

Higher standards needed for reports of microbial species in human cancers

Steven Salzberg^{1,*,#}, Minghao Chia^{2,3,*}, Alfred Tay⁴, Noel de Miranda⁵, Vincent Smit⁵, Jelle Wesseling⁵, Rafael Irizarry⁶, Barry Marshall⁴, Eske Willerslev⁷, Jacques Neefjes^{8,#}, Niranjan Nagarajan^{2,9,#}

¹Center for Computational Biology and Departments of Biomedical Engineering, Computer Science, and Biostatistics, Johns Hopkins University, Baltimore, Maryland 21211, United States

²Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), Singapore, Republic of Singapore

³Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), Singapore, Republic of Singapore

⁴Marshall Centre, School of Biomedical Sciences, The University of Western Australia, Perth 6009, Australia

⁵Department of Pathology, Leiden University Medical Center, Leiden 2333 ZA, Netherlands

⁶Department of Data Science, Dana Faber Cancer Institute, Boston, Massachusetts 02215, United States

⁷Centre for GeoGenetics, University of Copenhagen, Øster Voldgade 5-7, DK-1350, Denmark

⁸Department of Cell and Chemical Biology, Leiden University Medical Center, Leiden 2333 ZA, Netherlands

⁹Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117596, Singapore

*Co-first authors

#Corresponding authors

Preface

The presence of microbiota in human tumors has been reported by many groups based on bioinformatic searches of DNA sequence databases. The source of these microbial sequences in atypical sites is difficult to determine with certainty, but they could derive from nucleic acids acquired during the sampling, storage/handling and processing of samples, as observed earlier in studies of ancient DNA. Another source of spurious microbial signals is contamination of microbial reference genomes causing human reads to be wrongly classified. We present a series of quality controls and validation approaches summarized in a checklist, that should be considered in any study attempting to find microbes in tumor tissues. This provides a constructive path forward to improve the rigor and standards of studies exploring the role of microbes in human cancers.

Introduction

Multiple sequencing-based studies have reported the presence of bacterial, fungal, viral, and archaeal DNA in human tissues across diverse cancer types¹⁻⁶. In many of these studies, microbes were detected by searching through the non-human DNA (or RNA) sequences

generated by high-throughput sequencing experiments, and comparing those sequences to microbial genome databases using various computational tools^{7–9}. A recent retraction¹⁰ as well as the discussion it generated has raised the question of whether some of these findings need to be re-examined, and is being debated intensely in the fields of genomics, microbiology, immunology and oncology^{11–14}. In this perspective, we discuss the need for more rigorous standards for reporting the presence of microbes in human cancers, to avoid incorrect findings that undermine scientific literature and lead to futile investments of time and research effort.

Dramatic reductions in the cost of sequencing over the last decade and the corresponding availability of large cancer genomic datasets¹⁵ has enabled studies that have sought to explore the presence of microbes in diverse cancer types. Some of these efforts have helped to shortlist specific microbial species or strains for further molecular characterization of how these microbes could contribute to disease phenotypes such as pro-inflammatory metabolite production by *Fusobacterium nucleatum* C2¹⁶ or PI3K signalling by *Peptostreptococcus anaerobius*¹⁷. However, many cancer microbiome studies rely primarily or solely on the analysis of sequencing data to arrive at their conclusions, and often without sufficient controls (reviewed in Cullin et al¹⁸ and Knippel et al¹⁹). While various higher-quality reports have sought to address concerns on the reliability of findings by sequencing negative controls^{20–22}, detecting microbes with *in situ* imaging^{22–24}, isolating microbes from tissues^{20,21}, and detecting microbial antigens within tumors²⁵, the application of these ideas is by no means uniform, and correspondingly studies vary highly in terms of their quality and reliability of findings. Even some relatively well-controlled investigations can still suffer from one or more shortcomings (and often all three), including: (a) data from positive and negative control experiments were limited or missing^{24,26}, (b) the identified organisms have metabolic requirements and niches that are far removed from the body sites they were reported in²³, and (c) the microbial species reported within tumors were not independently confirmed through other assays⁵, which raises questions about the origin of these non-human DNA sequences. For example, despite efforts to control for contaminants, Pushalkar et al²³ reported high prevalence and abundance of *Elizabethkingia* (typically found in soil and hospital environments²⁷) in human pancreatic ductal adenocarcinomas (PDA), which was not subsequently observed in a separate larger dataset²⁸. This begs the question, where did the *Elizabethkingia* DNA come from in the first study? Was it a real signal or simply an environmental contaminant? Overall, such cases underscore the importance of establishing standardized checks and guidelines in the field to enhance the reproducibility and reliability of findings.

Of note, the idea that both viruses and bacteria can contribute to the development of cancers by triggering DNA mutations, or by activating oncogenic signaling pathways in pre-malignant cells is well known^{29,30}. For bacteria, this is most established in gastric tumors³¹, with increasing evidence emerging for colorectal cancers³² and gallbladder cancer³³, where tissues are exposed to abundantly present microbes in the gastrointestinal tract. Tissue-resident bacteria can thus provide one step in a multi-step process of transformation leading to

tumors¹⁶. Furthermore, emerging evidence suggests that bacteria such as *Bacteroides*, *Bifidobacterium* and *Fusobacterium* in the colon can infiltrate tumors^{16,34,35}, or produce antigens or metabolites that affect the outcome of immunotherapy^{36–38}. Similarly, gut microbial production of metabolites and cross-talk with the immune system could influence the risk for cancer and other diseases in distant organs^{39,40}.

Although the role of microbial colonization and function in some cancers is clearly important, the notion of *cancer microbiomes* in human tumor tissues at places not in contact with the external environment is fundamentally distinct. The term *microbiome* is generally used to refer to a community of microorganisms (e.g. bacteria, fungi, viruses) that colonize and interact in an environment. *Cancer microbiome*, in turn, is used to refer to the analogous concept of multi-species communities detected in a tumor tissue that presumably have similar interactions. In the human body, microbiomes are present on epithelial surfaces such as the intestines, the skin, and the oral cavity. However, internal organs and tissues are believed to rarely harbor microbes, with some entirely devoid of them (e.g. brain⁴¹ and blood⁴²) except during an acute infection. In fact, preventing infiltration of microbes in tissues is a major task of the immune system, which responds with strong inflammatory responses to eliminate microbes when they appear. Failure to do so may result in sepsis that is often lethal to the host.

Challenges faced by the emerging field of cancer microbiome studies

In recent years, a series of publications have reported the presence of microbiomes that include a wide range of bacterial, fungal, and viral species in tumor tissues from organs that should be microbiome-free^{43–45}. These studies were primarily based on bioinformatic analyses of DNA/RNA sequences from tumors and sometimes from matched normal tissues for microbe discovery. In a typical experiment, human DNA/RNA sequences were removed from shotgun-sequenced data for tissue samples, and the remaining sequences were aligned to databases containing bacterial, fungal, and viral genomes. When source material was collected from preserved or banked samples, as they generally were, the authors were often unable to rigorously test whether the microbes identified computationally could be detected through other means inside the original tissues, without the risk for contamination^{5,10}.

In addition, signatures for the activation of the immune system have rarely been tested, which is concerning because lipopolysaccharides (LPS) and bacterial DNA/RNA are strong inducers of immune responses, a point that is generally ignored in such studies. LPS in the brain can induce the expression of an array of chemokines to attract neutrophils, monocytes and T cells⁴⁶. LPS can then activate many immune cells including human CD8+ T cells⁴⁷. Local LPS enrichment in tissue is thus an excellent and specific immune stimulator. Of note, higher doses of LPS can induce septic shock and death due to the massive release of cytokines by activated immune cells. Furthermore, microbes can be easily detected directly by immunohistochemistry, with straightforward histopathological assessment. Few studies have reported immunodetection of LPS in cancer tissues²⁸, and fewer still have sought to replicate it in an independent cohort⁴⁸, raising the key question: what is the true origin of microbial

signals seen in sequencing data and can they be robustly and systematically validated across studies in tumor tissues from organs that are not otherwise known to have microbiomes?

Recent experience illustrates that in most human tissue sequencing assays, the danger of nucleic acid contamination by external sources is high, and any analysis focusing on the non-human reads in these experiments can easily discover mostly or entirely contaminant DNA/RNA sequences. For example, microbiomes have been reported in several low-biomass tissues, including the placenta⁴⁴, brain⁴⁹, ocular surface⁵⁰, and blood⁵¹, though the conclusions of these studies have been vigorously challenged^{41,42,52}. Considering how human cancer tissues are prepared, microbial nucleic acids can enter at all points in the process, including surgery, transfer of material between the operating theater and the pathology department, tissue fixation (if conducted), cutting tissues into sections, storage and transport, laboratory processing for nucleic acid extraction, DNA/RNA library preparation and sequencing⁵³. This can yield findings of large numbers of microbes that are simply the result of contamination during one or more of these steps. Many recent papers on the search for cancer microbiomes do not properly control for this possibility and might therefore be reporting on environmental or database contaminants rather than biologically meaningful results^{12,54,55}.

In the recent past, the field of ancient DNA studies faced similar challenges to what cancer microbiome studies face today. Early studies in that field, when processing samples with relatively minor amounts of endogenous DNA, suffered from severe problems of contamination⁵⁶. As a consequence, a number of extravagant claims were reported as putative DNA findings, only to be revealed later as resulting from contamination⁵⁷. The field developed strict criteria to overcome such false reports, including independent replication of results in more than one laboratory⁵⁸. As a result, the field is thriving today, with notable successes such as the sequencing of the Neanderthal genome⁵⁹, the Denisovan genome⁶⁰, and many others involving ancient plants, animals, and microbes^{61–63}.

Below, we describe problems and suggest solutions in four distinct areas that will improve the reliability and reproducibility of studies of microbes in cancer tissues. This is critical because confirming microbial signals in cancer samples is more challenging, as certain controls used in ancient DNA studies, such as quantifying DNA damage patterns⁶⁴, are not always applicable. In addition, while several studies and reviews have highlighted the care that is needed for low-biomass microbiome studies^{53,65}, cancer tissue studies have additional considerations, including the ability to collect specific controls, the routine sequencing of normal tissue or blood samples for tumor-normal comparisons, the availability of clinical infrastructure for histology based validation (particularly in relation to host cells and their response markers), and the increasing use of large datasets from consortium studies to discover cancer-associated microbiome signatures⁶⁶. Given the importance of these points for this emerging field, we have developed a specific set of recommendations below, as well as an easily referred to checklist (**Table 1**), to guide future studies.

Tissue samples should be consistently handled using aseptic techniques and processed in clean rooms where feasible

Minimize sample contamination during sample collection, handling and storage. In general, the best way to avoid microbial contaminant signals from confounding and overwhelming analysis of cancer tissues is to minimize their presence in the first place. As contaminants can enter at all stages of the tissue collection and processing workflow it is critical to maintain aseptic techniques at all points (**Figure 1**). For example, while surgery to collect tissue specimens may follow strict aseptic procedures, what happens after the tissue is harvested is equally important and sample contamination during transport, storage and handling before it enters a genomics lab is often the hardest to account for in terms of contamination (**Table 1**, Q1, Q3). Correspondingly, few studies have tried to minimize contamination in these stages and indeed for archived samples this is perhaps the greatest concern^{10,67}. Given the exquisite sensitivity of deep sequencing-based analysis, even relatively minor contamination amounts in these stages can turn a tumor tissue that lacks microbes into one that readily displays signals for diverse bacteria⁶⁷.

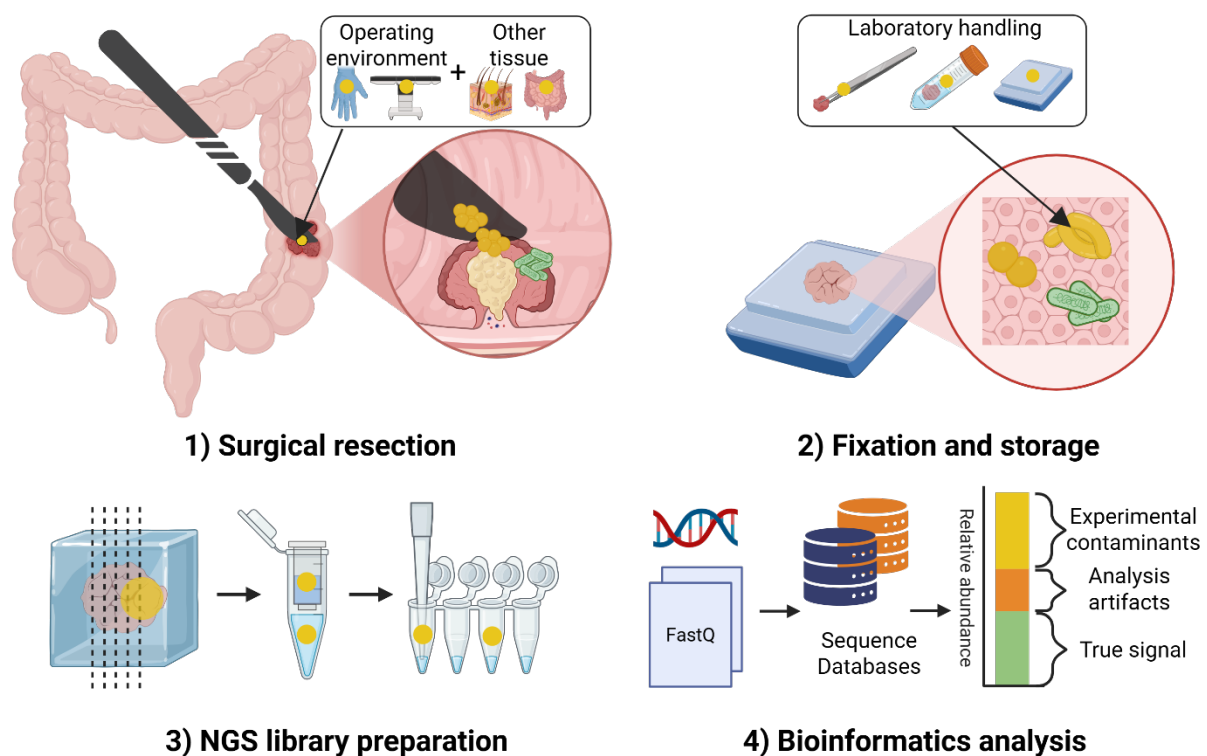


Figure 1. Possible sources of contamination (yellow) and analysis artefacts (orange) that can confound interpretation of true microbial signatures (green) in tumor tissues. Created in BioRender: <https://BioRender.com/9pnlnor>

Tissue samples should be processed in specialized clean facilities. This is a lesson that was learnt the hard way in studies of ancient DNA and correspondingly they employ stringent measures to enable high-fidelity analyses. In particular, tissue sampling, DNA extraction, and library building takes place in clean rooms that are physically isolated from laboratories where amplification of DNA takes place⁶⁸. This is because DNA amplification products provide among the greatest risks for contamination as they are produced in massive amounts and permeate

the laboratory, often floating in the air. For this reason, clean labs for ancient DNA analysis have separate ventilation systems, positive air pressure, and nightly UV irradiation to minimize contamination from the outside. Laboratory surfaces are cleaned daily by bleach, as are lab equipment. Similar precautions are also commonly adopted in molecular pathology labs. Full-body suits and disposable cloth are worn in clean labs (**Figure 2**) and daily movement of personnel between laboratories is always up the contamination gradient, i.e. from the clean lab towards the more contaminated laboratories⁵⁶. While such handling of tumor/normal tissues for microbial detection would be ideal, sadly no studies have incorporated this level of stringency. We acknowledge that access to clean rooms may not be available to many groups. Nevertheless, many hospitals have similar capabilities in pathology labs, and working in Biosafety Level 2 (BSL-2) labs with strict aseptic techniques would be an intermediate solution that is affordable and accessible, while still helping to reduce contamination risks (**Table 1**, Q6).



Figure 2. Images of a clean lab where ancient DNA samples are processed with appropriate measures to minimize contamination, including full body suits and physical isolation of areas where tissues are processed, and nucleic acids are amplified.

Internal controls should be systematically built in at every stage to assess and account for microbial contamination

Negative controls should be collected at every stage of sample handling. While reducing the risk of microbial contamination is an important strategy for improving the success rate of identifying genuine microbial signals, internal controls are needed in addition at every stage of sample handling to maximize the chances that contaminants are detected and removed from consideration. These include negative controls for sampling (>1 per sampling environment), storage (>1 per environment), DNA extraction, library preparation and sequencing (>1 per reagent and sequencing batch; **Table 1**, Q2, Q4, Q9) that are handled and processed under identical conditions as tumor samples, and can be sequenced to identify

contaminant signals (replicates can be used to assess consistency). Potential sources for negative controls include: (i) adjacent normal tissues sampled at the same time as the tumor and handled in the same batch (ideal for capturing upstream contamination sources), (ii) blank paraffin blocks as controls for formalin fixed and paraffin embedded (FFPE) samples, that are handled and processed in the same way as the tumor tissue, (iii) environmental swabs collected in the sample collection, storage and tissue processing environments, (iv) blank samples collected in the nucleic acid processing environment (e.g. PBS), and (v) tissue blanks from an FFPE block containing the tumor tissue (**Table 1**, Q3, Q4, Q5, Q9). Several studies have previously highlighted the utility of such negative controls^{67,69}, showing that sequencing of either blank (no tissue) paraffin blocks or extraction controls (no paraffin) readily detects microbial DNA reads originating from common gut, skin and oral microbes⁷⁰.

Independent replication of key results is critical. It is noteworthy that clean blank controls containing reagents, but no tissue samples, are by themselves not a guarantee for avoiding false positive results because of sample contamination and/or carrying effects. A blank control is a measure of how clean the reagents are, not the sample. Intriguingly, even if the reagents are contaminated it will not always show itself in the blank controls because minute amounts of DNA can bind to plastic surfaces and only be released by sample DNA being added and working as a carrier releasing the contaminants. The best way to assess if any positive finding derives from contamination of reagents is thus through independent replication of results by another laboratory, the rationale being that any such contaminant is unlikely to repeat itself across multiple labs (**Table 1**, Q7). Importantly, however, to avoid the case where a positive finding was due to microbial contamination of the sample itself, any replication study should be conducted on separate samples⁵⁶. At the very least, replication studies need to be conducted with different DNA isolation and library preparation kits to rule out the strong influence of contaminants from these sources (**Table 1**, Q8).

Positive controls should be included to assess sensitivity and robustness of data. While negative controls help reduce noise from contaminant sources, an equally important aspect is having positive controls for assessing (i) the sensitivity of detection methods, and (ii) to evaluate the impact of batch effects on the data (**Table 1**, Q14, Q16). By ensuring that microbial cell lysis methods (particularly for gram-positive bacteria and fungi) are optimized through positive control spike-ins/mock communities, studies can ensure that they improve their signal-to-noise ratio⁷¹. In addition, as large studies typically have multiple batches, and as comparison of data across batches can be significantly affected by sample processing differences (e.g. DNA extraction protocols), positive control samples are essential to evaluate and account for these differences⁷².

Bioinformatics analysis must be conservative and address sources of contamination

Analysis of legacy cancer genomic databases should be approached with caution. The growing number of whole-genome shotgun (WGS) and RNA sequencing datasets available in public archives, such as those from The Cancer Genome Atlas¹⁵ (TCGA), has sparked efforts to mine

these datasets for microbial signatures (**Table 2**). Notable publications from recent years include profiling a large cohort for evidence of a common blood microbiome⁴², finding cross-cohort microbial signatures in inflammatory bowel disease (IBD)⁷³ and mining sequences from metastatic tumors for microbiomes⁵. While bioinformatic analysis of existing sequence datasets is an efficient way of potentially gaining valuable insights into microbial signatures on a large scale, care must be taken to ensure that the conclusions are robust and account for possible false positive signals. Investigators must be especially cautious because most publicly available datasets were not collected specifically for microbiome analysis and therefore lack appropriate negative controls, or were handled in ways that could inadvertently introduce contaminants, as discussed above. Failure to account for such artifacts can result in spurious claims of microbiomes in tissues previously thought to be sterile, as has been claimed previously for the placenta⁴⁴ and for blood from healthy individuals⁵¹. Improper analysis of microbial reads in public datasets can also lead to mistaken claims that certain microbial species are associated with cancer. This is especially relevant for studies investigating fungal signatures in tumors, as they make up an even smaller proportion of microbial signals compared to bacteria. In one recent example, Aykut et al⁶ reported that pancreatic ductal adenocarcinomas (PDACs) had a higher burden of *Malassezia* spp, a genera of fungi which commonly colonizes the skin due to their metabolic dependence on host-derived lipids⁷⁴. However, a subsequent re-examination of their data showed that these associations were inconsistent and based on <10 fungal reads on average in PDACs, suggesting that those findings were confounded by background contamination, compounded by the absence of appropriate negative controls⁷⁵. In another example, Narunsky-Haziza et al¹ utilized TCGA data to define intra-tumoral fungal signatures (“mycotypes”) associated with different survival risks. However, many reads were incorrectly classified as fungal due to the misincorporation of vector or human sequences into fungal genomes within classification databases, or due to incomplete removal of human reads⁵⁴. We recognize that obtaining internal controls for historical sequencing datasets might be impossible, and we propose adopting the following data quality control steps to reduce the impact of contaminants on biological conclusions.

Genomes for contamination sources should be included during taxonomic classification.

Sequence databases used for taxonomic classification of microbial reads should contain not only the microbial genomes being searched, but also the host genome (as well as other potential contamination sources, e.g. plant pollen or dust mites), to reduce the false positive assignment of reads to microbial genomes (**Table 1**, Q17). This might seem obvious, but some published studies have assumed that a pre-filtering step in which reads are aligned to the human genome will be 100% effective at removing human reads, an assumption that is rarely correct. Inclusion of the host genome in the database is particularly important in the analysis of samples with relatively low microbial biomass, as is typical for human cancer tissue samples. For example, as shown recently¹², failure to include the human genome in a database resulted in millions of human reads being incorrectly identified as being microbial in a now-retracted study that claimed to find microbiome signatures in TCGA sequence data collected across 32 cancer types¹⁰. In addition, the use of more complete references such as telomere-to-

telomere assemblies (T2T CHM13) or the human pangenome⁷⁶ is likely to be more effective at removing human reads than commonly used references such as hg38⁷⁷.

High-quality microbial reference genomes should be used. Microbial sequence databases should be built from high-quality reference genomes, e.g. complete genomes from NCBI's RefSeq database. This reduces the chance that microbial reads from one species will be misclassified as another due to assembly errors in these genomes. If metagenome-assembled genomes (MAGs) must be used, extra care should be taken to ensure that they are of high quality and not chimeric^{78,79} (**Table 2**) and even then, any assignments based on MAGs should be recognized as having lower confidence.

Alignment properties of reads should be used to identify false positive signals. Microbial signatures of biological importance should be validated by independently aligning microbial reads to their relevant reference genomes, especially if these signatures are unexpected or novel (**Table 1**, Q18). This approach has previously been used in metagenomic profiling of ancient DNA to distinguish true positive species⁸⁰, and in the use of metagenomic sequencing to diagnose brain infections⁸¹. The central idea is that if a microbial species is truly present, then any reads sequenced from that species should originate from a relatively uniform distribution of locations on the reference genome. This contrasts with false-positive read assignments, which often accumulate unevenly in low-complexity regions, regions with high sequence identity to other genomes, or in regions where the assembly contains a contaminant from another genome. Low-complexity matches can be identified computationally by counting the number of unique k-mers in the matching reads⁸², and genome coverage metrics can be readily computed after alignment using common bioinformatics software⁸³ (**Table 2**).

Environmental contamination signals must be identified and removed. Environmental contaminants in metagenomic data can be distinguished from biological signals because abundances of the former are often negatively correlated with those of the latter (**Table 1**, Q12, Q19). As a corollary, signals from contaminant microbes from the same source tend to correlate positively with each other, and abundances of many contaminant microbes vary inversely with library concentration, which allows for their identification and removal (**Table 2**). This approach of identifying contaminants has been described in various studies and has proven successful in identifying contaminant signatures in low microbial biomass samples such as placenta⁵², blood⁴² and bronchoalveolar fluid⁸⁴. We emphasize that these quality control methods do not substitute for appropriate negative controls, but they can be valuable for metagenomic analyses, particularly when internal controls are not available. Also note that while contamination removal methods typically reduce the number of false positives, they do not guarantee that what remains are all true positives.

Information relevant to contaminant removal should be reported. Pertinent details for contamination avoidance and removal should be included in all publications to facilitate reproducibility and critical evaluation of the methods used (**Table 1**, Q13, Q20). For example, samples are often sequenced in barcoded pools, and incorrect assignment of reads across barcodes (*barcode hopping*) can readily introduce false signals, an issue that requires

particular care⁸⁵. Additionally, even if human reads cannot be released (e.g. due to privacy concerns), microbial reads should be made freely available with information on which taxa they were assigned to and read counts at all reported taxonomic levels should be provided per sample. Details on how the samples were collected and what was done to minimize microbial contamination should be provided. Where possible, extraction and sequencing kit lot numbers should be recorded to help identify experimental batch-specific contamination signals⁴.

Care should be taken for analysis of data across batches. Large public datasets that were sequenced for other purposes should be used with caution because cancers and controls are often sequenced in batches, and tissues are also processed in batches. The confounding is often so strong that computational batch correction techniques cannot alleviate the problem. In fact, applying these might exacerbate the strength of a false signal¹². Different laboratory techniques, such as the use of reagents or primer sequences that change over time, can also introduce artificial sequencing artifacts. These artifacts may not align with the human genome and could vary across tissue types or cancer status, leading to false signals. Such variability may mistakenly suggest different microbial profiles when, in fact, the differences are merely technical artifacts rather than true biological variation. Therefore, careful consideration of batch effects and technical variability is crucial when analyzing these datasets to avoid misinterpreting the data. Ideally, data should be analyzed within batches that are highly similar in how they were collected and processed. If data must be analyzed across clearly distinct batches, batch-correction steps may be needed, but these have to be used with great caution and with positive controls that help assess the success of this procedure (Table 1, Q15, Q16). Typically, despite these sanity checks, the risk that the signals identified are a function of batch correction remain, and careful validation in additional cohorts or through experimental approaches is essential before greater weight is lent to such results.

We recognize that studies that leverage legacy datasets will necessarily be restricted in being able to adopt several of the solutions proposed here. As a corollary, we propose that the conclusions that can be drawn solely from bioinformatic analysis of such datasets should be limited to ruling out the abundant presence of microbes. Claims of establishing the presence of microbes, or more generally a microbiome, should require extensive validation and internal controls in additional cohorts based on the approaches presented here (Table 1, Q21).

Additional detection techniques must be employed to support and clarify sequencing-based signals for microbes in cancer tissues

Targeted validation is essential for sequencing-based signals. Given the many confounding sources that can affect sequencing-based signals for microbes in cancer tissues, complementary detection techniques can be essential to validate signals (Table 1, Q22, Q24). For example, reverse transcription quantitative PCR (RT-qPCR) is widely used as a sensitive method for measuring microbial RNA abundances and to indicate the presence of live microbes in tissues^{86,87}. When used with appropriate controls, such as no-template and no–reverse transcriptase controls, qPCR can be a simple and effective method for confirming

specific microbial signals even in low-biomass environments. However, qPCR positivity still does not rule out the possibility of contamination from intact microbial cells introduced during surgical resection or laboratory handling. Similar challenges arise for culture-based confirmation, though it provides more direct evidence for live microbes (with appropriate negative controls).

Spatial information can clarify the nature of signals in tissues. Ideally, microbial signatures detected by sequencing-based methods should be further validated by *in situ* hybridization (ISH) of fluorescent or chromogenic probes to taxa-specific RNA markers in tissue sections (Table 1, Q10, Q11, Q23). Commonly used RNA markers for microbial detection include 16S/23S⁸⁸ and 28S⁸⁹ ribosomal RNAs for bacteria and fungi respectively, due to their high intracellular abundances and hypervariable regions whose sequences are genus or species specific^{90,91}. Less frequently, microbial mRNAs have also been targeted⁹² to detect some species e.g. *Mycobacterium tuberculosis*, which is more technically challenging but can be useful in cases where their 16S rRNAs lack species specificity⁹³. Skepticism about specific microbial signatures is warranted if they are detected by metagenomic sequencing or qPCR but not by RNA based imaging, or if the RNA signal is not localized to expected areas or cells in the tissue. Imaging based techniques have the additional advantage that they enable assessment of the frequency and distribution of specific bacteria in a tissue sample, and whether these patterns resemble contamination (e.g. if signals are peripheral to the tissue), localized infection or generalized dissemination in the tissue.

Detection of non-DNA/RNA microbial signals should be attempted. Since DNA is the major source of contamination in 16S rRNA and metagenomic sequencing experiments, *in-situ* staining and imaging of other biomolecules of microbial origin or signals of microbial infiltration are also powerful complementary methods to further verify microbial signals within tissues or cells (Table 1, Q22, Q23, Q24). Specifically, we know that the immune system is geared to detect microbes in the human body, and bacterial-derived lipopolysaccharides (LPSs) are one of the most reliable activators of immune responses⁹⁴. Consequently, strong immune cell infiltration should be detected when diverse microbes are present in tumor tissues. One possibility is that microbes are intracellular, but bacterial DNA and secreted LPSs nonetheless activate the innate immune system, resulting in local inflammatory responses. Thus a simple check for inflammatory markers or the presence of microbial cell-wall components (e.g. LPS for gram-negative or lipoteichoic acid (LTA) for gram-positive bacteria, as well as others for fungi⁹⁵) with immunohistochemistry (IHC) can serve as a basic control. Only a few recent studies have employed this control, such as a recent application detecting LPS-containing bacteria in breast cancer tissues²⁸. Reliable and well-established technologies such as IHC can be readily employed by multiple labs to provide independent (positive or negative) data on the presence of bacteria in the tumor or surrounding tissues, leveraging on spatial information to clarify the biological relevance of signals.

Conclusion

The application of ancient DNA analysis techniques, without adequate controls to exclude modern contamination, led to a series of publications spanning over at least eight years that had limited or no scientific value. Over time, the research community addressed these issues by implementing stringent quality controls, which significantly enhanced the reliability of subsequent studies, and recognition of the contributions of this field to science in the form of a Nobel Prize. Cancer microbiome research currently faces a similar challenge, and there have been growing calls for more rigorous approaches^{65,96}. Unfortunately, claims are still being published in lower quality studies based primarily on computational analyses of datasets, often lacking appropriate controls and validation (e.g. through qPCR, immunohistochemistry or culture), and thorough consideration of the limitations of the study (**Table 1**, Q25). This raises the concern that many microbial species detected in such studies may be artifacts of environmental or experimental contamination, rather than true biological findings. They can also impact the perception of the whole field (including of higher quality studies that incorporate more controls and validation work), when their poorly supported claims get retracted or are not validated in subsequent studies.

Overall, it is important to remember that detecting microbial DNA/RNA does not necessarily indicate the presence of a microbiome or for that matter even viable cells, and the relationship between microbes and cancer can be highly complex. It is not always clear whether the detected changes in the microbial profile of a tissue are a cause or consequence of the disease, creating a "chicken and egg" scenario. Nevertheless, even if microbial cells are convincingly shown to infect or infiltrate tissues after cancer initiation, this could have important consequences for the tumor microenvironment, immune responses and ultimately treatment outcomes. Therefore, while numerous studies on cancer microbiomes have already been published, readers should critically assess the methodologies used when evaluating published conclusions, while future studies are warranted with greater care to advance this exciting field.

The lack of rigorous quality control in many studies—compounded by inconsistent demands from reviewers and editors—risks wasting valuable time, talent, and resources on follow-up research that may not be reproducible due to flawed or poorly controlled data. To advance the field and derive meaningful biological and clinical insights, we propose several essential quality control measures that should be applied to all reports on microbes in cancer and other human tissues (summarized in a checklist; **Table 1**). The basic principle here is that the field needs to view reports of detection of microbes in diverse cancer tissue samples with care and appropriate skepticism, with the requirement that multiple carefully developed lines of evidence are needed in sites where the burden of microbes is expected to be very low (in the best-case scenario; **Table 1**, Q26). While research on ancient DNA analyses realized the need for rigorous controls essential to arrive at reliable conclusions nearly 10 years ago, similar controls are equally essential in the cancer microbiome field. These measures will strengthen the reliability of the findings and support progress in this important area of research.

Table 1: Checklist for reports on microbial species in human cancers.

Q#	Checklist question	Rationale	Response	Comments
Sample acquisition				
1	What was done to reduce contamination during sample acquisition?	Important control point; difficult to identify contaminants introduced here	Aseptic techniques post tissue harvest: ☐	
2	Were sampling controls collected in the same environment as the sample and were they sequenced?	Samples in each environment likely introduce a unique set of contaminant species	>1 control per sampling environment: ☐ Total:	
3	Were matched adjacent normal tissues collected and sequenced?	Excellent control to identify shared contaminants with tumor tissues, but may exclude true signals as well	Yes: ☐	
Sample handling and storage				
3	What was done to reduce contamination during handling and storage?	An aspect that is often ignored; labs where DNA is amplified are a great risk for contaminants	Aseptic techniques: ☐ Physical separation from DNA processing labs: ☐	
4	Were handling or storage controls collected and sequenced?	Can help identify contaminant signals	>1 control per storage environment: ☐ Total:	
5	For FFPE samples, were tissue blanks processed and sequenced?	Good control for FFPE samples	>1 control per FFPE batch: ☐ Total:	
Sample processing for nucleic acid analysis				
6	Were samples processed in a laboratory environment that minimizes nucleic acid contamination?	Laboratory contamination can easily overwhelm the signal in low biomass samples	Clean room: ☐ BSL2 lab w/ aseptic techniques: ☐	

			Physical separation from DNA amplification rooms: ☐	
7	Were samples processed in two independent labs for validating signals?	Important sanity check; does not rule out upstream contamination	Yes: ☐	
8	Were samples processed in the same lab with two different DNA isolation/library prep kits for validating signals?	Alternative, if access to independent labs was not feasible	Yes: ☐	
9	Were blank control samples collected and sequenced for each sequencing run and batch of reagents?	Reagent batches can have unique contaminant signals	≥ 1 per library prep & sequencing batch: ☐ Total:	
Sample processing for imaging-based analysis				
10	Were “no probe” controls included?	Signals can be due to autofluorescence	Yes: ☐	
11	Were negative control probes included in ISH experiments?	Checks for non-specific binding; probes for known contaminants can reveal distribution of environmental signals	Yes (details in comments): ☐	
Sequencing				
12	Were all libraries quantified before they were sequenced?	Useful information to identify contaminant species	Data is available: ☐	
13	What was done to minimize the risk of barcode hopping?	Can introduce significant read contamination	No multiplexing: ☐ Dual-indexing: ☐	
14	Were positive controls used to test the extent of barcode hopping?	Essential to determine the extent to which results may be affected	Estimated rate of barcode hopping:	
Batch information				

15	Has information for sample acquisition centers, date of acquisition, sample processing center and date, library preparation center and date, and sequencing batch been provided?	Key information to analyze control samples, evaluate sources of contamination and potentially account for batch effects on abundance profiles	Yes, in full: ☐ Partial (missing data explained in comments): ☐	
16	What was done to minimize the impact of batch variation across samples that are jointly analyzed?	Important for statistical analysis with data from multiple batches	No analysis across batches: ☐ Batch-to-batch variation was assessed to be minimal: ☐ Batch correction was performed w/ appropriate positive controls: ☐	
Bioinformatic analysis of sequencing data				
17	Was taxonomic classification done with databases that include the highest quality reference sequences available for humans (T2T, pangenome) and potential contamination sources?	Conservative approach to reduce false positive assignments to microbial taxa	Yes (details in comments): ☐	
18	Was the genomic distribution of reads and quality of alignments used to assess potential identification of microbial taxa?	False-positive signals often cluster unevenly and can be due to low-quality alignments	Yes (details in comments): ☐	
19	Were potential contaminant species identified and computationally subtracted from the data?	Important step that leverages control samples to conservatively identify contaminant signals	Yes (details in comments): ☐	

20	Is data relevant for contamination removal available?	Essential for reproducibility and reuse of data with newer bioinformatic methods	Taxonomic profiles (tumor and control): ☐ DNA quantification: ☐ Library quantification: ☐	
21	Were detected microbes validated based on sequencing data from another lab or another cohort?	Can be a quick sanity check if the data is already available	Yes (details in comments): ☐	
Evidence from non-shotgun-metagenomic sources				
22	Were detected microbes validated in the same samples based on targeted approaches?	Targeted validation can help rule out bioinformatic artefacts	FISH: ☐ IF: ☐ PCR: ☐ Others (details in comments): ☐	
23	Were detected microbes validated to be present in or near tumor cells?	Spatial mapping of microbes in tissue can help clarify the nature of detected signals	FISH: ☐ IF: ☐ IHC: ☐ Others (details in comments): ☐	
24	What is the evidence for viable or metabolically active microbial cells?	The viability and/or metabolic activity of cells can help reduce concerns related to contamination by dead cells or nucleic acids	Cell culture: ☐ Targeted RNA detection: ☐ Spatial metabolomics: ☐	
Other considerations				
25	Were limitations of the measures taken to reduce and identify	Contamination at any stage can impact the ability to draw	Yes (details in comments): ☐	

	contamination acknowledged?	conclusions from the data		
26	Are there multiple lines of evidence that consistently detect specific microbes in a set of samples?	Orthogonal sources of validation are essential to avoid artefacts of any one assay	Yes (details in comments): ☐	

448

449 **Table 2: Publicly available resources for mining microbial signatures associated with cancer**
450 **and normal tissues.**

No.	Name	Type	Salient Features	Caveats/Recommendations
1	The Cancer Genome Atlas Program (TCGA) ¹⁵	Omics Dataset	WGS, Exome and RNA-seq data for >20,000 tumors and matched normal tissues from 33 cancer types	Lacks negative controls; sample handling was not geared towards minimizing microbial contamination; limited experimental information (e.g. library concentrations)
2	CRC microbiome explorer ⁹⁷	Omics Dataset	FFPE RNA-seq for >900 colorectal tumors; contaminant genera filtered	Lacks negative controls; sample handling was not geared towards minimizing microbial contamination; limited experimental information (e.g. library concentrations)
3	AC-ICAM ⁹⁸	Omics Dataset	WGS, RNA-seq & 16S rRNA sequencing of 348 fresh-frozen primary colon tumors and matched normal tissue	Lacks negative controls; sample handling was not geared towards minimizing microbial contamination; limited experimental information (e.g. library concentrations)
4	Battaglia et al ⁵	Omics Dataset	WGS, RNA-seq & 16S rRNA sequencing of >4000 metastatic tumors from 26 tissue types, with negative controls	Taxonomic profiles may contain false positives due to the use of databases containing MAGs that are not of high quality; limited experimental information (e.g. library concentrations)
5	Nejman et al ²⁸	Omics Dataset	16S rRNA sequencing of >1500 tumors and matched normal tissues from 7 cancer types, with negative controls	Species level classification using 16S sequences alone might produce false positives due to sequence similarities ⁹⁹

6	Gihawi et al ¹²	Analysis Results	Taxonomic profiles for 17,625 TCGA samples accounting for computational analysis artefacts	Taxonomic profiles may still contain contaminants due to the absence of negative controls or due to false positive species calls
7	Eisenhofer et al ⁵³	Analysis Results	Meta-analysis resulting in a list of potential contaminant taxa	Published lists of contaminant genera are non-exhaustive
8	Salter et al ¹⁰⁰	Analysis Results	16S sequencing of samples and negative controls with batch information, yielding a list of potential contaminant taxa	Published lists of contaminant genera are non-exhaustive
9	KrakenUniq ⁸²	Software	Taxonomic classifier that identifies likely false positives using unique k-mers	Thresholds for unique k-mers should be empirically determined for different sample types
10	Decontam ¹⁰¹	Software	Detects likely contaminant taxa based on analysis of taxonomic profiles	Leverages information from DNA/RNA input concentrations and/or matched negative controls
11	Squeegee ¹⁰²	Software	De novo identification of contaminant taxa in the absence of negative controls and DNA/RNA input concentrations	Poorer sensitivity for low abundance contaminants ¹⁰² ; potentially more computationally expensive compared to alternatives
12	checkM ⁷⁸	Software	To assess completeness & contamination in MAGs	Use of high-quality MAGs with no/low chimerism is recommended as genomic references for taxonomic classification

13	GUNC ⁷⁹	Software	Detects and quantifies chimerism in MAGs	Use of high-quality MAGs with no/low chimerism is recommended as genomic references for taxonomic classification
14	MMUPHin ⁷³	Software	Performs batch correction of taxonomic profiles	Batch effects must not be highly confounded with the variable(s) of interest; should be used with care, only when essential, and with appropriate positive controls to assess the impact of batch correction

451

References

1. 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. 10.1016/j.cell.2022.09.005.
2. Wang, J., Wang, Y., Li, Z., Gao, X., and Huang, D. (2021). Global analysis of microbiota signatures in four major types of gastrointestinal cancer. *Front. Oncol.* 11, 685641. 10.3389/fonc.2021.685641.
3. Cai, M., Kandalai, S., Tang, X., and Zheng, Q. (2022). Contributions of Human-Associated Archaeal Metabolites to Tumor Microenvironment and Carcinogenesis. *Microbiol. Spectr.* 10, e0236721. 10.1128/spectrum.02367-21.
4. Dohman, 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. 10.1016/j.chom.2020.12.001.
5. Battaglia, T.W., Mimpen, I.L., Traets, J.J.H., van Hoeck, A., Zevenij, L.J., Geurts, B.S., de Wit, G.F., 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. 10.1016/j.cell.2024.03.021.
6. Aykut, B., Pushalkar, S., Chen, R., Li, Q., Abengozar, R., Kim, J.I., Shadaloey, S.A., Wu, D., Preiss, P., Verma, N., et al. (2019). The fungal mycobiome promotes pancreatic oncogenesis via activation of MBL. *Nature* 574, 264–267. 10.1038/s41586-019-1608-2.
7. Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20, 257. 10.1186/s13059-019-1891-0.
8. Hong, C., Manimaran, S., Shen, Y., Perez-Rogers, J.F., Byrd, A.L., Castro-Nallar, E., Crandall, K.A., and Johnson, W.E. (2014). PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome* 2, 33. 10.1186/2049-2618-2-33.
9. Kim, D., Song, L., Breitwieser, F.P., and Salzberg, S.L. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* 26, 1721–1729. 10.1101/gr.210641.116.
10. Poore, G.D., Kopylova, E., Zhu, Q., Carpenter, C., Fraraccio, S., Wandro, S., Kosciolk, T., Janssen, S., Metcalf, J., Song, S.J., et al. (2020). Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* 579, 567–574. 10.1038/s41586-020-2095-1.
11. Sepich-Poore, G.D., McDonald, D., Kopylova, E., Guccione, C., Zhu, Q., Austin, G., Carpenter, C., Fraraccio, S., Wandro, S., Kosciolk, T., et al. (2024). Robustness of

- 489 cancer microbiome signals over a broad range of methodological variation. *Oncogene*
490 43, 1127–1148. 10.1038/s41388-024-02974-w.
- 491 12. Gihawi, A., Ge, Y., Lu, J., Puiu, D., Xu, A., Cooper, C.S., Brewer, D.S., Pertea, M., and
492 Salzberg, S.L. (2023). Major data analysis errors invalidate cancer microbiome findings.
493 *MBio* 14, e0160723. 10.1128/mbio.01607-23.
- 494 13. Scott, A.J., Alexander, J.L., Merrifield, C.A., Cunningham, D., Jobin, C., Brown, R.,
495 Alverdy, J., O’Keefe, S.J., Gaskins, H.R., Teare, J., et al. (2019). International Cancer
496 Microbiome Consortium consensus statement on the role of the human microbiome
497 in carcinogenesis. *Gut* 68, 1624–1632. 10.1136/gutjnl-2019-318556.
- 498 14. Robinson, K.M., Crabtree, J., Mattick, J.S.A., Anderson, K.E., and Dunning Hotopp, J.C.
499 (2017). Distinguishing potential bacteria-tumor associations from contamination in a
500 secondary data analysis of public cancer genome sequence data. *Microbiome* 5, 9.
501 10.1186/s40168-016-0224-8.
- 502 15. The Cancer Genome Atlas Program (TCGA) - NCI
503 <https://www.cancer.gov/ccg/research/genome-sequencing/tcga>.
- 504 16. Zepeda-Rivera, M., Minot, S.S., Bouzek, H., Wu, H., Blanco-Míguez, A., Manghi, P.,
505 Jones, D.S., LaCourse, K.D., Wu, Y., McMahon, E.F., et al. (2024). A distinct
506 *Fusobacterium nucleatum* clade dominates the colorectal cancer niche. *Nature* 628,
507 424–432. 10.1038/s41586-024-07182-w.
- 508 17. Long, X., Wong, C.C., Tong, L., Chu, E.S.H., Ho Szeto, C., Go, M.Y.Y., Coker, O.O., Chan,
509 A.W.H., Chan, F.K.L., Sung, J.J.Y., et al. (2019). *Peptostreptococcus anaerobius*
510 promotes colorectal carcinogenesis and modulates tumour immunity. *Nat. Microbiol.*
511 4, 2319–2330. 10.1038/s41564-019-0541-3.
- 512 18. Cullin, N., Azevedo Antunes, C., Straussman, R., Stein-Thoeringer, C.K., and Elinav, E.
513 (2021). Microbiome and cancer. *Cancer Cell* 39, 1317–1341.
514 10.1016/j.ccell.2021.08.006.
- 515 19. Knippel, R.J., Drewes, J.L., and Sears, C.L. (2021). The cancer microbiome: recent
516 highlights and knowledge gaps. *Cancer Discov.* 11, 2378–2395. 10.1158/2159-
517 8290.CD-21-0324.
- 518 20. Fu, A., Yao, B., Dong, T., Chen, Y., Yao, J., Liu, Y., Li, H., Bai, H., Liu, X., Zhang, Y., et al.
519 (2022). Tumor-resident intracellular microbiota promotes metastatic colonization in
520 breast cancer. *Cell* 185, 1356–1372.e26. 10.1016/j.cell.2022.02.027.
- 521 21. Parhi, L., Alon-Maimon, T., Sol, A., Nejman, D., Shhadeh, A., Fainsod-Levi, T., Yajuk, O.,
522 Isaacson, B., Abed, J., Maalouf, N., et al. (2020). Breast cancer colonization by
523 *Fusobacterium nucleatum* accelerates tumor growth and metastatic progression. *Nat.*
524 *Commun.* 11, 3259. 10.1038/s41467-020-16967-2.
- 525 22. Riquelme, E., Zhang, Y., Zhang, L., Montiel, M., Zoltan, M., Dong, W., Quesada, P.,

526 Sahin, I., Chandra, V., San Lucas, A., et al. (2019). Tumor microbiome diversity and
527 composition influence pancreatic cancer outcomes. *Cell* 178, 795-806.e12.
528 10.1016/j.cell.2019.07.008.

529 23. Pushalkar, S., Hundeyin, M., Daley, D., Zambirinis, C.P., Kurz, E., Mishra, A., Mohan, N.,
530 Aykut, B., Usyk, M., Torres, L.E., et al. (2018). The pancreatic cancer microbiome
531 promotes oncogenesis by induction of innate and adaptive immune suppression.
532 *Cancer Discov.* 8, 403–416. 10.1158/2159-8290.CD-17-1134.

533 24. Galeano Niño, J.L., Wu, H., LaCourse, K.D., Kempchinsky, A.G., Baryames, A., Barber,
534 B., Futran, N., Houlton, J., Sather, C., Sicinska, E., et al. (2022). Effect of the
535 intratumoral microbiota on spatial and cellular heterogeneity in cancer. *Nature* 611,
536 810–817. 10.1038/s41586-022-05435-0.

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

540 26. Coker, O.O., Nakatsu, G., Dai, R.Z., Wu, W.K.K., Wong, S.H., Ng, S.C., Chan, F.K.L., Sung,
541 J.J.Y., and Yu, J. (2019). Enteric fungal microbiota dysbiosis and ecological alterations
542 in colorectal cancer. *Gut* 68, 654–662. 10.1136/gutjnl-2018-317178.

543 27. Choi, M.H., Kim, M., Jeong, S.J., Choi, J.Y., Lee, I.-Y., Yong, T.-S., Yong, D., Jeong, S.H.,
544 and Lee, K. (2019). Risk factors for elizabethkingia acquisition and clinical
545 characteristics of patients, south korea. *Emerging Infect. Dis.* 25, 42–51.
546 10.3201/eid2501.171985.

547 28. Nejman, D., Livyatan, I., Fuks, G., Gavert, N., Zwang, Y., Geller, L.T., Rotter-Maskowitz,
548 A., Weiser, R., Mallel, G., Gigi, E., et al. (2020). The human tumor microbiome is
549 composed of tumor type-specific intracellular bacteria. *Science* 368, 973–980.
550 10.1126/science.aay9189.

551 29. Zapatka, M., Borozan, I., Brewer, D.S., Iskar, M., Grundhoff, A., Alawi, M., Desai, N.,
552 Sültmann, H., Moch, H., PCAWG Pathogens, et al. (2020). The landscape of viral
553 associations in human cancers. *Nat. Genet.* 52, 320–330. 10.1038/s41588-019-0558-9.

554 30. Rubin, H. (2011). The early history of tumor virology: Rous, RIF, and RAV. *Proc Natl*
555 *Acad Sci USA* 108, 14389–14396. 10.1073/pnas.1108655108.

556 31. Polk, D.B., and Peek, R.M. (2010). *Helicobacter pylori*: gastric cancer and beyond. *Nat.*
557 *Rev. Cancer* 10, 403–414. 10.1038/nrc2857.

558 32. Wong, S.H., and Yu, J. (2019). Gut microbiota in colorectal cancer: mechanisms of
559 action and clinical applications. *Nat. Rev. Gastroenterol. Hepatol.* 16, 690–704.
560 10.1038/s41575-019-0209-8.

561 33. Scanu, T., Spaapen, R.M., Bakker, J.M., Pratap, C.B., Wu, L., Hofland, I., Broeks, A.,
562 Shukla, V.K., Kumar, M., Janssen, H., et al. (2015). *Salmonella* Manipulation of Host

563 Signaling Pathways Provokes Cellular Transformation Associated with Gallbladder
564 Carcinoma. *Cell Host Microbe* 17, 763–774. 10.1016/j.chom.2015.05.002.

565 34. Rubinstein, M.R., Wang, X., Liu, W., Hao, Y., Cai, G., and Han, Y.W. (2013).
566 *Fusobacterium nucleatum* promotes colorectal carcinogenesis by modulating E-
567 cadherin/ β -catenin signaling via its FadA adhesin. *Cell Host Microbe* 14, 195–206.
568 10.1016/j.chom.2013.07.012.

569 35. Serna, G., Ruiz-Pace, F., Hernando, J., Alonso, L., Fasani, R., Landolfi, S., Comas, R.,
570 Jimenez, J., Elez, E., Bullman, S., et al. (2020). *Fusobacterium nucleatum* persistence
571 and risk of recurrence after preoperative treatment in locally advanced rectal cancer.
572 *Ann. Oncol.* 31, 1366–1375. 10.1016/j.annonc.2020.06.003.

573 36. Jiang, S.-S., Xie, Y.-L., Xiao, X.-Y., Kang, Z.-R., Lin, X.-L., Zhang, L., Li, C.-S., Qian, Y., Xu,
574 P.-P., Leng, X.-X., et al. (2023). *Fusobacterium nucleatum*-derived succinic acid induces
575 tumor resistance to immunotherapy in colorectal cancer. *Cell Host Microbe* 31, 781-
576 797.e9. 10.1016/j.chom.2023.04.010.

577 37. Vétizou, M., Pitt, J.M., Daillère, R., Lepage, P., Waldschmitt, N., Flament, C.,
578 Rusakiewicz, S., Routy, B., Roberti, M.P., Duong, C.P.M., et al. (2015). Anticancer
579 immunotherapy by CTLA-4 blockade relies on the gut microbiota. *Science* 350, 1079–
580 1084. 10.1126/science.aad1329.

581 38. Mager, L.F., Burkhard, R., Pett, N., Cooke, N.C.A., Brown, K., Ramay, H., Paik, S., Stagg,
582 J., Groves, R.A., Gallo, M., et al. (2020). Microbiome-derived inosine modulates
583 response to checkpoint inhibitor immunotherapy. *Science* 369, 1481–1489.
584 10.1126/science.abc3421.

585 39. Björk, J.R., Bolte, L.A., Maltez Thomas, A., Lee, K.A., Rossi, N., Wind, T.T., Smit, L.M.,
586 Armanini, F., Asnicar, F., Blanco-Miguez, A., et al. (2024). Longitudinal gut microbiome
587 changes in immune checkpoint blockade-treated advanced melanoma. *Nat. Med.* 30,
588 785–796. 10.1038/s41591-024-02803-3.

589 40. Hsu, C.L., and Schnabl, B. (2023). The gut-liver axis and gut microbiota in health and
590 liver disease. *Nat. Rev. Microbiol.* 21, 719–733. 10.1038/s41579-023-00904-3.

591 41. Bedarf, J.R., Beraza, N., Khazneh, H., Özkurt, E., Baker, D., Borger, V., Wüllner, U., and
592 Hildebrand, F. (2021). Much ado about nothing? Off-target amplification can lead to
593 false-positive bacterial brain microbiome detection in healthy and Parkinson’s disease
594 individuals. *Microbiome* 9, 75. 10.1186/s40168-021-01012-1.

595 42. Tan, C.C.S., Ko, K.K.K., Chen, H., Liu, J., Loh, M., SG10K_Health Consortium, Chia, M.,
596 and Nagarajan, N. (2023). No evidence for a common blood microbiome based on a
597 population study of 9,770 healthy humans. *Nat. Microbiol.* 8, 973–985.
598 10.1038/s41564-023-01350-w.

599 43. Emery, D.C., Shoemark, D.K., Batstone, T.E., Waterfall, C.M., Coghill, J.A., Cerajewska,
600 T.L., Davies, M., West, N.X., and Allen, S.J. (2017). 16S rRNA Next Generation

Sequencing Analysis Shows Bacteria in Alzheimer's Post-Mortem Brain. *Front. Aging Neurosci.* 9, 195. 10.3389/fnagi.2017.00195.

44. Aagaard, K., Ma, J., Antony, K.M., Ganu, R., Petrosino, J., and Versalovic, J. (2014). The placenta harbors a unique microbiome. *Sci. Transl. Med.* 6, 237ra65. 10.1126/scitranslmed.3008599.

45. Alonso, R., Pisa, D., Fernández-Fernández, A.M., and Carrasco, L. (2018). Infection of fungi and bacteria in brain tissue from elderly persons and patients with alzheimer's disease. *Front. Aging Neurosci.* 10, 159. 10.3389/fnagi.2018.00159.

46. Thomson, C.A., McColl, A., Graham, G.J., and Cavanagh, J. (2020). Sustained exposure to systemic endotoxin triggers chemokine induction in the brain followed by a rapid influx of leukocytes. *J. Neuroinflammation* 17, 94. 10.1186/s12974-020-01759-8.

47. Komai-Koma, M., Gilchrist, D.S., and Xu, D. (2009). Direct recognition of LPS by human but not murine CD8+ T cells via TLR4 complex. *Eur. J. Immunol.* 39, 1564–1572. 10.1002/eji.200838866.

48. de Miranda, N.F., Smit, V.T., van der Ploeg, M., Wesseling, J., and Neefjes, J. (2023). Absence of Lipopolysaccharide (LPS) expression in Breast Cancer Cells. *BioRxiv*. 10.1101/2023.08.28.555057.

49. Branton, W.G., Ellestad, K.K., Maingat, F., Wheatley, B.M., Rud, E., Warren, R.L., Holt, R.A., Surette, M.G., and Power, C. (2013). Brain microbial populations in HIV/AIDS: α -proteobacteria predominate independent of host immune status. *PLoS ONE* 8, e54673. 10.1371/journal.pone.0054673.

50. Willis, K.A., Postnikoff, C.K., Freeman, A., Rezonzew, G., Nichols, K., Gaggar, A., and Lal, C.V. (2020). The closed eye harbours a unique microbiome in dry eye disease. *Sci. Rep.* 10, 12035. 10.1038/s41598-020-68952-w.

51. D'Aquila, P., Giacconi, R., Malavolta, M., Piacenza, F., Bürkle, A., Villanueva, M.M., Dollé, M.E.T., Jansen, E., Grune, T., Gonos, E.S., et al. (2021). Microbiome in Blood Samples From the General Population Recruited in the MARK-AGE Project: A Pilot Study. *Front. Microbiol.* 12, 707515. 10.3389/fmicb.2021.707515.

52. de Goffau, M.C., Lager, S., Sovio, U., Gaccioli, F., Cook, E., Peacock, S.J., Parkhill, J., Charnock-Jones, D.S., and Smith, G.C.S. (2019). Human placenta has no microbiome but can contain potential pathogens. *Nature* 572, 329–334. 10.1038/s41586-019-1451-5.

53. 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. 10.1016/j.tim.2018.11.003.

54. Ge, Y., Lu, J., Puiu, D., Revsine, M., and Salzberg, S.L. (2024). Comprehensive analysis of microbial content in whole-genome sequencing samples from The Cancer Genome

638 Atlas project. *BioRxiv*. 10.1101/2024.05.24.595788.

639 55. Breitwieser, F.P., Perte, M., Zimin, A.V., and Salzberg, S.L. (2019). Human
640 contamination in bacterial genomes has created thousands of spurious proteins.
641 *Genome Res.* 29, 954–960. 10.1101/gr.245373.118.

642 56. Willerslev, E., and Cooper, A. (2005). Ancient DNA. *Proc. Biol. Sci.* 272, 3–16.
643 10.1098/rspb.2004.2813.

644 57. Hebsgaard, M.B., Phillips, M.J., and Willerslev, E. (2005). Geologically ancient DNA:
645 fact or artefact? *Trends Microbiol.* 13, 212–220. 10.1016/j.tim.2005.03.010.

646 58. Cooper, A., and Poinar, H.N. (2000). Ancient DNA: do it right or not at all. *Science* 289,
647 1139. 10.1126/science.289.5482.1139b.

648 59. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N.,
649 Li, H., Zhai, W., Fritz, M.H.-Y., et al. (2010). A draft sequence of the Neandertal
650 genome. *Science* 328, 710–722. 10.1126/science.1188021.

651 60. Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G.,
652 Jay, F., Prüfer, K., de Filippo, C., et al. (2012). A high-coverage genome sequence from
653 an archaic Denisovan individual. *Science* 338, 222–226. 10.1126/science.1224344.

654 61. Liang, R., Li, Z., Lau Vetter, M.C.Y., Vishnivetskaya, T.A., Zanina, O.G., Lloyd, K.G.,
655 Pfiffner, S.M., Rivkina, E.M., Wang, W., Wiggins, J., et al. (2021). Genomic
656 reconstruction of fossil and living microorganisms in ancient Siberian permafrost.
657 *Microbiome* 9, 110. 10.1186/s40168-021-01057-2.

658 62. Willerslev, E., Hansen, A.J., Binladen, J., Brand, T.B., Gilbert, M.T.P., Shapiro, B., Bunce,
659 M., Wiuf, C., Gilichinsky, D.A., and Cooper, A. (2003). Diverse plant and animal genetic
660 records from Holocene and Pleistocene sediments. *Science* 300, 791–795.
661 10.1126/science.1084114.

662 63. Ramos-Madrigal, J., Smith, B.D., Moreno-Mayar, J.V., Gopalakrishnan, S., Ross-Ibarra,
663 J., Gilbert, M.T.P., and Wales, N. (2016). Genome Sequence of a 5,310-Year-Old Maize
664 Cob Provides Insights into the Early Stages of Maize Domestication. *Curr. Biol.* 26,
665 3195–3201. 10.1016/j.cub.2016.09.036.

666 64. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P.L.F., and Orlando, L. (2013).
667 mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage
668 parameters. *Bioinformatics* 29, 1682–1684. 10.1093/bioinformatics/btt193.

669 65. de Goffau, M.C., Lager, S., Salter, S.J., Wagner, J., Kronbichler, A., Charnock-Jones,
670 D.S., Peacock, S.J., Smith, G.C.S., and Parkhill, J. (2018). Recognizing the reagent
671 microbiome. *Nat. Microbiol.* 3, 851–853. 10.1038/s41564-018-0202-y.

672 66. Sepich-Poore, G.D., Zitvogel, L., Straussman, R., Hasty, J., Wargo, J.A., and Knight, R.
673 (2021). The microbiome and human cancer. *Science* 371. 10.1126/science.abc4552.

674 67. Cruz-Flores, R., López-Carvallo, J.A., Cáceres-Martínez, J., and Dhar, A.K. (2022).

675 Microbiome analysis from formalin-fixed paraffin-embedded tissues: Current
676 challenges and future perspectives. *J. Microbiol. Methods* 196, 106476.
677 10.1016/j.mimet.2022.106476.

678 68. Zhu, K., He, H., Tao, L., Ma, H., Yang, X., Wang, R., Guo, J., and Wang, C.-C. (2024).
679 Protocol for a comprehensive pipeline to study ancient human genomes. *STAR*
680 *Protocols* 5, 102985. 10.1016/j.xpro.2024.102985.

681 69. Lam, S.Y., Ioannou, A., Konstanti, P., Visseren, T., Doukas, M., Peppelenbosch, M.P.,
682 Belzer, C., and Fuhler, G.M. (2021). Technical challenges regarding the use of formalin-
683 fixed paraffin embedded (FFPE) tissue specimens for the detection of bacterial
684 alterations in colorectal cancer. *BMC Microbiol.* 21, 297. 10.1186/s12866-021-02359-
685 z.

686 70. CSB5/FFPE_kitome: Metagenomic sequencing conducted on blank paraffin blocks or
687 extraction controls. https://github.com/CSB5/FFPE_kitome.

688 71. Turlousse, D.M., Narita, K., Miura, T., Ohashi, A., Matsuda, M., Ohyama, Y.,
689 Shimamura, M., Furukawa, M., Kasahara, K., Kameyama, K., et al. (2022).
690 Characterization and demonstration of mock communities as control reagents for
691 accurate human microbiome community measurements. *Microbiol. Spectr.* 10,
692 e0191521. 10.1128/spectrum.01915-21.

693 72. Leek, J.T. (2014). svaseq: removing batch effects and other unwanted noise from
694 sequencing data. *Nucleic Acids Res.* 42, e161. 10.1093/nar/gku864.

695 73. Ma, S., Shungin, D., Mallick, H., Schirmer, M., Nguyen, L.H., Kolde, R., Franzosa, E.,
696 Vlamakis, H., Xavier, R., and Huttenhower, C. (2022). Population structure discovery in
697 meta-analyzed microbial communities and inflammatory bowel disease using
698 MMUPHin. *Genome Biol.* 23, 208. 10.1186/s13059-022-02753-4.

699 74. Vijaya Chandra, S.H., Srinivas, R., Dawson, T.L., and Common, J.E. (2020). Cutaneous
700 malassezia: commensal, pathogen, or protector? *Front. Cell. Infect. Microbiol.* 10,
701 614446. 10.3389/fcimb.2020.614446.

702 75. Fletcher, A.A., Kelly, M.S., Eckhoff, A.M., and Allen, P.J. (2023). Revisiting the intrinsic
703 mycobiome in pancreatic cancer. *Nature* 620, E1–E6. 10.1038/s41586-023-06292-1.

704 76. Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J.K.,
705 Monlong, J., Abel, H.J., et al. (2023). A draft human pangenome reference. *Nature*
706 617, 312–324. 10.1038/s41586-023-05896-x.

707 77. Forbes, M., Ng, D.Y.K., Boggan, R.M., Frick-Kretschmer, A., Durham, J., Lorenz, O.,
708 Dave, B., Lassalle, F., Scott, C., Wagner, J., et al. (2025). Benchmarking of human read
709 removal strategies for viral and microbial metagenomics. *BioRxiv*.
710 10.1101/2025.03.21.644587.

711 78. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015).

712 CheckM: assessing the quality of microbial genomes recovered from isolates, single
713 cells, and metagenomes. *Genome Res.* 25, 1043–1055. 10.1101/gr.186072.114.

714 79. Orakov, A., Fullam, A., Coelho, L.P., Khedkar, S., Szklarczyk, D., Mende, D.R., Schmidt,
715 T.S.B., and Bork, P. (2021). GUNC: detection of chimerism and contamination in
716 prokaryotic genomes. *Genome Biol.* 22, 178. 10.1186/s13059-021-02393-0.

717 80. Pochon, Z., Bergfeldt, N., Kirdök, E., Vicente, M., Naidoo, T., van der Valk, T., Altınışık,
718 N.E., Krzewińska, M., Dalén, L., Götherström, A., et al. (2023). aMeta: an accurate and
719 memory-efficient ancient metagenomic profiling workflow. *Genome Biol.* 24, 242.
720 10.1186/s13059-023-03083-9.

721 81. Salzberg, S.L., Breitwieser, F.P., Kumar, A., Hao, H., Burger, P., Rodriguez, F.J., Lim, M.,
722 Quiñones-Hinojosa, A., Gallia, G.L., Tornheim, J.A., et al. (2016). Next-generation
723 sequencing in neuropathologic diagnosis of infections of the nervous system. *Neurol.*
724 *Neuroimmunol. Neuroinflamm.* 3, e251. 10.1212/NXI.0000000000000251.

725 82. Breitwieser, F.P., Baker, D.N., and Salzberg, S.L. (2018). KrakenUniq: confident and fast
726 metagenomics classification using unique k-mer counts. *Genome Biol.* 19, 198.
727 10.1186/s13059-018-1568-0.

728 83. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis,
729 G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The
730 Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.
731 10.1093/bioinformatics/btp352.

732 84. Willner, D., Daly, J., Whiley, D., Grimwood, K., Wainwright, C.E., and Hugenholtz, P.
733 (2012). Comparison of DNA extraction methods for microbial community profiling
734 with an application to pediatric bronchoalveolar lavage samples. *PLoS ONE* 7, e34605.
735 10.1371/journal.pone.0034605.

736 85. Costello, M., Fleharty, M., Abreu, J., Farjoun, Y., Ferriera, S., Holmes, L., Granger, B.,
737 Green, L., Howd, T., Mason, T., et al. (2018). Characterization and remediation of
738 sample index swaps by non-redundant dual indexing on massively parallel sequencing
739 platforms. *BMC Genomics* 19, 332. 10.1186/s12864-018-4703-0.

740 86. Magalhães, A.P., França, Â., Pereira, M.O., and Cerca, N. (2019). RNA-based qPCR as a
741 tool to quantify and to characterize dual-species biofilms. *Sci. Rep.* 9, 13639.
742 10.1038/s41598-019-50094-3.

743 87. Kapoor, V., Pitkänen, T., Ryu, H., Elk, M., Wendell, D., and Santo Domingo, J.W. (2015).
744 Distribution of human-specific bacteroidales and fecal indicator bacteria in an urban
745 watershed impacted by sewage pollution, determined using RNA- and DNA-based
746 quantitative PCR assays. *Appl. Environ. Microbiol.* 81, 91–99. 10.1128/AEM.02446-14.

747 88. Fuchs, B.M., Syutsubo, K., Ludwig, W., and Amann, R. (2001). In situ accessibility of
748 *Escherichia coli* 23S rRNA to fluorescently labeled oligonucleotide probes. *Appl.*
749 *Environ. Microbiol.* 67, 961–968. 10.1128/AEM.67.2.961-968.2001.

- 750 89. Rickerts, V., Khot, P.D., Myerson, D., Ko, D.L., Lambrecht, E., and Fredricks, D.N.
751 (2011). Comparison of quantitative real time PCR with Sequencing and ribosomal
752 RNA-FISH for the identification of fungi in formalin fixed, paraffin-embedded tissue
753 specimens. *BMC Infect. Dis.* 11, 202. 10.1186/1471-2334-11-202.
- 754 90. Glöckner, F.O., Yilmaz, P., Quast, C., Gerken, J., Beccati, A., Ciuprina, A., Bruns, G.,
755 Yarza, P., Peplies, J., Westram, R., et al. (2017). 25 years of serving the community
756 with ribosomal RNA gene reference databases and tools. *J. Biotechnol.* 261, 169–176.
757 10.1016/j.jbiotec.2017.06.1198.
- 758 91. McDonald, D., Jiang, Y., Balaban, M., Cantrell, K., Zhu, Q., Gonzalez, A., Morton, J.T.,
759 Nicolaou, G., Parks, D.H., Karst, S.M., et al. (2024). Greengenes2 unifies microbial data
760 in a single reference tree. *Nat. Biotechnol.* 42, 715–718. 10.1038/s41587-023-01845-
761 1.
- 762 92. Fenhalls, G., Stevens-Muller, L., Warren, R., Carroll, N., Bezuidenhout, J., Van Helden,
763 P., and Bardin, P. (2002). Localisation of mycobacterial DNA and mRNA in human
764 tuberculous granulomas. *J. Microbiol. Methods* 51, 197–208. 10.1016/s0167-
765 7012(02)00076-3.
- 766 93. Loukil, A., Kirtania, P., Bedotto, M., and Drancourt, M. (2018). FISHing Mycobacterium
767 tuberculosis Complex by Use of a rpoB DNA Probe Bait. *J. Clin. Microbiol.* 56.
768 10.1128/JCM.00568-18.
- 769 94. Liu, J., Kang, R., and Tang, D. (2024). Lipopolysaccharide delivery systems in innate
770 immunity. *Trends Immunol.* 45, 274–287. 10.1016/j.it.2024.02.003.
- 771 95. Oumarou Hama, H., Aboudharam, G., Barbieri, R., Lepidi, H., and Drancourt, M.
772 (2022). Immunohistochemical diagnosis of human infectious diseases: a review. *Diagn.*
773 *Pathol.* 17, 17. 10.1186/s13000-022-01197-5.
- 774 96. Austin, G.I., and Korem, T. (2024). Planning and Analyzing a Low-Biomass Microbiome
775 Study: A Data Analysis Perspective. *J. Infect. Dis.* 10.1093/infdis/jiae378.
- 776 97. Zhao, L., Grimes, S.M., Greer, S.U., Kubit, M., Lee, H., Nadauld, L.D., and Ji, H.P. (2021).
777 Characterization of the consensus mucosal microbiome of colorectal cancer. *NAR*
778 *Cancer* 3, zcab049. 10.1093/narcan/zcab049.
- 779 98. Roelands, J., Kuppen, P.J.K., Ahmed, E.I., Mall, R., Masoodi, T., Singh, P., Monaco, G.,
780 Raynaud, C., de Miranda, N.F.C.C., Ferraro, L., et al. (2023). An integrated tumor,
781 immune and microbiome atlas of colon cancer. *Nat. Med.* 29, 1273–1286.
782 10.1038/s41591-023-02324-5.
- 783 99. Barb, J.J., Oler, A.J., Kim, H.-S., Chalmers, N., Wallen, G.R., Cashion, A., Munson, P.J.,
784 and Ames, N.J. (2016). Development of an Analysis Pipeline Characterizing Multiple
785 Hypervariable Regions of 16S rRNA Using Mock Samples. *PLoS ONE* 11, e0148047.
786 10.1371/journal.pone.0148047.

- 787 100. Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P.,
788 Parkhill, J., Loman, N.J., and Walker, A.W. (2014). Reagent and laboratory
789 contamination can critically impact sequence-based microbiome analyses. *BMC Biol.*
790 *12*, 87. 10.1186/s12915-014-0087-z.
- 791 101. Davis, N.M., Proctor, D.M., Holmes, S.P., Relman, D.A., and Callahan, B.J. (2018).
792 Simple statistical identification and removal of contaminant sequences in marker-
793 gene and metagenomics data. *Microbiome* *6*, 226. 10.1186/s40168-018-0605-2.
- 794 102. Liu, Y., Elworth, R.A.L., Jochum, M.D., Aagaard, K.M., and Treangen, T.J. (2022). De
795 novo identification of microbial contaminants in low microbial biomass microbiomes
796 with Squeegie. *Nat. Commun.* *13*, 6799. 10.1038/s41467-022-34409-z.