Article:
Dohlman, A.B., Mjelle, R., Wood, H.M. et al. (2026) Biodiversity and biogeography of the multi-kingdom cancer microbiome. Cell. ISSN: 0092-8674

https://doi.org/10.1016/j.cell.2026.04.015

Biodiversity and biogeography of the multi-kingdom cancer microbiome

Anders B. Dohlman[1,2,3]*, Robin Mjelle[4], Henry M. Wood[5], Kevin Jiang[1,2,3], Alaina Shumate[1,2,3], Iris Lee[1,2,3], Gianmarco Piccinno[7], Garazi Serna[8], Abdul-Rakeem Yakubu[1,2,3]*, Paolo Nuciforo[8], Phil Quirke[5-6], Curtis Huttenhower[9], Nicola Segata[7], Matthew Meyerson[1,2,3,10]*

[1] Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA

[2] Department of Genetics, Harvard Medical School, Boston, MA, USA

[3] Cancer Program, Broad Institute of MIT and Harvard, Boston, MA, USA

[4] Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway

[5] Pathology and Data Analytics, Leeds Institute of Medical Research at St. James's, University of Leeds, Leeds, UK

[6] National Institute for Health and Care Research (NIHR) Leeds Biomedical Research Centre, Leeds, UK

[7] Department of Cellular, Computational and Integrative Biology (CiBIO), University of Trento, Trento, Italy

[8] Molecular Oncology Group, Vall d'Hebron Institute of Oncology, Barcelona, Spain

[9] Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

[10] Lead contact

* Correspondence:

andersb_dohlman@dfci.harvard.edu

matthew_meyerson@dfci.harvard.edu

Summary

Microorganisms represent an important component of the tumor microenvironment, but conflicting reports have left the extent of microbial prevalence across cancer types unclear, necessitating more robust methods for characterizing tumor-associated microbiomes. We built and benchmarked a host-subtraction and classification pipeline to identify microbiota in whole-genome sequencing data and applied it to 16,369 high-depth tumor whole genomes from the UK 100,000 Genomes Project. After decontamination, microbial signatures were indistinguishable from background in most cancer types. However, in orodigestive tumors, we detected multi-kingdom polymicrobial communities including bacteria, fungi, viruses, archaea, and in some cases, *Trichomonas*, a protozoan parasite. These communities varied by tumor site and subtype, with increased microbial colonization of microsatellite-instable and POLE/POLD1-mutated tumors, supported by a correlation between microbial load and tumor mutation burden observed across cancer types. This analysis helps to resolve pan-cancer microbial structure and links the tumor microbiome to host phenotype and tumor genomic context.

Keywords

cancer, tumor microbiome, biogeography, decontamination, multi-kingdom, bacteria, fungi, viruses, archaea, mutation burden

Introduction

The human microbiome is composed of many trillion cells, including bacteria, fungi, viruses, and archaea[1–4]. In healthy individuals, these microorganisms are found at anatomical barrier surfaces, including the skin, upper respiratory tract, urogenital tract, and are most concentrated in the orodigestive tract[1,2,5], with substantial variation in community structure across body sites[4,6–12]. These commensal microbiota play essential roles in human metabolism, digestion, and immunity[4,13–15]. Perturbations to these microbial communities have been associated with various human health conditions[16–23], including cancer[24–27], where variation in gut microbial composition has been reported to be linked with response to immunotherapy in both melanoma and lung cancer[28–33] and to chemotherapy efficacy in pancreatic cancer[34].

Beyond systemic interactions with their hosts, microbiota can directly colonize tumors, exerting influence on both tumor cells and their surrounding microenvironment[24,35–39]. To date, some of the most robust evidence for tumor-associated microbiota comes from cancers arising at body sites known to harbor microbes under normal physiological conditions, including human papillomavirus (HPV) in oral and cervical cancer[26]; Epstein-Barr virus (EBV) in nasopharyngeal[40] and gastric[41] cancers; hepatitis B and C viruses in liver cancer[42]; and *Helicobacter pylori* in gastric cancer[43,44]. In colorectal cancer (CRC), multiple bacteria have been reproducibly linked to the disease, including *Fusobacterium nucleatum*[35,36], enterotoxigenic *Bacteroides fragilis*[45,46], and *pks*+ *Escherichia coli*[47–52]. Collectively, microbial infections are estimated to account for at least 13% of global cancer burden[53].

For cancers occurring at non-barrier sites, there have been inconsistent reports regarding the scope of microbial colonization. Pan-cancer surveys of tumor tissues have yielded widely divergent findings regarding the presence, distribution, and cancer-type specificity of microbial communities[39,54–63], contributing to significant uncertainty in the field. Many of these prior efforts have relied on large-scale tumor sequencing datasets such as The Cancer Genome Atlas (TCGA)[57,59,60,62,64], the UK 100,000 Genomes Project (100kGP)[61], and metastatic cancers from the Hartwig Medical Foundation[65], each containing thousands of sequenced tumor samples. While these datasets harbor verifiable tumor-associated microbial sequences[24,57,61,62,64], their utility for capturing pan-cancer microbial structure has been hindered by the challenges of sequence misclassification, contamination, batch effects, and the absence of gold-standard negative controls[57,58,64], underscoring the need for more rigorous measures to distinguish true, tumor-specific microbial signatures from false positives.

To address technical false positives arising from misclassification of human sequences, our group previously developed PathSeq[66,67], a computational host-subtraction pipeline designed for detecting microbiota in low-biomass tissues. Applying this approach allowed our discovery of *Fusobacterium nucleatum* enrichment in colorectal cancer[35] while similar methods were deployed by others to discover Merkel cell polyomavirus in Merkel cell carcinoma[68]. Such host subtraction measures are essential not just for mitigating spurious microbial signals but also for preventing inadvertent exposure of patient-identifiable genomic information[69]. Applying PathSeq to TCGA, members of our group later used developed and benchmarked models to distinguish tumor-associated microbial signals from laboratory contamination, enabling pan-cancer characterization of tumor-associated bacteria[57] and fungi[59].

However, host filtering approaches are inherently limited by the incompleteness of the host reference genome used to perform subtraction. Moreover, prior microbial analyses have not been comprehensively verified in independent, comparably large-scale tumor sequencing cohorts of primary tumor tissue. Today, it is possible to study the cancer microbiome at much greater resolution, due to the availability of thousands of high-coverage whole-genome sequences (WGS) of cancer DNA from the UK 100,000 Genomes Project (100kGP)[70–74] as well as the completed telomere-to-telomere human reference genome T2T-CHM13[75,76] for host subtraction.

To address the longstanding challenge of reliably detecting microbiota in tumor sequencing data and to resolve ongoing uncertainties regarding the distribution of tumor-associated microbial communities, we implemented and benchmarked PathSeq-T2T, an updated filtering pipeline that incorporates the complete telomere-to-telomere human genome reference (T2T-CHM13) for host subtraction and multiple metagenomic classifiers, enabling high-confidence detection of microbial signals. We then applied this pipeline to the pan-cancer 100kGP database, which includes 16,369 high-depth tumor whole genomes from 15,237 participants. The resulting atlas of decontaminated microbial profiles allowed a far more detailed view of the cancer microbiome than was previously possible.

Results

A host-subtraction pipeline for detecting microbiota in low-biomass human tissues

In tissue samples with low microbial biomass, the incomplete removal of human sequences can produce false-positive microbial signatures[58] and may expose patient-identifiable genomic information[69]. To improve the detection of microbial signals in sequenced tumor tissues, we developed PathSeq-T2T, a host-subtraction and microbial detection pipeline that leverages the complete T2T-CHM13 human reference genome to isolate and then classify non-human reads from bulk tumor whole-genome sequencing data **(Figure S1A; STAR Methods)**. PathSeq-T2T performs quality and complexity filtering on reads not mapped to the standard human genome reference (GRCh38), together with additional subtractive host read filtering against polymorphic human immune loci (e.g., MHC genes), known breakpoint junctions, vector sequences, and T2T-CHM13. PathSeq-T2T then screens the remaining putative non-human sequences with three microbial classification tools: Kraken2[77] (which uses k-mer-counting to classify individual reads), MetaPhlAn4[78] (which detects clade-specific marker genes), and Sylph[79] (which measures k-mer containment within microbial genomes). This multi-classifier approach is included to allow cross-validation of microbial signals. After classification, PathSeq-T2T calculates the observed microbial reads per million starting reads (RPM), allowing quantification of microbial tumor burden.

Benchmarking PathSeq-T2T with in silico mixtures of microbial and human DNA

To assess PathSeq-T2T's ability to detect trace microbial signals in human tissue, we generated sequences from an *in silico* microbial community containing 17 species of bacteria, fungi, and archaea, then mixed them with human sequences at varying dilutions. This resulted in nine, tenfold serial dilution conditions containing pre-defined read counts for each species **(Table S1)**. These began at a microbe-to-host ratio of 1:1 and extended to a final dilution condition of $1:10^8$ which contained just one microbial read pair. We also included controls containing only human or only microbial reads.

PathSeq-T2T was consistently effective in removing ground-truth human sequences across the in silico dilution series, leaving zero human reads in every condition **(Figure 1A, Figure S1B)**. By comparison, host filtering with just the GRCh38-unaligned read set (an approach used in studies conducted prior to the release of T2T-CHM13[55,60,64]) resulted in the retention of 5 million human reads (>17,000 RPM) in the human-only condition. Meanwhile, filtering with the original

PathSeq pipeline retained 5,128 human reads (~17 RPM) in the human-only condition. Thus, PathSeq-T2T substantially outperforms previous methods of host subtraction.

After host filtering, PathSeq-T2T correctly identified the absence of microbial sequences in the human-only control. In conditions containing microbial reads, PathSeq-T2T retained on average 79.8% ± 7.6% of microbial sequences **(Figure 1B)**. Collateral subtraction of true microbial reads primarily affected microbial sequences with low quality and/or complexity and duplicated reads **(Figure S1C)**. The removal of these sequences resulted in a substantial reduction in read misclassifications **(Figure S1D)**.

Overall, the final filtered read counts in the *in silico* mixtures closely matched expected microbe-to-host ratios. The largest absolute subtraction in human reads occurred during the hg38 alignment step, whereas the largest fractional reduction occurred during T2T-CHM13 subtraction **(Figure 1C, Table S2)**. Among conditions containing microbial reads, 98.3 ± 0.4% of PathSeq-T2T-filtered reads were correctly assigned to on-target microbial taxa, while the remaining reads could not be uniquely classified **(Figure 1D)**. Notably, in the $1{:}10^8$ dilution condition, PathSeq-T2T retained the lone microbial read pair among >300 million starting reads and correctly classified it as *Veillonella rogosae* after removing all human reads, indicating robust host subtraction and high sensitivity at low microbial biomass.

Benchmarking PathSeq-T2T with in vitro mixtures of microbial and human DNA

Although PathSeq-T2T performed effectively on our *in silico* microbe-host mixtures, such data are unable to capture environmental contamination introduced by laboratory manipulations during sample preparation, DNA extraction, and sequencing. We next generated mixtures of microbial and human DNA *in vitro*, using the same experimental design as before. Human and microbial DNA were extracted from a colorectal cancer cell line and microbial community standard, respectively, then mixed in triplicate across the same serial dilution ladder (1:1 to $1{:}10^8$). We then sequenced each mixture to a target of 30X human genome coverage (See STAR Methods). We also included host-only, microbe-only, and template-free extraction controls.

Consistent with the *in silico* mixtures, PathSeq-T2T filtering of *in vitro* mixtures achieved clear separation of the dilution series, with on-target microbial read counts largely matching expected ratios **(Figure 1E, Table S3)**. Moreover, PathSeq-T2T outperformed host filtering with GRCh38 alone or with the original PathSeq pipeline **(Figure 1F)**, demonstrating effective human

read removal and microbial detection on real sequencing data. Similar to our *in silico* mixtures, a subset of *in vitro* sequencing reads could not be uniquely classified, especially at lower dilutions.

In addition to these on-target microbial sequences, we detected evidence of contamination across nearly all *in vitro* experimental conditions, even after sterile handling during DNA extraction and mixing **(Figure 1G)**. This included off-target, skin-associated taxa[12] such as *Cutibacterium acnes*, *Staphylococcus epidermidis*, and *Malassezia restricta* **(Figure 1H)**, which were not included in the community standard. Each of these off-target species was also detected in template-free negative controls, confirming their status as contaminants **(Figure 1H)**. While contamination from these species varied across conditions, their read counts were highly correlated across samples, suggesting they were introduced in unison. The proportion of off-target microbial sequences generally increased with successive dilutions, despite some variability at lower dilution conditions **(Figure 1J)**. This mirrors prior observations that contamination disproportionately affects low-biomass samples, even as the absolute level of contamination remains consistent[80–82].

We also observed low levels of on-target microbial sequences in human-only samples that did not receive aliquots of microbial DNA **(Figure 1G)**. This is a phenomenon that can occur due to cross-contamination between sample conditions[81,82] or due to barcode swapping on the flow cell[82–84]. Collectively, these sequences accounted for 82 to 246 reads per sample (0.1 to 0.3 RPM). We therefore determined 0.1 RPM as the threshold for expected background that could be expected from cross-contamination in otherwise sterile samples.

Testing PathSeq-T2T on cancer cell lines

We next sought to evaluate PathSeq-T2T performance on genomically distinct cancer cell lines, which are not expected to harbor microbiota. We tested microsatellite-stable (MSS; n = 12) or microsatellite-unstable (MSI; n = 8) colorectal cancer cell lines with whole-genome sequencing data available from The Cancer Cell Line Encyclopedia[85]. No differences were observed between MSS and MSI cell lines, indicating that the presence of high mutation burden did not critically impact host filtering or microbial detection **(Figure S1E)**. However, we did observe a substantial enrichment of microbial sequences in cell lines with DNA libraries prepared using PCR **(Figure S1F)**. Examining the microbial composition of these cell lines, we observed variation in taxa between library types. Both groups included common contaminants such as *Cutibacterium, Pseudomonas, Sphingomonas, Variovorax,* and *Bradyrhizobium* **(Figure S1G)**.

Testing PathSeq-T2T on tumor and germline samples from 100kGP

Having validated and benchmarked PathSeq-T2T with *in silico* and *in vitro* host-microbe mixtures as cancer cell lines, we began to evaluate the pipeline's performance on real-world, tumor whole-genome sequencing data from the UK 100,000 Genomes Project (100kGP) **(Table S4)**. Patient biospecimens sequenced by this project include both tumor tissues (*n* = 16,369, ~100X coverage) and matched samples collected for the purpose of germline profiling (*n* = 15,249, ~30X coverage), including peripheral blood and saliva. As an initial performance analysis prior to studying the entire 100kGP cancer database, we compared tumor and germline samples with a range of expected microbial biomass. Among germline samples, we compared saliva (*n* = 490), which is typically microbe-rich, and peripheral blood (*n* = 14,127), which is believed to be microbe-low or sterile under normal conditions[86]. Among primary tumor sequences, we compared colorectal tumors (CRC, *n* = 2,482) and glioblastoma multiforme brain tumors (GBM, *n* = 295).

On these sample types, PathSeq-T2T yielded expected sample-type- and cancer-type-specific variation in the proportion of sequences retained after filtering **(Figure 2A-B)**. Filtering removed >99.99% of reads in blood and GBM, while retaining a median of 3.87% and 0.02% of reads from saliva and CRC samples, respectively. Among germline samples, we detected microbial reads at much higher levels in saliva (median 27,500 RPM) than blood (1.7 RPM), while among tumor samples, we detected more microbial reads in CRC samples (145 RPM) than GBM (0.7 RPM). This indicated that PathSeq-T2T was correctly retaining microbial sequences in sample types known to harbor microbiota, while mostly eliminating them in sample types that are sterile or microbe-low.

To further understand the sequences that remained in these sample types after host filtering, we examined their microbial composition. The taxa found in saliva samples was highly distinct from those found in blood **(Figure 2C)**, with saliva frequently being dominated by *Prevotella, Streptococcus, Rothia,* and *Veillonella*, which are typical of the oral cavity. Similarly, colorectal cancer (CRC) tissues exhibited a distinct microbial composition compared to glioblastoma (GBM) **(Figure 2D)**, containing high rates of *Bacteroides, Phocaeicola,* and *Fusobacterium*, which are known to be found in CRC tissues[8]. Blood and GBM samples typically harbored similar genera, including *Achromobacter*, *Sphingomonas, Pseudomonas, Variovorax, Bradyrhizobium,* and *Cutibacterium*. As many of these taxa are commonly associated with human skin, the environment, and laboratory reagents[80,81], we suspected that many of the microbiota detected in blood and GBM samples were contaminants.

To further interrogate this possibility, we compared species prevalence between sample types. Between saliva and blood samples, most detected species were unique to saliva, while a smaller subset of species were found at nearly equal rates in both sample types **(Figure 2E, Figure S2A-B)**. A similar pattern was observed comparing CRC and GBM **(Figure 2F, Figure S2C-D)**. However, no species were found to be unique to blood or GBM, suggesting that taxa found in these microbe-low sample types were likely contaminants shared across cancer types. Consistent with this, species equiprevalent in saliva/blood (*n* = 125) and in CRC/GBM (*n* = 161) overlapped substantially (*n* = 110; *Fisher's p* = 1.6e-98), suggesting a common pool of contaminants **(Figure S2E)**. Examining these equiprevalent species further supported their status as contaminants. In blood, the most prevalent such taxa were *Pseudomonas aeruginosa* and *Variovorax paradoxus* **(Figure 2G)**, while among CRC/GBM-equiprevalent taxa, prominent species included *Cutibacterium acnes* and *Bradyrhizobium sp. BTAi1* **(Figure 2H)**. By comparison, the most frequently detected species in saliva and CRC samples were canonically oral[11] and colorectal[8] microbiota **(Figure S2F-G).** These results showed that, in addition to recovering biologically plausible and sample-type-specific microbiota, PathSeq-T2T was also detecting contaminants that appeared to be present across sample types. This demonstrated the need for additional decontamination steps beyond host filtering alone.

Decontamination of 100kGP microbiomes

We next examined microbial sequences across the entire 100kGP dataset, which amounted to a total of 16,369 tumor and 15,249 germline samples spanning 28 distinct cancer types. Since contaminants seemed to appear at similar rates across all sample types, we designed a pan-cancer equiprevalence (PCE) score to quantify how evenly distributed a species was across distinct cancer types in the 100kGP cohort. This score was calibrated such that a species would receive a score of 0 if it was detected in 100% of samples from just one cancer type, and a score of 1 if it was detected at equal frequency in every cancer type (see STAR Methods). Applying this score, we found that many of the species that were equiprevalent in saliva/blood or CRC/GBM also appeared to be equiprevalent across the entire 100kGP cohort **(Figure 2I, Figure S2H-I, Table S5)**.

To assess the ability of pan-cancer equiprevalence to distinguish between environmental contamination and true tumor-associated taxa, we examined species known to be associated with cancer **(Figure 2J)**. This recapitulated expected cancer-type-specific microbial colonization patterns. *Helicobacter pylori* (PCE = 0.32), a known driver of gastric cancer, was enriched in gastric tumors; *Fusobacterium nucleatum* (PCE = 0.55), an oral pathogen associated with

colorectal cancer, was enriched in both oral squamous and colorectal cancers; and the human papillomavirus 16 species group (HPV; PCE = 0.25), implicated in oropharyngeal carcinogenesis, was specific to that disease. By comparison, skin-associated taxa *Cutibacterium acnes* (PCE = 0.90), *Staphylococcus epidermidis* (PCE = 0.78), and *Malassezia restricta* (PCE = 0.86) were equiprevalent across cancer types, indicating that 100kGP samples harbored similar contaminants to those found in our *in vitro* benchmarking experiment. Across the dataset, the most equiprevalent species were *Pseudomonas aeruginosa* (0.91), *Stutzerimonas stutzeri* (0.91), *Acidovorax temperans* (0.91), *Pseudomonas oleovorans* (0.91), and *Sphingobium yanoikuyae* (0.90) **(Figure S2J)**, indicating that these species were also likely contaminants.

As known contaminants consistently had PCE scores above 0.7 **(Figure 2I)**, we selected this value as a threshold for identifying and removing contaminant taxa. Interestingly, examining microbial genomes assembled from sequenced tumor tissues, we observed that species exceeding this threshold shared near-identical, strain-level average nucleotide identity (>99.9%) despite coming from completely different samples **(Figure S2K, Table S6)**, suggesting a common, exogenous contamination source. Other non-contaminant species exhibited greater genomic diversity between samples, indicating a more distant common ancestor.

Consistent with our analysis of cancer cell lines, putative contaminants were also frequently enriched in tumor and germline samples whose DNA libraries had undergone PCR amplification **(Figure S2J-K)**, indicating contamination associated with PCR reagents and/or amplification of contaminating microbial DNA. Contamination was also pronounced in formalin-fixed, paraffin-embedded (FFPE) samples compared to fresh-frozen tissues **(Figure S2L)**. We therefore restricted downstream analyses to fresh-frozen solid tumors sequenced from PCR-free libraries.

Tumor microbiota are largely restricted to cancers of the orodigestive tract

We next examined microbial abundance before and after decontamination. Prior to removing contamination, most cancer sites exhibited microbial signals above background **(Figure S3A-C)**, with the abundance of contaminant species generally consistent across cancer sites **(Figure S3D-F)**. However, after decontamination, only a subset of cancer types – specifically colorectal, oropharyngeal, esophageal, and gastric cancers – consistently retained microbial signal exceeding background (RPM > 0.1). This finding was reproduced by Kraken2, MetaPhlAn4, and Sylph **(Figure 3A-C)**, which were highly concordant in their abundance estimates **(Figure S3G).**

Comparing species prevalence across cancer sites, we observed that tumors arising from oropharyngeal, esophageal, gastric, and colorectal sites also harbored DNA sequences from many cancer-site-specific taxa, while other cancers contained relatively few **(Figure 3D, Figure S3H)**. Broadly, microbial abundance at these sites aligned with prior estimates of microbial density along the digestive tract[1,2,5], with colorectal tumors showing the highest abundance (IQR: 18–285 RPM), followed by oropharyngeal tumors (3–189 RPM), and then esophageal (0.3–24 RPM) and gastric cancers (0.4–19 RPM).

These cancer-associated microbial communities were typically polymicrobial **(Figure 3E)**, with more than one species detected in most colorectal (94.7%), oropharyngeal (77.0%), esophageal (59.8%), and gastric (62.7%) tumors. Their microbial composition mirrored known patterns of site-specific biogeography and biodiversity observed in healthy individuals[4,6–8,10] as well as in prior studies of orodigestive cancers[35,36,57,87–92] **(Figure 3F)**. Colorectal tumors commonly featured members of the *Bacteroidaceae* and *Lachnospiraceae* families, with *Fusobacteriaceae* enriched in both colorectal and oropharyngeal cancers. *Prevotellaceae* and *Streptococcaceae* were present across all four sites but were especially abundant in oropharyngeal and esophageal tumors. Gastric cancers were unique in harboring species from the *Lactobacillaceae* and *Helicobacteraceae* families. Meanwhile, the viral family *Papillomaviridae* (i.e., HPV-16) was uniquely detected in oropharyngeal tumors **(Figure S3I)**. Cervical cancers, which also typically harbor HPV-16/18, were not included in the 100kGP cohort.

Sparse microbial signals in non-orodigestive tumors

Outside of orodigestive cancer types, microbial signals were highly sparse. Examining the subset of samples from other cancer types that were microbe-high (>10 RPM), we observed some signals that were biologically plausible **(Figure S3G)**. For example, skin cancers[93,94] occasionally harbored Merkel cell polyomavirus (2.9% of cases) **(Figure S3J)** and *Staphylococcus aureus* (6.0%)[95–97]. Some pancreatic tumors[37,38,98,99] harbored gut-typical taxa (*Veillonellaceae*, *Fusobacteriaceae,* and *Enterobacteriaceae*), potentially reflecting proximity to the digestive tract, while a subset of bladder tumors[100–102] harbored species previously observed in the genitourinary tract, including *Peptoniphilus genitalis* and *Actinotignum timonense/shaali*[103,104].

Across other cancer types, microbe-high samples more often resembled acute infection rather than resident communities. For example, one lung cancer contained DNA from *Streptococcus pneumoniae* (10 RPM) while another contained *Haemophilus influenzae* (25 RPM), both common causes of pneumonia[105]. The remaining microbe-high samples from non-

orodigestive tissues were typically dominated by *Herpesviridae* species, which are mostly widespread among humans and acquired early in life[106–110]. Among *Herpesviridae*, HSV, EBV, CMV, and HHV-7 displayed varying degrees of cancer-type specificity **(Figure S3I)**, while HHV-6A and HHV-6B were more broadly distributed **(Figure S3J)**.

Thus, although microbial signals were generally sparse in non-orodigestive cancers, some harbored biologically plausible taxa. These signals could represent acute infections or potentially more prolonged colonization through mechanisms that have not yet been determined. However, they could also have arisen due to cross-contamination from adjacent organs, barcode swapping, or sample mislabeling.

Biogeography of oropharyngeal, gastroesophageal, and colorectal cancers

Since tumors of the orodigestive tract consistently harbored abundant, polymicrobial communities, we next sought to more thoroughly characterize and quantify these communities across anatomic sites. Concordant with prior biogeography studies of healthy tissues[1,2,5], we observed the highest microbial load and diversity in tumors of the lower gastrointestinal tract, followed by those of the oral cavity **(Figure 4A, Figure S4A)**, while microbial abundance and diversity were generally lower in esophageal and gastric sites.

Using microbial read counts to estimate bacterial density **(Figure S4B)**, we predicted a median burden of 20-30 bacteria per 1,000 human cells in oral and sinonasal cancers, while bacteria-host cell ratios were lower in esophageal and gastric sites (1-3 bacteria per 1,000 cells). In the lower gastrointestinal tract, microbe-host cell ratios varied by anatomic site, and generally declined progressively from the cecum (227 per 1,000), through the ascending (251 bacteria per 1,000), transverse (229 per 1,000), and descending colon (119 per 1,000), reaching their lowest rates in the rectum (77 per 1,000). However, these bacteria-to-host ratios were highly variable within individual sites and sometimes even exceeded 1:1, consistent with the substantial heterogeneity observed in spatial analyses[111] and tumor-associated biofilms[112].

We next examined variation in community structure across sites. Principal coordinate analysis showed that oropharyngeal, gastroesophageal, and colorectal tumors harbored distinct microbial communities ($p < 0.001$) **(Figure 4B)**. Cancers of the oropharynx and gastroesophageal tract showed some site-specific variation but primarily clustered together, consistent with their anatomic proximity, while cancers of the lower gastrointestinal tract clustered separately.

Consistent with this, we found both distinct and shared microbial taxa at each site **(Figure 4C, Figure S4C).** Of the 543 species detected in at least 1% of samples from any of these three sites, the largest fraction was unique to colorectal cancer (42.0%), while fewer were unique to oropharyngeal (11.8%) and gastroesophageal (2.6%) cancers. The degree of species overlap between these sites was greatest between oropharyngeal and gastroesophageal cancers (8.5%) followed by oropharyngeal and colorectal (7.9%), while very few species (0.9%) were shared exclusively between gastroesophageal and colorectal cancers. A core set of species (26.3%) were detected across all three sites.

This variation was mirrored in relative abundance profiles **(Figure 4D)**, which further highlighted the biodiversity and polymicrobial nature of these cancer types. *Prevotella* was the most abundant genus in both oropharyngeal and gastroesophageal tumors, with a mean relative abundance of 19.4% and 17.7%, respectively, often accompanied by *Haemophilus, Streptococcus,* and *Leptotrichia*. Oropharyngeal cancers also frequently contained high relative abundances of *Fusobacterium* (9.9%)*, Capnocytophaga* (5.3%)*,* and *Porphyromonas* (2.6%), while a subset were dominated by *Alphapapillomavirus*/HPV (7.2%). Gastroesophageal tumors exhibited higher relative levels of *Veillonella* (6.6%)*, Selenomonas* (6.6%)*,* and *Helicobacter* (5.3%)*.* Meanwhile, colorectal cancers had high average relative abundances of *Bacteroides* (27.3%) and *Phocaeicola* (10.6%), while also frequently containing *Fusobacterium* (5.7%)*, Escherichia* (4.1%)*,* and *Segatella* (3.7%).

Examining species prevalence along the orodigestive axis **(Figure 4E, Figure S4D)**, we observed enrichment of the periodontal disease associated species[113] *Porphyromonas gingivalis, Treponema socranskii*, and *Capnocytophaga gingivalis* in oropharyngeal cancer, while in gastric and esophageal cancers, we observed site-specific enrichment for *Helicobacter pylori* and acid-tolerant species including *Limosilactobacillus fermentum* and *Lactobacillus gasseri*[114,115]. Taxa shared between oropharyngeal and colorectal cancers included several oral-typical pathogens like *Fusobacterium animalis, Parvimonas micra, Solobacterium moorei,* and *Prevotella intermedia*, whereas *Bacteroides fragilis, Escherichia coli,* and *Akkermansia muciniphila* were largely colorectal-specific and varied between the proximal and distal colon. *Streptococcus anginosus,* which has been shown to promote gastric cancer in mice[116], was detected at similar rates across orodigestive sites.

Overall, microbial load and composition along the orodigestive axis broadly recapitulated known biogeography and biodiversity. However, the frequent presence of oral-typical species in

non-oral tumor sites suggests a partial breakdown of normal tissue-site specificity in cancer. Thus, while anatomic location appears to be the primary determinant of tumor microbiome composition, cancer-associated changes to the local microenvironment during cancer initiation or progression may encourage colonization by allochthonous species.

Orodigestive cancers harbor a multi-kingdom community of bacteria, fungi, archaea, viruses, and sometimes parasites

Viruses and bacteria have long been known to infect certain tumor tissues[26,35,36,40,41,43,53,68], while parasite infections have been linked to higher rates of bladder cancer in endemic areas[117]. More recently, studies have suggested that other domains, including fungi[59,60], also colonize tumor tissues. Therefore, we sought to characterize the multi-kingdom microbial composition of tumors in the 100kGP cohort, including fungi, archaea, viruses, and parasites.

Consistent with our previous findings from TCGA tumor tissues[59], we observed fungi across oropharyngeal, gastroesophageal, and colorectal cancers, including *Candida albicans*, *Nakaseomyces glabratus* (formerly *Candida glabrata*), and *Saccharomyces cerevisiae* (**Figure 4F**). Each of these fungi exhibited pronounced site-specific colonization patterns: the most prevalent site for *C. albicans* was in esophageal (7.8%) cancers and in tumors arising at the gastroesophageal junction and gastric cardia (6.9%), while *N. glabratus* was enriched in non-cardia gastric cancers, with greatest prevalence in the antrum/pylorus (5.3%). The yeast *S. cerevisiae* was also detected in tumors of the gastric antrum/pylorus (5.3%) as well as throughout the lower gastrointestinal tract.

Examining other microbial kingdoms, we found enrichment of DNA sequences from archaea in tumors of the lower gastrointestinal tract. *Methanobrevibacter smithii* **(Figure 4G)**, the predominant archaeon in the human gut[10,118,119], was most prevalent in tumors of the splenic flexure (25.9%) and descending colon (25.6%), with the lowest prevalence in the cecum (8.6%). *M. smithii* is a methane-producing species that has been discussed as a factor in irritable bowel syndrome (IBS)[120] and colorectal cancer[119]. We also detected the related archaea *Methanobrevibacter sp. TLL-48-HuF1* and *Candidatus Methanomassiliicoccus intestinalis* **(Figure S4E)**. Less is known about these other archaeal species, but both were originally isolated from human fecal samples and are reportedly methanogenic[121,122]. We did not detect any archaea in oropharyngeal or gastroesophageal sites.

In addition to fungi and archaea, we detected DNA sequences from the parasite *Trichomonas* in oropharyngeal cancers (2.6%), rectal cancers (0.8%), and cancers of the ascending colon (1.8%) **(Figure 4H)**. *Trichomonas* are single-celled flagellated protozoan parasites and include the causative agents of trichomoniasis (*T. vaginalis*), a sexually transmitted infection that typically affects the urogenital tract[123], with links to cervical cancer[124], and *T. tenax*, which is associated with necrotizing gingivitis. Many *Trichomonas* infections are asymptomatic[123]. In the 100kGP cohort, *Trichomonas* was most abundant in oropharyngeal cancers, at rates up to 12.75 RPM, corresponding to roughly one parasite per 1,000 human cells. *Trichomonas* presence was confirmed using both Kraken2 and MetaPhlAn4 (no protozoa database was available for Sylph). The affected oropharyngeal cancer cases occurred almost exclusively in men with HPV-negative disease.

Finally, we detected site-specific viral signals across oropharyngeal, gastroesophageal, and colorectal cancers **(Figure 4I)**. HPV-16 was found exclusively in oropharyngeal cancers (13%), while herpesviruses HHV-6A and HHV-6B were distributed throughout cancers of the lower gastrointestinal tract. CMV and EBV were detected in multiple sites, but EBV was most prevalent in gastric cancers of the fundus/body (13%). Interestingly, we also observed specificity of HHV-7 to gastric tumors, with highest prevalence in the gastroesophageal junction/cardia (10%) and the fundus/body (8.7%). Latent HHV-7 has previously been reported in gastric mucosa[125] and has been associated with EBV reactivation leading to mononucleosis[126], however its role in gastric cancer remains largely unexplored.

Overall, these findings show that while bacteria are by far the most abundant and diverse microbial kingdom associated with orodigestive tumors, these cancers may also become colonized by a broader, multi-kingdom microbial community that includes viruses, fungi, archaea, and in some cases, parasites.

Microbial load is elevated in hypermutated genomic subtypes of cancer

In colorectal cancer, the tumor-associated microbiome is a key modulator of the tumor microenvironment[27]. Prior studies have reported differences in microbial composition and diversity between anatomical sites within the colon, both in healthy individuals[4,6–10] and in the context of cancer[127,128], with the greatest differences observed between proximal (right-sided) and distal (left-sided) tumor sites.

Other studies have indicated that microbial colonization patterns may differ by genomic subtype, most notably by the observed enrichment of *Fusobacterium nucleatum* in microsatellite-instable colorectal cancer[129,130]. Genomically, colorectal cancers can become hypermutated due to deficiencies in DNA mismatch repair or due to mutations in the proofreading domains of DNA polymerase ε (*POLE*) or δ (*POLD1*)[73,131–133]. These characteristics respectively define microsatellite-instable (MSI; approximately 15% of colorectal cancer cases) and *POLE/POLD1*-mutated (polymerase-mutant; 1-2% of cases) genomic subtypes. Hypermutated subtypes appear most frequently in the proximal colon and are typically more immune-infiltrated, with a more favorable prognosis. In contrast, non-hypermutated cancers, described as microsatellite-stable (MSS; 80-85% of cases), are more common in the distal colon and rectum and are frequently driven by chromosomal instability[132,133].

Leveraging the large size of the colorectal cancer cohort in 100kGP along with detailed clinical annotations and tumor genotyping provided by Cornish et al.[73], we performed an in-depth characterization of colorectal microbiomes and their phenotypic and genotypic correlates. To begin, we performed a multivariable analysis (PERMANOVA) which revealed that tumor anatomic site ($p < 0.001$) and genomic subtype ($p < 0.001$) were the two strongest independent predictors of tumor microbiome composition in colorectal cancer **(Figure 5A)**. These features together accounted for the majority of explainable variance in composition; however, we also observed weaker but statistically significant correlations with patient age ($p < 0.05$) and history of radiotherapy ($p < 0.05$).

To investigate how microbial colonization patterns differ by anatomic site and genomic subtypes, we next examined tumor microbial abundance and diversity. Consistent with Cornish et al., who reported elevated bacterial load in MSI colorectal cancers[73], we found that overall bacterial load was significantly greater in hypermutated colorectal cancer cases compared to non-hypermutated cases, regardless of genomic mechanism. Compared to MSS cases, we observed 3.9-fold greater microbial density (± 1.1) in MSI cases ($p = 4.61e-28$) and 6.5-fold greater density (± 1.6) in polymerase-mutant cases ($p = 2.35e-3$) **(Figure 5B)**. This trend persisted even after stratifying hypermutated subtypes by tumor site **(Figure S5B)**. Proximal and distal MSI tumors showed 3.2-fold ($p = 4.26e-13$) and 2.9-fold ($p = 5.28e-3$) enrichment in bacterial abundance compared to MSS, respectively, while proximal polymerase-mutant tumors harbored an 8.8-fold enrichment compared to proximal MSS ($p = 2.09e-3$). The enrichment of microbiota in MSI and polymerase-mutant tumors appeared to be driven mostly by bacteria **(Figure S5C)**, however we also observed a modest enrichment of fungal sequences in MSI compared to MSS (FC = 1.5; *p*

= 1.8e-4) **(Figure S5D)**. No significant differences in viral or archaeal load were observed **(Figure S5D-E)**. Interestingly, despite increased microbial load in hypermutated colorectal cancers, we observed a decrease in microbial diversity in MSI cases (p = 4.64e-3) **(Figure 5D, Figure S5F)**, suggesting that this pattern was driven by a limited number of species rather than by a proportional expansion across the entire microbial community.

To validate these findings, we used decontaminated microbial profiles from The Cancer Microbiome Atlas[57] (derived from TCGA whole-genome sequencing), which confirmed the enrichment of bacterial load in MSI colorectal cancers (FC = 5.3 ± 1.78; p = 2.43e-2). Examining gastric cancers in TCGA, we also observed enrichment of microbial load in MSI cases (9.7-fold ±1.8; p = 2.25e-4) **(Figure 5C)**, recapitulating a recent finding by Booth et al.[92] and suggesting a trend that was generalizable beyond colorectal cancer.

Tumor mutation burden independently predicts tumor microbial load

As MSI and polymerase-mutant tumors are characterized by an elevated mutation rate, enrichment of microbiota in these subtypes led us to hypothesize a broader relationship between microbial colonization and tumor mutation burden. Strikingly, microbial load in colorectal cancers was indeed correlated with tumor mutation burden within both MSS (R = 0.132; p = 2.36e-7) and MSI cases (R = 0.126; p = 1.64e-2) **(Figure 5E)**. This association persisted after adjusting for the effect of genomic subtype on tumor mutation burden and the effect of tumor site on microbial load (R = 0.081; p = 2.34e-3) **(Figure 5F)**. Quantitatively, each 10-fold increase in tumor mutation burden corresponded to an average 4.7-fold increase in microbial abundance (±1.4 SE) among MSS tumors and a 3.6-fold increase (±1.2 SE) in MSI tumors, suggesting a “dose-dependent” relationship.

We next asked whether the previously observed enrichment of microbiota in hypermutated subtypes could be explained entirely by tumor mutation burden rather than by subtype-specific biology. To test this, we fit a linear model predicting microbial abundance from genomic subtype, adjusting for tumor mutation burden and tumor site. When tumor mutation burden was included in this model, we found that MSI (p = 0.40) and polymerase-mutant (p = 0.75) subtypes were no longer predictive of microbial load, even as tumor site remained predictive (p = 2.3e-8) **(Figure 5G)**. Moreover, a model incorporating tumor mutation burden, tumor site, and subtype performed no better than a model including only tumor mutation burden and tumor site (p = 0.67), signifying that aside from tumor mutation burden, subtype-intrinsic features could not explain additional variation in microbial abundance.

Having observed microbial enrichment in MSI gastric cancers, we wondered if the relationship between tumor mutation burden and microbial abundance generalized beyond colorectal cancer. Indeed, oropharyngeal cancers demonstrated a significant positive correlation between tumor mutation burden and microbial load in both 100kGP (R = 0.145; p = 2.4e-2) and TCGA data (R = 0.154; p = 4.3e-2) **(Figure S5G-H)**. The same pattern was not observed in 100kGP gastroesophageal cancers, but microbial load in TCGA gastric cancers strongly correlated with tumor mutation burden (R = 0.251; p = 5.2e-4) **(Figure S5I-J)**. Together, these findings implicate tumor mutation burden as a key predictor of microbial colonization across subtypes, cohorts, and cancer types.

Microbial composition varies independently by anatomic site and tumor mutation burden

Having established that microbial load correlates with both anatomic site and tumor mutation burden, we next asked whether these features were also associated with differences in microbial composition. To delineate the effects of these two features, we analyzed associations between species-level relative abundances and tumor mutation burden (high vs. low) within each anatomic site. We then performed the complementary analysis, evaluating associations between species composition and anatomic site (proximal vs. distal), stratifying by tumor mutation burden **(Figure 5H, Table S8)**.

This analysis revealed several species that were significantly associated with tumor mutation burden in colorectal cancer, including *Fusobacterium nucleatum* (q = 4.8e-11) and *Bacteroides fragilis* (q = 3.3e-5), which have previously been linked to MSI[129,130,134]. Other *Fusobacterium* species were also enriched in tumors with high mutation burden, including *F. vincentii* (q = 2.7e-8), *F. pseudoperiodonticum* (q = 1.5e-5), *F. polymorphum* (q = 9.5e-4), and *F. animalis* (q = 8.6e-12), a subclade of which has been shown to exhibit tropism for colorectal tumors[135]. Species not previously tied to hypermutated subtypes were also enriched, including *Solobacterium moorei* (q = 1.4e-8), *Selenomonas sputigena* (q = 2.7e-8), *Parvimonas micra* (q = 4.0e-8), *Hungatella* hathewayi (q = 2.0e-6), and *Dialister pneumosintes* (q = 5.4e-5). Notably, many species enriched in hypermutated tumors were oral-typical and/or periodontal disease-associated, suggesting a relationship between oral-typical microbiota and mutation burden in colorectal cancer.

Conversely, tumors with low mutation burden appeared to be enriched for gut commensals with reportedly anti-inflammatory properties[136–138], including *Phocaeicola vulgatus* (q = 7.5e-6), *Faecalibacterium prausnitzii* (q = 1.7e-6), and *Bacteroides uniformis* (q = 4.9e-6). We additionally

observed enrichment of *Escherichia coli* (*q* = 3.3e-4), *Veillonella atypica* (*q* = 2.1e-4), and *Veillonella parvula* (*q* = 1.4e-3) in tumors with low mutation burden.

Meanwhile, another group of species were mutation burden-independent and instead varied by anatomic site. For example, *Clostridium perfringens* (*q* = 5.8e-29) and *Roseburia intestinalis* (*q* = 1.1e-16) were significantly associated with proximal colorectal tumors, while *Porphyromonas asaccharolytica* (*q* = 1.4e-15)*, Ruthenibacterium lactatiformans* (*q* = 3.4e-16)*,* and *Akkermansia muciniphila* (*q* = 1.2e-13) were associated with distal tumors.

Some species varied both by mutation burden and by anatomic site, suggesting an interaction between these features to influence microbial composition. *S. sputigena* and *F. animalis* were more abundant in hypermutated tumors of the proximal colon, while *P.* micra was proportionally greater in hypermutated tumors of the distal colon **(Figure 5I)**. Conversely, *V. parvula* and *V. atypica* were enriched in tumors in the proximal colon with low mutation burden, while *E. coli* was more associated with distal tumors with low mutation burden **(Figure 5J)**.

Several species associated with mutation burden additionally appeared to accumulate with advancing tumor stage **(Figure 5K)**. While *F. animalis* appeared to be enriched in hypermutated cases as early as stage I, others like *S. sputigena, P. micra, P. intermedia, and S. moorei* were more enriched at later stages, potentially indicating “early” and “late” colonizers. The relative increase in these species by stage was accompanied by a corresponding decrease in commensals such as *P. vulgatus, F. prausnitzii,* and *B. uniformis* **(Figure 5L)**.

Notably, many of the species associated with tumor mutation burden, anatomic location, or cancer stage had overlap with reproducible fecal biomarkers of colorectal cancer recently reported by Piccinno et al.[127]. For example, Piccinno et al.’s observation that fecal detection of *V. parvula* and *V. atypica* predicted the presence of proximal colorectal tumors mirrors our observation that these species are enriched in proximal tumors themselves. This result supports the idea that these species may be contributing to fecal signatures by shedding from the tumor into stool, linking tumor microbiome composition to fecal biomarker candidates.

Akkermansia muciniphila is depleted in early-onset tumors of the distal colon

Early-onset colorectal cancer is a rising public health concern, with incidence in individuals under 50 years of age doubling in many countries since the 1990s[139–141]. A recent study reported enrichment of SBS88 — a mutational signature caused by colibactin, a genotoxin produced by pks+ *E. coli*[49] — in early-onset colorectal cancer[142], raising the possibility that gut microbiota may

be predictive of colorectal cancer onset in younger adults. Noting that patient age was a predictor of microbial composition in colorectal tumors **(Figure 5A)**, we sought to examine microbial correlates with age of onset.

We compared early- (<50y) and average-onset (≥50) colorectal cancers, stratified by anatomic site and restricting to microsatellite-stable (MSS) cases. Clinically, early-onset colorectal cancer is typically MSS and occurs most frequently in the distal colon[143]. In proximal colorectal cancers **(Figure S5K)**, we observed variation of *Streptococcus sanguinis* (p = 8.33e-3), *Thomasclavelia ramosa* (p = 7.01e-3), and *Vescimonas fastidiosa* (p = 1.08e-2) by age of onset. These species are generally considered commensals, but *S. sanguinis* is a common cause of endocarditis[144].

We next compared tumor microbial composition between early- and average-onset cases arising in the distal colon, where early-onset cancers are more common **(Figure S5L)**. *Akkermansia muciniphila* was significantly depleted in early-onset cases (FC = 12.5; p = 3.4e-3) **(Figure S5M)**, as was *Hungatella hathewayi* to a lesser extent (FC = 2.4, p = 3.3e-3). *H. hathewayi* is a widely carried gut commensal noted for its ability to efficiently degrade glycosaminoglycans[149], while *A. muciniphila* is a mucin-degrading commensal that has been linked to intestinal barrier function, metabolic health, and reduced inflammation[150], and has shown consistent enrichment in the gut microbiomes of centenarians[151].

To provide independent validation of this association, we examined microbial profiles from TCMA/TCGA **(Figure S5N)**. Consistent with 100kGP, *A. muciniphila* showed no difference in COAD tumors. However, in READ tumors, *A. muciniphila* was similarly depleted in early-onset cases (FC = 42.8), mirroring our observation in distal CRCs from the 100kGP cohort, though this did not reach statistical significance (p = 0.158), likely reflecting the low number of early-onset rectal cancers in the TCGA cohort (n = 10).

Together, these data indicate that *A. muciniphila* may be consistently depleted in early-onset CRCs, raising the possibility that loss of mucin-associated commensals might contribute to and/or reflect patterns of barrier dysfunction and broader mucosal perturbations in younger patients.

Limitations of the study

Although we observed taxa with a range of cell architectures, our microbial abundance estimates may be underestimates due to incomplete lysis of microbial cell walls during nucleic acid extraction. Furthermore, our analysis does not disambiguate “mixed-evidence” species, whose reads may reflect a mixture of contamination and true signal. This may lead to additional underestimates of microbial signal in skin cancer, for example, as some of the contaminant taxa we detected were common skin commensals.

Finally, while the tumor genomes analyzed in this study were sequenced to high coverage (~100X), detection of microbiota in these samples is nevertheless bounded by assay-specific limits of detection. Normalized to sequencing depth, we determined our species-level detection threshold to be approximately 0.1 reads per million (RPM). For bacteria, this corresponds to approximately one microbial cell per ~7,000 tumor cells or per ~50 ng of tumor DNA, which is comparable to prior thresholds used for estimating tumor microbiome abundance using quantitative amplification of bacterial 16S rRNA[39]. Nevertheless, the key conclusions of this study will need to be validated using orthogonal assays to determine the extent to which these patterns generalize beyond bulk tumor sequencing data.

Discussion

This study establishes the biodiversity and biogeography of the cancer microbiome by addressing persistent methodological challenges in tumor microbiome profiling. Robust, quantitative characterization of tumor-associated microbial signals has historically been elusive due to incomplete host subtraction, batch effects, and widespread contamination that can mask real tissue-associated microbial signal in low-biomass samples.

To address these challenges, we developed and benchmarked PathSeq-T2T, which uses the complete human reference genome for host filtering and leverages multiple metagenomic profilers for quantitative estimation and cross-validation of microbial signals. Experimental benchmarks using human-microbe DNA mixtures revealed empirical limits of detection and identified contamination patterns, enabling us to better interpret microbial signals in large-scale tumor sequencing datasets. We subsequently applied this pipeline to 16,369 high-depth tumor whole genomes from the UK 100,000 Genomes Project (100kGP), which enabled the largest and most comprehensive pan-cancer analysis of tumor-associated microbial communities to date.

Although contamination was present in this dataset, we were able to identify and remove microbial contaminants using the principle of equiprevalence, which posits that contaminants will generally be present at similar rates across distinct sample types. While this approach does not replace gold-standard negative controls (which are seldom available for large-scale sequencing initiatives), it has previously proven effective in decontaminating other large-scale sequencing datasets[57].

Our analysis of the 100kGP cohort affirmed that most solid cancer types lack a tumor-associated microbiome that is detectable in bulk sequencing data, consistent with and extending other recent studies of tumor genomes from 100kGP[61] and TCGA[57,62]. A clear exception was observed in orodigestive cancers — including from oropharyngeal, esophageal, gastric, and colorectal sites — each of which harbored polymicrobial communities that displayed site-specific abundance, biogeography, and biodiversity consistent with microbial surveys of healthy individuals[1–11]. These findings support a model in which microbial colonization of tumors occurs predominantly at mucosal and epithelial barrier surfaces that are already permissive to commensal or pathogenic colonization.

The conclusions of these collective studies differ from earlier reports of distinct microbial communities present across all[55,56,60] or many[39] cancer types. Such discrepancies may in some cases be the result of methodological challenges which have been discussed extensively elsewhere[54,56,58,61,62], but in other cases may be due to differences in analytic approach (e.g. bulk PCR-free sequencing versus PCR-amplified or spatially-resolved assays), thresholds of detection, and/or strategies for handling contamination (See *Limitations of the study*). These debates underscore the need for standardized contamination control and transparent reporting[82,152].

Notwithstanding, our analysis revealed multi-kingdom microbial communities in orodigestive cancers, including fungi, archaea, viruses, and in rare cases, protozoan parasites. These results extend prior knowledge of microbial involvement in cancer beyond bacteria and viruses, reinforce emerging evidence for fungal presence in orodigestive cancers[59,60], and further demonstrate that tumor-associated microbial communities can also accommodate archaea and protozoa. Future work will be needed to determine what role, if any, species from these newly implicated domains play in cancer development and progression.

Extending prior observations of microbial enrichment in hypermutated genomic subtypes [129,130,134], we additionally found that tumor mutation burden is a predictor of microbial abundance, diversity, and composition in orodigestive tumors. Although bacteria (e.g. *pks+ E. coli*) have been linked to tumor mutational signatures[47–50], elevated TMB is unlikely to be the result of increased microbial colonization, as most hypermutated tumors arise in the setting of genome alterations in well-defined molecular pathways, notably DNA repair and DNA polymerase mutations. Instead, the association of high microbial abundance with tumor mutation burden may be mediated by tumor-immune interactions associated with increased neoantigen load. These include altered inflammatory pathways and expression of immune checkpoints[153–157], leading to local immunosuppression or exhaustion that creates a permissive niche for microbial expansion. In an alternative but not mutually exclusive scenario, tumor-associated microbiota may themselves contribute to suppression of anti-tumor immunity, as has been reported for *Fusobacterium* interactions with TIGIT[158], providing benefit to neoantigen-rich tumors. Additional work will be required to further validate the microbial abundance and mutation burden correlation and discern models for its selective benefit for tumors and/or microbes.

We also observed depletion of *Akkermansia muciniphila* in early-onset colorectal cancers of the distal colon. This mucosa-associated species has attracted interest as a potential probiotic agent, due to reported links to mucosal barrier function[150], longevity[151], and response to immunotherapy[29,33]. Accordingly, *A. muciniphila* warrants further study in the context of colorectal cancer and age of onset, including the role of the protective mucin layer in preventing exposure to luminal genotoxins. Confirmation in larger cohorts of younger colorectal cancer patients is necessary to assess the potential relevance of *A. muciniphila* to the increasing rate of colorectal cancer in young adults.

In conclusion, our analysis helps to resolve persistent uncertainties stemming from conflicting reports on the scope and distribution of microbial colonization across cancer sites, while identifying phenotypic and genotypic determinants of tumor microbiome community structure. We hope that the methodology, dataset, and insights presented here provide a foundation for future functional and mechanistic investigations of tumor-associated microbes and help nominate cancers most likely to benefit from prospective biomarker discovery and microbiome-focused therapeutic development.

Resource availability

Lead contact

Further requests for data should be directed to the lead contact, Dr. Matthew Meyerson (matthew_meyerson@dfci.harvard.edu).

Materials availability

This study did not generate new, unique reagents.

Data and code availability

The 100kGP microbiome profiles generated by PathSeq-T2T are accessible from within the Genomics England research environment, which requires an application to access, found here: https://www.genomicsengland.co.uk/join-us. Microbial profiles and sample metadata needed to reproduce the analyses described in this study can be found at the following path: /re_gecip/shared_allGeCIPs/adohlman/gel_manuscript. Due to data privacy considerations, Genomics England does not allow the export of sample- and patient-level data tables from the research environment.

The code used to perform the analyses and generate the figures associated with this manuscript are available in Jupyter Notebook format, which are accessible from within the Genomics England research environment. These can be found in the following directory: /re_gecip/shared_allGeCIPs/adohlman/gel_manuscript/analysis. The PathSeq-T2T pipeline is available on GitHub (https://github.com/abdohlman/pathseq-t2t/).

Acknowledgments

This work was supported by the Damon Runyon Cancer Research Foundation (grant no. DRG-2504-23 to A.B.D.); Cancer Grand Challenges OPTIMISTICC, funded by Cancer Research UK (grant no. A27140 to A.B.D., H.W., K.J., G.S., P.N., P.Q., C.H., and M.M.); the NIHR Leeds Biomedical Research Centre (grant no. NIHR203331 to H.W. and P.Q.); and Yorkshire Cancer Research (grant no. L386 to P.Q.). P.Q. is an NIHR Senior Investigator. N.S. is supported by the European Union's Horizon 2020 programme ONCOBIOME (grant no. 825410); the European Research Council (grant nos. ERC-StG MetaPG-716575 and ERC-CoG microTOUCH-101045015); the U.S. National Cancer Institute (grant no. 1U01 CA230551); the Premio Internazionale Lombardia e Ricerca 2019; and MIUR PRIN 2017 (grant no. 2017J3E2W2). N.S.

and C.H. are additionally supported through the Cancer Grand Challenges PROSPECT team (Cancer Research UK grant nos. CGCATF-2023/100036 and CGCATF-2023/100041; National Cancer Institute grant nos. OT2CA297680 and 1OT2CA297205-01; the Bowelbabe Fund for Cancer Research UK; and Institut National du Cancer). C.H. is further supported by Prescient Metabiomics. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.

This research was made possible through access to data in the National Genomic Research Library, which is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The National Genomic Research Library holds data provided by patients and collected by the NHS as part of their care. The National Genomic Research Library is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. We also thank Carrie Cibulskis and Matthew Brown for their contributions, as well as all of the research participants and their families for making this work possible.

Author contributions

A.B.D. conceived of the study, performed the formal data analysis, created the figures, wrote and edited the manuscript. R.M. performed metagenome assemblies and genome analyses and contributed to data interpretation. H.W. performed preliminary analyses of colorectal and gastric cancer microbiomes in the 100kGP cohort and helped curate metadata for these cancer types. K.J. performed the *in vitro* microbe-host mixture experiments. A.S. generated the *in silico* microbe-host mixture data. I.L. helped to reanalyze TCGA data. G.P. provided and advised on oral and fecal biomarker annotations used in the study. P.N. and G.S. performed external validation analyses and assisted in interpretation of the results. A-R.Y. provided analysis and interpretation of unclassified sequences. P.Q. advised on preliminary analyses of colorectal and gastric cancer microbiomes. C.H. and N.S. advised on the analytical approach and data interpretation. M.M. conceived of the study, supervised the analyses, reviewed and edited the manuscript.

Declaration of interests

A.B.D. is an inventor on US patent application no. PCT/US2023/029634, submitted by Cornell University and Duke University related to the use of fungi for classifying and treating gastrointestinal and lung tumors. M.M. receives research support from Bayer, holds equity in and consults for Bayer**,** Delve Bio, Isabl, and Karyoverse, and is an inventor on patents licensed to

Bayer and LabCorp, all outside of the current work. M.M. is also an inventor on an issued patent and several patent applications regarding *Fusobacterium* in colorectal cancer, which are not licensed. N.S. is a founder and shareholder of PreBiomics Srl and is on the scientific advisory board of ZOE Ltd and received consultancy fees from them. C.H. is a member of the Seres Therapeutics and Empress Therapeutics scientific advisory boards.

Declaration of generative AI and AI-assisted technologies

During the preparation of this work, authors used ChatGPT to assist in writing code used for performing data analyses and visualizations. After the manuscript was written, ChatGPT was used to assist in editing, such as checking for spelling and grammatical errors. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Supplemental information

Table S1. Pre-defined species-level read counts for *in silico* microbe-human dilution series.

Table S2. *In silico* read counts after each PathSeq-T2T filtering step.

Table S3. *In vitro* read counts after each PathSeq-T2T filtering step.

Table S4. Summary of samples in the Genomics England cohort.

Table S5. Species equiprevalence scores and contamination classifications.

Table S6. Summary of genome assembly results from orodigestive cancer types.

Table S7. Site-specific species associations in oropharyngeal, gastroesophageal, and colorectal cancers.

Table S8. Joint analysis of CRC tumor anatomic site and TMB subtype.

Figure titles and legends

Figure 1. A host subtraction and classification pipeline for identifying microorganisms in low-biomass human tissues

(A) Retention rate of *in silico* human sequences in the human-only condition after each host filter, expressed as reads per million primary reads.

(B) Retention rate of *in silico* microbial reads after each host filter, across microbe-host mixture conditions (*n* = 9) ranging from a host-microbe ratio of 1:1 (*red*) to 1:10^8 (*blue*).

(C) Total read retention rate after each successive PathSeq-T2T filtering step for each condition in the *in silico* dilution series. Horizontal gray lines indicate the starting microbial proportion in each condition.

(D) Relative abundance of human (*blue*), on-target microbial (*green*), off-target microbial (*orange*), and not uniquely classified (*gray*) reads after each host filtering step across *in silico* dilution conditions. Reads are "on-target" if assigned to species/genera/families present in microbial community standard, and "off-target" otherwise.

(E) Total read retention rate after each successive PathSeq-T2T filtering step for each condition in the *in vitro* dilution series (*n* = 3 replicates each).

(F) Relative abundance of human (*blue*), on-target microbial (*green*), off-target microbial (orange), and not uniquely classified (*gray*) reads after each host filtering step across *in vitro* dilution conditions.

(G) Abundance in reads per million (RPM) of on-target microbial sequences in each condition after PathSeq-T2T filtering of *in vitro* microbe-host DNA mixtures.

(H) Abundance of off-target (putative contaminants) *Cutibacterium acnes, Staphylococcus epidermidis,* and *Malassezia restricta* across *in vitro* dilution conditions.

(I) Relative abundance of species detected in negative controls (*n* = 3) with no human or microbial DNA added, processed in parallel with *in vitro* conditions dilutions.

(J) Percentage of off-target (contaminant) reads across successive dilution conditions.

Microbial counts are from Kraken2 unless stated otherwise.

Figure S1. A host subtraction pipeline for identifying microorganisms in low-biomass human tissues (Related to Figure 1)

(A) Schematic of the PathSeq-T2T pipeline used to study microbial signals in low-biomass tissues. Tumor whole-genome sequencing (WGS; delivered as hg38-aligned BAMs) were (1) filtered to select unmapped reads; (2) processed using quality, complexity, and host k-mer filtering; and then (3) filtered using T2T-CHM13. Last (4) reads passing these filtering steps are classified using Kraken2[77], MetaPhlAn4[78], and Sylph[79]. In parallel, raw hg38-unmapped reads were assembled for microbial genome reconstruction using MEGAHIT[159] (See **Table S6**).

(B) Retention rate of *in silico* human sequences in the across microbe-host mixture conditions (1:1 to $1:10^8$) after each host filter.

(C) Breakdown of *in silico* microbial reads removed at each PathSeq-T2T filtering step (microbe-only condition), expressed as a percentage of total removed microbial reads.

(D) Misclassification rate of reads removed or retained by PathSeq-T2T, expressed as misclassifications per million microbial reads.

(E) Total read retention rate after each successive PathSeq-T2T filtering step for colorectal cancer cell lines annotated as microsatellite-instable (MSI; *n* = 8) or microsatellite-stable (MSS; *n* = 12).

(F) Total read retention rate after each successive PathSeq-T2T filtering step for colorectal cell lines prepared using PCR (*n* = 10) or PCR-free (*n* = 10) methods.

(G) Relative abundance of microbial genera identified in colorectal cancer cell lines by PathSeq-T2T, representing likely contaminants.

** p < 0.05, ** p < 0.01, *** p < 0.001, **** p < 0.0001 by Wilcoxon's signed-rank test. Microbial counts are from Kraken2 unless stated otherwise.*

Figure 2. Identification and removal of contamination from the 100kGP cohort

(A) Median microbial reads per million (RPM) retained in saliva (blue; *n* = 490) and blood (red; *n* = 14,127) after each PathSeq-T2T filtering step. Error bars indicate interquartile range.

(B) Median RPM retained in colorectal cancer (CRC; brown; *n* = 2,482) and glioblastoma (GBM; green; *n* = 295) tumors after each PathSeq-T2T filtering step. Error bars indicate interquartile range.

(C) Relative abundance of genera detected in saliva (top) and blood (bottom). Samples are ordered by germline sequence delivery date.

(D) Relative abundance of genera detected in CRC (top) and GBM (bottom). Samples are ordered by germline sequence delivery date.

(E) Prevalence (i.e. percentage of samples) of species detected in blood samples (x-axis) and saliva samples (y-axis). Saliva harbors saliva-specific species (blue), while equiprevalent species (red) are found at similar frequencies in saliva and blood.

(F) Prevalence of species detected in GBM samples (x-axis) and CRC samples (y-axis). CRC samples harbor CRC-specific species (blue), while equiprevalent species (red) are found at similar frequencies in CRC and GBM.

(G) Bar plot showing ten most blood-prevalent species, among those equiprevalent in saliva/blood.

(H) Bar plot showing ten most GBM-prevalent species, among those equiprevalent in CRC/GBM.

(I) Histogram showing pan-cancer equiprevalence (PCE) scores of species detected in 100kGP tumor samples. Species equiprevalent in blood/saliva or CRC/GBM are labeled in red, other species in blue. Species with PCE scores greater than 0.7 (gray shaded area) were classified as contaminants. Distribution includes species with >5% prevalence in at least one cancer type.

(J) Prevalence of selected species of across cancer types, labeled with PCE scores. Known cancer-associated species (*H. pylori, F. nucleatum, HPV*) had low PCE scores, while contaminants (*C. acnes, S. epidermidis, M. restricta*) had high PCE scores.

Microbial counts are from Kraken2 unless stated otherwise.

Figure S2. Identification and removal of contamination from the 100kGP cohort (Related to Figure 2)

(A) Prevalence (i.e. percentage of samples) of species detected in blood (x-axis) and saliva samples (y-axis) using MetaPhlAn4[78]. Saliva samples harbor saliva-specific species (blue), while equiprevalent species (red) are found at equal frequency in saliva and blood.

(B) Prevalence of species detected in blood (x-axis) and saliva samples (y-axis) using Sylph[79].

(C) Prevalence of species detected in GBM (*x-axis*) and CRC samples (*y-axis*) using MetaPhlAn4[78] (*left*) and Sylph[79] (*right*). CRC samples harbor CRC-specific species (*blue*), while equiprevalent species (*red*) are found at equal frequency in CRC/GBM.

(D) Prevalence of species detected in GBM (x-axis) and CRC samples (y-axis) using Sylph[79].

(E) Venn diagram depicting the degree of overlap between species that were equiprevalent in saliva and blood and equiprevalent in CRC and GBM.

(F) Bar plot showing the ten most saliva-prevalent species.

(G) Bar plot showing the ten most CRC-prevalent species.

(H) Histogram of PCE scores of microbial species detected in 100kGP tumor samples by MetaPhlAn4[78] (*left*) and Sylph[79] (*right*). Species that were equiprevalent in blood and saliva or in CRC and GBM (*red*) typically had high PCE scores, while other species (*blue*) did not. Species with PCE > 0.7 (*gray shaded area*) were classified as contaminants. Distribution includes species detected at >5% prevalence in at least one cancer type.

(I) Histogram of PCE scores of species detected in 100kGP tumor samples by Sylph[79].

(J) Prevalence of species with the highest PCE scores in the 100kGP cohort, across cancer types. Most of these species are environmental and do not typically colonize humans.

(K) Within-species median average nucleotide identity (ANI) of microbial genomes obtained through metagenomic assembly, sorted by median ANI. Genomes from species with high PCE scores (*red*) have strain-level sequence similarity (>99.99%), suggesting a common contaminant source.

(L) Volcano plot showing enrichment of equiprevalent species (*red*) in germline samples that underwent PCR-based ($log_2OR > 0$) versus PCR-free preparation ($log_2OR < 0$). Statistical significance was determined using Fisher's exact test (*y-axis*),

(M) Volcano plot showing enrichment of equiprevalent species (*red*) in tumor samples that underwent PCR-based ($log_2OR > 0$) versus PCR-free preparation ($log_2OR < 0$). Statistical significance was determined using Fisher's exact test (*y-axis*),

(N) Volcano plot showing enrichment of equiprevalent species (*red*) in tumor samples that were stored as FFPE ($log_2OR > 0$) versus fresh-frozen ($log_2OR < 0$).

Microbial counts are from Kraken2 unless stated otherwise.

Figure 3. Tumors of the orodigestive tract harbor a microbiome but most other cancer types do not

(A) Cumulative distribution of Kraken2[77] sample-level microbial abundance by cancer site after decontamination. Few non-orodigestive cancers exceeded background (*shaded gray area*; RPM < 0.1).

(B) Cumulative distribution of MetaPhlAn4[78] sample-level microbial abundance by cancer site after decontamination.

(C) Cumulative distribution of Sylph[79] sample-level microbial abundance by cancer site after decontamination.

(D) Species prevalence in each cancer type (*y-axis*) compared to their prevalence in brain (*x-axis*) after decontamination. Few non-orodigestive cancers exhibit cancer-site specific microbial signals.

(E) Letter-value (boxen) plots showing number of species detected per sample (RPM > 0.1) after decontamination. Few non-orodigestive cancers consistently harbor more than one species.

(F) Average family-level relative abundance of microbe-high samples (>10 RPM) from each cancer site after decontamination. Only a small fraction of samples from non-orodigestive cancers met this criterion and many were dominated by common *Herpesviridae*.

Microbial counts are from Kraken2 unless stated otherwise.

Figure S3. Tumors of the orodigestive tract harbor a microbiome but most other cancer types do not (Related to Figure 3)

(A) Cumulative distribution of Kraken2[77] sample-level microbial abundance by cancer site, prior to decontamination. Nearly all samples and cancer types exceed the background threshold (shaded gray area, RPM < 0.1).

(B) Cumulative distribution of MetaPhlAn4[78] sample-level microbial abundance by cancer site, prior to decontamination.

(C) Cumulative distribution of Sylph[78] sample-level microbial abundance by cancer site, prior to decontamination.

(D) Cumulative distribution of Kraken2[77] sample-level contamination abundance by cancer site. Cancer types mostly harbor similar rates of contamination.

(E) Cumulative distribution of MetaPhlAn4[78] sample-level contamination abundance by cancer site.

(F) Cumulative distribution of Sylph[78] sample-level contamination abundance by cancer site.

(G) Spearman correlation (Rho) between microbial abundance estimates from Kraken2, MetaPhlAn4, and Sylph across cancer types.

(H) Prevalence of species in each cancer type (*y-axis*) compared to their prevalence in brain (*x-axis*), prior to decontamination. Species classified as contaminants are shown in red.

(I) Prevalence of the *Papillomaviridae* species HPV-16 and HPV-18 across cancer sites.

(J) Prevalence of the *Polyomaviridae* species MCPyV across cancer sites.

(K) Prevalence of the *Herpesviridae* species HSV-1, EBV, CMV, and HHV-7 across cancer sites.

(L) Prevalence of the *Herpesviridae* species HHV-6A and HHV-6B across cancer sites.

** p < 0.05, ** p < 0.01, *** p < 0.001, **** p < 0.0001 by Spearman's rank correlation test. Microbial counts are from Kraken2 unless stated otherwise.*

Figure 4. Biogeography and biodiversity of oropharyngeal, gastroesophageal, and colorectal cancers (Related to Figure 4)

(A) Microbial abundance (RPM) by across anatomical sites in oropharyngeal (*blue*), gastroesophageal (*orange*), and colorectal (*green*) cancers.

(B) Principal coordinates analysis (PCoA) of samples from oropharyngeal, gastroesophageal, and colorectal cancers.

(C) Radial plot showing species-level barycentric centroids calculated from cancer site-specific abundances ($\log_{10}$ RPM). Abundances in oropharyngeal, gastroesophageal, and colorectal cancers were used to define a triangle in site-space, the centroid of which was mapped to polar coordinates (*angle*: site bias; *radius*: magnitude of enrichment).

(D) Relative abundance of microbial genera in oropharyngeal (*top left*), gastroesophageal (*top right*), and colorectal cancers (*bottom*).

(E) Prevalence of bacterial species (*P. gingivalis, H. pylori, F. animalis, B. fragilis, E. coli, S. anginosus*) across anatomical sites of the orodigestive tract.

(F) Prevalence of fungal species (*C. albicans, N. glabratus, S. cerevisiae*) across anatomical sites.

(G) Prevalence of the archaeon *M. smithii* across anatomical sites of the orodigestive tract.

(H) Prevalence of the protozoan genus *Trichomonas* across anatomical sites.

(I) Prevalence of viruses (HPV-16, HHV-7, EBV, HHV-6A, and HHV-6B) across anatomical sites.

Microbial counts are from Kraken2 unless stated otherwise.

Figure S4. Biogeography and biodiversity of oropharyngeal, gastroesophageal, and colorectal cancers (Related to Figure 4)

(A) Diversity of microbiota (Shannon index) across anatomical sites of the orodigestive tract.

(B) Estimated bacterial cells per thousand human cells across anatomical sites.

(C) Venn diagram showing numbers of overlapping species detected in oropharyngeal, gastroesophageal, and colorectal cancers, including species detected in >1% of samples from at least one site.

(D) Prevalence of bacterial species (*T. socransii, C. gingivalis, L. fermentum, L. gasseri, P. histicola, P. micra, S. moorei, P. intermedia, V. parvula, S. sputigena,* and *A. muciniphila*) across anatomical sites of the orodigestive tract.

(E) Prevalence of the archaea *M. sp. TLL-48-HuF1* and *Candidatus Methanomassiliicoccus intestinalis* across anatomical sites.

Microbial counts are from Kraken2 unless stated otherwise.

Figure 5. The tumor microbiome varies by tumor site, mutation burden, and age of onset.

(A) Bar plot showing the percentage of variance explained identifying phenotypic and genomic features of colorectal cancers associated with microbial composition (*p*-value by PERMANOVA).

(B) Microbial abundance in reads per million (RPM) in microsatellite-stable (MSS; *gray*) microsatellite-instable (MSI; *red*), polymerase-mutated (POL; *blue*) subtypes.

(C) Microbial abundance in MSS (*gray*) and MSI (*red*) colorectal (COAD) and gastric (STAD) cancers from The Cancer Microbiome Atlas (TCMA)[57].

(D) Microbial diversity (Shannon index) in MSS (*gray*), MSI (*red*), and POL (*blue*) colorectal cancer subtypes.

(E) Pearson correlation between tumor mutation burden (TMB; *x-axis*) and microbial abundance (RPM; *y-axis*) in MSS (*gray*), MSI (*red*), and POL (*blue*) subtypes. Microbial load and TMB are correlated in MSS and MSI but not POL.

(F) Correlation between TMB, after regressing out the effect of genomic subtype (*x-axis*) and microbial abundance, after regressing out the effect of anatomic site (*y-axis*).

(G) Microbial abundance by tumor subtype (MSS/MSI/POL), after regressing out the influence of TMB on microbial abundance. Tumor subtype is no longer predictive of microbial load.

(H) Scatter plot showing joint associations between relative abundance of CRC-associated species and tumor site (proximal vs. distal) and TMB status (TMB high vs. TMB low). Axes show the composite Wilcoxon test statistic, calculated for tumor site, blocked by subtype (x-axis) and TMB status, blocked by tumor site (*y-axis*). Species are colored according to their association with tumor site (*green*), TMB status (*red*), or both (*orange*), based on a combined *p*-values (Fisher's method). Species associated with TMB status are also reproducible fecal biomarkers for CRC[127] (*triangles*) and frequently include oral-typical microbes (*bold*).

(I) Relative abundance of TMB-high-associated species (*S. sputigena, F. animalis,* and *P. micra*) by tumor site and genomic subtype.

(J) Relative abundance of TMB-low-associated species (*V. parvula, P. vulgatus,* and *E. coli*) by tumor site and genomic subtype.

(K) Relative abundance of TMB-high-associated species (*F. animalis, S. sputigena, P. micra, P. intermedia,* and *S. moorei*) by CRC tumor stage and genomic subtype. Data are presented as median ± interquartile range.

(L) Relative abundance of TMB-low-associated species (*P. vulgatus, F. prausnitzii, and B. uniformis*) by CRC tumor stage and genomic subtype.

** p < 0.05, ** p < 0.01, *** p < 0.001, **** p < 0.0001 by Wilcoxon's signed-rank test unless specified otherwise. Microbial counts are from Kraken2 unless stated otherwise.*

Figure S5. The tumor microbiome varies by tumor site, mutation burden, and age of onset (Related to Figure 5)

(A) Microbial abundance (RPM) in MSS (gray), MSI (red), and POL-mutant (blue) colorectal cancer subtypes, stratified by anatomic site.

(B) Bacterial abundance (RPM) in MSS, MSI, and POL-mutant (blue) colorectal cancer subtypes.

(C) Fungal abundance (RPM) in MSS, MSI, and POL-mutant (blue) colorectal cancer subtypes.

(D) Viral abundance (RPM) in MSS, MSI, and POL CRC subtypes.

(E) Archaea abundance (RPM) in MSS, MSI, and POL CRC subtypes.

(F) Microbial diversity (Shannon index) in MSS, MSI, and POL-mutant (blue) colorectal cancer subtypes, stratified by anatomic site.

(G) Pearson correlation between tumor mutation burden (TMB; *x-axis*) and microbial abundance (RPM; *y-axis*) in oropharyngeal cancers from 100kGP.

(H) Pearson correlation between tumor mutation burden (TMB; *x-axis*) and microbial abundance (RPM; *y-axis*) in oropharyngeal cancers from TCMA.

(I) Pearson correlation between tumor mutation burden (TMB; *x-axis*) and microbial abundance (RPM; *y-axis*) in gastroesophageal cancers from 100kGP.

(J) Pearson correlation between tumor mutation burden (TMB; *x-axis*) and microbial abundance (RPM; *y-axis*) in gastroesophageal cancers from TCMA.

(K) Volcano plot showing differences in microbial composition between average-onset (AO; *green*) and early-onset (EO; *purple*) in proximal colorectal tumors (MSS only).

(L) Volcano plot showing differences in microbial composition between average-onset (AO; *green*) and early-onset (EO; *purple*) in distal colorectal tumors (MSS only).

(M) Relative abundance of *Akkermansia muciniphila* in EO versus AO colorectal cancer in 100kGP, stratified by tumor site.

(N) Relative abundance of *Akkermansia muciniphila* in EO versus AO colorectal cancer in TCMA[57], stratified by tumor site.

** p < 0.05, ** p < 0.01, *** p < 0.001, **** p < 0.0001 by Wilcoxon's signed-rank test unless specified otherwise. Microbial counts are from Kraken2 unless stated otherwise.*

STAR Methods

In silico microbe-human DNA mixtures

To model sequencing of microbe-human mixtures, we simulated 151bp paired-end reads from a defined microbiome community and the human reference, then mixed them at predefined ratios. The microbial composition and relative abundance of these mixtures was designed to mirror the ZymoBIOMICS Gut Microbiome Standard (Zymo Research, Cat. #D6331), which contains 17 species at varying proportions, spanning bacteria (e.g. *F. nucleatum, B. fragilis, E. coli, A. muciniphila*), fungi (*C. albicans and S. cerevisiae)*, and archaea (*M. smithii*). Microbial genomes were obtained from the vendor-provided reference bundle (https://s3.amazonaws.com/zymo-files/BioPool/D6331.refseq.zip), while human reads were simulated from the CHM13v2.0 assembly (RefSeq accession GCF_009914755.1).

Illumina sequencing output was simulated using InSilicoSeq[160] v2.0.1 (NovaSeq error model), then down-sampled using seqtk v1.2-r94 (https://github.com/lh3/seqtk) using a fixed random seed (100). For each mixture, simulated microbial and human reads were combined at fixed, 10-fold microbe:host ratios ranging from 1:1 to $1:10^8$, keeping total read counts constant at 300 million per sample **(Table S1)**, approximating ~30X human WGS depth. The resulting simulated sequencing outputs of microbe-human mixtures were aligned to GRCh38 (NCI Genomic Data Commons resource: https://api.gdc.cancer.gov/data/254f697d-310d-4d7d-a27b-27fbf767a834) using BWA-MEM with GDC parameters (`-T 0`) to generate GRCh38-aligned BAMs. These BAMs were then processed using PathSeq-T2T for host filtering and downstream microbial profiling.

In vitro microbe-human DNA mixtures

Microbial and human DNA were mixed *in vitro* using microbial DNA derived from the ZymoBIOMICS Gut Microbiome Standard (Zymo Research, Cat. #D6331) and human DNA derived from colorectal cancer cell line HCT116 (ATCC, Manassas, VA, USA), cultured in McCoy's 5A medium supplemented with 10% FBS and 1% penicillin-streptomycin. Microbial DNA was extracted using the ZymoBIOMICS DNA Miniprep Kit (Zymo Research, Cat. #D4300). Human DNA was extracted using the DNeasy Blood & Tissue Kit (QIAGEN, Cat. #69504). All DNA extractions were performed following the manufacturers' instructions. DNA concentrations were quantified using Qubit™ dsDNA Quantification Assay Kit (Thermo Fisher Scientific, Cat. #Q32851).

Purified microbial and human DNA was then combined to generate an eight-point 10-fold serial dilution series in triplicate (n = 24), spanning microbe:human DNA ratios ranging from 1:1 to $1:10^8$. To assess technical and biological contamination, we also included human-only (n = 3) and template-free extraction negative controls (n = 3). The human-only negative control contained purified HCT116 gDNA with no microbial input and was sequenced in parallel with the dilution series; this control allowed us to assess the degree of cross-contamination between conditions, either due to the inadvertent exchange of DNA between experimental conditions[81,82] or due to misassignment of barcodes during the sequencing process[82–84]. The no-template extraction control contained aliquots of purified, nuclease-free water which underwent human and microbial DNA extraction in parallel then mixed in equal proportions; this control allowed us to assess the introduction of contamination during the sample preparation and DNA extraction process. All extractions and mixtures were performed under sterile conditions. Mixtures and human-only

controls were prepared at a minimum final concentration of 15 ng/μL in a 100 μL volume, then sent for sequencing together with no-template extraction controls.

Library preparation and sequencing were performed by the Broad Institute's Genomics Platform. PCR-free whole-genome sequencing (WGS) libraries were prepared from the microbial-human mixtures and human-only gDNA and sequenced to ~30X human coverage. Template-free negative controls were prepared as non-tagmented libraries and analyzed using mid-output (1.5 Gb) metagenomic WGS. Samples undergoing 30X WGS were demultiplexed, aggregated, and aligned using Illumina DRAGEN and delivered as GRCh38-aligned CRAM files, then subjected to host-filtering by PathSeq-T2T. Sequencing output for the template-free negative controls was delivered as FASTQ files and screened for microbial sequences using Kraken2[77].

Analysis of sequenced colorectal cancer cell lines

Whole genome sequencing data from colorectal cancer cell lines were obtained from The Cancer Cell Line Encyclopedia (CCLE)[85] via the Terra workspace. We selected cell lines classified as microsattelite-instable (n = 8) or microsatellite stable (n = 12) by Cellosaurus[161] and annotated by CCLE as being from PCR-Free (*n* = 10) or non-PCR-free (*n* = 10) libraries. Sequencing data from these cell lines were analyzed with PathSeq-T2T using default parameters.

Host filtering and microbial classification with PathSeq-T2T

Whole-genome sequencing data in this study were host-filtered using PathSeq-T2T, a pipeline we designed that extends our previous host-filtering tool, PathSeq[66,67]. This pipeline is designed to process sequencing data pre-aligned to GRCh38, which is the standard delivery format for sequencing data. First, GRCh38-aligned data are pre-filtered using samtools[162], first removing all reads mapped in proper pairs to the reference (parameters: -f 3) to retain unmapped/discordant reads. Reads aligned using BWA-MEM are additionally subjected an alignment score filter (-e '[AS]>35'). Reads properly paired but aligning to viral decoy sequences (often included as alternative contigs in reference genomes to reduce false positives in variant calling) are retained and cleared of alignment tags using the "RevertSam" function from Picard[163].

After prefiltering, both unaligned and viral-mapped sequences are then subjected to quality, complexity, and additional host-filtering using GATK PathSeqFilterSpark (parameters: --min-clipped-read-length 60 --host-is-aligned true). This step masks and/or removes reads with low base quality (Phred < 15) and low complexity (entropy < 0.3), duplicates, as well as sequences containing one or more k-mers matching (*i*) the human reference GRCh38, (*ii*) sequences from

the highly variable human major histocompatibility complex from the Immuno Polymorphism Database, (*iii*) cloning vector sequences from NCBI UniVec, (*iv*) known human transcripts from Gencode (human v25), and (*v*) human breakpoint sequences from GenBank (KY503218, KY5808060). This results in a set of high-quality, high-complexity, host-filtered paired and singleton reads (the mates of some reads are removed at this step, due to the lower quality of some reverse reads).

Next, resulting paired and singleton reads are then aligned (independently for pairs and singletons) to the complete human reference genome T2T-CHM13[76] with BWA MEM using the Genomic Data Commons (GDC) default parameters (-T 0). Properly-paired or singleton reads aligning to T2T-CHM13 (i.e. BWA alignment score > 35) are classified as human and removed, leaving a final set of high-quality, putatively non-human reads.

Next, reads passing host-filtering and quality-control are processed using Kraken2[77], MetaPhlAn4[78], and Sylph, complementary microbial classification tools commonly used for metagenomic analyses. Kraken2[77] is a k-mer-based tool that tends to prioritize sensitivity over specificity, allowing detection of microbial signals even when read counts are very low. MetaPhlAn4[78] relies on microbial marker-genes for classification and prioritizes specificity over sensitivity, enabling high-confidence microbial calls provided that sufficient reads are present to cover marker genes. Sylph[79] is another k-mer based tool which measures microbial genome containment and is demonstrates high specificity.

In this study, Kraken2 classification was performed using the “PlusPF” k-mer database (release 2024-06-05; obtained from https://benlangmead.github.io/aws-indexes/k2). This reference contains k-mers recovered from complete and chromosome-level genome assemblies of bacteria, archaea, viruses, fungi, and protozoa as well as the human genome and UniVec vector sequences obtained from RefSeq. PathSeq-T2T uses default Kraken2 parameters, with the exception of the minimum k-mer fraction, which is increased to 15% (--confidence 0.15) from the default (0.0) to increase specificity, which was found to be optimal for accuracy by Wright et al.[164]. Because Kraken2 concatenates paired reads prior to classification, host-filtered paired and unpaired reads are classified separately; afterwards, counts are merged. Throughout this manuscript, given microbial abundance estimates are from Kraken2, unless specified otherwise.

MetaPhlAn4 classification in this study was performed using the pre-built CHOCOPhlAn SGB database (mpa_vJun23_CHOCOPhlAn_202403), which contains bacterial and archaeal

marker genes. In PathSeq-T2T and in this study, default parameters are used, except for the minimum read length (--read_min_len 60) and output format (-t rel_ab_w_read_stats), which are added to enable microbial read count estimates based on marker gene coverage. Because MetaPhlAn4 does not take read pairing into consideration, paired and unpaired reads are merged prior to classification.

Sylph classification was performed using prebuilt reference databases, including references for bacteria (gtdb-r226-c200-dbv1.syldb), fungi (fungi-refseq-2025-10-11-c200-dbv1.syldb), and viruses (uhgv_c200_dbv1.syldb). In PathSeq-T2T and tin this study, default parameters are used with the exception of "--estimate-read-counts" and "--estimate-unknown", which are used for downstream abundance estimation. Like Kraken2, Sylph handles paired and unpaired reads differently; therefore these reads were processed separately and counts were merged afterwards.

Finally, read counts from Kraken2, MetaPhlAn4, and Sylph are normalized to per-sample sequencing depth using primary read counts obtained from samtools' flagstat, performed on the original, unfiltered bam file. This value is used as a denominator to calculate the microbial reads per million primary reads (RPM). This normalization against the total primary reads (human and microbial) quantifies the microbial-to-human DNA ratio, providing an approximation of the absolute microbial abundance in the sequenced tumor portion. This value be used to calculate microbe-host cell ratios after normalizing to relative genome size.

Detection of microbial signals in 100,000 Genomes Project (100kGP) cohort

The 100kGP cohort includes tumor and normal sample pairs from cancer patients recruited to the 100,000 Genomes Project (100kGP v18 release) through 13 Genomic Medicine Centres in the United Kingdom. We analyzed the entire 100kGP cohort, consisting of 16,369 tumor samples and 15,249 matched germline sequencing experiments from a total of 15,237 participants. Written informed consent was obtained by Genomics England from all participants. Tumor samples were sequenced to ~100X coverage and consisted of primary (*n* = 15,041), metastatic (*n* = 963), and recurrent (*n* = 365) lesions **(Table S2)**. Germline samples were sequenced to ~30X coverage and consisted primarily of blood (*n* = 14,127) with additional saliva (*n* = 490) and other non-diseased tissues. Most tumor samples were fresh-frozen (*n* = 15,011) and sequenced from PCR-free libraries (*n* = 14,420), with the remaining samples predominantly formalin-fixed, paraffin-embedded (FFPE; *n* = 614) and sequenced using PCR-amplified libraries (*n* = 1,949), respectively. Sequencing data for both sample types were delivered as BAM files

aligned to GRCh38 using the DRAGEN pipeline, which we analyzed using PathSeq-T2T implemented in “Double Helix”, the high-performance computing (HPC) system within 100kGP Research Environment’s virtual desktop. After decontamination, downstream microbiome analyses were restricted to primary and recurrent fresh-frozen tumor samples from solid cancer sites with a minimum of 50 samples available (i.e. bladder, brain, breast, colorectal, esophageal, gastric, kidney, liver, lung, oropharyngeal, ovarian, pancreatic, prostate, skin, testicular, and uterus) prepared from PCR-free libraries (*n* = 10,698).

Decontamination of 100kGP tumor microbiomes

Decontamination was performed on host- and quality-filtered reads produced by PathSeq-T2T, using species-level abundance estimates from Kraken2, MetaPhlAn4, and Sylph. Our approach was based on previously reported observations[57,165] that reagent and sample-handling contaminants in cancer sequencing datasets tend to appear broadly and uniformly across distinct sample types (e.g. tissue and blood), whereas species that are truly present in tumor samples show sample-type-specific biological variation.

To decontaminate 100kGP tumor microbiomes, we computed a pan-cancer equiprevalence (PCE) score that quantifies species ubiquity across cancer types. This score was obtained by calculating the coefficient of variation (CV) of species prevalence across cancer types, rescaled to [0, 1] based on the maximum/minimum theoretical possible CV values a species could be given, such that PCE = 1 for species with perfectly equal distribution across cancer types and PCE = 0 for detection confined to a single cancer type. Prevalence within each cancer type (100kGP “study name”) was defined as the percentage of samples with RPM > 0.1. To reduce variance, PCE was only calculated using 29 clearly defined cancer types with ≥50 primary or recurrent tumors.

As expected, common environmental contaminants and species that were equiprevalent in blood and saliva or in CRC and GBM also clustered at high PCE **(Figure 2I and S2F)**. Based on this observation that nearly all of these species had PCE > 0.7, we selected 0.7 as a cutoff for identifying and removing contaminants from the dataset. The only species excepted from this threshold were viral taxa known to infect humans (HPV-16, HPV-18, MCPyV, HSV-1, EBV, CMV, HHV-6A, HHV-6B, HHV-7), many of which are widely carried in the population.

Assembly of microbial genomes

To further validate microbial signals and assess contaminant signals, we performed metagenome assembly on unmapped tumor sequencing reads from orodigestive cancer types, based on the observation that tumors arising at these sites were unique in harboring significant microbial signal. Tumor WGS reads not mapped to the human genome reference (GRCh38) were extracted from the WGS BAM files using samtools[162] (v.1.16.1) with parameter "-f 4", then converted to FASTQ with the samtools bam2fq function. The unmapped FASTQ files were used as input to trim_galore (v. 0.6.10) with parameters "--illumina --stringency 5 --length 75 --max_n 2 --trim-n -j 25". The trimmed reads were then assembled into contigs using MEGAHIT[159] (v. 1.2.9) with default parameters. The contigs were used to create a bowtie2[166] index to which the trimmed reads were realigned using bowtie2 with default parameters. The aligned contigs were then converted to BAM, sorted, and indexed using samtools (v. 1.16.1). Finally, the BAM files with reads mapped to contigs were used as input to MetaBAT2 for genome assembly, with the contig FASTA and the sorted BAM file of reads mapped to contigs as inputs. MetaBAT2 was run with the default minimum contig length of 2,500 bp. Quality control of the final assembled genomes was performed using CheckM2 (parameters: predict -x fa --force --lowmem). ANI analysis was performed using pyani (v0.3.0-alpha) with the -m ANIm method. SGBs with at least 10 unique genomes and completeness >50% based on CheckM2 results were included in the analysis.

Analysis of pan-cancer absolute and relative microbial abundance

To quantify microbial load, we summed species-level RPM values after removing contaminant species (PCE > 0.7). To allow efficient data visualization, sites were recoded from cancer types according to their anatomic site of origin (e.g. "Glioblastoma multiforme" and "Low grade glioma" were assigned to "Brain"). The threshold for "background" microbial signals was set to 0.1 RPM, which was determined based on our observation that sequenced human-only controls typically contained microbial read counts at approximately this level **(Figure 1G)**. This threshold was also used for determining species-level presence/absence.

The relative abundance analysis **(Figure 3E)** was calculated by aggregating decontaminated species-level read counts to the family level for "microbe-high" samples (RPM > 10). To visualize the average relative abundances at the level of cancer site, we averaged sample-specific relative abundances before re-normalizing to the sum of family-level averages.

Analysis of orodigestive tumor microbiomes

For analysis of orodigestive cancer types, samples were categorized as oropharyngeal (oral and sinonasal cancers; *n = 251*), gastroesophageal (esophageal and gastric cancers; *n* =

199), and colorectal (colon and rectal cancers; *n* = 2,422). Curated clinical and anatomic annotations of gastroesophageal and colorectal cancers were provided by Booth et al.[92] and Cornish et al.[73], respectively. Total microbial load was calculated by summing species-level RPM values after removing contaminant species (PCE > 0.7). Diversity was measured using Shannon's entropy.

The site-specific analysis of species-level associations across oropharyngeal, gastroesophageal, and colorectal cancers **(Figure 4D)** was performed by calculating a Kruskall-Wallis test of site-specific enrichments **(Table S7)** which adjusted for multiple hypothesis tests using the Benjamini-Hochberg correction. For radial visualization, the geometric mean abundance (average $\log_{10}$RPM) was calculated for each species and site, creating a 3-vector defining a triangular plane whose vertices represented each site and we calculated the barycentric centroid of that triangle. This centroid was then mapped into polar coordinates, with angle encoding site bias and radius encoding the magnitude of site selectivity.

Principal coordinates analysis and PERMANOVA

Principal coordinates analyses (PCoA) and PERMANOVA analyses were performed in R (v. 4.3.3) using the package "vegan" (v. 2.6.10) with genus-level read counts. Distance matrices for PCoA and PERMANOVA were calculated using the Bray-Curtis dissimilarity index. PERMANOVA was used to identify features predictive of microbial composition and was run using vegan's "adonis2" function using the "margin" setting to identify features that independently influence microbial composition.

Validation using tumor microbiome data from The Cancer Microbiome Atlas

To validate findings regarding 100kGP tumor genomes, we used PathSeq-filtered sequences from The Cancer Microbiome Atlas (TCMA)[167], a decontaminated tumor microbiome database based on tumor whole-genome sequencing from The Cancer Genome Atlas (TCGA). These sequences were reanalyzed using downstream PathSeq-T2T filtering steps. They were then decontaminated using the same methods as the original TCMA paper. Mutation burden for correlative analyses were obtained and calculated using the PanCanAtlas[168]a Mutations file (mc3.v0.2.8.PUBLIC.maf.gz). Pan-cancer annotations for microsatellite-instability status were taken from MANTIS[169].

Regression analyses of tumor mutation burden and microbial load

For our analysis of CRC tumor microbiomes, we compared tumor mutation burden (TMB) which we measured using $\log_{10}$ mutations per megabase (SNVs and indels). This was provided by Cornish et al.[73] along with CRC genomic subtyping (MSS/MSI/POL) and anatomical site (Proximal/Distal/Rectal) annotations, which were used as covariates. Microbial load was measured using $\log_{10}$RPM representing the sum of species-level RPM values after removing contaminants (PCE > 0.7).

We used ordinary least squares (OLS) with three complementary approaches: (*i*) within-subtype regressions to characterize subtype-specific correlations **(Figure 5E)**, and (*ii*) residualization to estimate a tumor subtype- and site-adjusted correlations **(Figure 5F-G)**, (*iii)* a nested model comparison to test whether genomic subtype adds predictive value beyond TMB and anatomic site. Model designs in R-formula format are given below.

Model 1: Bivariate models, trained separately for each subtype (MSS, MSI, and POL).

log_RPM ~ log_TMB_per_mb

Model 2: Computing correlations for residualized TMB and microbial load.

1. log_RPM ~ C(anatomic_site) $\Rightarrow$ RPM_resid
2. log_TMB_per_mb ~ C(subtype) $\Rightarrow$ TMB_resid
3. RPM_resid ~ TMB_resid

Model 3: Nested models predicting microbial load, compared with ANOVA F-test.

1. log_RPM ~ log_TMB_per_mb + C(anatomic_site) + C(subtype)
2. log_RPM ~ log_TMB_per_mb + C(anatomic_site)

Regression analyses were performed using python's "statsmodels" module (v. 0.13.5).

Bivariate analysis of CRC tumor site and TMB subtype

To jointly model the influence of CRC tumor site and TMB subtype (TMB-high versus TMB-low), we performed a blocked analysis of species-level relative abundance **(Figure 5H)**. For our analysis of CRC microbiome variation by tumor site and subtype, we defined CRC samples with $\log_{10}$TMB > 1.25 as "High-TMB", and samples with $\log_{10}$TMB ≤ 1.25 as "Low-TMB", and categorized CRCs arising in the cecum, ascending colon, transverse colon, or hepatic flexure as

"proximal" and CRCs arising in the splenic flexure, descending colon, sigmoid colon, rectosigmoid junction, or rectum as "distal".

Post-decontamination species-level relative abundances were calculated for each cancer. For species detected in >10% of CRC samples, we used a Wilcoxon rank-sum test to evaluate site and subtype associations for each species. This test was used to test four pairwise comparisons: (1) proximal vs. distal, within TMB-low; (2) proximal vs. distal, within TMB-high; (3) TMB-high vs. TMB-low, within proximal; and (4) TMB-high vs. TMB-low, within distal. This blocking was implemented to control confounding effects, since TMB-high subtypes occur more frequently in the proximal colon while TMB-low subtypes are more common distal CRCs. Test statistics for both "proximal vs. distal" comparisons were averaged, as were statistics for both "TMB-high vs. TMB-low" comparisons, to obtain a composite Wilcoxon test statistic. The *p*-values for matched comparisons were combined using Fisher's method, then corrected for multiple comparisons using Benjamini-Hochberg. Candidate CRC biomarker species and oral-typical annotations were provided by Piccinno et al.[127].

Analysis of microbial composition and early-onset CRC

To study microbial signals associated with CRC age of onset, we evaluated species-level relative abundances between early-onset (EO) and average-onset (AO) CRC. We defined EO-CRC as cases for which patient age was <50 years at the time of tumor sampling and AO-CRC as cases for which patients were ≥50 years of age at sampling. We restricted our analysis of age of onset to MSS cancers. Due to low sample numbers of EO-CRC cases, comparisons were not possible for high-TMB subtypes (MSI/POL). We performed relative abundance comparisons for both proximal MSS and distal MSS using Wilcoxon's rank-sum test on relative abundances, as described above. Validation analyses were performed using decontaminated CRC tumor microbiome profiles from The Cancer Microbiome Atlas[57] (TCMA; available at https://doi.org/10.7924/r4bk1j35s).

References

1. Sender, R., Fuchs, S., and Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLoS Biol *14*, e1002533. https://doi.org/10.1371/journal.pbio.1002533.

2. Berg, R.D. (1996). The indigenous gastrointestinal microflora. Trends in Microbiology *4*, 430–435. https://doi.org/10.1016/0966-842X(96)10057-3.

3. Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A.B., Brady, A., Creasy, H.H., McCracken, C., Giglio, M.G., et al. (2017). Strains, functions and dynamics in the expanded Human Microbiome Project. Nature *550*, 61–66. https://doi.org/10.1038/nature23889.

4. Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J.H., Chinwalla, A.T., Creasy, H.H., Earl, A.M., FitzGerald, M.G., Fulton, R.S., et al. (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214. https://doi.org/10.1038/nature11234.

5. Yang, K., Li, G., Li, Q., Wang, W., Zhao, X., Shao, N., Qiu, H., Liu, J., Xu, L., and Zhao, J. (2025). Distribution of gut microbiota across intestinal segments and their impact on human physiological and pathological processes. Cell & Bioscience *15*, 47. https://doi.org/10.1186/s13578-025-01385-y.

6. Segata, N., Haake, S.K., Mannon, P., Lemon, K.P., Waldron, L., Gevers, D., Huttenhower, C., and Izard, J. (2012). Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. Genome Biology *13*, R42. https://doi.org/10.1186/gb-2012-13-6-r42.

7. Stearns, J.C., Lynch, M.D.J., Senadheera, D.B., Tenenbaum, H.C., Goldberg, M.B., Cvitkovitch, D.G., Croitoru, K., Moreno-Hagelsieb, G., and Neufeld, J.D. (2011). Bacterial biogeography of the human digestive tract. Sci Rep *1*, 170. https://doi.org/10.1038/srep00170.

8. Donaldson, G.P., Lee, S.M., and Mazmanian, S.K. (2016). Gut biogeography of the bacterial microbiota. Nat Rev Microbiol *14*, 20–32. https://doi.org/10.1038/nrmicro3552.

9. Hillman, E.T., Lu, H., Yao, T., and Nakatsu, C.H. (2017). Microbial Ecology along the Gastrointestinal Tract. Microbes Environ *32*, 300–313. https://doi.org/10.1264/jsme2.ME17017.

10. Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S.R., Nelson, K.E., and Relman, D.A. (2005). Diversity of the Human Intestinal Microbial Flora. Science *308*, 1635–1638. https://doi.org/10.1126/science.1110591.

11. Baker, J.L., Mark Welch, J.L., Kauffman, K.M., McLean, J.S., and He, X. (2024). The oral microbiome: diversity, biogeography and human health. Nat Rev Microbiol *22*, 89–104. https://doi.org/10.1038/s41579-023-00963-6.

12. Byrd, A.L., Belkaid, Y., and Segre, J.A. (2018). The human skin microbiome. Nat Rev Microbiol *16*, 143–155. https://doi.org/10.1038/nrmicro.2017.157.

13. Oliphant, K., and Allen-Vercoe, E. (2019). Macronutrient metabolism by the human gut microbiome: major fermentation by-products and their impact on host health. Microbiome *7*, 91. https://doi.org/10.1186/s40168-019-0704-8.

14. Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J.-M., Kennedy, S., et al. (2013). Richness of human gut microbiome correlates with metabolic markers. Nature *500*, 541–546. https://doi.org/10.1038/nature12506.

15. Donald, K., and Finlay, B.B. (2023). Early-life interactions between the microbiota and immune system: impact on immune system development and atopic disease. Nat Rev Immunol *23*, 735–748. https://doi.org/10.1038/s41577-023-00874-w.

16. Hillier, S.L., Nugent, R.P., Eschenbach, D.A., Krohn, M.A., Gibbs, R.S., Martin, D.H., Cotch, M.F., Edelman, R., Pastorek, J.G., and Rao, A.V. (1995). Association between bacterial vaginosis and preterm delivery of a low-birth-weight infant. The Vaginal Infections and Prematurity Study Group. N Engl J Med *333*, 1737–1742. https://doi.org/10.1056/NEJM199512283332604.

17. Fettweis, J.M., Serrano, M.G., Brooks, J.P., Edwards, D.J., Girerd, P.H., Parikh, H.I., Huang, B., Arodz, T.J., Edupuganti, L., Glascock, A.L., et al. (2019). The vaginal microbiome and preterm birth. Nat Med *25*, 1012–1021. https://doi.org/10.1038/s41591-019-0450-2.

18. Buffie, C.G., Bucci, V., Stein, R.R., McKenney, P.T., Ling, L., Gobourne, A., No, D., Liu, H., Kinnebrew, M., Viale, A., et al. (2015). Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. Nature *517*, 205–208. https://doi.org/10.1038/nature13828.

19. Hsiao, A., Ahmed, A.M.S., Subramanian, S., Griffin, N.W., Drewry, L.L., Petri, W.A., Haque, R., Ahmed, T., and Gordon, J.I. (2014). Members of the human gut microbiota involved in recovery from Vibrio cholerae infection. Nature *515*, 423–426. https://doi.org/10.1038/nature13738.

20. Garrett, W.S., Lord, G.M., Punit, S., Lugo-Villarino, G., Mazmanian, S.K., Ito, S., Glickman, J.N., and Glimcher, L.H. (2007). Communicable ulcerative colitis induced by T-bet deficiency in the innate immune system. Cell *131*, 33–45. https://doi.org/10.1016/j.cell.2007.08.017.

21. Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N., Schirmer, M., Avila-Pacheco, J., Poon, T.W., Andrews, E., Ajami, N.J., Bonham, K.S., Brislawn, C.J., et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature *569*, 655–662. https://doi.org/10.1038/s41586-019-1237-9.

22. Karlsson, F.H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C.J., Fagerberg, B., Nielsen, J., and Bäckhed, F. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature *498*, 99–103. https://doi.org/10.1038/nature12198.

23. Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature *490*, 55–60. https://doi.org/10.1038/nature11450.

24. Kostic, A.D., Chun, E., Robertson, L., Glickman, J.N., Gallini, C.A., Michaud, M., Clancy, T.E., Chung, D.C., Lochhead, P., Hold, G.L., et al. (2013). Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. Cell Host Microbe *14*, 207–215. https://doi.org/10.1016/j.chom.2013.07.007.

25. Nomura, A., Stemmermann, G.N., Chyou, P.-H., Kato, I., Perez-Perez, G.I., and Blaser, M.J. (1991). Helicobacter pylori Infection and Gastric Carcinoma among Japanese Americans in Hawaii. New England Journal of Medicine *325*, 1132–1136. https://doi.org/10.1056/NEJM199110173251604.

26. zur Hausen, H. (2002). Papillomaviruses and cancer: from basic studies to clinical application. Nat Rev Cancer *2*, 342–350. https://doi.org/10.1038/nrc798.

27. El Tekle, G., and Garrett, W.S. (2023). Bacteria in cancer initiation, promotion and progression. Nat Rev Cancer *23*, 600–618. https://doi.org/10.1038/s41568-023-00594-2.

28. 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.

29. Routy, B., Le Chatelier, E., 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.

30. 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.

31. Park, J.S., Gazzaniga, F.S., Wu, M., Luthens, A.K., Gillis, J., Zheng, W., LaFleur, M.W., Johnson, S.B., Morad, G., Park, E.M., et al. (2023). Targeting PD-L2–RGMb overcomes microbiome-related immunotherapy resistance. Nature *617*, 377–385. https://doi.org/10.1038/s41586-023-06026-3.

32. Gazzaniga, F.S., and Kasper, D.L. (2025). The gut microbiome and cancer response to immune checkpoint inhibitors. J Clin Invest *135*. https://doi.org/10.1172/JCI184321.

33. Derosa, L., Routy, B., Thomas, A.M., Iebba, V., Zalcman, G., Friard, S., Mazieres, J., Audigier-Valette, C., Moro-Sibilot, D., Goldwasser, F., et al. (2022). Intestinal Akkermansia muciniphila predicts clinical response to PD-1 blockade in patients with advanced non-small-cell lung cancer. Nat Med *28*, 315–324. https://doi.org/10.1038/s41591-021-01655-5.

34. Tintelnot, J., Xu, Y., Lesker, T.R., Schönlein, M., Konczalla, L., Giannou, A.D., Pelczar, P., Kylies, D., Puelles, V.G., Bielecka, A.A., et al. (2023). Microbiota-derived 3-IAA influences chemotherapy efficacy in pancreatic cancer. Nature *615*, 168–174. https://doi.org/10.1038/s41586-023-05728-y.

35. Kostic, A.D., Gevers, D., Pedamallu, C.S., Michaud, M., Duke, F., Earl, A.M., Ojesina, A.I., Jung, J., Bass, A.J., Tabernero, J., et al. (2012). Genomic analysis identifies association of

Fusobacterium with colorectal carcinoma. Genome Res *22*, 292–298. https://doi.org/10.1101/gr.126573.111.

36. Castellarin, M., Warren, R.L., Freeman, J.D., Dreolini, L., Krzywinski, M., Strauss, J., Barnes, R., Watson, P., Allen-Vercoe, E., Moore, R.A., et al. (2012). Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. Genome Res *22*, 299–306. https://doi.org/10.1101/gr.126516.111.

37. 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.

38. Riquelme, E., Zhang, Y., Zhang, L., Montiel, M., Zoltan, M., Dong, W., Quesada, P., Sahin, I., Chandra, V., San Lucas, A., et al. (2019). Tumor Microbiome Diversity and Composition Influence Pancreatic Cancer Outcomes. Cell *178*, 795-806.e12. https://doi.org/10.1016/j.cell.2019.07.008.

39. 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.

40. Raab-Traub, N. (2002). Epstein-Barr virus in the pathogenesis of NPC. Semin Cancer Biol *12*, 431–441. https://doi.org/10.1016/s1044579x0200086x.

41. Shibata, D., and Weiss, L.M. (1992). Epstein-Barr virus-associated gastric adenocarcinoma. Am J Pathol *140*, 769–774.

42. Arzumanyan, A., Reis, H.M.G.P.V., and Feitelson, M.A. (2013). Pathogenic mechanisms in HBV- and HCV-associated hepatocellular carcinoma. Nat Rev Cancer *13*, 123–135. https://doi.org/10.1038/nrc3449.

43. Parsonnet, J., Friedman, G.D., Vandersteen, D.P., Chang, Y., Vogelman, J.H., Orentreich, N., and Sibley, R.K. (1991). Helicobacter pylori infection and the risk of gastric carcinoma. N Engl J Med *325*, 1127–1131. https://doi.org/10.1056/NEJM199110173251603.

44. Uemura, N., Okamoto, S., Yamamoto, S., Matsumura, N., Yamaguchi, S., Yamakido, M., Taniyama, K., Sasaki, N., and Schlemper, R.J. (2001). Helicobacter pylori Infection and the Development of Gastric Cancer. New England Journal of Medicine *345*, 784–789. https://doi.org/10.1056/NEJMoa001999.

45. Wu, S., Rhee, K.-J., Albesiano, E., Rabizadeh, S., Wu, X., Yen, H.-R., Huso, D.L., Brancati, F.L., Wick, E., McAllister, F., et al. (2009). A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. Nat Med *15*, 1016–1022. https://doi.org/10.1038/nm.2015.

46. Dejea, C.M., Fathi, P., Craig, J.M., Boleij, A., Taddese, R., Geis, A.L., Wu, X., DeStefano Shields, C.E., Hechenbleikner, E.M., Huso, D.L., et al. (2018). Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria. Science *359*, 592–597. https://doi.org/10.1126/science.aah3648.

47. Nougayrède, J.-P., Homburg, S., Taieb, F., Boury, M., Brzuszkiewicz, E., Gottschalk, G., Buchrieser, C., Hacker, J., Dobrindt, U., and Oswald, E. (2006). Escherichia coli induces DNA double-strand breaks in eukaryotic cells. Science *313*, 848–851. https://doi.org/10.1126/science.1127059.

48. Wilson, M.R., Jiang, Y., Villalta, P.W., Stornetta, A., Boudreau, P.D., Carrá, A., Brennan, C.A., Chun, E., Ngo, L., Samson, L.D., et al. (2019). The human gut bacterial genotoxin colibactin alkylates DNA. Science *363*, eaar7785. https://doi.org/10.1126/science.aar7785.

49. Pleguezuelos-Manzano, C., Puschhof, J., Rosendahl Huber, A., van Hoeck, A., Wood, H.M., Nomburg, J., Gurjao, C., Manders, F., Dalmasso, G., Stege, P.B., et al. (2020). Mutational signature in colorectal cancer caused by genotoxic pks+ E. coli. Nature *580*, 269–273. https://doi.org/10.1038/s41586-020-2080-8.

50. Arthur, J.C., Perez-Chanona, E., Mühlbauer, M., Tomkovich, S., Uronis, J.M., Fan, T.-J., Campbell, B.J., Abujamel, T., Dogan, B., Rogers, A.B., et al. (2012). Intestinal inflammation targets cancer-inducing activity of the microbiota. Science *338*, 120–123. https://doi.org/10.1126/science.1224820.

51. Xue, M., Kim, C.S., Healy, A.R., Wernke, K.M., Wang, Z., Frischling, M.C., Shine, E.E., Wang, W., Herzon, S.B., and Crawford, J.M. (2019). Structure elucidation of colibactin and its DNA cross-links. Science *365*, eaax2685. https://doi.org/10.1126/science.aax2685.

52. Carlson, E.S., Haslecker, R., Lecchi, C., Aguilar Ramos, M.A., Vennelakanti, V., Honaker, L., Stornetta, A., Millán, E.S., Johnson, B.A., Kulik, H.J., et al. (2025). The specificity and structure of DNA cross-linking by the gut bacterial genotoxin colibactin. Science *390*, eady3571. https://doi.org/10.1126/science.ady3571.

53. de Martel, C., Georges, D., Bray, F., Ferlay, J., and Clifford, G.M. (2020). Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis. Lancet Glob Health *8*, e180–e190. https://doi.org/10.1016/S2214-109X(19)30488-7.

54. 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.

55. 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). RETRACTED ARTICLE: Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature *579*, 567–574. https://doi.org/10.1038/s41586-020-2095-1.

56. Sepich-Poore, G.D., McDonald, D., Kopylova, E., Guccione, C., Zhu, Q., Austin, G., Carpenter, C., Fraraccio, S., Wandro, S., Kosciolek, T., et al. (2024). Robustness of cancer microbiome signals over a broad range of methodological variation. Oncogene *43*, 1127–1148. https://doi.org/10.1038/s41388-024-02974-w.

57. Dohlman, A.B., Arguijo Mendoza, D., Ding, S., Gao, M., Dressman, H., Iliev, I.D., Lipkin, S.M., and Shen, X. (2021). The cancer microbiome atlas: a pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants. Cell Host Microbe *29*, 281-298.e5. https://doi.org/10.1016/j.chom.2020.12.001.

58. 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.

59. Dohlman, A.B., Klug, J., Mesko, M., Gao, I.H., Lipkin, S.M., Shen, X., and Iliev, I.D. (2022). A pan-cancer mycobiome analysis reveals fungal involvement in gastrointestinal and lung tumors. Cell *185*, 3807-3822.e12. https://doi.org/10.1016/j.cell.2022.09.015.

60. Narunsky-Haziza, L., Sepich-Poore, G.D., Livyatan, I., Asraf, O., Martino, C., Nejman, D., Gavert, N., Stajich, J.E., Amit, G., González, A., et al. (2022). Pan-cancer analyses reveal cancer-type-specific fungal ecologies and bacteriome interactions. Cell *185*, 3789-3806.e17. https://doi.org/10.1016/j.cell.2022.09.005.

61. Gihawi, A., Wood, H.M., Clark, J., O'Grady, J., Eeles, R.A., Wedge, D.C., Jakobsdottir, G.M., Magiorkinis, G., Schache, A.G., Masterson, L., et al. (2025). The landscape of microbial associations in human cancer. Science Translational Medicine *17*, eads6166. https://doi.org/10.1126/scitranslmed.ads6166.

62. Ge, Y., Lu, J., Puiu, D., Revsine, M., and Salzberg, S.L. (2025). Comprehensive analysis of microbial content in whole-genome sequencing samples from The Cancer Genome Atlas project. Sci Transl Med *17*, eads6335. https://doi.org/10.1126/scitranslmed.ads6335.

63. Ghaddar, B.C., Blaser, M.J., and De, S. (2025). Revisiting the cancer microbiome using PRISM. Preprint at bioRxiv, https://doi.org/10.1101/2025.01.21.634087 https://doi.org/10.1101/2025.01.21.634087.

64. Robinson, K.M., Crabtree, J., Mattick, J.S.A., Anderson, K.E., and Dunning Hotopp, J.C. (2017). Distinguishing potential bacteria-tumor associations from contamination in a secondary data analysis of public cancer genome sequence data. Microbiome *5*, 9. https://doi.org/10.1186/s40168-016-0224-8.

65. Battaglia, T.W., Mimpen, I.L., Traets, J.J.H., Hoeck, A. van, Zeverijn, L.J., Geurts, B.S., Wit, G.F. de, Noë, M., Hofland, I., Vos, J.L., et al. (2024). A pan-cancer analysis of the microbiome in metastatic cancer. Cell *187*, 2324-2335.e19. https://doi.org/10.1016/j.cell.2024.03.021.

66. Kostic, A.D., Ojesina, A.I., Pedamallu, C.S., Jung, J., Verhaak, R.G.W., Getz, G., and Meyerson, M. (2011). PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nat Biotechnol *29*, 393–396. https://doi.org/10.1038/nbt.1868.

67. 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.

68. Feng, H., Shuda, M., Chang, Y., and Moore, P.S. (2008). Clonal integration of a polyomavirus in human Merkel cell carcinoma. Science *319*, 1096–1100. https://doi.org/10.1126/science.1152586.

69. Guccione, C., Patel, L., Tomofuji, Y., McDonald, D., Gonzalez, A., Sepich-Poore, G.D., Sonehara, K., Zakeri, M., Chen, Y., Dilmore, A.H., et al. (2025). Incomplete human reference

genomes can drive false sex biases and expose patient-identifying information in metagenomic data. Nat Commun *16*, 825. https://doi.org/10.1038/s41467-025-56077-5.

70. Turnbull, C., Scott, R.H., Thomas, E., Jones, L., Murugaesu, N., Pretty, F.B., Halai, D., Baple, E., Craig, C., Hamblin, A., et al. (2018). The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. BMJ *361*, k1687. https://doi.org/10.1136/bmj.k1687.

71. Turnbull, C. (2018). Introducing whole-genome sequencing into routine cancer care: the Genomics England 100 000 Genomes Project. Annals of Oncology *29*, 784–787. https://doi.org/10.1093/annonc/mdy054.

72. Kinnersley, B., Sud, A., Everall, A., Cornish, A.J., Chubb, D., Culliford, R., Gruber, A.J., Lärkeryd, A., Mitsopoulos, C., Wedge, D., et al. (2024). Analysis of 10,478 cancer genomes identifies candidate driver genes and opportunities for precision oncology. Nat Genet *56*, 1868–1877. https://doi.org/10.1038/s41588-024-01785-9.

73. Cornish, A.J., Gruber, A.J., Kinnersley, B., Chubb, D., Frangou, A., Caravagna, G., Noyvert, B., Lakatos, E., Wood, H.M., Thorn, S., et al. (2024). The genomic landscape of 2,023 colorectal cancers. Nature *633*, 127–136. https://doi.org/10.1038/s41586-024-07747-9.

74. National Genomic Research Library v5.1, Genomics England (2020). 10.6084/m9.figshare.4530893/7.

75. Rhie, A., Nurk, S., Cechova, M., Hoyt, S.J., Taylor, D.J., Altemose, N., Hook, P.W., Koren, S., Rautiainen, M., Alexandrov, I.A., et al. (2023). The complete sequence of a human Y chromosome. Nature *621*, 344–354. https://doi.org/10.1038/s41586-023-06457-y.

76. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. Science *376*, 44–53. https://doi.org/10.1126/science.abj6987.

77. 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.

78. Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L.J., Thompson, K.N., Zolfo, M., Manghi, P., Dubois, L., Huang, K.D., Thomas, A.M., et al. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nat Biotechnol *41*, 1633–1644. https://doi.org/10.1038/s41587-023-01688-w.

79. Shaw, J., and Yu, Y.W. (2025). Rapid species-level metagenome profiling and containment estimation with sylph. Nat Biotechnol *43*, 1348–1359. https://doi.org/10.1038/s41587-024-02412-y.

80. 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.

81. Eisenhofer, R., Minich, J.J., Marotz, C., Cooper, A., Knight, R., and Weyrich, L.S. (2019). Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. Trends Microbiol *27*, 105–117. https://doi.org/10.1016/j.tim.2018.11.003.

82. Fierer, N., Leung, P.M., Lappan, R., Eisenhofer, R., Ricci, F., Holland, S.I., Dragone, N., Blackall, L.L., Dong, X., Dorador, C., et al. (2025). Guidelines for preventing and reporting contamination in low-biomass microbiome studies. Nat Microbiol *10*, 1570–1580. https://doi.org/10.1038/s41564-025-02035-2.

83. Effects of Index Misassignment on Multiplexing and Downstream Analysis.

84. Larsson, A.J.M., Stanley, G., Sinha, R., Weissman, I.L., and Sandberg, R. (2018). Computational correction of index switching in multiplexed sequencing libraries. Nat Methods *15*, 305–307. https://doi.org/10.1038/nmeth.4666.

85. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity. Nature *483*, 603. https://doi.org/10.1038/nature11003.

86. Tan, C.C.S., Ko, K.K.K., Chen, H., Liu, J., Loh, M., SG10K_Health Consortium, Chia, M., and Nagarajan, N. (2023). No evidence for a common blood microbiome based on a population study of 9,770 healthy humans. Nat Microbiol *8*, 973–985. https://doi.org/10.1038/s41564-023-01350-w.

87. Zhao, H., Chu, M., Huang, Z., Yang, X., Ran, S., Hu, B., Zhang, C., and Liang, J. (2017). Variations in oral microbiota associated with oral cancer. Sci Rep *7*, 11773. https://doi.org/10.1038/s41598-017-11779-9.

88. Hayes, R.B., Ahn, J., Fan, X., Peters, B.A., Ma, Y., Yang, L., Agalliu, I., Burk, R.D., Ganly, I., Purdue, M.P., et al. (2018). Association of Oral Microbiome With Risk for Incident Head and Neck Squamous Cell Cancer. JAMA Oncol *4*, 358–365. https://doi.org/10.1001/jamaoncol.2017.4777.

89. Börnigen, D., Ren, B., Pickard, R., Li, J., Ozer, E., Hartmann, E.M., Xiao, W., Tickle, T., Rider, J., Gevers, D., et al. (2017). Alterations in oral bacterial communities are associated with risk factors for oral and oropharyngeal cancer. Sci Rep *7*, 17686. https://doi.org/10.1038/s41598-017-17795-z.

90. Yang, W., Chen, C.-H., Jia, M., Xing, X., Gao, L., Tsai, H.-T., Zhang, Z., Liu, Z., Zeng, B., Yeung, S.-C.J., et al. (2021). Tumor-Associated Microbiota in Esophageal Squamous Cell Carcinoma. Front Cell Dev Biol *9*, 641270. https://doi.org/10.3389/fcell.2021.641270.

91. Ferreira, R.M., Pereira-Marques, J., Pinto-Ribeiro, I., Costa, J.L., Carneiro, F., Machado, J.C., and Figueiredo, C. (2018). Gastric microbial community profiling reveals a dysbiotic cancer-associated microbiota. Gut *67*, 226–236. https://doi.org/10.1136/gutjnl-2017-314205.

92. Booth, M.E., Wood, H.M., Travis, M.A., Genomics England Research Consortium, Quirke, P., and Grabsch, H.I. (2025). The relationship between the gastric cancer microbiome and clinicopathological factors: a metagenomic investigation from the 100,000 genomes project and The Cancer Genome Atlas. Gastric Cancer *28*, 358–371. https://doi.org/10.1007/s10120-025-01588-9.

93. Bender, M.J., McPherson, A.C., Phelps, C.M., Pandey, S.P., Laughlin, C.R., Shapira, J.H., Medina Sanchez, L., Rana, M., Richie, T.G., Mims, T.S., et al. (2023). Dietary tryptophan

metabolite released by intratumoral *Lactobacillus reuteri* facilitates immune checkpoint inhibitor treatment. Cell *186*, 1846-1862.e26. https://doi.org/10.1016/j.cell.2023.03.011.

94. He, Y., Li, L., Li, Y., Wang, X., Qian, L., Yang, J., and Jiang, M. (2025). Mendelian randomization study reveals causal association between skin microbiome and skin cancers. Sci Rep *15*, 21590. https://doi.org/10.1038/s41598-025-07265-2.

95. Krueger, A., Mohamed, A., Kolka, C.M., Stoll, T., Zaugg, J., Linedale, R., Morrison, M., Soyer, H.P., Hugenholtz, P., Frazer, I.H., et al. (2022). Skin Cancer-Associated S. aureus Strains Can Induce DNA Damage in Human Keratinocytes by Downregulating DNA Repair and Promoting Oxidative Stress. Cancers *14*, 2143. https://doi.org/10.3390/cancers14092143.

96. Kullander, J., Forslund, O., and Dillner, J. (2009). Staphylococcus aureus and squamous cell carcinoma of the skin. Cancer Epidemiol Biomarkers Prev *18*, 472–478. https://doi.org/10.1158/1055-9965.EPI-08-0905.

97. Krueger, A., Zaugg, J., Chisholm, S., Linedale, R., Lachner, N., Teoh, S.M., Tuong, Z.K., Lukowski, S.W., Morrison, M., Soyer, H.P., et al. (2022). Secreted Toxins From Staphylococcus aureus Strains Isolated From Keratinocyte Skin Cancers Mediate Pro-tumorigenic Inflammatory Responses in the Skin. Front. Microbiol. *12*. https://doi.org/10.3389/fmicb.2021.789042.

98. Eckhoff, A.M., Fletcher, A.A., Kelly, M.S., Dohlman, A., McIntyre, C.A., Shen, X., Iyer, M.K., Nussbaum, D.P., and Allen, P.J. (2024). Comprehensive Assessment of the Intrinsic Pancreatic Microbiome. Ann Surg. https://doi.org/10.1097/SLA.0000000000006299.

99. 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.

100. Ginwala, R., Bukavina, L., Sindhani, M., Nachman, E., Peri, S., Franklin, J., Drevik, J., Christianson, S., Geynisman, D.M., Kutikov, A., et al. (2025). Bladder cancer microbiome and its association with chemoresponse. Front Oncol *15*, 1506319. https://doi.org/10.3389/fonc.2025.1506319.

101. Bučević Popović, V., Šitum, M., Chow, C.-E.T., Chan, L.S., Roje, B., and Terzić, J. (2018). The urinary microbiome associated with bladder cancer. Sci Rep *8*, 12157. https://doi.org/10.1038/s41598-018-29054-w.

102. Zhang, Y., Lin, H., Liang, L., Jin, S., Lv, J., Zhou, Y., Xu, F., Liu, F., and Feng, N. (2024). Intratumoral microbiota as a novel prognostic indicator in bladder cancer. Sci Rep *14*, 22198. https://doi.org/10.1038/s41598-024-72918-7.

103. Wolfe, A.J., and Brubaker, L. (2019). Urobiome Updates: Advances in Urinary Microbiome Research. Nat Rev Urol *16*, 73–74. https://doi.org/10.1038/s41585-018-0127-5.

104. Lotte, R., Lotte, L., and Ruimy, R. (2016). *Actinotignum schaalii* (formerly *Actinobaculum schaalii*): a newly recognized pathogen—review of the literature. Clinical Microbiology and Infection *22*, 28–36. https://doi.org/10.1016/j.cmi.2015.10.038.

105. Torres, A., Cilloniz, C., Niederman, M.S., Menéndez, R., Chalmers, J.D., Wunderink, R.G., and van der Poll, T. (2021). Pneumonia. Nat Rev Dis Primers *7*, 25. https://doi.org/10.1038/s41572-021-00259-0.

106. Langenberg, A.G.M., Corey, L., Ashley, R.L., Leong, W.P., and Straus, S.E. (1999). A Prospective Study of New Infections with Herpes Simplex Virus Type 1 and Type 2. New England Journal of Medicine *341*, 1432–1438. https://doi.org/10.1056/NEJM199911043411904.

107. Young, L.S., Yap, L.F., and Murray, P.G. (2016). Epstein–Barr virus: more than 50 years old and still providing surprises. Nat Rev Cancer *16*, 789–802. https://doi.org/10.1038/nrc.2016.92.

108. Griffiths, P., and Reeves, M. (2021). Pathogenesis of human cytomegalovirus in the immunocompromised host. Nat Rev Microbiol *19*, 759–773. https://doi.org/10.1038/s41579-021-00582-z.

109. Zerr, D.M., Meier, A.S., Selke, S.S., Frenkel, L.M., Huang, M.-L., Wald, A., Rhoads, M.P., Nguy, L., Bornemann, R., Morrow, R.A., et al. (2005). A Population-Based Study of Primary Human Herpesvirus 6 Infection. New England Journal of Medicine *352*, 768–776. https://doi.org/10.1056/NEJMoa042207.

110. Verbeek, R., Vandekerckhove, L., and Van Cleemput, J. Update on human herpesvirus 7 pathogenesis and clinical aspects as a roadmap for future research. J Virol *98*, e00437-24. https://doi.org/10.1128/jvi.00437-24.

111. Galeano Niño, J.L., Wu, H., LaCourse, K.D., Kempchinsky, A.G., Baryiames, A., Barber, B., Futran, N., Houlton, J., Sather, C., Sicinska, E., et al. (2022). Effect of the intratumoral microbiota on spatial and cellular heterogeneity in cancer. Nature *611*, 810–817. https://doi.org/10.1038/s41586-022-05435-0.

112. Queen, J., Cing, Z., Minsky, H., Nandi, A., Southward, T., Ferri, J., McMann, M., Iyadorai, T., Vadivelu, J., Roslani, A., et al. (2025). Fusobacterium nucleatum is enriched in invasive biofilms in colorectal cancer. npj Biofilms Microbiomes *11*, 81. https://doi.org/10.1038/s41522-025-00717-7.

113. Socransky, S.S., Haffajee, A.D., Cugini, M.A., Smith, C., and Kent, R.L. (1998). Microbial complexes in subgingival plaque. J Clin Periodontol *25*, 134–144. https://doi.org/10.1111/j.1600-051x.1998.tb02419.x.

114. Bao, Y., Zhang, Y., Zhang, Y., Liu, Y., Wang, S., Dong, X., Wang, Y., and Zhang, H. (2010). Screening of potential probiotic properties of *Lactobacillus fermentum* isolated from traditional dairy products. Food Control *21*, 695–701. https://doi.org/10.1016/j.foodcont.2009.10.010.

115. Selle, K., and Klaenhammer, T.R. (2013). Genomic and phenotypic evidence for probiotic influences of Lactobacillus gasseri on human health. FEMS Microbiol Rev *37*, 915–935. https://doi.org/10.1111/1574-6976.12021.

116. Fu, K., Cheung, A.H.K., Wong, C.C., Liu, W., Zhou, Y., Wang, F., Huang, P., Yuan, K., Coker, O.O., Pan, Y., et al. (2024). Streptococcus anginosus promotes gastric inflammation,

atrophy, and tumorigenesis in mice. Cell *187*, 882-896.e17. https://doi.org/10.1016/j.cell.2024.01.004.

117. Fried, B., Reddy, A., and Mayer, D. (2011). Helminths in human carcinogenesis. Cancer Letters *305*, 239–249. https://doi.org/10.1016/j.canlet.2010.07.008.

118. Miller, T.L., and Wolin, M.J. (1982). Enumeration of Methanobrevibacter smithii in human feces. Arch Microbiol *131*, 14–18. https://doi.org/10.1007/BF00451492.

119. Borrel, G., Brugère, J.-F., Gribaldo, S., Schmitz, R.A., and Moissl-Eichinger, C. (2020). The host-associated archaeome. Nat Rev Microbiol *18*, 622–636. https://doi.org/10.1038/s41579-020-0407-y.

120. Triantafyllou, K., Chang, C., and Pimentel, M. (2014). Methanogens, methane and gastrointestinal motility. J Neurogastroenterol Motil *20*, 31–40. https://doi.org/10.5056/jnm.2014.20.1.31.

121. Low, A., Lee, J.K.Y., Gounot, J.-S., Ravikrishnan, A., Ding, Y., Saw, W.-Y., Tan, L.W.L., Moong, D.K.N., Teo, Y.Y., Nagarajan, N., et al. (2022). Mutual Exclusion of Methanobrevibacter Species in the Human Gut Microbiota Facilitates Directed Cultivation of a Candidatus Methanobrevibacter Intestini Representative. Microbiology Spectrum *10*, e00849-22. https://doi.org/10.1128/spectrum.00849-22.

122. Borrel, G., Harris, H.M.B., Parisot, N., Gaci, N., Tottey, W., Mihajlovski, A., Deane, J., Gribaldo, S., Bardot, O., Peyretaillade, E., et al. (2013). Genome Sequence of “Candidatus Methanomassiliicoccus intestinalis” Issoire-Mx1, a Third Thermoplasmatales-Related Methanogenic Archaeon from Human Feces. Genome Announcements *1*, 10.1128/genomea.00453-13. https://doi.org/10.1128/genomea.00453-13.

123. Poole, D.N., and McClelland, R.S. (2013). Global epidemiology of Trichomonas vaginalis. Sex Transm Infect *89*, 418–422. https://doi.org/10.1136/sextrans-2013-051075.

124. Hamar, B., Teutsch, B., Hoffmann, E., Hegyi, P., Váradi, A., Nyirády, P., Hunka, Z., Ács, N., Lintner, B., Hermánné, R.J., et al. (2023). Trichomonas vaginalis infection is associated with increased risk of cervical carcinogenesis: A systematic review and meta-analysis of 470 000 patients. International Journal of Gynecology & Obstetrics *163*, 31–43. https://doi.org/10.1002/ijgo.14763.

125. Gonelli, A., Boccia, S., Boni, M., Pozzoli, A., Rizzo, C., Querzoli, P., Cassai, E., and Di Luca, D. (2001). Human herpesvirus 7 is latent in gastric mucosa. J Med Virol *63*, 277–283. https://doi.org/10.1002/1096-9071(200104)63:4%253C277::aid-jmv1002%253E3.0.co;2-k.

126. Chiu, H.H., Lee, C.Y., Lee, P.I., Lin, K.H., and Huang, L.M. (1998). Mononucleosis syndrome and coincidental human herpesvirus-7 and Epstein-Barr virus infection. Arch Dis Child *78*, 479–480. https://doi.org/10.1136/adc.78.5.479.

127. Piccinno, G., Thompson, K.N., Manghi, P., Ghazi, A.R., Thomas, A.M., Blanco-Míguez, A., Asnicar, F., Mladenovic, K., Pinto, F., Armanini, F., et al. (2025). Pooled analysis of 3,741 stool metagenomes from 18 cohorts for cross-stage and strain-level reproducible microbial biomarkers of colorectal cancer. Nat Med, 1–14. https://doi.org/10.1038/s41591-025-03693-9.

128. Flemer, B., Lynch, D.B., Brown, J.M.R., Jeffery, I.B., Ryan, F.J., Claesson, M.J., O'Riordain, M., Shanahan, F., and O'Toole, P.W. (2017). Tumour-associated and non-tumour-associated microbiota in colorectal cancer. https://doi.org/10.1136/gutjnl-2015-309595.

129. Mima, K., Nishihara, R., Qian, Z.R., Cao, Y., Sukawa, Y., Nowak, J.A., Yang, J., Dou, R., Masugi, Y., Song, M., et al. (2016). Fusobacterium nucleatum in colorectal carcinoma tissue and patient prognosis. Gut *65*, 1973–1980. https://doi.org/10.1136/gutjnl-2015-310101.

130. Hamada, T., Zhang, X., Mima, K., Bullman, S., Sukawa, Y., Nowak, J.A., Kosumi, K., Masugi, Y., Twombly, T.S., Cao, Y., et al. (2018). Fusobacterium nucleatum in Colorectal Cancer Relates to Immune Response Differentially by Tumor Microsatellite Instability Status. Cancer Immunol Res *6*, 1327–1336. https://doi.org/10.1158/2326-6066.CIR-18-0174.

131. Muzny, D.M., Bainbridge, M.N., Chang, K., Dinh, H.H., Drummond, J.A., Fowler, G., Kovar, C.L., Lewis, L.R., Morgan, M.B., Newsham, I.F., et al. (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330–337. https://doi.org/10.1038/nature11252.

132. Dienstmann, R., Vermeulen, L., Guinney, J., Kopetz, S., Tejpar, S., and Tabernero, J. (2017). Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. Nat Rev Cancer *17*, 79–92. https://doi.org/10.1038/nrc.2016.126.

133. Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., et al. (2015). The consensus molecular subtypes of colorectal cancer. Nat Med *21*, 1350–1356. https://doi.org/10.1038/nm.3967.

134. Takashima, Y., Kawamura, H., Okadome, K., Ugai, S., Haruki, K., Arima, K., Mima, K., Akimoto, N., Nowak, J.A., Giannakis, M., et al. (2024). Enrichment of Bacteroides fragilis and enterotoxigenic Bacteroides fragilis in CpG island methylator phenotype-high colorectal carcinoma. Clin Microbiol Infect *30*, 630–636. https://doi.org/10.1016/j.cmi.2024.01.013.

135. Zepeda-Rivera, M., Minot, S.S., Bouzek, H., Wu, H., Blanco-Míguez, A., Manghi, P., Jones, D.S., LaCourse, K.D., Wu, Y., McMahon, E.F., et al. (2024). A distinct Fusobacterium nucleatum clade dominates the colorectal cancer niche. Nature *628*, 424–432. https://doi.org/10.1038/s41586-024-07182-w.

136. Wang, C., Xiao, Y., Yu, L., Tian, F., Zhao, J., Zhang, H., Chen, W., and Zhai, Q. (2021). Protective effects of different Bacteroides vulgatus strains against lipopolysaccharide-induced acute intestinal injury, and their underlying functional genes. J Adv Res *36*, 27–37. https://doi.org/10.1016/j.jare.2021.06.012.

137. Sokol, H., Pigneur, B., Watterlot, L., Lakhdari, O., Bermúdez-Humarán, L.G., Gratadoux, J.-J., Blugeon, S., Bridonneau, C., Furet, J.-P., Corthier, G., et al. (2008). Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. Proc Natl Acad Sci U S A *105*, 16731–16736. https://doi.org/10.1073/pnas.0804812105.

138. Fabersani, E., Portune, K., Campillo, I., López-Almela, I., la Paz, S.M., Romaní-Pérez, M., Benítez-Páez, A., and Sanz, Y. (2021). Bacteroides uniformis CECT 7771 alleviates

inflammation within the gut-adipose tissue axis involving TLR5 signaling in obese mice. Sci Rep *11*, 11788. https://doi.org/10.1038/s41598-021-90888-y.

139. Siegel, R.L., Torre, L.A., Soerjomataram, I., Hayes, R.B., Bray, F., Weber, T.K., and Jemal, A. (2019). Global patterns and trends in colorectal cancer incidence in young adults. Gut *68*, 2179–2185. https://doi.org/10.1136/gutjnl-2019-319511.

140. Vuik, F.E., Nieuwenburg, S.A., Bardou, M., Lansdorp-Vogelaar, I., Dinis-Ribeiro, M., Bento, M.J., Zadnik, V., Pellisé, M., Esteban, L., Kaminski, M.F., et al. (2019). Increasing incidence of colorectal cancer in young adults in Europe over the last 25 years. https://doi.org/10.1136/gutjnl-2018-317592.

141. Sung, H., Siegel, R.L., Laversanne, M., Jiang, C., Morgan, E., Zahwe, M., Cao, Y., Bray, F., and Jemal, A. (2025). Colorectal cancer incidence trends in younger versus older adults: an analysis of population-based cancer registry data. The Lancet Oncology *26*, 51–63. https://doi.org/10.1016/S1470-2045(24)00600-4.

142. Díaz-Gay, M., dos Santos, W., Moody, S., Kazachkova, M., Abbasi, A., Steele, C.D., Vangara, R., Senkin, S., Wang, J., Fitzgerald, S., et al. (2025). Geographic and age variations in mutational processes in colorectal cancer. Nature, 1–11. https://doi.org/10.1038/s41586-025-09025-8.

143. Sinicrope, F.A. (2022). Increasing Incidence of Early-Onset Colorectal Cancer. New England Journal of Medicine *386*, 1547–1558. https://doi.org/10.1056/NEJMra2200869.

144. Martini, A.M., Moricz, B.S., Ripperger, A.K., Tran, P.M., Sharp, M.E., Forsythe, A.N., Kulhankova, K., Salgado-Pabón, W., and Jones, B.D. (2020). Association of Novel Streptococcus sanguinis Virulence Factors With Pathogenesis in a Native Valve Infective Endocarditis Model. Front Microbiol *11*, 10. https://doi.org/10.3389/fmicb.2020.00010.

145. Nijjer, S., and Dubrey, S.W. (2010). Streptococcus sanguis endocarditis associated with colonic carcinoma. BMJ Case Rep *2010*, bcr09.2009.2311. https://doi.org/10.1136/bcr.09.2009.2311.

146. Macaluso, A., Simmang, C., and Anthony, T. (1998). Streptococcus sanguis bacteremia and colorectal cancer. South Med J *91*, 206–207. https://doi.org/10.1097/00007611-199802000-00016.

147. Thomas, R., Gupta, V., and Kwan, B. (2018). Second look at Streptococcus sanguinis and the colon. BMJ Case Rep *2018*, bcr2018224799. https://doi.org/10.1136/bcr-2018-224799.

148. Association of Streptococcus sanguinis Infection With Colorectal Carcinoma (2020). Consultant360. https://www.consultant360.com/article/infectious-diseases/bacterial-infections/association-streptococcus-sanguinis-infection.

149. Rawat, P.S., Li, Y., Zhang, W., Meng, X., and Liu, W. (2022). Hungatella hathewayi, an Efficient Glycosaminoglycan-Degrading Firmicutes from Human Gut and Its Chondroitin ABC Exolyase with High Activity and Broad Substrate Specificity. Appl Environ Microbiol *88*, e0154622. https://doi.org/10.1128/aem.01546-22.

150. Ottman, N., Reunanen, J., Meijerink, M., Pietilä, T.E., Kainulainen, V., Klievink, J., Huuskonen, L., Aalvink, S., Skurnik, M., Boeren, S., et al. (2017). Pili-like proteins of Akkermansia muciniphila modulate host immune responses and gut barrier function. PLOS ONE *12*, e0173004. https://doi.org/10.1371/journal.pone.0173004.

151. Chen, S., Zhang, Z., Liu, S., Chen, T., Lu, Z., Zhao, W., Mou, X., and Liu, S. (2024). Consistent signatures in the human gut microbiome of longevous populations. Gut Microbes *16*, 2393756. https://doi.org/10.1080/19490976.2024.2393756.

152. Li, Y., Ma, A., Johnson, E., Eng, C., De, S., Jiang, S., Li, Z., Spakowicz, D., and Ma, Q. (2025). The new microbiome on the block: challenges and opportunities of using human tumor sequencing data to study microbes. Nat Methods *22*, 1788–1799. https://doi.org/10.1038/s41592-025-02807-y.

153. Matsushita, H., Vesely, M.D., Koboldt, D.C., Rickert, C.G., Uppaluri, R., Magrini, V.J., Arthur, C.D., White, J.M., Chen, Y.-S., Shea, L.K., et al. (2012). Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting. Nature *482*, 400–404. https://doi.org/10.1038/nature10755.

154. Gubin, M.M., Zhang, X., Schuster, H., Caron, E., Ward, J.P., Noguchi, T., Ivanova, Y., Hundal, J., Arthur, C.D., Krebber, W.-J., et al. (2014). Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. Nature *515*, 577–581. https://doi.org/10.1038/nature13988.

155. McGranahan, N., Furness, A.J.S., Rosenthal, R., Ramskov, S., Lyngaa, R., Saini, S.K., Jamal-Hanjani, M., Wilson, G.A., Birkbak, N.J., Hiley, C.T., et al. (2016). Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. Science *351*, 1463–1469. https://doi.org/10.1126/science.aaf1490.

156. Le, D.T., Uram, J.N., Wang, H., Bartlett, B.R., Kemberling, H., Eyring, A.D., Skora, A.D., Luber, B.S., Azad, N.S., Laheru, D., et al. (2015). PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. New England Journal of Medicine *372*, 2509–2520. https://doi.org/10.1056/NEJMoa1500596.

157. Rizvi, N.A., Hellmann, M.D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J.J., Lee, W., Yuan, J., Wong, P., Ho, T.S., et al. (2015). Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. Science *348*, 124–128. https://doi.org/10.1126/science.aaa1348.

158. 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.

159. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph | Bioinformatics | Oxford Academic https://academic.oup.com/bioinformatics/article/31/10/1674/177884.

160. Gourlé, H., Karlsson-Lindsjö, O., Hayer, J., and Bongcam-Rudloff, E. (2019). Simulating Illumina metagenomic data with InSilicoSeq. Bioinformatics *35*, 521–522. https://doi.org/10.1093/bioinformatics/bty630.

161. Bairoch, A. (2018). The Cellosaurus, a Cell-Line Knowledge Resource. Journal of Biomolecular Techniques : JBT *29*, 25. https://doi.org/10.7171/jbt.18-2902-002.

162. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

163. Picard toolkit (2019).

164. Wright, R.J., Comeau, A.M., and Langille, M.G.I. (2023). From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools. Microbial Genomics *9*, 000949. https://doi.org/10.1099/mgen.0.000949.

165. Davis, N.M., Proctor, D.M., Holmes, S.P., Relman, D.A., and Callahan, B.J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome *6*, 226. https://doi.org/10.1186/s40168-018-0605-2.

166. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

167. Dohlman, A.B., Arguijo Mendoza, D., Ding, S., Gao, M., Dressman, H., Iliev, I.D., Lipkin, S.M., and Shen, X. (2021). The cancer microbiome atlas: a pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants. Cell Host Microbe *29*, 281-298.e5. https://doi.org/10.1016/j.chom.2020.12.001.

168. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet *45*, 1113–1120. https://doi.org/10.1038/ng.2764.

169. Kautto, E.A., Bonneville, R., Miya, J., Yu, L., Krook, M.A., Reeser, J.W., and Roychowdhury, S. (2016). Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. Oncotarget *8*, 7452. https://doi.org/10.18632/oncotarget.13918.

Figure 1 A host subtraction pipeline for detecting microorganisms in low-biomass human tissues

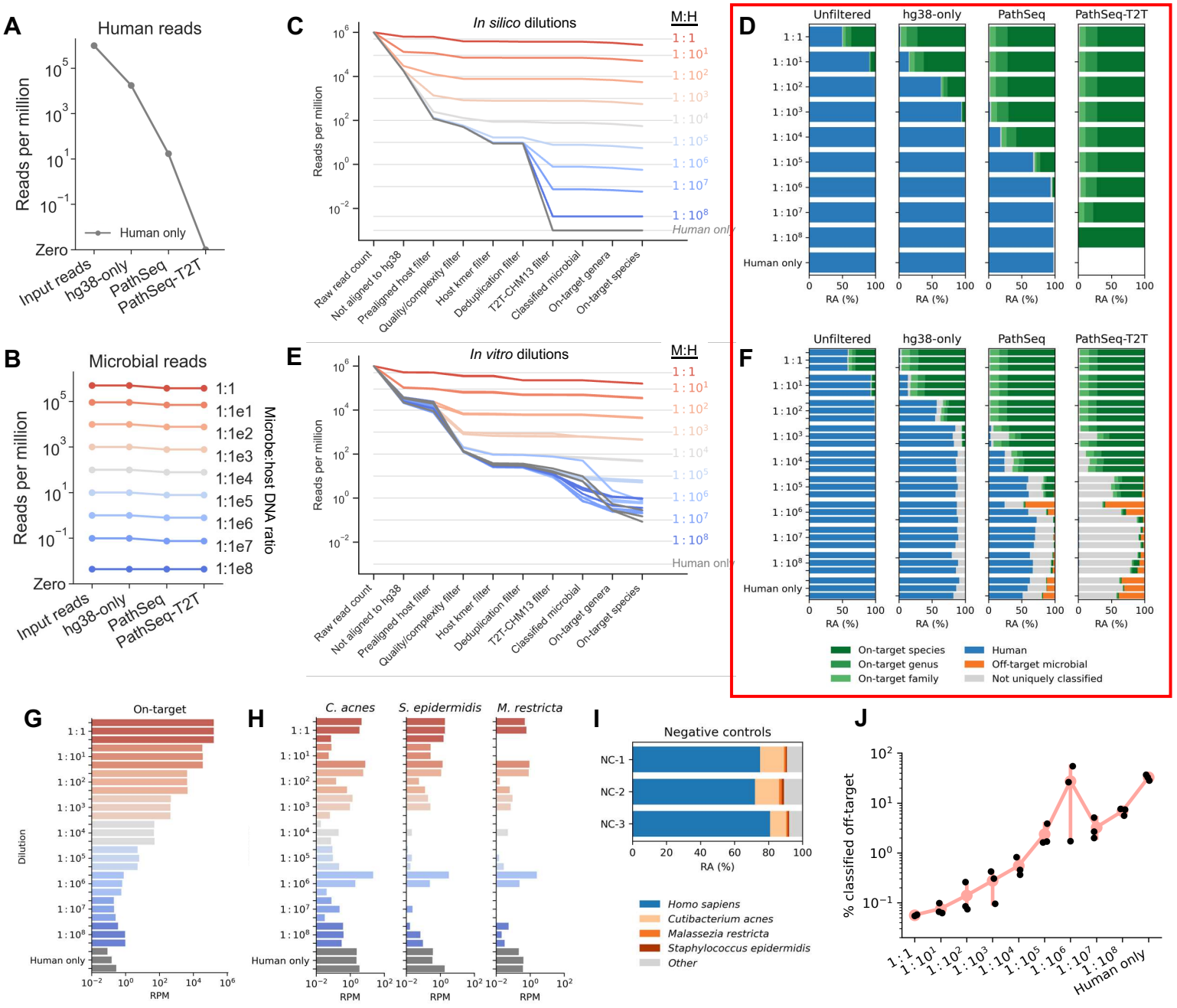
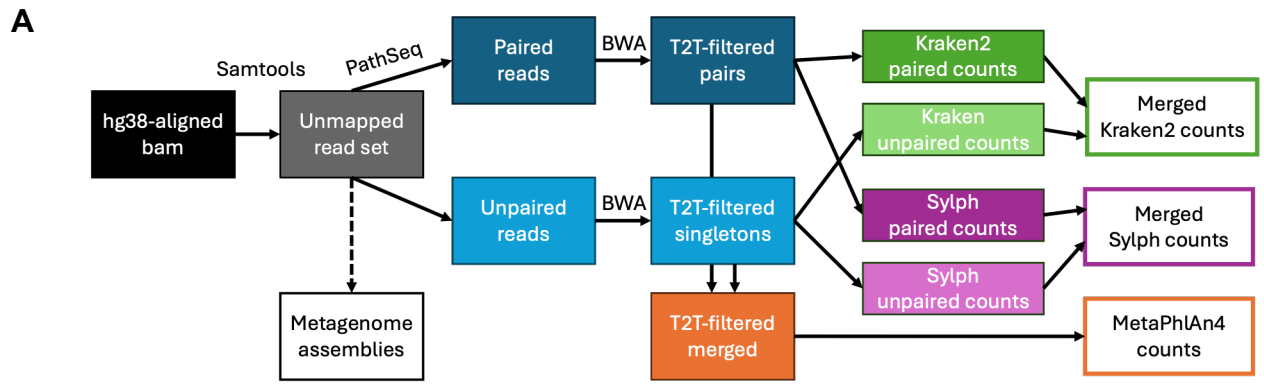
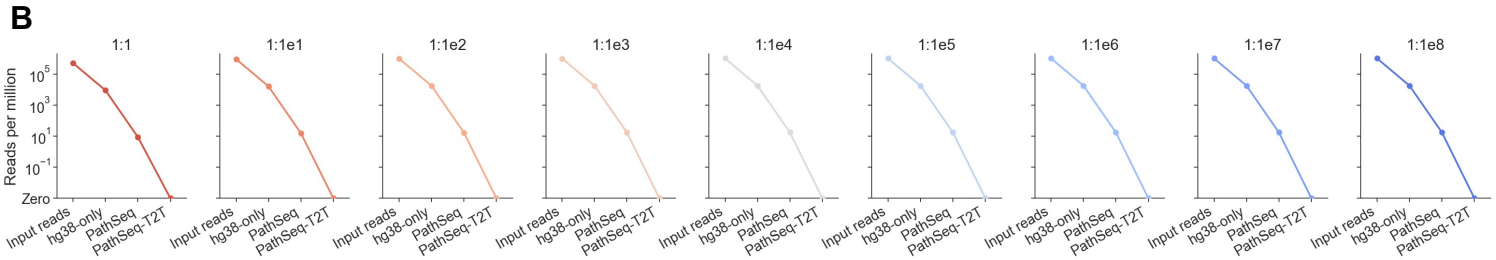
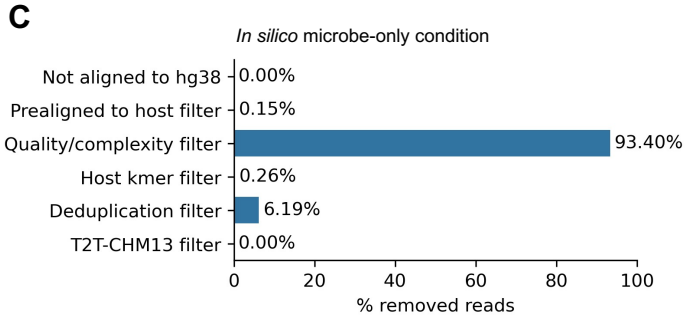
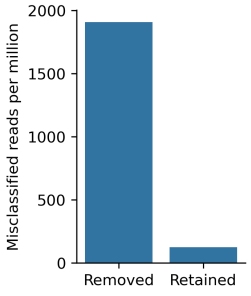


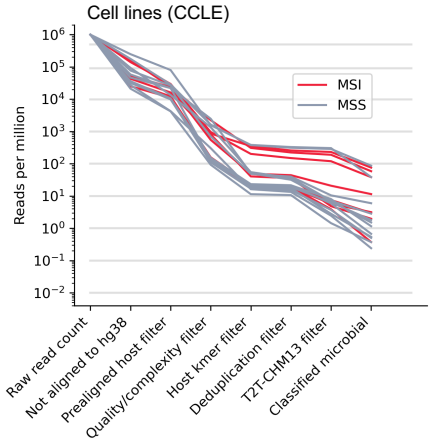
Figure S1 A host subtraction pipeline for detecting microorganisms in low-biomass human tissues

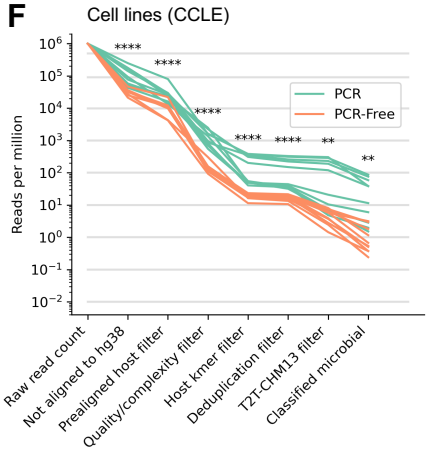












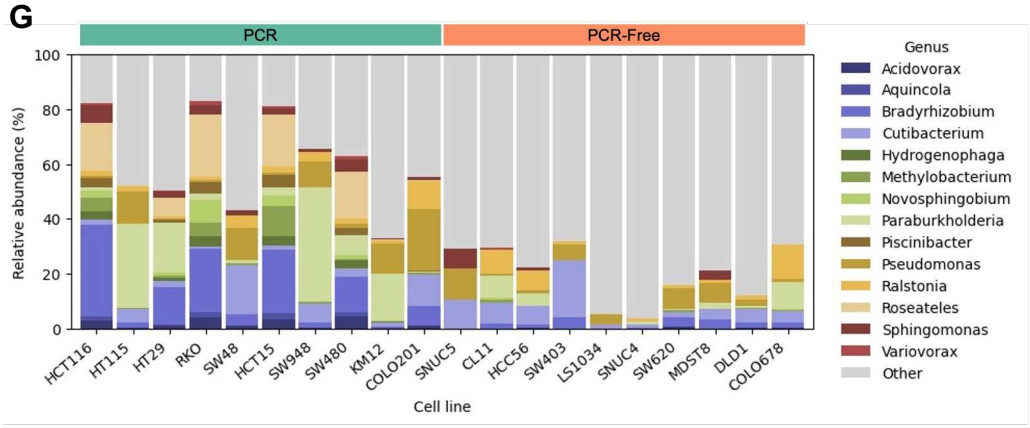


Figure 2 Identification and removal of contamination from the 100kGP cohort

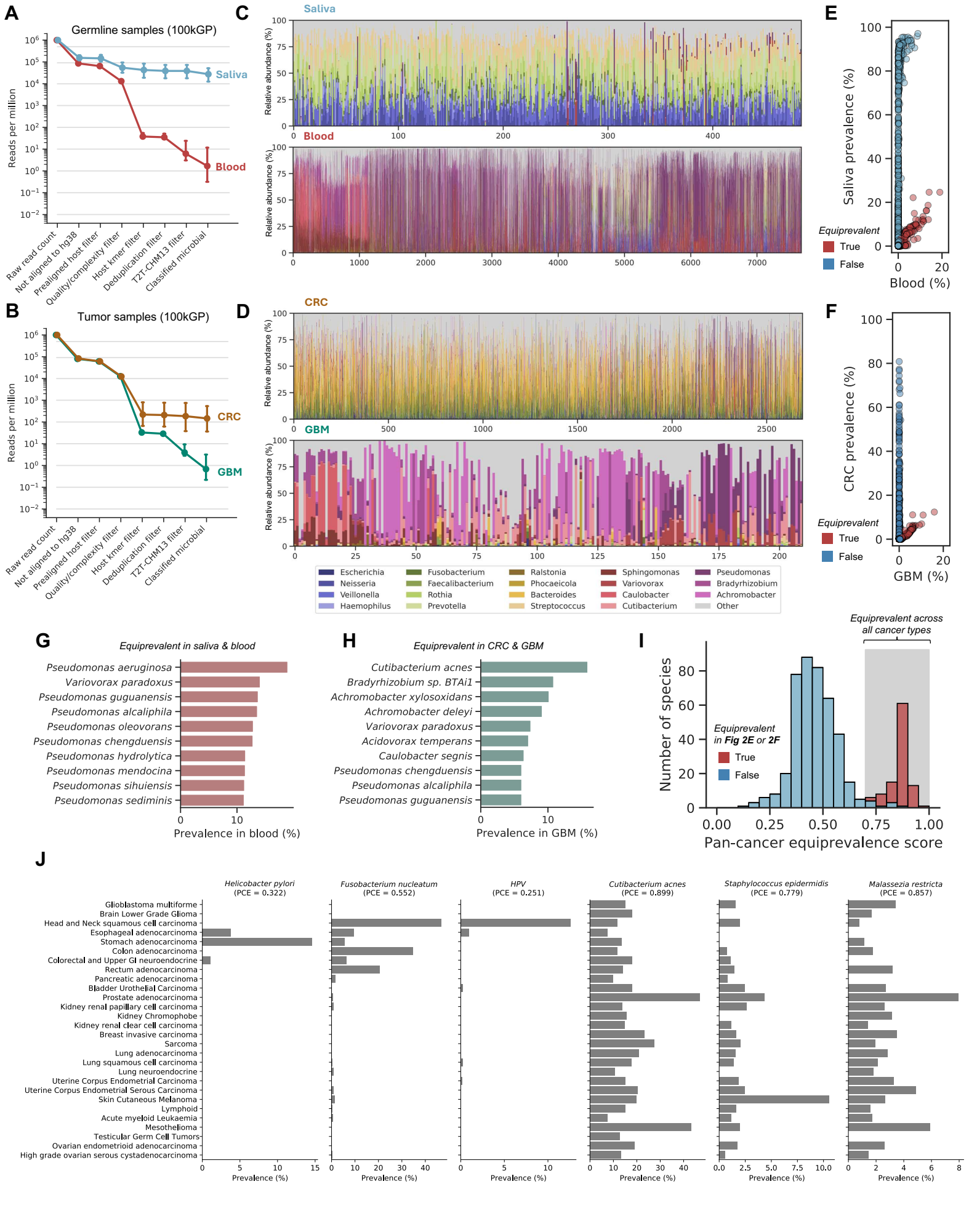
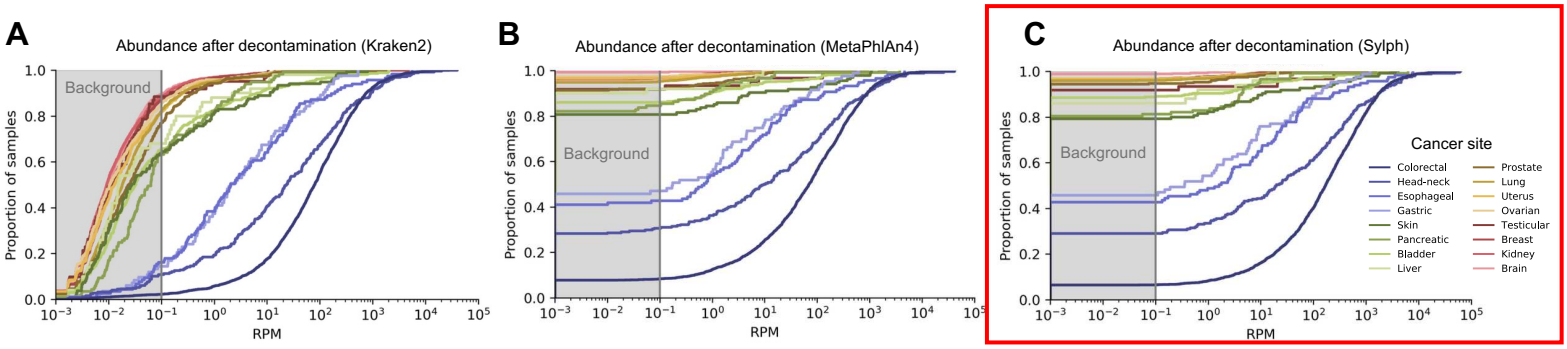
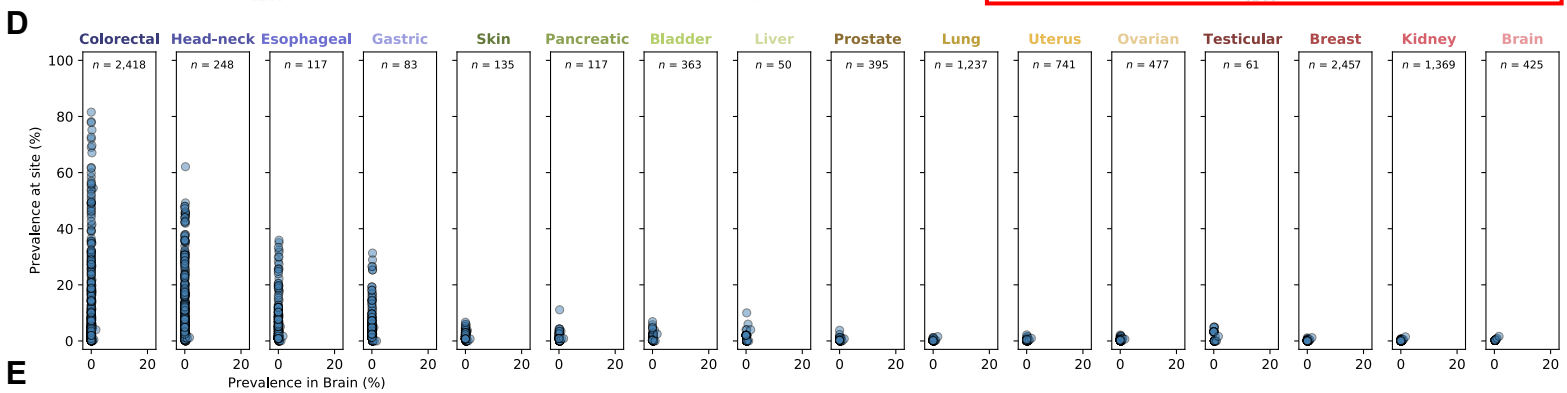
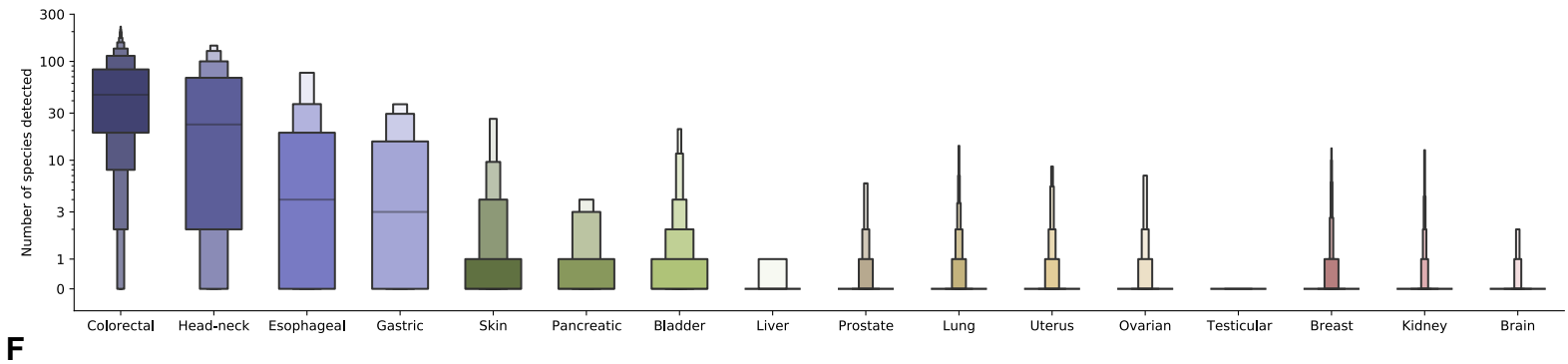


Figure S2 Identification and removal of contamination from the 100kGP cohort

Figure 3 Tumors of the oro-digestive tract harbor a microbiome but most other cancer types do not







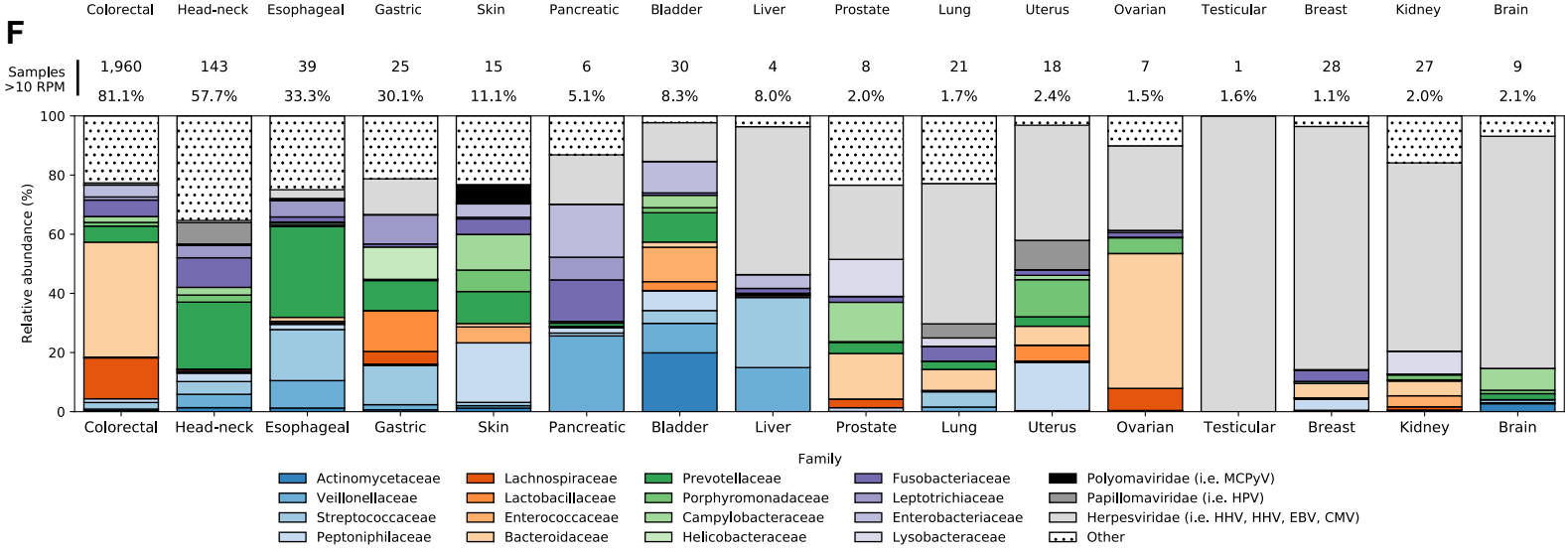


Figure S3 Tumors of the oro-digestive tract harbor a microbiome but most other cancer types do not

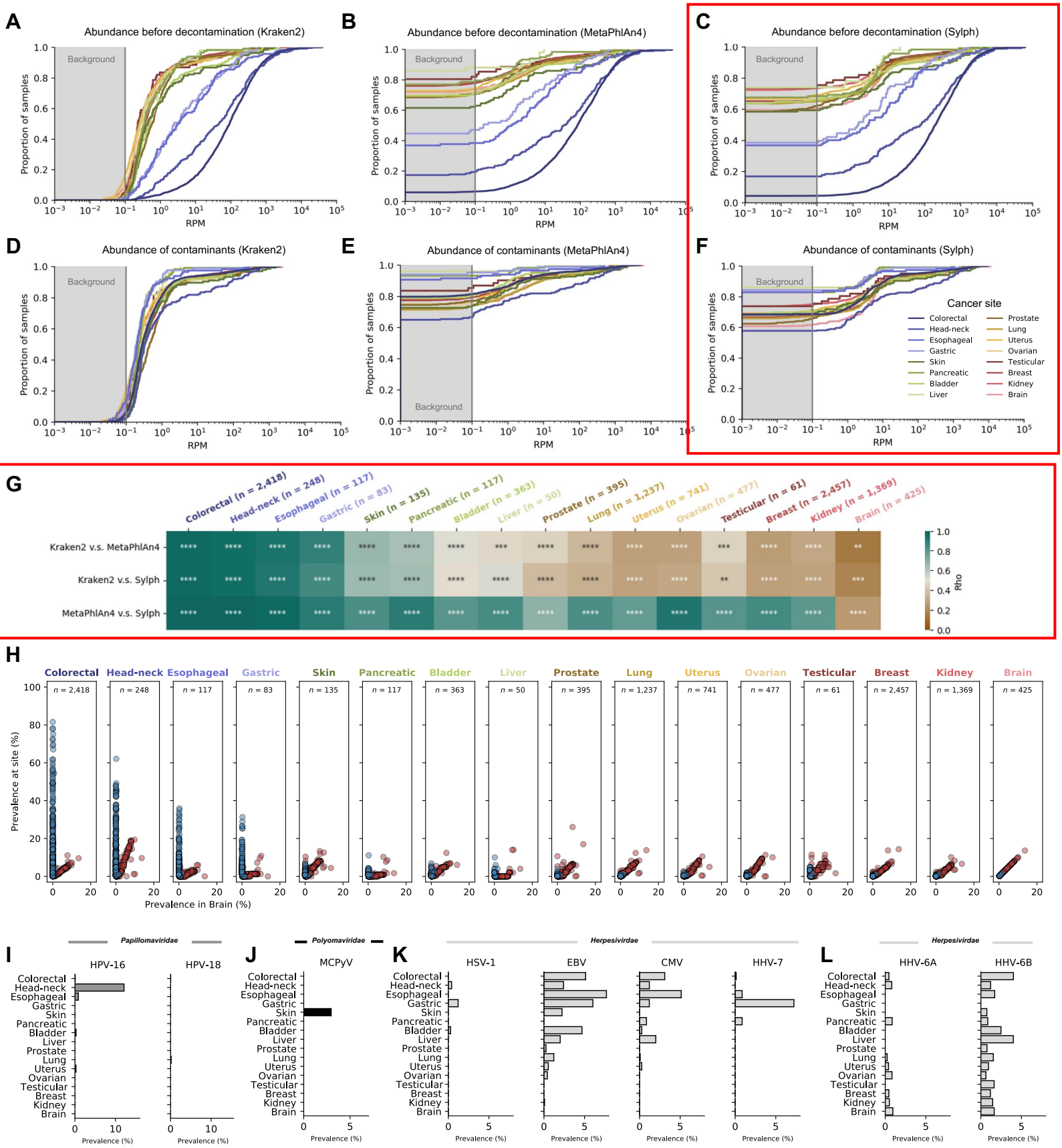


Figure 4 Biogeography and biodiversity of oropharyngeal, gastroesophageal, and colorectal cancers

Figure S4 Biogeography and biodiversity of oropharyngeal, gastroesophageal, and colorectal cancers

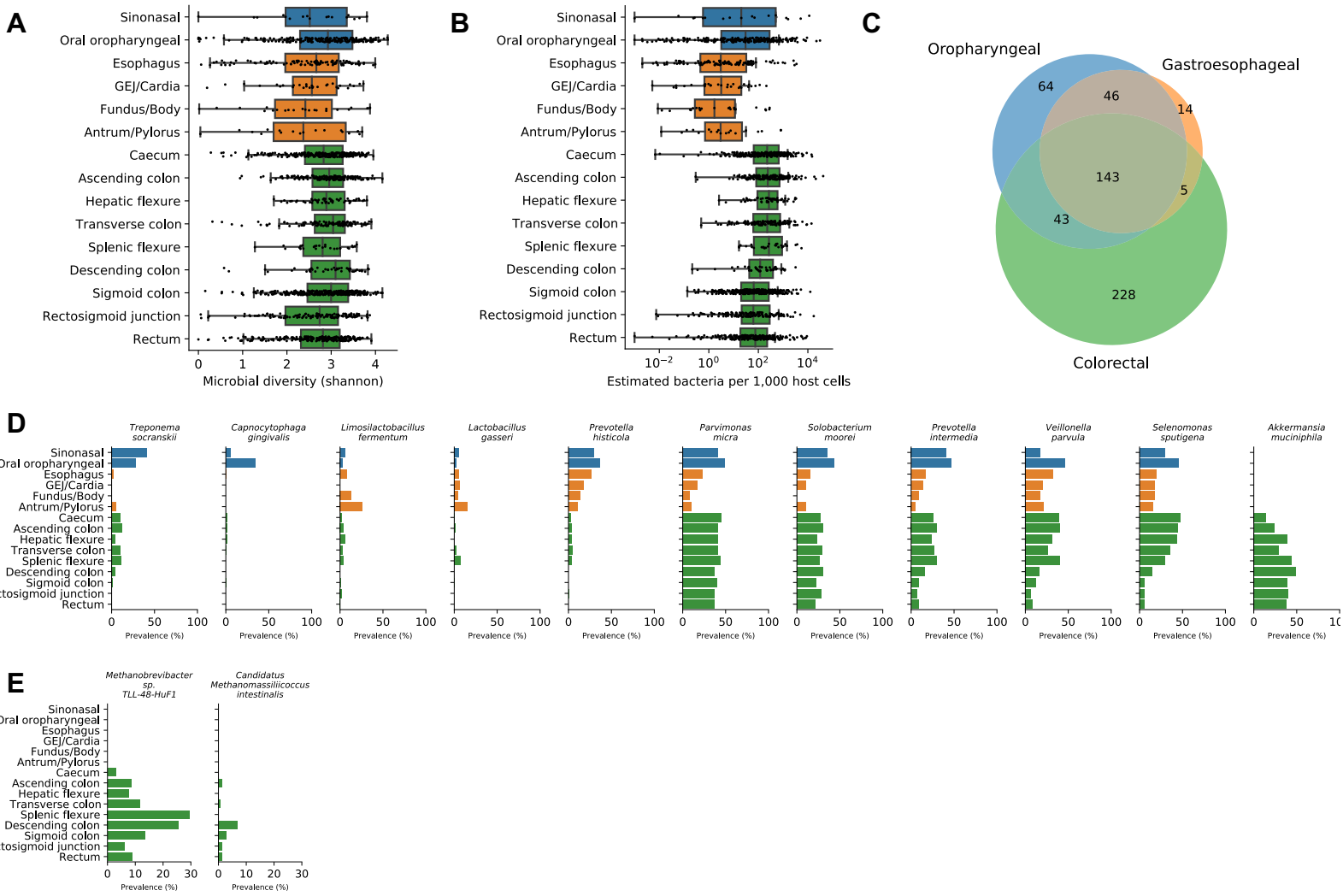


Figure 5 The tumor microbiome varies by tumor site, mutation burden, and age of onset

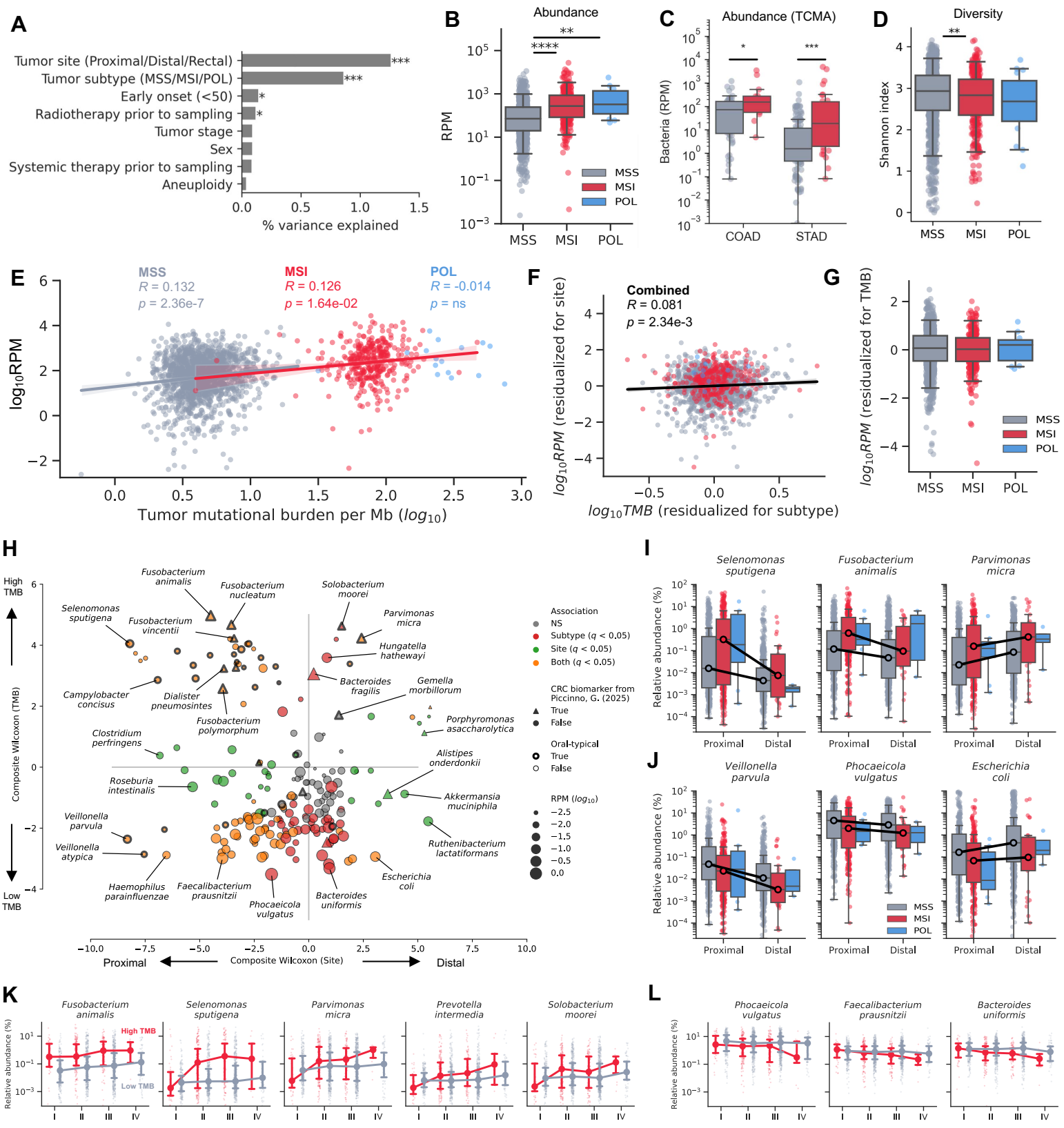


Figure S5 The tumor microbiome varies by tumor site, mutation burden, and age of onset