Major data analysis errors invalidate cancer microbiome findings

Abraham Gihawi,[1] Yuchen Ge,[2,3] Jennifer Lu,[2,3] Daniela Puiu,[2,3] Amanda Xu,[2] Colin S. Cooper,[1] Daniel S. Brewer,[1,4] Mihaela Pertea,[2,3,5] Steven L. Salzberg[2,3,5,6]

AUTHOR AFFILIATIONS See affiliation list on p. .

ABSTRACT We re-analyzed the data from a recent large-scale study that reported strong correlations between DNA signatures of microbial organisms and 33 different cancer types and that created machine-learning predictors with near-perfect accuracy at distinguishing among cancers. We found at least two fundamental flaws in the reported data and in the methods: (i) errors in the genome database and the associated computational methods led to millions of false-positive findings of bacterial reads across all samples, largely because most of the sequences identified as bacteria were instead human; and (ii) errors in the transformation of the raw data created an artificial signature, even for microbes with no reads detected, tagging each tumor type with a distinct signal that the machine-learning programs then used to create an apparently accurate classifier. Each of these problems invalidates the results, leading to the conclusion that the microbiome-based classifiers for identifying cancer presented in the study are entirely wrong. These flaws have subsequently affected more than a dozen additional published studies that used the same data and whose results are likely invalid as well.

IMPORTANCE Recent reports showing that human cancers have a distinctive microbiome have led to a flurry of papers describing microbial signatures of different cancer types. Many of these reports are based on flawed data that, upon re-analysis, completely overturns the original findings. The re-analysis conducted here shows that most of the microbes originally reported as associated with cancer were not present at all in the samples. The original report of a cancer microbiome and more than a dozen follow-up studies are, therefore, likely to be invalid.

B acteria and viruses have been implicated as the cause of multiple types of cancer, including human papillomavirus for cervical cancer (1), *Helicobacter pylori* for stomach cancer (2), and *Fusobacterium nucleatum* for colon cancer (3), among others. However, until a few years ago, little evidence indicated that a complex microbiome—a mixture of various bacteria and viruses—might affect the etiology of other cancer types. This changed after a large-scale analysis of 17,625 samples from the Cancer Genome Atlas (TCGA) reported that, in the sequence data from 33 types of cancer, a distinctive microbial signature was present in 32 of the cancer types (4). These signatures were remarkably accurate at discriminating between each tumor type and all other cancers. For 15 cancer types, signatures were created that could distinguish between tumor and normal tissue, and for 20 cancer types, signatures were developed to identify tumors based on microbial DNA found in the blood of those patients. The machine-learning models created in this study had surprisingly high accuracy, with most models ranging from 95 to 100% accurate.

However, despite efforts taken by Poore et al. to remove contaminating species and to avoid common biases like batch effects, we were concerned because many of the machine-learning models reported in the study were based on genera that did not make sense in the context of human disease. The models included species that had never been reported in humans, and that were associated only with extreme environments, ocean-dwelling species, plants, or other non-human environments.

Multiple studies over the past decade have reported that the problem of contamination is not limited to the physical samples themselves: in addition, genome databases are contaminated (in a different sense of the term) with large numbers of mislabeled sequences. The biggest problem, as reported in one study (5), is that human DNA has contaminated the assembled genomes of thousands of bacteria. An even larger study showed that cross-contamination with the wrong species is ubiquitous, affecting over 2 million entries in the GenBank database (6). These contamination events are predominantly present in draft genomes, where some sequences ("contigs") that originated from humans or other non-microbial species are mislabeled with the name of a bacterial, fungal, or other microbial species. Database contamination can, in turn, lead to misclassification of human reads that match a contaminated non-human genome.

This contamination problem is of particular concern when using a metagenomics analysis method to classify reads that are derived from a human sample and that have a relatively small number of microbial reads (7–9). Recent studies overturning findings of a human placental microbiome (10) and a blood microbiome (11) illustrate the perils of working with low biomass samples, which is precisely the scenario encountered in the cancer microbiome study (4), where original samples were collected from human tumors and normal tissue, in which the vast majority of the sequenced reads were human. Poore et al. (4) reported that 7.2% of the raw reads were classified as non-human, and we were concerned that a substantial fraction of those reads were, in fact, human. Our results below confirm that this concern was legitimate.

RESULTS

We re-analyzed all of the raw and normalized taxonomic classification data from the Poore et al. study (4), which included read counts that were summarized at the genus level. This included the counts for 1,993 genera for each of the 17,625 samples. Their raw count matrix was created by processing the data with Kraken, a metagenomics classification method developed originally in one of our labs (12, 13). In addition, we downloaded and re-analyzed 1,255 of the original TCGA samples from three cancer types: bladder urothelial carcinoma (BLCA), head and neck squamous cell carcinoma (HNSC), and breast invasive carcinoma (BRCA) (Table 1).

Filtered "non-human" reads contained millions of human reads

As described by Poore et al. (4), their analysis began with reads that did not align to known human reference genomes based on the mapping information in the raw BAM files from TCGA. Those BAM files were the results of using programs such as bwa (14) or Bowtie2 (15, 16) to align the reads against a version of the human reference genome, either GRCh37 (hg19) or GRCh38 depending on the date when the samples were processed. This alignment process is imperfect, and many human reads can fail to align from a typical sample. Thus, if one simply downloads the reads that do not align

TABLE 1 Number of cancer data sets downloaded from TCGA for re-analysis, from BLCA, HNSC, and BRCA[a]

Cancer type	Number of samples downloaded			Total read count	Initially unmapped reads
	WGS	RNA-seq	Total		
BLCA	277	406	683	207,716,524,420	5,025,973,203 (2.4%)
HNSC	334	0	334	258,961,253,944	3,573,898,240 (1.4%)
BRCA	238	0	238	324,824,097,837	1,532,210,153 (0.5%)

[a]The last column shows the number of reads that did not align to the human genome in the TCGA raw BAM files. WGS: whole-genome shotgun sequencing.

with the human genome, as Poore et al. did, many of the reads retrieved in this way will still be human.

To illustrate, we re-aligned all of the initially unmapped reads from the 1,255 BLCA, HNSC, and BRCA samples shown in the last column of Table 1. In the original data sets, the proportion of reads that originally did not map to the human genome was 2.4%, 1.4%, and 2.7%, respectively. We re-aligned these unmapped reads to the complete CHM13 human genome (17) using Bowtie2 (16) and identified 981,451,972 (19.5%), 519,222,095 (14.5%), and 785,947,157 (51%) additional reads that matched human. This equates to an average number of human reads per sample of 1.39 million, 1.55 million, and 3.3 million in the BLCA, HNSC, and BRCA data sets, respectively.

Thus, in each of these data sets, the strategy of relying on the mapping information in the raw BAM files, as Poore et al. did, leaves on average 1.4–3.3 million human reads in each sample.

Bacterial read counts were inflated by many orders of magnitude

The presence of millions of human reads in the samples means that in the primary analysis step in Poore et al., where they matched all of these reads to a microbial database, any microbial genome that contained short regions of human sequences could generate large numbers of false-positive matches; i.e., reads that were reported to match a bacterial genome when in fact the reads were from human DNA. As mentioned above, thousands of draft genomes do indeed contain small amounts of human DNA sequences that are erroneously labeled as bacterial (5).

The TCGA read data were analyzed with the Kraken program (13), a very fast algorithm that assigns reads to a taxon using exact matches of 31 base pairs (bp) or longer. The Kraken program is highly accurate, but it depends critically on the database of genomes to which it compares each read. Poore et al. used a database containing 59,974 microbial genomes, of which 5,503 were viruses and 54,471 were bacteria or archaea, including many draft genomes. Notably, their Kraken database did not include the human genome, nor did it include common vector sequences. This dramatically increased the odds for human DNA sequences present in the TCGA reads to be falsely reported as matching microbial genomes. This problem can be mitigated by including the human genome and using only complete bacterial genomes in the Kraken database.

Re-analysis of bladder cancer samples

We re-analyzed 156 bladder cancer samples (all whole-genome shotgun [WGS] primary tumor and normal tissue samples from BLCA) by matching them against a curated, publicly available Kraken database (18) that contained only finished bacterial genomes as well as viruses, eukaryotic pathogens, the human genome, and commonly used laboratory vectors (see Materials and Methods). None of the bacterial genomes in this database were draft genomes. We first re-filtered the unmapped reads by aligning them to the human CHM13 reference genome, and we only analyzed the reads remaining after this second filtering step (see Materials and Methods). Note that even with two rounds of alignment against the human genome, many of the reads in each sample (an average of 174,883, or 5%) were still classified as human by the Kraken program using our database.

Figure 1; Table S1 show the top 20 most-abundant microbial genera as reported in the Poore et al. study for BLCA, compared to the read counts found in our analysis.

As shown in the figure, the top genera from Poore et al. were *Streptococcus*, *Mycobacterium*, and *Staphylococcus*, with average read counts per sample of 560,000, 411,000, and 241,000, respectively. In our re-analysis of the same samples, we found far fewer reads in these genera: an average of 36, 6, and 266 reads, respectively, values that are 16,000, 67,000, and 900 times smaller. Table S2 shows the top 20 genera found in our analysis, which had abundances ranging from 10 to 447 reads per sample.

As we describe below, the vast majority of the excessive counts in the Poore et al. study were apparently due to human reads in the filtered data that were incorrectly labeled as bacterial. Because filtering the raw reads only against GRCh37 or GRCh38

did not remove all human reads, the input to their metagenomics pipeline included 1.4–3.3 million human reads per sample, and these reads explain the dramatic over-counts shown on the left side of Fig. 1.

We also compared the read counts for the genera that were given the highest weights in the machine-learning models created by Poore et al. Tables S3 and S4 show the average read counts for the 20 top-weighted genera in models that classified bladder cancer tumors vs other tumor types and tumors vs normal tissue. In our analysis, nearly all counts for these genera averaged between 0 and 1, with a maximum value of just 18 reads (for *Campylobacter*). Nearly half of the top-weighted genera had average read counts below 10 in the Poore et al. data as well, although several had counts in the thousands. Below we explain how genera with raw counts near zero were selected by the machine-learning models.

Re-analysis of head and neck cancer and breast cancer samples

We conducted the same re-analysis on 334 HNSC samples and 238 BRCA samples (see Materials and Methods). As with the BLCA samples, we filtered to remove reads matching the CHM13 human genome and then used the Kraken program to match all reads against a curated database of microbes, common vectors, and the human genome. Note that even after filtering by alignment to the human genome, an average of 227,272 (5%) and 34,806 (1%) of the reads in the HNSC and BRCA samples, respectively, were classified as human by Kraken.
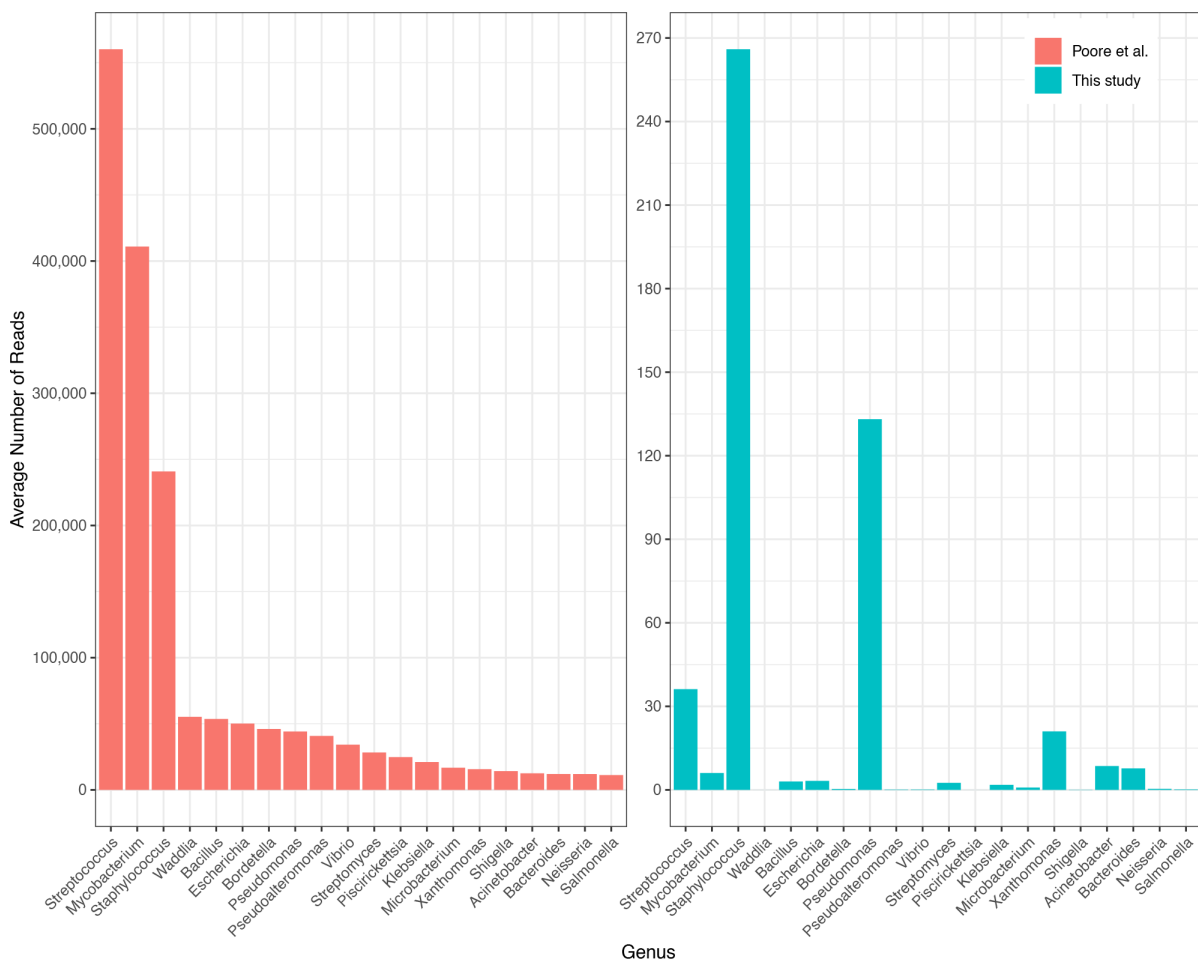

FIG 1   Average number of reads per sample in bladder cancer (BLCA) in the top 20 most-abundant genera reported in Poore et al. (left), averaged across 156 whole-genome sequencing samples. On the right are the counts for the same samples and the same genera, in the same order, as computed in our re-analysis. Note that the *y*-axis scales are different by a factor of 2,000. The *x*-axis shows genus names.

Figure S1 and S2, and Tables S5 and S6 show the read counts for the 20 most abundant genera in the HNSC data (Fig. S1) and BRCA data (Fig. S2) as computed by Poore et al. contrasted with our read counts for the same genera in the same samples. As with the bladder cancer results shown in Fig. 1, the average read counts reported by Poore et al. in both cancer types were consistently hundreds or thousands of times higher than found in our analysis. For most of these putatively abundant genera in all three cancer types, we found average read counts near zero, while Poore et al. reported read counts ranging from tens of thousands to over one million. As we demonstrate below, the vast majority of these over-counts are human reads that were erroneously assigned to bacteria.

Nearly all of the raw numbers in the Poore et al. study are incorrect and far too high

We then looked more broadly at all of the read counts, for all genera, for the BLCA, HNSC, and BRCA whole-genome samples, focusing on the non-zero counts reported by Poore et al. Note that in the original raw data matrix, 21,074,259 (60%) of the entries were zero and 14,071,985 (40%) were non-zero. Each cell in the matrix represents a sample/genus pair, i.e., the count of the number of reads from a given sample that were assigned to a given genus.

Table 2 summarizes our comparisons. For every non-zero genus in every sample, we compared the number of reads reported by Poore et al. to the number we found in our analysis. The table focuses on sample/genus pairs where the read counts were at least 10, on the assumption that smaller values likely represent noise or contamination.

As shown in the table, in the BLCA samples the number of reads reported by Poore et al. was at least 10 times larger than our results for 98.5% of the data entries. If we looked only at samples and genera where Poore et al. found ≥100 reads, their value was more than 10 times too large in 99.4% of all cases. The results were similar for HNSC, where 92.9% of values were at least 10 times too high, and in BRCA where 97.9% of the read counts were inflated at least 10-fold.

We also computed how many of the non-zero read counts were at least approximately the same as the value as determined in our re-analysis. In the BLCA samples, only 90 out of 37,258 (0.2%) were within 50% of the counts that we found in the same samples. Equivalently, fewer than 1 in 400 non-zero values in the bladder cancer data were within 50% of the value found upon re-analysis. The HNSC and BRCA read counts were only marginally better, with just 1.0% and 0.2%, respectively, within 50% of the correct value. Thus, the vast majority of the non-zero data in Poore et al.'s original data matrix—the data upon which all of their results were based—appears to be wrong, by very large amounts.

How human reads create the false appearance of bacteria

The likely reason for these vast over-counts is that human reads were erroneously categorized as bacterial by Poore et al. The number of human reads matching bacteria

TABLE 2 Microbial read totals found by Poore et al. (4) for three cancer types, compared to counts computed in a re-analysis using a database with only complete bacterial genomes[a]

Cancer type	Read counts reported in Poore et al.	Sample/ genus pairs	Poore et al. value	
			≥10 times too high	Within 50% of re-estimate
BLCA samples	≥10 reads	37,258	36,714 (98.5%)	90 (0.2%)
(156 total)	≥100 reads	16,969	16,869 (99.4%)	5 (0.03%)
HNSC samples	≥10 reads	114,640	106,544 (92.9%)	1,190 (1.0%)
(334 total)	≥100 reads	55,259	52,689 (95.3%)	274 (0.5%)
BRCA samples	≥10 reads	83,476	81,736 (97.9%)	172 (0.2%)
(238 total)	≥100 reads	46,025	45,686 (99.3%)	11 (0.02%)

[a]BLCA: bladder cancer; HNSC: head and neck cancer; and BRCA: breast cancer.

was unrelated to the actual presence of bacteria in the tumor sample; instead, it was determined by the database itself, in which many draft bacterial genomes contained mislabeled human sequences.

To illustrate how such high read counts can appear when few or even no reads from a bacterial genus are present, we did a deep analysis of two genera, *Streptococcus* and *Waddlia*, in one primary tumor sample, s2707 (case ID TCGA-DK-A1AB), from the BLCA data. We chose these genera because they were reported to be among the most abundant in Poore et al., as shown in Fig. 1; Fig. S1 and S2. Sample s2707 was reported by Poore et al. to have 327,985 *Streptococcus* reads and 20,673 *Waddlia* reads. When we aligned s2707 to our Kraken database, which only contains complete bacterial genomes, we found just one read labeled as *Streptococcus* and none labeled as *Waddlia*.

We then extracted all reads from s2707 that did not match the human genome, which comprised 11,997,726 unmapped reads. Next, we built a custom Kraken database containing all 10,270 *Streptococcus* genomes (including draft genomes) available in GenBank as of 2016. (We chose 2016 because Poore et al. downloaded all bacteria for their database in June 2016 [4].) We built a second Kraken database that contained all four of the *Waddlia* genomes (three of which are draft genomes) that are publicly available. We then ran KrakenUniq (13) to map the ~12 million unaligned reads from s2707 against both databases and found that 1,434,287 read pairs were classified as *Streptococcus* and 197,811 as *Waddlia*, respectively. This finding demonstrates that it is indeed possible, starting with the unaligned reads from a cancer sample (s2707), to find large numbers of reads from each of these genera when aligning against a database built entirely from bacterial genomes, as long as that database does not contain the human genome.

To confirm that the over-counts were due to human reads that erroneously matched bacteria, we then extracted all reads labeled as either *Streptococcus* or *Waddlia* in the Kraken analyses above and aligned them to the CHM13 human genome using Bowtie2 (16). This step revealed that 98.1% and 98.9% of the reads (respectively) matched human DNA. Thus, the Kraken matches were nearly all false positives, caused by the presence in the database of bacterial genomes that erroneously contained human sequences.

Finally, to emphasize the effect of omitting the human genome from the Kraken database, as Poore et al. did, we created two more databases: one containing the 10,270 *Streptococcus* genomes plus human, and one with the four *Waddlia* genomes plus human, using the CHM13 human genome in both cases. We then classified all reads from sample s2707 again. When classified against the first database, the number of *Streptococcus* reads dropped from 1,434,287 to 10,792, a 132-fold decline. When using the second database, the number of *Waddlia* reads dropped from 197,811 to 174, a decline of more than 1100-fold.

Normalization of the reads erroneously created a distinct signature for each cancer

The second major error in the Poore et al. study occurred during the normalization of the raw read counts. Poore et al. used normalized rather than raw data to build all of their machine-learning classifiers, in order to remove batch effects (4). In the process of converting the raw counts to normalized values, many of the cancer types (e.g., all tumor samples for one cancer type, or all healthy samples for another cancer type, etc.) were erroneously tagged with distinct values, marking the cancer samples even when the raw values were not informative. The machine-learning programs were then able to use these artificial tags to create near-perfect classifiers. We examined the top genera used in many of these classifiers and found numerous examples of this erroneous marking, a few of which are shown here.

First, consider the values for *Hepandensovirus* in adrenocortical carcinoma (ACC). All of the ACC cancer samples had raw read counts of zero for this virus, but during normalization, 71 of the 79 samples (90%) were assigned the value 3.078874655 by Poore et al. Out of all 17,625 samples across all cancer types (including 13,883 primary tumor

samples), only 77 other samples had a value equal to or smaller than this value in the normalized data. In the raw data, however, 17,624 samples had zero *Hepandensovirus* reads, and one sample had two reads.

As illustrated in Fig. 2, the extremely non-random distribution of normalized values— all but one of which started as raw values of zero—makes it easy for a machine-learning classifier to separate the ACC samples from other cancers. If we call the normalized *Hepandensovirus* value H_N, then if the model splits the samples using the simple rule $H_N > 3.078874655$, it will label 71/79 (90%) of the positive samples correctly, and only make 77/17,625 (0.4%) errors (Fig. 2). This explains why *Hepandensovirus* was the highest-weighted feature for the machine-learning model distinguishing ACC from other cancers in the most stringent decontamination (MSD) data set, despite the fact that only 1/17,624 samples had any reads at all matching this virus.

We observed a similar pattern in the normalized values of another genus, *Thiorhodospira*, for the Kidney chromophobe (KICH) tumor samples (Fig. 3). *Thiorhodospira* was the highest-weighted feature for the machine-learning classifier that distinguished KICH from normal tissue in several different models [including the Full data set, the "likely contaminants removed" data set, and the "all putative contaminants removed" (APCR) data set]. The TCGA-KICH data contained 51 tumor samples and 41 normal tissue samples, and in the raw data, 85 samples had read counts of zero, and 7 samples (four cancer, three normal) had counts of 1 for *Thiorhodospira*, meaning that it had almost no utility as a discriminating feature. In the normalized data, though, the cancer samples were assigned an almost perfectly disjoint set of values from the normal tissues, as shown in Fig. 3. Thus once again, the normalization process created an artificial signal separating the tumor from normal tissue.

Another example is *Nitrospira*, which was a highly weighted genus for the machine-learning models in 13 different cancer types in Poore et al., including lung squamous cell cancer (LUSC) where it was the top-ranked genus. Figure 4 shows the normalized counts of *Nitrospira* reads, after Voom-SNM normalization, in the LUSC samples compared to all other cancer types. In the figure, the frequencies of *Nitrospira* in LUSC are shifted to the right, i.e., they have larger values on average than other cancers. This explains why the machine-learning model gave *Nitrospira* the highest weight; however, in the raw data, there is no such shift to the right. Thus, the Voom-SNM normalization process created a signature of lung cancer even though no such signature was present in the original read counts.
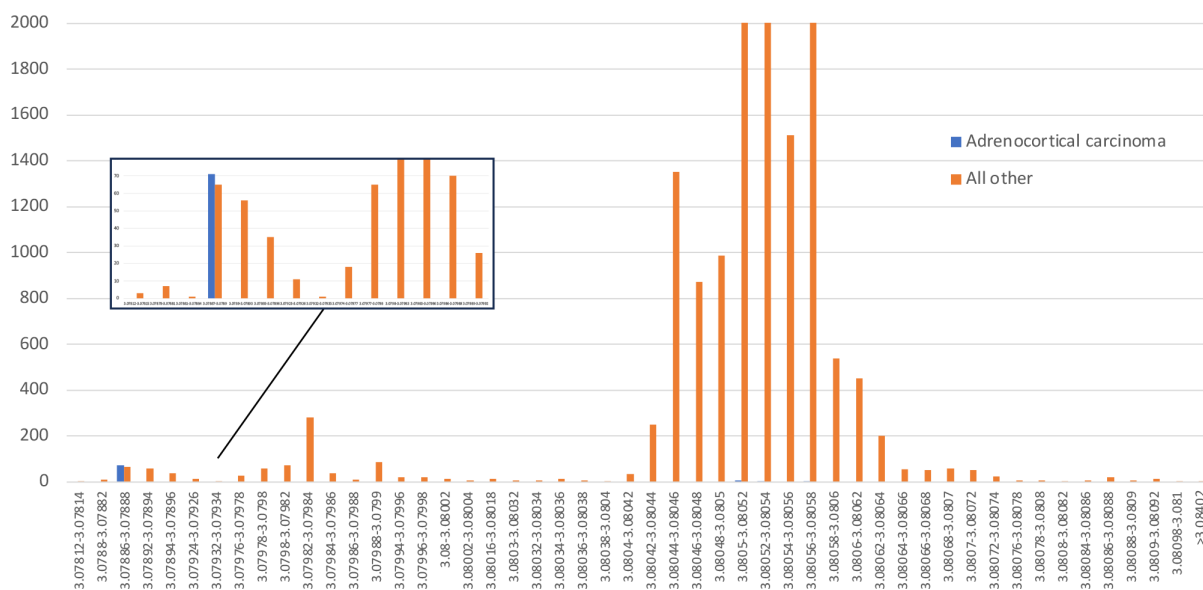

FIG 2 Distribution of normalized counts for *Hepandensovirus* for adrenocortical carcinoma (blue) vs all other samples (orange). The inset shows a zoomed-in view of the distribution for the small values. All raw values were zero.
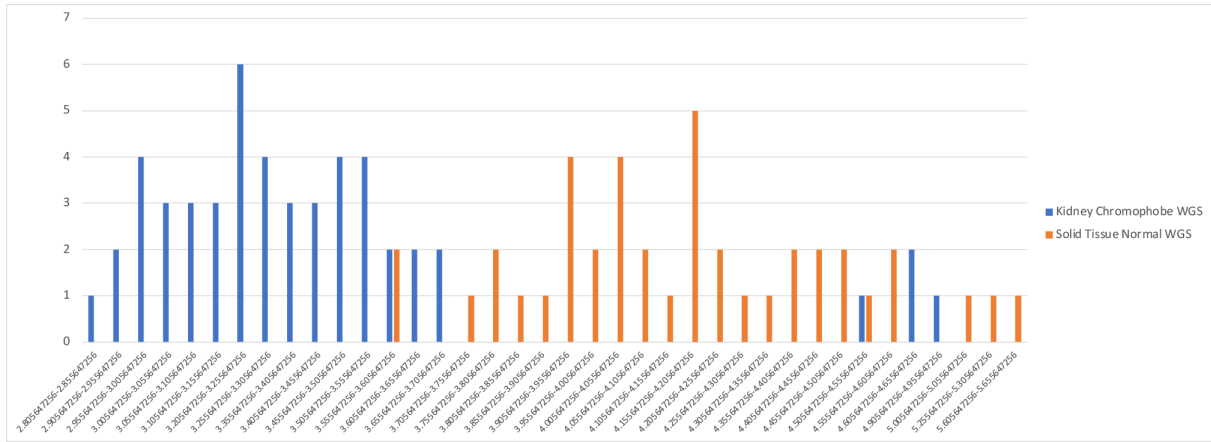
FIG 3 Distribution of normalized counts for *Thiorhodospira* reads in KICH cancer (blue) and normal (orange) samples. Nearly all raw values were zero except for seven samples with a raw count of 1.

We observed this phenomenon again in HNSC, where the genus with the highest weight in the MSD data set was *Mulikevirus*. This genus had the highest weight both for distinguishing tumor from normal tissue and for distinguishing HNSC from all other cancers. All 906 HNSC samples, including tumor, blood, and normal tissue, had zero reads in the raw data for *Mulikevirus*, making this virus useless at discriminating between tumor and normal samples.

However, in the Voom-SNM normalized data, values for the 70 normal samples were set to lower values than any of the tumor samples, as shown in Fig. 5. In particular, 38 samples had the identical normalized value of 3.07584214, 18 others had the value 3.07585718, and 5 had the value 3.076237397. The vast majority of the 693 tumor samples had larger values, as shown in the figure. Thus, a machine-learning model using this genus alone would have very high accuracy, which explains the very high weight given by the model to *Mulikevirus*, despite the fact that all the raw read counts were zero.

Replicating highly accurate classifiers on information-free raw data demonstrates flaws in the normalization process

Given that individual genera such as *Hepandensovirus* were erroneously tagged with tumor-type specific values, we wanted to explore how this tagging would affect the performance of machine-learning classifiers on a larger selection of tumor types and taxa. To investigate this question, we extracted a completely empty microbial-sample
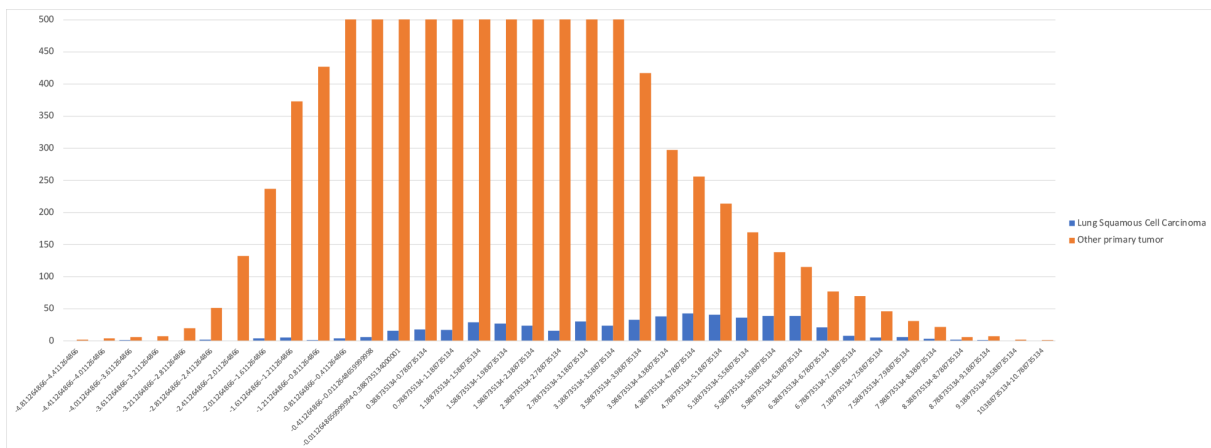

FIG 4 Distribution of normalized read counts in the APCR data set for *Nitrospira* reads found in lung squamous cell carcinoma (blue) and all other cancer types (orange). For clarity, the *y*-axis is truncated at 500, but the peak of the distribution for other cancers is at 1,389.
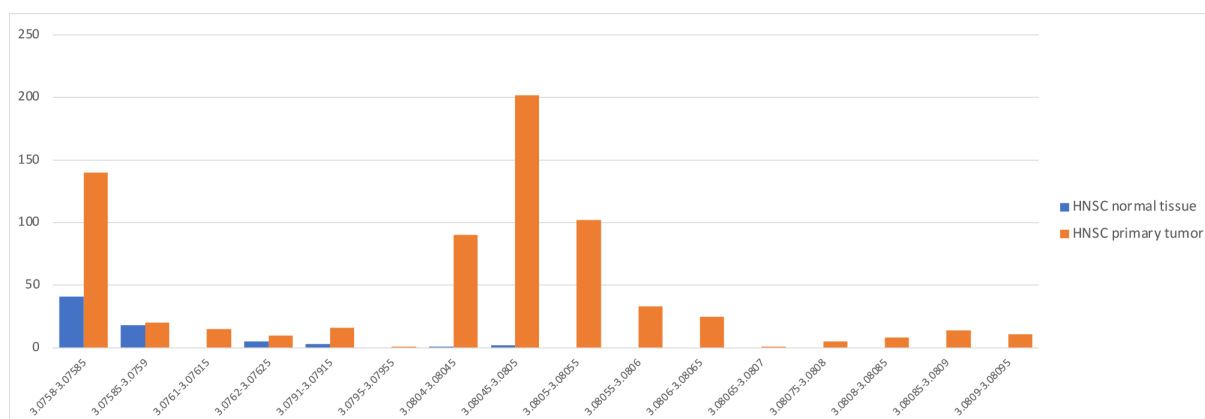
FIG 5 Distribution of normalized counts for *Mulikevirus* reads in head and neck squamous cell cancer (orange) and normal (blue) samples. All raw values were zero.

matrix (all zeros) from the raw Kraken classification data provided in the Poore et al. study. To obtain the empty matrix, we filtered the data to retain only genera present in fewer than 50 samples, and then removed any samples with non-zero values for any genus. This produced a matrix containing 16,567 samples and 170 genera in which all values were zero. No machine-learning classifier can use such data to discriminate among cancer types, because every entry in the matrix is identical.

We then populated each cell in the empty matrix with its corresponding value from the Voom-SNM normalized data. For this experiment, we used the Voom-SNM data from the "MSD" data set, which included only 66 of the 170 taxa in our initial empty matrix. We then filtered to retain only primary tumor samples (N = 12,803) so that we could attempt to build classifiers discriminating each cancer type from all others.

We then applied the original code provided by Poore et al. (4) to classify one tumor type vs all others and created classifiers for all 32 cancer types using this 12,803 × 66 matrix. Accuracies for these classifiers are shown in Fig. 6. Nearly all the models obtained very high accuracy, including a median (across all cancer types) sensitivity of 0.94, a median specificity of 0.9, and a median negative predictive value of 1.0. Several models obtained high positive predictive values (PPV) as well, including those for stomach adenocarcinoma (PPV = 0.65), ovarian serous cystadenocarcinoma (PPV = 0.91), and glioblastoma multiforme (PPV = 0.92). Comparing the model performance between these models and those reported in Poore et al., 14 out of 32 models had equal or improved accuracy as measured by area under the sensitivity-specificity curve.

Thus, despite the fact that the original raw data contained values of zero for all genera and all samples, remarkably high classification accuracy was obtained by the machine-learning classifiers, similar to the performance reported by Poore et al. All of the signals in these recreated models, therefore, must be artifactual, arising purely from the Voom-SNM normalization process, which the machine-learning methods exploited to create highly accurate classifiers despite the absence of any true signal.

We conclude that the Voom-SNM normalization, at least in the manner employed by Poore et al., inadvertently attached prior information about the tumor type to the normalized data. Note that we do not know precisely where Poore et al. went wrong in applying the normalization code, but because we have the original read count data and the resulting normalized data, we know that the transformation created the artificial signals that we describe here.

This result not only casts doubt on the claim that tumor types can be distinguished based on a microbial signature but it also raises concerns about the machine-learning models that distinguished between tumor and normal tissue and those based on microbial reads detected in blood samples.

Sensitivity	Specificity	AUC	PPV	NPV	
1.00	0.84	0.97	0.03	1.00	Uterine Carcinosarcoma
0.96	0.87	0.97	0.04	1.00	Mesothelioma
0.60	0.97	0.79	0.05	1.00	Cholangiocarcinoma
0.77	0.91	0.94	0.08	1.00	Skin Cutaneous Melanoma
0.94	0.90	0.98	0.08	1.00	Thymoma
0.94	0.86	0.96	0.09	1.00	Pancreatic Adenocarcinoma
1.00	0.92	0.98	0.09	1.00	Kidney Chromophobe
1.00	0.96	0.99	0.09	1.00	Lymphoid Neoplasm Diffuse Large B Cell Lymphoma
0.89	0.82	0.92	0.11	1.00	Kidney Renal Papillary Cell Carcinoma
0.89	0.84	0.94	0.11	1.00	Rectum Adenocarcinoma
0.89	0.78	0.91	0.14	0.99	Lung Squamous Cell Carcinoma
0.92	0.87	0.96	0.16	1.00	Cervical Squamous Cell Carcinoma And Endocervical Adenocarcinoma
0.97	0.95	0.99	0.17	1.00	Testicular Germ Cell Tumours
0.96	0.94	0.98	0.18	1.00	Pheochromocytoma and Paraganglioma
0.86	0.84	0.93	0.19	0.99	Bladder Urothelial Carcinoma
0.95	0.80	0.92	0.19	1.00	Prostate Adenocarcinoma
0.95	0.80	0.95	0.19	1.00	Thyroid Carcinoma
0.85	0.82	0.91	0.20	0.99	Lung Adenocarcinoma
0.83	0.90	0.94	0.20	0.99	Liver Hepatocellular Carcinoma
0.89	0.92	0.96	0.21	1.00	Sarcoma
0.87	0.84	0.93	0.23	0.99	Head and Neck Squamous Cell Carcinoma
0.79	0.86	0.92	0.26	0.98	Colon Adenocarcinoma
1.00	0.98	1.00	0.26	1.00	Adrenocortical Carcinoma
0.90	0.80	0.93	0.30	0.99	Breast Invasive Carcinoma
0.96	0.83	0.95	0.31	1.00	Kidney Renal Clear Cell Carcinoma
0.98	0.97	1.00	0.37	1.00	Esophageal Carcinoma
1.00	0.99	1.00	0.44	1.00	Uveal Melanoma
0.93	0.95	0.99	0.47	1.00	Brain Lower Grade Glioma
0.93	0.94	0.98	0.54	0.99	Uterine Corpus Endometrial Carcinoma
1.00	0.97	1.00	0.65	1.00	Stomach Adenocarcinoma
0.96	0.99	1.00	0.81	1.00	Ovarian Serous Adenocarcinoma
0.98	1.00	1.00	0.92	1.00	Glioblastoma Multiforme

FIG 6 Accuracies for one-vs-all tumor classification models obtained from a selection of samples and genera with zero classified reads prior to normalization. Each row shows the accuracies of a classifier that distinguished one cancer type from all other cancer types in the table. AUC: maximum measured area under the sensitivity-specificity curve. PPV: positive predictive value. NPV: negative predictive value.

Multiple other studies rely on the same flawed data

Since the publication of the study by Poore et al., more than a dozen studies have downloaded and used the Poore et al. data to find additional associations with the cancer microbiome, associations that in each case are likely to be invalid, because the underlying data are invalid. These include the following studies.

Hermida et al. (19) built predictive models for cancer prognosis in multiple cancer types using the Voom-SNM normalized data from Poore et al., which they used as the basis for creating machine-learning models to predict overall survival and progression-free survival for different cancer types. As shown above, the Voom-SNM data were flawed, introducing a distinctive signature into each cancer type even when the original read counts were all zeros. Thus, no classifiers based on this data can be considered valid.

A 2023 study by Parida et al. (20) reported finding distinct microbial communities in breast tumors from Asian, Black, and white women, based on the raw data matrix downloaded from Poore et al., which as described above has vastly over-inflated counts for nearly all genera. A number of the taxa highlighted as being important in this study are extremophiles (e.g., *Halonatronum* and *Salinarchaeum*), which are unlikely to be present in human samples.

Mao et al. (21) used the Voom-SNM normalized microbiome data from Poore et al. to create a predictive model of survival in breast cancer, based on the abundances of 94 genera. The study claims that its 15-microbe signature can predict overall survival and progression-free survival, but the model includes genera that are not known to exist in humans. For example, one genus is *Methanothermus*, an extremophile archaeon that lives in deep-sea hydrothermal vents at very high temperatures. This genus is extremely unlikely to be present in human breast cancers, and indeed no reads from this genus were found in our re-analysis.

Multiple other studies, including Luo et al. (22), Zhu et al. (23), Chen et al. (24), Chen et al. (25), Lim et al. (26), Bentham et al. (27), Kim et al. (28), Xu et al. (29), and Li et al. (30), have also utilized the Voom-SNMdata from Poore et al. to explore various aspects of the tumor microbiome and its potential associations with cancer. However, given the aforementioned flaws and inaccuracies in the Voom-SNM data, caution should be exercised when interpreting the results of any of these studies.

DISCUSSION

The original findings of a strong association between microbial species and 33 different cancer types were based on a large collection of DNA and RNA sequencing samples taken from human cancers and from matched normal tissues, which in turn was processed by a sophisticated machine-learning method to create highly accurate classifiers that could distinguish among tumor types and could distinguish tumor from normal tissue (4). Many of these classifiers used bacterial and viral genera that were not known to exist in humans, and therefore raised questions about their plausibility (31); however, this observation alone was not a fatal flaw. It did lead us to explore the machine-learning models more closely, though, in an effort to determine why organisms such as non-human extremophile microbes appeared as key features in the classifiers.

After re-analyzing all of the raw and transformed data, and after downloading and re-analyzing the original reads from more than 1,200 tumor and normal samples, we identified two major errors: first, the raw read counts were vastly overestimated for nearly every bacterial species, often by a factor of 1,000 or more. The likely cause of these overestimates was that the metagenomics database included thousands of draft genomes, which are known to be contaminated with human sequences. Consequently, as we showed above, millions of human reads were erroneously assigned to bacterial or archaeal genera. Second, the process of transforming the raw read counts into normalized values erroneously tagged many of the genera with values that were unique to specific cancer types. It is possible that this information leakage occurred during supervised normalization. When these values were fed to machine-learning classifiers, the algorithms discovered these artificial tags and built highly accurate classifiers, often using features (genera) that in the raw data had zero discriminative power. This error seems to have involved every tumor type and many genera that had zero or near-zero read counts across all of the human samples.

Either of these two errors suffices to invalidate the conclusions of the Poore et al. study and of the other studies that relied upon the same data. The original data matrix of raw read counts contained millions of wildly inaccurate values, and the normalized data compounded this error by tagging the cancer types with distinctive normalized values. Our conclusion after re-analysis is that the near-perfect association between microbes and cancer types reported in the study is, simply put, a fiction.

MATERIALS AND METHODS

We downloaded raw reads from the Genome Data Commons at the U.S. National Cancer Institute (gdc.cancer.gov) for three types of cancer from the TCGA project: bladder cancer, head and neck cancer, and breast cancer. These were chosen randomly from the 15 cancer types in Poore et al. for which both cancer and normal tissue samples were available. For TCGA-BLCA, we downloaded read data from 683 samples, which included

277 WGS samples and 406 RNA-seq samples (Table 1). We focused our re-analysis on the WGS samples, which included 129 primary tumor and 27 solid-tissue normal samples. All reads had been previously aligned by the TCGA project to either GRCh38 or GRCh37/hg19 using bwa (14). We extracted all unmapped reads and re-aligned them against the CHM13 human genome using Bowtie2 (16) to remove additional human reads and created new files for further downstream analysis. Note that for the in-depth analysis of sample s2707, we used the unmapped reads after alignment to hg19, as was done in the Poore et al. study.

For TCGA-HNSC, we downloaded the raw reads from 334 WGS samples, which included 24 solid-tissue normal, 140 blood-derived normal, and 170 primary tumor samples. As with BLCA, we focused the analysis on WGS samples. For TCGA-BRCA, we downloaded the unmapped reads from all 238 available WGS samples in TCGA (Table 2), which included 114 primary tumors, 106 blood-derived normal, 16 solid-tissue normal, and 2 metastatic samples. For both HNSC and BRCA, we ran the same two-pass filtering as for the BLCA samples, re-aligning all unmapped reads against CHM13. The average read lengths for the BLCA, HNSC, and BRCA samples from the WGS samples used here were 67, 83, and 98 bp, respectively.

Using these two-pass filtered files, for all samples in the BLCA, HNSC, and BRCA data, we ran KrakenUniq (13) against a customized database built from all complete genomes of bacteria and viruses from RefSeq that contained 46,711 bacterial genomes (5,981 species), 13,011 viral genomes (9,905 species), and 604 archaeal genomes (295 species). It also included a collection of 246 eukaryotic pathogens from EuPathDB (32), the UniVec set of standard laboratory vectors from NCBI (https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/), and the GRCh38 human genome. This 384 GB KrakenUniq database is available for download from https://benlangmead.github.io/aws-indexes/k2. Files containing lists of all species, genera, and NCBI accession numbers in this database are available as Data files 1 to 3. All supplemental files and tables from this study are available at https://github.com/yge15/Cancer_Microbiome_Reanalyzed.

Tables S8 to S10 contain the read counts at the genus level for all non-zero bacteria, archaea, and viruses found in our re-analysis of the BLCA, HNSC, and BRCA data and summarized in Table 2. Note that even though these numbers are far smaller than those reported in Poore et al., they likely still contain some false positives and should be regarded as upper bounds on the actual number of reads from each genus. Table S11 contains the top 25 genera identified by the machine-learning classifiers created in the Poore et al. study, downloaded from http://cancermicrobiome.ucsd.edu/CancerMicrobiome_DataBrowser. These include all the classifiers for the APCR and MSD data sets (separate classifiers were created for each data set) and for those data sets, the table includes the top genera used for classifying one cancer type vs all others and for distinguishing tumor from normal tissue.

application of biomarker bacterial genera in prostate cancer. All other authors declare no conflicts.

AUTHOR AFFILIATIONS

[1]Norwich Medical School, University of East Anglia, Norwich, United Kingdom
[2]Center for Computational Biology, Johns Hopkins University, Baltimore, Maryland, USA
[3]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, USA
[4]Earlham Institute, Norwich Research Park, Colney Lane, Norwich, United Kingdom
[5]Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, USA
[6]Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland, USA

AUTHOR ORCIDs

Steven L. Salzberg http://orcid.org/0000-0002-8859-7432

AUTHOR CONTRIBUTIONS

Abraham Gihawi, Conceptualization, Formal analysis, Investigation, Software, Visualization, Writing – original draft, Writing – review and editing | Yuchen Ge, Data curation, Formal analysis, Software, Writing – review and editing | Jennifer Lu, Data curation, Formal analysis, Software, Writing – review and editing | Daniela Puiu, Data curation, Formal analysis, Software | Amanda Xu, Data curation | Colin S. Cooper, Conceptualization, Methodology, Writing – review and editing | Daniel S. Brewer, Conceptualization, Methodology, Writing – review and editing | Mihaela Pertea, Conceptualization, Methodology, Writing – review and editing, Supervision | Steven L. Salzberg, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Supervision, Visualization, Writing – original draft, Writing – review and editing

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Data File S1 (mBio01607-23-s0001.txt). All species in the Kraken database used in this study.

Data File S2 (mBio01607-23-s0002.txt). All genera in the Kraken database used in this study.

Supplemental tables and figures (mBio01607-23-s0003.docx). Tables S1-S7 and Figures S1 and S2, captions to Tables S8-S10, and legends for Data Files S1 and S2.

Table S8 (mBio01607-23-s0004.xlsx). Read counts for 156 bladder cancer samples.

Table S9 (mBio01607-23-s0005.xlsx). Read counts for 334 head and neck cancer samples.

Table S10 (mBio01607-23-s0006.xlsx). Read counts for 238 breast cancer samples.

Table S11 (mBio01607-23-s0007.xlsx). Top-ranked features for machine learning models for cancer classification.

REFERENCES

1. Bosch FX, Lorincz A, Muñoz N, Meijer C, Shah KV. 2002. The causal relation between human papillomavirus and cervical cancer. J Clin Pathol 55:244–265. https://doi.org/10.1136/jcp.55.4.244

2. Warren JR, MarshallB. 1983. Unidentified curved bacilli on gastric epithelium in active chronic gastritis. Lancet 1:1273–1275. https://doi.org/10.1016/S0140-6736(83)92719-8

3. Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, Barnes R, Watson P, Allen-Vercoe E, Moore RA, Holt RA. 2012. *Fusobacterium nucleatum* infection is prevalent in human colorectal carcinoma. Genome Res 22:299–306. https://doi.org/10.1101/gr.126516.111

4. Poore GD, Kopylova E, Zhu Q, Carpenter C, Fraraccio S, Wandro S, Kosciolek T, Janssen S, Metcalf J, Song SJ, Kanbar J, Miller-Montgomery S, Heaton R, Mckay R, Patel SP, Swafford AD, Knight R. 2020. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nat 579:567–574. https://doi.org/10.1038/s41586-020-2095-1

5. Breitwieser FP, Pertea M, Zimin AV, Salzberg SL. 2019. Human contamination in bacterial genomes has created thousands of spurious proteins. Genome Res 29:954–960. https://doi.org/10.1101/gr.245373.118

6. Steinegger M, Salzberg SL. 2020. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in genbank. Genome Biol 21:115. https://doi.org/10.1186/s13059-020-02023-1

7. Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, Salamat SM, Somasekar S, Federman S, Miller S, Sokolic R, Garabedian E, Candotti F, Buckley RH, Reed KD, Meyer TL, Seroogy CM, Galloway R, Henderson SL, Gern JE, DeRisi JL, Chiu CY. 2014. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. N Engl J Med 370:2408–2417. https://doi.org/10.1056/NEJMoa1401268

8. Salzberg SL, Breitwieser FP, Kumar A, Hao H, Burger P, Rodriguez FJ, Lim M, Quiñones-Hinojosa A, Gallia GL, Tornheim JA, Melia MT, Sears CL, Pardo CA. 2016. Next-generation sequencing in neuropathologic diagnosis of infections of the nervous system. Neurol Neuroimmunol Neuroinflamm 3:e251. https://doi.org/10.1212/NXI.0000000000000251

9. Li Z, Breitwieser FP, Lu J, Jun AS, Asnaghi L, Salzberg SL, Eberhart CG. 2018. Identifying corneal infections in formalin-fixed specimens using next generation sequencing. Invest Ophthalmol Vis Sci 59:280–288. https://doi.org/10.1167/iovs.17-21617

10. Kennedy KM, de Goffau MC, Perez-Muñoz ME, Arrieta M-C, Bäckhed F, Bork P, Braun T, Bushman FD, Dore J, de Vos WM, Earl AM, Eisen JA, Elovitz MA, Ganal-Vonarburg SC, Gänzle MG, Garrett WS, Hall LJ, Hornef MW, Huttenhower C, Konnikova L, Lebeer S, Macpherson AJ, Massey RC, McHardy AC, Koren O, Lawley TD, Ley RE, O'Mahony L, O'Toole PW, Pamer EG, Parkhill J, Raes J, Rattei T, Salonen A, Segal E, Segata N, Shanahan F, Sloboda DM, Smith GCS, Sokol H, Spector TD, Surette MG, Tannock GW, Walker AW, Yassour M, Walter J. 2023. Questioning the fetal microbiome illustrates pitfalls of low-biomass microbial studies. Nat 613:639–649. https://doi.org/10.1038/s41586-022-05546-8

11. Tan CCS, Ko KKK, Chen H, Liu J, Loh M, Consortium SGKH, Chia M, Nagarajan N. 2023. No evidence for a common blood microbiome based on a population study of 9,770 healthy humans. Nat Microbiol 8:973–985. https://doi.org/10.1038/s41564-023-01350-w

12. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46. https://doi.org/10.1186/gb-2014-15-3-r46

13. Breitwieser FP, Baker DN, Salzberg SL. 2018. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. Genome Biol 19:198. https://doi.org/10.1186/s13059-018-1568-0

14. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinf 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324

15. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25. https://doi.org/10.1186/gb-2009-10-3-r25

16. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923

17. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, Aganezov S, Hoyt SJ, Diekhans M, Logsdon GA, Alonge M, Antonarakis SE, Borchers M, Bouffard GG, Brooks SY, Caldas GV, Chen N-C, Cheng H, Chin C-S, Chow W, de Lima LG, Dishuck PC, Durbin R, Dvorkina T, Fiddes IT, Formenti G, Fulton RS, Fungtammasan A, Garrison E, Grady PGS, Graves-Lindsay TA, Hall IM, Hansen NF, Hartley GA, Haukness M, Howe K, Hunkapiller MW, Jain C, Jain M, Jarvis ED, Kerpedjiev P, Kirsche M, Kolmogorov M, Korlach J, Kremitzki M, Li H. 2022. The complete sequence of a human genome. Sci 376:44–53. https://doi.org/10.1126/science.abj6987

18. Lu J, Salzberg SL. 2018. Removing contaminants from databases of draft genomes. PLoS Comput Biol 14:e1006277. https://doi.org/10.1371/journal.pcbi.1006277

19. Hermida LC, Gertz EM, Ruppin E. 2022. Predicting cancer prognosis and drug response from the tumor microbiome. Nat Commun 13:2896. https://doi.org/10.1038/s41467-022-30512-3

20. Parida S, Siddharth S, Xia Y, Sharma D. 2023. Concomitant analyses of intratumoral microbiota and genomic features reveal distinct racial differences in breast cancer. NPJ Breast Cancer 9:4. https://doi.org/10.1038/s41523-023-00521-6

21. Mao AW, Barck H, Young J, Paley A, Mao JH, Chang H. 2022. Identification of a novel cancer microbiome signature for predicting prognosis of human breast cancer patients. Clin Transl Oncol 24:597–604. https://doi.org/10.1007/s12094-021-02725-3

22. Luo M, Liu Y, Hermida LC, Gertz EM, Zhang Z, Li Q, Diao L, Ruppin E, Han L. 2022. Race is a key determinant of the human intratumor microbiome. Cancer Cell 40:901–902. https://doi.org/10.1016/j.ccell.2022.08.007

23. Zhu G, Su H, Johnson CH, Khan SA, Kluger H, Lu L. 2021. Intratumour microbiome associated with the infiltration of cytotoxic CD8+ T cells and patient survival in cutaneous melanoma. Eur J Cancer 151:25–34. https://doi.org/10.1016/j.ejca.2021.03.053

24. Chen F, Yang J, Guo Y, Su D, Sheng Y, Wu Y. 2023. 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. https://doi.org/10.3389/fimmu.2023.1140995

25. Chen C, Cai Y, Liu Y, Chen S, Li Y, Zhang F, Zhang M, Lu Z, Ying P, Huang J, Fan L, Cai X, Ning C, Wang W, Jiang Y, Zhang H, Yang S, Wang Z, Wang X, Zhang S, Huang C, Xu B, Fu Z, Song Q, Jin M, Chen K, Chen H, Dai M, Miao X, Yang X, Zhu Y, Tian J. 2022. Pan-cancer analysis of microbiome quantitative trait loci. Cancer Res 82:3449–3456. https://doi.org/10.1158/0008-5472.CAN-22-1854

26. Lim DM, Lee H, Eom K, Kim YH, Kim S. 2023. Bioinformatic analysis of the obesity paradox and possible associated factors in colorectal cancer

using TCGA cohorts. J Cancer 14:322–335. https://doi.org/10.7150/jca.80977

27. Bentham R, Litchfield K, Watkins TBK, Lim EL, Rosenthal R, Martínez-Ruiz C, Hiley CT, Bakir MA, Salgado R, Moore DA, Jamal-Hanjani M, ConsortiumTR, Swanton C, McGranahan N. 2021. Using DNA sequencing data to quantify T cell fraction and therapy response. Nat 597:555–560. https://doi.org/10.1038/s41586-021-03894-5

28. Kim YK, Kwon EJ, Yu Y, Kim J, Woo S-Y, Choi H-S, Kwon M, Jung K, Kim H-S, Park HR, Lee D, Kim YH. 2022. Microbial and molecular differences according to the location of head and neck cancers. Cancer Cell Int 22:135. https://doi.org/10.1186/s12935-022-02554-6

29. Xu Y, Zhao J, Ma Y, Liu J, Cui Y, Yuan Y, Xiang C, Ma D, Liu H. 2023. The microbiome types of colorectal tissue are potentially associated with the prognosis of patients with colorectal cancer. Front Microbiol 14:1100873. https://doi.org/10.3389/fmicb.2023.1100873

30. Li Y, Zhang D, Wang M, Jiang H, Feng C, Li Y-X. 2023. Intratumoral microbiota is associated with prognosis in patients with adrenocortical carcinoma. iMeta 2:e102. https://doi.org/10.1002/imt2.102

31. Gihawi A, Cooper CS, Brewer DS. 2023. Caution regarding the specificities of pan-cancer microbial structure. Microb Genom 9. https://doi.org/10.1099/mgen.0.001088

32. Warrenfeltz S, Basenko EY, Crouch K, Harb OS, Kissinger JC, Roos DS, Shanmugasundram A, Silva-Franco F. 2018. EuPathDB: the eukaryotic pathogen genomics database resource. Methods Mol Biol 1757:69–113. https://doi.org/10.1007/978-1-4939-7737-6_5