RNA: $r$ = -0.28, $p = 1.6\times10^{-8}$, protein: $r$ = -0.31, $p = 7.5\times10^{-9}$).

Next, we employed sparse CCA, a complimentary method with the advantage to identify cross-omics association modules. We observed that several microbe-RNA modules were also captured by microbe-protein and/or microbe-methylation analysis. For instance, in breast invasive carcinoma (BRCA), microbes including *Ruegeria* (coefficient = -0.33), *Simplevirus* (coefficient = -0.41), and genes including *ESR1* (coefficient = 0.07), *IRS1* (coefficient = 0.04) and *BCL2* (coefficient = 0.08) were members of the fifth sparse CCA component for microbe-RNA associations, and the same set of microbes and genes also contributed (absolute coefficient = 0.02-0.09) to the third component of microbe-protein associations (**Figure 4F, Table S9**). In COAD, LUSC and BLCA, we also found two microbes (*Luteibacter* and *Lachnoclostridium*) that may affect the same features in different omics (**Figure 4F, Table S9**).

Integrative analysis above further supported the idea that the tumor microbiome may modulate host molecular features via the microbe-methylation-RNA-protein axis, so we next performed the sequential mediation analysis, a statistical method that can test if one variable (microbe) affects another (RNA/protein) via a mediator (methylation). In BRCA, we connected the *Simplexvirus* genus of herpes simplex virus (HSV) with methylation, RNA, and protein changes for *IRS1* and *ESR1* (**Figure 4G, Figure S7** and **Table S10**). We speculated that *Simplexvirus* might enhance both the gene and protein expression of *IRS1* and *ESR1* via hypomethylating two corresponding CpG islands marked by cg04751089 and cg24900983, respectively. Moreover, *Blautia* and *Luteibacter* might contribute to the hypomethylation of *SYK* (promoter), *YAP1* (promoter), and *DUSP4* (gene body) to modulate their expression in LGG (**Figure S7**). In prostate adenocarcinoma (PRAD) and BRCA, similar microbe-methylation-RNA-protein axis was found for *Nitrospira-FASN* and *Ruegeria-BCL2* (**Figure S7**). For pan-cancer key microbes, *Lachnoclostridium* was associated with the hypermethylation of cg07746998 (*ESR1*) and cg04794420 (*ERBB3*), and their corresponding gene/protein levels in LUSC and BLCA, respectively (**Figure 4G**). Collectively, through multi-omics integrative analysis, our

results indicated that microbes could potentially affect host gene and/or protein expression via altering genome methylation.

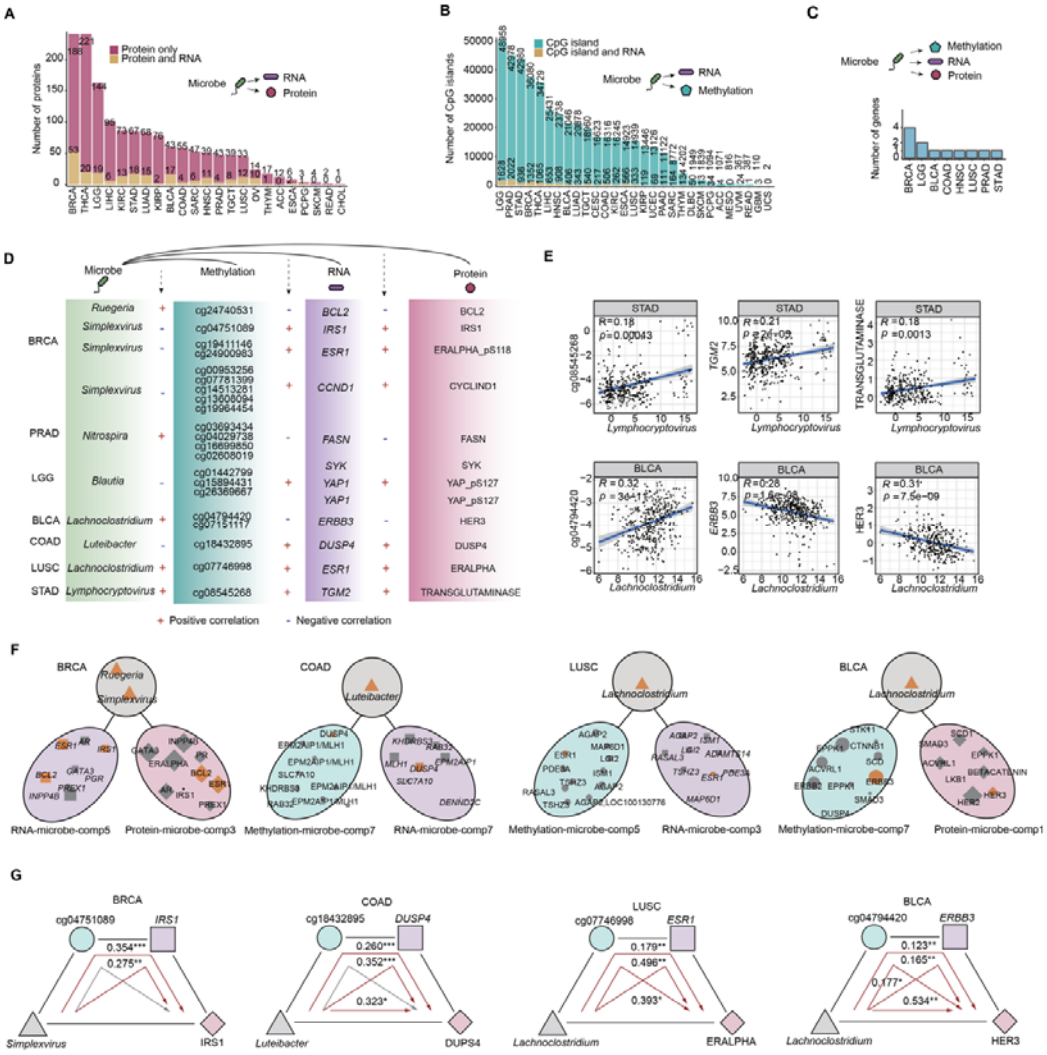



**Figure 4. Overlap of multi-omics between specific microbes and host molecular aberrations. A-C)** Bar plots show the number of overlaps for associations among RNA-protein (A), methylation-RNA (B) and methylation-RNA-protein (C). **D)** Associations between microbes and host genes with complete microbe-methylation-RNA-protein paths. **E)** Scatterplots show two multi-omics association examples from D. The Spearman's rank correlation coefficients and *p* values are showed. **F)** Microbes associated with the same sets of genes across sparse CCA components (comp) of different omics. CpG islands have been linked to genes. The size of rectangular nodes indicates the absolute value of coefficients from sparse CCA. **G)** Sequential mediation analysis results for tumor microbes and host genes. Each path represents a potential microbe-methylation-RNA-protein axis from D (also in Figure S7). The

proportional effect size (shown in number) and *p* values (*, <0.05; **, <0.01; ***, <0.001) are showed in each plot, and the red line represents significant path with *p* values < 0.05 and non-zero proportional effect size.

**Investigating the clinical impact of microbes with pan-cancer associations**

We next sought to characterize the clinical impact of these microbes having recurrent pan-cancer associations. Since the gut microbiota can shape the response to immune checkpoint inhibitors [28,29], we suspected that the tumor microbiome may also be associated with immunotherapy response. One previous study reported that the T cell-inflamed gene expression profile (GEP) score can be utilized to classify responders versus non-responders to PD-1 checkpoint blockades [30]. Hence, we calculated GEP scores for all samples and tested their associations with the tumor microbiome. Our analysis revealed significant positive associations between the GEP score and six pan-cancer microbes across multiple cancer types (**Figure 5A**), especially for *Lachnoclostridium* ($r$ = 0.31-0.70), *Flammeovirga* ($r$ = 0.30-0.54), and *Terrabacter* ($r$ = 0.23-0.42).

All TCGA tumors have been divided into six immune subtypes (C1 to C6) with clinical relevance in a previous study [22], and we next examined if our pan-cancer key microbes contributed to tumor immune subtypes. The C5 (immunologically quiet) subtype featured by few lymphocytes is mainly reported in LGG [22]; thus, we first compared the abundance of immune-related microbes between C3 (inflammatory), C4 (lymphocyte depleted) and C5 in LGG (**Figure 5B**). As expected, the abundance of *Lachnoclostridium* and *Flammeovirga* significantly reduced in C5 compared with C3 ($p$ = 0.007 and 0.024, respectively) and C4 (both $p$ < 0.001) in LGG. Moreover, C1 (wound healing), C2 (IFN-γ dominant) and C3 (inflammatory) are present in most cancers [22], and we found significant differences in the abundance of *Lachnoclostridium* between these immune subtypes in multiple cancer types (**Figure 5C**). Note that *Lachnoclostridium* was also positively correlated with the *IFNG* expression (C2 immune subtype marker) in 18 cancer types (**Table S12**).

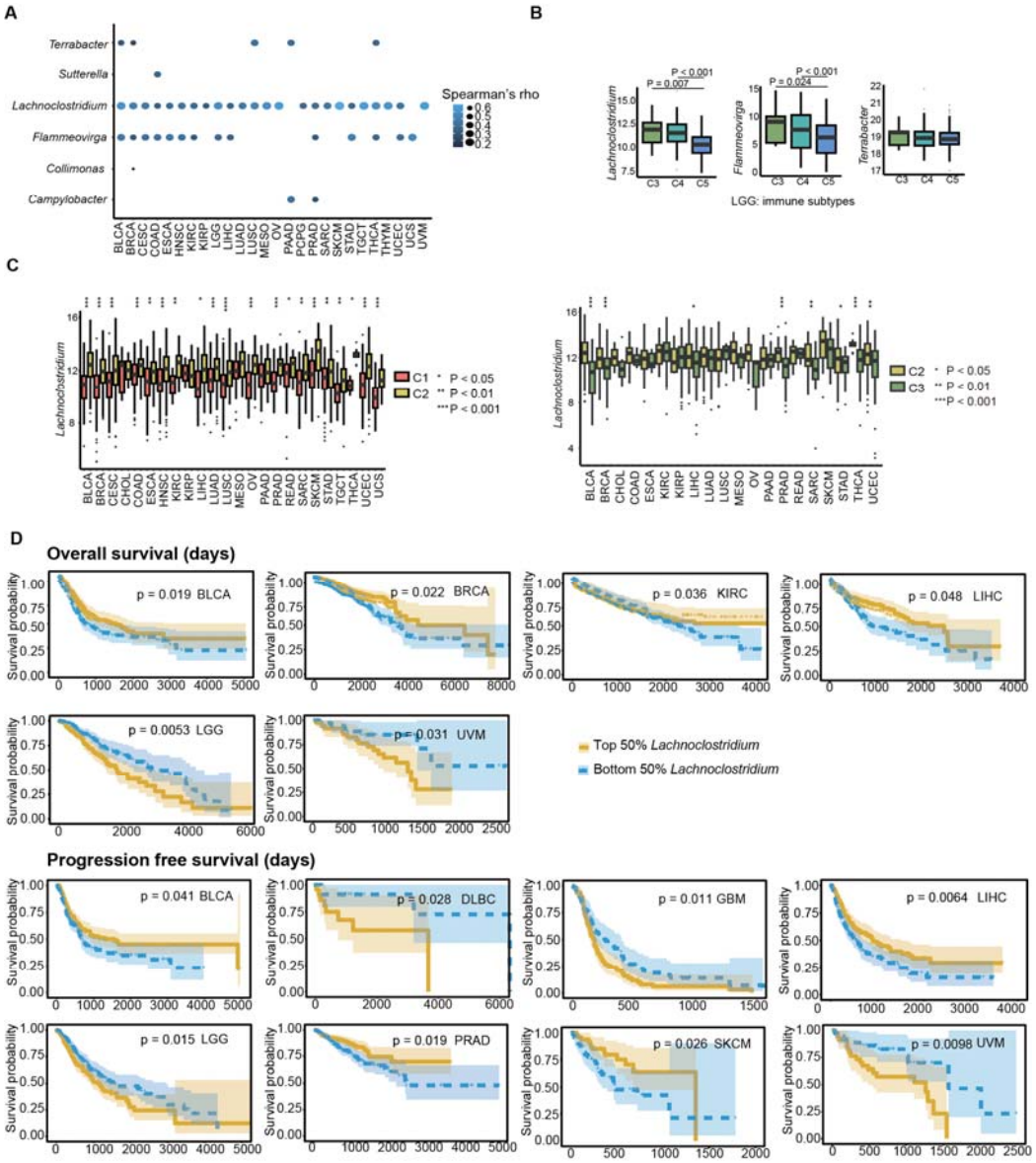



**Figure 5. Clinical impact of pan-cancer key microbes. A)** Associations between microbes and GEP scores. Significant associations are shown with circles (Spearman: FDR < 0.1; lasso: FDR < 0.1, and stability selection > 0.6). **B-C)** Associations between *Lachnoclostridium* and immune subtypes in LGG (B) and pan-cancer (C). C1, wound healing; C2, IFN-γ dominant; C3, inflammatory; C4, lymphocyte depleted; C5, immunologically quiet. Wilcoxon tests were used to determine statistical significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). **D)** Kaplan-Meier plots show the potential clinical impact of *Lachnoclostridium* in OS and PFS across multiple cancer types. $P < 0.05$ was considered as significant.

Finally, these findings with immunotherapy response and immune subtype prompted us to investigate the association between pan-cancer key microbes and prognosis. We defined top and bottom groups based on the median abundance of pan-cancer key microbes and compared the overall survival (OS) and progression-free survival (PFS) between these two groups. Stratification of patients based on the abundance of *Lachnoclostridium* was significantly associated with OS and PFS in multiple cancer types (**Figure 5D**), including positive correlations with OS in BLCA, BRCA, kidney renal clear cell carcinoma (KIRC) and LIHC, and PFS in BLCA, LIHC, PRAD and skin cutaneous melanoma (SKCM), and negative correlations in LGG (OS: $p = 0.015$, PFS: $p = 0.02$) and GBM (glioblastoma multiforme; PFS: $p = 0.0098$), UVM (uveal melanoma; OS: $p = 0.021$, PFS: $p = 0.0098$) and DLBC (diffuse Large B-cell Lymphoma; PFS: $p = 0.028$). We also identified significant associations with OS and/or PFS for other pan-cancer key microbes (**Figure S8**). Together, these results suggested that the tumor microbiome has potential clinical utility such as patient stratification and prediction of therapeutic response and prognosis.

## Discussion

In this study, we linked the tumor microbiome and host molecular features at the pan-cancer and multi-omics level. We screened the correlations and further characterized six microbes with recurrent pan-cancer associations in detail. Note that there were also numerous cancer-type-specific host-microbiome associations that could be explored via our MOMAC web portal. Additionally, we have undertaken extensive effort to validate our findings and correlate them with clinical features. The prevalence and importance of the tumor microbiome has been indicated in multiple previous studies [31-34]; our systematic study here can improve our understanding of the host-microbe interplay in cancer and suggest potential molecular mechanisms.

Our observations confirmed several known host-microbe associations. The HBV genus is causally related to LIHC[35]. Previous studies have discovered that HBV could integrate at the *TERT* and *MLL4* genes, increasing their expression, which contributes

to carcinogenesis [35-37]. In our study, we found that the *TERT* expression and the *MLL4* hypomethylation was positively and negatively associated with the HBV abundance, respectively. However, some cancer-related genes such as *PTPRD*, *UNC5D*, *NRG3*, *CTNND2* and *AHRR* that are altered in HBV-infected liver cancer were not identified in our study, suggesting future studies are required to collect more samples and/or improve the method for host-microbe association study. In STAD, *Helicobacter pylori* infection is known to trigger a proinflammatory response and enhance p53 degradation in host cells due to the presence of CagA protein [38,39], which is also consistent with our analysis. Moreover, previous study reports that the promoter hypermethylation of *MGMT* is associated with *Helicobacter pylori* CagA-positive strains [40], and we also find positive correlation between overabundant *Helicobacter* and the hypermethylation of *MGMT*. Thus, our analysis recaptures known host-microbe interplays and further suggest the possibility of finding new host-microbe associations through MOMAC.

From our pan-cancer findings, we speculate that the tumor microbiome may contribute to tumor progression and development via affecting the host immune responses and cancer-related signaling pathways. A recent study reports that tumor microbes belonging to the Lachnospiraceae family can enhance the activation of CD8+ T cells [13]. Similarly, in COAD, we discovered positive association between *Lachnoclostridim* (within the Lachnospiraceae family) and T lymphocytes. Besides *Lachnoclostridium*, we discovered other microbes associated with immune cells in multiple cancers, including *Flammeovirga*, *Terrabacter* and *Campylobacter*. Of note, the *Campylobacter* has also been reported recently in PDAC to be related to immune cells [41]. Furthermore, we also identified two microbes associated with fibroblasts and pathways related to extracellular matrix. Hence, non-tumor cells in the TME might be important targets for pan-cancer microbes and future studies should be conducted to validate and study such interplays *in vivo* and *in vitro*.

Furthermore, our multi-omics integration using multiple algorithms shows strong

evidence of associations between tumor microbes and host molecular profiles. The CpG island hypermethylation and hypomethylation are common epigenetic signals that regulate gene expression [42]. Previous studies have shown that the gut microbiome has an impact on host DNA methylation, which in turn contributes to colon cancer development [43]. Here we also tested the probability of the tumor microbiome modulating host gene expression via CpG island methylation by sequential mediation analysis. Our results indicate that the microbe-methylation-RNA-protein axis can be observed for multiple cancer-related genes across a wide range of cancer types, although how tumor microbes induce host DNA methylation changes requires more studies. Previous studies demonstrated that butyrate produced by gut microbes can alter DNA methylation in cell lines [44], so if certain toxins produced by tumor microbes have similar functions should be tested experimentally in the future.

In conclusion, we conduct the first pan-cancer multi-omics association study between host molecular features and the tumor microbiome, and identify potential associations with putative molecular mechanisms within the TME which can be further explored in our MOMAC web portal.

**Limitations of the study**

In this study, we discovered candidate associations using the TCGA cohort, one of the most comprehensive tumor sequencing efforts, while additional large multi-omics cohorts should nonetheless be used in future studies to validate our findings and explore potential environmental and/or genetic factors that may shape these associations. We mainly focused on investigating microbes with recurrent pan-cancer associations here. Nevertheless, there are more cancer-type-specific associations that are intriguing and may be crucial for cancer biology awaiting characterization. Moreover, we associated the tumor microbiome to host cell types inferred from bulk RNA-Seq, while future work should consider the usage of additional single-cell RNA sequencing and spatially resolved transcriptomics data to directly link the associations between the tumor microbiome and host cells in the cancer ecosystem. Finally, while

suggesting potential molecular mechanism, our analysis cannot establish causal links, which demands future functional studies to provide experimental evidence of how the tumor microbiome can alter host DNA methylation, gene and/or protein expression.

## Methods

### TCGA data integration and processing

We closely examined tumor microbiome profiles of 32 cancer types collected from Dohlman et al [1] (n = 10,245) and Poore et al. [2] (n = 9,751), and batch-corrected data and "Kraken-TCGA-Voom-SNM-All-Putative-Contaminants-Removed-Data" for the tumor microbiome were downloaded from them, respectively. The resulting two microbial tables include 224 fungi, 1,206 bacteria, 138 viruses and 62 archaea, all of which were decontaminated. For host molecular features, RNA (RESM_tpm, n = 10,535), RPPA (n = 7,744), DNA methylation 450K only (n = 9,639) can be downloaded from the UCSC Xena-TCGA Pan-Cancer platform [45]. Gene names, CpG islands and proteins were paired through the TCGA portal and the official manifest file for Illumina Infinium HumanMethylation450 v1.2 BeadChip. We selected primary tumor and RNA-sequenced samples for further data analysis, resulting in a sample size of 9,600 and 7,936 for Poore et al. and Dohlman et al. datasets, respectively. Note, to ensure data integrity, duplicated samples and multiple samples from the same individual were kept, with only one chosen at random for association testing. Sample sizes for all 32 cancer types can be found in **Figure S2A**. Note that adrenocortical carcinoma (ACC) and sarcoma (SARC) did not have any eukaryotic microbe profiles.

For host gene expression data, we utilized the R package biomaRt to retain only protein-coding genes [46]. Then, for sparse CCA analysis, to reduce the number of features, we kept only genes expressed in >95% of samples and having variances within the top 75% of all genes by cancer type. For each cancer type's host methylation, protein, and microbial abundance data, we likewise retained features with variance among the top 75% of all features for the sparse CCA analysis.

## Cell type data collection and processing

First, for initial discovery, we collected bulk RNA-Seq deconvolution profiles for 10,490 TCGA tumors using Kassandra, including abundance estimation of 20 different cell types [47]. We then took an extensive search of resources to validate our findings, including leukocyte fraction data for 9,783 samples in 30 non-hematologic cancer types based on deconvolution of DNA methylation markers [22], and the tumor-infiltrating lymphocyte fraction data for 4,210 samples in 13 cancer types based on AI processing of pathological images [23]. Since the fibroblast fraction has not been directly evaluated by other methods, we further collected tumor purity data for 10,787 samples in 30 non-hematologic cancer types. After intersecting with the leukocyte fraction and the cancer microbiome data, we calculated an estimated fibroblast fraction for 9,071 samples, which should be positively correlated with the fibroblast fraction in the TME as:

$$Estimated\ fibroblast\ fraction \propto 1 - Tumor\ purity - Leukocyte\ fraction$$

When validating key microbe-cell associations, we included *Bacteroides* and *Neisseria* as common bacteria and pathogen controls, respectively, as they were not associated with cell types in the initial test and the average abundance was similar to that of pan-cancer immune-related microbes.

## Sparse canonical correlation analysis

*Algorithm description*. The sparse CCA algorithm is a multivariate analysis method designed to identify the maximum correlation between variables by linearly projecting two sets of data onto a shared latent space, with regularization to achieve sparsity [48]. The sparse CCA method has been used in identifying host-microbe associations in literature before [49]. Given two data matrices with the same number of samples but different number of features, $\mathbf{X}_{n\times p}$ and $\mathbf{Y}_{n\times q}$, where $n$ denoted the number of observations (e.g, tumor samples), $p$ and $q$ were the number of features (e.g., gene and protein expression, or microbial abundance, etc.), sparse CCA involved finding $u$ and

$v$, which were loading vectors, that maximize $cor(Xu, Yv)$. In mathematical terms, this optimization problem can be formulated as:

$$\text{maximize}_{u,v} u^T X^T Y v$$

$$\text{subject to } u^T X^T X u \leq 1, v^T Y^T Y v \leq 1, ||u||_1 \leq \lambda_1, \text{and } ||v||_1 \leq \lambda_2$$

where parameters $\lambda_1$ and $\lambda_2$ correspond to the $L^1$ norm penalty terms that control the relative weight of the regularization term in the loss function and impacted the solution of the sparse CCA model. These parameters can be adjusted as needed.

*Grid search for hyperparameters*. The range of penalty parameters was constrained to be between 0.1-0.4 using the PMA R package (v1.2.1) [48], where a smaller value represents fewer features and a higher level of sparsity. We employed the leave-one-out cross-validation (LOOCV) method to perform a grid search on these hyperparameters $(\lambda_1, \lambda_2)$, with the objective of determining hyperparameter settings that would yield the highest Pearson's correlation coefficient between the first pair of canonical variate scores derived from $\mathbf{X}_{n\times p}$ and $\mathbf{Y}_{n\times q}$ datasets. The correlation was a metric for gauging the efficacy of the chosen hyperparameters in capturing the underlying relationship between the two datasets, ultimately guiding the selection of optimal hyperparameters that maximize the correlation among the canonical variates.

*Identification of correlated components*. We used sparse CCA to obtain highly correlated CCA components for different $\mathbf{X}_{n\times p}$-$\mathbf{Y}_{n\times q}$ pairs, including microbe-RNA, microbe-methylation, and microbe-protein for each cancer type. Note that, the sparse CCA between fungi and host features were ran separately since the quantification and normalization of fungus abundance were different than other microbes. We utilized optimized hyperparameters to fit a sparse CCA model on each pair of data, and computed the first 10 mutually uncorrelated latent feature spaces (referred to as components [50]), each encompassing highly correlated data from $\mathbf{X}_{n\times p}$ and $\mathbf{Y}_{n\times q}$. Then, we evaluated the statistical significance of all components using the LOOCV method, and used the BH procedure to correct for multiple hypothesis testing,

defining significant components as those with FDR < 0.1. We employed the R package foreach to construct a parallel computing framework for sparse CCA.

**Spearman's rank correlation and lasso regression analysis**

After filtering out low variable host molecular characters using the methods described above, we used a custom script to calculate the Spearman's rank correlation between tumor microbes and host molecular features in R. The total number of Spearman's rank correlation tests is as follows:

$$\sum_{i,j}^{m,n}(t_j f_i)$$

where $f_i$ is the number of features for every host omics phenotype (m = 4, cell types, gene expression, CpG islands and proteins), and $t_j$ is the number of microbes in each cancer type (n = 1,406 for ACC and SARC; n = 1,630 for other cancer types). The resulting *p* values were adjusted using the BH method by cancer and host feature type.

Furthermore, we conducted regression analyses to investigate the association between host molecular features and tumor microbe abundances, using the latter as predictors and the former as outcomes, as described in a recent study [49]. To mitigate the risk of overfitting, we applied lasso penalized regression model, implemented using the R package glmnet version 4.1.6 [51]. The lasso regression employs regularization to minimize the following equation:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

where *n* is the number of samples, *p* is the number of microbes, *y* is host molecular profiles, *x* is tumor microbes and $\lambda$ is the regularization parameter. To determine the optimal $\lambda$ value, we utilized LOOCV. Then, *p* values were calculated for the coefficient of each microbe ($\beta_j$) using the R package hdi version 0.1-9. We next used the BH method for multiple test correction by cancer and host feature type.

Further, to control for false discoveries, we implemented randomized lasso for stability selection. This approach ran lasso regression 50 times with randomly selected subsets of samples and calculated the fraction of times a feature being selected. A cutoff of 0.6 was used to select stable host-microbe associations. The stability selection was carried out using the R package stabs version 0.6-4 [52].

Finally, we screened all associations and created a set of candidate associations. These associations were considered significant if their BH FDR was less than 0.1 for both the Spearman's rank correlation test and lasso regression test, and if their stability selection value was greater than 0.6. Note that, since host molecular profiles from UCSC Xena have been normalized across pan-cancer, occasionally, some host features in a given cancer type might have constant values (i.e., *y* has no variance), which were excluded in the association analysis. In addition, association analysis was not performed when the usable number of samples was less than 10 in our study.

**Sequential mediation analysis**

To integrate multi-omics data and uncover potential biological association pathways, we used R packages bruceR (0.8.10) and lavaan (0.6-15) [53] to perform sequential mediation analysis as described in a recent study [54]. Sequential mediation analysis leverages linear models and bootstrapping techniques to assess if independent variable X (e.g, tumor microbiome) influences dependent variable Y via a seris of mediator variables $\mathrm{M}_1 \cdots \mathrm{M}_\mathrm{n}$, such as host methylation (HostM), RNA expression (HostR) and/or protein expression (HostP). We first conducted the mediation analysis to test the microbe-methylation-RNA axis, by fitting two linear models [55]:

$$\mathrm{HostM} = a_1 \cdot \mathrm{Microbe} + \epsilon$$

$$\mathrm{HostR} = a_2 \cdot \mathrm{Microbe} + b_1 \cdot \mathrm{HostM} + \epsilon$$

Upon fitting the linear models, the parameters $a_1$ and $a_2$ were obtained, and their product $(a_1 \cdot a_2)$ represents the average indirect effect of Microbe on HostR. Concurrently, $a_2$ indicates the average direct influence of Microbe on HostR, adjusting for the average indirect effect mediated by Microbe through HostM.

Likewise, for mediation analysis that includes multiple mediator variables, such as the microbe-methylation-RNA-protein sequential mediation analysis model, the tumor microbiome impacts host methylation (HostM), subsequently affecting host gene expression (HostR), and ultimately influencing host protein expression (HostP). The significance and proportional effect sizes of the mediation effects were determined by fitting the following three linear equations:

$$\mathrm{HostM} = a_1 \cdot \mathrm{Microbe} + \epsilon$$

$$\mathrm{HostR} = a_2 \cdot \mathrm{Microbe} + b_1 \cdot \mathrm{HostM} + \epsilon$$

$$\mathrm{HostP} = a_3 \cdot \mathrm{Microbe} + b_2 \cdot \mathrm{HostM} + c_1 \cdot \mathrm{HostR} + \epsilon$$

The product of $(a_1 \cdot b_1 \cdot c_1)$ signifies the average indirect effect of Microbe on HostP mediated by HostM and HostR. The parameter $a_3$ represents the average direct effect of Microbe on HostP, adjusting for the effects via HostM and HostR.

**Gene set and pathway analysis**

We identified gene sets and pathways enriched in genes associated with pan-cancer key microbes and pathogen bacteria. We linked CpG islands to genes using the Illumina official manifest file. We transformed gene SYMBOL to ENTREZID using R packages clusterProfiler and org.Hs.eg.db [56], and performed GO (gene ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway enrichment analysis with BH correction (pvalueCutoff = 0.25). To visualize host-microbe association pathways, we employed R packages ggraph and scatterpie to construct network graphs between host features and microbes. We also used barplot (showCategory = 10) and cnetplot (showCategory = 5 or 10) in the clusterProfiler package to show the enrichment results.

**GEP score analysis**

The T cell-inflamed gene expression profile (GEP) score was calculated based on the sum of 18 inflammatory genes, including *CCL5*, *CD27*, *CD274* (PD-L1), *CD276* (B7-H3), *CD8A*, *CMKLR1*, *CXCL9*, *CXCR6*, *HLA-DQA1*, *HLA-DRB1*, *HLA-E*, *IDO1*,

*LAG3*, *NKG7*, *PDCD1LG2* (PDL2), *PSMB10*, *STAT1*, and *TIGIT* as described in literature before [30]. Further, we tested the associations between GEP scores and microbes based on Spearman's rank correlation and lasso regression tests. Then, we filtered out significant associations (Spearman: FDR < 0.1; lasso: FDR < 0.1, and stability selection > 0.6).

**Immune subtype analysis**

Six immune subtypes spanning multiple tumor types have been classified for 9,127 tumors in TCGA as described in literature [22]. After overlapping with tumors having microbiome profiles, 9,033 samples were included in the analysis. We compared the abundance of lymphocytes-related microbes between different immune subtypes using the Wilcoxon rank sum test.

**Survival analysis**

We obtained curated survival phenotypic data including overall survival (OS) and progression-free survival (PFS) from the UCSC Xena server. We excluded samples with missing OS and PFS data (10,702 samples were retained) [57]. After paired with the tumor microbiome data, 9,483 (Poore et al. dataset) and 7,800 (Dohlman et al. dataset) samples in 32 cancer types were kept for further analysis. Then, samples were divided into high (top 50%) and low (bottom 50%) groups based on the median abundance of microbes in each cancer type. For MOMAC, the prognostic impact of all microbes on OS and PFS were estimated using Kaplan-Meier analysis and hazard ratios were computed using the R package survminer version0.4.9. In this study, we showed the impact on OS and PFS for six pan-cancer key microbes and $p$ values < 0.05 were treated as significant.

**The MOMAC web portal**

The web portal can be accessed at https://comics.med.sustech.edu.cn/momac. The platform is composed of five main parts: MOMAC Search, Advanced Search, Survival, Statistics and Download. Significant candidate associations can be queried

in “MOMAC Search” by inputting gene, microbe and/or cancer type names. The “Advanced Search” function can be used to select candidate associations with more filter parameters. The “Statistics” page shows basic statistics of candidate associations and the “Survival” page provides survival analysis for a given microbe as described above. Furthermore, all summary statistics of our association analysis can be acquired in the “Download” page.

**Data and code availability**

The summary statistics of associations tests have been uploaded to the open-access MOMAC web portal (https://comics.med.sustech.edu.cn/momac), which are available for download to all users. The publicly available microbiome data can be found for fungi (https://github.com/knightlab-analyses/mycobiome/tree/master/Final_files) and other microbes (ftp://ftp.microbio.me/pub/cancer_microbiome_analysis). All code and supplementary data are available via GitHub at https://github.com/comics-bio/pan-cancer-host-microbiome-associations. The corresponding author (Shimin Shuai) can be contacted for further additional information.

**Ethics approval and consent to participate**

Not applicable.

**Competing interests**

The authors have declared no conflict of interest.

**Acknowledgements**

C.M., C.S., J.L. and S.S. were partially supported by the National Natural Science Foundation of China (82270239) and the Science and Technology Program of Shenzhen (JCYJ20220530115207016). C.S. and J.Q. were supported by the Guangdong Province Science and Technology Project (2020B1111170014), the Guangdong Basic and Applied Basic Research Foundation (2020A1515010586), the

Science and Technology Program of Shenzhen (JCYJ20190809144005609 and JCYJ20210324131212034), the Third People's Hospital of Shenzhen (G2021014), and the Shenzhen Key Laboratory of Biochip (ZDSYS201504301534057). Computational analysis of this work was supported by the Center for Computational Science and Engineering at Southern University of Science and Technology.

## Contributions

C.M. and S.S. conceived the project; C.M., C.S. and S.S. analyzed the data, interpreted the results, and wrote the manuscript; J.L. designed the web portal; J.Q. and S.S. supervised the project. All authors read and approved the manuscript.

## References

1. Dohlman, A.B., Klug, J., Mesko, M., Gao, I.H., Lipkin, S.M., Shen, X., and Iliev, I.D. (2022). A pan-cancer mycobiome analysis reveals fungal involvement in gastrointestinal and lung tumors. Cell *185*, 3807-3822 e3812. 10.1016/j.cell.2022.09.015.
2. Poore, G.D., Kopylova, E., Zhu, Q., Carpenter, C., Fraraccio, S., Wandro, S., Kosciolek, T., Janssen, S., Metcalf, J., Song, S.J., et al. (2020). Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature *579*, 567-574. 10.1038/s41586-020-2095-1.
3. Nejman, D., Livyatan, I., Fuks, G., Gavert, N., Zwang, Y., Geller, L.T., Rotter-Maskowitz, A., Weiser, R., Mallel, G., Gigi, E., et al. (2020). The human tumor microbiome is composed of tumor type-specific intracellular bacteria. Science *368*, 973-+. 10.1126/science.aay9189.
4. Narunsky-Haziza, L., Sepich-Poore, G.D., Livyatan, I., Asraf, O., Martino, C., Nejman, D., Gavert, N., Stajich, J.E., Amit, G., Gonzalez, A., et al. (2022). Pan-cancer analyses reveal cancer-type-specific fungal ecologies and bacteriome interactions. Cell *185*, 3789-3806 e3717. 10.1016/j.cell.2022.09.005.
5. Zhang, N., Kandalai, S., Zhou, X., Hossain, F., and Zheng, Q. (2023).

Applying multi-omics toward tumor microbiome research. iMeta *2*. 10.1002/imt2.73.

6. Sepich-Poore, G.D., Zitvogel, L., Straussman, R., Hasty, J., Wargo, J.A., and Knight, R. (2021). The microbiome and human cancer. Science *371*. 10.1126/science.abc4552.

7. Fu, A., Yao, B., Dong, T., Chen, Y., Yao, J., Liu, Y., Li, H., Bai, H., Liu, X., Zhang, Y., et al. (2022). Tumor-resident intracellular microbiota promotes metastatic colonization in breast cancer. Cell *185*, 1356-1372 e1326. 10.1016/j.cell.2022.02.027.

8. Geller, L.T., Barzily-Rokni, M., Danino, T., Jonas, O.H., Shental, N., Nejman, D., Gavert, N., Zwang, Y., Cooper, Z.A., Shee, K., et al. (2017). Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. Science *357*, 1156-1160. 10.1126/science.aah5043.

9. Reginato, E., Wolf, P., and Hamblin, M.R. (2014). Immune response after photodynamic therapy increases anti-cancer and anti-bacterial effects. World J Immunol *4*, 1-11. 10.5411/wji.v4.i1.1.

10. Russo, E., Taddei, A., Ringressi, M.N., Ricci, F., and Amedei, A. (2016). The interplay between the microbiome and the adaptive immune response in cancer development. Ther Adv Gastroenter *9*, 594-605. 10.1177/1756283x16635082.

11. Hiam-Galvez, K.J., Allen, B.M., and Spitzer, M.H. (2021). Systemic immunity in cancer. Nat Rev Cancer *21*, 345-359. 10.1038/s41568-021-00347-z.

12. Cullin, N., Azevedo Antunes, C., Straussman, R., Stein-Thoeringer, C.K., and Elinav, E. (2021). Microbiome and cancer. Cancer Cell *39*, 1317-1341. 10.1016/j.ccell.2021.08.006.

13. Zhang, X., Yu, D., Wu, D., Gao, X., Shao, F., Zhao, M., Wang, J., Ma, J., Wang, W., Qin, X., et al. (2023). Tissue-resident Lachnospiraceae family bacteria protect against colorectal carcinogenesis by promoting tumor immune surveillance. Cell Host & Microbe *31*, 418-432.e418. 10.1016/j.chom.2023.01.013.

14. Xia, X., Wu, W.K.K., Wong, S.H., Liu, D., Kwong, T.N.Y., Nakatsu, G., Yan, P.S., Chuang, Y.M., Chan, M.W., Coker, O.O., et al. (2020). Bacteria pathogens drive host colonic epithelial cell promoter hypermethylation of tumor suppressor genes in colorectal cancer. Microbiome *8*, 108. 10.1186/s40168-020-00847-4.

15. Gao, S., Zhu, D.J., Zhu, J., Shen, L.Q., Zhu, M., and Ren, X.F. (2021). Screening Hub Genes of Hepatocellular Carcinoma Based on Public Databases. Comput Math Method M *2021*. 10.1155/2021/7029130.

16. Qiang, R., Zhao, Z.T., Tang, L., Wang, Q., Wang, Y.H., and Huang, Q. (2021). Identification of 5 Hub Genes Related to the Early Diagnosis, Tumour Stage, and Poor Outcomes of Hepatitis B Virus-Related Hepatocellular Carcinoma by Bioinformatics Analysis. Comput Math Method M *2021*. 10.1155/2021/9991255.

17. Tang, Y.Q., Zhang, Y.Q., and Hu, X. (2020). Identification of Potential Hub Genes Related to Diagnosis and Prognosis of Hepatitis B Virus-Related Hepatocellular Carcinoma via Integrated Bioinformatics Analysis. Biomed Res Int *2020*. 10.1155/2020/4251761.

18. Costa, L., Corre, S., Michel, V., Le Luel, K., Fernandes, J., Ziveri, J., Jouvion, G., Danckaert, A., Mouchet, N., Da Silva Barreira, D., et al. (2020). USF1 defect drives p53 degradation during Helicobacter pylori infection and accelerates gastric carcinogenesis. Gut *69*, 1582-1591. 10.1136/gutjnl-2019-318640.

19. Yang, J., Song, H., Cao, K., Song, J., and Zhou, J. (2019). Comprehensive analysis of Helicobacter pylori infection-associated diseases based on miRNA-mRNA interaction network. Brief Bioinform *20*, 1492-1501. 10.1093/bib/bby018.

20. Luoma, A.M., Suo, S., Wang, Y., Gunasti, L., Porter, C.B.M., Nabilsi, N., Tadros, J., Ferretti, A.P., Liao, S., Gurer, C., et al. (2022). Tissue-resident memory and circulating T cells are early responders to pre-surgical cancer immunotherapy. Cell *185*, 2918-2935 e2929. 10.1016/j.cell.2022.06.018.

21. Petitprez, F., de Reynies, A., Keung, E.Z., Chen, T.W., Sun, C.M., Calderaro, J., Jeng, Y.M., Hsiao, L.P., Lacroix, L., Bougouin, A., et al. (2020). B cells are associated with survival and immunotherapy response in sarcoma. Nature *577*, 556-560. 10.1038/s41586-019-1906-8.

22. Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Ou Yang, T.H., Porta-Pardo, E., Gao, G.F., Plaisier, C.L., Eddy, J.A., et al. (2018). The Immune Landscape of Cancer. Immunity *48*, 812-830 e814. 10.1016/j.immuni.2018.03.023.

23. Saltz, J., Gupta, R., Hou, L., Kurc, T., Singh, P., Nguyen, V., Samaras, D., Shroyer, K.R., Zhao, T., Batiste, R., et al. (2018). Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. Cell Rep *23*, 181-193 e187. 10.1016/j.celrep.2018.03.086.

24. Smith, J., Sen, S., Weeks, R.J., Eccles, M.R., and Chatterjee, A. (2020). Promoter DNA Hypermethylation and Paradoxical Gene Activation. Trends Cancer *6*, 392-406. 10.1016/j.trecan.2020.02.007.

25. Cho, S.Y., Oh, Y., Jeong, E.M., Park, S., Lee, D., Wang, X., Zeng, Q., Qin, H., Hu, F., Gong, H., et al. (2020). Amplification of transglutaminase 2 enhances tumor-promoting inflammation in gastric cancers. Exp Mol Med *52*, 854-864. 10.1038/s12276-020-0444-7.

26. Lei, Z., Chai, N., Tian, M., Zhang, Y., Wang, G., Liu, J., Tian, Z., Yi, X., Chen, D., Li, X., et al. (2018). Novel peptide GX1 inhibits angiogenesis by specifically binding to transglutaminase-2 in the tumorous endothelial cells of gastric cancer. Cell Death Dis *9*, 579. 10.1038/s41419-018-0594-x.

27. Chow, N.H., Chan, S.H., Tzai, T.S., Ho, C.L., and Liu, H.S. (2001). Expression profiles of ErbB family receptors and prognosis in primary transitional cell carcinoma of the urinary bladder. Clin Cancer Res *7*, 1957-1962.

28. Andrews, M.C., Duong, C.P.M., Gopalakrishnan, V., Iebba, V., Chen, W.S., Derosa, L., Khan, M.A.W., Cogdill, A.P., White, M.G., Wong, M.C., et al.

(2021). Gut microbiota signatures are associated with toxicity to combined CTLA-4 and PD-1 blockade. Nat Med *27*, 1432-+. 10.1038/s41591-021-01406-6.

29. Simpson, R.C., Shanahan, E.R., Batten, M., Reijers, I.L.M., Read, M., Silva, I.P., Versluis, J.M., Ribeiro, R., Angelatos, A.S., Tan, J., et al. (2022). Diet-driven microbial ecology underpins associations between cancer immunotherapy outcomes and the gut microbiome. Nat Med. 10.1038/s41591-022-01965-2.

30. Cristescu, R., Mogg, R., Ayers, M., Albright, A., Murphy, E., Yearley, J., Sher, X., Liu, X.Q., Lu, H., Nebozhyn, M., et al. (2019). Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy (vol 363, eaar3593, 2019). Science *363*. 10.1126/science.aax1384.

31. Mao, A.W., Barck, H., Young, J., Paley, A., Mao, J., and Chang, H. (2022). Identification of a novel cancer microbiome signature for predicting prognosis of human breast cancer patients. Clin Transl Oncol *24*, 597-604. 10.1007/s12094-021-02725-3.

32. Zhou, B., Shi, L., Jin, M., Cheng, M., Yu, D., Zhao, L., Zhang, J., Chang, Y., Zhang, T., and Liu, H. (2022). Caulobacter and Novosphingobium in tumor tissues are associated with colorectal cancer outcomes. Front Oncol *12*, 1078296. 10.3389/fonc.2022.1078296.

33. Huang, Y., Zhu, N., Zheng, X., Liu, Y.H., Lu, H.P., Yin, X.C., Hao, H.J., Tan, Y., Wang, D.J., Hu, H., et al. (2022). Intratumor Microbiome Analysis Identifies Positive Association Between Megasphaera and Survival of Chinese Patients With Pancreatic Ductal Adenocarcinomas. Front Immunol *13*. 10.3389/fimmu.2022.785422.

34. Riquelme, E., Zhang, Y., Zhang, L., Montiel, M., Zoltan, M., Dong, W., Quesada, P., Sahin, I., Chandra, V., San Lucas, A., et al. (2019). Tumor Microbiome Diversity and Composition Influence Pancreatic Cancer Outcomes. Cell *178*, 795-806 e712. 10.1016/j.cell.2019.07.008.

35. Zapatka, M., Borozan, I., Brewer, D.S., Iskar, M., Grundhoff, A., Alawi, M.,

Desai, N., Sultmann, H., Moch, H., Pathogens, P., et al. (2020). The landscape of viral associations in human cancers. Nat Genet *52*, 320-330. 10.1038/s41588-019-0558-9.

36. Saigo, K., Yoshida, K., Ikeda, R., Sakamoto, Y., Murakami, Y., Urashima, T., Asano, T., Kenmochi, T., and Inoue, I. (2008). Integration of hepatitis B virus DNA into the myeloid/lymphoid or mixed-lineage leukemia (MLL4) gene and rearrangements of MLL4 in human hepatocellular carcinoma. Hum Mutat *29*, 703-708. 10.1002/humu.20701.

37. Li, X., Zhang, J., Yang, Z., Kang, J., Jiang, S., Zhang, T., Chen, T., Li, M., Lv, Q., Chen, X., et al. (2014). The function of targeted host genes determines the oncogenicity of HBV integration in hepatocellular carcinoma. J Hepatol *60*, 975-984. 10.1016/j.jhep.2013.12.014.

38. Wei, J., Nagy, T.A., Vilgelm, A., Zaika, E., Ogden, S.R., Romero-Gallo, J., Piazuelo, M.B., Correa, P., Washington, M.K., El-Rifai, W., et al. (2010). Regulation of p53 tumor suppressor by Helicobacter pylori in gastric epithelial cells. Gastroenterology *139*, 1333-1343. 10.1053/j.gastro.2010.06.018.

39. Wei, J., Noto, J.M., Zaika, E., Romero-Gallo, J., Piazuelo, M.B., Schneider, B., El-Rifai, W., Correa, P., Peek, R.M., and Zaika, A.I. (2015). Bacterial CagA protein induces degradation of p53 protein in a p14ARF-dependent manner. Gut *64*, 1040-1048. 10.1136/gutjnl-2014-307295.

40. Sepulveda, A.R., Yao, Y., Yan, W., Park, D.I., Kim, J.J., Gooding, W., Abudayyeh, S., and Graham, D.Y. (2010). CpG methylation and reduced expression of O6-methylguanine DNA methyltransferase is associated with Helicobacter pylori infection. Gastroenterology *138*, 1836-1844. 10.1053/j.gastro.2009.12.042.

41. Ghaddar, B., Biswas, A., Harris, C., Omary, M.B., Carpizo, D.R., Blaser, M.J., and De, S. (2022). Tumor microbiome links cellular programs and immunity in pancreatic cancer. Cancer Cell *40*, 1240-1253 e1245. 10.1016/j.ccell.2022.09.009.

42. Suzuki, M.M., and Bird, A. (2008). DNA methylation landscapes: provocative

insights from epigenomics. Nat Rev Genet *9*, 465-476. 10.1038/nrg2341.

43. Allen, J., and Sears, C.L. (2019). Impact of the gut microbiome on the genome and epigenome of colon epithelial cells: contributions to colorectal cancer development. Genome Med *11*, 11. 10.1186/s13073-019-0621-2.

44. Wippermann, A., Rupp, O., Brinkrolf, K., Hoffrogge, R., and Noll, T. (2017). Integrative analysis of DNA methylation and gene expression in butyrate-treated CHO cells. J Biotechnol *257*, 150-161. 10.1016/j.jbiotec.2016.11.020.

45. Goldman, M.J., Craft, B., Hastie, M., Repecka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A.N., et al. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol *38*, 675-678. 10.1038/s41587-020-0546-8.

46. Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc *4*, 1184-1191. 10.1038/nprot.2009.97.

47. Zaitsev, A., Chelushkin, M., Dyikanov, D., Cheremushkin, I., Shpak, B., Nomie, K., Zyrin, V., Nuzhdina, E., Lozinsky, Y., Zotova, A., et al. (2022). Precise reconstruction of the TME using bulk RNA-seq and a machine learning algorithm trained on artificial transcriptomes. Cancer Cell *40*, 879-894 e816. 10.1016/j.ccell.2022.07.006.

48. Witten, D.M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics *10*, 515-534. 10.1093/biostatistics/kxp008.

49. Priya, S., Burns, M.B., Ward, T., Mars, R.A.T., Adamowicz, B., Lock, E.F., Kashyap, P.C., Knights, D., and Blekhman, R. (2022). Identification of shared and disease-specific host gene-microbiome associations across human diseases using multi-omic integration. Nat Microbiol *7*, 780-795. 10.1038/s41564-022-01121-z.

50. Ash, J.T., Darnell, G., Munro, D., and Engelhardt, B.E. (2021). Joint analysis

of expression levels and histological images identifies genes associated with tissue morphology. Nat Commun *12*, 1609. 10.1038/s41467-021-21727-x.

51. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw *33*, 1-22.

52. Hofner, B., Boccuto, L., and Goker, M. (2015). Controlling false discoveries in high-dimensional situations: boosting with stability selection. BMC Bioinformatics *16*, 144. 10.1186/s12859-015-0575-3.

53. Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software *48*, 1-36. DOI 10.18637/jss.v048.i02.

54. Yan, Z., Chen, B., Yang, Y., Yi, X., Wei, M., Ecklu-Mensah, G., Buschmann, M.M., Liu, H., Gao, J., Liang, W., et al. (2022). Multi-omics analyses of airway host-microbe interactions in chronic obstructive pulmonary disease identify potential therapeutic interventions. Nat Microbiol *7*, 1361-1375. 10.1038/s41564-022-01196-8.

55. VanderWeele, T.J. (2016). Mediation Analysis: A Practitioner's Guide. Annu Rev Public Health *37*, 17-32. 10.1146/annurev-publhealth-032315-021402.

56. Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284-287. 10.1089/omi.2011.0118.

57. Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J., Benz, C.C., Levine, D.A., Lee, A.V., et al. (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell *173*, 400-416 e411. 10.1016/j.cell.2018.02.052.