Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Technical replicates and multiple sequencing approaches reveal weak and inconsistent microbiome signals in tumour and blood samples

Vanessa DeClercq
Dalhousie University
André M Comeau
Dalhousie University
Alessi Kwawukume
Dalhousie University
Rowan Murphy
Dalhousie University
Nidhi R Parmar
Dalhousie University
Dylan P Quinn
Dalhousie University
Robyn Wright
Dalhousie University
Alison Wallace
Dalhousie University
Morgan GI Langille
morgan.g.i.langille@dal.ca

Dalhousie University

Research Article

Keywords: tumour, blood, microbiome, bioinformatics, 16S rRNA gene sequencing, metagenomic sequencing

Posted Date: June 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9902816/v1

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

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Microbiome on September 14th, 2026. See the published version at https://doi.org/10.1186/s40168-026-02529-z.

Abstract

Background

Unique microbial profiles have been described in tumours, blood, and other low-biomass samples. However, recent controversy has highlighted the challenges of utilizing these microbiomes for cancer diagnosis and treatment. This work aims to characterize current challenges with low-biomass samples and identify potential solutions for identifying robust microbial biomarkers using lung cancer samples.

Methods

Samples from multiple body sites (lung cancer tumour, adjacent lung tissue, blood, and saliva) were collected from patients undergoing curative surgery for non-small cell lung cancer (NSCLC). DNA extracted from samples underwent a variety of molecular techniques, including full-length amplicon sequencing, nested amplicon sequencing, metagenomic sequencing (MGS), and digital PCR (dPCR). Microbiome profiles were analyzed by sample type, in duplicate, and in relation to controls.

Results

Full-length 16S gene sequencing failed to generate useable data for nearly all tumour, adjacent tissue, and blood samples. While nested short-read 16S sequencing successfully generated sequence data these samples contained very few taxa, exhibited substantial variability across replicates, and were compositionally indistinguishable from negative controls. In contrast, saliva and positive control samples showed high diversity and strong reproducibility. Absolute quantification by dPCR confirmed extremely low bacterial biomass in tumour, adjacent tiussue, and blood samples, comparable to negative controls. Lastly, metagenomic sequencing of tumour, adjacent tissue, and blood samples detected very few taxa with minimal overlap with 16S data, whereas saliva and positive controls showed significant overlap of detected genera.

Conclusions

Current microbiome methodologies present challenges with samples of low microbial biomass, making it difficult to detect signals that are distinct or above levels found in collection and processing environments. Of concern is the lack of precision and reproducibility with samples of low microbial biomass. Technical replicates, alternative sequencing strategies, and absolute quantification of microbial DNA should be used to confirm the reliability of microbiome sequencing data from low biomass samples. This study provides important insights into site-specific microbiomes and challenges when assessing the tumour microbiome.

Background

With globally rising rates of cancer, there is an urgent need for novel biomarkers for early detection, outcome prediction or targeted therapeutics to improve cancer survival rates. The evolution of high-throughput sequencing and intensified interest in the microbiome has resulted in an increased understanding of the microbiome in human health and its role in cancer. For instance, previous research has established a role of the gut microbiota in modulating therapeutic response to cancer treatments [1–6] and has demonstrated altered microbial communitas in those with lung cancer [7–9]. The microbial communities of the gut are the most well-studied human microbiome [10, 11]; however, interest in other body sites such as tumours and blood has intensified over recent years.

In recent years, several reports have identified intra-tumoural microorganisms or specific microbiota as key players in tumour development or progress [12–16]. In 2020, a collaborative multi-country study on over 1500 tumour and adjacent normal tissues used amplicon sequencing, quantitative PCR, and imaging techniques (immunohistochemistry and fluorescence in situ hybridization) to demonstrate the presence of bacterial DNA in tumours and that different proportions of bacterial taxa were noted in specific cancer types [17]. This work was followed by a meta-analysis that showed significant differences in the relative abundance of specific taxa between tumour and adjacent normal tissues [15]. However, tumour microbiome data has been inconsistent between studies, and much uncertainty remains about specific tumour microbes as biomarkers[12, 13, 15, 16, 18, 19] making it difficult to draw conclusions. Early studies in mouse models suggest that intra-tumoural microbiota may be due to translocation through the bloodstream [20], and one study suggested the potential for the human blood microbiome as a biomarker for diagnosing specific cancers [21]. However, this work has subsequently been retracted due to concerns about the genome database used and downstream data processing [14, 22]. Furthermore, microbial blood profiles on thousands of samples show no evidence for a consistent core blood microbiome of healthy humans [23]. Together, these findings bring into question the ability to identify a robust microbial signature in blood and tumour samples that could be used as a biomarker for early detection or prognosis of cancer outcomes.

Compared to microbially rich samples such as stool, the detection of microbiota in body sites such as tumours and blood poses additional challenges due to their low ratio of microbial to host DNA (low microbial biomass) and high risk of contamination. Bacteria can be introduced in the laboratory environment and sequencing process (e.g. sampling supplies/tools, laboratory reagents and kits), as well as introduction of biases from genome databases [14, 24, 25]. Recent work has further suggested that apparent microbial signals in low-biomass tissues may reflect technical artefacts, with metrics such as DNA fragment length proposed to distinguish true biological signal from background noise [26].

Given the ongoing debates/controversies, further research is needed to address current challenges in detecting reliable microbial signatures from samples of low microbial biomass. Thus, this work aimed to address current challenges in analyzing tumour and blood samples to identify cancer-specific microbial biomarkers by utilizing multiple microbiome techniques and datasets.

Material and Methods

Study Design and Data Sources

This observational prospective clinical study includes samples and data collected from 70 patients with non-small cell lung cancer (NSCLC) that were undergoing curative resection. The Research Ethics board at the Nova Scotia Health Authority granted ethics approval (REB# 1026340) and participants provided written informed consent prior to data/sample collection. Inclusion/exclusion criteria have been previously published [27].

Sample Collection

Tumour tissue, adjacent normal lung tissue, and blood were collected at the time of surgery. Saliva samples were collected preoperatively using a standardized in-hospital protocol as previously described [27]. Additionally, swabs of surgical instruments were collected and used as negative controls in the sequencing analysis. All samples were transported immediately to the laboratory for processing and storage. Tumour, adjacent tissues, saliva, and negative control samples were frozen at -80°C until microbial analysis. Blood samples were collected in ethylenediaminetetraacetic acid (EDTA) tubes, centrifuged at 1500×*g* for 10 minutes at 4^{O}C, then the buffy coat was

collected in 1.5 ml tubes and stored at -80°C until further analyses. Additional positive controls included an in-house human-derived stool control and the ZymoBIOMICS™ Microbial Community Standard (Zymo Research, cat #D6300).

Amplicon Sequencing, Bioinformatic and Statistical Analyses

Samples were profiled using multiple techniques at the Integrated Microbiome Resource (IMR) at Dalhousie University. DNA was extracted from saliva using the QIAamp PowerFecal Pro DNA Kit (QIAGEN) and the QIAGEN Blood and Tissue kit for all other sample types, as per the manufacturer's guidelines. Library preparation and sequencing were completed as originally described in Comeau et al.[28] and updated as in Comeau and Filloramo[29] and Comeau and Kwawukume [30], with the exception that the SPK3.0 kit was utilized on the PacBio Vega. Briefly, for amplicon sequencing, extracted DNA underwent PCR amplification for the full-length 16S rRNA gene using the 27F (AGRGTTYGATYMTGGCTCAG) and 1492R (RGYTACCTTGTTACGACTT) primer set and was subsequently sequenced on a Sequel II or Vega (PacBio) to generate HiFi reads. Due to the inability to generate full-length sequencing data and to overcome low microbial DNA in certain sample types, a nested PCR approach was applied. For this method, the full-length PCR products above were used as the template for a second round of PCR, using primers targeting the V6-V8 regions (A956F = TYAATYGGANTCAACRCC; A1401R = CRGTGWGTRCAAGGRGCA) and subsequently sequenced on the MiSeq (Illumina) using 2x300bp PE chemistry.

As outlined in the Microbiome Helper Standard Operating Procedure [28] (https://github.com/LangilleLab/microbiome_helper/wiki/Microbiome-Helper-2-Marker-gene-workflow), 16S rRNA gene sequencing data was processed using QIIME2 (version 2024.1) [31] to give estimates of alpha-diversity, beta-diversity, and relative abundance of individual amplicon sequence variants (ASVs). Full-length and nested 16S sequencing reads were denoised and trimmed using the DADA2 [32] and Deblur [33] pipelines, respectively. ASVs were classified taxonomically using the naive bayes scikit-learn classifier [34] trained on the full-length 16S rRNA gene SILVA v138 database [35]. ASVs were removed if they were classified as mitochondria or chloroplasts, were unclassified at the domain level, or had a frequency of less than 0.1% of the mean sample depth, and all samples with < 1,000 reads were removed.

All statistical analyses were conducted using R Version 4.4.0. Rarefied data at the ASV level were used for assessments of alpha- and beta-diversity metrics. Kruskal-Wallis tests and permutational multivariate analysis of variance (PERMANOVA; the adonis2 function with version 2.6–6.1 of the vegan package [36] with 1,000 permutations were used for alpha- and beta-diversity, respectively. All diversity plots were generated using ggplot2 and an alpha value of 0.05 was chosen for determining significance. All code used for analysis can be found at https://github.com/vdeclercq/DeClercq_2026_microbiome_tumour_blood_lung_cancer.git. Raw sequencing data is available in the NCBI SRA under project PRJEB105384.

Metagenomic Sequencing

A subset of samples were sequenced using the NextSeq2000 (Illumina) including: blood (n = 3), tumour (n = 3), adjacent tissue (n = 2) and saliva (n = 3) samples, and a positive control (stool sample). Paired-end reads (2x150 bp) were processed using KneadData v0.12.4 [37] with Trimmomatic v0.40 [38] and Bowtie 2 v2.5.4 [39] with a database containing the PhiX genome, the human genome (GRCh38 p14), and UniVec sequences (sequences for common vectors, adapters, linkers, and primers) to remove low quality and contaminating reads. We classified reads taxonomically using Kraken 2 v2.17.1 [40] with a custom database containing genomes from NCBI RefSeq V214 (all bacterial, archaeal and viral genomes as well as complete eukaryotic genomes and UniVec_Core sequences; 803 GB) built as described in [41] and a confidence threshold of 0. Kraken 2 taxonomic classifications were verified by mapping reads back to reference genomes using GeCoCheck v0.0.5, as described in [42]. We used the GeCoCheck script filter_kraken_outraw.py to keep

species-level taxonomic classifications with at least 1 read mapping to the reference genome and then KrakenTools [43] scripts were used to generate Kraken report files. Bracken [44] was run on the Kraken report files at the genus level with a read threshold of 10. The blood, tumour, adjacent tissue and positive control (stool) samples all had nested 16S sequencing that was performed on the same biological samples. We used a custom Python script to compare these matched pairs of samples and to also compare the saliva samples sequenced using metagenomics with three randomly-matched (using the Python package "random") nested 16S saliva samples. Because some genus names differ between the metagenomic NCBI taxonomic classifications and the 16S SILVA V138 taxonomic classifications, the genera were examined manually and matched (where possible) between the databases. Python packages pandas and matplotlib were used for plotting.

Digital PCR

A subset of study samples were analyzed by the QIAcuity digital PCR system (QIAGEN). Sample DNAs were quantified using the Quant-iT fluorescent method on the Qubit reader (both ThermoFisher). dPCR was then conducted with the PanBacteria1 universal 16S probe using the QIAcuity UCP (ultra-clean) Probe PCR kit in 12 uL reaction volumes in 8.5k Nanoplates.

Results

Samples were obtained from 46 female and 24 male patients with NSCLC between the ages of 52 and 82 years (mean age 68 years). Multiple sample types were collected from each participant and analyzed by numerous methods including full-length amplicon sequencing, nested amplicon sequencing, metagenomic sequencing, and dPCR (Fig. 1). The number of samples run by each method are detailed in Table 1.

Full-length Amplicon Sequencing

Initially, DNA from all sample types underwent full-length amplification of the 16S rRNA gene. After processing the data for read quality, low sample depth, contaminants and unclassified ASVs, all tumour, blood, and negative controls, as well as all but 1 adjacent sample had no remaining reads and were excluded for further analysis (Table 1). Conversely, almost all microbially rich samples (saliva and positive controls) were retained after processing, with the saliva profiles having typical oral bacteria, with the most abundant genera being *Streptococcus, Prevotella, Veilonella, Rothia*, and *Haemophilus* (**Figure S1**).

Table 1
Sample Characteristics

	Adjacent	Blood	Neg CTL	Pos CTL	Saliva	Tumour	Overall
16S Full-Length PCR+Sequencing							
Samples, n	26	12	7	2	69	14	130
# of sequences/ sample, mean (range)	660 (1-5919)	248 (92–450)	65 (7-224)	5062 (4627–5497)	46286 (75-175952)	315 (0-1634)	41081 (7-175952)
Samples passed filtering, n (%)	1 (4)	0 (0)	0 (0)	2 (100)	68 (99)	0 (0)	71 (55)
16S Nested PCR+Sequencing							
Samples, n	60	51	6	6	18	47	188
# of sequences/ sample, mean (range)	44949 (9-139171)	42471 (1-139222)	51066 (3-73520)	19099 (91-62233)	40101 (13101–86152)	348088 (7-102667)	40648 (1-139222)
Samples passed filtering, n (%)	54 (90)	40 (78)	5 (83)	5 (83)	18 (100)	41 (87)	163 (87)
Digital PCR							
Samples, n	6	6	9	3	6	6	36
16S (cp/uL), mean (range)	6 (2–10)	39 (0–85)	7 (0–32)	29165 (23494->32000)	28697 (21595->32000)	21 (3–92)	7226 (0->32000)
Metagenomic Sequencing							
Samples, n	2	3	0	1	3	3	12
# of sequences/ sample, mean (range)	32,348,220 (29,656,604 – 35,039,836)	27,926,124 (22,127,566 – 32,844,494)	-	18,066,562	25,376,375 (17,159,020–34,063,930)	31,744,793 (30,210,676 – 34,678,066)	28,158,740 (17,159,020–35,039,836)
% Human reads *	99.9	99.8	-	3.5	44.9	99.9	78.1
# reads passed filtering, and mapped to	159	363	-	6,114,052	3,103,140	162	-

	Adjacent	Blood	Neg CTL	Pos CTL	Saliva	Tumour	Overall
16S Full-Length PCR+Sequencing							
a genus (mean)							

* Note that this percentage contains all reads that mapped to the Bowtie 2 reference database that contained the GRCh38 p14 human genome, phiX genome, and UniVec sequences.

Nested Amplicon Sequencing

Using nested amplicon sequencing, low numbers of ASVs were observed in blood (median 3, IQR 1,4), tumour (median 5, IQR 2,9), and adjacent (median 5, IQR 3,10) tissues as well as in negative controls (median 23, IQR 4,28) (Fig. 2A). The lowest number of ASVs was observed in blood (median 3, IQR 1,4) and was similar to tumour. In contrast, saliva (median 64, IQR 52,70) and stool positive control (median 78, IQR 19,79) samples had more ASVs than blood, adjacent and tumour tissues. The ZymoMock positive control had 19 ASVs in both samples, which match the expected genera (Fig. 3). Beta-diversity analysis showed that the community composition of saliva samples was distinct from blood, adjacent and tumour tissues (Fig. 2B), while all the latter sample types clearly clustered together and had overlapping communities with the negative controls.

Investigation of the top 10 most abundant genera in each sample highlighted variability per sample and across sample types (Fig. 3). While there were many shared taxa among the blood, tumour, and adjacent tissues, many of these taxa also overlapped with the negative control samples (Fig. 3). For instance, the two genera *Cutibacteria* and *Methylobacterium-Methylorubrum* were in the top 5 most abundant genera in the negative control, tumour, blood and adjacent tissue samples. There was also substantial variability from sample to sample within these sample types illustrated by the low prevalence of the top 10 most abundant genera (Table 2**)**. For example, the most abundant genera in blood were *Massilia* followed by a member of the Rhizobiaceae family, but they were only found in 25% and 18% of blood samples, respectively. On the other hand, the most abundant genera in saliva samples were *Prevotella* and *Haemophilus*, both found in 100% of saliva samples.

The low prevalence of genera in some sample types such as the blood and tumour samples suggests a lack of common core microbiome in these sample types and questions the reliability of using them to detect consistent biomarkers. To assess sequencing reproducibility within these sample types, a subset of samples were sequenced in duplicate (i.e. technical replicates of the PCR and sequencing of the same extracted DNA). Similarity in community composition of saliva and positive control samples (human stool and ZymoBIOMICS™ Microbial Community Standard) showed that their duplicates cluster together whereas the duplicates for the other sample types are more disparate (Fig. 4). The median weighted UniFrac distances between duplicates for each sample type support this with saliva (0.064) and positive control (0.027), being much less than the adjacent (0.496), blood (0.655), and tumour (0.546) samples (Fig. 4B). These results suggest duplicate blood, tumour, and adjacent tissue samples are highly dissimilar in community composition and this is clearly visible when comparting the top 10 genera across duplicates (Fig. 4C).

Digital PCR

Considering sequencing cannot adequately quantify absolute abundance of bacterial DNA, dPCR was used to measure 16S rRNA gene copies. The saliva and positive controls (both human stool and ZymoBIOMICS standard) had much higher copy numbers of the 16S gene (mean 28,697 and 29,165 cp/µL, respectively) compared with the blood (39 cp/µL), tumour (21 cp/µL), adjacent tissue (6 cp/µL), and negative control samples (7 cp/µL) (Fig. 5).

Metagenomic Sequencing (MGS)

Finally, a subset of samples were analyzed by MGS to compare bacterial reads observed with the nested amplicon sequencing. We found that ~ 99.9% of reads in adjacent, blood, and tumour samples were human, phiX or sequencing-related vectors, while this was the case for only ~ 45% or 3.5% of reads in saliva or stool positive control samples, respectively (Table 1). After applying GeCoCheck, where metagenomic taxonomic classifications are verified by mapping reads back to reference genomes [GeCoCheck, as above], only 12 genera (*Bacillus*, *Listeria*, *Burkholderia*, *Ligilactobacillus*, *Campylobacter*, *Staphylococcus*, *Lactobacillus*, *Escherichia*, *Clostridium*, *Cutibacterium*, *Acinetobacter* and *Paraburkholderia*) were identified in total in the adjacent, blood, and tumour samples, with an average of 4.5 [adjacent], 5.7 [blood], or 3.7 [tumour] genera per sample and only 159 [adjacent], 363 [blood] or 162 [tumour] reads classified per sample, respectively (Table 1). Conversely, 349 and 177 genera per sample were observed in stool positive control and saliva samples, respectively. Overlap between the genera identified in these samples by nested 16S and metagenomic sequencing was minimal or non-existent, while saliva samples from different participants showed much more similarity between the different sequencing methods (Fig. 6).

Discussion

There is great interest in identifying microbial biomarkers that can enhance early cancer detection and improve patient outcomes. While much work has been done on the gut microbiome, the literature on microbiomes from other body-sites and their potential for cancer biomarkers is still evolving. Findings from this study could not produce consistent or reliable microbiome profiles from blood, tumour, and adjacent tissues using current sequencing methodologies. Further, this study illustrates how sequencing can produce false signals from low microbial biomass samples that may go unnoticed without proper controls. On the other hand, the same methodologies produce reproducible results from samples that are bacterially rich, such as saliva, and warrant further investigation for use as targets for disease diagnostics or monitoring.

For low microbial biomass samples, additional methods and techniques may be necessary for auxiliary investigations. Early work by Nejman *et al.* used multiple methods including amplicon sequencing, quantitative PCR, and imaging techniques (immunohistochemistry and fluorescence *in situ* hybridization) to demonstrate the presence of bacterial DNA in tumours (Nejman et al., 2020). However, even in these studies, bacterial DNA was not found in all tumours; only 17% of lung tumours had more bacteria than the 99th percentile of the negative controls when assessed by qPCR using 16S rRNA gene primers. More recently, Sun *et al.* applied both culturomics and 16S rRNA gene sequencing (V3-V4 region) to identify tumor-resident microorganisms from 18 NSCLC patients [45]. Lung homogenates were cultured and 12 species of bacteria were identified from the purified isolates. More than half were from the *Staphylococcus* genus; others included *Bacillus, Proprionibacterium, Acinetobacter, Microbacterium* and *Tessaracoccus*. Of these, only *Bacillus* was identified in the top 10 most abundant genera by both culture and culture-independent methods, and this was only observed in one sample pair. Furthermore, the microbial profiles varied widely between tumor samples, similar to the current study. The heterogeneity in microbial composition of lung tissue observed by Sun *et al.* as well as the current study, challenges the idea of a core microbiome in these types of samples and questions the practicality of current methodologies in identifying robust microbial signatures from these sample types.

The microbial heterogeneity of these low biomass samples has been demonstrated previously. For example, Tan *et al.* reported no evidence for a consistent core blood microbiome after analyzing close to 10,000 blood samples from healthy individuals [23]. Similar to the current study, they showed that the prevalence of microbiota varies greatly between samples and the most prevalent bacteria, *Cutibacterium acnes*, was detected in < 5% of samples. *Cutibacterium* was also detected in the blood, tumour, adjacent tissue, and negative control samples in the current study. Gihawi *et al.*

[18] also showed that only microbial profiles from colorectal cancer - out of 22 cancer types, including lung - could be used to reliably distinguish between tumour types.

Although technical replicates are not traditionally utilized in microbiome studies because of the consistent signal in normal high microbial load samples, this study suggests their benefit in low biomass samples like blood, tumour, and even other non-human microbiome applications. Inconsistency across technical replicates highlights the stochastic nature of PCR and sequencing when template microbial DNA is not in high enough abundance. While nested PCR can produce sequencing data in some of these challenging samples, its use should be cautioned in samples with low microbial biomass. An additional recommendation is that dPCR may be able to provide an early warning metric that can be used to flag samples that may not provide reliable signal. Lastly, this study strongly supports the use of negative controls in low microbial biomass samples to try and distinguish biologically relevant microbes from technical artifacts, in agreement with previous recommendations [46, 47].

The primary limitation of this study is a relatively small sample size in a single clinical cohort of a specific cancer. Others have shown lack of clinical signal in lung cancer microbiomes from much larger cohorts (McElderry, et al), and this study supports that, but focuses more on the underlying sequencing approach and consistency of analyses as the driving reason behind that lack of clinical signal. Expanding the technical approaches employed in this study to other cohorts and other cancers and body sites would be beneficial to determine the best practices for identifying reliable biomarkers in human microbiome studies from low biomass samples.

Conclusions

Together this data shows the challenges in using current high-throughput sequencing technologies to identify robust microbial signatures from low microbial biomass samples. With current methods, resultant microbiome profiles appear indiscernible from background noise inherent to the collection and processing environments. This study provides important insights into the current issues of working with specific microbiomes of low biomass and emphasizes the importance of using multiple controls and methods when assessing low biomass samples as cancer biomarkers.

Declarations

Ethics approval and consent to participate

The Research Ethics board at the Nova Scotia Health Authority granted ethics approval (REB# 1026340) and participants provided written informed consent prior to data/sample collection.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This project was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (2016–05039) and the Dalhousie Medical Research Foundation (DMRF) Health Influencers Grant (R35846). None of the funding sources took part in the design, collection, analysis, interpretation of the data, or writing the manuscript.

Author Contribution

ML, AW, and VD contributed to the study concept and design. RM and AW contributed to the sample acquisition. AC, AK, NP, DQ, RW, ML, and VD contributed to data generation, analysis, and interpretation of data. VD wrote the first draft and substantively revised it based on feedback from other authors. All authors reviewed the manuscript and approved the final version to be published.

Acknowledgement

The authors thank all the study participants who donated their personal health information and biological samples to this project.

Data Availability

Code used for analysis can be found at https://github.com/vdeclercq/DeClercq_2026_microbiome_tumour_blood_lung_cancer.git. Raw sequencing data is available in the NCBI SRA under project PRJEB105384.

References

1. Vétizou M, Pitt JM, Daillère R, Lepage P, Waldschmitt N, Flament C, et al. Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. Science (1979). 2015;350:1079–84. https://doi.org/10.1126/science.aad1329.
2. Sivan A, Corrales L, Hubert N, Williams JB, Aquino-Michaels K, Earley ZM, et al. Commensal Bifidobacterium promotes antitumor immunity and facilitates anti–PD-L1 efficacy. Science. 2015;350:1084. https://doi.org/10.1126/SCIENCE.AAC4255.
3. Routy B, Le Chatelier E, Derosa L, Duong CPM, Alou MT, Daillère R, et al. Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. Science (1979). 2018;359:91–7. https://doi.org/10.1126/science.aan3706.
4. Cascone T, Leung CH, Weissferdt A, Pataer A, Carter BW, Godoy MCB, et al. Neoadjuvant chemotherapy plus nivolumab with or without ipilimumab in operable non-small cell lung cancer: the phase 2 platform NEOSTAR trial. Nat Med. 2023;29:593–604. https://doi.org/10.1038/S41591-022-02189-0.
5. Morita A, Ichihara E, Inoue K, Fujiwara K, Yokoyama T, Harada D, et al. Impacts of probiotics on the efficacies of immune checkpoint inhibitors with or without chemotherapy for patients with advanced non-small-cell lung cancer. Int J Cancer. 2024. https://doi.org/10.1002/ijc.34842.
6. Alkan Şen G, Şentürk Öztaş N, Değerli E, Can G, Turna H, Özgüroğlu M. Effect of antibiotic treatment on immune checkpoint inhibitors efficacy in patients with advanced non-small cell lung cancer. Lung Cancer. 2023;184. https://doi.org/10.1016/j.lungcan.2023.107347.
7. Derosa L, Routy B, Thomas AM, Iebba V, Zalcman G, Friard S, et al. Intestinal Akkermansia muciniphila predicts clinical response to PD-1 blockade in patients with advanced non-small-cell lung cancer. Nat Med. 2022;28:315–24. https://doi.org/10.1038/S41591-021-01655-5.
8. Shoji F, Yamaguchi M, Okamoto M, Takamori S, Yamazaki K, Okamoto T, et al. Gut microbiota diversity and specific composition during immunotherapy in responders with non-small cell lung cancer. Front Mol Biosci. 2022;9. https://doi.org/10.3389/FMOLB.2022.1040424.

9. Wei YF, Huang MS, Huang CH, Yeh YT, Hung CH. Impact of Gut Dysbiosis on the Risk of Non-Small-Cell Lung Cancer. Int J Environ Res Public Health. 2022;19. https://doi.org/10.3390/IJERPH192315991.
10. Zitvogel L, Derosa L, Routy B, Loibl S, Heinzerling L, de Vries IJM, et al. Impact of the ONCOBIOME network in cancer microbiome research. Nature Medicine 2025 31:4. 2025;31:1085–98. https://doi.org/10.1038/s41591-025-03608-8.
11. Piccinno G, Thompson KN, Manghi P, Ghazi AR, Thomas AM, Blanco-Míguez A, et al. Pooled analysis of 3,741 stool metagenomes from 18 cohorts for cross-stage and strain-level reproducible microbial biomarkers of colorectal cancer. Nature Medicine 2025 31:7. 2025;31:2416–29. https://doi.org/10.1038/s41591-025-03693-9.
12. Repass J, Iorns E, Denis A, Williams SR, Perfito N, Errington TM. Replication Study: Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. Elife. 2018;7:e25801. https://doi.org/10.7554/eLife.25801.
13. Wen Q, Wang S, Fu S, Zhou X, Min Y, Lang J, et al. Intratumoral microbiota composition in women's cancers: a systematic review and meta-analysis. Front Oncol. 2025;15:1544786. https://doi.org/10.3389/fonc.2025.1544786.
14. Gihawi A, Ge Y, Lu J, Puiu D, Xu A, Cooper CS, et al. Major data analysis errors invalidate cancer microbiome findings. mBio. 2023;14. https://doi.org/10.1128/MBIO.01607-23/SUPPL_FILE/MBIO.01607-23-S0007.XLSX.
15. Najafi S, Abedini F, Azimzadeh Jamalkandi S, Shariati P, Ahmadi A, Gholami Fesharaki M. The composition of lung microbiome in lung cancer: a systematic review and meta-analysis. BMC Microbiol. 2021;21:315. https://doi.org/10.1186/s12866-021-02375-z.
16. McElderry JP, Zhang T, Zhao W, Hoang PH, Anyaso-Samuel S, Sang J, et al. Microbiome analysis of 940 lung cancers in never-smokers reveals lack of clinically relevant associations. Nature Communications 2025 17:1. 2025;17:192-. https://doi.org/10.1038/s41467-025-66780-y.
17. Nejman D, Livyatan I, Fuks G, Gavert N, Zwang Y, Geller LT, et al. The human tumor microbiome is composed of tumor type–specific intracellular bacteria. Science (1979). 2020;368:973–80. https://doi.org/10.1126/science.aay9189.
18. Gihawi A, Wood HM, Clark J, O'Grady J, Eeles RA, Wedge DC, et al. The landscape of microbial associations in human cancer. Sci Transl Med. 2025;17:6166. https://doi.org/10.1126/SCITRANSLMED.ADS6166.
19. Gihawi A, Cooper CS, Brewer DS. Caution regarding the specificities of pan-cancer microbial structure. Microb Genom. 2023;9. https://doi.org/10.1099/MGEN.0.001088.
20. Abed J, Maalouf N, Manson AL, Earl AM, Parhi L, Emgård JEM, et al. Colon Cancer-Associated Fusobacterium nucleatum May Originate From the Oral Cavity and Reach Colon Tumors via the Circulatory System. Front Cell Infect Microbiol. 2020;10. https://doi.org/10.3389/FCIMB.2020.00400.
21. Poore GD, Kopylova E, Zhu Q, Carpenter C, Fraraccio S, Wandro S, et al. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature. 2020;579:567–74. https://doi.org/10.1038/s41586-020-2095-1.
22. Poore GD, Kopylova E, Zhu Q, Carpenter C, Fraraccio S, Wandro S, et al. Retraction Note: Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature. 2024;631:694. https://doi.org/10.1038/s41586-024-07656-x.
23. Tan CCS, Ko KKK, Chen H, Liu J, Loh M, Chia M, et al. No evidence for a common blood microbiome based on a population study of 9,770 healthy humans. Nat Microbiol. 2023;8:973–85. https://doi.org/10.1038/s41564-023-01350-w.
24. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. Gut Pathog. 2016;8:24. https://doi.org/10.1186/s13099-016-0103-7.
25. Selway CA, Eisenhofer R, Weyrich LS. Microbiome applications for pathology: challenges of low microbial biomass samples during diagnostic testing. Journal of Pathology: Clinical Research. 2020;6:97–106. https://doi.org/10.1002/cjp2.151.

26. Zhang Y, Mead EA, Ni M, Ksiezarek M, Liu Y, Cao L, et al. Critical assessment of intratumor and low-biomass microbiome using long-read sequencing. bioRxiv. 2026;:2026.02.02.703393. https://doi.org/10.64898/2026.02.02.703393.
27. DeClercq V, Murphy RE, Zinck N, Langille MGI, Wallace A. Association between the oral microbiome and clinical indicators of resectable non-small cell lung cancer (NSCLC). J Oral Microbiol. 2026;18:2675065. https://doi.org/10.1080/20002297.2026.2675065.
28. Comeau AM, Douglas GM, Langille MGI. Microbiome Helper: a Custom and Streamlined Workflow for Microbiome Research. mSystems. 2017;2. https://doi.org/10.1128/mSystems.00127-16.
29. Comeau AM, Filloramo G. Preparing multiplexed 16S/18S/ITS amplicon SMRTbell libraries with the Express TPK2.0 for the PacBio Sequel2 v1. protocols.io. 2023. https://doi.org/10.17504/PROTOCOLS.IO.J8NLKK89WL5R/V1.
30. Comeau AM, Kwawukume A. Preparing multiplexed 16S/18S/ITS amplicons for the Illumina MiSeq. protocols.io. 2023. https://doi.org/10.17504/PROTOCOLS.IO.4R3L277K3G1Y/V1.
31. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature biotechnology. 2019;37:852–7. https://doi.org/10.1038/s41587-019-0209-9.
32. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods 2016 13:7. 2016;13:581–3. https://doi.org/10.1038/nmeth.3869.
33. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, et al. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. mSystems. 2017;2. https://doi.org/10.1128/mSystems.00191-16.
34. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-Learn: Machine Learning in Python. J Mach Learn Res. 2011;12 null:2825–2830.
35. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, et al. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res. 2014;42:D643–8. https://doi.org/10.1093/NAR/GKT1209.
36. Oksanen J, Simpson G, Blanchet F, Kindt R, Legendre P, Minchin P, et al. vegan: Community Ecology Package, version 2.6-4. vegan_2.6-4.tar.gz. 2022. https://cran.r-project.org/web/packages/vegan/index.html. Accessed 30 Jan 2025.
37. Huttenhower Lab. Kneaddata. https://huttenhower.sph.harvard.edu/kneaddata. 2016.
38. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114. https://doi.org/10.1093/BIOINFORMATICS/BTU170.
39. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9. https://doi.org/10.1038/nmeth.1923.
40. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20:257. https://doi.org/10.1186/s13059-019-1891-0.
41. Wright RJ, Comeau AM, Langille MGI. From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools. Microb Genom. 2023;9:000949. https://doi.org/10.1099/MGEN.0.000949.
42. Wright RJ, Fisher BR, Comeau AM, Langille MGI. From classification to confirmation: verifying taxonomic classifications by mapping metagenomic reads to reference genomes. Microb Genom. 2026;12:001739. https://doi.org/10.1099/MGEN.0.001739.
43. Lu J, Rincon N, Wood DE, Breitwieser FP, Pockrandt C, Langmead B, et al. Metagenome analysis using the Kraken software suite. Nature Protocols 2022 17:12. 2022;17:2815–39. https://doi.org/10.1038/s41596-022-00738-y.
44. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. PeerJ Comput Sci. 2017;3:e104. https://doi.org/10.7717/PEERJ-CS.104.

45. Sun Y, Tong J, Tan Y, Liu Y, Li D, Yu H, et al. Isolation, culture, and in vitro functional validation of intratumor bacteria from lung cancer patients. BMC Microbiol. 2025;25:558. https://doi.org/10.1186/s12866-025-04237-4.
46. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. Trends Microbiol. 2019;27:105–17. https://doi.org/10.1016/j.tim.2018.11.003.
47. Kim D, Hofstaedter CE, Zhao C, Mattei L, Tanes C, Clarke E, et al. Optimizing methods and dodging pitfalls in microbiome research. Microbiome. 2017;5. https://doi.org/10.1186/S40168-017-0267-5.

Table 2

Table 2 is available in the Supplementary Files section.

Figures

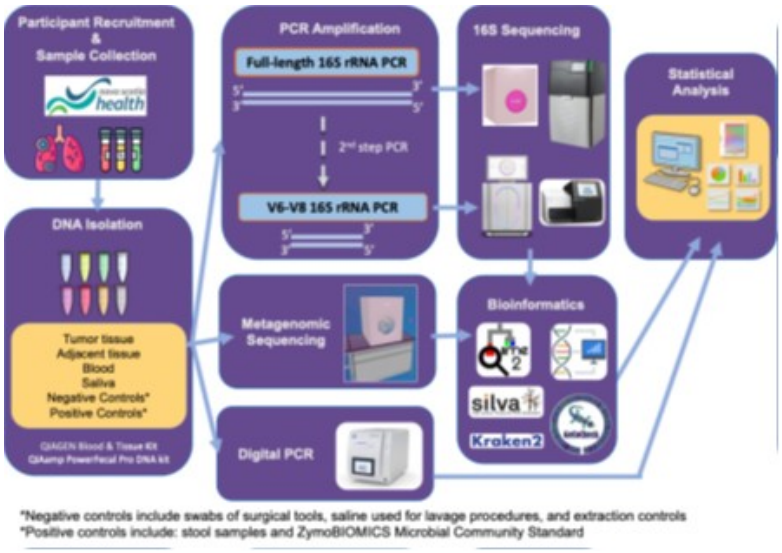


Figure 1

Overview of sample collection, microbiome profiling, and bioinformatic analyses.

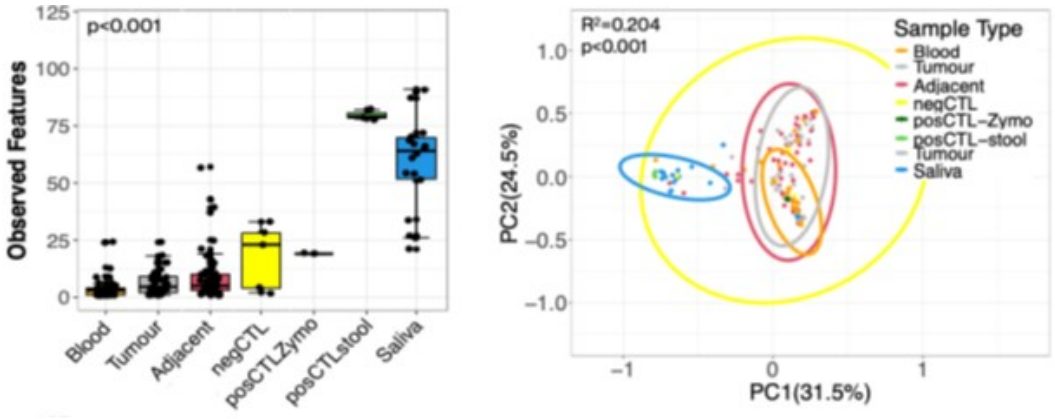


Figure 2

Microbial richness and differences in community composition between sample types that underwent nested amplicon sequencing. (**A**) Alpha-diversity is represented by observed ASVs (richness). P-value was obtained from the Kruskal-

Wallis rank-sum test. (**B**) Beta-diversity analyses represented by Principal Coordinates Analysis plots based on weighted UniFrac distances. R^2 and p-values are the result of a PERMANOVA test.

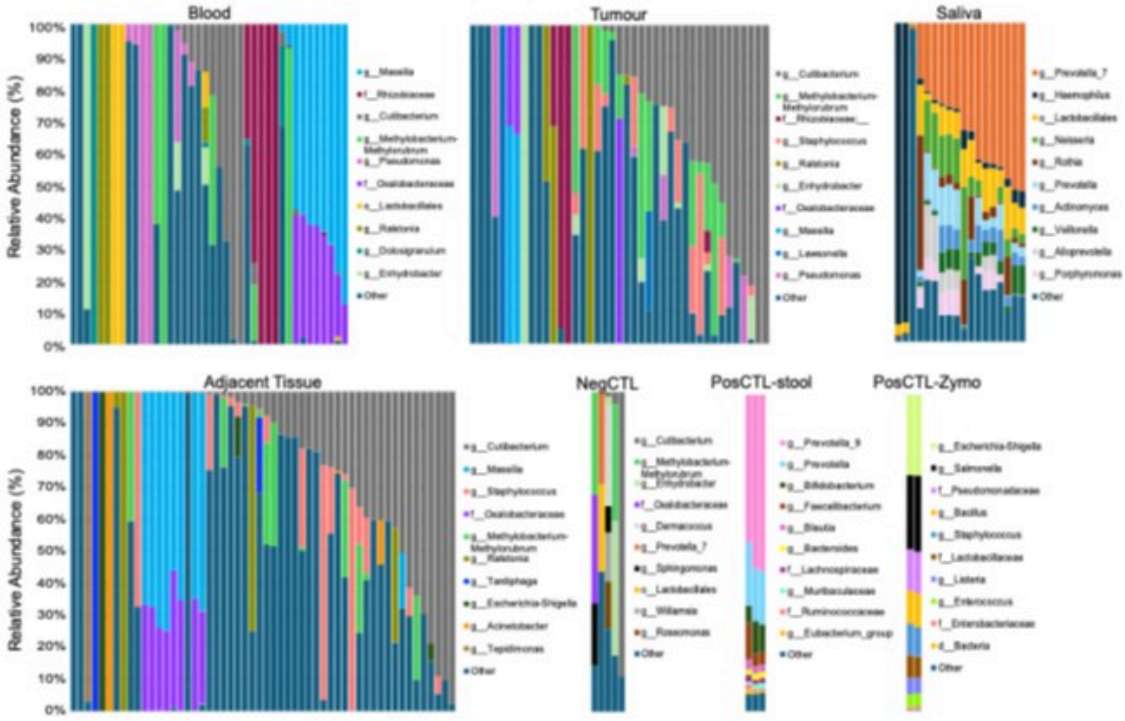


Figure 3

Stacked bar chart of the top 10 most abundant genera in different sample types that underwent nested amplicon sequencing. Each vertical bar represents a single sample and samples in each panel are ordered by the most abundant genera in that sample type.

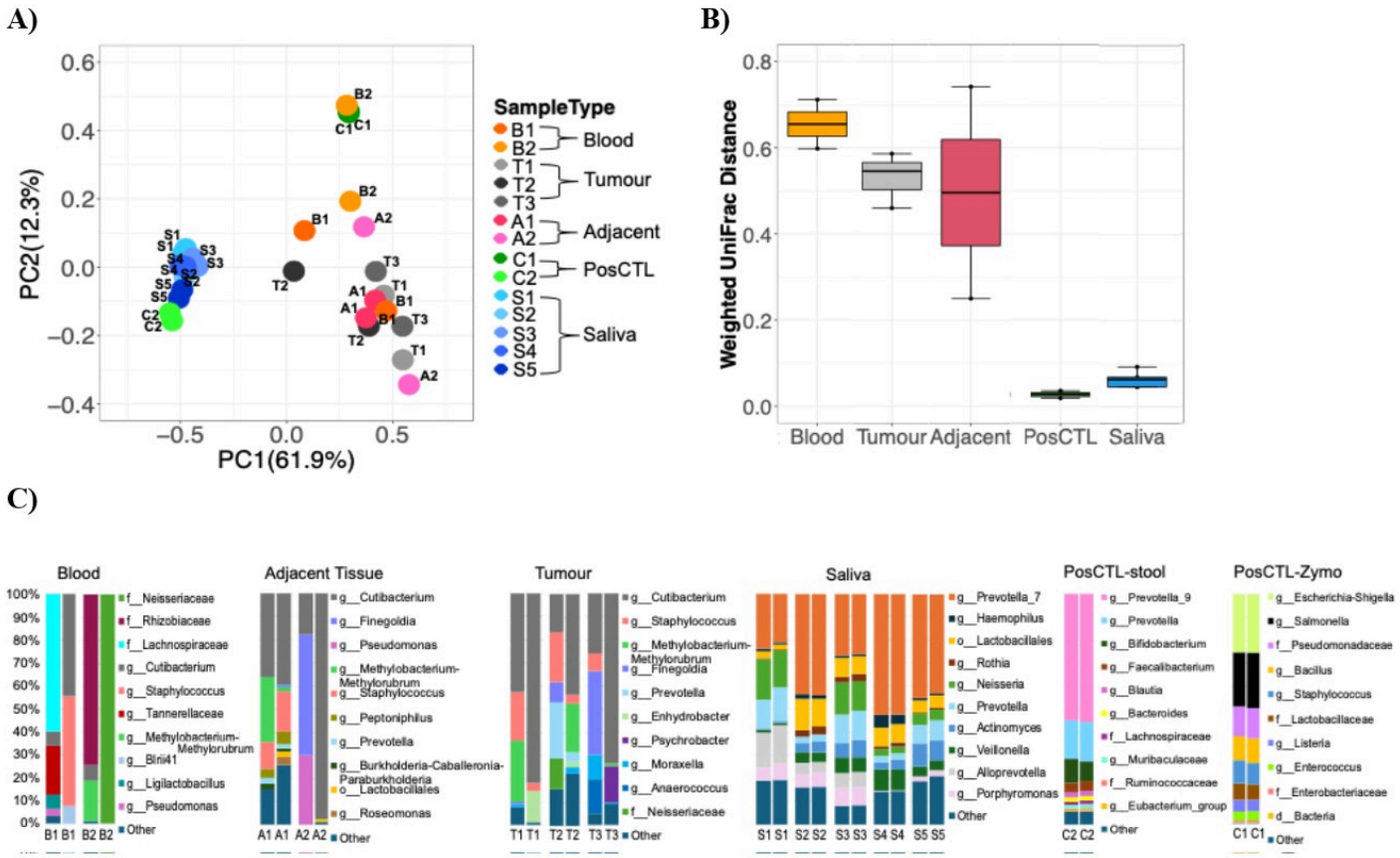


Figure 4

Similarity in community composition between duplicate samples. (**A**) Principal Coordinates Analysis plots based on weighted UniFrac distances between duplicate samples. C1 represents the ZymoBIOMICS™ Microbial Community Standard and C2 represents the human stool positive control. (**B**) Median weighted UniFrac distances between duplicate samples for each sample type. **(C)** Stacked bar chart of a subset of samples that were sequenced in duplicate. Charts show the top 10 most abundant genera in duplicates that underwent nested amplicon sequencing. Technical duplicates are ordered each other.

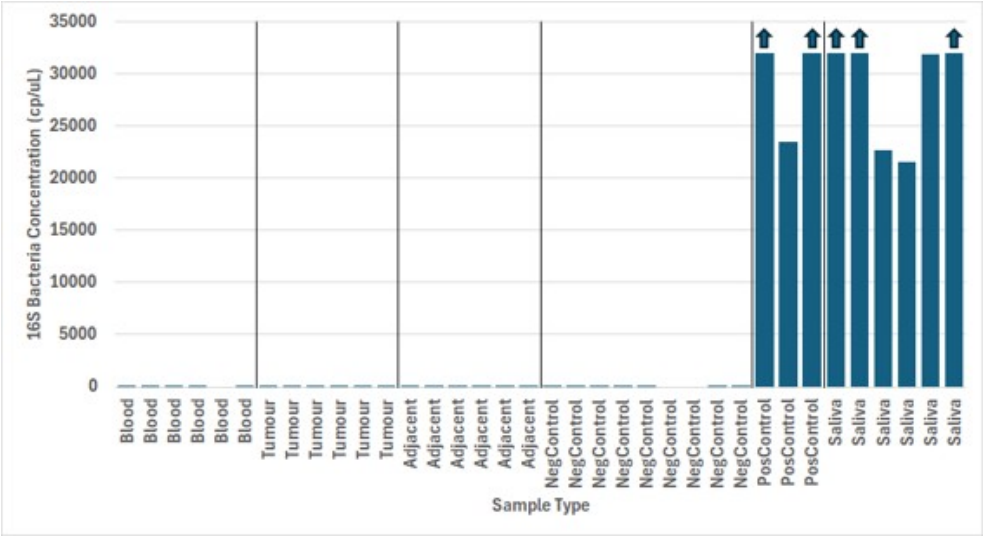


Figure 5

Concentration of bacteria across sample type. Concentration of bacteria was quantified by digital PCR and is represented as the copy number (cp) of 16S rRNA gene per microlitre (mL) per reaction. The arrows indicate samples with >32000 cp/uL (beyond the limit of detection).

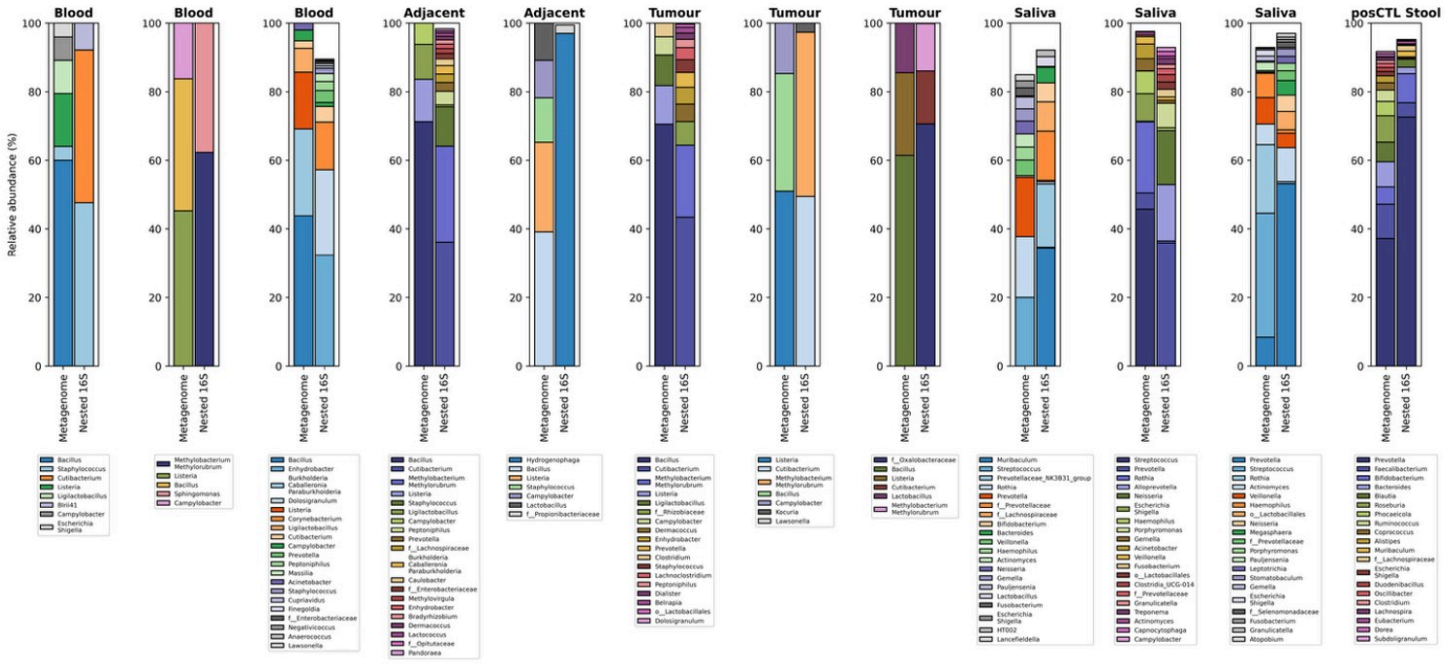


Figure 6

Shared genera across samples sequenced using metagenomics and nested 16S amplicons. Blood, tumour, adjacent and posCTL samples use the same biological material across sample pairs while MGS saliva samples are randomly matched with nested 16S samples. Stacked bar plots show relative abundance within each sample and are filtered to include only the top 20 genera, where necessary. Note that colour schemes are separate for each subplot.

Supplementary Files

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

- Table2.docx
- FigureS1.docx