A pan-cancer analysis of the microbiome in metastatic cancer

Battaglia, Thomas W.; Mimpen, Iris L.; Traets, Joleen J.H.; van Hoeck, Arne; Zeverijn, Laurien J.; Geurts, Birgit S.; de Wit, Gijs F.; Vos, Joris L.; Wessels, Lodewyk; More Authors

DOI
10.1016/j.cell.2024.03.021

Publication date
2024

Document Version
Final published version

Published in
Cell

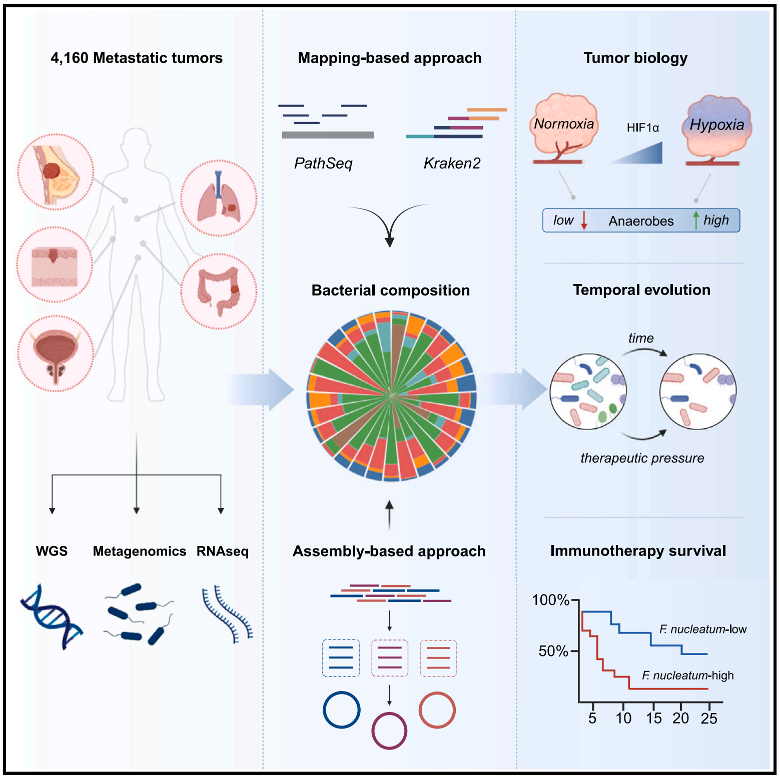
Citation (APA)
Battaglia, T. W., Mimpen, I. L., Traets, J. J. H., van Hoeck, A., Zeverijn, L. J., Geurts, B. S., de Wit, G. F., Vos, J. L., Wessels, L., & More Authors (2024). A pan-cancer analysis of the microbiome in metastatic cancer. *Cell*, *187*(9), 2324-2335.e19. https://doi.org/10.1016/j.cell.2024.03.021

Cell

Resource

A pan-cancer analysis of the microbiome in metastatic cancer

Graphical abstract



Highlights

- Tumor-resident bacteria DNA is detectable in a pan-cancer metastasis cohort
- Assembly of tumor-derived bacterial DNA reveals species-level genomic characterization
- Bacterial diversity is associated with cellular and molecular tumor immunity features
- Fusobacterium DNA abundance indicates poor immunotherapy response in an NSCLC cohort

Authors

Thomas W. Battaglia, Iris L. Mimpen, Joleen J.H. Traets, ..., Lodewyk Wessels, Joris van de Haar, Emile Voest

Correspondence

e.voest@nki.nl

In brief

Characterization of microbiome genomes at the species level in over 4,000 metastatic tumor biopsies identifies the distribution and diversity features of tumor-resident bacterial DNA at a pan-cancer scale, highlighting the associations between microbial community dynamics and tumor immunity and immunotherapy efficacy.

Battaglia et al., 2024, Cell *187*, 2324–2335
April 25, 2024 © 2024 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cell.2024.03.021

CellPress
OPEN ACCESS

Cell

Resource

A pan-cancer analysis of the microbiome in metastatic cancer

Thomas W. Battaglia,[1,2,3,13] Iris L. Mimpen,[1,3,13] Joleen J.H. Traets,[4] Arne van Hoeck,[3,8] Laurien J. Zeverijn,[1,3] Birgit S. Geurts,[1,3] Gijs F. de Wit,[1,3] Michaël Noë,[5] Ingrid Hofland,[6] Joris L. Vos,[4,7] Sten Cornelissen,[6] Maartje Alkemade,[6] Annegien Broeks,[6] Charlotte L. Zuur,[4,8,9] Edwin Cuppen,[10,11] Lodewyk Wessels,[2,3,12,14] Joris van de Haar,[1,3,14] and Emile Voest[1,3,14,15,*]

[1]Division of Molecular Oncology & Immunology, the Netherlands Cancer Institute, Amsterdam 1066 CX, the Netherlands
[2]Division of Molecular Carcinogenesis, the Netherlands Cancer Institute, Amsterdam 1066 CX, the Netherlands
[3]Oncode Institute, the Netherlands Cancer Institute, Amsterdam 1066 CX, the Netherlands
[4]Division of Tumor Biology & Immunology, the Netherlands Cancer Institute, Amsterdam 1066 CX, the Netherlands
[5]Department of Pathology, Antoni van Leeuwenhoek/the Netherlands Cancer Institute, Amsterdam, the Netherlands
[6]Core Facility Molecular Pathology & Biobanking, the Netherlands Cancer Institute, Amsterdam 1066 CX, the Netherlands
[7]Head and Neck Service and Immunogenomic Oncology Platform, Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[8]Department of Head and Neck Surgery and Oncology, the Netherlands Cancer Institute, 1066 CX Amsterdam, the Netherlands
[9]Department of Otorhinolaryngology Head and Neck Surgery, Leiden University Medical Center, Leiden, the Netherlands
[10]Center for Molecular Medicine, University Medical Centre Utrecht, Utrecht 3584CX, the Netherlands
[11]Hartwig Medical Foundation, Science Park, Amsterdam 1098XH, the Netherlands
[12]Faculty of EEMCS, Delft University of Technology, Delft 2628 CD, the Netherlands
[13]These authors contributed equally
[14]These authors contributed equally
[15]Lead contact
*Correspondence: e.voest@nki.nl

https://doi.org/10.1016/j.cell.2024.03.021

SUMMARY

Microbial communities are resident to multiple niches of the human body and are important modulators of the host immune system and responses to anticancer therapies. Recent studies have shown that complex microbial communities are present within primary tumors. To investigate the presence and relevance of the microbiome in metastases, we integrated mapping and assembly-based metagenomics, genomics, transcriptomics, and clinical data of 4,160 metastatic tumor biopsies. We identified organ-specific tropisms of microbes, enrichments of anaerobic bacteria in hypoxic tumors, associations between microbial diversity and tumor-infiltrating neutrophils, and the association of Fusobacterium with resistance to immune checkpoint blockade (ICB) in lung cancer. Furthermore, longitudinal tumor sampling revealed temporal evolution of the microbial communities and identified bacteria depleted upon ICB. Together, we generated a pan-cancer resource of the metastatic tumor microbiome that may contribute to advancing treatment strategies.

INTRODUCTION

An intricate and dynamic crosstalk exists between the host immune system and the commensal microbiota comprising bacteria, viruses, fungi, and archaea. Microbes assist in the education of the developing immune system, maintain mucosal barrier integrity, and produce short-chain fatty acids (SCFAs) through gut microbial fermentation. As such, recent studies have begun to deconvolve the composition and function of the microbiome in relation to cancer initiation and response to therapy.[1–4] The gut microbiome, which harbors the largest diversity of microbial species, is able to modulate responses to immune checkpoint blockade (ICB) and traditional chemotherapies[5–8] and have been implicated in instances of molecular mimicry with tumor neoantigens.[9] Furthermore, early evidence has indicated that fecal microbiota transplants (FMTs) may enhance clinical responses to ICB.[10,11]

Microbial populations in primary tumors have been demonstrated to be distinct from those in non-malignant tissue,[12] can reside intracellularly,[12,13] are capable of tumor microenvironment (TME) reprogramming[14] and can be successfully presented on human leukocyte antigen (HLA)-I and HLA-II complexes.[13] However, the biological and clinical relevance of the microbiome in metastases is still unclear. As metastases are responsible for the majority of solid cancer deaths[15] and are key targets for systemic anti-cancer therapies, there is an urgent need to understand how tumor-resident bacteria can reshape tumor biology, immune infiltration, and responsiveness to treatment.

Bacteria can induce both an immunosuppressed or inflamed microenvironment, and understanding these processes in tumors may help explain the heterogeneity of response to ICB. Yet, several questions still need to be addressed: is the microbial community in the primary tumor different from those in metastases? Are bacteria preferentially localized in specific tumor types or rather in specific organs? How resilient is the microbial population during therapeutic interventions?

Given the complexity of identifying intracellular bacteria and the highly complex structure of the microbiome, we used a bioinformatics approach to investigate characteristics, evolution, and relevance of the metastatic cancer microbiome in patients. We analyzed biopsies from 4,160 metastatic tumors across various cancers originating from 26 tissues with matched genomics and transcriptomic profiling. Moreover, we assessed the persistence of the metastatic microbiome over time and under therapeutic pressure in a unique cohort of 185 paired biopsies of metastatic cancers.

RESULTS

Mining the microbiome of metastases

In a range of metastatic cancer types, we first characterized the tumor-resident microbiome communities at the genus level by extracting unmapped reads and applying two independent computational methods, Kraken2 and PathSeq. We further characterized these communities at the species level by employing a metagenomic assembly-based approach. We explored 4,164 whole-genome sequencing (WGS), pre-treatment tumor biopsies from the Hartwig Medical Foundation (Hartwig) cohort[16] that included 26 cancer types. Using a mapping-based approach and filtering genera for technical contaminants (as previously described[17]) and infrequently observed genera (STAR Methods), we cataloged 165 genera among 3,526 samples of which 49% (81/165) were gram-negative and 68% (112/165) were anaerobic/facultative-anaerobes (Figure 1A; STAR Methods). Amplicon sequencing (16S rRNA) on a subset of matched tumor tissue correlated with WGS-based bacterial profiles (Figure S2A). Furthermore, we were able to assemble 514 near-high-quality and medium-quality metagenomic-assembled genomes (MAGs, as defined by others[18,19]) from tumor-derived bacterial reads (Figures 1B and S3A–S3G; Table S2). The most prevalent cancer types were breast, colorectal, lung, prostate, and melanoma, while liver, lymph node, and lung were the most frequent metastatic sites from which the tumor biopsies were obtained (Figures 1B and S2B). The predominant treatment class directly following tumor biopsy was chemotherapy (28%), followed by immunotherapy (12%) and targeted therapy (12%). In total, 9.5×10^{10} unmapped sequences were processed at an average of 2.3×10^{7} reads per sample with 4.7×10^{4} genus-level resolved reads (Figure S2C).

Our analysis revealed differences in the number of bacterial-derived reads as a fraction of human-mapped reads (fractional reads) among cancer types and detected a higher fraction in uterine and renal cancers and a lower load in primary central nervous system (CNS)-derived cancers (Figure 1C). Moreover, we found higher fractional reads in colorectal and renal metastases (Figure S2D). Biodiversity metrics, including the total number of genera per sample (richness) and relative differences in abundance within a community (evenness), revealed a generally rich and balanced community across multiple cancer types, (Figure S2E), with the exception of CNS cancers, which had a less-rich community. Moreover, colorectal metastases had the richest diversity, while head and neck metastases had more dominant genera.

Shaping the metastatic tumor microbiome

We first sought to determine which factors shaped metastatic tumor microbiome composition. Interestingly, we found that the microbial community composition was more strongly impacted by the anatomical site of the biopsied lesion as compared with the primary tumor tissue of origin (Figure 2A). This may suggest a permissiveness of bacteria to organ-specific niches. As our metastatic tumor cohort provides the opportunity to deconvolve the impact of cancer type versus metastatic localization, we next sought to assess cancer-type-specific microbial populations. To do this, we computed pairwise microbial community similarities between all cancer types and determined community-level differences while accounting for localization site and other potential confounders (Figure 2B; Table S3). Specifically, we observed primary CNS cancers, mesothelial cancers, and head and neck cancers to differ strongly in their microbial compositions compared with other cancer types. Between metastatic localizations (Figure S4A), we found colorectal biopsies to have the strongest deviance between localizations.

To further interrogate potential factors that could shape the metastatic microbiome, we hypothesized that hypoxic tumors may harbor more anaerobic bacteria as these bacteria are well adapted for low-oxygen environments. To test this, we measured the degree of hypoxia using the hallmark hypoxia gene signature in a set of metastatic tumors with matched RNA sequencing (RNA-seq) data (n = 2,358, Figures S4B and S4C; Table S4) and performed gene set enrichment analysis (GSEA). For anaerobic bacteria, we found a significant positive enrichment for hypoxia (normalized enrichment score [NES] = 1.72, $p < 0.001$) and a negative enrichment for response to oxygen (NES = -1.76, $p < 0.001$), while we observed the opposite effect for aerobic bacteria (Figure 2C). As an independent validation of our findings, we analyzed an external cohort of human papillomavirus (HPV)-negative head and neck tumors (n = 31 patients[20]) and found an enrichment of anaerobic bacteria in more hypoxic tumors, as determined by hypoxia-inducible factor-1 alpha (HIF-1α) staining (Figures 2D and 2E). Furthermore, we were able to confirm this association using our assembly-based MAG catalog (Figure S3H).

We next hypothesized that distinct cancer phenotypes driven by genomic alterations have[21] preferred community compositions; therefore, we performed community detection analysis[21,22] (Figure S4E) on a pan-cancer set of microsatellite instability (MSI) or stable (MSS) tumors (n = 191). Our analysis revealed two discrete clusters that significantly separated according to MSI status (logit model adjusted for cancer type and metastatic localization, p = 0.03, Figure 2F), whereby cluster A was composed mainly of MSI patients and cluster B of MSS. Furthermore, we found that MSS patients with a cluster A-type

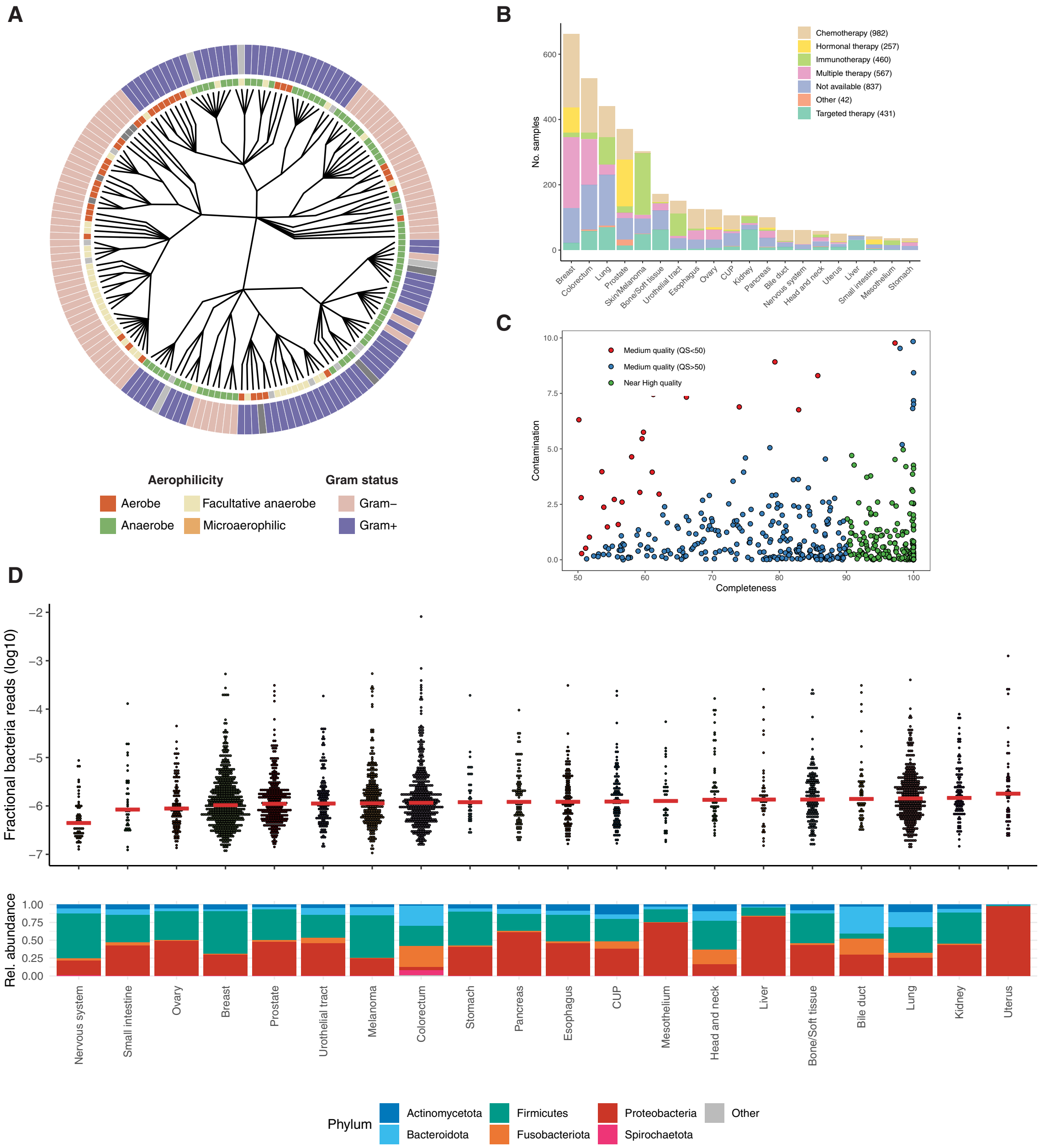


Figure 1. Computationally profiling the tumor microbiome of 4,160 metastatic cancer samples

(A) Phylogenetic tree of genera detected from metastatic tumors using whole-genome sequencing (WGS) data annotated by aerophilicity (inner ring, aerobic [red], anaerobic [green], facultative anaerobic [yellow], or microaerophilic [orange]) and Gram staining status (outer ring, gram positive [purple] or gram negative [pink]).

(B) Overview of cancer types and treatments within the cohort. The bar plot indicates the number of samples per primary tumor location, whereas colors denote the class of treatment given right after the biopsy.

(C) Overview of the metagenomic-assembled genome (MAG) completeness and contamination after assembly and bin splitting (STAR Methods). Colors correspond to MAG's annotated as near-high quality (green, $n = 242$), quality score > 50 (blue, $n = 248$), and quality score < 50 (red, $n = 24$).

(D) Boxplots of the fraction of bacterial-derived reads from the total number of human-mapped reads (fractional bacterial reads), stratified by primary tumor location (top) and the phylum-level average relative composition of the bacterial community within a primary tumor location (bottom). The red line indicates median for each group.

See Figures S1–S3.

composition had a significant increase in interferon (IFN)γ signaling ($p = 0.003$), resembling an MSI phenotype and points to a potential association between distinctive tumor phenotypes and microbiome community structure in a pan-cancer setting. In summary, these results reveal biological elements driving microbial compositions of tumor metastases.

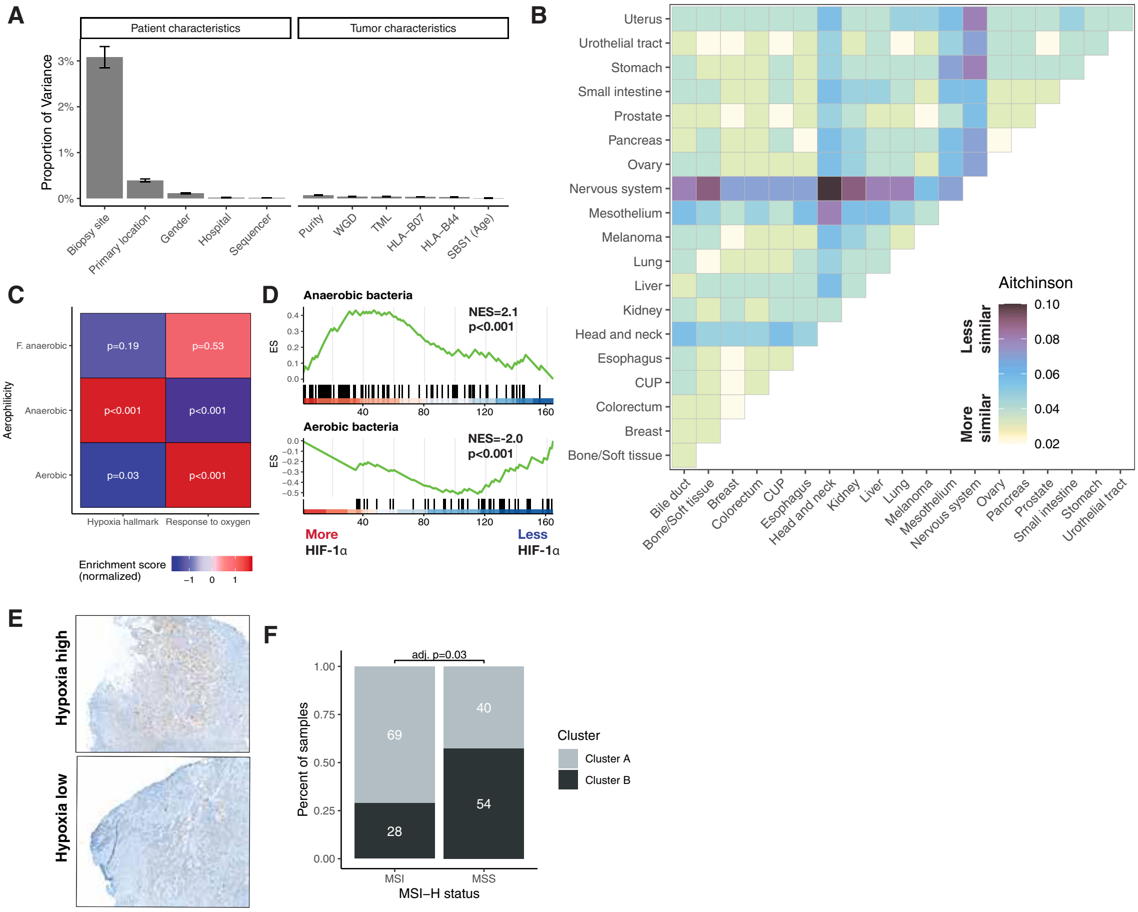


Figure 2. Characteristics of the tumor microbiome in metastatic cancer

(A) Bar plots of the proportion of bacterial abundance variance explained by patient characteristics. Bar size represents the mean variance per genus and error bars show standard deviation.

(B) Pairwise cancer-type-specific community Aitchison dissimilarities. *p* values were generated using the permutational multivariate ANOVA (PERMANOVA) adonis test adjusted for biopsy site, sequencing platform, and hospital of origin.

(C) Normalized enrichment scores (NESs) derived from GSEA analysis relating bacteria and tumor scores of hypoxia or response to oxygen in aerobic, anaerobic, or facultative anaerobic (*F. anaerobic*) bacteria. *p* values were generated using *fgsea*.

(D and E) (D) Same as (C) but represented by GSEA running enrichment scores (ESs) of anaerobic (top) or aerobic (bottom) bacteria on an independent validation cohort, with hypoxia determined by HIF-1α staining as depicted in (E).

(F) Community typing of microsatellite-stable (MSS, $n = 94$) or microsatellite-instable (MSI, $n = 97$) tumors, assigned into two groups (cluster A [gray] and cluster B [black]) using graph-based clustering (*bluster*) on center log ratios (clr). Significance was determined using a logit model of cluster groupings by MSI status after correction by primary tumor location, biopsy site, and hospital of origin.

See Figure S4.

Tumor-resident microbial communities are associated with tumor biology

After establishing a link between general tumor characteristics and the tumor microbiome, we next investigated whether microbial communities can influence host immunity and hence the TME. We first investigated if the presence of gram-negative bacteria in metastases was correlated with expression of the family of Toll-like receptors (TLRs). Lipopolysaccharides (LPSs), a major component of the cell wall in gram-negative bacteria, are potent ligands for TLR4 and lead to activation of the TLR4/MD2 complex, thereby inducing the production of proinflammatory cytokines, such as interleukin (IL)-6, tumor necrosis factor alpha (TNF-α), and type-I IFNs.[23] We hypothesized that LPS, derived from dead or living bacteria, may be an important factor of TLR4 signaling in metastases. To test this, we defined an LPS load and found a significant association between LPS load and TLR4 signaling (linear mixed effect [LME]-$p = 0.02$) but not gram-positive lipoteichoic acid (LTA) load (LME-$p = 0.18$) (Figure S5A). Conversely, we found an association between LTA load and TLR2 signaling (LME-$p = 0.04$, Figure S5B).

Furthermore, we hypothesized that bacterial compositions would be broadly correlated with tumor gene expression. In line with this, we found a significant correspondence between tumor microbiome composition and several MSigDB hallmark signatures (Mantel's test $p = 0.05$) and immune marker gene

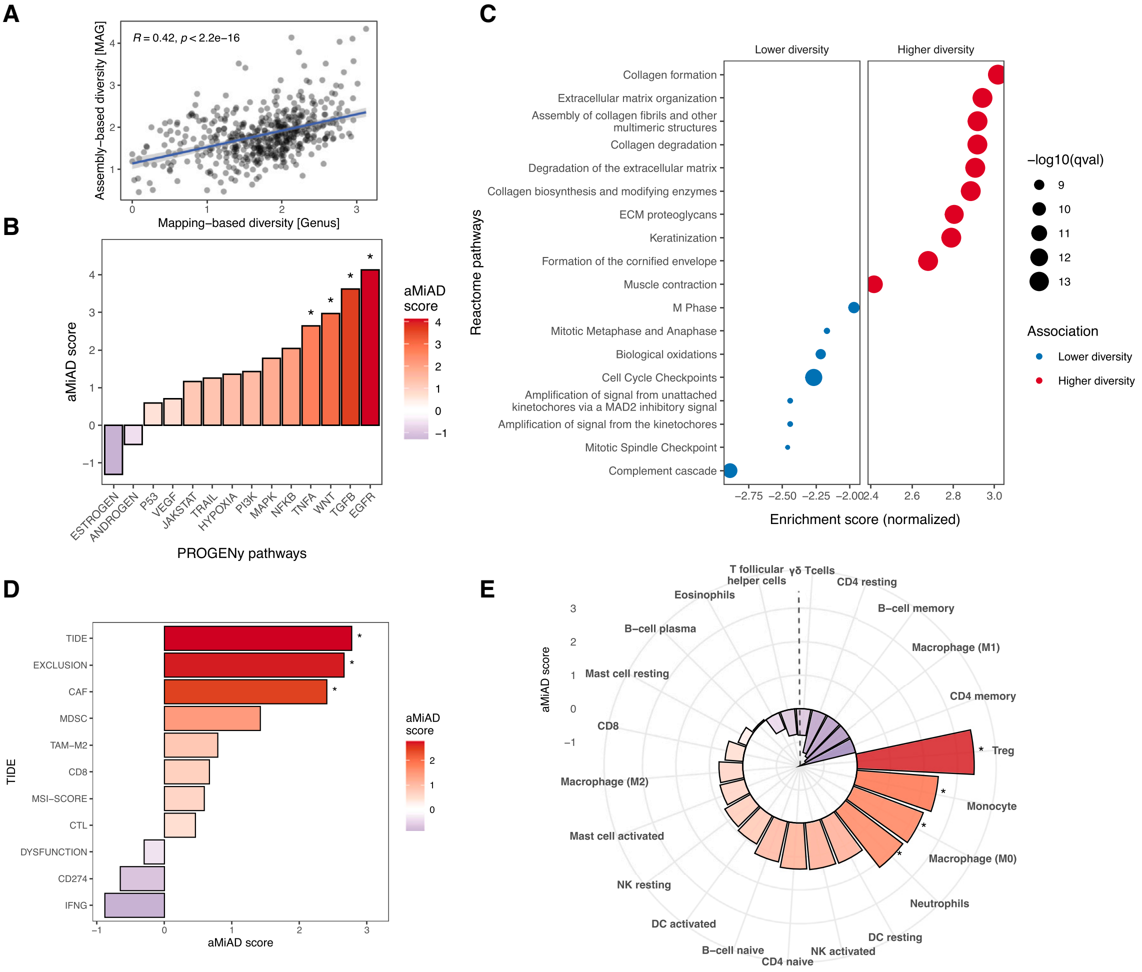


Figure 3. Associations of the microbiome with tumor physiology

(A) Shannon diversity estimates of mapping-based (genus) (x axis) and assembly-based (MAG) (y axis) approaches in a set of tumor microbiomes ($n = 500$). p value was computed using Spearman correlation.

(B) Bar plot of adaptive microbiome α-diversity-based association analysis (aMiAD) scores associating tumor community diversity and pathway responsive gene activity inference (PROGENy) pathway quantifications. The shade of the bar indicates the aMiAD scores for each feature. Positive associations are colored red, and negative associations are colored blue. Asterisks correspond to p values < 0.05. p values were generated with aMiAD after correcting for primary tumor location, metastatic localization site, sequencing platform, and hospital of origin with 5,000 permutations.

(C) Dot plot showing the top Reactome pathways significantly enriched (red) or depleted (blue) in tumors by community diversity, using GSEA on Reactome pathways (STAR Methods). Dot size represent false discovery rate (FDR)-corrected p values (q value).

(D and E) (D) Associations between community diversity and tumor immune dysfunction and exclusion (TIDE) signatures or (E) CIBERSORT deconvolved immune cell types as described in (B).

See Figures S5 and S6.

expression (Mantel's test $p = 0.002$) variation. Given this intersection and our observation that MAG-based diversity estimates correlated with mapping-based estimates (Figure 3A), we then proceeded to associate bacterial diversity with the expression of curated tumor-specific responsive pathways to determine pan-cancer characteristics correlated with the presence and abundance of bacteria. We observed an activation of epidermal growth factor receptor (EGFR), transforming growth factor β (TGF-β), and TNF-α signaling in tumors with a high-diversity microbiome (Figures 3B and S5C). We then performed an unbiased gene expression comparison to determine which processes were activated or inhibited with increased bacterial diversity. Interestingly, we found a strong enrichment of multiple pathways involved in extracellular matrix (ECM) organization and antimicrobial peptides (AMPs) in tumors with a higher bacterial diversity (Figures 3C and S5D).

Bacteria and their byproducts can influence innate and adaptive immune cell populations, yet it remains unclear which microbes can reshape immune cell context and alter immune signaling in tumor metastases. Therefore, we first associated tumor immune evasion mechanisms with bacterial diversity. We found a positive association of bacterial diversity with expression of signatures for cancer-associated fibroblast (CAF) infiltration and immune exclusion (Figures 3D and S5E). Following this,

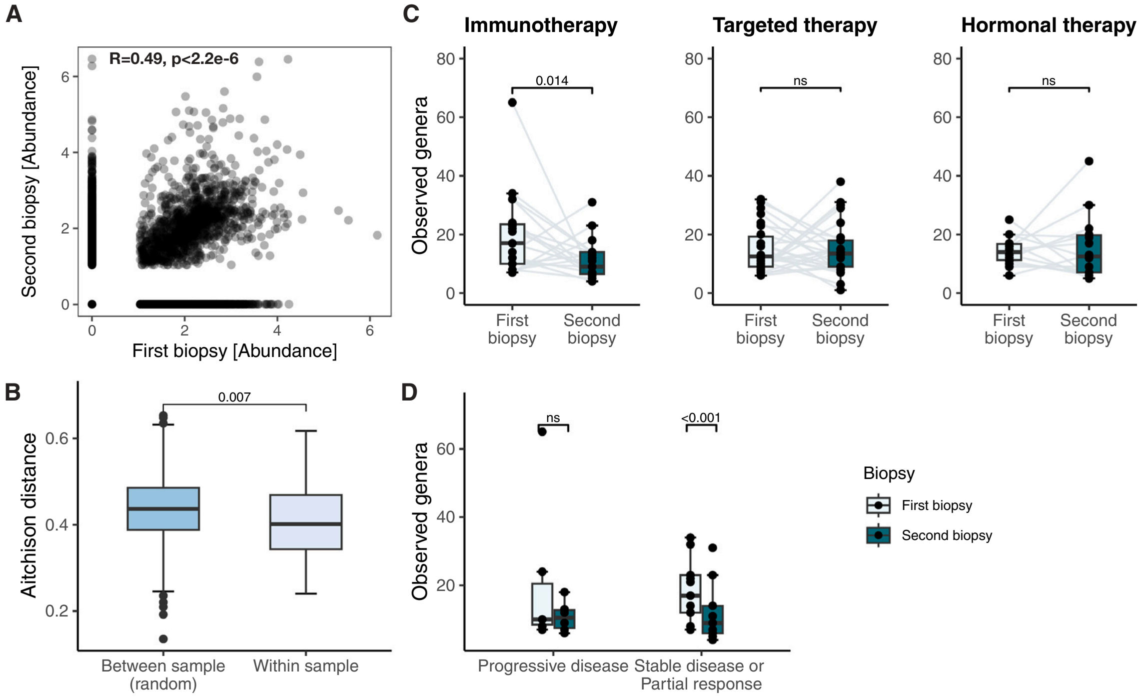


Figure 4. Dynamics of the tumor microbiome over the course of systemic anticancer treatment

(A) Scatterplot of the abundance ($\log_{10}$) between tumor pairs between the first (x axis) and second (y axis) biopsy. p values shown were computed using Spearman correlation.

(B) Boxplots of the Aitchison community dissimilarities between paired biopsies derived from the same lesion ($n = 60$) compared with random sampling tumor pairs ($n = 500$). p values were calculated using the non-parametric Wilcoxon signed-rank test.

(C) Boxplot of bacterial richness derived from paired biopsies of individual patients sampled before (first biopsy, [light green]) and after treatment (second biopsy, [dark green]), according to the type of treatment received.

(D) Boxplot of bacterial richness of patients receiving immunotherapy between the first and second biopsy, stratified based on whether the patients were non-responsive (progressive disease [PD] according to RECIST1.1; left) or responsive (stable disease [SD], partial response [PR], or complete response [CR]; right). The paired p values shown in (C) and (D) were determined using the *betta_random* framework, adjusted for time between biopsies, primary tumor location, and lesion status.

See also Figure S7.

we next examined deconvolved tumor-infiltrating immune cell compositions and found an extensive enrichment of innate immune cells driven by neutrophils, natural killer (NK) cells, and macrophages, as well as regulatory T cells (Tregs), in tumors with high microbial diversity (Figures 3E and S5F), highlighting a potential innate immune conditioning. Importantly, there is emerging evidence showing the functional importance of neutrophils in tumor immunity, progression, and response to immunotherapy.[24,25]

Lastly, we sought to identify individual bacteria-immune cell associations. To do this, we correlated bacteria abundance with immune cell functional states[26] (Figure S6) and found multiple bacteria associated with cellular states of fibroblasts, endothelial cells, macrophages, and CD4 T cells. Specifically, we found *Bifidobacterium* to be positively associated with NK cells (state S2), which has been previous shown to enhanced NK cell function in the TME.[27]

In summary, our data revealed strong associations between the microbiome and immune system but also reveal a deeper biological understanding of host-microbe interactions in metastatic tumors.

Temporal evolution of microbial communities in metastases

We next sought to better understand the impact of metastatic heterogeneity and the persistence of tumor-resident microbes over time. Therefore, we investigated a set of 185 pairs (370 repetitive tumor biopsies obtained from 173 individual patients) with a median time between biopsies of 7.5 months[28] and comparable tumor purities (R = 0.41, $p < 0.001$, Figure S7A). Quantitatively, bacterial abundances were generally well-correlated between paired biopsies (Figures 4A and S7B) (median Spearman R = 0.52, interquartile range [IQR], 0.45–0.61). We then investigated community similarity between pairs derived from the same ($n = 61$ pairs) or random ($n = 250$ pairs) and observed a significantly higher similarity of tumor pairs derived from the same lesion (Figures 4B and S7C), highlighting statistically significant community preservation. Furthermore, we profiled strain-level diversity of tumor microbiomes and subsequently assessed the degree of shared similarity for each tumor pair. In doing so, we found a significant increase in the degree of similarity in paired biopsies compared with random pairs (Figure S3I).

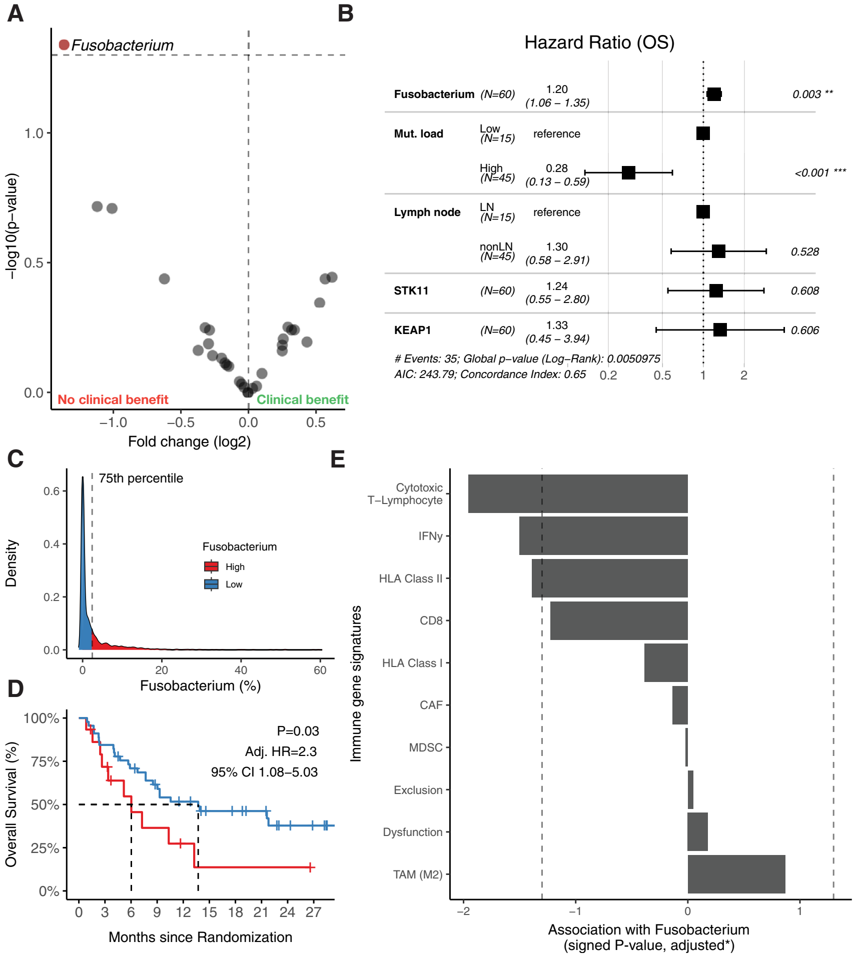


Figure 5. *Fusobacterium* presence is negatively associated with response to ICB in NSCLC

(A) Differential abundant genera in NSCLC patients experiencing a durable clinical benefit (DCB, $n = 23$) or not (NCB, $n = 37$) after treatment with ICB. p values were generated using ANCOMBC, with biopsy site as fixed effects and sequencing platform and hospital of origin as random effects. Significant bacteria ($p < 0.05$) are colored red.

(B) Multivariate Cox proportional-hazards model of overall survival (OS) modeling the abundance of *Fusobacterium*, genome-wide mutational load, lymph node metastatic location status, and presence of resistance genes, STK11 and KEAP1.

(C) Relative abundance distribution of *Fusobacterium* across the metastatic pan-cancer cohort ($n = 3{,}576$) and assigned into either a high (>75th percentile relative abundance [Fuso-high, red]) or low (<75th percentile [Fuso-low, blue]) degree of *Fusobacterium*. Dotted line corresponds to the 75th percentile cutoff.

(D) Overall survival (OS) of NSCLC patients treated with immunotherapy assigned to Fuso-high (red) or Fuso-low (blue) groups. p value, adjusted hazard ratio (HR), and 95% confidence interval (CI) were generated using a multivariate Cox proportional-hazards model as described in (B).

(E) Bar plot representing the degree of cytolytic activity,[30] IFNG signatures,[31] and MHC class I/II[32] and TIDE immune expression in all metastatic lung tumors ($n = 231$) harboring a high or low degree of *Fusobacterium*. Signed p values shown were generated using an LME model correcting for genome-wide mutation burden, tumor subtype, biopsy site, sequencing platform, and hospital origin.

See also Figure S8.

Immunotherapy can reshape tumor microbial populations

We then wanted to explore the impact of therapeutic pressures on the bacterial community. Therefore, we compared changes in bacterial richness before (first biopsy) and after (second biopsy) tumors were exposed to either immunotherapy ($n = 33$), targeted therapy ($n = 34$), or hormonal therapy ($n = 19$) (Figure S7D). We found a significant decrease in bacterial richness exclusively in patients treated with ICB. This was seen by both mapping and assembly-based diversity estimates (Figures 4C and S3J). Moreover, we found that chemotherapies with known antibiotic mode of actions (such as, doxorubicin/epirubicin and bleomycin) trended toward a decrease in bacterial richness after treatment (Figures S7E–S7G).

Furthermore, in a small subset, we found a significant decrease in richness associated with a favorable response (best overall response [BOR], partial response [PR]/stable disease [SD], $n = 14$) compared with non-responsive patients (progressive disease [PD], $n = 9$) (Figure 4D). Based on these observations, we next analyzed (1) which bacteria were recurrently depleted after ICB in responsive patients and (2) whether these bacteria had a lower prevalence in non-responsive patients at pre-treatment. In this with this, we found several bacteria, including *Actinomyces*, *Bacteroidetes*, *Bifidobacterium*, and *Prevotella* to be frequently depleted (Figures S7H and S7I). Moreover, it has been demonstrated that bacterial peptides derived from these genera were presented on tumor HLA machinery.[13] Taken together, our analyses showed an immunotherapy-directed effect on the metastatic tumor microbiome that remodels community structure and is more pronounced in responsive patients.

Fusobacterium is associated with a poor response to ICB in NSCLC

Finally, we next sought to determine pre-treatment bacterial populations affiliated with a lack of response to ICB given the capacity of some microbes to shape an immunosuppressive microenvironment.[14] To do this, we performed an unbiased analysis testing whether bacteria were differentially abundant between patients with and without durable clinical benefit (DCB, progression-free survival [PFS] > 6 months) in an ICB-monotherapy cohort of NSCLC patients ($n = 63$).[29] After adjustment for potential confounders, we found that the abundance of *Fusobacterium* was negatively associated with DCB (ANCOMBC $p = 0.045$) (Figure 5A). A multivariate Cox proportional-hazards model showed that reduced overall survival (OS) and PFS was significantly associated to the continuous abundance of *Fusobacterium* after correcting from genome-wide mutational burden (Figures 5B and S8A). Using our pan-cancer resource, we then designated all tumors into Fuso-high or Fuso-low groups based on an upper quartile relative abundance cutoff comparable to previously defined thresholds[3] (Figure 5C). Using these designations, we were able to corroborate a worse OS and PFS in Fuso-high ICB-treated NSCLC tumors (Figures 5D, S8B, and S8C). Species-level quantification of *Fusobacterium nucleatum* (*F. nucleatum*) recapitulated genus-level designations (Figure S3K). Furthermore, we found that Fuso-high tumors had significantly lower cytotoxic, IFNG, and major histocompatibility complex (MHC) class II gene expression signatures (Figure 5E).

Taken together, our analyses revealed clear effects of immunotherapy treatment and efficacy on the composition of the metastatic tumor microbiome and demonstrate that the presence of *F. nucleatum* is associated with reduced effectiveness of ICB in metastatic NSCLC.

DISCUSSION

In this study, we integrated metagenomics, genomics, and transcriptomics to profile the metastatic tumor microbiome and—at the same time—resolve physiological characteristics of the tumor and its TME to provide biological insight and guide the development of bacteria-oriented strategies to complement and enhance cancer (immuno)therapy.

Our study shows that the composition of the tumor-resident microbiome is associated with cancer hallmarks such as hypoxia and inflammation and may influence immune cell infiltration and immunotherapy efficacy. We obtained insights showing that the metastatic microbiome is (1) to an important extent composed by anaerobic bacteria, (2) influenced by anatomical site, (3) partly maintained between metastatic lesions over time, (4) can be reshaped after treatment with immunotherapy, (5) and could predict ICB treatment efficacy. This pan-cancer metastatic cancer cohort provides a resource to study bacterial communities and their potential to modulate anti-tumor immunity, extending previous findings to the metastatic setting and across multiple cancer types.[12,17]

We determined an association between a low-oxygen TME and the abundance of anaerobic bacteria in tumors. Our data do not establish whether tumor-resident anerobic bacteria play a causal role in the development of a hypoxic TME or whether their presence simply reflects suitable conditions for anaerobic colonization. However, hypoxia has been shown to be a common feature of bacterial-infected microenvironments as an innate defense mechanism to regulate pathogen-specific responses.[33,34] Moreover, it has been suggested that metastatic organ sites may be more permissible to microbial colonization, as shown in the context of colorectal cancer liver metastases.[35] Furthermore, we found preferred community compositions between MSS and MSI patients and highlight the potential for metastatic tumor-derived community clusters, analogous to enterotypes in the gut, that may further stratify patients.

Our integration with matched transcriptomics allowed us to determine how resident bacteria are associated with tumor physiology. We observed associations between intratumoral microbes and the activation of innate immune sensing pathways, suggestive of TME changes through direct recognition of bacterial ligands. TLR-specific therapies are increasingly recognized in drug development due to their ability to act as immunomodulators,[36,37] and locally induced LPS/TLR4 activation may resemble the effects of existing TLR-agonist therapies and contribute to changes in the TME.

We also found that tumors with a high microbial diversity were characterized by an activation of TGF-β and TNF-α signaling, which are well conserved cytokines that are induced by microbial-derived products[38,39] and play central roles at the interface of host-microbe interactions in the intestine. Although we were not able to interrogate the causality of these associations, the implications of the findings are important for our understanding of host-microbe interactions in the tumor and warrant further research. Moreover, we found an enrichment of gene sets related to ECM remodeling, suggestive of a potential functional role of microbial-ECM interactions in metastases. These results are in line with previous studies showing bacterial-induced ECM modifications through changes in the expression of matrix metalloproteinases and actin cytoskeleton remodeling factors to promote survival in metastases.[40,41]

Furthermore, we found how the presence of specific bacteria was associated with the tumor-infiltrating immune compartment. Broadly, we found signatures of tumor immune evasion mediated by CAF infiltration and immune exclusion in tumors with higher microbial diversity. In line with this, we found a potent association between tumor microbial diversity and infiltrating neutrophils and macrophages, shown by two independent cell-type quantification methods.[42,43] This is in line with the primary function of neutrophils to clear pathogenic infections, in part due to the release of neutrophil extracellular traps (NETs), which in turn have recently been implicated in response to ICB via NET-based shielding of tumor cells from cytotoxic killing.[24] Therefore, bacterial colonization may induce neutrophil infiltration and thereby contribute to an immunosuppressive TME.

Our findings also show that immunotherapy augments the metastatic tumor microbiome and reduces bacterial diversity, which was even more pronounced in responsive patients. Recent studies have demonstrated that various bacteria are capable of invading tumor cells and that their peptides can be presented on the tumor cells' HLA machinery,[12,13] making it conceivable that these bacteria are a potential source of immunogenic peptides hence a substrate of cytotoxic immune responses. Indeed, we determined genera including *Actinomyces*, *Bacteroidetes*, and *Prevotella* to recurrently dropout in responsive patients, whereas these genera have been previously shown to produce HLA-presented peptides.[13] Conversely, bacteria may also stimulate a tolerogenic response[14] that can be co-opted by the tumor as an immune escape mechanism, further highlighting the importance of characterizing the local tumor microbiome. Indeed, we found that the presence of *Fusobacterium* was negatively associated with responsiveness to ICB treatment in NSCLC. This gram-negative anaerobic oral commensal has been associated with the progression and initiation of colorectal cancers[3] and can directly interact with tumor-infiltrating lymphocytes to suppress their activities.[44,45] Moreover, indirect interactions mediated by the secretion of bacterial-derived metabolites or outer membrane vesicles[46] may also contribute to an immunosuppressive phenotype. Therefore, it would be of interest to further investigate the potential targetable role of *F. nucleatum* in suppressing antitumor immunity beyond colorectal cancers.

In conclusion, we present the first large-scale, pan-cancer atlas of intratumor microbiomes in metastatic cancers with matched WGS, transcriptomics, clinical, and treatment outcome data. Our analyses show how intratumoral communities vary among anatomical sites, depend on the primary tumor type, associate with immune cell infiltration, and correlate with treatment responses, especially in the context of immunotherapy. We expect that our data will serve as an important community resource to enable future studies on the complex, potentially targetable roles of the intratumor microbiome in metastatic cancers.

Limitations of the study

Our study also has limitations: (1) where previous studies have repeatedly shown a compelling difference between matched healthy vs. primary tumor communities,[12,17] our cohort is comprised exclusively of metastatic lesions and hence cannot directly assess potential differences with healthy communities. (2) Despite our high sequencing depth (>100× coverage), the possibility exists of under sampling lowly abundant bacteria given the low biomass. (3) Although we have provided examples of the utility of our resource, our findings are at this stage strictly correlative but can help to inform hypothesis-driven experiments and subsequently dissect causality, potentially in tumor or immune cell co-cultures. (4) Our bioinformatics approach allowed us to perform high-dimensional analyses of the established bacterial genera in the context of tumor biology on a vast scale, but in-depth follow-up studies using orthogonal imaging-based approaches may shed more light on the biomass due to the limitations in computing bacterial load. Moreover, spatial localization and cellular interactions of the many microbes residing in metastatic cancers can provide key insights into these host-microbe interactions. (5) Despite our ability to reconstruct high-quality metagenomes from tumors, this approach will inevitably undersample the diversity of microbes. Nonetheless, it can provide a detailed and high-confidence taxonomic and functional characterization of resident bacteria beyond existing reference databases.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
    - ○ Lead contact
    - ○ Materials availability
    - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
    - ○ Data accession
    - ○ Library preparation and sequencing
    - ○ Bioinformatics: Host analysis
    - ○ Microbial profiling pipeline
    - ○ Contamination assessment
    - ○ Bioinformatics: Microbial analysis
    - ○ IMCISION hypoxia validation cohort
    - ○ 16S rRNA amplification and sequencing
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2024.03.021.

ACKNOWLEDGMENTS

We would like to thank Paul Wolfe, Sandra van den Broek, and Moezammin Baksi for their computational support accessing WGS data; Daniel Vis for bioinformatic and statistical support; and Jelle Vooijs for initial exploratory analyses for this project. Mrs. Anneke Hoogendijk and de Foundation Weteringschans have supported this research with their generous gifts. Furthermore, we would also like to thank Laura A. Smit for her assistance with immunohistochemistry staining of HIF-1α, Marius Messemaker for his help with 16S sequencing, and Stefania Mengoli for her help with qPCR. We would like to acknowledge the NKI-AVL Core Facility Molecular Pathology & Biobanking (CFMPB) for supplying NKI-AVL Biobank material and/or lab support as well as the NKI Genomics Core Facility (NKI-GCF). This work was supported by the NWO Gravitation program (NWO 2012–2022) (to E.V.) and by Oncode Institute (to L.W. and E.V.).

AUTHOR CONTRIBUTIONS

Conceptualization, T.W.B., J.v.d.H., L.W., and E.V.; methodology, T.W.B., J.v.d.H., A.v.H., and I.L.M; writing – original draft, T.W.B., J.v.d.H., L.W., and E.V.; funding acquisition, E.V. and L.W.; data curation, B.S.G., L.J.Z., and G.F.d.W; resources, E.C. and A.v.H.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 11, 2022
Revised: November 30, 2023
Accepted: March 18, 2024
Published: April 9, 2024

REFERENCES

1. Slowicka, K., Petta, I., Blancke, G., Hoste, E., Dumas, E., Sze, M., Vikkula, H., Radaelli, E., Haigh, J.J., Jonckheere, S., et al. (2020). Zeb2 drives invasive and microbiota-dependent colon carcinoma. Nat. Cancer *1*, 620–634. https://doi.org/10.1038/s43018-020-0070-2.
2. Kadosh, E., Snir-Alkalay, I., Venkatachalam, A., May, S., Lasry, A., Elyada, E., Zinger, A., Shaham, M., Vaalani, G., Mernberger, M., et al. (2020). The gut microbiome switches mutant p53 from tumour-suppressive to oncogenic. Nature *586*, 133–138. https://doi.org/10.1038/s41586-020-2541-0.
3. Bullman, S., Pedamallu, C.S., Sicinska, E., Clancy, T.E., Zhang, X., Cai, D., Neuberg, D., Huang, K., Guevara, F., Nelson, T., et al. (2017). Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer. Science *358*, 1443–1448. https://doi.org/10.1126/science.aal5240.
4. Derosa, L., Routy, B., Fidelle, M., Iebba, V., Alla, L., Pasolli, E., Segata, N., Desnoyer, A., Pietrantonio, F., Ferrere, G., et al. (2020). Gut Bacteria Composition Drives Primary Resistance to Cancer Immunotherapy in Renal Cell Carcinoma Patients. Eur. Urol. *78*, 195–206. https://doi.org/10.1016/j.eururo.2020.04.044.

5. Gopalakrishnan, V., Spencer, C.N., Nezi, L., Reuben, A., Andrews, M.C., Karpinets, T.V., Prieto, P.A., Vicente, D., Hoffman, K., Wei, S.C., et al. (2018). Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. Science *359*, 97–103. https://doi.org/10.1126/science.aan4236.

6. Matson, V., Fessler, J., Bao, R., Chongsuwat, T., Zha, Y., Alegre, M.L., Luke, J.J., and Gajewski, T.F. (2018). The commensal microbiome is associated with anti–PD-1 efficacy in metastatic melanoma patients. Science *359*, 104–108. https://doi.org/10.1126/science.aao3290.

7. 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. https://doi.org/10.1126/science.aah5043.

8. Routy, B., Le Chatelier, E.L., Derosa, L., Duong, C.P.M., Alou, M.T., Daillère, R., Fluckiger, A., Messaoudene, M., Rauber, C., Roberti, M.P., et al. (2018). Gut microbiome influences efficacy of PD-1–based immunotherapy against epithelial tumors. Science *359*, 91–97. https://doi.org/10.1126/science.aan3706.

9. Fluckiger, A., Daillère, R., Sassi, M., Sixt, B.S., Liu, P., Loos, F., Richard, C., Rabu, C., Alou, M.T., Goubet, A.G., et al. (2020). Cross-reactivity between tumor MHC class I–restricted antigens and an enterococcal bacteriophage. Science *369*, 936–942. https://doi.org/10.1126/science.aax0701.

10. Baruch, E.N., Youngster, I., Ben-Betzalel, G., Ortenberg, R., Lahat, A., Katz, L., Adler, K., Dick-Necula, D., Raskin, S., Bloch, N., et al. (2021). Fecal microbiota transplant promotes response in immunotherapy-refractory melanoma patients. Science *371*, 602–609. https://doi.org/10.1126/science.abb5920.

11. Davar, D., Dzutsev, A.K., McCulloch, J.A., Rodrigues, R.R., Chauvin, J.M., Morrison, R.M., Deblasio, R.N., Menna, C., Ding, Q., Pagliano, O., et al. (2021). Fecal microbiota transplant overcomes resistance to anti–PD-1 therapy in melanoma patients. Science *371*, 595–602. https://doi.org/10.1126/science.abf3363.

12. 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*, 973–980. https://doi.org/10.1126/science.aay9189.

13. Kalaora, S., Nagler, A., Nejman, D., Alon, M., Barbolin, C., Barnea, E., Ketelaars, S.L.C., Cheng, K., Vervier, K., Shental, N., et al. (2021). Identification of bacteria-derived HLA-bound peptides in melanoma. Nature *592*, 138–143. https://doi.org/10.1038/s41586-021-03368-8.

14. Pushalkar, S., Hundeyin, M., Daley, D., Zambirinis, C.P., Kurz, E., Mishra, A., Mohan, N., Aykut, B., Usyk, M., Torres, L.E., et al. (2018). The Pancreatic Cancer Microbiome Promotes Oncogenesis by Induction of Innate and Adaptive Immune Suppression. Cancer Discov. *8*, 403–416. https://doi.org/10.1158/2159-8290.CD-17-1134.

15. Dillekås, H., Rogers, M.S., and Straume, O. (2019). Are 90% of deaths from cancer caused by metastases? Cancer Med. *8*, 5574–5576. https://doi.org/10.1002/cam4.2474.

16. Priestley, P., Baber, J., Lolkema, M.P., Steeghs, N., de Bruijn, E., Shale, C., Duyvesteyn, K., Haidari, S., van Hoeck, A., Onstenk, W., et al. (2019). Pan-cancer whole-genome analyses of metastatic solid tumours. Nature *575*, 210–216. https://doi.org/10.1038/s41586-019-1689-y.

17. 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. https://doi.org/10.1038/s41586-020-2095-1.

18. Nayfach, S., Shi, Z.J., Seshadri, R., Pollard, K.S., and Kyrpides, N.C. (2019). New insights from uncultivated genomes of the global human gut microbiome. Nature *568*, 505–510. https://doi.org/10.1038/s41586-019-1058-x.

19. Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. *35*, 725–731. https://doi.org/10.1038/nbt.3893.

20. Vos, J.L., Elbers, J.B.W., Krijgsman, O., Traets, J.J.H., Qiao, X., van der Leun, A.M., Lubeck, Y., Seignette, I.M., Smit, L.A., Willems, S.M., et al. (2021). Neoadjuvant immunotherapy with nivolumab and ipilimumab induces major pathological responses in patients with head and neck squamous cell carcinoma. Nat. Commun. *12*, 7348. https://doi.org/10.1038/s41467-021-26472-9.

21. Pons, P., and Latapy, M. (2005). Computing Communities in Large Networks Using Random Walks. In Computer and Information Sciences – ISCIS 2005, 20th International Symposium. Lecture Notes in Computer Science, pp. 284–293. https://doi.org/10.1007/11569596_31.

22. Holmes, I., Harris, K., and Quince, C. (2012). Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. PLoS One *7*, e30126. https://doi.org/10.1371/journal.pone.0030126.

23. Ciesielska, A., Matyjek, M., and Kwiatkowska, K. (2021). TLR4 and CD14 trafficking and its influence on LPS-induced pro-inflammatory signaling. Cell. Mol. Life Sci. *78*, 1233–1261. https://doi.org/10.1007/s00018-020-03656-y.

24. Teijeira, Á., Garasa, S., Gato, M., Alfaro, C., Migueliz, I., Cirella, A., de Andrea, C., Ochoa, M.C., Otano, I., Etxeberria, I., et al. (2020). CXCR1 and CXCR2 Chemokine Receptor Agonists Produced by Tumors Induce Neutrophil Extracellular Traps that Interfere with Immune Cytotoxicity. Immunity *52*, 856–871.e8. https://doi.org/10.1016/j.immuni.2020.03.001.

25. Coffelt, S.B., Wellenstein, M.D., and de Visser, K.E. (2016). Neutrophils in cancer: neutral no more. Nat. Rev. Cancer *16*, 431–446. https://doi.org/10.1038/nrc.2016.52.

26. Luca, B.A., Steen, C.B., Matusiak, M., Azizi, A., Varma, S., Zhu, C., Przybyl, J., Espín-Pérez, A., Diehn, M., Alizadeh, A.A., et al. (2021). Atlas of clinically distinct cell states and ecosystems across human solid tumors. Cell *184*, 5482–5496.e28. https://doi.org/10.1016/j.cell.2021.09.014.

27. Rizvi, Z.A., Dalal, R., Sadhu, S., Kumar, Y., Kumar, S., Gupta, S.K., Tripathy, M.R., Rathore, D.K., and Awasthi, A. (2021). High-salt diet mediates interplay between NK cells and gut microbiota to induce potent tumor immunity. Sci. Adv. *7*, eabg5016. https://doi.org/10.1126/sciadv.abg5016.

28. van de Haar, J., Hoes, L.R., Roepman, P., Lolkema, M.P., Verheul, H.M.W., Gelderblom, H., de Langen, A.J., Smit, E.F., Cuppen, E., Wessels, L.F.A., et al. (2021). Limited evolution of the actionable metastatic cancer genome under therapeutic pressure. Nat. Med. *27*, 1553–1563. https://doi.org/10.1038/s41591-021-01448-w.

29. van de Haar, J., Mankor, J.M., Hummelink, K., Monkhorst, K., Smit, E.F., Wessels, L.F.A., Cuppen, E., Aerts, J.G.J.V., and Voest, E.E. (2024). Combining genomic biomarkers to guide immunotherapy in non-small cell lung cancer. Clin. Cancer Res., OF1–OF12. https://doi.org/10.1158/1078-0432.CCR-23-4027.

30. Wakiyama, H., Masuda, T., Motomura, Y., Hu, Q., Tobo, T., Eguchi, H., Sakamoto, K., Hirakawa, M., Honda, H., and Mimori, K. (2018). Cytolytic Activity (CYT) Score Is a Prognostic Biomarker Reflecting Host Immune Status in Hepatocellular Carcinoma (HCC). Anticancer Res. *38*, 6631–6638. https://doi.org/10.21873/anticanres.13030.

31. Ayers, M., Lunceford, J., Nebozhyn, M., Murphy, E., Loboda, A., Kaufman, D.R., Albright, A., Cheng, J.D., Kang, S.P., Shankaran, V., et al. (2017). IFN-γ–related mRNA profile predicts clinical response to PD-1 blockade. J. Clin. Invest. *127*, 2930–2940. https://doi.org/10.1172/JCI91190.

32. Zeng, D., Ye, Z., Shen, R., Yu, G., Wu, J., Xiong, Y., Zhou, R., Qiu, W., Huang, N., Sun, L., et al. (2021). IOBR: Multi-Omics Immuno-Oncology Biological Research to Decode Tumor Microenvironment and Signatures. Front. Immunol. *12*, 687975. https://doi.org/10.3389/fimmu.2021.687975.

33. Schaffer, K., and Taylor, C.T. (2015). The impact of hypoxia on bacterial infection. FEBS J. *282*, 2260–2266. https://doi.org/10.1111/febs.13270.

34. Worlitzsch, D., Tarran, R., Ulrich, M., Schwab, U., Cekici, A., Meyer, K.C., Birrer, P., Bellon, G., Berger, J., Weiss, T., et al. (2002). Effects of reduced mucus oxygen concentration in airway Pseudomonas infections of cystic fibrosis patients. J. Clin. Invest. *109*, 317–325. https://doi.org/10.1172/JCI13870.

35. Bertocchi, A., Carloni, S., Ravenda, P.S., Bertalot, G., Spadoni, I., Lo Cascio, A.L., Gandini, S., Lizier, M., Braga, D., Asnicar, F., et al. (2021). Gut vascular barrier impairment leads to intestinal bacteria dissemination and colorectal cancer metastasis to liver. Cancer Cell *39*, 708–724.e11. https://doi.org/10.1016/j.ccell.2021.03.004.

36. Anwar, M.A., Shah, M., Kim, J., and Choi, S. (2019). Recent clinical trends in Toll-like receptor targeting therapeutics. Med. Res. Rev. *39*, 1053–1090. https://doi.org/10.1002/med.21553.

37. Opal, S.M., Laterre, P.F., Francois, B., LaRosa, S.P., Angus, D.C., Mira, J.P., Wittebole, X., Dugernier, T., Perrotin, D., Tidswell, M., et al. (2013). Effect of Eritoran, an Antagonist of MD2-TLR4, on Mortality in Patients With Severe Sepsis: The ACCESS Randomized Trial. JAMA *309*, 1154–1162. https://doi.org/10.1001/jama.2013.2194.

38. Martin-Gallausiaux, C., Béguet-Crespel, F., Marinelli, L., Jamet, A., Ledue, F., Blottière, H.M., and Lapaque, N. (2018). Butyrate produced by gut commensal bacteria activates TGF-beta1 expression through the transcription factor SP1 in human intestinal epithelial cells. Sci. Rep. *8*, 9742. https://doi.org/10.1038/s41598-018-28048-y.

39. van der Bruggen, T., Nijenhuis, S., van Raaij, E., Verhoef, J., and van Asbeck, B.S. (1999). Lipopolysaccharide-Induced Tumor Necrosis Factor Alpha Production by Human Monocytes Involves the Raf-1/MEK1-MEK2/ERK1-ERK2 Pathway. Infect. Immun. *67*, 3824–3829. https://doi.org/10.1128/IAI.67.8.3824-3829.1999.

40. Parhi, L., Alon-Maimon, T., Sol, A., Nejman, D., Shhadeh, A., Fainsod-Levi, T., Yajuk, O., Isaacson, B., Abed, J., Maalouf, N., et al. (2020). Breast cancer colonization by Fusobacterium nucleatum accelerates tumor growth and metastatic progression. Nat. Commun. *11*, 3259. https://doi.org/10.1038/s41467-020-16967-2.

41. 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.e26. https://doi.org/10.1016/j.cell.2022.02.027.

42. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457. https://doi.org/10.1038/nmeth.3337.

43. Danaher, P., Warren, S., Dennis, L., D'Amico, L., White, A., Disis, M.L., Geller, M.A., Odunsi, K., Beechem, J., and Fling, S.P. (2017). Gene expression markers of Tumor Infiltrating Leukocytes. J. Immunother. Cancer *5*, 18. https://doi.org/10.1186/s40425-017-0215-8.

44. Gur, C., Ibrahim, Y., Isaacson, B., Yamin, R., Abed, J., Gamliel, M., Enk, J., Bar-On, Y., Stanietsky-Kaynan, N., Coppenhagen-Glazer, S., et al. (2015). Binding of the Fap2 Protein of Fusobacterium nucleatum to Human Inhibitory Receptor TIGIT Protects Tumors from Immune Cell Attack. Immunity *42*, 344–355. https://doi.org/10.1016/j.immuni.2015.01.010.

45. Abed, J., Emgård, J.E.M., Zamir, G., Faroja, M., Almogy, G., Grenov, A., Sol, A., Naor, R., Pikarsky, E., Atlan, K.A., et al. (2016). Fap2 Mediates Fusobacterium nucleatum Colorectal Adenocarcinoma Enrichment by Binding to Tumor-Expressed Gal-GalNAc. Cell Host Microbe *20*, 215–225. https://doi.org/10.1016/j.chom.2016.07.006.

46. Engevik, M.A., Danhof, H.A., Ruan, W., Engevik, A.C., Chang-Graham, A.L., Engevik, K.A., Shi, Z., Zhao, Y., Brand, C.K., Krystofiak, E.S., et al. (2021). Fusobacterium nucleatum Secretes Outer Membrane Vesicles and Promotes Intestinal Inflammation. mBio *12*, e02706-20. https://doi.org/10.1128/mBio.02706-20.

47. Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S.M., Betley, J., Fraser, L., Bauer, M., et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J. *6*, 1621–1624. https://doi.org/10.1038/ismej.2012.8.

48. Manders, F., Brandsma, A.M., Kanter, J. de, Verheul, M., Oka, R., Roosmalen, M.J. van, Roest, B. van der, Hoeck, A. van, Cuppen, E., and Boxtel, R. van (2022). MutationalPatterns: the one stop shop for the analysis of mutational processes. Bmc Genomics 23, 134. https://doi.org/10.1186/s12864-022-08357-3

49. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetJ *17*, 10–12. https://doi.org/10.14806/ej.17.1.200.

50. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

51. Liao, Y., Smyth, G.K., and Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res *41*, e108. https://doi.org/10.1093/nar/gkt214.

52. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res *43*, e47. https://doi.org/10.1093/nar/gkv007.

53. Schubert, M., Klinger, B., Klünemann, M., Sieber, A., Uhlitz, F., Sauer, S., Garnett, M.J., Blüthgen, N., and Saez-Rodriguez, J. (2018). Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat. Commun. *9*, 20. https://doi.org/10.1038/s41467-017-02391-6.

54. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innov *2*, 100141. https://doi.org/10.1016/j.xinn.2021.100141.

55. Sturm, G., Finotello, F., and List, M. (2020). Bioinformatics for Cancer Immunotherapy, Methods and Protocols. Methods Mol. Biol. *2120*, 223–232. https://doi.org/10.1007/978-1-0716-0327-7_16.

56. Jiang, P., Gu, S., Pan, D., Fu, J., Sahu, A., Hu, X., Li, Z., Traugh, N., Bu, X., Li, B., et al. (2018). Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. Nat Med *24*, 1550–1558. https://doi.org/10.1038/s41591-018-0136-1.

57. Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biol. *20*, 257. https://doi.org/10.1186/s13059-019-1891-0.

58. Walker, M.A., Pedamallu, C.S., Ojesina, A.I., Bullman, S., Sharpe, T., Whelan, C.W., and Meyerson, M. (2018). GATK PathSeq: a customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts. Bioinformatics *34*, 4287–4289. https://doi.org/10.1093/bioinformatics/bty501.

59. Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Res *27*, 824–834. https://doi.org/10.1101/gr.213959.116.

60. Nissen, J.N., Johansen, J., Allesøe, R.L., Sønderby, C.K., Armenteros, J.J.A., Grønbech, C.H., Jensen, L.J., Nielsen, H.B., Petersen, T.N., Winther, O., et al. (2021). Improved metagenome binning and assembly using deep variational autoencoders. Nat. Biotechnol. *39*, 555–560. https://doi.org/10.1038/s41587-020-00777-4.

61. Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ *7*, e7359. https://doi.org/10.7717/peerj.7359.

62. Chklovski, A., Parks, D.H., Woodcroft, B.J., and Tyson, G.W. (2023). CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. Nat. Methods *20*, 1203–1212. https://doi.org/10.1038/s41592-023-01940-w.

63. Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2019). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy

Database. Bioinformatics *36*, 1925–1927. https://doi.org/10.1093/bioinformatics/btz848.

64. Asnicar, F., Thomas, A.M., Beghini, F., Mengoni, C., Manara, S., Manghi, P., Zhu, Q., Bolzan, M., Cumbo, F., May, U., et al. (2020). Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. Nat. Commun. *11*, 2500. https://doi.org/10.1038/s41467-020-16366-7.
65. Olm, M.R., Crits-Christoph, A., Bouma-Gregson, K., Firek, B.A., Morowitz, M.J., and Banfield, J.F. (2021). inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat Biotechnol *39*, 727–736. https://doi.org/10.1038/s41587-020-00797-0.
66. Peschel, S., Müller, C.L., Mutius, E., von, Boulesteix, A.-L., and Depner, M. (2020). NetCoMi: network construction and comparison for microbiome data in R. Brief Bioinform. bbaa290-. https://doi.org/10.1093/bib/bbaa290.
67. Koh, H. (2018). An adaptive microbiome α-diversity-based association analysis method. Sci Rep-uk *8*, 18026. https://doi.org/10.1038/s41598-018-36355-7.
68. Mallick, H., Rahnavard, A., McIver, L.J., Ma, S., Zhang, Y., Nguyen, L.H., Tickle, T.L., Weingart, G., Ren, B., Schwager, E.H., et al. (2021). Multivariable association discovery in population-scale meta-omics studies. PLoS Comput. Biol. *17*, e1009442. https://doi.org/10.1371/journal.pcbi.1009442.
69. Ghazi, A.R., Sucipto, K., Rahnavard, A., Franzosa, E.A., McIver, L.J., Lloyd-Price, J., Schwager, E., Weingart, G., Moon, Y.S., Morgan, X.C., et al. (2022). High-sensitivity pattern discovery in large, paired multiomic datasets. Bioinformatics *38*, i378–i385. https://doi.org/10.1093/bioinformatics/btac232.
70. Willis, A., and Bunge, J. (2015). Estimating diversity via frequency ratios. Biometrics *71*, 1042–1049. https://doi.org/10.1111/biom.12332.
71. Lin, H., and Peddada, S.D. (2020). Analysis of compositions of microbiomes with bias correction. Nat Commun *11*, 3514. https://doi.org/10.1038/s41467-020-17041-7.
72. Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. *37*, 852–857. https://doi.org/10.1038/s41587-019-0209-9.
73. McMurdie, P.J., and Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. Plos One *8*, e61217. https://doi.org/10.1371/journal.pone.0061217.
74. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the Tidyverse. J. Open Source Softw. *4*, 1686. https://doi.org/10.21105/joss.01686.
75. Barnett, D., Arts, I., and Penders, J. (2021). microViz: an R package for microbiome data visualization and statistics. J. Open Source Softw. *6*, 3201. https://doi.org/10.21105/joss.03201.
76. Sturm, G., Finotello, F., and List, M. (2020). Immunedeconv: An R Package for Unified Access to Computational Methods for Estimating Immune Cell Fractions from Bulk RNA-Sequencing Data. In Bioinformatics for Cancer Immunotherapy. Methods in Molecular Biology, *2120* (Springer), pp. 223–232. https://doi.org/10.1007/978-1-0716-0327-7_16.
77. Jiang, P., Gu, S., Pan, D., Fu, J., Sahu, A., Hu, X., Li, Z., Traugh, N., Bu, X., Li, B., et al. (2018). Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. Nat. Med. *24*, 1550–1558. https://doi.org/10.1038/s41591-018-0136-1.
78. Gihawi, A., Ge, Y., Lu, J., Puiu, D., Xu, A., Cooper, C.S., Brewer, D.S., Pertea, M., and Salzberg, S.L. (2023). Major data analysis errors invalidate cancer microbiome findings. mBio *14*, e0160723. https://doi.org/10.1128/mbio.01607-23.
79. Gihawi, A., Cooper, C.S., and Brewer, D.S. (2023). Caution regarding the specificities of pan-cancer microbial structure. Microb. Genom. *9*, mgen001088. https://doi.org/10.1099/mgen.0.001088.
80. Sepich-Poore, G.D., Kopylova, E., Zhu, Q., Carpenter, C., Fraraccio, S., Wandro, S., Kosciolek, T., Janssen, S., Metcalf, J., Song, S.J., et al. (2023). Reply to: Caution Regarding the Specificities of Pan-Cancer Microbial Structure. Preprint at bioRxiv. https://doi.org/10.1101/2023.02.10.528049.
81. Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J., and Walker, A.W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. *12*, 87. https://doi.org/10.1186/s12915-014-0087-z.
82. McDonald, D., Jiang, Y., Balaban, M., Cantrell, K., Zhu, Q., Gonzalez, A., Morton, J.T., Nicolaou, G., Parks, D.H., Karst, S., et al. (2023). Greengenes2 enables a shared data universe for microbiome studies. Preprint at bioRxiv. https://doi.org/10.1101/2022.12.19.520774.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Mouse anti-human HIF-1α (54/ HIF-1α)	BD Transduction Laboratories	Cat#610959; RRID: AB_398272
Biological samples		
Human HNSCC tumor material	Vos et al.[20]	https://clinicaltrials.gov/study/NCT03003637
Human NSCLC tumor material	This paper	Netherlands Cancer Institute
Chemicals, peptides, and recombinant proteins		
10mM dNTPs	Roche	Cat#11969064001
Q5® Reaction Buffer Pack	New England Biolabs (NEB)	Cat#B9027S
Q5® Hot Start High-Fidelity DNA Polymerase	New England Biolabs (NEB)	Cat#M0493L
Critical commercial assays		
AllPrep DNA/RNA FFPE Kit	Qiagen	Cat#80234
AMPure XP Reagent	Beckman Coulter	Cat#A63881
Deposited data		
Mapping-based metagenomics phyloseq object	This paper	https://github.com/twbattaglia/MicrobeDS
Mapping and assembly-based metagenomics	This paper	Zenodo: https://doi.org/10.5281/zenodo.10777510
Microbe reference genomes	PathSeq customized bundle	Zenodo: https://doi.org/10.5281/zenodo.10777510
Human and microbe reference genomes	Kraken2 bundle (RefSeq)	https://www.ncbi.nlm.nih.gov/refseq/
T2T-CHM13v2.0	NCBI	GenBank: GCF_009914755.1
Hartwig WGS data	Priestley et al.[16]	https://www.hartwigmedicalfoundation.nl/en/data/data-access-request/
Hartwig RNA-seq data	Priestley et al.[16]	https://www.hartwigmedicalfoundation.nl/en/data/data-access-request/
Hartwig clinical metadata	Priestley et al.[16]	https://www.hartwigmedicalfoundation.nl/en/data/data-access-request/
Human Microbiome Project WGS raw data	NCBI	dbGaP: phs000228
IMCISION	Vos et al.[20]	https://clinicaltrials.gov/study/NCT03003637
Oligonucleotides		
16S 515F: GTGYCAGCMGCCGCGGTAA	Caporaso et al.[47]	Integrated DNA Technologies; IDT
16S 806R: GTGYCAGCMGCCGCGGTAA	Caporaso et al.[47]	Integrated DNA Technologies; IDT
Software and algorithms		
Python version 3.8	Python Software Foundation	https://www.python.org/
R version 4.1.3	R Core Team	https://www.R-project.org/
Hartwig - Pipeline5	Priestley et al.[16]	https://github.com/hartwigmedical/pipeline5
Nextflow (v22.04.4)		https://www.nextflow.io/
MutationalPatterns (v3.4.1)	Manders et al.[48]	https://bioconductor.org/packages/release/bioc/html/MutationalPatterns.html
Cutadapt (v2.10)	Martin et al.[49]	https://pypi.org/project/cutadapt/
STAR (v 2.7.10a)	Dobin et al.[50]	https://github.com/alexdobin/STAR
Subread (v 2.0.1)	Liao et al.[51]	https://subread.sourceforge.net/
Limma (v3.50.3)	Ritchie et al.[52]	https://bioconductor.org/packages/release/bioc/html/limma.html
PROGENy (v1.16)	Schubert et al.[53]	https://bioconductor.org/packages/release/bioc/html/progeny.html
clusterProfiler (v4.2.2)	Wu et al.[54]	https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Immunedeconv (v2.10)	Sturm et al.[55]	https://github.com/omnideconv/immunedeconv
TIDEpy (v1.3.8)	Jiang et al.[56]	https://pypi.org/project/tidepy/
EcoTyper	Luca et al.[26]	https://github.com/digitalcytometry/ecotyper
Samtools (v1.10)		https://github.com/samtools/samtools
Kraken2 (v2.0.8)	Wood et al.[57]	https://github.com/DerrickWood/kraken2
Bracken (v2.5.3)	Wood et al.[57]	https://github.com/jenniferlu717/Bracken
PathSeq (v4.1.4.1)	Walker et al.[58]	https://software.broadinstitute.org/pathseq/
metaSPAdes (v3.15.5)	Nurk et al.[59]	https://github.com/ablab/spades
VAMB (v3.0.2)	Nissen et al.[60]	https://pypi.org/project/vamb/
MetaBAT2 (v2.15)	Kang et al.[61]	https://bitbucket.org/berkeleylab/metabat/src/master/
MAGpurify (v2.1.2)	Nayfach et al.[18]	https://pypi.org/project/magpurify/
CheckM2 (v1.0.1)	Chklovski et al.[62]	https://pypi.org/project/CheckM2/
GTDB-TK (v2.3.0)	Chaumeil et al.[63]	https://pypi.org/project/gtdbtk/
PhyloPhlAn (v3.0.3)	Asnicar et al.[64]	https://pypi.org/project/PhyloPhlAn/
CoverM (v0.6.1)		https://github.com/wwood/CoverM
inStrain (v1.8.0)	Olm et al.[65]	https://pypi.org/project/inStrain/
NetCoMi (v.1.1.0)	Peschel et al.[66]	https://github.com/stefpeschel/NetCoMi
aMiAD (v2.0)	Koh et al.[67]	https://github.com/hk1785/aMiAD
MaAsLin2 (v1.8.0)	Mallick et al.[68]	https://www.bioconductor.org/packages/release/bioc/html/Maaslin2.html
HAllA (v0.8.2.0)	Ghazi et al.[69]	https://pypi.org/project/HAllA/
breakaway (v4.8.4)	Willis et al.[70]	https://cran.r-project.org/web/packages/breakaway/index.html
ANCOMBC (v2.1.1)	Lin et al.[71]	https://www.bioconductor.org/packages/release/bioc/html/ANCOMBC.html
QIIME2 (v2023.2)	Bolyen et al.[72]	https://qiime2.org/
phyloseq (v1.38.0)	McMurdie et al.[73]	https://bioconductor.org/packages/release/bioc/html/phyloseq.html
tidyverse (v1.3.1)	Wickham et al.[74]	https://cran.r-project.org/web/packages/tidyverse/index.html
microViz (0.10.7)	Barnett et al.[75]	https://david-barnett.github.io/microViz/
Biorender	Biorender	https://www.biorender.com
Other		
Analysis code	This paper	https://github.com/twbattaglia/tumor-microbiome
Somatic variant calling	Priestley et al.[16]	https://github.com/hartwigmedical/hmftools
Curated bacterial genomes	This paper	Table S1
Bacteria composition of Hartwig samples	This paper	Table S2

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Emile Voest (e.voest@nki.nl).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- De-identified patient whole genome sequencing (WGS), clinical metadata, and transcriptomics data from the Hartwig Medical Foundation are freely available for academic use through standardized procedures. Request forms can be found at https://www.hartwigmedicalfoundation.nl/en/data/data-acces-request/. Data are publicly available as of the date of publication. DOIs are listed in the key resources table.

- All original code has been deposited at: https://github.com/twbattaglia/tumor-microbiome and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Within the Hartwig Medical Foundation cohort, patients with advanced cancer without any further treatment options were included as part of the CPCT-02 (NCT01855477) and DRUP (NCT02925234) clinical studies, which were approved by the medical ethical committees (METC) of the University Medical Center Utrecht and the Netherlands Cancer Institute, respectively. Informed consent was given for WGS as well as data sharing for research purposes. Detailed patient information was collected by the Hartwig Medical Foundation. Patient characteristics, including primary tumor location, biopsy site, age group and sex are detailed in Table S4.

METHOD DETAILS

Data accession

The Hartwig Medical Foundation (Hartwig) is a spin out of a study by the Center for Personalized Cancer Treatment (CPCT). The goal of this study was to systemically collect clinical data on the treatment of patients with metastatic cancer and simultaneously biopsy a metastatic lesion for whole genome sequencing. Together this has created the current valuable cohort of the Hartwig Medical Foundation. Both initiatives (Hartwig Medical Foundation and CPCT) were founded by the senior author of this manuscript. All Hartwig Medical Foundation data were accessed and processed using the Google Cloud Platform (GCP). Collectively, we gathered 4,160 tumor-derived whole genome sequencing CRAM files, sequenced at a median sequencing depth of 106x, from the Hartwig's Google Cloud bucket under the data request DR-043. Clinical data was annotated to ensure metastatic samples fell into the appropriate biopsy site categories outlined in previous research.[16]

To maintain reproducibility and more rapidly gather results, files were batched processed using a Nextflow pipeline connected to the Google Cloud Platform using the Cloud Life Sciences API. Each CRAM file was processed on an independent Docker containerized Compute Instance with 8-16 vCPU's and 30-100 GB of memory for pre-processing and profiling, respectively, using pre-emptible instances and local solid-state drives. The computational wall time for processing all samples was approximately 4 months.

Library preparation and sequencing

Samples were prepared as previously described in a uniform and centralized location.[16] Briefly, DNA was isolated from biopsy and blood on an automated setup (QiaSymphony) according to supplier's protocols (Qiagen). Blood was extracted using the DSP DNA Midi kit and QIAsymphony DSP DNA Mini kit for tissue. Before starting DNA isolation from tissue, the biopsy was dissolved in 100 microliter Nuclease-free water by using the Qiagen TissueLyzer and split in two equal fractions for parallel automated DNA and RNA isolation (QiaSymphony). RNA was extracted using the QIAGEN QIAsymphony RNA kit, prepared with the KAPA RNA Hyper + RiboErase HMR and underwent paired end sequenced on the Illumina NextSeq550 platform (2x75bp) or Illumina NovaSeq6000 platform (2x150bp).

Bioinformatics: Host analysis

Whole genome profiling

Whole genome sequencing and analysis was performed by the Hartwig Medical Foundation as previously described,[16] whereby reads were mapped to the reference genome GRCh37, somatic single nucleotide variants and insertions and deletions were called using Strelka, and purity, ploidy, structural variant (SV) and copy number somatic analysis was performed using the Hartwig's in-house tools GRIDSS, PURPLE and LINX, respectively. Samples with <20% tumor purity were removed before analysis. Mutational signatures were determined using MutationalPatterns with strict refitting and a max delta of 0.004.

Transcriptome profiling

RNA-seq data was filtered using Cutadapt for reads > 35nt and depletion of TruSeq adapters. Subsequent reads were aligned to GRCh38 (Gencode v35) using STAR (2-pass mode) and quantified using featureCounts within the Subread package. Reads counts were then normalized to transcripts-per-million (TPM) before use with downstream pathway activity tools, or were normalized using Limma-Voom for before differential abundance testing.

Tumor-derived gene set variation analysis (GSVA) pathway activities were computed using edgeR normalized gene expression profiles. The pathways were derived from MSigDB as follows: Hallmark Hypoxia (M5891), Cellular response to increased oxygen levels (M23439), TLR4 signaling (M13874) and TLR2 signaling (M23239). The activity of 14 distinct signaling pathways were computed using Pathway RespOnsive GENe activity inference (PROGENy) using the tool, EaSIeR, on raw gene expression counts. And gene set enrichment analysis (GSEA) was performed on $\log_2$ fold-changes derived from Limma Voom differential abundance testing using, clusterProfiler.

Tumor immune infiltration

Immune cell deconvolution was performed on TPM-normalized counts by CIBERSORT (absolute) using the immunedeconv framework.[76] Tumor Immune Dysfunction and Exclusion (TIDE) signatures[77] were computed using the tool, tidepy on TPM normalized gene counts. Immune cell states were deconvolved using EcoTyper[26] using default settings on TPM normalized counts. Significant correlations were found between expression-based PD-L1 TIDE scores and IHC staining on matched tumor tissue (Figure S5G).

Microbial profiling pipeline

Aggregating non-microbial sequences for analysis

We gathered all sequencing pairs that did not successfully align to the human genome (build hg19) from the Hartwig's bioinformatic pipeline. To do this we used samtools to filter CRAM files for reads in which both read pairs did not map sufficiently (SAM flag 12) and produced a BAM file for downstream analyses. Unmapped read quality profiles were analyzed using FastQC.

Kraken2 microbial profiling pipeline

Unmapped BAM files were converted into FASTQ format using Picard and quality filtering using Cutadapt with a Phred quality score > 15, minimum length > 50nt and a depletion of Illumina TruSeq Index adapters. Unmapped, quality filtered sequences were then profiled with the fast k-mer based metagenomic taxonomic classification tool, Kraken2 (v2.0.8-beta). Briefly, Kraken2[57] uses a database of short genomic substrings (k-mers) labelled with their lowest common ancestor (LCA). Each read's set of 31nt k-mers are rapidly mapped to this database to determine the appropriate taxonomic lineage, most reliably at the Genus level. A Kraken2 reference database was created from >23,000 human, bacterial, archaeal and viral NCBI RefSeq genomes and was access on 03 May 2020 (RefSeq release 99). Previous analyses found this method performed as well as alignment-based approaches against endogenous tumor microbiome data.[17] Following Kraken2 profiling, genus-level abundances were re-estimated with Bracken2 (v2.5.3) which recomputes read assignments using a Bayesian model.

PathSeq microbial profiling pipeline

To increase the sensitivity of our approach and systemically remove potential false positives generated by Kraken2, we also processed the unmapped reads with the pathogen discovery tool, PathSeq.[58] Unlike Kraken2, PathSeq relies on a computational subtraction approach in which unmapped reads undergo successive steps of host removal, quality filtering and sequence complexity filtering before aligning reads against a large reference database of viral, bacterial and archaeal RefSeq genomes using BWA. We chose to create our own custom database of RefSeq genomes derived from human-associated microbial catalogues to (1) decrease overall computational time and remove database-associated false positives. The human-associated database contained 6,328 genomes and was composed of bacterial and viral genomes that were outlined in 5 human-associated microbial surveys (Table S1).

Metagenomic assembly of unmapped reads

We employed a multi-sample, bin-splitting approach for metagenomic assembly of bacterial genomes.[60] Unmapped, quality filtered reads were first individually assembled with metaSPAdes (v3.15.5) and contigs were concatenated using VAMB (v3.0.2) with a minimum length of 2,000 bp resulting in a catalog of 2.27 million contigs. Depth of coverage was calculated by mapping reads back to contig catalog using minimap2 (v2.24) and applying the function *jgi_summarize_bam_contig_depths* from MetaBAT2 (v2.15). Thereafter, assemblies were binned using VAMB (v3.0.2) using recommended settings resulting in 1205 metagenomic assembled genomes (MAG) > 200Kbp using an NVIDIA T4 GPU. Bins were further refined using MAGpurify[18] to remove contigs with outlier GC, tetra-nucleotide content and taxonomic discordant contigs, as well as human contamination. Completeness and contamination were computed using CheckM2 (v1.0.1). Quality score (QS) of each metagenome-assembled genome (MAG was estimates as follows: completeness – (5 × contamination). Bins were classified into high and medium quality MAG based on standards set by minimum information about a metagenome-assembled genome (MIMAG)[19]: near-high quality high: >90% completeness and <5% contamination; medium (QS>50): $\geq$ 50% completeness and <10% contamination and QS > 50, medium (QS<50): $\geq$ 50% completeness and <10% contamination and QS<50. MAGs were assigned taxonomy using the *classifywf* module of GTDB-TK (v2.3.0, R214) and PhyloPhlAn (v3.0.3) using the *phylophlan_metagenomic* module against the SGB.Jul20 database. Reads were remapped against the near-high quality and medium quality assemblies using the *genome* module of CoverM (v0.6.1). Coverage statistics were then computed using inStrain (v1.8.0). For the set of paired biopsies, the *compare* module was used and genome wide consensus-based ANI (conANI) scores were computed for each sample pair.

Contamination assessment

Discussions surrounding the computational methodology of preceding cancer microbiome studies,[17] centered around a substantial degree of human-derived false positives and insufficient normalization techniques,[78–80] have been raised. Therefore, we will specifically highlight how our computational work distinguishes from those previously described and how it addresses these concerns.

Integration of two microbial profiling bioinformatics methods

We first evaluated the possibility of false positives derived from computational methodologies. In our testing, we found that both Kraken2 and PathSeq each had their respective disadvantages when it came to accurately identifying simulated microbial compositions. More specifically, we found that while Kraken2 would overestimate the diversity of a small simulated community (Figure S1A), PathSeq would underestimate (Figure S1B) given its strict quality and mapping cutoff's. This was evident in simulated analysis by which the strain-level genomes of the simulated bacteria were removed from both reference databases as an attempt to mimic the potential community in vivo. We found that intersection of the two methods, by which Kraken2 profiles were filtered given

PathSeq's detected bacteria, had the best accuracy of recapitulating the simulated community (Figure S1C). Therefore, we found that the combination of the two methods reduced the possibility of false positive bacterial presence.

Moreover, we wanted to confirm that our methodology was not specifically optimized for tumor-derived low biomass samples, but can also be generalized to existing fecal gut microbiome datasets. Therefore, we applied our pipeline to a subset of 200 samples derived from the skin, urogenital or gut within the Human Microbiome Project cohort (Figure S1D). A compositional overview of the microbial communities showed a similar structural difference as compared to existing metagenomic methodologies, by which the gut-derived microbial communities clusters apart from both skin and urogenital communities, thus showing the robustness of our approach.

Curated human-associated bacterial reference database

Next, we sought to remove false positives derived from the reference databases. To do this, we made use of two methodologies. The first includes the strict inclusion of human-associated bacteria reference genomes in the PathSeq reference database. To this end we made use of 6,328 genomes from 5 studies of who bacterial species had been deposited in the RefSeq database. Moreover, these studies include metagenomic-assembled genomes (MAG), which have been shown to increase the characterization of the human microbiome. Secondly, we included the human-reference genome within our Kraken2 reference database given the unmapped reads of the tumor microbiome are potentially human-derived but with insufficient mapping quality. Experiments with simulated reads derived from the human genome against the Kraken2 standard database without the human reference genome included had a high degree of false mappings to multiple bacteria, while this effect was completely removed with the Kraken2 standard dataset that included the human genome (data not shown). We found a substantial percent of human-associated reads after Kraken2 mapping, suggesting that the lack of correction of this artefact may potentially bias the microbial composition results. Additionally, we found Kraken2 to detected a higher degree of community richness compared to PathSeq (Figure S1E), but we found a majority of these bacteria to be sourced from non-human habitats as annotated by Genomes Online Database (GOLD) (Figure S1F).

Furthermore, we re-ran our pipeline on a subset of samples (n=20), with an additional T2T-CHM13 decontamination step. As expected, we find fewer microbial reads in the samples decontaminated using the T2T-CHM13 genome (Figure S1L), however, we still find a strong correlation of the bacterial abundances between the two methods (Figure S1M). We hypothesize that this can be attributed to (1) the additional depletion step of low-quality reads we perform which removes many of the poorly mapped human reads and (2) that our dual-approach computational profiling method removes Kraken2 mappings that may otherwise inflate microbial read count. Additionally, we simulated reads (n=1,000,000) from the T2T-CHM13 (v2.0) genome and determined read count and coverage against our MAG catalog. We found a very low mapping rate of 0.24% (2433/1000000) highlighting a lack of human-derived overlap with our MAG catalog. Moreover, we found that less than 0.20% of reads mapped to any MAG species (Figure S1N).

Depletion of known contaminants

The third step was to correct for false detection of bacteria by removing genera that were detected but have been determined to be potential laboratory-derived contaminants derived from DNA extraction kits and laboratory reagents. To do this, we made use of the published list of genera[81] and removed both Likely contaminants and Mixed Evidence contaminants. Furthermore, we found that these contaminants made up a large portion of the abundance and prevalence across our cohort (Figure S1G), suggesting their inclusion would bias total community levels.

Modeling batch effects

Next, we sought to account for false associations derived from technical confounders. To do this we made use of linear mixed effects models (LME) by which we account for sequencer type and hospital location as random effects. The assessment of variance explained by hospital location and sequencer showed a high degree of proportional variance (Figure S1H). Specifically for the hospital location, this can be in part related to the fact that some subsets of cancer were received from a few hospital locations due to their specialization. These effects could be reduced after applying batch correction methods from Limma (removeBatchEffect). Moreover, we found that using an LME with the confounder accounted for as random effects on uncorrected counts achieved similar model performance as a linear model on batch-corrected counts (Figure S1I). Therefore, by making use of linear mixed effects models, we were able to reduce the number of false associations attributed to technical variance.

Assessment of matched blood-derived microbial communities

Finally, we investigated the potential for our pipeline and methodologies to determine microbial compositions on matched blood-based sequencing data, which serve as a negative control. To do this, we applied our pipeline to a subset of 255 blood-derived WGS data with matched tumor microbial profiles. We found that the richness of the blood-derived microbial community was significantly lower compared to that of the tissue-based community (Figure S1J). We found that the bacteria that had a high prevalence in the blood-based communities were generally found to be laboratory-derived contaminants that would be depleted during our pre-processing steps (Figure S1K). Moreover, we found that bacteria that had high prevalence generally were classified as pathogenic and who contain species that are involved in blood-born infections.

Bioinformatics: Microbial analysis

Sample and bacterial filtering

Unmapped whole genome sequencing (WGS) samples from 4,164 underwent microbial profiling using the methods described in detail above, and of these, 4,115 (98%) had a successful microbial community profile generated. Next, samples without sufficient clinical information, such as a lack of metastatic localization site or primary tumor site of origin were removed. Following this, we preserved only metastatic localization sites or primary tumor site of origin in which at least 25 samples were available to retain sufficient power for linear modeling. Next, we retained samples in which a library of >500 reads were achieved (n=3576).

Bacterial physiologies

Bacterial physiologies, including Gram staining status and aerophilicity taxonomic annotations were generated using bugphyzz (v0.0.1.3) (Table S1). Briefly the Gram staining and aerophilicity annotations were extracted propagated at the genus-level using taxPPro (v0.1.0). Only annotations with a score > 0.5 were retained. Those with a score <0.5 were classified as 'Not available'.

Alpha and beta diversity estimations

The Shannon diversity index (alpha diversity) estimates, were computed using mia (v1.2.7) on a rarefied table (n=1500, seed=918). Beta-diversity estimates was computed using NetCoMi (v1.1.0) using the modified central log ratio transformation (mclr) in Euclidean space to generate Aitchison distances. To compute pairwise cancer-type dissimilarities, a distance matrix was used as an input for *betadisper* to compute difference in medians. Marginal pairwise P-values were generated using a PERMANOVA adonis test (*adonis2*) from vegan (v2.6-2) to account for cancer type, biopsy site, hospital location and sequencer type.

Integration between bacterial profiles and Hallmarks or Ecotype expression was performed using Procrustes analysis using the function *protest* within the vegan package (2.6-2) on a random selection of 1000 samples. Bacterial profiles were first filtered for 1% prevalence and subsequently clr-transformed. Next Bray-Curtis and Jaccard dissimilarities were computed from Hallmark zscore and Ecotype state abundance quantifications, respectively. A Mantel test with 9999 permutations was used to estimate significance between distance matrices.

Alpha diversity associations with tumor immune and pathway features were performed with Adaptive Microbiome α-diversity-based Association Analysis (aMiAD) (v2.0). Briefly, rarefied diversity estimates of Shannon, Richness and Gini Simpson were used as an input. TIDE or PROGENy residuals after correcting for cancer type, biopsy location or CD45 expression (as in the case with TIDE to account for total immune infiltration), was set as the variable of interest with 5000 permutations.

MSI vs. MSS community cluster detection

Microbial samples derived from the DRUP study (n=191) were used as an input. DRUP-specific samples were included as there is an enrichment of dMMR/MSI-H tumors derived from cohorts of patients treated with ICB. Cancer types were only included if there existed a corresponding set of MSI and MSS samples. Bacterial profiles were then centered-log ratio transformed with a pseudocount and multiple graphs were create running the short random walks algorithm from bluster (v1.4.0) with varying NNGraphParam (k=2, 10, 25, 50). After running purity estimations, a NNGraphParam of k=20 was selected and samples were classified as Cluster A (n=109) or Cluster B (n=82). A logit model predicting MSI status from cluster group after correcting for cancer type, biopsy site, and hospital location was used.

Hypoxia enrichment analysis

To determine an enrichment of anaerobic bacteria in hypoxia tumors, we first generated pvalues and ranking metrics for each individual bacteria and tumor hypoxia quantification using GSVA quantifications. To do this, we first performed differential expression analysis using Microbiome Multivariate Association with Linear Models (MaAsLin2) (v1.8.0) on GSVA pathway quantifications. Our model was constructed with hypoxia, cancer type and biopsy location as fixed effects and hospital location and sequencer type as random effects. MaAsLin2 parameters were set to a linear modeling with a clr-transformation. Then, we extracted the t-statistic derived from each bacterial-specific linear model and applied GSEA with 9999 permutations from clusterProfiler (v4.2.2) using annotations as described above (bacterial physiologies). For the IMCISION validation cohort, HIF-1α was modelled for each bacteria with fixed effects set as timepoint (Week 0 or Week 4), tumor localization, and major pathological response, while random effects were set to patient ID.

Host-microbe data integration

To determine associations between bacterial populations and high dimensional datasets, we made use of the Hierarchical All-against-All Association Testing (HAllA) (v0.8.2.0). Briefly, bacterial profiles were filtered for 5% prevalence and subsequently transformed to relative compositions. Next a general linearize model with Gaussian was fit with cancer type, biopsy site hospital location and sequencer type set as covariates to generate covariate-independent residuals. The same procedure was performed for EcoTyper state abundances, but included cancer type, biopsy site, and CD45 infiltration. Next HAllA was performed using all-against-all parameters, with default parameters.

Longitudinal samplings analysis

To assess the longitudinal presence of microbial populations within a patient', between metastases, we made use of a cohort of paired biopsies included within our resource (n=250). Of these pairs, 199 had sufficient microbial profiling information for downstream analysis. To ensure differences we not driven by technical artefacts, we removed pairs in which biopsies were sequenced on differing sequencing technologies.

Beta diversity estimates were generated by NetCoMi using the mclr transformation in Euclidean space to generate Aitchison distances and binary Jaccard dissimilarities were computed using *vegdist* from the vegan package. Genus-level richness estimates were computed using breakaway (v4.8.4) using a cutoff of 50. Statistical modelling of the observed and unobserved total diversity was performed using the *betta* model with random effects. Fixed effects include lesion status, weeks between biopsies and matched biopsy site location and random effects was patient pairing ID.

Differential abundance testing

ANCOMBC (v2.1.1) was used determine differentially abundant genera between responsive and non-responsive NSCLC patients treated with ICB, using the *ANCOMBC2* function. Fixed parameters included durable clinical benefit (PFS>6 months) and lymph node biopsy localization status, and random effects were set to sequencing platform and hospital origin. Structural zero detection and pseudo-counts sensitivity analysis was included and a sample prevalence cutoff of 20% was used.

IMCISION hypoxia validation cohort

An external cohort of 31 HPV-negative stage II-IVB HNSCC patients treated with ICB (NCT03003637)[20] were used to validate the presence of hypoxia and anaerobic bacteria infiltration. Briefly, transcriptomics data from 31 patients across 2 timepoints (Baseline and Week 4, n=62) underwent microbial profiling as described above. Hypoxia quantification was computed by immunohistochemistry staining of HIF-1α on FFPE primary tumor samples on a BenchMark Ultra autostainer (Ventana Medical Systems). Briefly, the HIF-1α was scored by determining the percentage of tumor cells that express HIF-1α. Transcriptomic-derived hypoxia signatures correlated well with HIF-1α staining's (Figure S4D).

16S rRNA amplification and sequencing

DNA was isolated from FFPE tumor slides (10 μm) with the AllPrep DNA/RNA FFPE Kit (Qiagen, 80234) using the QIAcube, according manufacturer's instructions. Prior to isolation, tumor percentages were scored by a pathologist to indicate the most tumor-dense regions for isolation on H&E slides using Slidescore (www.slidescore.com).

A detailed protocol for performing 16S rRNA amplification and sequencing has previously been described.[47] Briefly, extracted DNA was used for 16S amplification of the V3-V4 region using the primers 515F (5'-GTGYCAGCMGCCGCGGTAA-3') and 806R (5'-GGACTACNVGGGTWTCTAAT-3') adapted to incorporate the Illumina adapters and a sample barcode sequence. The amplification mix contained 1X Q5 Buffer (NEB, B9027S), 200μM dNTP (Roche, 11969064001), 0.5 μM forward and reverse primers, 0.5 units Q5 (NEB, M0493L) and 3μl of template DNA (30ng) in a 25μl reaction using the following cycling conditions: denaturation 3 min at 98°C, followed by a set of 30 cycles (98°C for 50 s, 55°C for 30 s and 72°C for 90 s) and final elongation (5 min 72°C). Constructed libraries were purified using 60μl of magnetic beads (AMPure XP Reagent; Beckman Coulter, A63881) per 75μl PCR reaction. Amplicons were pooled in equimolar concentrations and checked for quality prior to sequencing and sequenced using the MiSeq Nano (2x150 bp) platform. Amplicon sequencing data was processed and analyzed using QIIME2[72] (v2023.2). Briefly, forward reads were denoised with DADA2 (q2-dada2) for quality filtering and the depletion of chimeric, and erroneous reads using default parameters. Reads were subsequently assigned taxonomic annotations (q2-feature-classifier) using a naive Bayes classification method against the Greengenes2 database.[82]

QUANTIFICATION AND STATISTICAL ANALYSIS

Downstream analysis and plots were performed with R version v4.1.3. Packages used in analysis include phyloseq 1.38.0, ggplot2 3.4.2, patchwork 1.1.1, taxPPro 0.1.0, ggthemes 4.2.4, NetCoMi 1.1.0, breakaway 4.8.4, bluster 1.4.0, microViz 0.10.7, mia 1.2.7, clusterProfiler 4.2.2, miaViz 1.2.1, survminer 0.4.9, microbiome 1.19.1, survival 3.2-13, tidyverse 1.3.1, ReactomePA 1.38.0, scater 1.22, bugphyzz 0.0.1.3, ggpubr 0.4.0, RColorBrewer 1.1-3, ggsci 3.0.0.

Supplemental figures

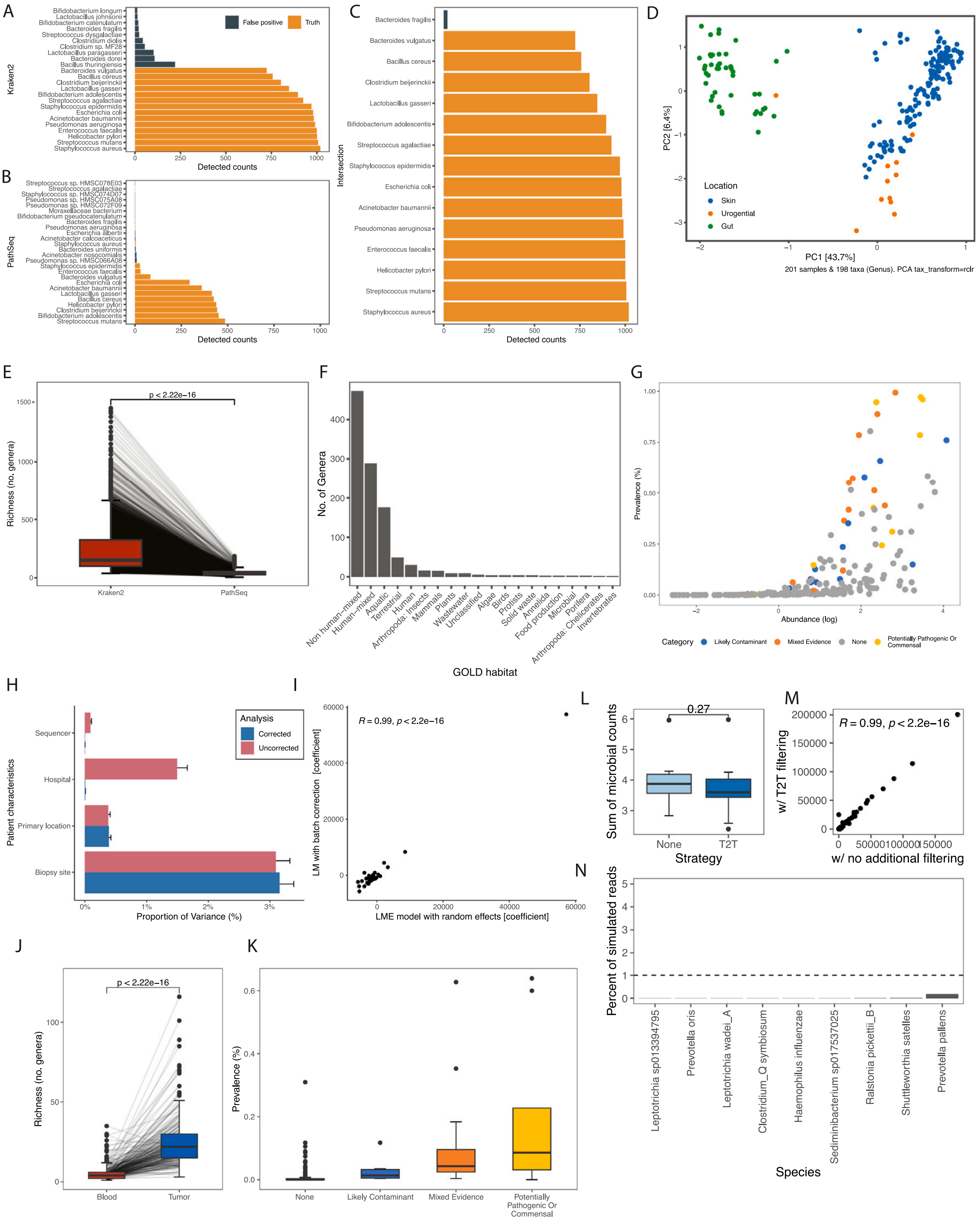


Figure S1. Mapping-based computational methods for bacterial detection, related to Figure 1

(A–C) Intersection between Kraken2 and PathSeq profiles on a simulated bacterial community ($n = 14$). Microbial profiles derived from a simulated set of 14 strains profiled against Kraken2 (A) or PathSeq (B) and whose strains were depleted from the reference database and the intersection between the two methods (C). Colors represent bacteria within (orange) or outside (black) the simulated community.

(legend continued on next page)

(D) Principal-component analysis (PCA) of the healthy microbiome communities derived from the Human Microbiome Project (HMP) (*N* = 200) after being processed with our standardized microbial profiling tool and filtering methodology. Colors represent samples derived from different source locations (skin [blue], urogenital [orange], and gut [green]).
(E) Richness estimates as determined by Kraken2's standard databases (red) or curated PathSeq database (blue). *p* values shown were determined using a Wilcoxon rank-sum test.
(F) Genomes Online Database (GOLD) annotation of Kraken2-specific genera detected.
(G) Abundance (x axis) and prevalence (y axis) of contaminant bacteria within the metastatic tumor microbiome cohort (*n* = 4,115) across detected genera (*n* = 220). Colors indicate the annotation of the genera as a likely contaminant (blue), potential commensal (yellow), or mixed evidence (orange).
(H) Variance explained of bacterial communities by patient and sample covariates on uncorrected (red) or batch-corrected (blue) microbial counts. Error bars represent std.
(I) Correlation of model coefficients between a mixed-effects model account for technical confounders (sequencer type and hospital location) and a fixed linear model with batch-corrected counts. *p* value was calculated using a Spearman rank correlation.
(J) Number of genera detected (richness) in blood-based (red) or tumor (blue) whole-genome sequencing (WGS) data after applying our standardized filtering methodology. *p* value was computed using a paired Wilcoxon rank-sum test.
(K) Prevalence of the genera in blood-derived communities according to contamination annotations.
(L) Per-sample library sizes of microbial reads after applying the Kraken2 + PathSeq intersection profiling methodology. Boxplot of samples decontaminated with either hg19 alignment (light blue) or with hg19 + T2T-CHM13 (dark blue).
(M) Per-genus abundance correlations between the two strategies. *p* value shown was calculated using a Spearman rank correlation.
(N) Percentage of reads mapped to our MAG catalog from reads simulated from the T2T-CHM13 genome (*n* = 1,000,000).

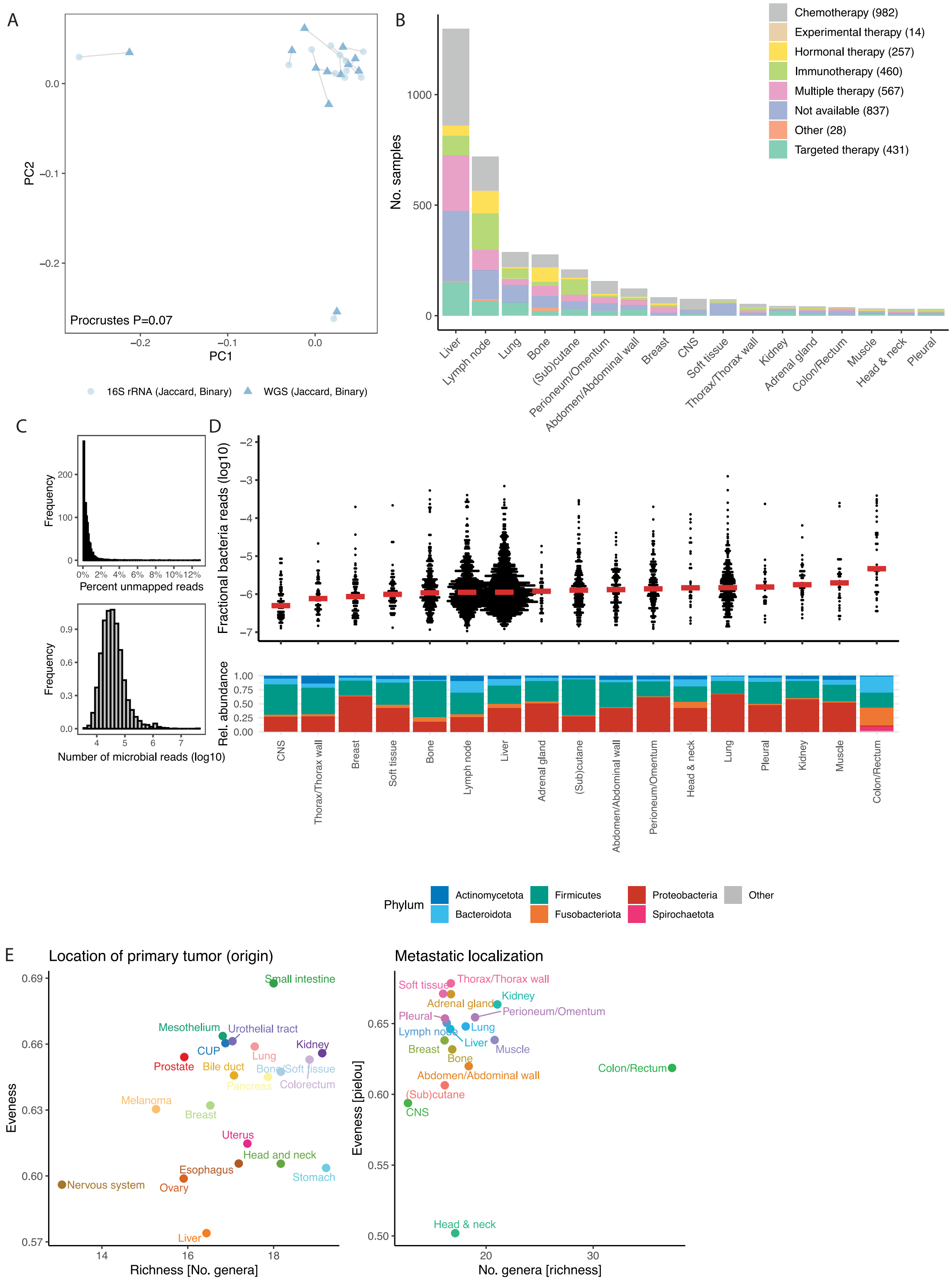


Figure S2. Overview of bacterial-derived reads in metastatic tumor microbiomes, related to Figure 1

(A) Procrustes analysis of WGS-based and established 16S rRNA microbial profiles in matched tumors. Each point represents a single tumor's microbial composition derived from either 16S rRNA profiling (circle) or WGS-based profiling (triangle). Procrustes rotations were performed on two distance matrices of Jaccard dissimilarities between patient's tumor microbial profiles. The *p* value was generated using the protest function with 9,999 permutations.

(legend continued on next page)

(B) Overview of metastatic localization sites and treatments within the cohort. The bar plot indicates the number of samples per localization site, where colors denote the class of treatment given right after the biopsy.

(C) Histogram of the percent of unmapped WGS reads (top) or total number of microbial-mapped reads (bottom).

(D) Boxplots of the fraction of bacterial-derived reads from the total number of human-mapped reads (fractional bacterial reads), stratified by metastatic localization site (top) and the phylum-level average relative composition of the bacterial community within a site (bottom). Red line indicates group median.

(E) Average rarefied estimates of community richness (x axis) and evenness (y axis) by primary tumor location of origin (left) and metastatic localization site (right).

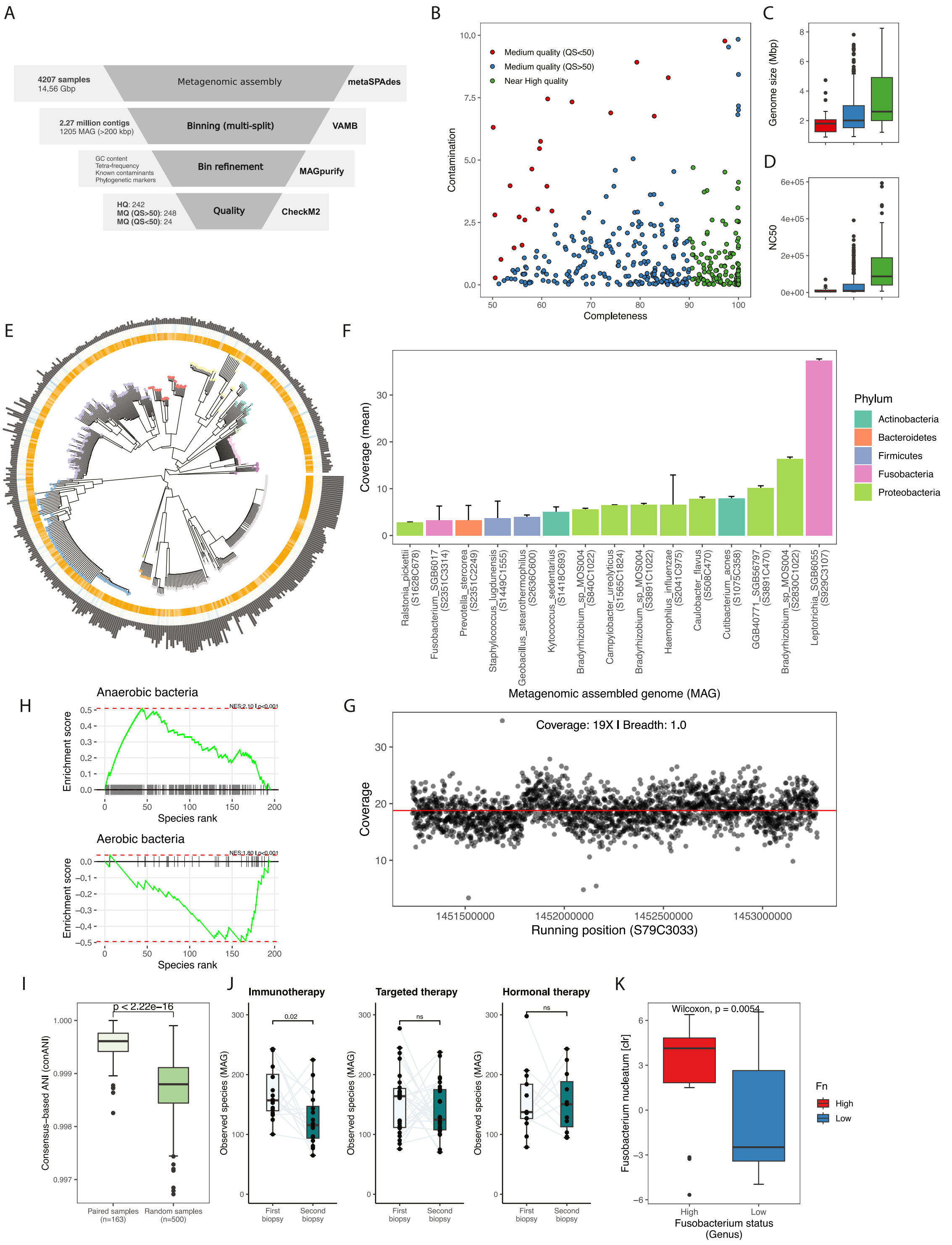


Figure S3. Metagenomic-assembly of metastatic tumor microbiomes, related to Figure 1
(A) Overview of computational workflow to generate metagenomic-assembled genomes (MAGs) using a multi-sample, co-binning approach.
(B) Overview of the MAG completeness and contamination after co-assembly and binning. Colors correspond to MAGs annotated as near-high quality (green), quality score > 50 (blue), and quality score < 50 (red).
(C and D) Distribution of genome size (C) and contig NC50 (D) determined from MAG with medium or near high-quality MAGs.

(legend continued on next page)

(E) Phylogenetic tree of MAGs using the Genome Taxonomy Database (GTDB). Inner (orange) and outer (blue) rings correspond to MAG completeness and contamination, respectively. Outer bar plot corresponds to MAG genome size.

(F) Per-MAG mean coverage across metastatic tumor microbiomes (n = 4,207) as computed by CoverM. Taxonomy annotations were provided by PhyloPhlAn using the SGB.Jul20 database. Colors represent MAG phylum-level annotation.

(G) Mapping coverage across a MAG genome (S57C3033) in a single sample using a 1-kb window, with an average coverage of 19× and a breadth of 1.0.

(H) Species-level gene set enrichment analysis (GSEA) of anerobic (top) or aerobic (bottom) bacteria enrichment in hypoxic tumors, as defined through HIF-1α staining.

(I) Consensus-based average nucleotide identity (conANI) as computed by inStrain on metagenomic-assembled genomes (MAGs) between tumors derived from the matched tumors pairs (light green, [left]) or unmatched random pairs (dark green, [right]). p values shown are derived from Wilcoxon rank-sum test.

(J) Species-level richness estimates from our MAG catalog between matched biopsies before and after treatment. p values shown were computed using linear effect model with repeated samplings found within the tool *Breakaway*.

(K) Genus-level designations of *Fusobacterium* high (red) or low (blue) quantified by clr transformed assembly-based counts of *Fusobacterium nucleatum*. p value shown was computed using a Wilcoxon rank-sum test.

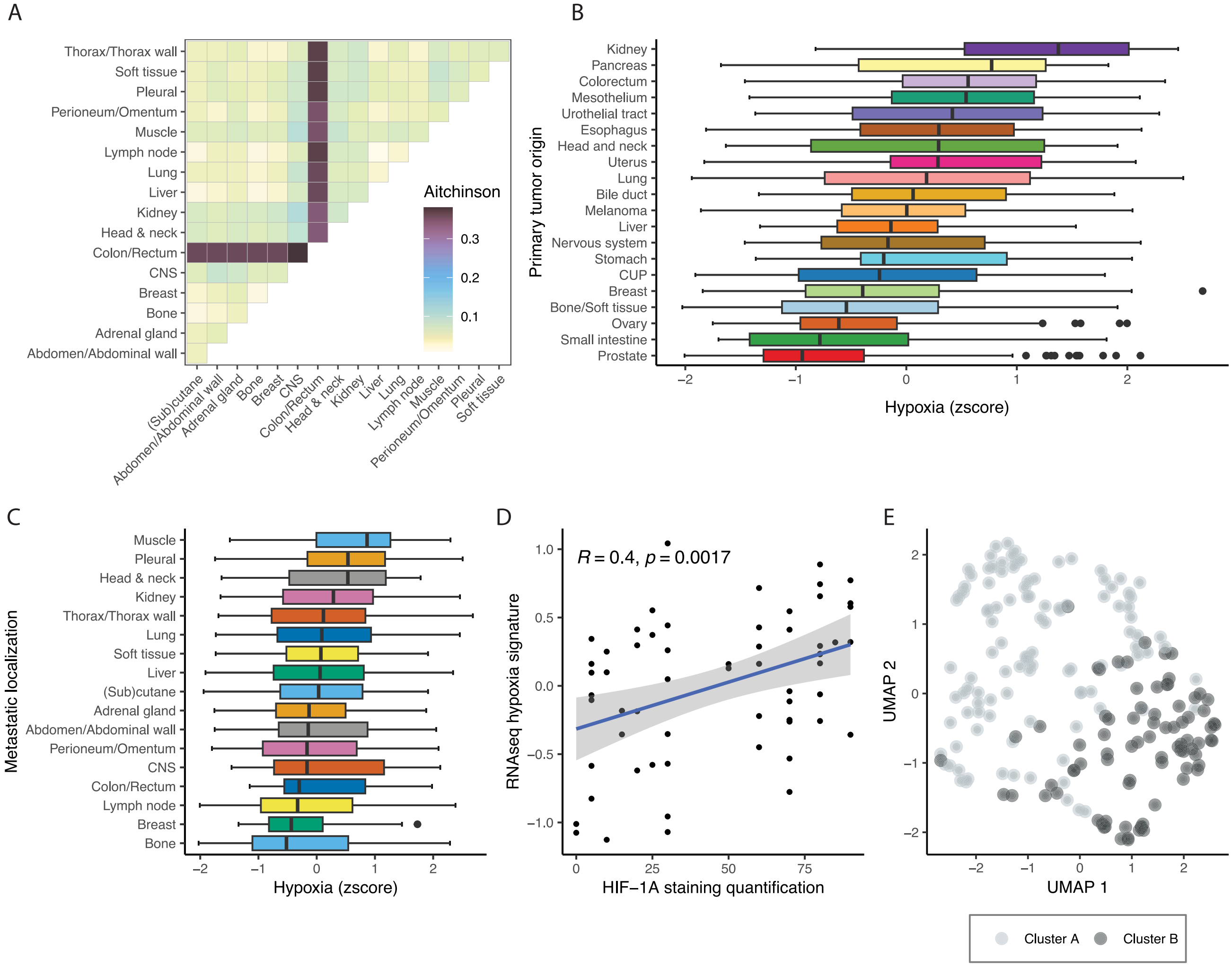


Figure S4. Community-based bacterial similarity between cancer types, related to Figure 2

(A) Pairwise metastatic localization site community Aitchison dissimilarities. *p* values were generated using the permutational multivariate ANOVA (PERMANOVA) adonis test after adjusted for primary tumor tissue origin, sequencing platform, and hospital of origin.

(B and C) (B) Boxplot of hypoxia activity across primary tumor location origin or (C) metastatic localization derived from gene set variation analysis (GSVA) scores.

(D) Correlation between transcriptomic-derived hypoxia scores and HIF-1α staining. *p* value shown was derived from Spearman correlation.

(E) Uniform manifold approximation and projection (UMAP) of center log ratios (clr) of microbial communities and colored according to community typing assignment (cluster A [gray] or a cluster B [black]).

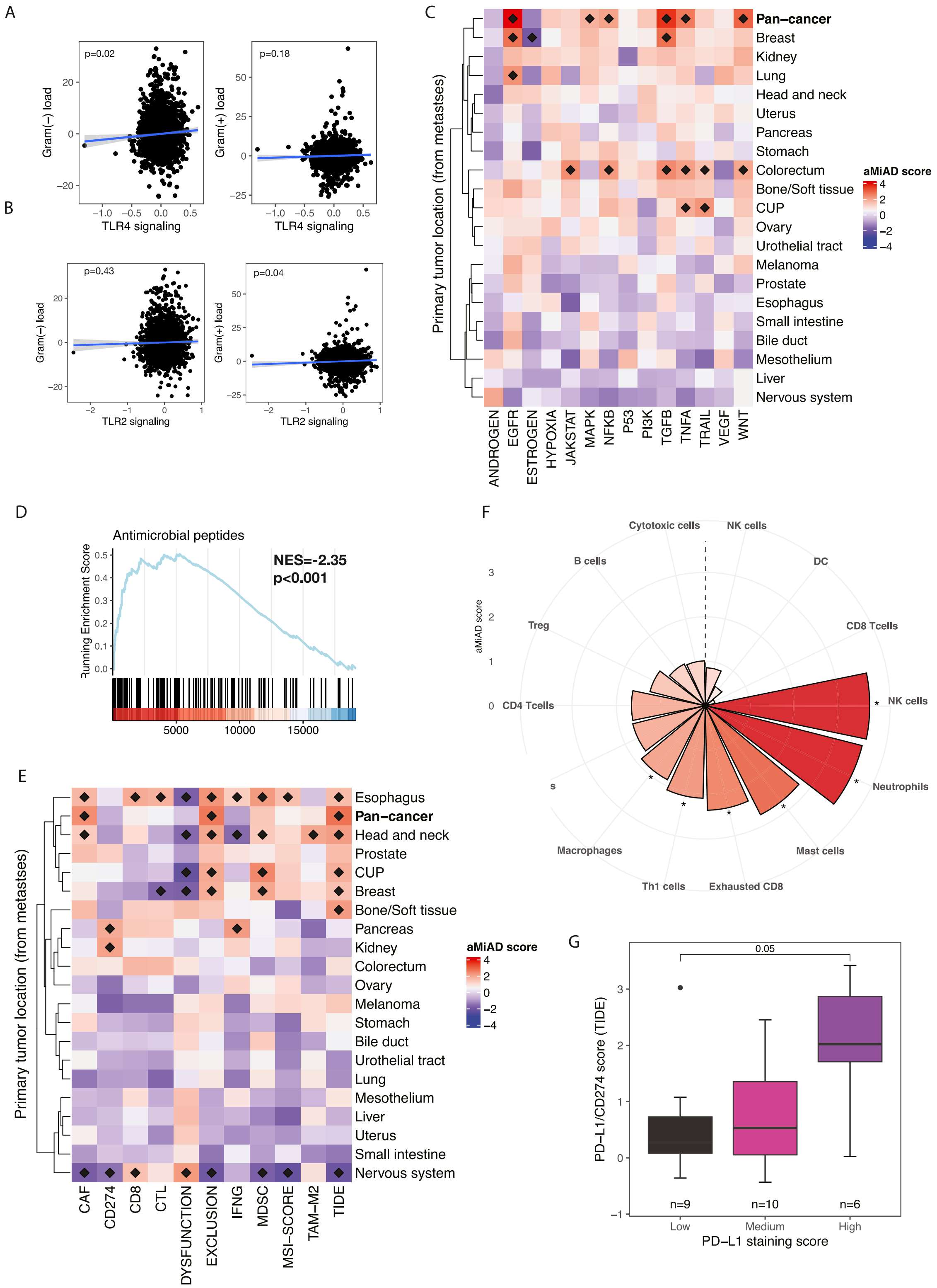


Figure S5. Associations between bacterial compositions and immune and tumor biology, related to Figure 3

(A) Correlation of residuals of TLR4 signaling and Gram-negative (LPS, left) or Gram-positive (LTA, right) load. *p* values were generated from an LME regression model after correcting for primary tumor location, metastatic localization site, sequencing platform, and hospital of origin.

(legend continued on next page)

(B and C) (B) Same as (A) but with TLR2 signaling; (C) heatmap of adaptive microbiome α-diversity-based association analysis (aMiAD) associations to PROGENy pathway quantifications across all tumor types (pan-cancer) or in individual cancer types. The shade of the bar indicates the aMiAD scores for each feature. Scores represent an enrichment (red) or depletion (blue) of the pathway to tumor microbiome community diversity. Associations with an FDR $p < 0.05$ is represented by diamonds.

(D) Running enrichment score of antimicrobial peptides gene sets in tumors with a more diverse tumor microbiome community (Shannon diversity). *p* value shown was computed using a GSEA.

(E) Same as (C) but with TIDE immune dysfunction gene signatures.

(F) Radar plot of aMiAD scores associated to the Danaher immune cell gene sets. Asterisks correspond to *p* values < 0.05.

(G) Correspondence of transcriptomic-derived TIDE-derived CD274 (PD-L1) score and PD-L1 immunohistochemistry (IHC) staining of low (<1% [black]), medium (1%–50% [light purple]), or high (>50% [dark purple]). *p* value shown was computed using a Wilcoxon rank-sum test.

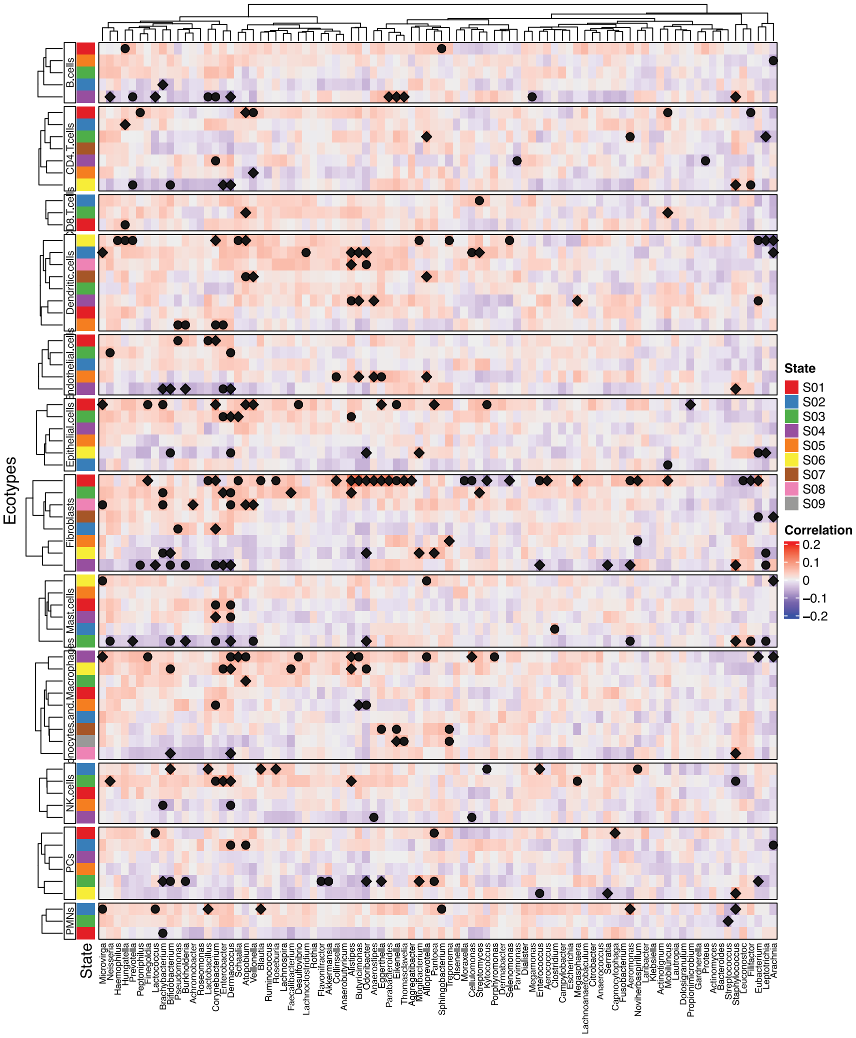


Figure S6. Hierarchical all-against-all (HAllA) correlations between bacteria and tumor-infiltrating immune states, related to Figure 3

Pairwise correlation between bacterial abundance and EcoTyper state abundances using hierarchical all-against-all (HAllA) analysis. Distinct colors represent cell states as previously described.[26] Values are shaded according to Spearman correlation coefficients. Red indicates a positive correlation, and blue indicates a negative correlation. *q* values < 0.10 are represented by solid diamonds, and circles represent *q* value < 0.05. *q* values shown were computed using HAllA with Spearman correlations and Bonferroni correction (STAR Methods).

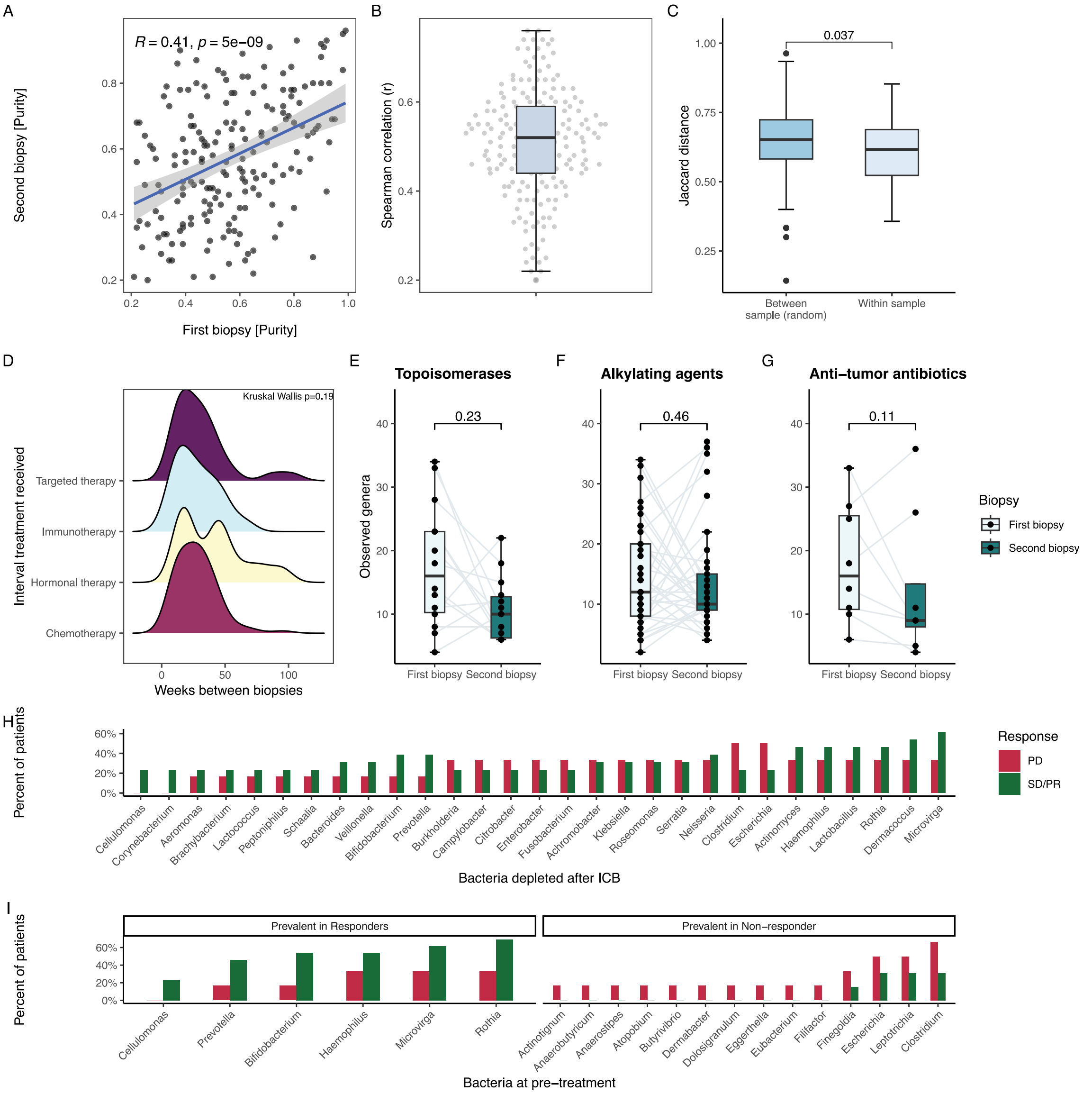


Figure S7. Dynamics of bacterial compositions over time and after anti-cancer therapy, related to Figure 4

(A) Scatterplot of the inferred tumor purity between tumor pairs between the first (x axis) and second (y axis) biopsy. *p* value shown was determined by Spearman correlation.

(B) Distribution of the per-tumor pair Spearman correlation.

(C) Boxplots of the Jaccard community dissimilarities between paired biopsies derived from the same lesion (within sample) and random sampling tumor pairs (between sample). *p* value shown was computed with Wilcoxon rank-sum test.

(D) Ridgeline plot of the weeks between tumor pair biopsies between different mode of treatments. Kruskal-Wallis one-way analysis of variance was used to compute the *p* value shown.

(E–G) (E) Boxplots of the number of observed genera between the first (light green) and second biopsy (dark green) after tumors were treated with chemotherapies, including classes of topoisomerases (E), alkylating agents (F), or anti-tumor antibiotics (G). The paired *p* value shown was determined using the *betta_random* framework, with adjustment for time between biopsies, primary tumor location, and lesion status.

(H) Bar plots of the frequency of responsive (green) or non-responsive (red) patients in which a bacterium was determined to be depleted after treatment with immunotherapy.

(I) Bar plots of the frequency of responsive (green) or non-responsive (red) patients in which a bacterium was present at pre-treatment.

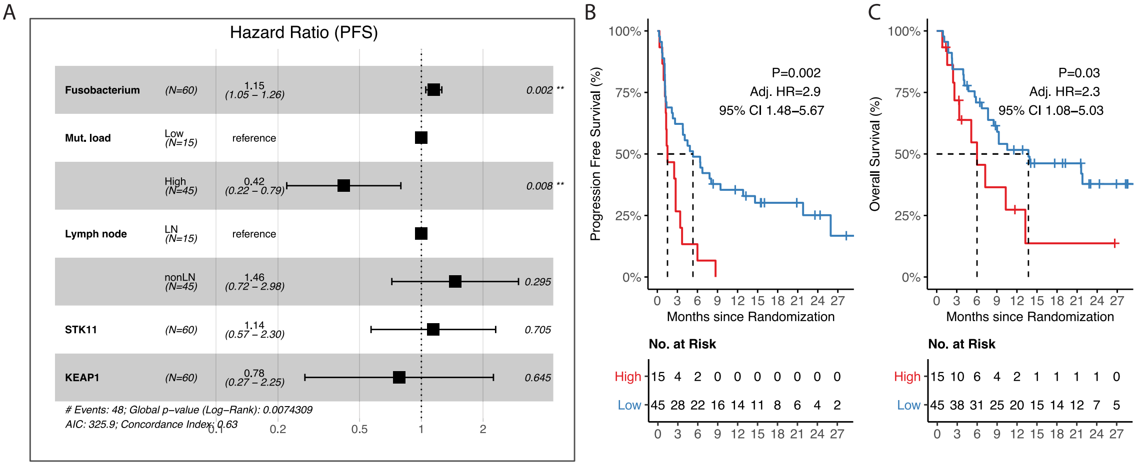


Figure S8. Fusobacterium as a biomarker for poor response to immunotherapy, related to Figure 5

(A) Multivariate Cox proportional hazard model of PFS, including *Fusobacterium* relative abundance, tumor mutational burden (no. of mutations per Mbp), lymph node metastatic localization, and STK11/KEAP1 resistance mutations.

(B and C) (B) Progression-free survival (PFS) and overall survival (OS) (C) of NSCLC patients treated with immunotherapy assigned to Fuso-high (red) or Fuso-low (blue) groups. Adjusted *p* values and HRs shown were computed with a multivariate Cox proportional hazard model and adjusted for confounders as described in (A).