Comprehensive analysis of microbial content in whole-genome sequencing samples from The Cancer Genome Atlas project

Yuchen Ge[1,2,*], Jennifer Lu[1,3], Daniela Puiu[1,2], Mahler Revsine[1,4], and Steven L. Salzberg[1,2,3,4,*]

[1]Center for Computational Biology, Johns Hopkins University, Baltimore, Maryland, United States
[2]Department of Biomedical Engineering, Johns Hopkins University
[3]Department of Pathology, Johns Hopkins School of Medicine
[4]Department of Computer Science, Johns Hopkins University
[5]Department of Biostatistics, Johns Hopkins University
*Correspondence to: salzberg@jhu.edu, yge15@jhmi.edu

Abstract
In recent years, a growing number of publications have reported the presence of microbial species in human tumors and of mixtures of microbes that appear to be highly specific to different cancer types. Our recent re-analysis of data from three cancer types revealed that technical errors have caused erroneous reports of numerous microbial species reportedly found in sequencing data from The Cancer Genome Atlas (TCGA) project. Here we have expanded our analysis to cover all 5,734 whole-genome sequencing (WGS) data sets currently available from The Cancer Genome Atlas (TCGA) project, covering 25 distinct types of cancer. We analyzed the microbial content using updated computational methods and databases, and compared our results to those from two major recent studies that focused on bacteria, viruses, and fungi in cancer. Our results expand upon and reinforce our recent findings, which showed that the presence of microbes is far smaller than had been previously reported, and that most species identified in TCGA data are either not present at all, or are known contaminants rather than microbes residing within tumors. As part of this expanded analysis, and to help others avoid being misled by flawed data, we have released a dataset that contains detailed read counts for bacteria, viruses, archaea, and fungi detected in all 5,734 TCGA samples, which can serve as a public reference for future investigations.

Introduction
A number of recent studies have used the vast sequencing resource created by The Cancer Genome Atlas (TCGA) project to explore the potential role of microbial species in cancer. Although most of the TCGA data was collected with the goal of studying human genetic variation or gene expression, microbes present in the tumors–including viruses, bacteria, and fungi–might also be captured as an incidental side effect of sequencing experiments. Identifying microbes in a human tumor sample, in which the vast majority of the biomass is expected to be human in origin, requires great care in order not to be misled by contaminants, sequencing vectors, or other artifacts that might also be present in the data. In this study, our objective was to conduct an exhaustive and meticulous survey of microbial communities across thousands of whole-genome sequencing (WGS) samples from the TCGA project, with the goal of identifying any microbes within these samples. By making our results publicly available, we hope to spur additional research that may amplify or alternatively refute recent findings of microbiomes in a wide variety of tumor types.

We also compare our findings to two recent studies that used much of the same TCGA data and described findings that were, in some instances, substantially affected by contamination. Those studies and others that have relied on their data have implicated the microbiome in various aspects of cancer, from modulating the tumor microenvironment to influencing treatment

responses. In the first study, Poore *et al*. [1] analyzed 17,625 samples from TCGA, and reported that they were able to use machine learning algorithms to construct highly discriminative microbial signatures in 32 of 33 types of cancer. Their classifiers, which used combinations of bacteria, archaea, and viruses, were remarkably accurate, obtaining 95-100% accuracy at discriminating each of the 32 cancer types against all others. They reported additional models with similar accuracy at distinguishing tumors from matched normal samples in 15 cancer types. In the second study, Narunsky-Haziza *et al*. [2] analyzed 17,401 samples from the TCGA project and other sources with the goal of identifying associations between cancer and fungi, which the first study had not considered. They reported distinctive fungal signatures of cancer in most of the 35 cancer types they considered [2]. Our analysis here looks at many of the same TCGA samples in an effort to replicate some of these findings.

One source of over-counts when analyzing human samples for microbial content is data contamination in public genome databases. As described previously, the inadvertent inclusion of human DNA within microbial genomes has affected thousands of genomes [3]. When creating a microbial database to use in a microbiome study, it is crucial to be aware of this issue and to take rigorous steps to remove these computational contaminants, which otherwise will substantially skew the results of metagenomic analyses. Such contamination events can be especially problematic when working with low biomass samples, where the microbial content is expected to be a very small proportion of the total sample. This is precisely the scenario encountered when "mining" human DNA sequencing projects such as TCGA for microbial content.

We should emphasize further that even a tiny amount of contamination in a genome can lead to enormous over-counts of bacterial species, for the following reason. The most common source of human contamination in bacterial genome databases is high-copy human repeats such as Alus, LINEs, and SINEs, as described earlier [3]. Any WGS sample from human tissue is likely to contain large numbers of reads from these widespread repeats. Thus if one scans a human DNA sample against a bacterial genome that is contaminated with even one of these human repeats, many human reads will appear to match the bacterium. For a typical TCGA sample, it would not be surprising to find tens or even hundreds of thousands of reads incorrectly matching a bacterial genome in this circumstance [4].

Vector contamination, in which reads deriving from vectors such as manufacturer-specific sequencing primers make their way into a genome assembly, further compounds the challenges associated with metagenomic analyses. Sequences originating from vectors have inadvertently found their way into genome databases, where they might be labeled as bacteria, fungi, plants, or animals. As we describe below, some fungal genome sequences in public databases are contaminated with vector or adaptor sequences, which can lead to large numbers of false positive matches to a sample that was sequenced using the same vectors.

Results

We analyzed the microbial content of 5,734 WGS samples from TCGA, which comprised all of the available WGS samples as of late 2023. Despite the fact that TCGA also includes large numbers of RNA sequencing (RNA-seq) experiments, we excluded them because they used poly-A selection to capture messenger RNA. Bacterial transcripts do not have long poly-A tails [5] and will not be captured, except very rarely, with poly-A selection protocols. Thus any bacterial

sequences found in a human RNA-seq experiment are almost certain to be contaminants, and the inclusion of human RNA-seq data in a search for bacterial signatures, as has been done in the past [1], simply does not make sense.

We removed human sequences from the TCGA data by mapping the reads against both the GRCh38 reference genome and the CHM13 human genome (see Methods). As shown in Table 1, after removal of human sequences, the number of reads remaining in most samples was relatively small, averaging 2.6 million reads per sample (0.48% of the total). Across all samples, a total of 15 billion reads remained after two-pass filtering. Of these, we identified 2.44 billion as vector contaminants after classification with Kraken.

We used KrakenUniq [6] to classify the filtered reads against a microbial database containing 50,651 genomes representing 30,355 species including bacteria, archaea, viruses, and fungi (see Methods). Across all samples, we identified 11,349 non-eukaryotic microbial species that occurred at least once. Table S1 reports the number of sequencing reads detected for every species in each of the samples. Table S2 includes normalized counts for the data in Table S1, where read counts are divided by the total number of reads sequenced for each sample, in millions. Table S3 further normalizes the counts by dividing by the genome size of each species, in kilobases. The vast majority of the entries in these tables are zero: out of >65 million entries, only 1,882,423 (2.9%) have non-zero values, and only 214,338 (0.3%) have raw read counts of 10 or greater. We then classified reads that were not identified in the previous step against a comprehensive fungal genome database containing 557 species (see Methods). Table S4 reports the raw read counts for all 5,734 samples for each of these fungal species, with normalized counts reported in Tables S5-S6.

Interestingly, a number of samples showed indications of an acute infection with a single pathogen. For example, although only a handful of samples contained more than 100 reads from *Helicobacter pylori*, one sample contained >175,000 reads. That sample was from stomach cancer (STAD) data but the source was normal tissue rather than a tumor. Similarly, only 13 samples contained >10 reads from *Alphapapillomavirus* 9, a species that causes cervical cancer [7], but the highest count of 110 reads occurred in a tumor sample from the cervical cancer (CESC) dataset. (Note that for small genomes, such as the 8-kilobase papillomavirus genome, many fewer reads will be detected in a shotgun sequencing run as compared to bacterial genomes that are hundreds of times larger.) For *Fusobacterium nucleatum*, which has been associated with colon cancer, only 123 samples had >100 reads, but three samples had >100,000 reads with one having 372,665 reads, found in a tumor sample from head and neck cancer (HNSC).

Despite these examples of possible infections, we did not find evidence for a community of microbes residing in the samples from any cancer type.

Table 1. Overview of the 5,734 WGS samples analyzed in this study. Cancer type abbreviations are listed in Supplementary Table S7. Column 4 shows how many reads remained after removing reads that mapped to the GRCh38 human genome. Column 6 shows the number of reads remaining after a second pass further removed reads matching the CHM13 genome. Columns 8 and 10 show how many reads from column 6 were identified by Kraken as either human or vector.

Cancer type	Total # samples	Avg read count (millions)	Unmapped reads after mapping to GRCh38 (thousands)		Unmapped reads after mapping to GRCh38+CHM13 (thousands)		Kraken-identified human reads (thousands)		Kraken-identified vector reads (thousands)	
BLCA	288	284	5,983	(2.10%)	4,829	(1.70%)	199.5	(0.07%)	14	(0.00%)
BRCA	245	690	1,738	(0.25%)	1,532	(0.22%)	35.0	(0.01%)	418	(0.06%)
CESC	130	375	3,347	(0.89%)	2,675	(0.71%)	113.9	(0.03%)	83	(0.02%)
COAD	262	354	5,195	(1.47%)	4,496	(1.27%)	183.9	(0.05%)	745	(0.21%)
DLBC	14	926	381	(0.04%)	381	(0.04%)	0.1	(0.00%)	363	(0.04%)
ESCA	115	349	3,546	(1.02%)	2,620	(0.75%)	88.2	(0.03%)	584	(0.17%)
GBM	117	777	898	(0.12%)	758	(0.10%)	108.6	(0.01%)	28	(0.00%)
HNSC	335	388	5,334	(1.38%)	4,348	(1.12%)	235.4	(0.06%)	279	(0.07%)
KICH	100	869	454	(0.05%)	453	(0.05%)	0.2	(0.00%)	436	(0.05%)
KIRC	87	705	1,054	(0.15%)	1,054	(0.15%)	0.1	(0.00%)	400	(0.06%)
KIRP	77	866	340	(0.04%)	340	(0.04%)	0.2	(0.00%)	324	(0.04%)
LAML	110	742	3,165	(0.43%)	2,205	(0.30%)	264.0	(0.04%)	899	(0.12%)
LGG	185	444	3,546	(0.80%)	2,804	(0.63%)	98.6	(0.02%)	2	(0.00%)
LIHC	108	905	391	(0.04%)	391	(0.04%)	0.2	(0.00%)	368	(0.04%)
LUAD	577	473	4,808	(1.02%)	4,185	(0.88%)	126.3	(0.03%)	1,964	(0.42%)
LUSC	100	906	194	(0.02%)	194	(0.02%)	0.1	(0.00%)	2	(0.00%)
OV	121	747	454	(0.06%)	454	(0.06%)	0.3	(0.00%)	346	(0.05%)
PRAD	272	304	6,680	(2.20%)	5,152	(1.69%)	305.7	(0.10%)	15	(0.01%)
READ	120	283	6,620	(2.34%)	5,704	(2.02%)	230.2	(0.08%)	969	(0.34%)
SARC	82	733	181	(0.02%)	181	(0.02%)	0.1	(0.00%)	159	(0.02%)
SKCM	320	311	3,826	(1.23%)	3,134	(1.01%)	116.7	(0.04%)	632	(0.20%)
STAD	299	368	5,292	(1.44%)	3,877	(1.05%)	284.7	(0.08%)	28	(0.01%)
THCA	1,248	784	676	(0.09%)	527	(0.07%)	20.6	(0.00%)	43	(0.01%)
UCEC	320	383	4,590	(1.20%)	3,811	(1.00%)	290.7	(0.08%)	461	(0.12%)
UVM	102	156	3,599	(2.31%)	2,548	(1.64%)	103.1	(0.07%)	2	(0.00%)
Total	5,734	3,099,077	18,420,684	(0.59%)	14,998,914	(0.48%)	713,538	(0.02%)	2,444,357	(0.08%)
Average	229	540	3,213	0.59%	2,616	0.48%	124	0.02%	426	0.08%

Worth noting is that even after alignment against two human genomes, an average of ~124,000 reads per sample were still classified as human by the Kraken program. This highlights the necessity of including the human genome in any metagenomics database, even if the data are pre-filtered to remove human reads. Figure 1 shows the breakdown of bacterial, viral, and fungal read counts across the 25 cancer types.

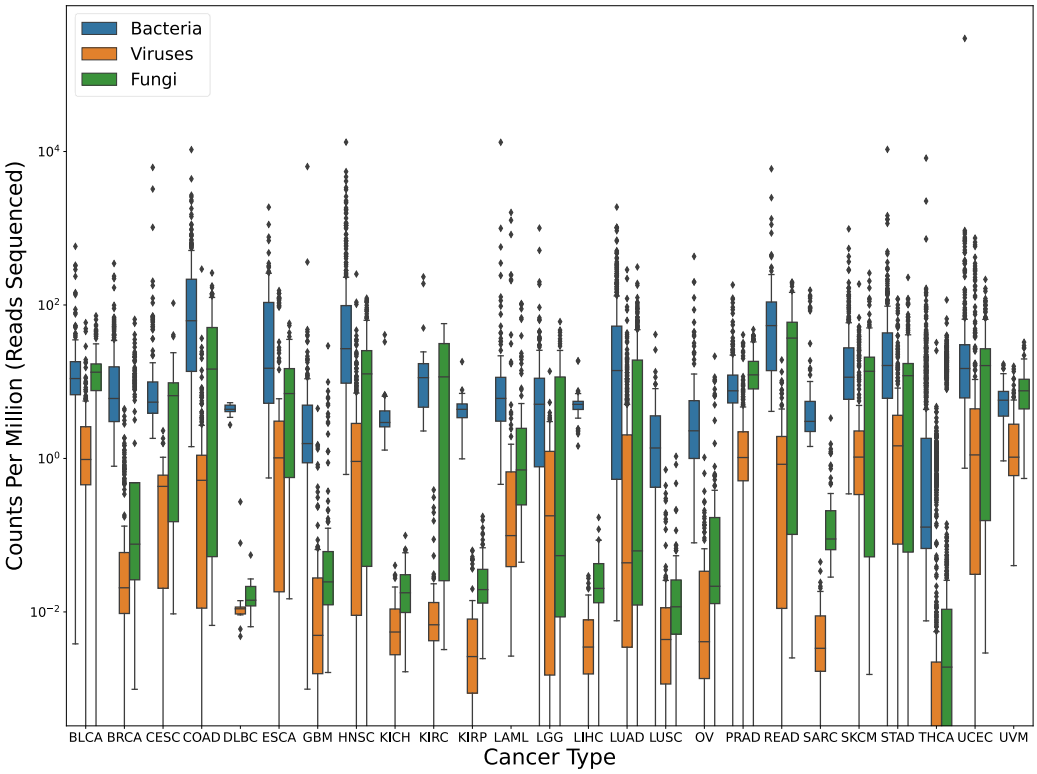

Figure 1. Microbial content found in TCGA data from 25 cancer types. Box plots show the inter-quartile range of read counts for the samples from each cancer type, with a horizontal line indicating the mean value. On average, bacterial reads (blue) had much higher counts than fungal (green), followed by viral (orange) across the board. Read counts were normalized by the number of reads sequenced in each sample, in millions; i.e., a value of 10 indicates 10 reads for every million reads in the sequencing run. Note that no adjustments were made for genome size here, which explains the much smaller read counts for viruses. For clarity, the y-axis uses a log scale.

Abundant bacteria, viruses, and fungi are likely to be contaminants

The most abundant species across all 5,734 samples, in our analysis, were *Delftia acidovorans*, *Rothia mucilaginsa*, *Human mastadenovirus C*, *Pseudomonas sp. J380*, *Escherichia coli*, *Bacteroides fragilis*, *Prevotella intermedia*, and other Delftia and Rothia species. Each of these is known to be present on human skin, in the oral cavity, or in the gut, and each was found in large numbers of both tumor and normal samples, ranging from 378 samples containing *Human mastadenovirus C* to 1,968 samples with *D. acidovorans* and 4,440 samples with *E. coli*. The mundane and thus most likely explanation is that all of these findings represent contamination, with possible sources including the sequencing facilities and the multiple people handling each of the samples. The most frequently observed fungal species, appearing in 2,312 samples, was *Saccharomyces cerevisiae*, a widely-used model organism that is not a human pathogen and that frequently appears as a cross-contaminant in sequencing centers.

Associations between microbes and cancer types

To determine if any of the microbial species detected in our analysis might be associated with cancer, we examined the most abundant microbes in each of the cancer types, using the normalized values in Table S3, and also looked specifically for bacteria and viruses that have known associations with cancer. The top species for each cancer type are shown in **Supplementary Figure S1**.

We noticed that after normalization, a few known cancer-associated microbes appeared relatively abundant, as expected: HPV was the most abundant species in the cervical cancer samples, *Bacteroides fragilis* was the top species for rectal adenocarcinoma (READ) and the fourth-most abundant microbe in colon adenocarcinoma (COAD), and *H. pylori* was the 17th-most common microbe in stomach adenocarcinoma (STAD). However, the vast majority of the abundant species across the 25 cancer types appear to represent simple contamination. For example, the bacterium *Cutibacterium acnes*, a common human skin bacterium and a common contaminant in sequencing projects, was abundant in most of the cancer types. Another abundant species was *Rosellinia necatrix partitivirus 8*, a virus that infects plant fungi and has no known associations with human disease [8,9].

A small number of abundant species might represent acute infections from the original patient samples. *Metamycoplasma salivarium*, an oral pathogen that may be pathogenic in immunocompromised individuals, was the top species in esophageal carcinoma [10], and this finding might merit further investigation. Overall, though, the microbes present in these samples are consistent with contamination, and we found no evidence to support claims of a microbiome–i.e., a community of microbes–in any of the cancer types.

Comparison of bacterial and viral read counts to previous reports

In a recent re-analysis [4] of TCGA sequence data from three cancer types–bladder cancer, breast cancer, and head and neck cancer–we found that an earlier study by Poore *et al*. of the same data [1] had reported read counts that were far too high. In particular, we found that >95% of the read counts in the earlier study were inflated by at least a factor of 10 [4]. Using the more-comprehensive data here, we can now confirm the previous findings and extend them to all 25 cancer types for which WGS data is available.

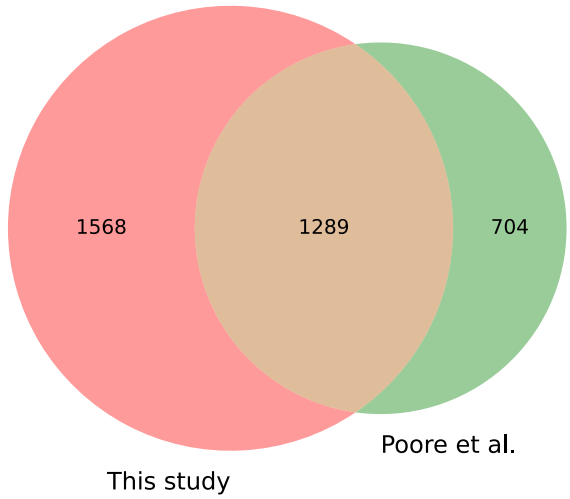

Figure 2. Comparison of bacterial, viral, and archaeal genera found in 4,550 TCGA samples versus those found in an earlier study. We found a total of 2,857 genera while a previous study [1] found 1,993, with 1,289 found in both studies.

The Poore *et al*. study [1] reported read counts for 1,993 microbial genera, including bacteria, archaea, and viruses. Out of the 5,734 WGS samples analyzed in this study, we identified 4,550 samples that matched those from the Poore *et al*. study; the remaining samples were added to TCGA more recently and thus were excluded from this comparison. Our analysis found reads from 2,857 genera across the same 4,550 samples, which included 1,289 found in the previous study. The union of the two sets yielded 3,561 genera found in one or both studies (Figure 2).

Supplementary Tables S8 and S9 contain read counts for all 4,550 samples and for the 3,561 genera found in either study, with Table S8 containing the results for this study and Table S9 showing the corresponding counts from Poore *et al*. We compared the findings by computing the ratio of counts (S9/S8), replacing any zero values with 1 to avoid division by zero. We then analyzed the ratios of read counts for all cells with a count ≥10 in at least one study, which comprised 2,197,510 entries.

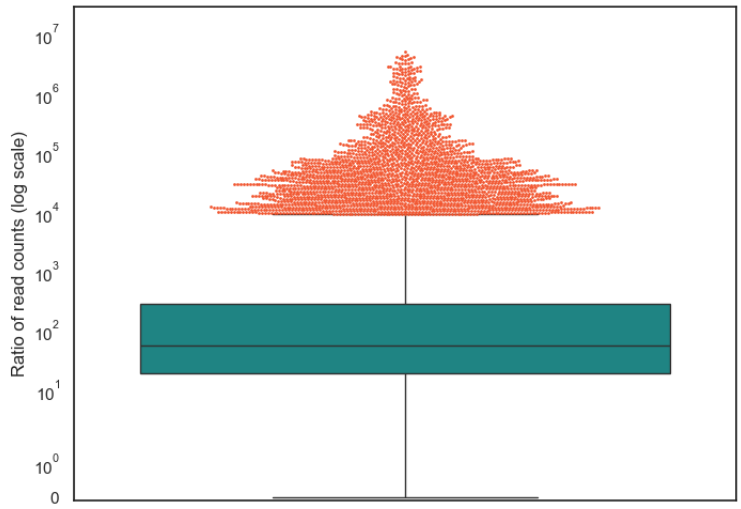

Figure 3. Distribution of ratios between read counts found by Poore et al. (Poore et al. 2020) and this study, for all samples in which either study found at least 10 microbial reads. The boxplot shows the median (56) and interquartile range. The top 5% of points, comprising 109,863 values, are shown as a cloud above the value 9225.

This analysis showed that the microbial read counts reported by Poore *et al*. were vastly higher than the counts found in this study. The median ratio was 56; i.e., half of the read counts in the earlier study were at least 56 times too high, and 90% of the values were at least 11 times too high. As shown in **Figure 3**, the top 5% (109,863) of the Poore *et al*. read counts are more than 9225-fold too high. The primary reason for these extreme over-estimates, as explained previously [4], was the use of a database containing thousands of incomplete ("draft") bacterial genomes, which themselves were contaminated with human sequences. As a result, millions of human reads in the TCGA data were mistakenly identified by Poore *et al*. as bacterial reads.

To provide another comparison, we identified the top 10 most-abundant microbial genera across all 4,550 WGS samples reported in the earlier study and compared them to our read counts for the same genera. As shown in **Figure 4**, the top three genera in Poore *et al*. were *Streptococcus, Mycobacterium, and Staphylococcus*, with average read counts per sample of 1,780,000, 1,400,000, and 922,000, respectively. In our re-analysis of the same samples, we found average read counts of 1129, 31, and 39 reads in those genera, values that range from 1,500 to 45,000 times smaller.

Only 2.3% of the values in our study were higher than those in the previous study. Among the 1,568 genera found exclusively in our study, the average read count per sample was just 0.81, suggesting that most are either false positives or low-level contaminants. The most abundant species found exclusively in our analysis was *Schaalia turicensis* (previously called *Actinomyces turicensis* [11]), a bacterium that is commonly found in the oral and gut microbiome.

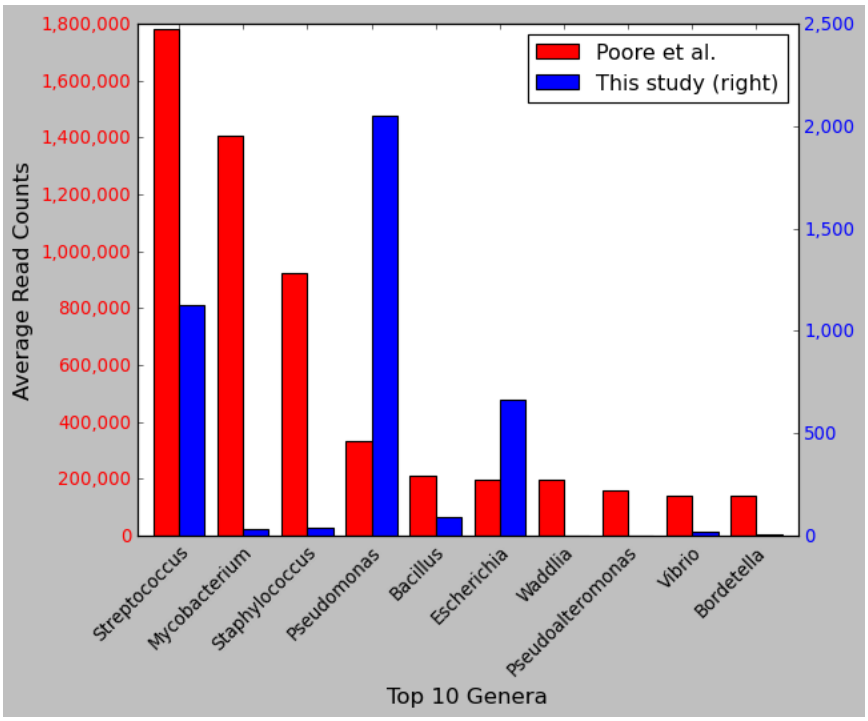

Figure 4. Comparison of read counts for the top 10 microbial genera from a previous study. Shown are the top ten most-abundant genera ranked by the average number of reads reported in Poore *et al*. [1] across 4,550 WGS samples from 25 cancer types. The left axis shows read counts as reported by Poore *et al*., and the right axis shows read counts as computed in our re-analysis of the same samples and the same species. Note that the y-axis scales differ by a factor of ~700. The x-axis shows genus names.

Comparison of fungal content to previous reports

In another recent study using the TCGA data, Narunsky-Haziza *et al*. [2] reported a strong association between mixtures of fungal species, which the authors called a "mycobiome," and multiple cancer types. Out of the 5,734 WGS samples analyzed here, we identified 4,271 that were identical to those used in the Narunsky-Haziza *et al*. study. We re-analyzed these samples for fungal content using an additional database that contained 557 fungal species, including all 224 of the species included in the Narunsky-Haziza study (see Methods). The full set of read counts for the 4,271 samples used in both studies can be found in Supplementary Table S10 and S11, where the counts in Table S11 were copied from Narunsky-Haziza *et al*. [2] Note that because we used a superset of the fungal species, we identified many reads from species not found in the previous study; however, these were observed at very low levels, with an average read count of 3.8 reads per sample in species unique to our analysis.

Although our average read counts were in rough agreement, we found highly divergent results for a small number of species that were estimated in Narunsky-Haziza *et al*. to be highly abundant. These are illustrated in **Figure 5**, which compares the maximum read counts from any sample for the top 10 most-abundant fungal species from Narunsky-Haziza *et al*. These include 2,013,180 reads for *Ramularia collo-cygni*, 656,503 reads for *Trichosporon asahii*, 101,344 reads for Candida albicans, and 54,641 reads for *Malassezia globosa*. In contrast, our counts for the same samples were 332, 4622, 266, and 4, values that range from 142-fold to 13,660-fold smaller.

Notably, the species with large over-counts were considered particularly important by Narunsky-Haziza *et al*., and were used by them to define three fungi-driven "mycotypes," labeled F1 (*Malassezia*, *Ramularia*, and *Trichosporon*), F2 (*Candida*, *Aspergillus*), and F3 (multiple genera including *Yarrowia*), which they reported were associated with distinct immune responses and overall survival [2]. Below we explain how contamination in the genomes themselves led to these over-counts. We suspect that if the raw counts are corrected, these mycotypes and their association with cancer will likely disappear.

Table 2. Contaminants found in fungal genomes in GenBank. Columns 1-3 show the species name, GenBank accession number of the genome, and accession number of the scaffold in which the contaminant was found. Columns 4-6 show the starting location of the contaminant within the scaffold, its length, and the correct source of the DNA sequence.

Organism	Accession	Scaffold Accession	Start	Length	Contamination Source
M. globosa	GCF_000181695.1	NW_001849834.1	1	897	Human
M. globosa	GCF_000181695.1	NW_001849877.1	1	557	Human
T. asahii	GCF_000293215.1	NW_014040855.1	924633	80	Vector
T. asahii	GCF_000293215.1	NW_014040868.1	203979	68	Vector
R. collo-cygni	GCF_900074925.1	NW_019716264.1	394,960	77	Vector
R. collo-cygni	GCF_900074925.1	NW_019716256.1	147,8654	75	Vector
C. albicans	GCF_000182965.3	NC_032096.1	1,888,747	8,469	Human

Contamination in the *Malassezia globosa* genome

The highest read count for *M. globosa* as reported by Narunsky-Haziza *et al*. was 54,641 reads from sample h2540, a blood-derived normal sample from the head and neck cancer (HNSC) dataset, in which our re-analysis found only 4 *M. globosa* reads. We subsequently aligned all reads (without filtering) from sample h2540 against the *M. globosa* reference genome (see Methods) and found a very large number of matches, nearly all aligning to just two locations: an 897-bp contig and a 557-bp contig (**Table 2**). We then used BLAST [12,13] to confirm that both contigs were human sequences that were mis-labeled as *M. globosa*. Note that although both of these contigs are still part of the *M. globosa* genome assembly (GenBank accession GCF_000181695.1), the individual records for them have been removed by NCBI because they were determined to be contaminated.

Contamination in the *Trichosporon asahii* genome

Trichosporon asahii also had an unusually high read count in the Narunsky-Haziza *et al*. study, which reported a maximum count of 656,503 reads in sample h1948. Upon aligning all reads (unfiltered) from that sample to the *Trichosporon* genome, we found an even higher number of matches, over 54 million; however, 99.99% of the matches hit the same 80-bp interval in the genome. We investigated and found that the 80-bp sequence is a perfect match to an Illumina sequencing vector (accession NCB00360.1); i.e., it is a contaminant in the genome assembly. The vector contaminant occurs in the middle of a large, 2.7 Mbp scaffold (NCBI accession NW_014040855.1, see **Table 2**). Also worth noting here is that sample h1948 was a failed sequencing run, in which over 95.4% of the read pairs (160 million out of a total of 179 million) were vector.

We found similar results for other samples; e.g., Narunsky-Haziza *et al*. reported 302,882 *Trichosoporon* reads in sample h1325, a lung cancer tumor sample. When we aligned the entire set of reads (unfiltered) from this sample to *T. asahii*, we found 12,235,713 matching reads, and all except for 18 reads aligned to the same 80-bp vector contaminant mentioned above. Thus for this and other samples, the large numbers of apparent matches to *Trichosporon* appear to have been the result of vector contamination in the fungal genome sequence.

Contamination in the *Ramularia collo-cygni* genome

Sample h1948, a normal tissue sample from the lung cancer (LUAD) dataset, was reported by Narunsky-Haziza *et al*. to contain 2,013,180 reads from *R. collo-cygni*, whereas we found only 322 reads. We re-aligned the filtered reads to the *R. collo-cygni* genome using the same Bowtie2 parameters as used by Narunsky-Haziza *et al*., and found ~34M matching reads, all matching just two locations: a 77-bp sequence on RCC_scaffold10 and a 75-bp sequence on RCC_scaffold02 (**Table 2**). Both of these sequences are identical to an Illumina PCR primer, albeit a different primer from the one that we found in *Trichosporon*. Thus these apparent matches to *Ramularia* appear to have been the result of vector contamination.

Contamination in the *Candida albicans* genome

The highest read count reported by Narunsky-Haziza *et al*. for *C. albicans* was 101,344 reads from sample h949, a primary tumor sample from the head and neck cancer (HNSC) dataset. We attempted to replicate this finding by processing these reads using the host depletion pipeline described in the original study, which left only 93,555 reads, a number already lower than the reported count for *C. albicans*. We then aligned these reads to the *C. albicans* genome using Bowtie2 [14], which detected only 544 matching reads. Using the same procedure, we analyzed the sample with the second highest *C. albicans* read count, h5103, a normal tissue sample from the lung cancer (LUSC) dataset with 75,799 reported matches, and found only 1,373 hits. We were similarly unable to replicate high read counts for *C. albicans* in other samples.

To attempt to explain the far higher read counts reported by Narunsky-Haziza *et al*., we aligned the entire set of unfiltered reads from the HNSC sample and the LUSC sample to the *C. albicans* genome, which yielded 147,874 and 273,683 matches, respectively, values much closer to the original report. We then determined that nearly all these reads matched an 8.5Kb region of the *C. albicans* genome (**Table 2**) that was clearly human in origin. Thus it appears that the high read counts for *C. albicans* in Narunsky-Haziza *et al*. were due to human contamination in the genome itself, rather than the presence of fungal DNA in the tumor samples.

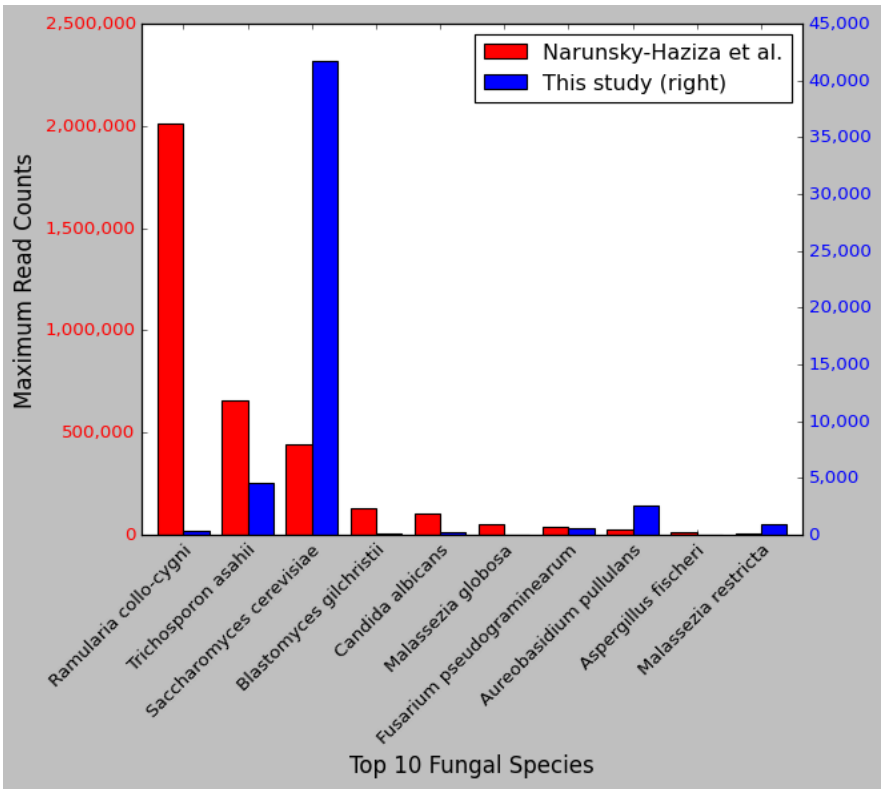

Figure 5. Comparison of read counts in fungal species. Shown are the top ten most-abundant fungal species ranked by the maximum number of reads reported in Narunsky-Haziza *et al*. [2], across 4,271 WGS samples from 25 cancer types. The left axis shows read counts as reported by Narunsky-Haziza *et al*., and the right axis shows read counts as computed in our re-analysis for the same samples and the same species. Note that the y-axis scales differ by a factor of 50. Note also that Narunsky-Haziza *et al*. reported a maximum of ~441K reads from one sample matching *S. cerevisiae*, a finding that we were unable to replicate. In our analysis of the same sample using reads filtered to remove human (see Methods), we found ~40K reads matching *S. cerevisiae*, as shown in the figure. These matches predominantly fell within a region of the yeast genome that contains the 18S rRNA gene and that has significant similarity to human, suggesting that these were human reads that slipped past the filtering step.

Discussion

Using updated computational methods and databases, we have analyzed the non-human content in a large collection of whole-genome sequencing data sets in TCGA covering 25 distinct types of cancer, and created a comprehensive dataset encompassing detailed read counts for bacteria, viruses, archaea, fungi, and other microbes in 5,734 samples from tumors and normal tissue. Our data show that the read counts reported in the previously-published dataset are greatly inflated, often exceeding the true counts by factors exceeding 1,000. As we explained previously [4], these over-counts can be attributed in part to the inclusion of numerous draft genomes (which themselves contain contaminants) in the database used for the metagenomic analysis. The work of Poore *et al*. [1] has already been used in at least a dozen other studies [15–26] that downloaded their read count data and then published findings based in part or entirely on that data, and another recent study [27] similarly based its results on the "mycobiome" data from Narunsky-Haziza *et al*. [2] Our analyses here used a much cleaner database, containing only complete bacterial genomes, which in turn yielded much lower and more accurate counts for the microbial species we identified. We hope that by providing a more accurate set of read counts for the same TCGA

samples, our new dataset can serve as a valuable public resource, enabling researchers to better distinguish genuine microbial signals from background noise and contaminants in future investigations.

We also extended our previous work to add a comparison with a recent study that focused on fungi in cancer, and that claimed to find highly specific fungal signatures in multiple cancer types [2] . Upon looking at the data published with that study, we discovered that several of the key fungal species used to create those signatures had excessively high read counts, and that those counts were the result of contamination in the fungal genome sequences. The contaminants included both human DNA and vector contaminants, either of which can lead to high numbers of false positives when doing metagenomic analysis.

Multiple other studies have recently reported microbiomes in cancer, including a widely publicized report of a fungal mycobiome in pancreatic cancer [28] that was only recently (in 2023) shown to be deeply flawed [29], likely due to mis-identification of fungal reads in the original data. Similarly, a 2020 study reported tumor type-specific microbes in seven types of cancer [30], but an attempt to replicate those findings in breast cancer found no evidence of microbes at all [31]. Taken together with the results described here, these reports suggest that claims regarding microbiomes and cancer need to be scrutinized more rigorously than they have been in the recent past, and that contamination of human samples with environmental microbes can easily be mistaken for a genuine signal.

Finally, we did serendipitously find a few individual microbes present in high abundance in some samples, and that these were either consistent with prior knowledge or worthy of further investigation. These findings can be assessed further using additional, independent experimental data.

Methods

We downloaded sequence data from the Genome Data Commons at the U.S. National Cancer Institute (gdc.cancer.gov) for 25 cancer types from the TCGA project. Metadata on all samples used in this study and previous studies, including unique sample identifiers from TCGA, can be found in Table S12. In total, data from 5,734 WGS samples were downloaded from the NCI data portal, which had aligned the reads to either hg19 or GRCh38 using bwa [32]. Of these samples, 2824 represented solid tumors, 569 were solid normal tissue, 64 were blood-derived cancer, and 2277 were blood-derived normal. We downloaded all reads that were unmapped by TCGA to their reference genome, which included the human genome (either hg19 or GRCh38, depending on the date of data collection) as well as human papilloma virus 16, HPV33, and Epstein-Barr virus. We then aligned the unmapped reads against the CHM13 human reference using Bowtie2 [14] to remove more human reads (**Table 1**).

We classified these two-pass filtered read sets using KrakenUniq [6] using its default parameters in paired-read mode, which treats each pair of reads as a single discontiguous sequence. We classified the two-pass filtered read pairs against a curated database, Microbial2023, containing all RefSeq bacterial, archaeal, and viral complete genomes, a collection of curated eukaryotic pathogen genomes from EuPathDB54, 10,798 vector sequences from UniVec and EmVec, the CHM13v2.0 genome, and the GRCh38.p14 human genome. In total, Microbial2023 contains

50,079 genomes (29,798 species) divided into 34,452 bacteria (16,150 species), 14,018 viruses (12,910 species), 534 archaea (422 species) and 503 eukaryotic microbes (316 species). The list of species and their GenBank accessions can be found in **Table S13**. The Microbial2023 database, which is 535GB in size, can be downloaded from https://benlangmead.github.io/aws-indexes/k2.

In order to conduct a more comprehensive search against fungi, we created a separate database containing all 572 fungal genomes (557 species) in the NCBI RefSeq database as of mid-2023, which we designate as Fungi_RefSeq. The list of species and their GenBank accessions can be found in **Table S14**. After processing each sample using Microbial2023, we extracted all reads that were either failed to match or were classified as fungi (taxid=4751), and screened them against Fungi_RefSeq using KrakenUniq with default parameters. These steps yield read counts for all 5,734 samples against the 557 fungal species.

To compare our bacterial, archaeal, and viral read counts to Poore et al.'s results, we mapped the sample metadata and species to Poore et al.'s, identifying 4,550 out of 5,734 TCGA samples that were analyzed in both studies, and 1,289 out of 1,993 microbial genera were reported in both studies. (Note that because Microbial2023 is a newer database, it has some species missing from the Poore et al. study. Conversely, because Microbial2023 only includes finished genomes while Poore et al. included draft genomes, some species and genera used in Poore et al. are missing from Microbial2023.) Similarly for Narunsky-Haziza et al.'s results, we matched 4,271 out of 5,734 TCGA samples. All 224 fungal species used in Narunsky-Haziza et al. were included in our re-analysis. A mapping of scientific name changes from RefSeq200 to RefSeq220 involving some of these fungi can be found in Table S15.

In our analysis of the high read counts in *M. globosa*, *T. asahii*, *R. collo-cygni*, and *C. albicans*, we aligned human-filtered reads from the samples against their respective reference genomes using the same Bowtie2 parameters as used by Narunsky-Haziza *et al*. (--end-to-end --very-sensitive -k 16 --np 1 --mp 1,1 --rdg 0,1 --rfg 0,1 --score-min L,0,-0.05).

Data Availability
All supplemental files and tables from this study are available at https://github.com/yge15/TCGA_Microbial_Content.

Acknowledgements. The authors wish to thank David Lipman, Mihaela Pertea, and Aleksey Zimin for helpful comments on earlier drafts of this manuscript. This work was supported in part by NIH under grants R35-GM130151 and R01-HG006677.

References

1. Poore, G. D. *et al.* Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579**, 567–574 (2020).
2. Narunsky-Haziza, L. *et al.* Pan-cancer analyses reveal cancer-type-specific fungal ecologies and bacteriome interactions. *Cell* **185**, 3789-3806.e17 (2022).

3. Breitwieser, F. P., Pertea, M., Zimin, A. V. & Salzberg, S. L. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res.* **29**, 954–960 (2019).
4. Gihawi, A. *et al.* Major data analysis errors invalidate cancer microbiome findings. *MBio* **14**, e0160723 (2023).
5. Dreyfus, M. & Régnier, P. The poly(A) tail of mRNAs. *Cell* **111**, 611–613 (2002).
6. Breitwieser, F. P., Baker, D. N. & Salzberg, S. L. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biol.* **19**, 198 (2018).
7. Chen, Z. *et al.* Evolution and taxonomic classification of human papillomavirus 16 (HPV16)-related variant genomes: HPV31, HPV33, HPV35, HPV52, HPV58 and HPV67. *PLoS One* **6**, e20183 (2011).
8. Sasaki, A., Miyanishi, M., Ozaki, K., Onoue, M. & Yoshida, K. Molecular characterization of a partitivirus from the plant pathogenic ascomycete Rosellinia necatrix. *Arch. Virol.* **150**, 1069–1083 (2005).
9. Chiba, S., Lin, Y.-H., Kondo, H., Kanematsu, S. & Suzuki, N. A novel victorivirus from a phytopathogenic fungus, Rosellinia necatrix, is infectious as particles and targeted by RNA silencing. *J. Virol.* **87**, 6727–6738 (2013).
10. Yacoub, E., Saed Abdul-Wahab, O. M., Al-Shyarba, M. H. & Ben Abdelmoumen Mardassi, B. The relationship between mycoplasmas and cancer: Is it fact or fiction ? Narrative review and update on the situation. *J. Oncol.* **2021**, 1–13 (2021).
11. Nouioui, I. *et al.* Genome-based taxonomic classification of the phylum Actinobacteria. *Front. Microbiol.* **9**, 2007 (2018).
12. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7**, 203–214 (2000).
13. Morgulis, A. *et al.* Database indexing for production MegaBLAST searches. *Bioinformatics* **24**, 1757–1764 (2008).
14. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
15. Hermida, L. C., Gertz, E. M. & Ruppin, E. Predicting cancer prognosis and drug response from the tumor microbiome. *Nat. Commun.* **13**, 2896 (2022).
16. Parida, S., Siddharth, S., Xia, Y. & Sharma, D. Concomitant analyses of intratumoral microbiota and genomic features reveal distinct racial differences in breast cancer. *NPJ Breast Cancer* **9**, 4 (2023).
17. Mao, A. W. *et al.* Identification of a novel cancer microbiome signature for predicting prognosis of human breast cancer patients. *Clin. Transl. Oncol.* **24**, 597–604 (2022).
18. Luo, M. *et al.* Race is a key determinant of the human intratumor microbiome. *Cancer Cell* **40**, 901–902 (2022).
19. Zhu, G. *et al.* Intratumour microbiome associated with the infiltration of cytotoxic CD8+ T cells and patient survival in cutaneous melanoma. *Eur. J. Cancer* **151**, 25–34 (2021).
20. Chen, F. *et al.* Integrating bulk and single-cell RNA sequencing data reveals the relationship between intratumor microbiome signature and host metabolic heterogeneity in breast cancer. *Front. Immunol.* **14**, 1140995 (2023).
21. Chen, C. *et al.* Pan-cancer analysis of microbiome quantitative trait loci. *Cancer Res.* **82**, 3449–3456 (2022).

22. Lim, D. M., Lee, H., Eom, K., Kim, Y. H. & Kim, S. Bioinformatic analysis of the obesity paradox and possible associated factors in colorectal cancer using TCGA cohorts. *J. Cancer* **14**, 322–335 (2023).
23. Bentham, R. *et al.* Using DNA sequencing data to quantify T cell fraction and therapy response. *Nature* **597**, 555–560 (2021).
24. Kim, Y. K. *et al.* Microbial and molecular differences according to the location of head and neck cancers. *Cancer Cell Int.* **22**, 135 (2022).
25. Xu, Y. *et al.* The microbiome types of colorectal tissue are potentially associated with the prognosis of patients with colorectal cancer. *Front. Microbiol.* **14**, 1100873 (2023).
26. Li, Y. *et al.* Intratumoral microbiota is associated with prognosis in patients with adrenocortical carcinoma. **2**, (2023).
27. Guan, S.-W., Lin, Q., Wu, X.-D. & Yu, H.-B. Weighted gene coexpression network analysis and machine learning reveal oncogenome associated microbiome plays an important role in tumor immunity and prognosis in pan-cancer. *J. Transl. Med.* **21**, (2023).
28. Aykut, B. *et al.* The fungal mycobiome promotes pancreatic oncogenesis via activation of MBL. *Nature* **574**, 264–267 (2019).
29. Fletcher, A. A., Kelly, M. S., Eckhoff, A. M. & Allen, P. J. Revisiting the intrinsic mycobiome in pancreatic cancer. *Nature* **620**, E1–E6 (2023).
30. Nejman, D. *et al.* The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science* **368**, 973–980 (2020).
31. de Miranda, N. F., Smit, V. T., Wesseling, J. & Neefjes, J. Tumor-type specific intracellular bacteria are undetectable in human breast cancer cells. (eLetter response to Nejman, D. *et al.* The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science* **368**, 973–980 (2020)).
32. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* (2009).

Supplementary figure captions

Figure S1. Top 10 microbial species in primary tumor samples for each of 24 cancer types. The X-axis shows the species names, sorted by maximum normalized counts, measured as reads per kilobase of genome per million reads sequenced. Notably Alphapapillomavirus 9 (α-9 HPV) appeared as the most abundant for CESC (cervical squamous cell carcinoma and endocervical adenocarcinoma). LAML (acute myeloid leukemia) is not included because TCAG contained no whole-genome sequencing tumor samples from that cancer type.

Supplementary table captions

Table S1. Raw counts of 5,734 TCGA WGS samples classified at the species level against the Kraken Microbial2023 database, excluding eukaryotes. The contents of Microbial2023 are described in the main text. This resulted in 11,349 species that have non-zero counts.

Table S2. Normalized counts of the values in Table S1, converted to counts per million reads sequenced.

Table S3. Normalized counts of the values in Table S2, converted to reads per kilobase of genome per million reads sequenced.

Table S4. Raw counts of 5,734 TCGA WGS samples classified at the species level against Fungi_RefSeq database. The contents of Fungi_RefSeq include all 557 fungal species in RefSeq as of late 2023.

Table S5. Normalized counts of TableS4, converted to counts per million reads sequenced.

Table S6. Normalized counts of TableS4, converted to reads per kilobase of genome per million reads sequenced.

Table S7. TCGA study abbreviations.

Table S8. Raw counts of 4,550 TCGA WGS samples classified at the genus level against Microbial2023 database, for samples used in both this study and in the Poore et al. study. This dataset contains counts for all 3,561 genera that had non-zero counts in either study.

Table S9. Raw counts from the Poore et al. study at the genus level, for samples used in both this study and Poore et al. This set contains 4,550 samples and 3,561 genera. Raw counts included here were taken directly from Poore et al. (2020).

Table S10. Raw counts of TCGA WGS samples classified at the species level against the Fungi_RefSeq database, for samples matched with the Narunsky-Haziza et al. study. This resulted in 4,271 samples and 557 species.

Table S11. Raw counts from the Narunsky-Haziza et al. study at the species level, for samples matched with this study. This resulted in 4,271 samples and 557 species. Raw counts included here were taken directly from Narunsky-Haziza et al. (2022).

Table S12. TCGA metadata for 5,734 samples, including unique IDs for this study, IDs used by Poore et al., IDs used by Narunsky-Haziza et al., and the original TCGA identifiers.

Table S13. List of species in the Microbial2023 Kraken database, with RefSeq accessions.

Table S14. List of species in the Fungi_RefSeq Kraken database, with RefSeq accessions.

Table S15. RefSeq v200 to v220 name conversion list, for 14 of the 557 fungal species whose names were changed between the two releases of RefSeq.