Article

Microbiome analyses of blood and tissues suggest cancer diagnostic approach

https://doi.org/10.1038/s41586-020-2095-1

Received: 7 June 2019

Accepted: 6 February 2020

Published online: 11 March 2020

Check for updates

Gregory D. Poore[1,12], Evguenia Kopylova[2,9,12], Qiyun Zhu[2], Carolina Carpenter[3], Serena Fraraccio[3], Stephen Wandro[3], Tomasz Kosciolek[2,10], Stefan Janssen[2,11], Jessica Metcalf[4], Se Jin Song[3], Jad Kanbar[5], Sandrine Miller-Montgomery[1,3], Robert Heaton[6], Rana Mckay[7], Sandip Pravin Patel[3,7], Austin D. Swafford[3] & Rob Knight[1,2,3,8]✉

Systematic characterization of the cancer microbiome provides the opportunity to develop techniques that exploit non-human, microorganism-derived molecules in the diagnosis of a major human disease. Following recent demonstrations that some types of cancer show substantial microbial contributions[1–10], we re-examined whole-genome and whole-transcriptome sequencing studies in The Cancer Genome Atlas[11] (TCGA) of 33 types of cancer from treatment-naive patients (a total of 18,116 samples) for microbial reads, and found unique microbial signatures in tissue and blood within and between most major types of cancer. These TCGA blood signatures remained predictive when applied to patients with stage Ia–IIc cancer and cancers lacking any genomic alterations currently measured on two commercial-grade cell-free tumour DNA platforms, despite the use of very stringent decontamination analyses that discarded up to 92.3% of total sequence data. In addition, we could discriminate among samples from healthy, cancer-free individuals (n = 69) and those from patients with multiple types of cancer (prostate, lung, and melanoma; 100 samples in total) solely using plasma-derived, cell-free microbial nucleic acids. This potential microbiome-based oncology diagnostic tool warrants further exploration.

Cancer is classically considered a disease of the human genome[12,13]. However, recent studies have shown that the microbiome makes substantial contributions to some types of cancer; in particular, contributions of the faecal microbiome to gastrointestinal cancers[1–10]. However, the extent and diagnostic implications of microbial contributions to different types of cancer remain unknown. The possibility of sample contamination during collection, processing, and sequencing limits these investigations, as procedural controls have rarely been implemented in cancer genomics projects. The use of recently developed tools[14–18] to minimize the contributions of contaminants to microbial signatures could enable the rational development of microbiome-based diagnostics.

To characterize the cancer-associated microbiome, we re-examined microbial reads from 18,116 samples across 10,481 patients and 33 types of cancer from the TCGA compendium of whole-genome sequencing (WGS; n = 4,831) and whole-transcriptome sequencing (RNA-seq; n = 13,285) studies. Microbial reads were previously identified in ad hoc analyses (including Epstein–Barr virus (EBV) in stomach adenocarcinoma[19] and human papillomavirus (HPV) in cervical cancer[20]) and have been systematically studied in small subsets of samples (for example, the viromes of 4,433 TCGA samples from 19 types of cancer[21] and the bacteriomes of 1,880 TCGA samples across 9 types of cancer[17]). Most TCGA sequencing data remain unexplored for microorganisms. Here we present, to our knowledge, the most comprehensive cancer microbiome data set yet created using two orthogonal microbial-detection pipelines, systematically measuring and mitigating technical variation and contamination. We use machine learning (ML) to identify microbial signatures that discriminate among types of cancer, and compare their performance.

Because TCGA processing did not control for microbial contamination and excluded healthy individuals, we performed an additional analysis on blood, the TCGA sample type most likely to contain adventitious microbial contamination, using gold-standard microbiology protocols[18,22]. We focused on commensurably benchmarking signatures from plasma-derived microbial DNA against clinically available cell-free tumour DNA (ctDNA) assays. Deep metagenomic sequencing on plasma samples from individuals with prostate, lung, or skin cancers (n = 100 total), and healthy, cancer- and HIV-free control participants (n = 69) suggested that cell-free microbial profiles could be used to achieve healthy-versus-cancer and cancer-versus-cancer discriminations. These findings suggest a new class of microbiome-based cancer diagnostic tools that may complement existing ctDNA assays for detecting and monitoring cancer.

[1]Department of Bioengineering, University of California San Diego, La Jolla, CA, USA. [2]Department of Pediatrics, University of California San Diego, La Jolla, CA, USA. [3]Center for Microbiome Innovation, University of California San Diego, La Jolla, CA, USA. [4]Department of Animal Sciences, Colorado State University, Fort Collins, CO, USA. [5]Department of Medicine, University of California San Diego, La Jolla, CA, USA. [6]Department of Psychiatry, University of California San Diego, La Jolla, CA, USA. [7]Moores Cancer Center, University of California San Diego Health, La Jolla, CA, USA. [8]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA. [9]Present address: Clarity Genomics, Beerse, Belgium. [10]Present address: Malopolska Centre of Biotechnology, Jagiellonian University in Krakow, Krakow, Poland. [11]Present address: Algorithmic Bioinformatics, Department of Biology and Chemistry, Justus Liebig University Gießen, Gießen, Germany. [12]These authors contributed equally: Gregory D. Poore, Evguenia Kopylova. ✉e-mail: robknight@ucsd.edu

TCGA cancer microbiome and its normalization

Of 6.4×10^{12} sequencing reads in TCGA, 7.2% were classified as non-human, of which 35.2% (2.5% of total reads) were assigned to bacteria, archaea, or viruses with 12.6% (0.9% of total reads) resolved at the genus level by Kraken[23], which matches short genomic substrings (*k*-mers) to taxa in a reference database (Fig. 1a, Supplementary Tables 1, 2, Extended Data Fig. 1a (TCGA study abbreviations)). After we had filtered samples for quality-controlled metadata (Fig. 1b) and normalized by sample number within a cancer type and sample type (Extended Data Fig. 1f, g), WGS provided significantly more microbial reads than RNA-seq experiments for primary tumour ($P = 2.08 \times 10^{-9}$), solid-tissue normal ($P = 1.26 \times 10^{-7}$), metastatic ($P = 0.0396$), and recurrent tumour samples ($P = 0.0336$; all two-sided Mann–Whitney *U*-tests). Fast *k*-mer-matching approaches are prone to false-positive results, so we performed slower, but potentially more specific, genome alignments of Kraken-positive, genus-level microbial reads (on which our findings are based) for four TCGA types of cancer (cervical squamous cell carcinoma (CESC), stomach adenocarcinoma (STAD), lung adenocarcinoma (LUAD), and ovarian serous cystadenocarcinoma (OV)) with known microbial relationships[5,19,20] and/or with paired proteomic data[24]. We found a low estimated false-positive rate of 1.09% (Supplementary Table 3), suggesting that the Kraken data were valid for downstream analyses.

TCGA expression and human genomic data are known to show substantial batch effects[25,26], which were replicated in metagenomic data (Fig. 1c, Extended Data Fig. 1b, d). Therefore, we implemented a pipeline that converted discrete taxonomical counts into log-counts per million (log-cpm) per sample using Voom[27], and performed supervised normalization (SNM; see Methods)[28]. Principal variance components analysis[29,30] showed that normalization reduced batch effects while increasing biological signal, including 'disease type' (that is, type of cancer), above the individual technical variables (Fig. 1d, e, Extended Data Fig. 1c, e).

Predicting among and within types of cancer

Using normalized data, we trained stochastic gradient-boosting ML models to discriminate between and within types and stages of cancer. The performance of these models was strong for discriminating (i) one cancer type versus all others ($n = 32$ types of cancer) and (ii) tumour versus normal ($n = 15$ types of cancer) (Fig. 1f, g, Extended Data Fig. 2a–f; all performance metrics can be found at http://cancermicrobiome.ucsd.edu/CancerMicrobiome_DataBrowser). Differences in sensitivities and specificities between types of cancer may be partially due to differences in class sizes, as there was a significant linear relationship in one-cancer-type-versus-all-others comparisons between the minority class size and AUROC (area under the receiver operating characteristic curve; $P = 0.0231$) and AUPR (area under the precision–recall curve; $P = 0.0089$) values (two-sided hypothesis tests of slope; Extended Data Fig. 2g, h). Cancer microbial heterogeneity may also contribute to this differential performance, although spatial examination of these historical tissue samples is beyond the scope of this study. Tissue-based microbial models performed well for discriminating between stage I and stage IV tumours ($n = 8$ types of cancer) for colon adenocarcinoma (COAD), STAD, and kidney renal clear cell carcinoma (KIRC), but not the other five cancers tested (Fig. 1h), nor for discriminating intermediate stages (data not shown). These results suggest that microbial community structure dynamics may not correlate with cancer stages as defined by host tissue for all types of cancer.

To evaluate the generalizability of our approach across data sets, we randomly sorted raw TCGA microbial counts into two batches, repeated all procedures on each independently, tested each independently trained model on the other half of the data, and found highly similar performance (Extended Data Fig. 3a). Discriminatory microbial signatures held when examining singular methodologies (WGS or RNA-seq) or sequencing centres that performed either WGS (Harvard Medical School; HMS) or RNA-seq (University of North Carolina; UNC) (Extended Data Fig. 3b–i), or using only genomic alignment-filtered Kraken data (Supplementary Table 3, Extended Data Fig. 4a–h).

For further validation, we applied SHOGUN[31], an alignment-based microbial taxonomic pipeline using a reduced, phylogenetically based, bacteria-only database on 13,517 TCGA samples (WGS, $n = 3,434$; RNA-seq, $n = 10,083$), covering every analysed type of cancer ($n = 32$), type of sample ($n = 7$), sequencing platform ($n = 6$), and sequencing centre ($n = 8$) in the Kraken-based analysis. The SHOGUN-derived data replicated the batch effects that had been identified in Kraken-derived data despite the use of a smaller, non-identical underlying database (Extended Data Fig. 4j–l). We input these data and a corresponding subset of Kraken-derived data (see Methods) independently into our normalization and ML pipelines and found no major differences in discriminatory performance between the data sets (Extended Data Fig. 4m–t). Together, the results imply that microbial communities are unique to each cancer type and that our approach of normalization and model training to distinguish cancers based on microbial profiles alone can be applied more broadly.

Biological relevance of microorganism profiles

Given the strong discrimination of microbial signatures, we sought evidence for their biological relevance using ecologically expected and/or clinically tested outcomes. To assess whether cancer-associated microorganisms are ecologically expected (that is, part of the 'native' organ-specific commensal community), we trained a Bayesian microbial-source tracking algorithm[32] on data from 217 samples across 8 body sites in the Human Microbiome Project 2 (HMP2) project[33] that had been processed with our microbial-detection and normalization pipelines to estimate the body-site contribution from 70 solid-tissue normal samples in the COAD cohort and 122 skin cutaneous melanoma (SKCM) primary tumours (see Methods). Stool was the primary known body-site contributor only to COAD profiles (average mean ± s.e.m. fractional contribution, 20.17 ± 2.55%; Fig. 2a), but not SKCM profiles (one-tailed Mann–Whitney *U*-test, $P = 0.0014$; Extended Data Fig. 5b), suggesting that part of the community had a local source.

Fusobacterium spp. are important in the development and progression of gastrointestinal tumours[1,19,34,35] and *Fusobacterium* was overabundant in primary tumours compared to solid-tissue normal samples (all $P \leq 8.5 \times 10^{-3}$) and especially to blood-derived normal samples (all $P \leq 3.3 \times 10^{-11}$; Fig. 2b). Pan-cancer analyses also showed an overabundance of *Fusobacterium* when comparing all broadly defined gastrointestinal (GI) cancers in TCGA ($n = 8$) against non-GI cancers ($n = 24$) in both primary tumour tissue ($P < 2.2 \times 10^{-16}$) and adjacent solid-tissue normal samples ($P = 0.031$; Fig. 2c, Extended Data Fig. 5a). Similar to previous investigations of STAD in TCGA[19], we found no differences in *Helicobacter pylori* between primary tumours and adjacent solid-tissue normal samples ($P = 0.72$, data not shown; all tests were two-sided Mann–Whitney *U*-tests).

We then confirmed clinically annotated TCGA viral infections and compared our microorganism-detection pipeline to studies that examined the TCGA virome using two different bioinformatic pipelines: (i) de novo metagenome assembly methods and (ii) read-based methods (PathSeq[36] algorithm)[19,21]. There was differential abundance of the Alphapapillomavirus genus between primary tumours in individuals who were clinically tested as 'positive' or 'negative' for HPV infection in CESC and head and neck squamous cell carcinoma (HNSC) samples (all $P \leq 3 \times 10^{-9}$, two-sided Mann–Whitney *U*-test; Fig. 2d, e). Blood-derived normal samples from patients with CESC were used as negative controls and were not statistically different ($P = 0.99$, two-sided Mann–Whitney *U*-test), and selective overabundance for Alphapapillomavirus held when comparing across all other types of cancer and sample types

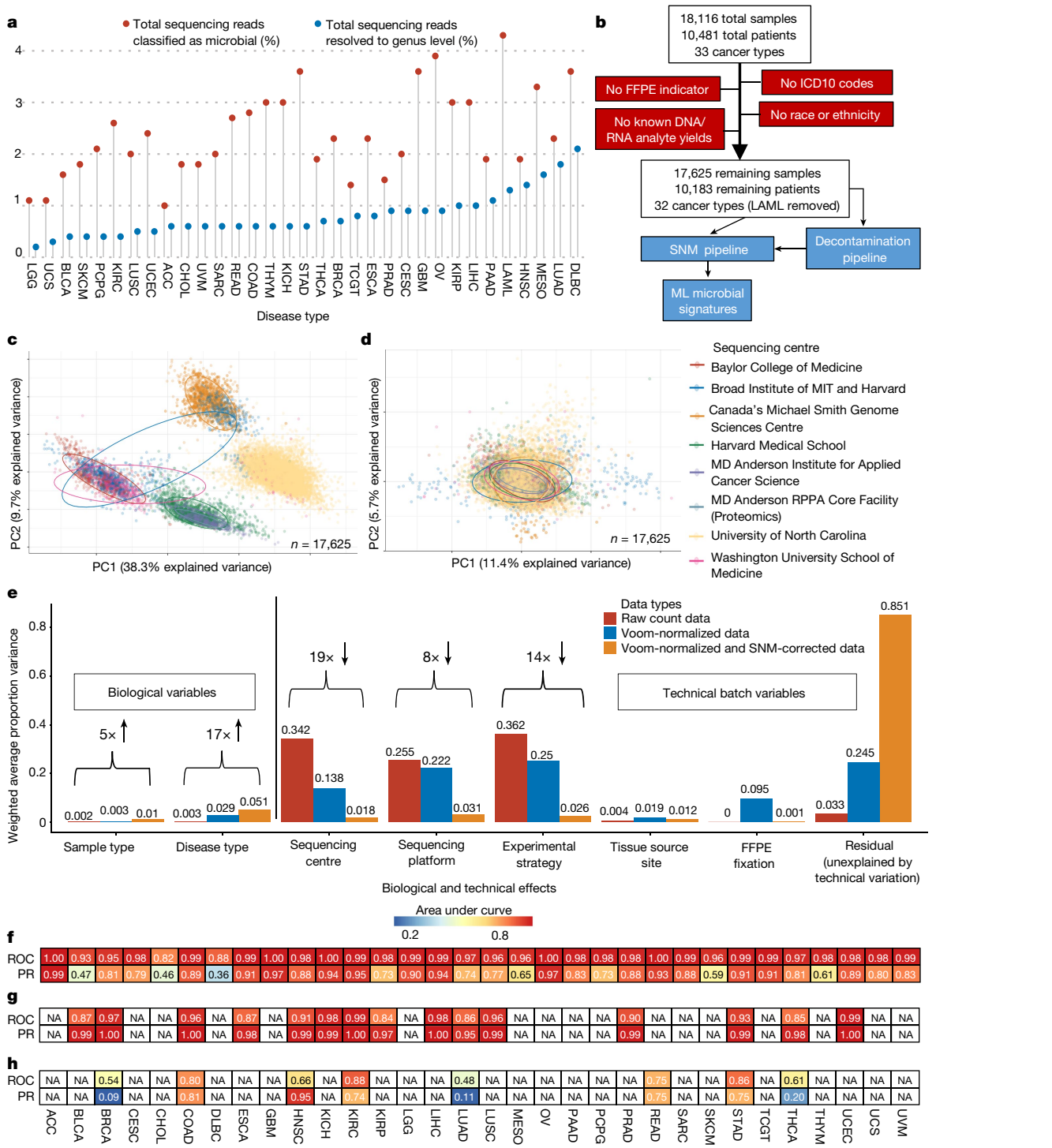


Fig. 1 | Approach and overall findings of the cancer microbiome analysis of TCGA. a, Lollipop plot showing the percentage of total sequencing reads identified by the microbial-detection pipeline, and those resolved at the genus level in TCGA data set by Kraken. LAML, acute myeloid leukaemia; PAAD, pancreatic adenocarcinoma; GBM, glioblastoma multiforme; PRAD, prostate adenocarcinoma; ESCA, oesophageal carcinoma; TCGT, testicular germ cell tumours; BRCA, breast invasive carcinoma; THCA, thyroid carcinoma; KICH, kidney chromophobe; THYM, thymoma; READ, rectum adenocarcinoma; SARC, sarcoma; UVM, uveal melanoma; CHOL, cholangiocarcinoma; ACC, adrenocortical carcinoma; UCEC, uterine corpus endometrial carcinoma; LUSC, lung squamous cell carcinoma; PCPG, pheochromocytoma and paraganglioma; BLCA, bladder urothelial carcinoma; UCS, uterine carcinosarcoma; LGG, brain lower grade glioma (Extended Data Fig. 1a). The number of samples included for each cancer type and sample type can be found in Supplementary Table 4. **b**, CONSORT-style diagram showing quality control processing and the number of remaining samples. FFPE, fixed-formalin paraffin-embedded. **c**, Principal components analysis (PCA) of Voom-normalized data, with cancer microbiome samples coloured by sequencing centre. **d**, PCA of Voom-SNM data. **e**, Principal variance components analysis of raw taxonomical count data, Voom-normalized data, and Voom-SNM data. **f**–**h**, Heatmaps of classifier performance metrics (AUROC (ROC) and AUPR (PR)) from red (high) to blue (low) for distinguishing between TCGA primary tumours (**f**), between tumour and normal samples (**g**), and between stage I and stage IV cancers (**h**). NA, fewer than 20 samples available in any ML class for model training.

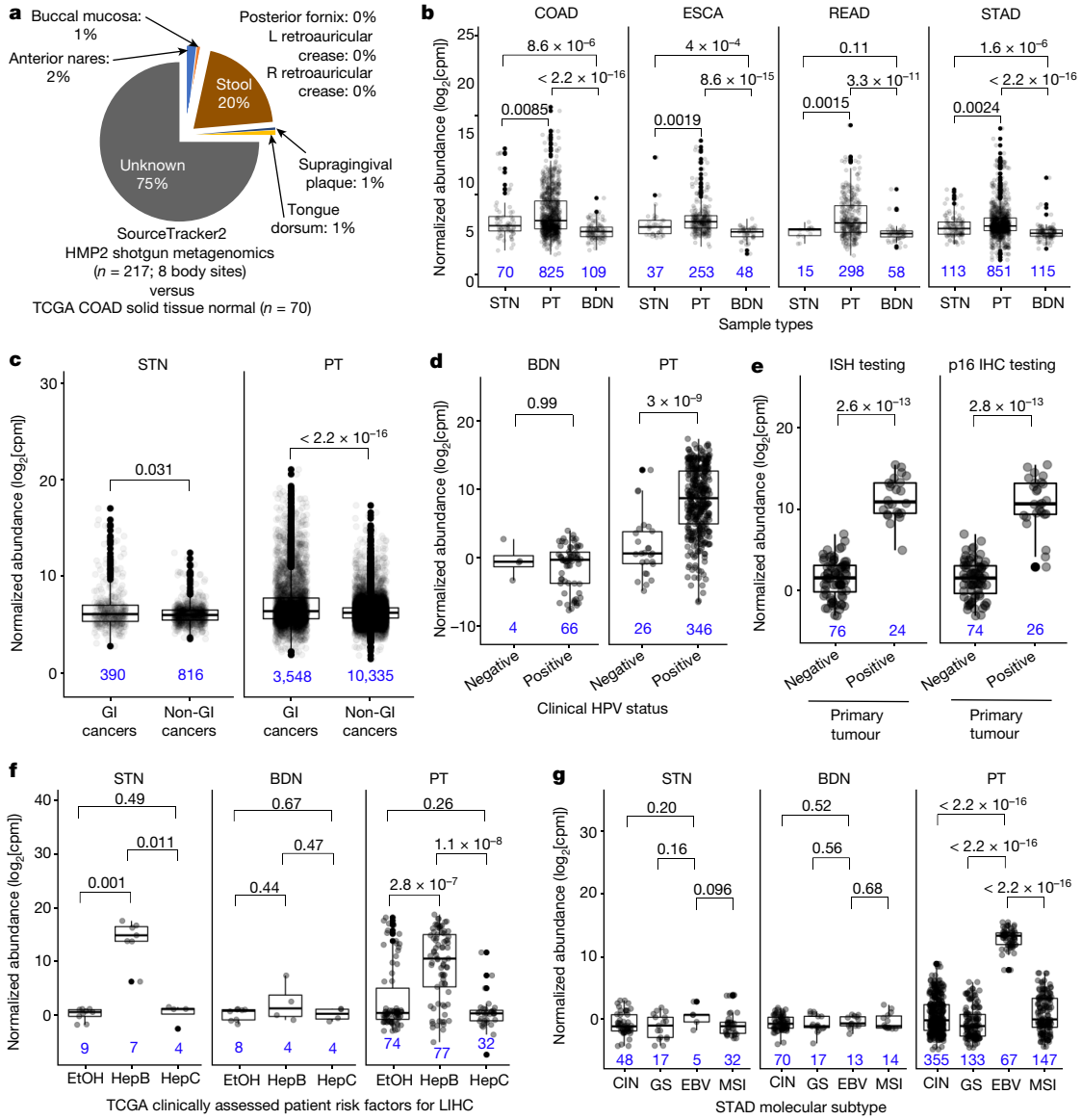


Fig. 2 | Ecological validation of viral and bacterial reads within the TCGA cancer microbiome data set. a, Average body site attribution for solid-tissue normal samples from patients with COAD ($n = 70$) using SourceTracker2[32] trained on the HMP2 data set. **b**, Differential abundances of the *Fusobacterium* genus for common gastrointestinal (GI) cancers associated with *Fusobacterium* spp.[1,19,34,35]. BDN, blood derived normal; STN, solid tissue normal; PT, primary tumour. **c**, Differential abundances of *Fusobacterium* among grouped GI cancers ($n = 8$: COAD, READ, CHOL, LIHC, PAAD, HNSC, ESCA, STAD) and non-GI cancers ($n = 24$) (see Methods). **d**, **e**, Normalized HPV abundances for HPV-infected patients with CESC (**d**) or HNSC (**e**), as clinically denoted in TCGA. ISH, in situ hybridization; IHC, immunohistochemistry. **f**, Normalized Orthohepadnavirus abundance in patients with LIHC with clinically adjudicated risk factors: HepB, prior hepatitis B infection; EtOH, heavy alcohol consumption; HepC, prior hepatitis C infection. **g**, Normalized EBV abundance in STAD integrative molecular subtypes: CIN, chromosomal instability; GS, genome stable; MSI, microsatellite unstable; EBV, EBV-infected samples. In all panels, blood-derived normal and/or solid-tissue normal data are shown as comparative negative controls; two-sided Mann–Whitney *U*-tests were used with multiple testing correction for more than two comparisons; box plots show median (line), 25th and 75th percentiles (box), and 1.5 × the interquartile range (IQR, whiskers). Blue numbers show sample sizes.

(Extended Data Fig. 5c). Patients with liver hepatocellular carcinoma (LIHC) and a prior history of hepatitis B had selective overabundance of the HBV genus (Orthohepadnavirus) in both primary tumours and adjacent solid-tissue normal samples compared to patients with LIHC and a prior history of alcohol consumption and hepatitis C (Hepacivirus genus) (Fig. 2f; primary tumour $P \leq 2.8 \times 10^{-7}$; solid-tissue normal $P \leq 0.011$); blood-derived normal samples were used as negative controls and were not statistically different ($P \geq 0.44$; all tests were two-sided Mann–Whitney *U*-tests). Also consistent with the previous reports[19], the genus for EBV (Lymphocryptovirus) was selectively overabundant in EBV-infected primary tumours compared to patients assigned to other STAD molecular subtypes (Fig. 2g; $P \leq 2.2 \times 10^{-16}$). Solid-tissue normal and blood-derived normal samples were used as negative controls and were not statistically different (blood, $P \geq 0.52$; tissue, $P \geq 0.096$; all tests were two-sided Mann–Whitney *U*-tests).

These data are consistent with information about feature importance provided by our models in one-cancer-type-versus-all-others distinctions. Namely, cancers with known microbial 'drivers' or 'commensals' provided initial evidence that the models were ecologically relevant; for example, Alphapapillomavirus genus was the most important feature for identifying CESC tumours; for COAD tumours, the *Faecalibacterium* genus; for LIHC tumours, the Orthohepadnavirus genus was the second most important feature (after the hepatotoxic *Microcystis* genus[37]). For additional hypothesis generation, we created an interactive website to

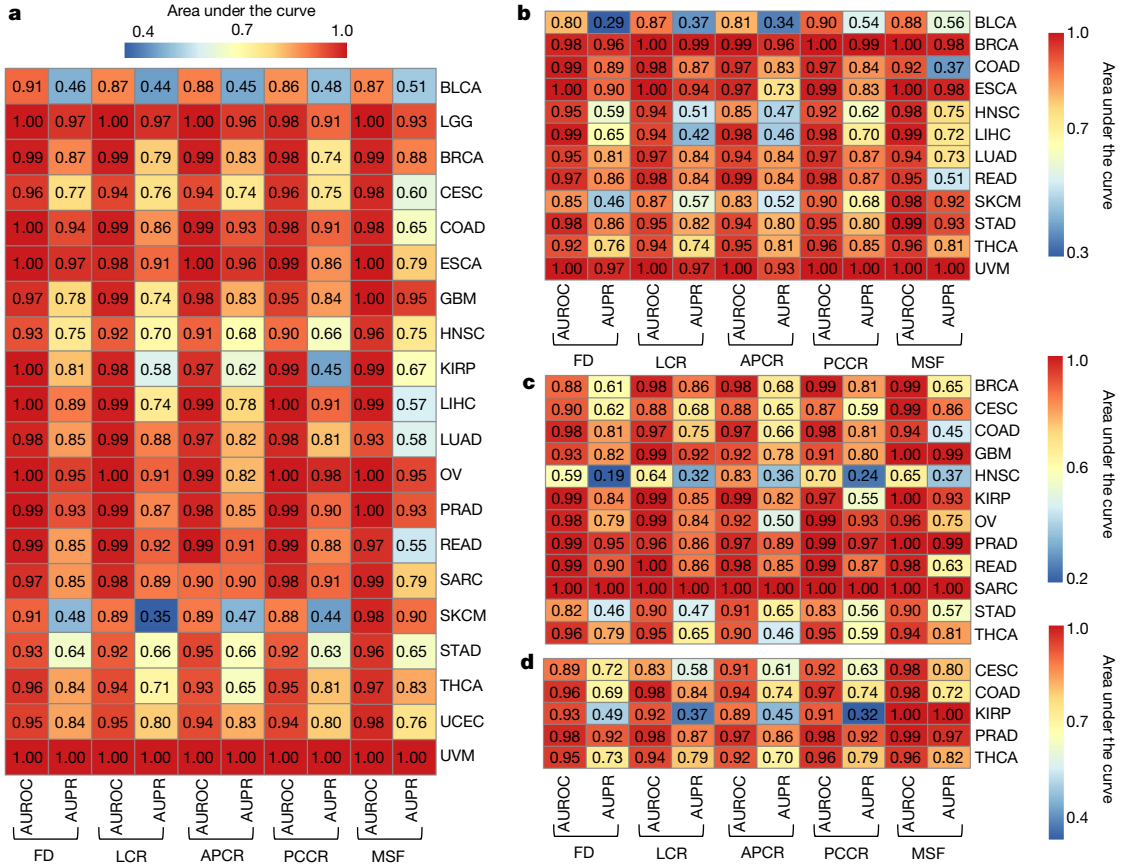


Fig. 3 | Classifier performance for cancer discrimination using mbDNA in blood and as a complementary diagnostic approach for cancer 'liquid' biopsies. a, Model performance heatmap analogous to Fig. 1f–h to predict one cancer type versus all others using blood mbDNA with TCGA study IDs on the right (Extended Data Fig. 1a); at least 20 samples were required in each ML minority class to be eligible. **b**, ML model performances predicting one cancer type versus all others using blood mbDNA for stage Ia–IIc cancers. **c**, **d**, ML model performances using blood mbDNA from patients without detectable primary tumour genomic alterations, per Guardant360 (**c**) and FoundationOne Liquid (**d**) ctDNA assays. FD, full data; LCR, likely contaminants removed by sequencing centre; APCR, all putative contaminants removed by sequencing centre; PCCR, plate–centre contaminants removed; MSF, most stringent filtering by sequencing centre. The number of samples included to evaluate the performance of each comparison can be found in the data browser confusion matrices at http://cancermicrobiome.ucsd.edu/CancerMicrobiome_DataBrowser.

enable the exploration of normalized microbial abundances in TCGA cancers and major sample types (Extended Data Fig. 5d, e; http://cancermicrobiome.ucsd.edu/CancerMicrobiome_DataBrowser). We provide raw and normalized microbial abundance data sets for public reuse, and anticipate the opportunity to integrate these with host multi-omic data to generate additional mechanistic hypotheses. Collectively, the findings provide ecological validation of our bioinformatic and normalization approaches for viral and bacterial data while extending the results to many more samples and microorganisms.

Measuring and mitigating contamination

We recognize the importance of measuring and mitigating the potential effects of contamination, in order to best characterize putative cancer-associated microorganisms[14–18]. Previous work identified just six contaminants in TCGA (*Staphylococcus epidermidis*, *Propionibacterium acnes*, *Ralstonia* spp., *Mycobacterium*, *Pseudomonas*, and *Acinetobacter*) based on common low-read abundances across types of cancer[17], but recent studies have shown that external contaminants more consistently have frequencies that are inversely correlated with sample analyte concentration and can be detected using a robust statistical framework[16,38].

Based on the latter approach, we used DNA and RNA concentrations calculated during TCGA sample processing (n = 17,625) and taxon read fractions (n = 1,993) to identify putative contaminants, and also removed genera typically found in 'negative blank' reagents[14] (n = 94 genera; see Methods). Extended Data Fig. 6a outlines the approaches taken from surgical resection to bioinformatic processing; we also spiked five types of pseudo-contaminants into the raw data set to track through decontamination, supervised normalization, and ML. Given known technical variation (Fig. 1c–e), we processed samples in batches by sequencing centre (n = 8) and removed taxa found to be a contaminant at any centre. This identified 283 putative contaminants, including 19.1% (n = 18 genera) of the reagent 'blacklist'[14]. After combining these two lists (n = 377 genera), we manually reviewed the literature (Supplementary Tables 6, 7) to re-allow pathobiont genera or mixed-evidence genera (both a pathogen and common contaminant; for example, *Mycobacterium*). This resulted in two data sets, one with likely contaminants removed and another with all putative contaminants removed. We also created a third 'most stringent filtering' data set that discarded about 92% of the total reads using a stricter filtering schema (see Methods; Extended Data Fig. 6b). Finally, we grouped samples into individual sequencing plates at each centre and removed all putative contaminants identified in any one 'plate–centre' batch (n = 351; Supplementary Table 8; see Methods), in addition to the aforementioned reagent blacklist (497 genera in total). Decontamination did not appear to differentially affect the types of sample or cancer under study (Extended Data Fig. 7).

a

Variable	Control	PC	LC	SKCM
Sample size (*n*)	69	59	25	16
Age (years) (mean ± s.d.)	45.00 ± 12.80	69.70 ± 9.63	69.50 ± 9.88	58.50 ± 13.00
Sex (% female)	24.6%	0%	68%	12.5%
Known condition(s)/ subtype(s) (*n*)	HIV-free (69)	HSPC (32) CRPC (27)	LUAD (15) LUSC (5) Sarcomatoid (1) Large cell (1) NOS (3)	SKCM (16)

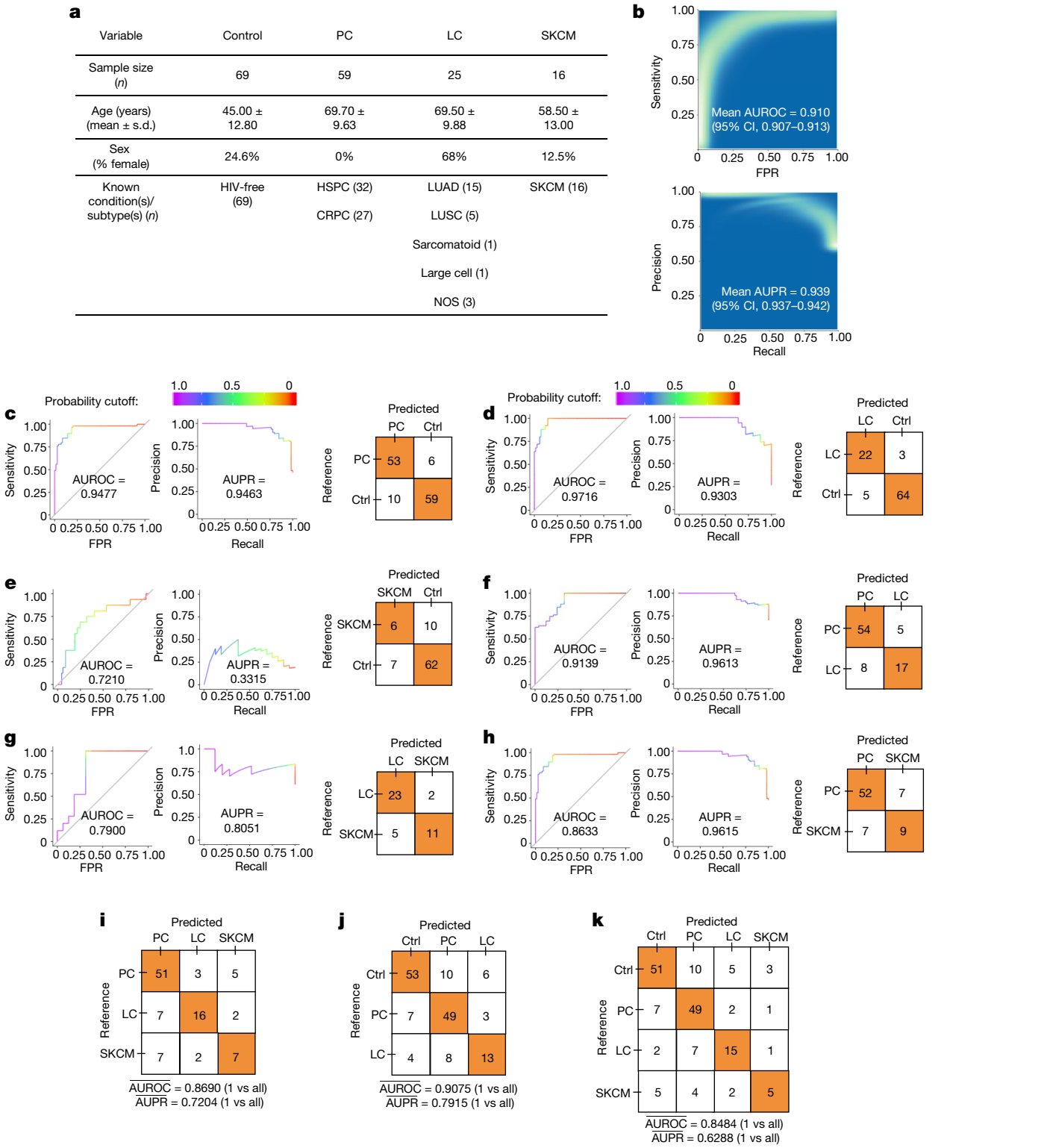


Fig. 4 | Performance of ML models to discriminate between types of cancer and healthy controls using plasma-derived, cell-free mbDNA. **a**, Demographics of samples analysed in the validation study. All patients had high-grade (stage III–IV) cancers of multiple subtypes and were aggregated into PC, LC, and SKCM groups. **b**, Bootstrapped performance estimates for distinguishing grouped cancer samples ($n = 100$) from non-cancer healthy controls ($n = 69$). Rasterized density plot of ROC (top) and PR (bottom) curve data from 500 iterations with different training–testing splits (70%–30%). **c**–**h**, LOO iterative ML performances between two classes: PC versus control (Ctrl; **c**), LC versus control (**d**), SKCM versus control (**e**), PC versus LC (**f**), LC versus SKCM (**g**), and PC versus SKCM (**h**). **i**–**k**, Multi-class ($n = 3$ or 4), LOO iterative ML performances to distinguish among types of cancer (**i**) and between mixed patients with cancer and healthy control individuals (**j**, **k**). Overall LOO ML performance was calculated as the mean of performances when comparing one versus all others (shown below as confusion matrices).

We stress that these in silico decontamination methods are not substitutes for implementing gold-standard microbiology practices on cancer samples, including sterile processing, sterile-certified reagents, negative blanks of reagents processed from start to finish, and multiple-sample pooling as 'positive' controls[18,22]. The in silico tools described here reflect the state of the art, but are not designed to detect abundant 'spikes' of contaminants or cross-contaminants. These latter contaminants should not drive uniform discriminatory signals between and within types of cancer collected over many centres and years, but may limit biological conclusions, particularly in small studies, if not controlled.

Another risk with stringent decontamination is that real signals that reflect commensal, tissue-specific microbial communities and concomitant cancer-predictive microbial profiles may be discarded. To evaluate this concern, we re-calculated the body-site attribution percentages for COAD solid-tissue normal samples ($n = 70$), and found that successively stringent decontamination improved recognition of concomitant tissues before they became unrecognizable (Extended Data Fig. 6c–f).

We recalculated all ML models shown in Fig. 1f–h and compared their performances before and after each decontamination approach (Extended Data Fig. 6g–l). Most models did not rely on spiked pseudo-contaminants (Extended Data Fig. 8a), although the lymphoid neoplasm diffuse large B cell lymphoma (DLBC) and mesothelioma (MESO) models (with very few available samples; Supplementary Table 4) appear to be exceptions and may be unreliable. As expected, comparisons where knowledge about the tissue type is informative (for example, COAD versus all other cancer types) generally performed less well with stringent decontamination, but within-tissue comparisons (for example, tumour versus normal) often performed equally well or better. These results suggest that stringent filtering may be desirable in certain comparisons, but a universal approach to decontamination may preclude biologically informative results.

Predictions using microbial DNA in blood

There is mounting evidence that blood-based microbial DNA (mbDNA) can be clinically informative in cancer[39–44], including those featuring blood-barrier or lymphatic disruptions (for example, COAD)[39], but it is unclear how broadly this applies. Using WGS data from TCGA blood samples, we applied our ML strategies to the full data sets and four decontaminated data sets and found that blood-borne mbDNA could discriminate between numerous types of cancer (Fig. 3a), regardless of the microbial taxonomic algorithm and database used for classification or when using only genomic-alignment-filtered Kraken data (Extended Data Fig. 4g, h, s, t). Retrospective analysis showed that few models included spiked pseudo-contaminants for predictions (Extended Data Fig. 8b); models that did (CESC, kidney renal papillary cell carcinoma (KIRP), LIHC) may be less trustworthy.

Spurred by these findings, we sought to benchmark our ML models against existing ctDNA assays, focusing on circumstances under which ctDNA assays fail: stage Ia–IIc cancers and tumours without detectable genomic alterations. After removing all blood-derived normal samples from patients harbouring stage III or IV cancers, we built new ML models and found that they were able to discriminate well between types of cancer using blood mbDNA (Fig. 3b). We further used gene lists from the Guardant360 and FoundationOne Liquid assays[45,46] to filter out TCGA patients with one or more targeted modification (about 70%; Extended Data Fig. 8c–e) and found that the same ML approach showed good discrimination for most remaining types of cancer (Fig. 3c, d).

These analyses are limited by the facts that ctDNA assays use plasma rather than whole blood[45,46], and that the distribution of mbDNA among blood compartments is unknown. It is impossible to tell whether mbDNA came from live or dead microorganisms, as RNA data were unavailable, or whether mbDNA is cell-free or in host leukocytes, as TCGA standard operating procedures (SOPs) allowed whole blood or buffy coat extraction (see Methods). It is also impossible to know the origin of blood mbDNA without examining primary specimens and, possibly, matched gut epithelia, as certain types of cancer may 'leak' mbDNA in unexpected ways (for example, gut bacterial translocation in leukaemia)[8,10]. There is likely to be a continuum of ideal decontamination, as the effect of decontamination on model performance varied across types of cancer, but our filtering was limited by (i) not having access to the primary specimens, (ii) genus-level taxonomic resolution, and (iii) not knowing which non-TCGA samples were concurrently processed.

Validating microbial signatures in blood

To demonstrate the real-world utility of these results while benchmarking against plasma-based ctDNA assays[45,46], we evaluated the use of plasma-derived, cell-free mbDNA signatures to discriminate among healthy individuals and multiple types of cancer in a validation study while implementing gold-standard microbiology controls for low-biomass studies[18,22]. Although plasma represents a distinct subset of whole blood that is not studied in TCGA, limiting direct comparability, it carries major advantages in archival stability (for example, freezability), biorepository availability, and biological interpretation (that is, non-living material). Our cohort included 69 cancer- and HIV-free individuals and 100 patients with one of three types of high-grade (stage III–IV) cancer: prostate cancer ($n = 59$; PC); lung cancer ($n = 25$; LC), and melanoma ($n = 16$; SKCM) (Fig. 4a). Without prior literature to estimate effect sizes, we used independent simulations on TCGA blood samples from matched types of cancer at The Broad Institute and HMS to estimate minimum sample sizes (Extended Data Fig. 9a; see Methods). Cell-free DNA was extracted from these plasma samples with extensive controls[18,22] (Extended Data Fig. 9b, c), and processed for whole metagenomic sequencing by a limited set of users, using a single library preparation method[47], in a single batch, in one deep-sequencing run. We performed human-read removal, classification of remaining reads by Kraken, stringent decontamination using both DNA concentrations and negative blanks[16], and Voom-SNM. Demographic comparisons and permutation analyses suggested necessary normalization for age and sex (Extended Data Fig. 9d, e, h–j; see Methods), and direct age regression performance showed mean absolute errors similar to the gut microbiome[48] (Extended Data Fig. 9g). 'Bootstrapping' the same ML protocol used in the TCGA analyses showed strong, generalizable discrimination between healthy control individuals and grouped patients with cancer (Fig. 4b; see Methods). Because of our small sample sizes, we performed leave-one-out (LOO) iterative ML on the normalized data and found high discriminatory performance in pairwise and multiclass comparisons between and among healthy samples and types of cancer except for the smallest SKCM cohort (Fig. 4c–k). Therefore, we iteratively subsampled PC and LC groups to match the SKCM cohort size and performed pairwise LOO discrimination of each type of cancer against subsampled healthy controls (Extended Data Fig. 9k; see Methods). The PC and LC cohorts were still separable at the same cohort size as SKCM (mean (95% confidence interval (CI)) AUROC = 0.891 (0.879–0.903); mean (95% CI) AUPR = 0.827 (0.815–0.839); 100 iterations), revealing universal deficits in SKCM performance. This deficit may have a biological basis, as SKCM was the second-worst performer in TCGA blood discriminations for four of five data sets tested (Fig. 3a), although this warrants further confirmation. To ensure that the microbial assignments by Kraken were valid, we repeated all bioinformatic, normalization, and ML steps using bacterial assignments from SHOGUN[31] and its separate database[49], which showed highly concordant performances (Extended Data Fig. 10). We anticipate refinement of the taxonomic assignments for cfDNA signatures as microbial databases improve[50]. The plasma microbial abundances detected can be explored at http://cancermicrobiome.ucsd.edu/CancerMicrobiome_DataBrowser (Extended Data Fig. 5d, e).

Discussion

Collectively, our data suggest that there are widespread associations between diverse types of cancer and specific microbiota. These microbial profiles appear to discriminate within and between most types of cancer, including when using blood-based mbDNA at low-grade tumour stages and in patients without any detectable genomic alterations on commercial ctDNA assays. These results often remain valid even after extensive internal validation checks and decontamination, which at times discards more than 90% of the total data. The high discriminatory performance among healthy control individuals and patients with multiple types of cancer using only cell-free mbDNA in plasma, while adopting more extensive internal and external contamination controls than TCGA, suggests that clinically relevant and retrospective testing using widely available samples would be feasible and generalizable. More work is needed to determine whether the observed nucleic acids come from live microorganisms, host cells, or lysed bacteria in the tumour microenvironment and blood. Notably, many technical and biological factors limit the analysis of retrospective cancer sequencing data for low-biomass microorganisms, and advances in this field will require collaborations between cancer biologists and microbiologists. Nonetheless, our results suggest that a new class of microbiome-based cancer diagnostic tools may provide substantial future value to patients.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2095-1.

1. Bullman, S. et al. Analysis of *Fusobacterium* persistence and antibiotic response in colorectal cancer. *Science* **358**, 1443–1448 (2017).
2. Dejea, C. M. et al. Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria. *Science* **359**, 592–597 (2018).
3. Geller, L. T. et al. Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. *Science* **357**, 1156–1160 (2017).
4. Gopalakrishnan, V. et al. Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. *Science* **359**, 97–103 (2018).
5. Jin, C. et al. Commensal microbiota promote lung cancer development via γδ T cells. *Cell* **176**, 998–1013.e16 (2019).
6. Ma, C. et al. Gut microbiome-mediated bile acid metabolism regulates liver cancer via NKT cells. *Science* **360**, eaan5931 (2018).
7. Matson, V. et al. The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients. *Science* **359**, 104–108 (2018).
8. Meisel, M. et al. Microbial signals drive pre-leukaemic myeloproliferation in a Tet2-deficient host. *Nature* **557**, 580–584 (2018).
9. Routy, B. et al. Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science* **359**, 91–97 (2018).
10. Ye, H. et al. Subversion of systemic glucose metabolism as a mechanism to support the growth of leukemia cells. *Cancer Cell* **34**, 659–673.e6 (2018).
11. The Cancer Genome Atlas Research Network et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
12. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
13. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
14. Salter, S. J. et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 87 (2014).
15. Glassing, A., Dowd, S. E., Galandiuk, S., Davis, B. & Chiodini, R. J. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog.* **8**, 24 (2016).
16. Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**, 226 (2018).
17. Robinson, K. M., Crabtree, J., Mattick, J. S. A., Anderson, K. E. & Dunning Hotopp, J. C. Distinguishing potential bacteria-tumor associations from contamination in a secondary data analysis of public cancer genome sequence data. *Microbiome* **5**, 9 (2017).
18. Eisenhofer, R. et al. Contamination in low microbial biomass microbiome studies: issues and recommendations. *Trends Microbiol.* **27**, 105–117 (2019).
19. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209 (2014).
20. The Cancer Genome Atlas Research Network. Integrated genomic and molecular characterization of cervical cancer. *Nature* **543**, 378–384 (2017).
21. Tang, K.-W., Alaei-Mahabadi, B., Samuelsson, T., Lindh, M. & Larsson, E. The landscape of viral expression and host gene fusion and adaptation in human cancer. *Nat. Commun.* **4**, 2513 (2013).
22. Minich, J. J. et al. KatharoSeq enables high-throughput microbiome analysis from low-biomass samples. *mSystems* **3**, e00218-17 (2018).
23. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
24. Zhang, H. et al. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell* **166**, 755–765 (2016).
25. Choi, J.-H., Hong, S.-E. & Woo, H. G. Pan-cancer analysis of systematic batch effects on somatic sequence variations. *BMC Bioinformatics* **18**, 211 (2017).
26. Lauss, M. et al. Monitoring of technical variation in quantitative high-throughput datasets. *Cancer Inform.* **12**, 193–201 (2013).
27. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
28. Mecham, B. H., Nelson, P. S. & Storey, J. D. Supervised normalization of microarrays. *Bioinformatics* **26**, 1308–1315 (2010).
29. Boedigheimer, M. J. et al. Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories. *BMC Genomics* **9**, 285 (2008).
30. Scherer, A. *Batch Effects and Noise in Microarray Experiments: Sources and Solutions* (Wiley, 2009).
31. Hillmann, B. et al. Evaluating the information content of shallow shotgun metagenomics. *mSystems* **3**, e00069-18 (2018).
32. Knights, D. et al. Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* **8**, 761–763 (2011).
33. Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* **16**, 276–289 (2014).
34. Yamamura, K. et al. Human microbiome *Fusobacterium nucleatum* in esophageal cancer tissue is associated with prognosis. *Clin. Cancer Res.* **22**, 5574–5581 (2016).
35. Hsieh, Y.-Y. et al. Increased abundance of *Clostridium* and *Fusobacterium* in gastric microbiota of patients with gastric cancer in Taiwan. *Sci. Rep.* **8**, 158 (2018).
36. Kostic, A. D. et al. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat. Biotechnol.* **29**, 393–396 (2011).
37. Svircev, Z. et al. Molecular aspects of microcystin-induced hepatotoxicity and hepatocarcinogenesis. *J. Environ. Sci. Health C Environ. Carcinog. Ecotoxicol. Rev.* **28**, 39–59 (2010).
38. Jervis-Bardy, J. et al. Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data. *Microbiome* **3**, 19 (2015).
39. Kwong, T. N. Y. et al. Association between bacteremia from specific microbes and subsequent diagnosis of colorectal cancer. *Gastroenterology* **155**, 383–390.e8 (2018).
40. Blauwkamp, T. A. et al. Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. *Nat. Microbiol.* **4**, 663–674 (2019).
41. Hong, D. K. et al. Liquid biopsy for infectious diseases: sequencing of cell-free plasma to detect pathogen DNA in patients with invasive fungal disease. *Diagn. Microbiol. Infect. Dis.* **92**, 210–213 (2018).
42. Burnham, P. et al. Urinary cell-free DNA is a versatile analyte for monitoring infections of the urinary tract. *Nat. Commun.* **9**, 2412 (2018).
43. De Vlaminck, I. et al. Temporal response of the human virome to immunosuppression and antiviral therapy. *Cell* **155**, 1178–1187 (2013).
44. Huang, Y.-F. et al. Analysis of microbial sequences in plasma cell-free DNA for early-onset breast cancer patients and healthy females. *BMC Med. Genomics* **11** (Suppl. 1), 16 (2018).
45. Bettegowda, C. et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci. Transl. Med.* **6**, 224ra24 (2014).
46. Clark, T. A. et al. Analytical validation of a hybrid capture-based next-generation sequencing clinical assay for genomic profiling of cell-free circulating tumor DNA. *J. Mol. Diagn.* **20**, 686–702 (2018).
47. Sanders, J. G. et al. Optimizing sequencing protocols for leaderboard metagenomics by combining long and short reads. *Genome Biol.* **20**, 226 (2019).
48. Huang S. et al. Human skin, oral, and gut microbiomes predict chronological age. *mSystems* **5**, e00630-19 (2020).
49. Zhu, Q. et al. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477 (2019).
50. Chiu, K.-P. & Yu, A. L. Application of cell-free DNA sequencing in characterization of bloodborne microbes and the study of microbe-disease interactions. *PeerJ* **7**, e7426 (2019).

Methods

TCGA data accession

All TCGA sequence data (Supplementary Tables 4, 5) were accessed via the Cancer Genomics Cloud (CGC) as sponsored by SevenBridges (https://cgc.sbgenomics.com)[51]. Details of how these samples were acquired and processed are comprehensively described elsewhere[52]. SOPs for TCGA were accessed via the NCI Biospecimen Research Database (https://brd.nci.nih.gov/brd/sop-compendium/show/701). Matched patient metadata, including molecular subtypes, were accessed via the CGC through both SevenBridges and the Institute for Systems Biology (ISB; https://isb-cgc.appspot.com/)[53], via the TCGA-Mutations R package[54], or were taken directly from the supplementary data of the respective TCGA publication[19,55]. Genomic alteration statuses for all TCGA patients were queried and downloaded via cBioPortal[56,57]. Gene panels for commercial ctDNA assays were accessed from company white papers for the Guardant360 assay (https://www.therapyselect.de/sites/default/files/downloads/guardant360/guardant360_specification-sheet_en.pdf) and the FoundationOne Liquid assay (https://assets.ctfassets.net/vhribv12lmne/3SPYAcbGdqAeMsOqMyKUog/d0eb51659e08d733bf39971e85ed940d/F1L_TechnicalInformation_MKT-0061-04.pdf). For TCGA metadata accession and transformation from hierarchical formats to flat tables, custom Python scripts (available on Github; https://github.com/biocore/tcga) were written to query SevenBridges's metadata ontology and organize the data where possible; for information not stored in that ontology, we used the ISB CGC R programming language API[53] to access its recent metadata release (tcga_201607_beta.Clinical_data).

Kraken TCGA microbial-detection pipeline

The SevenBridges CGC interface enabled rapid development of the bioinformatic pipeline for this project while ensuring its future reproducibility[51]. Bioinformatic tools were either loaded directly from the CGC platform (for example, samtools, Burrows–Wheeler aligner (BWA)) or uploaded and run as separate Docker containers (for example, QIIME, Kraken) in order to create customized 'app' workflows. These app workflows take sample BAM files as inputs and label which DNA or RNA reads within each sample are microbial. These app workflows can be publicly shared for reproducibility purposes, as needed. The computational analyses themselves were hosted on Amazon Web Services (AWS; https://aws.amazon.com/) through the CGC interface and most often used the AWS x1.16 EC2 compute instance, comprised of the following specifications: 64 vCPU, 174.5 ECU, 976 GB of memory, and 1,920 GB of instance storage. The computational wall-time was approximately 6 months using these specifications.

Sequencing reads that did not align to known human reference genomes (based on mapping information in the raw BAM files) were mapped against all known bacterial, archaeal, and viral microbial genomes using the ultrafast Kraken algorithm[23]. A total of 71,782 microbial genomes were downloaded using RepoPhlan (https://bitbucket.org/nsegata/repophlan) on 14 June 2016, of which 5,503 were viral and 66,279 were bacterial or archaeal. On the basis of prior literature, bacterial and archaeal genomes were filtered for quality scores of 0.8 or better[58], which left 54,471 of them for subsequent analysis, or a total of 59,974 microbial genomes.

As previously described in detail[23], the Kraken algorithm breaks each sequencing read into *k*-mers (we used default 31-mers) and exactly matches each *k*-mer against a database of microbial *k*-mers, which was built from the 59,974 microbial genomes described above before running the algorithm. The set of exact *k*-mer matches for a given read, in turn, provides a putative taxonomy assignment of the lowest common ancestor for that read, most accurately to the genus level, to which we summarized our data. The matching and classification operations are orders of magnitude faster than performing direct genome alignments. As a safeguard against false positives and to properly benchmark our pipeline, we took four types of cancer (COAD, CESC, OV and LUAD) and aligned the reads Kraken classified as microbial to the 59,974 microbial genomes using BWA[59], which is computationally more expensive but yields a result with higher specificity and taxonomic resolution (that is, to species and strain level). The four types of cancer that were directly aligned included CESC as a putative positive viral control (for HPV), STAD as a putative positive bacterial control (for *H. pylori*), and two others (LUAD, OV) based on microbial signatures in the literature and/or available mass-spectrometry proteomic information (data not shown)[5,24,60–62]. We found that 98.91% of reads that were classified to genus level or lower by Kraken (on which our main findings are based) also aligned with BWA to the microbial database (bacteria, archaea, viruses; see Supplementary Table 3), or a false-positive rate of 1.09%, suggesting that the genus-level, Kraken-labelled, pan-cancer microbial reads were sufficiently usable for further analyses.

SHOGUN TCGA bioinformatic processing

To evaluate the robustness of cancer type discriminations using different taxonomic identification algorithms, we used a previously published shallow shotgun taxonomic assignment approach (co-developed by Q.Z. & R.K.)[31] and a separate, phylogeny-centric database called Web of Life (WoL; co-developed by Q.Z. & R.K.; $n = 10{,}575$ bacterial and archaeal genomes; https://bitbucket.org/nsegata/repophlan)[49] on TCGA samples. SHOGUN utilizes computationally intensive direct genomic alignments for taxonomy assignments rather than an ultrafast *k*-mer-based approach like that used by Kraken. To reduce processing time for TCGA samples, reads classified as microbial in origin by Kraken were used as input for the SHOGUN[31] align function, which used Bowtie2[63] to map reads against the WoL database[49] to generate taxonomy profiles. In total, 13,517 samples (WGS: $n = 3{,}434$; RNA-seq: $n = 10{,}083$) were processed, covering all TCGA types of cancer ($n = 32$), sample types ($n = 7$), sequencing centres ($n = 8$), and sequencing platforms ($n = 6$) under study in the Kraken analysis, including 21 TCGA types of cancer ($n = 9{,}444$ samples) that had all samples in the Kraken analysis re-analysed by SHOGUN. Profiles were then collapsed to the genus level using QIIME 2[64]. Analyses were run on a local compute cluster that comprised 1,024 Intel Ivy-bridge compute cores, as well as 384 AMD compute cores, and 12 TB of total RAM with a 10 gigabit ethernet (GbE) compute network for approximately 5 months of computational wall-time; typical job submissions for a single cancer type used ~30 cores and ~250 GB of RAM.

Quantitative measurement and normalization of TCGA technical variation

Cognizant of how technical variation between TCGA sequencing centres ($n = 8$), sequencing platforms ($n = 6$), experimental strategy (WGS versus RNA-seq), and possible contamination could confound our results, we developed a pipeline to quantify and remove batch effects while maintaining or increasing the signal attributed to biological variables. In brief, we filtered out samples with poor metadata quality (that is, missing race or ethnicity, ICD10 codes, DNA/RNA analyte amounts, or FFPE status information); transformed our discrete taxonomical count data to approximately normally distributed, log-count per million (log-cpm) data using the Voom algorithm[27], which models and removes the data's heteroskedasticity; and lastly performed supervised normalization (SNM) on the data to remove all significant batch effects while preserving biological effects[28]. Voom is traditionally used in combination with limma[65] for differential expression (or abundance) analysis of discrete count data, but we used it only for the algorithmic transformation to 'microarray-like' data, which permitted subsequent SNM. The Voom and SNM model matrices were equivalent and built using sample type as the target biological variable ($n = 7$; for example, primary tumour tissue) owing to expected biological differences between them, for which signal should be preserved during the SNM; conversely, the following were modelled as technical covariates to be

mitigated during SNM: sequencing centre ($n = 8$), sequencing platform ($n = 6$), experimental strategy ($n = 2$), tissue source site ($n = 191$), and FFPE status ($n = 2$; 'yes' or 'no'). It was not possible to model disease type as the target biological variable owing to complete confounding between certain types of cancer and sequencing centres (that is, some types of cancer were only sequenced at one TCGA site). During the Voom transformation, weighted trimmed mean of M-values (TMM) normalization from the edgeR package[66] was used for most data ('full dataset', 'likely contaminants removed' data, 'plate–centre decontaminated' data, and 'all putative contaminants removed' data) while dropping unvarying features (filterByExpr() function; edgeR), as shown by limma's user guide (https://www.bioconductor.org/packages/devel/bioc/vignettes/limma/inst/doc/usersguide.pdf; chapter 15: RNA-seq data, p. 70). In other cases ('most stringently filtered' data, 'SHOGUN TCGA data', 'Kraken TCGA data matched to SHOGUN TCGA data', and both plasma microbiome data sets), quantile normalization was used because downstream SNM correction was not compatible with stringently filtered TMM normalized, feature-dropped data, as these data sets already had significantly reduced or low feature counts. With the exception of 'most stringently filtered' data, all quantile-normalized data sets were compared only to other quantile-normalized data sets. Principal components were calculated before and after SNM correction of the Voom-adjusted data, and principal variance components analysis (PVCA)[29,30] quantified these changes between raw count data, Voom-adjusted data, and Voom-SNM normalized data. The mathematical basis for PVCA is well described by the NIEHS (https://www.niehs.nih.gov/research/resources/software/biostatistics/pvca/index.cfm), and we set the one tunable parameter to 80%, based on their recommendation of 60–90%.

Using SourceTracker2 as a validation analysis to address contamination concerns

Shotgun sequencing data from the NIH's HMP2 Project[33], which swabbed eight body sites among 217 total samples, were downloaded and run using the same TCGA Kraken microbial-detection pipeline as described above, including against the same microbial database ($n = 59{,}974$ bacterial, archaeal, and viral metagenomes) for taxonomy assignments. HMP2 data were summarized at the genus level, per our TCGA cancer microbiome data, and then were used to train a Bayesian source tracking model (SourceTracker2; https://github.com/biota/sourcetracker2)[32]. Details of the Bayesian model have been previously described by our laboratory[32]. Using SourceTracker parlance, these HMP2 samples served as 'sources' while the Voom-SNM-normalized samples acted as 'sinks', and the SourceTracker algorithm was used to calculate the proportion of each source attributable to each sink. In lay terms, we estimated the proportion of body site from HMP2 data attributable to each Voom-SNM-normalized cancer microbiome sample using the Bayesian model. After (i) intersecting the genera in our cancer microbiome data set with those in HMP2, (ii) converting the $\log_2$(cpm)-normalized values to scaled relative abundances (scaled by 10^6 to give approximately 1 million total reads, as HMP2 data had 917,450 reads), and (iii) converting the data to BIOM table format[67], we applied the model to solid-tissue normal samples from the TCGA COAD cohort ($n = 70$) and on primary tumour SKCM samples ($n = 122$). SKCM primary tumour samples were chosen instead of solid-tissue normal samples as the best proxy of skin flora, as only one adjacent solid-tissue normal sample for SKCM was available (Supplementary Table 4). SourceTracker2 default settings (alpha1 = 0.001, alpha2 = 0.1, beta = 10, restarts = 10, draws_per_restart = 1, burnin = 100, delay = 1) were used for both runs. The outputs were calculated in terms of mean fractional contributions of each source to each sink; averages and standard errors of these values were subsequently calculated. Statistical differences between the faecal contributions to COAD and SKCM samples (Extended Data Fig. 5b) were calculated using a one-sided Mann–Whitney U-test. The above protocol was repeated for the four decontaminated data sets to generate Extended Data Fig. 6c–f.

TCGA ML methods

Stochastic gradient-boosting machine (GBM) learning models were trained, automatically tuned, and tested using the R programming language (https://www.r-project.org/), GBM package[68,69], and Caret package[70]. Training and testing occurred on separate, randomly selected, stratified sampling splits of 70% and 30% of the data, respectively, and a fixed random number seed was used to ensure reproducibility of the model results and comparability among models. During model training, the data were first centred and scaled for each sample to have mean zero and unit standard deviation; fourfold cross validation was used to create multiple subsets of the training data and to perform a basic grid search optimization of GBM parameters, including interaction depth (1, 2, or 3) and number of trees (50, 100, or 150), while maximizing AUROC of the final, fully trained model. Learning rate (shrinkage) was held constant at 0.1 and the number of minimum observations per node was fixed at 5. In cases of class imbalance, we upsampled the minority class during training to help the model to generalize, after observing that other methods (differential class weighting, downsampling the majority class, minority class interpolation) did not consistently improve performance (data not shown). Comparisons were not made when the minority class contained fewer than 20 samples in total, owing to the inability of the classifiers to be adequately trained and tested with so few samples. Final model performances, including ROC curves, PR curves, and confusion matrices (generated with a probability cutoff of 50% for class #1 versus class #2 discrimination), were generated by applying the final model to the unseen 30% holdout test set. ROC and PR curves as well as AUROC and AUPR values were calculated using the PRROC package[71] while confusion matrices were calculated using the Caret package[70]. Variable importance scores of the resultant, non-zero model features were estimated using the GBM and Caret packages[68–70]. The percentage contribution of a particular feature to the model's predictions was estimated by dividing that feature's variable importance score by the sum of all variable importance scores for a given model (compare Extended Data Fig. 8a, b).

TCGA ML benchmarking and generalizability

As a benchmarking and generalizability assessment, we split TCGA into two stratified data halves (across sequencing centre, sample type, and disease type) of raw Kraken-derived, genus-level microbial count data (split #1: $n = 8{,}814$; split #2: $n = 8{,}811$ samples), ran them both separately through our Voom-SNM protocol, built separate ML models on each normalized half, and then tested these tuned ML models on each other's normalized data. We then compared these model performances against a third ML model that was built on the full Voom-SNM-normalized data set ($n = 17{,}625$ samples) and used 50–50% training and testing splits. Final performance was compared across all three approaches using their respective 50% holdout test set AUROC and AUPR. For additional internal validation, we built models to predict one cancer type versus all others using just (i) RNA samples or (ii) DNA samples, as well as on (iii) samples from one sequencing centre that only did RNA-seq (UNC) or (iv) DNA-seq (HMS) (Extended Data Fig. 3).

TCGA decontamination analyses

Broadly speaking, there are two classes of possible contamination that affect next-generation sequencing data: external contamination (for example, reagents, investigators' or subjects' bodies, environmental contributions) and internal contamination (that is, cross-contamination between samples during processing or sequencing)[14,16]. Our overall decontamination approach attempts to (i) simulate contamination to estimate its contribution to predictive performance and/or model unreliability, (ii) mitigate external contamination as much as possible, and (iii) measure the degree of internal contamination using sensible

positive and negative controls. External contaminants were identified and removed using sample analyte concentrations for all TCGA samples (n = 17,625), as recently described[16,38], and by using a blacklist of microorganisms identified from reagents in sequencing kits similar to those used in TCGA[14]. Internal contaminants are particularly difficult to identify without having access to the primary samples or knowing which other samples (especially non-cancer samples) were run at the same time. As such, the only internal contaminants that were identified and removed as clear cross-contaminants were four reads assigned to the Ebolavirus genus (two reads from one TCGA-LGG sample at The Broad Institute and two reads from one TCGA-HNSC sample at HMS), almost certainly from concurrent studies on the 2014 West Africa outbreak at these same sequencing centres during the TCGA study collection period (2006–2016)[72,73], and four reads assigned to the Marburgvirus genus (from two TCGA-OV samples at The Broad Institute), also probably of similar origin or as false positives (that is, Ebolavirus and Marburgvirus are both of the Filoviridae family). Doing so is in line with our previously published work that removes microbial assignments that cannot be related to the biology at hand[74]. It is further unlikely that such cross-contaminants, especially of extremely low abundance, would drive uniform discriminatory signals between and within types of cancer collected over many centres and years. For other possible cross-contaminants, we relied on estimating their contribution using Bayesian analyses (described above) of ecologically expected communities rather than their identification and removal.

First, we spiked five pseudo-contaminants into the raw data set (Extended Data Fig. 6a, top right) to track them through decontamination, SNM, and ML. This included the following: (1) 1,000 reads across all samples from HMS; (2) 1,000 reads across all samples from HMS, Baylor College of Medicine, Washington University School of Medicine, and Canada's Michael Smith Genome Sciences Centre; (3) 1,000 reads across all samples from all sequencing centres; (4) 10^6 reads spiked across 100 randomly selected samples from HMS; and (5) 10^6 reads spiked across 1,000 randomly selected samples from all sequencing centres. The mean raw read count across all samples and taxa was 1,481.20, so pseudo-contaminants containing 1,000 reads can be considered 'low-level' background while those with 10^6 reads are considered 'high-abundance' spikes. If pseudo-contaminants are present in downstream ML models after training, three interpretations are available: evaluate the percent predictive contribution of the pseudo-contaminants via feature importance scores and decide whether it is negligible or not; eliminate any ranked model features below the pseudo-contaminant; or, most conservatively, flag the entire model as being unreliable.

As TCGA did not include any negative blank reagent tubes during sample processing, we next attempted to pair a microbial blacklist at the genus level that used similar reagents and/or library preparation kits. TCGA SOPs mainly used QIAGEN products (Qiagen, Valencia, CA) for extracting DNA and RNA in tissues (DNA/RNA AllPrep kit) and DNA in blood (QiaAmp Blood Midi Kit)[52]. Salter and colleagues[14] described such a list (n = 94 genera) for DNA extraction kits in metagenomic experiments, including from QiaAmp kits that used the same silica membrane-based DNA purification as those used in TCGA blood extractions, obtained across four years of 'negative blank' sequencing and three high-throughput sequencing centres. Additional putative external contamination was identified on the basis that sequences from contaminants generally have frequencies that are inversely correlated with sample analyte concentration[16,38]. A robust statistical framework recently validated this principle[16], providing the opportunity to exploit sample DNA or RNA concentrations recorded by TCGA as a means to identify putative contaminants. The two main assumptions of this framework are (i) the contaminants are added in uniform amounts across samples; and (ii) the amount of contaminant DNA or RNA is small relative to the true sample DNA or RNA (microbial or host). Filtering was then conducted using the associated decontam R package (https://github.com/benjjneb/decontam)[16] using the recommended hyperparameter threshold (P^* = 0.1) and a more stringent approach (P^* = 0.5). Note, P^* = 0.5 means that taxonomies are classified as 'contaminant' or 'not' if the contaminant model or non-contaminant model fit the distribution better. As we found that sequencing centre contributed substantial variation to the raw count data, we processed the data in batches corresponding to them, whereby a taxon identified as a contaminant at any centre was subsequently discarded for all centres (that is, batch.combine = "minimum" in decontam). Putative lists of contaminants (P^* = 0.1: n = 283 genera; P^* = 0.5: n = 1,818 genera) were then combined/intersected with the microbial blacklist (n = 94 genera) and subtracted from the full data set. Manual literature inspection of the smaller combined contaminant list (n = 377) re-allowed 89 genera that were potentially pathogens or commensals (Supplementary Table 6). This resulted in three new data sets: 'likely contaminants removed', 'all putative contaminants removed', and 'most stringent filtering'. As a further conservative measure, we took all TCGA sample barcodes (for example, TCGA-02-0001-01C-01D-0182-01; as shown on NCI's documentation https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/) and extracted all sequencing plate–sequencing centre combinations, as named by the barcode's last two sets of integers (that is, plate 0182 from centre 01, or 0182-01, in this example). As decontam calculates the equivalent of a linear regression between taxon read fractions and analyte concentrations for all samples in a batch to determine whether a given taxon is classified as a contaminant, we required more than 10 samples per plate–centre combination to qualify as a batch, giving 351 total plate–centre batches. P^* = 0.1 was used (default value), and, as before, if a taxon was identified as a contaminant in any one of the 351 batches (batch.combine = "minimum"), it was removed from the data set (n = 421 taxa removed; Supplementary Table 8). After intersecting with the microbial blacklist, a total of 497 genera were removed. This provided the fourth decontaminated data set, and all of them were then processed through the same SNM and ML pipelines described above.

Comparing ML performances between BWA, SHOGUN, and Kraken data

BWA filtering occurred against the same database used to generate the Kraken-based assignments (n = 59,974 microbial genomes (bacteria, archaea and viruses)). Then, filtered BWA microbial count data were batch corrected in the same way as the Kraken data via Voom-SNM, except that DNA and RNA data were normalized separately owing to confounding between experimental strategy and sequencing centre of the reduced sample count. Samples from the raw Kraken-derived data were then matched against samples processed by BWA and normalized in the same way as the BWA data. This resulted in a total of four normalized data sets: DNA BWA data, RNA BWA data, DNA Kraken-subsetted data, and RNA Kraken-subsetted data. All four normalized data sets were then inputted for ML and their performances were compared to each other (Extended Data Fig. 4a–h).

The 'Web of Life' database[49] used for SHOGUN taxonomy assignments did not contain viruses, and SHOGUN processed a subset of all TCGA samples evaluated by Kraken (13,517 versus 17,625 samples). Thus, to make a fair comparison between their downstream ML performances, raw Kraken count data were subsetted to remove all identified viruses and to match the same samples processed by SHOGUN. Both data sets were then identically normalized by Voom (using quantile normalization) and SNM algorithms (using the same biological and technical variables as in the main TCGA analysis described above) before being fed into the ML pipelines for discrimination between and within types of cancer.

Complementary diagnostic analyses

When evaluating the applicability of blood mbDNA to low-grade cancers, all patients with stage Ia-c and IIa-c classified tumours were grouped together and all others were discarded. For comparisons

against the Guardant360 and FoundationOne Liquid ctDNA assays, all TCGA patients with at least one genomic alteration evaluated on their coding gene panels were filtered out; this included whether mutations were considered to be passengers or drivers. Remaining patients were used for ML analyses as described above.

TCGA simulations to estimate required sample sizes for validation study

To estimate the number of required samples from prostate, lung, and skin cancer (melanoma) for discrimination, we performed empirical simulations on TCGA blood samples at two different sequencing centres (Broad, HMS) that were all sequenced on one type of platform (Illumina HiSeq). We first used Kraken-derived microbial count data and then repeated the simulations with SHOGUN-derived microbial count data. This most closely mimicked the expected real-world experimental conditions of the validation study.

First, all TCGA PRAD, LUAD, LUSC, and SKCM blood samples at Broad and HMS that were sequenced on Illumina HiSeq machines were subsetted from the raw Kraken data of microbial counts (Broad: $n = 99$; HMS: $n = 288$). Our lung cancer samples from author S.P.P. were of mixed origin so we combined LUAD and LUSC blood samples into a single non-small-cell lung cancer (NSCLC) umbrella disease type; however, this applied only to Broad samples, as all blood-derived lung cancer samples at HMS were LUAD in origin. This left a breakdown of samples as follows: HMS: 66 LUAD, 104 PRAD, 118 SKCM; Broad: 42 NSCLC (24 LUAD, 18 LUSC), 17 PRAD, 40 SKCM. Then, each raw count data set for HMS and Broad was independently normalized through Voom (using quantile normalization) and SNM algorithms, using disease type as the biological variable of interest and tissue source site as the technical variable, as all other technical factors were precluded by picking a single sequencing centre, data type, and platform.

The simulations were performed as follows on the normalized data sets: (1) random stratified sampling picked equal numbers of samples from the three classes; (2) one sample of the three-class subsample was left out; (3) an ML model was built on all the remaining samples in the subsample and applied on the left-out sample to make a prediction with a certain probability; (4) steps 2–3 were repeated until all samples had been iterated through; (5) using the list of observed classes and list of predicted classes along with their probabilities, multi-class performance metrics were estimated; (6) another stratified random sample was selected of the same sample size and steps 2–5 were repeated nine more times (a total of ten times) to estimate standard errors of the multi-class performance metrics; (7) steps 1–6 were repeated for individual class sample sizes of 5–40 with a step size of five samples. In cases where the stratified sampling size was larger than the number of samples in a class, all samples in that class were used. Collectively, this provided an estimate of the number of samples required to perform multi-cancer discrimination well (Extended Data Fig. 9a). The empirical performance estimates (mean AUROC, mean AUPR) suggest that having at least 15 samples per cancer class should be sufficient. Note that it was not possible to estimate an ideal sample size for healthy controls because TCGA did not include them.

Clinical cohort selection and IRB protocols numbers

Biobanked, frozen plasma samples from 169 patients were analysed as part of this study, all from UC San Diego. All studies were approved by the Institutional Review Board (IRB) at UC San Diego, and under their respective IRB-approved protocols, patients provided written informed consent for sample donation and study. All prostate cancer plasma samples ($n = 59$) came from R.M. under IRB protocol 131550. All lung cancer and melanoma plasma samples came from S.P.P. under IRB protocol 150348. All cancer- and HIV-free healthy control subjects ($n = 69$) came from a group led by R.H. under the following IRB protocol numbers: 130296, 091054, 172092, 151057, and 182064.

Plasma-derived, cell-free microbial DNA sample processing, and sequencing

Total circulating DNA was extracted from a volume of 250 µl plasma from each sample using the QIAamp Circulating Nucleic Acid Kit (QIAGEN) according to the manufacturer's instructions, and purified with AMPure XP SPRI paramagnetic beads (Beckman Coulter). Sequencing libraries were prepared from purified cfDNA using the KAPA HyperPlus Kit (Kapa Biosystems) with standard Illumina indexed adapters (IDT) as described[47]. Sample libraries were characterized using the Agilent 4200 TapeStation System (High Sensitivity DNA Kit) and quantified by qPCR using the NEBNext Library Quant Kit for Illumina (New England Biolabs). Paired-end 2 × 150-bp sequencing (S4 flow cell) was performed on a NovaSeq 6000 instrument (Illumina), and samples were pooled across all four lanes during sequencing.

Bioinformatic processing for plasma microbiome samples

A total of 21,600,141,264 reads were generated on the single NovaSeq 6000 sequencing run across all samples. Of these, 19,046,611,360 reads were assigned to human samples (that is, negative and positive controls removed), and 2.186% of the total reads were classified as non-human. Raw sequencing data were demultiplexed and adaptor-trimmed using Atropos[75]. Additional quality filtering was done using Trimmomatic with the following settings—(ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:7, MINLEN:50, TRAILING:20, AVGQUAL:20, SLIDINGWINDOW:20:20)[76]. An additional adaptor sequence consisting of a string of only G was added to the standard TruSeq3 adapters to remove trailing G stretches from the 5′ ends of reads. Read pairs were discarded if either mate mapped to the human genome (major-allele-SNP reference from 1000 Genomes Project) using Bowtie2 with the fast-local parameter set[62,77]. Paired-end reads were then merged using FLASH with the following parameters—(minimum overlap: 20, maximum overlap: 150, mismatch ratio: 0.01)[78].

The filtered, merged reads were then processed either by Kraken, using the same workflow and database ($n = 59{,}974$ microbial genomes) detailed above, or with SHOGUN as detailed here. Samples were processed on individual plasma microbiome samples (that is, on a per-sample-per-lane basis, as samples were pooled across all four sequencing flow cells during the run). After per-sample-per-lane taxonomy assignment by Kraken or SHOGUN, microbial counts across lanes were aggregated for each sample after hierarchical clustering procedures showed consistent grouping by sample IDs rather than by flow cell lane. For SHOGUN-derived data, both successfully merged and unmerged reads were used as input for the SHOGUN align function, using Bowtie2 to map reads against the WoL database to generate taxonomy profiles, which were then collapsed to the genus level using QIIME 2[64]. The taxonomy profiles of each sample were then filtered to remove all taxa whose relative abundance was less than 0.01%.

Plasma microbiome technical validation and data decontamination

To evaluate the performance of the sequencing run and bioinformatic microbial-detection pipelines, spiked wells and experimental serial dilutions of *Aliivibrio fischeri* (genus: *Aliivibrio*) included on the sequencing plate were examined against other sample types for differential abundance and in isolation for log-fold changes in abundance across dilutions. These technical positive controls are plotted in Extended Data Fig. 9b, c for both Kraken and SHOGUN-derived taxonomy assignments.

Three kinds of negative blank controls were included on the sequencing plate: (1) DNA extraction blanks, which had reagents from the DNA extraction stage through sequencing; (2) DNA library preparation blanks, which had reagents from the library preparation stage through sequencing; and (3) empty control wells, which had water added to them and then reagents during library preparation and would contain splashed and/or aerosolized microbial nucleic acids. As in the TCGA

analysis, decontam[16] was again used to decontaminate the plasma microbial data, except that it had access to both negative blank controls and DNA concentrations for all samples (excluding empty control wells for the latter). As a conservative measure, we used a $P^* = 0.5$ hyperparameter value for decontam for both 'prevalence' (that is, blank-based) and 'frequency' (that is, concentration-based) modes of decontamination; this hyperparameter value is equivalent to the most stringent decontamination in TCGA that discarded >90% of the total data. For prevalence mode, $P^* = 0.5$ will flag any taxon that is more prevalent in negative controls than biological ones as a contaminant; for frequency mode, $P^* = 0.5$ will flag any taxon whose model (that is, a regression model) fits a contaminant distribution better than a non-contaminant distribution using read fractions and DNA concentrations[16]. For Kraken count data, prevalence mode discarded 21 taxa and frequency mode discarded 1,261 taxa (out of 1,753 original assignments); for SHOGUN count data, prevalence mode discarded 57 taxa and frequency mode discarded 244 taxa (out of 1,181 original assignments). Decontaminated data for both Kraken and SHOGUN were fed into downstream normalization and ML pipelines.

Plasma microbiome data normalization, permutation testing, and ML

An attempt to predict age using raw microbial count data was performed using GBM ML models (architectures same as those described above for TCGA) and leave-one-out (LOO) iterative ML (Extended Data Fig. 9g).

To confirm the importance of normalizing for age and gender in this cohort, we performed a permutation analysis with 100 iterations for each factor and then simultaneously for both factors (Extended Data Fig. 9h–j). In brief, the following four steps were performed: (1) randomly swap age and/or sex labels among all samples; (2) run Voom-SNM on the raw data, using disease type as the biological variable of interest and permuted age and/or sex as the technical factors; (3) perform an ML analysis to discriminate grouped cancer samples from healthy controls using 70%–30% training–testing splits with a fixed random number seed and internal fourfold cross validation to obtain a two-class performance estimate (AUROC, AUPR); (4) repeat steps 1–3 for a total of 100 times to create a null performance distribution. Next, using correct, fixed age and/or sex assignments, we ran steps 2–3 a total of 100 times while randomly selecting the random number seed in step 3. Last, this performance distribution was directly compared to its null distribution for significance using a two-sided Mann–Whitney *U*-test. As all of these tests were extremely significant (all $P \leq 1.5 \times 10^{-13}$), we incorporated age and sex as technical factors in the Voom-SNM while holding disease type as the biological variable of interest. Note, all lung cancer samples were labelled with a consolidated disease type label during normalization regardless of pathological subtype, as done in the TCGA cancer simulations (described above). All negative blank and positive monoculture controls were removed before Voom-SNM.

ML on the Voom-SNM normalized plasma microbiome samples was done exactly as previously described for TCGA samples, except for the sampling schema, because the sample sizes were smaller by orders of magnitude. First, to estimate generalization of healthy versus grouped cancer discriminations, we 'bootstrapped' 70%–30% training–testing splits with fourfold cross-validation during training for 500 iterations. Sampling with replacement was allowed in that every training–testing split (that is, every iteration) was unique; however, in no case was a sample allowed to be both a training case and a testing case. Summary statistics on the resultant performance metrics from all 500 iterations estimated the AUROC and AUPR distributions and CIs (Fig. 4b, Extended Data Fig. 10a). Second, pairwise and multi-class discriminations between and among healthy controls and individual types of cancer were done with LOO ML. In other words, one sample was iteratively left out, a model was iteratively trained on the remaining samples with fourfold cross-validation for hyperparameter tuning, and a prediction was iteratively made on the left-out sample with a probability given by the model. The final list of actual classes for all samples was compared to the list of predicted classes and their probabilities to estimate AUROC and AUPR metrics, as described previously using the PRROC R package[71]. Multi-class performance was estimated by taking the mean of all one-versus-all-others comparisons, as reported by the multiClassSummary() function in the caret R package[70].

Iterative subsampling to evaluate the contribution of smaller samples sizes to the melanoma cohort performance (Extended Data Fig. 9k) was done as follows: (1) perform random stratified sampling of a single cancer type and healthy controls of 16 samples each (32 total); (2) perform LOO iterative ML and evaluate performance on those 32 samples for healthy versus cancer discrimination; (3) repeat steps 1–2 100 times to estimate performance standard errors; (4) repeat steps 1–3 for each of the three types of cancer. The same process was also done for iterative subsampling of PC and LC cohorts to study the impact of decreased sample size on their discrimination. Note that the entire melanoma cohort was used during each stratified subsampling, as the goal was to compare its cohort size to the other sample sizes.

Statistical analyses

All statistical analyses were done using R version 3.4.3. The ggpubr package (https://github.com/kassambara/ggpubr) performed nonparametric statistical testing between groups and accounted for multiple hypothesis testing correction when necessary. Note that *P* values less than 2.2×10^{-16} cannot be accurately calculated by R, so *P* values less than this are listed as $<2.2 \times 10^{-16}$; it is not a range of *P* values. Measurements were taken from distinct samples and not by repeatedly measuring samples. Sample size estimates for the validation study came from empirical simulations with TCGA blood samples and relied on the GBM package[68,69], Caret package[70], and MLmetrics package (https://github.com/yanyachen/MLmetrics) for performing ML and multi-class performance estimation. All other multi-class performance estimates were calculated using the Caret[70] and MLmetrics packages.

Website

The website referenced in this manuscript can be accessed at http://cancermicrobiome.ucsd.edu/CancerMicrobiome_DataBrowser. Links are directly available on the website to the data repository as well as the code hosted on GitHub (upper right side of the website). Additionally, there are five separate tabs (left-hand side of the website) for interactively plotting Kraken and SHOGUN-derived normalized microbial abundances in TCGA and the plasma microbiome data (raw counts or normalized data), as well as for interactively examining all model performances and ranked feature lists shown in this paper (approximately 600 models).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Pre-processed cancer microbiome data generated and analysed in this study (that is, summarized read counts at the genus taxonomic level) as well as the metadata are available at ftp://ftp.microbio.me/pub/cancer_microbiome_analysis/. Raw outputs of Kraken- or SHOGUN-processed TCGA sequencing data comprise hundreds of terabytes of files and are not directly available unless otherwise coordinated with the corresponding author. However, all raw TCGA data and the bioinformatics pipeline necessary to generate such raw outputs from Kraken can be accessed through SevenBridge's CGC. Each of the hundreds of ML models in this work generated a list of ranked features used to make predictions, and we provide the code to generate these lists, in addition to showing them on our website. Raw data for the plasma validation

study are available through the European Nucleotide Archive (accession IDs ERP119598 (HIV-free); ERP119596 (PC); ERP119597 (LC and SKCM)); these data and the SHOGUN-processed data for the plasma validation study are available in Qiita (https://qiita.ucsd.edu/)[79] under study IDs (12667 (HIV-free); 12691 (PC); 12692 (LC and SKCM)).

Code availability

All programming scripts used to access, manage, and run data on the CGC as well as development of the supervised normalization, decontamination, ML pipelines, and so forth can be found at our GitHub repository link: https://github.com/biocore/tcga. These can be applied directly to the summarized, genus-level count data given above. Our CGC pipeline is also publicly shareable and available upon reasonable request from the corresponding author.

51. Lau, J. W. et al. The Cancer Genomics Cloud: collaborative, reproducible, and democratized—a new paradigm in large-scale computational research. *Cancer Res*. **77**, e3–e6 (2017).
52. Hoadley, K. A. et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291–304.e6 (2018).
53. Reynolds, S. M. et al. The ISB Cancer Genomics Cloud: a flexible cloud-based platform for cancer genomics research. *Cancer Res*. **77**, e7–e10 (2017).
54. Ellrott, K. et al. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst*. **6**, 271–281.e7 (2018).
55. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
56. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. **2**, 401–404 (2012).
57. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal*. **6**, pl1 (2013).
58. Land, M. L. et al. Quality scores for 32,000 genomes. *Stand. Genomic Sci*. **9**, 20 (2014).
59. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
60. Greathouse, K. L. et al. Interaction between the microbiome and TP53 in human lung cancer. *Genome Biol*. **19**, 123 (2018).
61. Shanmughapriya, S. et al. Viral and bacterial aetiologies of epithelial ovarian cancer. *Eur. J. Clin. Microbiol. Infect. Dis*. **31**, 2311–2317 (2012).
62. Banerjee, S. et al. The ovarian cancer oncobiome. *Oncotarget* **8**, 36225–36245 (2017).
63. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
64. Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol*. **37**, 852–857 (2019).
65. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. **43**, e47 (2015).
66. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
67. McDonald, D. et al. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. **1**, 2047-217X-1-7 (2012).
68. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal*. **38**, 367–378 (2002).
69. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat*. **29**, 1189–1232 (2001).
70. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw*. **28**, 1–26 (2008).
71. Grau, J., Grosse, I. & Keilwagen, J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* **31**, 2595–2597 (2015).
72. Gire, S. K. et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369–1372 (2014).
73. Matranga, C. B. et al. Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. *Genome Biol*. **15**, 519 (2014).
74. Gonzalez, A. et al. Avoiding pandemic fears in the subway and conquering the platypus. *mSystems* **1**, e00050-16 (2016).
75. Didion, J. P., Martin, M. & Collins, F. S. Atropos: specific, sensitive, and speedy trimming of sequencing reads. *PeerJ* **5**, e3720 (2017).
76. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
77. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
78. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
79. Gonzalez, A. et al. Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods* **15**, 796-798 (2018).

Acknowledgements We acknowledge conversations with C. Sepich, C. Martino, R. Bejar, and H. Carter. G.D.P. has been supported by training grants from the National Institutes of Health during the course of this work (5T32GM007198-42; 5T32GM007198-43). S.F. is partially funded through trainee support from Merck KGaA in partnership with the Center for Microbiome Innovation at UC San Diego. Samples acquired for the validation cohort were collected under the following grants: R00 AA020235, R01 DA026334, P30 MH062513, P01 DA012065, and P50 DA026306. The Seven Bridges Cancer Genomics Cloud was used during the course of this work and has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Contract No. HHSN261201400008C, and ID/IQ Agreement No. 17X146 under Contract No. HHSN261201500003I. This work was supported in part by the Chancellor's Initiative in the Microbiome and Microbial Sciences (R.K., A.D.S., S.M.-M.) and by Illumina, Inc. through reagent donation in partnership with the Center for Microbiome Innovation at UC San Diego. We thank G. Humphrey and K. Sanders for sample processing, and G. Ackermann, A. Gonzalez, and J. DeReus for assistance with metadata curation and data handling.

Author contributions The research topic was developed by E.K., G.D.P., T.K., S.J., J.M., S.J.S., S.M.-M., A.D.S., S.P.P., and R.K. The TCGA microbial-detection pipeline was co-developed by E.K., S.J.S., J.M., J.K., and G.D.P. The supervised normalization pipeline was developed by G.D.P., the decontamination pipeline by G.D.P., A.D.S., and S.P.P., and the ML pipeline by G.D.P., A.D.S., T.K., and S.J. SourceTracker2 analyses, including re-running HMP2 shotgun metagenomic data through the microbial-detection pipeline, were completed by E.K., Q.Z., and G.D.P. Samples for the validation study were collected by R.H., R.M., and S.P.P., processed for sequencing by C.C., S.F., and G.D.P., bioinformatically analysed by E.K., S.W., and A.D.S., and then put through normalization and ML pipelines by G.D.P. and A.D.S. The cell-free microbial DNA extraction protocol was originally designed and refined by C.C., S.F., S.M.-M., and A.D.S. The original version of the manuscript was written by G.D.P., A.D.S., S.P.P., and R.K. All authors contributed to the final version of the manuscript.

Competing interests Clarity Genomics, the employer of E.K., did not provide funding for this study. G.D.P. and R.K. have jointly filed U.S. Provisional Patent Application Serial No. 62/754,696 and International Application No. PCT/US19/59647 on the basis of this work. G.D.P., R.K., and S.M.-M. have started a company to commercialize the intellectual property. R.K. is a member of the scientific advisory board for GenCirq, holds an equity interest in GenCirq, and can receive reimbursements for expenses up to US$5,000 per year. R.K., A.D.S., and S.M.-M. are directors at the Center for Microbiome Innovation at UC San Diego, which receives industry research funding for various microbiome initiatives, but no industry funding was provided for this cancer microbiome project.

Additional information

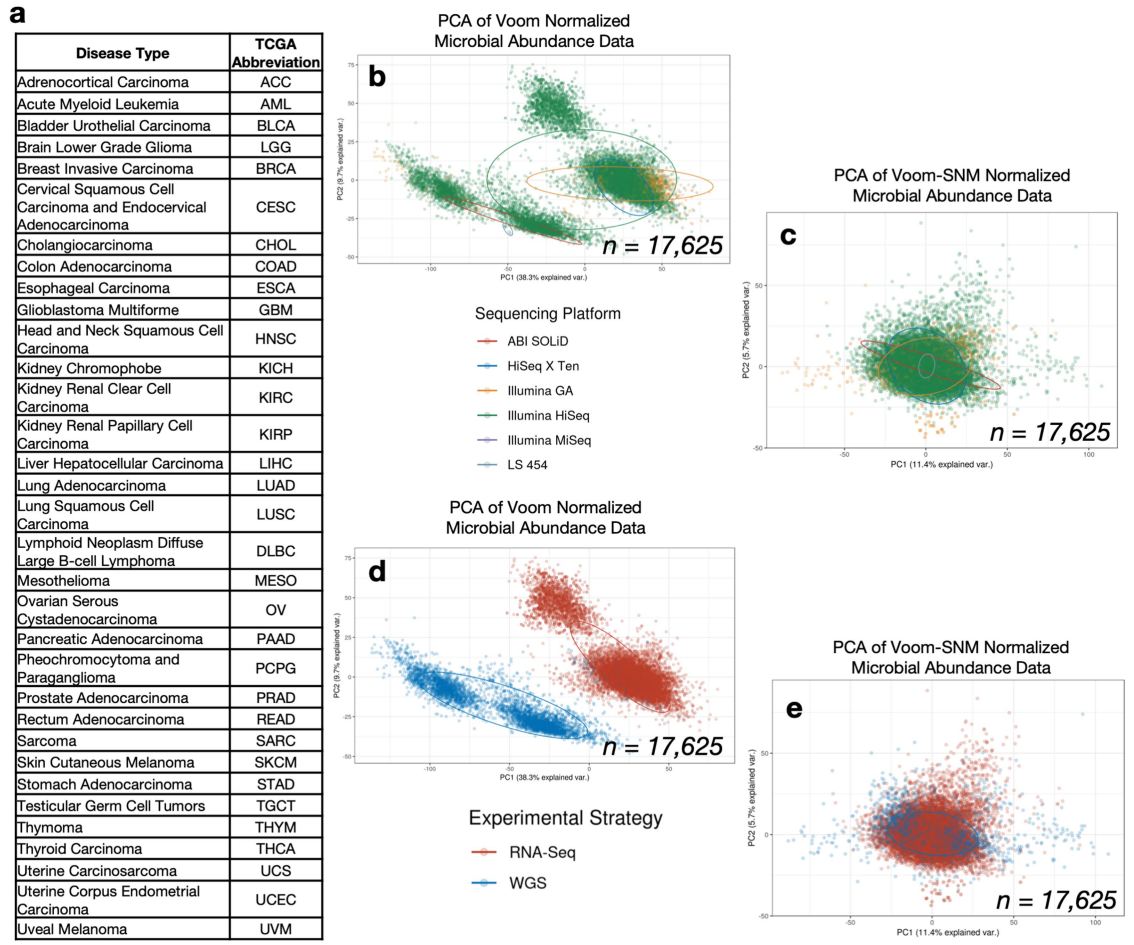
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2095-1.

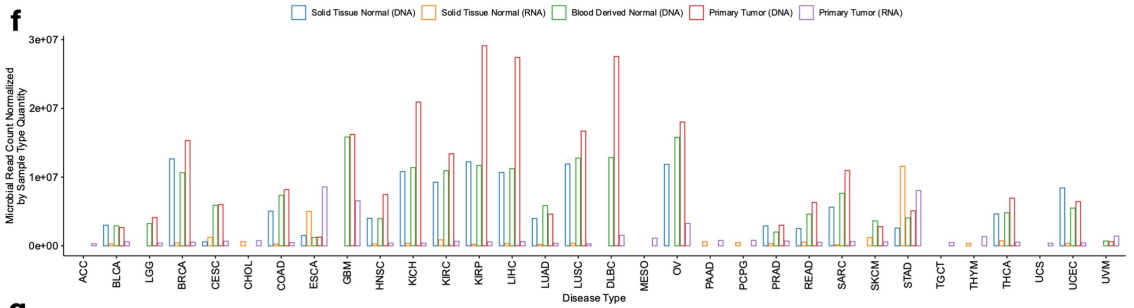
Correspondence and requests for materials should be addressed to R.K.

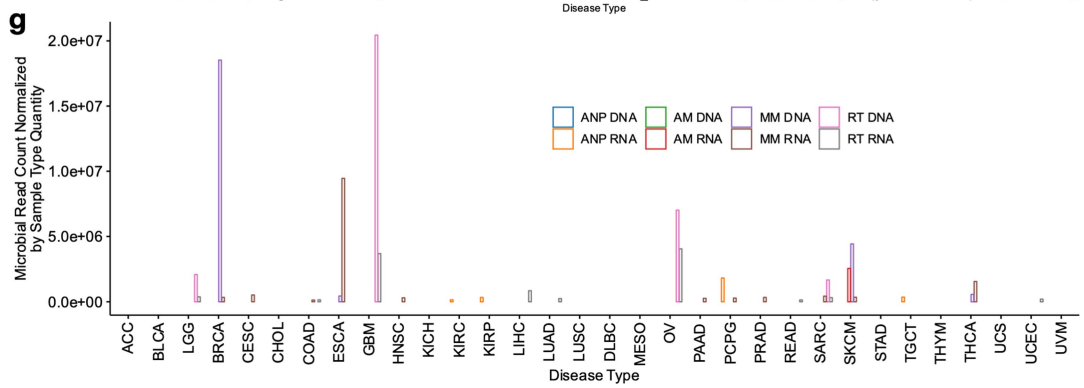
Peer review information *Nature* thanks Eran Elinav, Victor Velculescu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

a

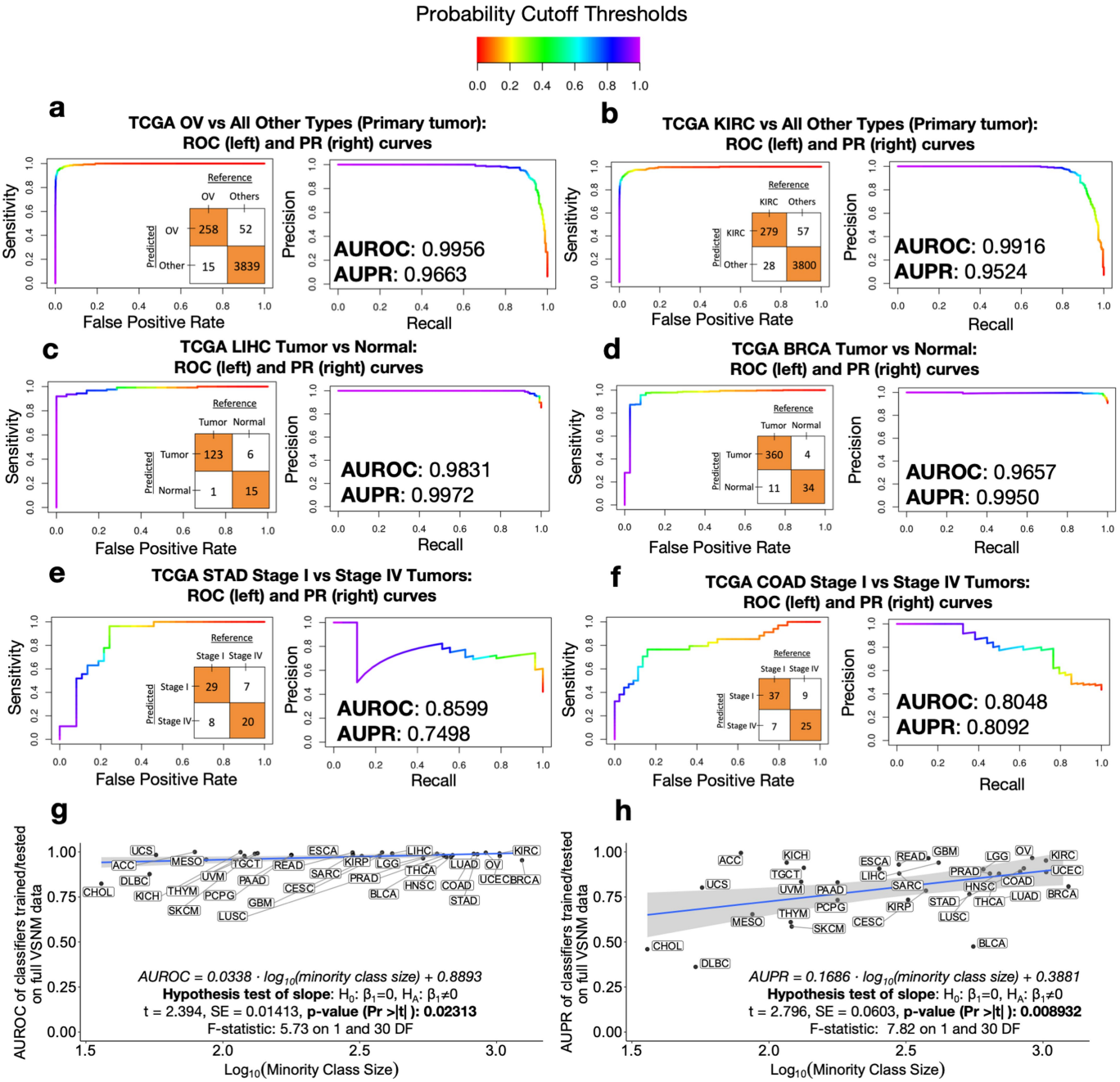
Disease Type	TCGA Abbreviation
Adrenocortical Carcinoma	ACC
Acute Myeloid Leukemia	AML
Bladder Urothelial Carcinoma	BLCA
Brain Lower Grade Glioma	LGG
Breast Invasive Carcinoma	BRCA
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	CESC
Cholangiocarcinoma	CHOL
Colon Adenocarcinoma	COAD
Esophageal Carcinoma	ESCA
Glioblastoma Multiforme	GBM
Head and Neck Squamous Cell Carcinoma	HNSC
Kidney Chromophobe	KICH
Kidney Renal Clear Cell Carcinoma	KIRC
Kidney Renal Papillary Cell Carcinoma	KIRP
Liver Hepatocellular Carcinoma	LIHC
Lung Adenocarcinoma	LUAD
Lung Squamous Cell Carcinoma	LUSC
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC
Mesothelioma	MESO
Ovarian Serous Cystadenocarcinoma	OV
Pancreatic Adenocarcinoma	PAAD
Pheochromocytoma and Paraganglioma	PCPG
Prostate Adenocarcinoma	PRAD
Rectum Adenocarcinoma	READ
Sarcoma	SARC
Skin Cutaneous Melanoma	SKCM
Stomach Adenocarcinoma	STAD
Testicular Germ Cell Tumors	TGCT
Thymoma	THYM
Thyroid Carcinoma	THCA
Uterine Carcinosarcoma	UCS
Uterine Corpus Endometrial Carcinoma	UCEC
Uveal Melanoma	UVM



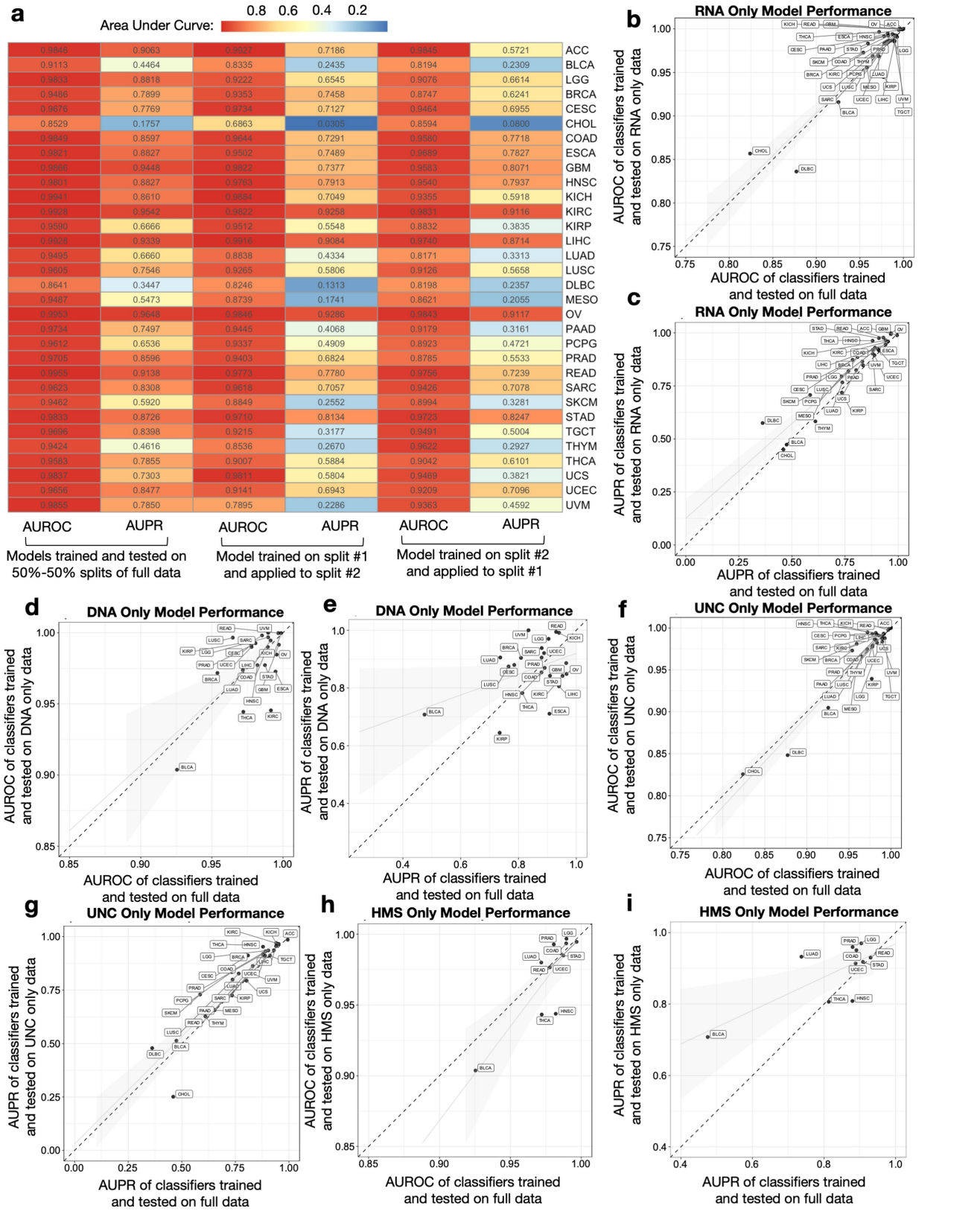




Extended Data Fig. 1 | Continued overview of the TCGA cancer microbiome. **a**, TCGA study abbreviations. **b**, PCA of Voom-normalized data, where colours represent sequencing platform of the sample and each dot denotes a cancer microbiome sample. **c**, PCA of the data following consecutive Voom-SNM supervised normalization, as labelled by sequencing platform. **d**, PCA of Voom-normalized data, where colours represent experimental strategy of the sample and each dot denotes a cancer microbiome sample. **e**, PCA of the data following consecutive Voom-SNM supervised normalization, as labelled by experimental strategy. **f**, **g**, Microbial reads counts as normalized by the quantity of samples within a given sample type across all types of cancer in TCGA after metadata quality control (Fig. 1b), including the three major sample types analysed in the paper (**f**) and the remaining sample types (**g**). ANP, additional, new primary; AM, additional metastatic; MM, metastatic; RT, recurrent tumour. For PCAs of raw and normalized data, $n = 17{,}625$; the number of samples per cancer type and per tissue type are shown in Supplementary Table 4.

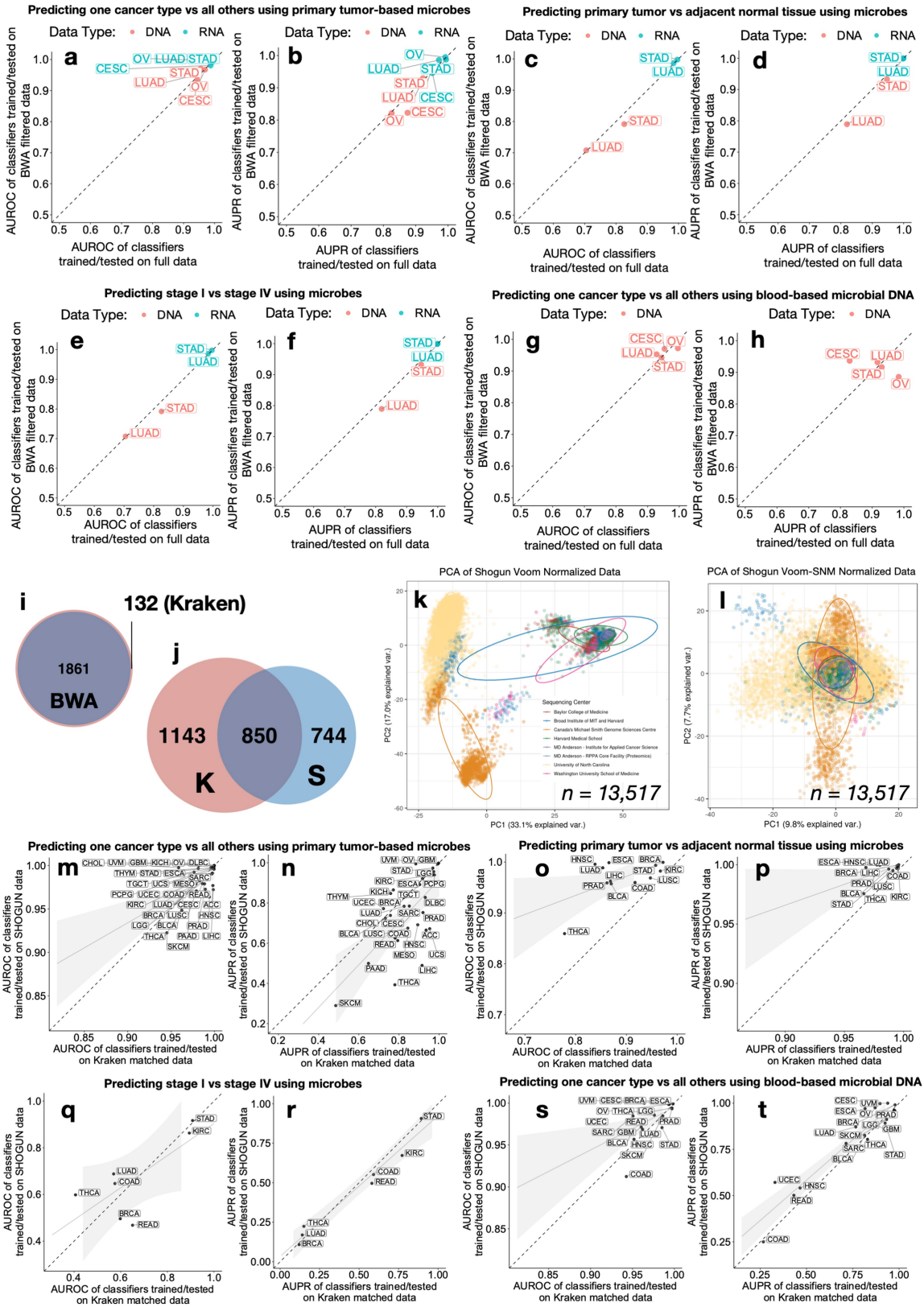


Extended Data Fig. 2 | Performance metrics details discriminating between and within TCGA types of cancer using microbial abundances. a–f, Expanded examples from the heatmaps in Fig. 1f–h. A colour gradient (top) denotes the probability threshold at any point along the ROC and PR curves. An inset confusion matrix is shown using a 50% probability threshold cutoff, which can be used to calculate sensitivity, specificity, precision, recall, positive predictive value, negative predictive values, and so forth at the corresponding point on the ROC and PR curves. **g**, **h**, Linear regressions of model performance, specifically AUROC (**g**) and AUPR (**h**), for discriminating between types of cancer in a one-cancer-type-versus-all-others manner, as a function of minority class size. Performances are shown for models using microorganisms detected in primary tumours, for which we had the greatest number of samples ($n = 13{,}883$) and types of cancer ($n = 32$) to compare. As AUROC and AUPR have domains of [0,1] and the minority class size varied from 20 to 1,238 samples, the latter is regressed on a $\log_{10}$ scale. Inset hypothesis tests and associated *P* values are based on the null hypothesis of there being no relationship between the dependent and independent variables (two-sided hypothesis test of slope). The number of samples included to evaluate performance of each comparison can be found in the data browser confusion matrices at http://cancermicrobiome.ucsd.edu/CancerMicrobiome_DataBrowser.



Extended Data Fig. 3 | See next page for caption.

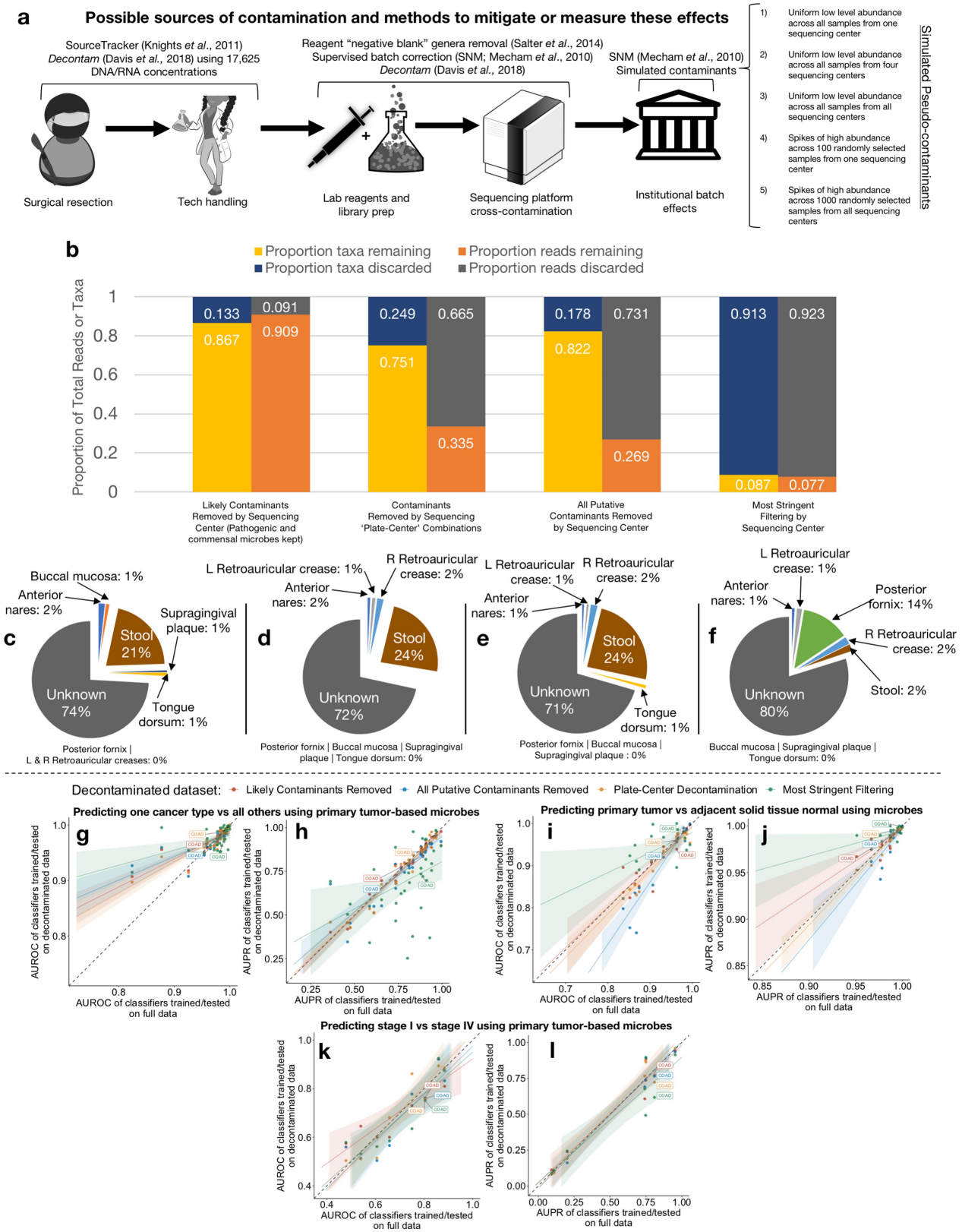
Extended Data Fig. 3 | Internal validation of ML model pipeline. a, Two independent halves of TCGA raw microbial count data were normalized and used for model training to predict one cancer type versus all others using tumour microbial DNA and RNA; each model was then applied to the other half's normalized data. This heatmap compares the performances of these models compared to training and testing on 50–50% splits of the full data set (split 1: n = 8,814 samples; split 2: n = 8,811 samples; total samples: n = 17,625). **b**, **c**, Model performance comparison when subsetting the full Voom-SNM data by primary tumour RNA samples (n = 11,741) across multiple sequencing centres to predict one cancer type versus all others (**b**, AUROC; **c**, AUPR). **d**, **e**, Model performance comparison when subsetting the full Voom-SNM data by primary tumour DNA samples (n = 2,142) across multiple sequencing centres to predict one cancer type versus all others (**d**, AUROC; **e**, AUPR). **f**, **g**, Model performance comparison when subsetting the full Voom-SNM data by samples from the UNC (n = 9,726), which only did RNA-seq, to predict one cancer type versus all others using primary tumour RNA samples (**f**, AUROC; **g**, AUPR). **h**, **i**, Model performance comparison when subsetting the full Voom-SNM data by samples from HMS (n = 898), which only did WGS, to predict one cancer type versus all others using primary tumour DNA samples (**h**, AUROC; **i**, AUPR). **b**–**i**, Generalized linear models with s.e. are shown in grey; dotted diagonal line denotes a perfect linear relationship; for sample size comparison, the full Voom-SNM data set contained 13,883 primary tumour samples.



Extended Data Fig. 4 | See next page for caption.

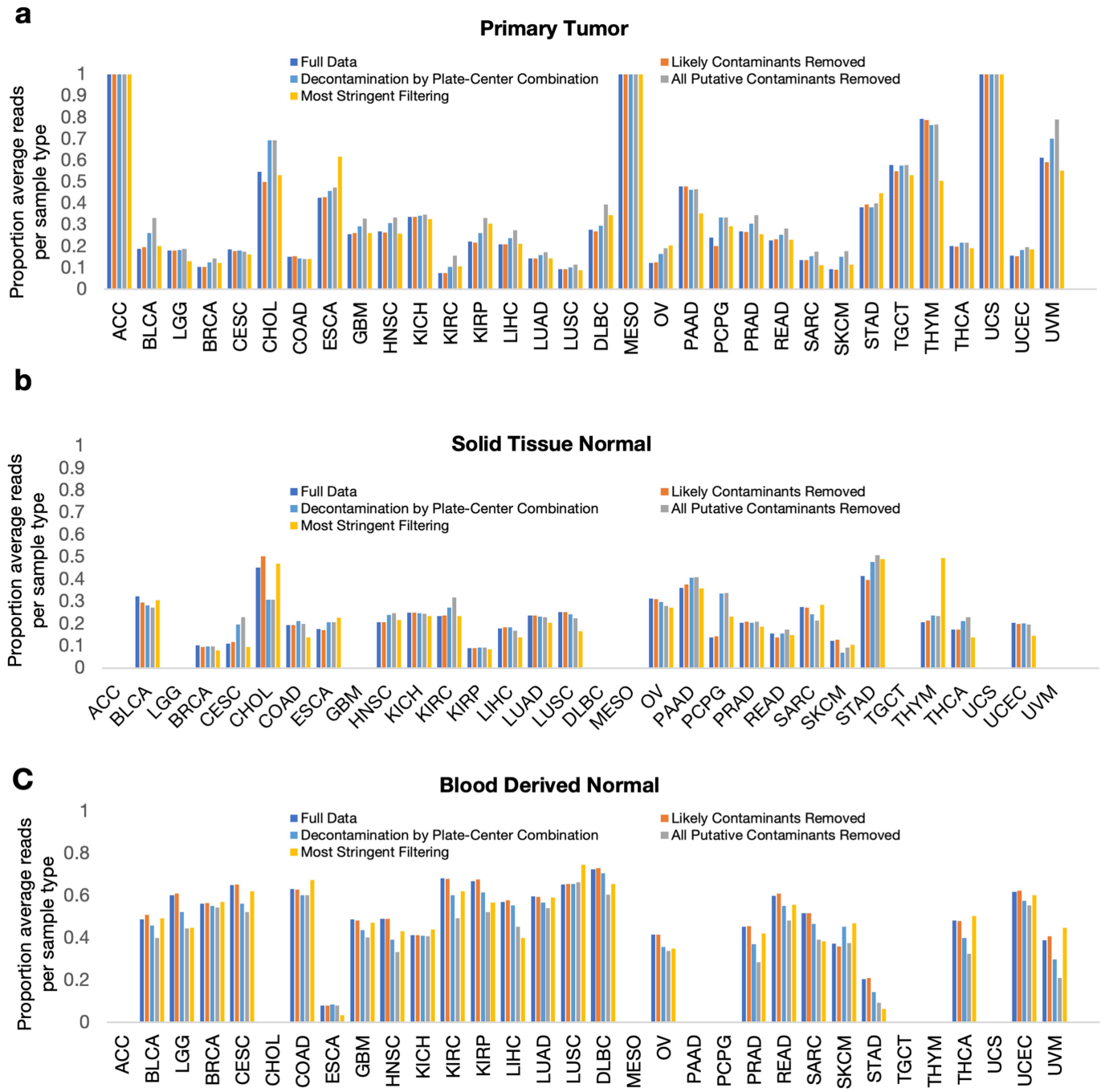
Extended Data Fig. 4 | Orthogonal validation of Kraken-derived TCGA cancer microbiome profiles and their ML performances. a–**h**, Four TCGA types of cancer (CESC, $n = 142$ (DNA) and $n = 309$ (RNA); STAD, $n = 322$ (DNA) and $n = 770$ (RNA); LUAD, $n = 351$ (DNA) and $n = 600$ (RNA); and OV, $n = 189$ (DNA) and $n = 850$ (RNA)) underwent additional filtering after Kraken-based taxonomy assignments via direct genome alignments (BWA[59]) using tumour microbial DNA and RNA. ML performances are compared between the normalized, BWA filtered data and matched, independently normalized Kraken data for one cancer type versus all others using primary tumour microorganisms (**a**, AUROC; **b**, AUPR), tumour-versus-normal discriminations (**c**, AUROC; **d**, AUPR), stage I versus stage IV tumour discriminations using primary tumour microorganisms (**e**, AUROC; **f**, AUPR), and one cancer type versus all others using blood-derived microorganisms (**g**, AUROC; **h**, AUPR) (see Methods). **i**, Venn diagram of the taxon count between the BWA filtered data and the Kraken full data. **j**–**t**, An orthogonal microbial-detection pipeline called SHOGUN[31] and a separate database[49] were run on a subset of TCGA samples ($n = 13,517$ total samples), normalized via Voom-SNM, analogous to its Kraken counterpart, and used for downstream ML analyses. **j**, Venn diagram of the SHOGUN-derived microbial taxa (S) and the Kraken-derived microbial taxa (K). Note that SHOGUN's database[49] does not include viruses whereas the Kraken database does. **k**, **l**, PCA of Voom (**k**) and Voom-SNM (**l**) normalized SHOGUN data, coloured by sequencing centre. **m**–**t**, ML performance comparisons between models trained and tested on SHOGUN data and matched Kraken data, using the same 70%–30% splits, for one cancer type versus all others using primary tumour microorganisms (**m**, AUROC; **n**, AUPR), tumour-versus-normal discriminations (**o**, AUROC; **p**, AUPR), stage I versus stage IV tumour discriminations using primary tumour microorganisms (**q**, AUROC; **r**, AUPR), and one cancer type versus all others using blood-derived microorganisms (**s**, AUROC; **t**, AUPR). For fair comparison, matched Kraken data were derived by removing all virus assignments in the raw Kraken count data and subsetting to the same 13,517 TCGA samples analysed by SHOGUN; these matched Kraken data were then normalized independently via Voom-SNM in the same way as the SHOGUN data (see Methods) and fed into downstream ML pipelines. For all ML performances, ≥ 20 samples in each class was required to be eligible. For regression subfigures, the dotted diagonal line denotes perfect performance correspondence; generalized linear models with s.e. ribbons are shown.

Extended Data Fig. 5 | Pan-cancer microbial abundances and an interactive website for TCGA cancer microbiome profiling and ML model inspection. **a**, Pan-cancer normalized abundances of *Fusobacterium* with a one-way ANOVA (Kruskal–Wallis) test for microbial abundances across types of cancer for each sample type. Sample sizes are inset in blue and box plots show median (line), 25th and 75th percentiles (box), and 1.5 × IQR (whiskers); TCGA study names are listed below. **b**, SourceTracker2 results for faecal contribution, as based on HMP2 data, for TCGA-COAD solid-tissue normal samples ($n = 70$) and TCGA-SKCM primary tumour samples ($n = 122$). Only one solid tissue normal sample was available for TCGA-SKCM (Supplementary Table 4), so primary tumours were used instead as the best proxy of expected skin flora. It is expected that colon samples should have higher faecal contribution than skin, so a one-sided Mann–Whitney *U*-test was used. As SourceTracker2 outputs the mean fractional contributions of each source (that is, HMP2) to each sink (that is, COAD, SKCM samples), the centre value of each bar plot is the mean of these values and the error bars denote the s.e.m. The sample sizes are shown below in blue. **c**, Pan-cancer normalized abundances of Alphapapillomavirus with a one-way ANOVA (Kruskal–Wallis) test for microbial abundances across types of cancer for each sample type. Sample sizes are inset in blue, and box plots show median (line), 25th and 75th percentiles (box), and 1.5 × IQR (whiskers); TCGA study names are listed below. TCGA studies that clinically tested patients for HPV infection are divided into negative and positive groups. **d**, Screenshot of interactive website showing plotting of Alphapapillomavirus normalized microbial abundances using Kraken-derived data. Plotting using SHOGUN-derived normalized microbial abundances is available on another tab of the website (left-hand side). **e**, Screenshot of interactive website of ML model inspection. Selecting the data type (for example, all likely contaminants removed), cancer type (for example, invasive breast carcinoma), and comparison of interest (for example, tumour versus normal) will automatically update the ROC and PR curves, as well as the confusion matrix (using a probability cutoff threshold of 50%) and the ranked model feature list. Website is accessible at http://cancermicrobiome.ucsd.edu/CancerMicrobiome_DataBrowser.

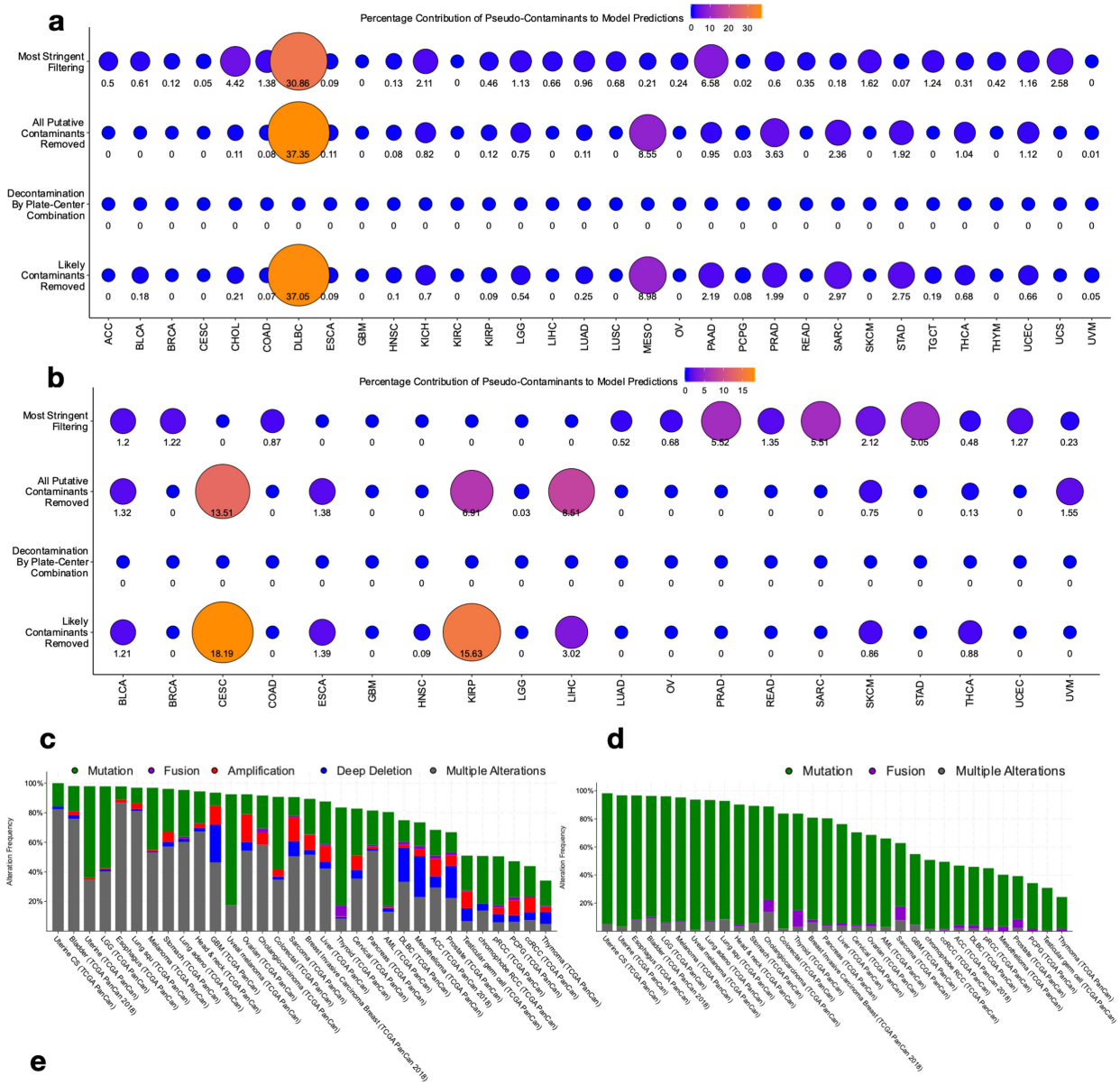


Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | The decontamination approach along with its results, benefits, and limitations on cancer microbiome data. a, Various approaches used to evaluate, mitigate, remove and/or simulate sources of contamination. **b**, The proportion of remaining taxa or microbial reads in TCGA after varying levels of decontamination. Decontamination by sequencing centre removed all taxa identified as a contaminant at any one sequencing centre ($n = 8$ batches); decontamination by plate–centre combinations removed all taxa identified as a contaminant on any single sequencing plate with more than ten TCGA samples on it ($n = 351$ batches). **c**–**f**, Body-site attribution prediction on the likely contaminants removed data set (**c**), the plate–centre decontaminated data set (**d**), the all putative contaminants removed data set (**e**), and the most stringent filtering data set (**f**). **g**–**l**, All of the models and concomitant performance values (AUROC and AUPR) were re-generated using the four decontaminated data sets described above (each labelled with a different colour as shown above). The AUROC and AUPR values obtained from models trained and tested on the decontaminated data sets are plotted against the AUROC or AUPR values from the full data set (Fig. 1f–h). The dashed diagonal line denotes a perfect linear relationship. Generalized linear models have been fitted to the AUROC and AUPR values of the corresponding data sets; s.e. of the linear fits are shown by the associated shaded regions. COAD ($n = 1{,}006$ total samples; Supplementary Table 4) model performances are identified throughout the Figures.



Extended Data Fig. 7 | Decontamination effects on proportion of average reads per sample type. The total read count (DNA and RNA) of each major sample type (primary tumour (**a**), solid-tissue normal (**b**), blood-derived normal (**c**)) was summed and divided by the total number of samples within each sample type. This normalized read count (per sample type) was then divided by the summed normalized read count across all sample types for each cancer type, thereby providing an estimate of the proportion of average reads per sample type per cancer type. This was repeated for all five data sets, as shown by the legend, to assess whether decontamination differentially impacted certain types of sample and/or cancer; relative stability in the percentages shown would suggest a lack of differential contamination. Minor sample types that were not further analysed in this paper by decontamination or ML (for example, additional metastatic lesions; $n = 4$ sample types; Extended Data Fig. 1g) are not shown and comprised only 3.80% of total TCGA samples. Note, in the special case that only one sample type existed for a given cancer type (primary tumour in ACC, MESO, UCS), then all bars will show that 100% of the normalized reads came from that one sample type. The number of samples examined for each cancer type and sample type are shown in Supplementary Table 4.

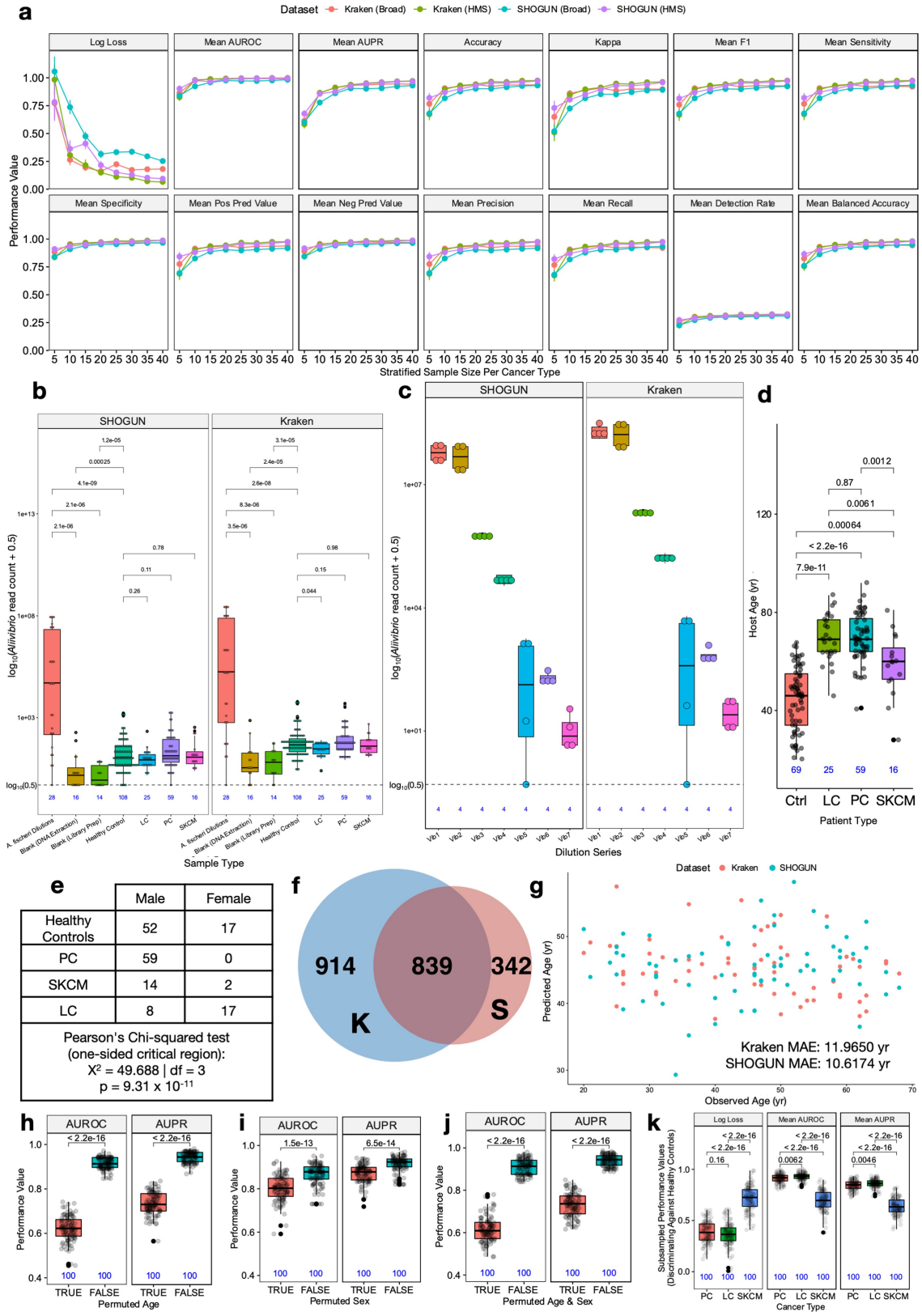


e

List of coding genes profiled for FoundationOne® Liquid ctDNA panel (n = 70 unique genes) for SNVs, Indels, Copy Number Alterations, and Rearrangements	**Entire coding sequence:** APC AR ATM BRCA1 BRCA2 CCND1 CD274 CDH1 CDK4 CDK6 CDK12 CDKN2A CHEK2 CRKL EGFR ERBB2 ERRFI1 FGFR1 FGFR2 FOXL2 KRAS MDM2 MET MYC MYCN NF1 PALB2 PDCD1LG2 PTEN PTPN11 RB1 SMO STK11 TP53 VEGFA **Select exons:** ABL1 AKT1 ALK ARAF BRAF BTK CTNNB1 DDR2 ESR1 EZH2 FGFR3 FLT3 GNA11 GNAQ GNAS HRAS IDH1 IDH2 JAK2 JAK3 KIT MAP2K1 MAP2K2 MPL MTOR MYD88 NPM1 NRAS PDGFRA PDGFRB PIK3CA RAF1 RET ROS1 TERT **Select Rearrangements:** ALK EGFR FGFR2 FGFR3 PDGFRA RET ROS
List of coding genes profiled for Guardant360® Assay ctDNA panel (n = 73 unique genes) for SNVs, Indels, Amplifications, and Fusions	**SNVs:** AKT1 ALK APC AR ARAF ARID1A ATM BRAF BRCA1 BRCA2 CCND1 CCND2 CCNE1 CDH1 CDK4 CDK6 CDKN2A CTNNB1 DDR2 EGFR ERBB2 ESR1 EZH2 FBXW7 FGFR1 FGFR2 FGFR3 GATA3 GNA11 GNAQ GNAS HNF1A HRAS IDH1 IDH2 JAK2 JAK3 KIT KRAS MAP2K1 MAP2K2 MAPK1 MAPK3 MET MLH1 MPL MTOR MYC NF1 NFE2L2 NOTCH1 NPM1 NRAS NTRK1 NTRK3 PDGFRA PIK3CA PTEN PTPN11 RAF1 RB1 RET RHEB RHOA RIT1 ROS1 SMAD4 SMO STK11 TERT TP53 TSC1 VHL **Indels:** ATM APC ARID1A BRCA1 BRCA2 CDH1 CDKN2A EGFR ERBB2 GATA3 KIT MET MLH1 MTOR NF1 PDGFRA PTEN RB1 SMAD4 STK11 TP53 TSC1 VHL **Amplifications:** AR BRAF CCND1 CCND2 CCNE1 CDK4 CDK6 EGFR ERBB2 FGFR1 FGFR2 KIT KRAS MET MYC PDGFRA PIK3CA RAF1 **Fusions:** ALK FGFR2 FGFR3 NTRK1 RET ROS1

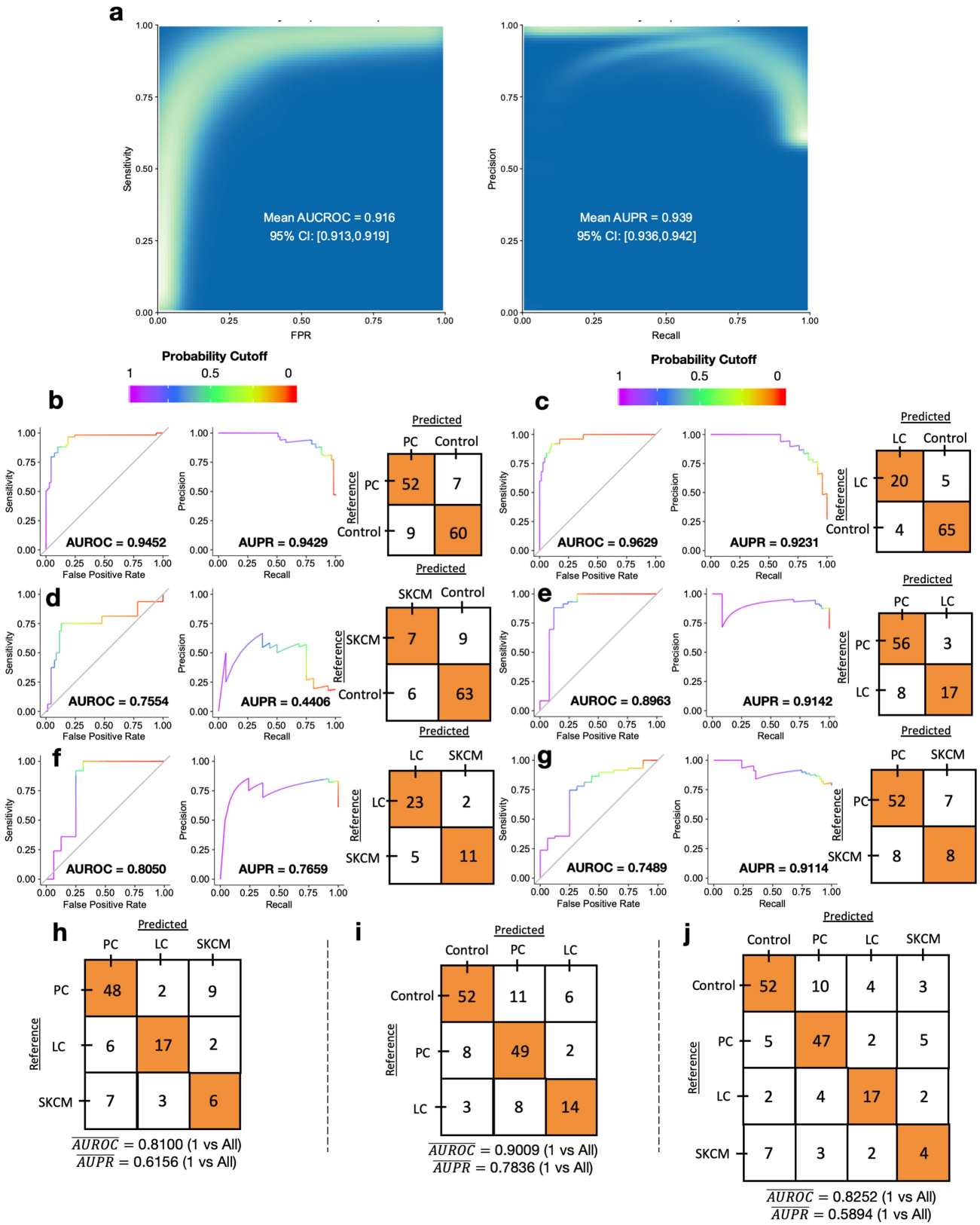
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Measuring spiked pseudo-contaminant contribution in downstream ML models and theoretical sensitivities of commercially available, host-based, ctDNA assays in patients from TCGA. a, **b**, Feature importance scores were calculated for all taxa used in models trained to discriminate one cancer type versus all others in all four decontaminated data sets (Extended Data Fig. 6b) using primary tumour microbial DNA or RNA (**a**), or using blood-derived mbDNA (**b**). These decontaminated data sets were spiked with pseudo-contaminants before the decontamination and normalization pipelines to evaluate their performance (see Methods), and the test set performances of the models shown are given in Extended Data Fig. 6g, h and Fig. 3a, respectively. Any spiked pseudo-contaminant(s) used by a model had their feature importance score(s) divided by the sum total of all feature importance scores in that model to estimate their percentage contribution towards making accurate predictions; the higher the score (out of 100), the less biologically reliable the model is. Note, zero means that no spiked pseudo-contaminants were used for making predictions by the model; none of the models generated on the plate–centre decontaminated data included spiked pseudo-contaminants as features. The number of samples included to evaluate performance of each comparison can be found in the data browser confusion matrices at http://cancermicrobiome.ucsd.edu/CancerMicrobiome_DataBrowser. **c**, **d**, Percentage distribution among TCGA studies of patients with one or more genomic alterations on FoundationOne Liquid ctDNA coding genes (**c**) or on Guardant360 ctDNA coding genes (**d**). The number of samples examined and raw data are available at https://www.cbioportal.org/. **e**, The specific list of coding genes for the FoundationOne and Guardant360 ctDNA assays and their examined alterations (source listed in the Methods).



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Supporting analysis for real-world, plasma-derived, cell-free microbial DNA analysis between and among healthy individuals and multiple types of cancer. a, Discriminatory simulations in TCGA used to empirically power the real-world validation study (Fig. 4; see Methods). Centre values for each stratified sample size are the means of the performances across ten iterations; error bars denote s.e.m. **b**, Evaluation of *Aliivibrio* genus abundance values (raw read counts) among positive control bacterial (*Aliivibrio*) monocultures, negative control blanks, and human sample types using Kraken and SHOGUN-derived data. **c**, *Aliivibrio* genus abundance (raw read counts) across bacterial monoculture dilutions. **d**, Age distribution among cancer-free healthy control individuals (Ctrl) and grouped patients with lung cancer (LC), prostate cancer (PC), or melanoma (SKCM). **e**, Gender distribution among patients with inset Pearson's χ^2 test (one-sided critical region). **f**, Venn diagram of taxon assignments between Kraken and SHOGUN, which used different databases. **g**, Iterative LOO ML regression of host age using Kraken (pink) or SHOGUN (aqua) raw microbial count data in healthy cancer-free individuals. Mean absolute errors (MAE) evaluated across all samples are shown. **h**–**j**, The effects of permuted age (**h**), sex (**i**), and age and sex (**j**) before Voom-SNM on ML performance to discriminate healthy individuals versus grouped patients with cancer using cell-free microbial DNA. One hundred permutations were used for each comparison (see Methods). **k**, Iterative subsampling of PC, LC, SKCM, and control groups to match SKCM cohort size ($n = 16$ samples), followed by LOO pairwise ML of each subsampled cancer type against subsampled healthy controls. One hundred permuted iterations were used to estimate discriminatory performance distributions and standard errors (see Methods). **b**, **c**, Note the $\log_{10}$ scale and 0.5 pseudo-count lower limit (dotted line). **b**–**d**, **h**–**k**, All hypothesis tests are two-sided Mann–Whitney *U*-tests with multiple testing correction when testing more than two comparisons; box plots show median (line), 25th and 75th percentiles (box), and 1.5 × IQR (whiskers). For all box plots and bar plots, sample sizes are shown in blue below.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | SHOGUN-derived ML performances to discriminate between types of cancer and healthy, cancer-free individuals using cell-free microbial DNA. a, Bootstrapped performance estimates for distinguishing grouped patients with cancer (n = 100) from cancer-free healthy control individuals (n = 69). ROC and PR curve data from 500 iterations with different training–testing splits (70%–30%) are shown on the rasterized density plot; mean values and 95% CI estimates are shown. **b**–**g**, LOO iterative ML performance between two classes: PC versus control (**b**), LC versus control (**c**), SKCM versus control (**d**), PC versus LC (**e**), LC versus SKCM (**f**), and PC versus SKCM (**g**). **h**–**j**, Multi-class (n = 3 or 4), LOO iterative ML performances to distinguish between types of cancer, as well as between patients with cancer and healthy cancer-free control individuals. Mean AUROC and AUPR, as calculated from one-versus-all-others AUROC and AUPR values, are shown below confusion matrices. **h**, LOO ML performance between the three types of cancer under study. **i**, LOO ML performance between the three sample types with at least 20 samples in the minority class (that is, the cutoff used in the TCGA analysis, Fig. 1f–h). **j**, LOO ML performance between all four sample types under study. For all subfigures with confusion matrix plots: LOO ML was used instead of single or bootstrapped training–testing splits because of small sample sizes; these confusion matrices also reflect the number of samples used for each comparison.

nature research

Corresponding author(s): Gregory D Poore, Rob Knight

Last updated by author(s): Dec 11, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

All TCGA sequence data were accessed via the Cancer Genomics Cloud (CGC) as sponsored by SevenBridges (https://cgc.sbgenomics.com). Matched patient metadata, including molecular subtypes (e.g. BRCA Pam50 subtypes), were accessed via the CGC through both SevenBridges CGC and the Institute for Systems Biology CGC (ISB; https://isb-cgc.appspot.com/), via the TCGAMutations R package (version 0.2.0), or were taken directly from their respective TCGA publication's supplemental data. Genomic alteration statuses for all TCGA patients were queried and downloaded via the cBioPortal (https://www.cbioportal.org/). Gene panels for commercial ctDNA assays were accessed from company white papers for the Guardant360® assay (https://www.therapyselect.de/sites/default/files/downloads/guardant360/guardant360_specification-sheet_en.pdf) and the FoundationOne® Liquid assay (https://assets.ctfassets.net/vhribv12lmne/3SPYAcbGdqAeMsOqMyKUog/d0eb51659e08d733bf39971e85ed940d/F1L_TechnicalInformation_MKT-0061-04.pdf).

For TCGA metadata accession and transformation from hierarchical formats to flat tables, custom Python scripts (Python 3) were written to query SevenBridges's metadata SPARQL ontology (https://opensparql.sbgenomics.com/#/console) and organize the data where possible; these scripts are in our Github linked below and the summarized metadata are also provided as supplemental data. For information not stored in that ontology, we used the ISB's CGC R programming language API (bigrquery R package version 1.0.0) to access its recent metadata release (tcga_201607_beta.Clinical_data).

For the Kraken-based microbial detection pipeline, a total of 71,782 microbial genomes were downloaded using RepoPhlan (https://bitbucket.org/nsegata/repophlan; Python 2) on 14 June 2016, of which 5,503 were viral and 66,279 were bacterial or archaeal. For the cancer microbiome pipeline, we create customized, shareable'app' workflows on the CGC that included bioinformatic tools hosted by them (e.g. samtools, BWA; versions are maintained by the CGC) or self-uploaded to the CGC and run as separate Docker containers (QIIME [version 1.9.1], Kraken [version 0.10.5-beta]). These 'app' workflows inputted raw TCGA BAM files hosted by the CGC and eventually outputted Kraken-derived microbial outputs at the genus level. Analyses done off the CGC cloud (i.e. on-site) to estimate Kraken's false positive rate using genome alignments used the following tools: BWA (version 0.7.16a-r1181) and Bowtie2 (2.2.9).

For the SHOGUN-based microbial detection pipeline, the recently published 'Web of Life' database (WoL; PMID: 31792218; https://biocore.github.io/wol/data/genomes/#) was used along with the SHOGUN taxonomy assignment algorithm (v 1.0.6; PMID: 30443602). TCGA sequencing reads that were marked as Kraken-positive for being microbial were reprocessed using the SHOGUN algorithm and

WoL database. Of note, this was much faster than reprocessing all 'non-human' reads in TCGA, as we originally did for Kraken, as some cancer types (e.g. TCGA-OV) had as much as 19.5% of their total sequencing content marked as 'non-human' by not aligning to human reference genomes. As done with Kraken, taxonomy assignments for SHOGUN-derived data were summarized at the genus level.

For the prospective validation cohort, we collected 169 frozen plasma samples that were previously collected (biobanked) under the following IRB protocol numbers (all at UC San Diego): 131550, 150348, 130296, 091054, 172092, 151057, and 182064. This included plasma samples from 69 healthy, HIV-, non-cancer individuals and 100 plasma samples from patients of three types of cancer (prostate cancer, lung cancer, and melanoma). From these plasma samples, cell-free DNA was extracted in a very controlled fashion with 58 controls, comprising 30 independent "negative blank" controls and 28 "positive" bacteria monocultures (Vibrio fishceri), further comprising of internal replicates and dilutions. After extraction, all samples were sequenced using paired-end 2×150 bp sequencing in a single run on a NovaSeq 6000 instrument (Illumina) while pooling across all four lanes during sequencing. Sequencing data were then processed bioinformatically and fed into downstream normalization and machine learning pipelines.

Data analysis

All analyses were performed using R (version 3.4.3). Supervised normalization was performed using the following packages: limma (for the voom algorithm; version 3.34.9), edgeR (version 3.20.9), and snm (version 1.26.0). Machine learning models were trained and tested using the R packages gbm (version 2.4.1), caret (version 6.0-80), and doMC (for parallel computing; version 1.3.5). AUC and PR curves and their concomitant areas under the curve were calculated using the PRROC package (version 1.3.1). Additional packages used for data formatting/manipulation included the following: purrr (version 0.3.2), dplyr (version 0.8.0.1), and tibble (version 2.1.1). Decontamination was primarily performed using the decontam package (version 1.1.2). Packages used for plotting and associated statistical analyses included: ggpubr (version 0.1.8.999), ggplot2 (version 3.1.1.9000), ggsci (version 2.9), ggrepel (version 0.8.1), and cowplot (version 0.9.3). BIOM tables were inputted and outputted using the biomformat package (version 1.5.0). For the Bayesian source tracking analysis, the updated SourceTracker2 algorithm (https://github.com/biota/sourcetracker2) was employed with default settings.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Pre-processed Kraken-derived and SHOGUN-derived cancer microbiome data generated and analyzed in this study (i.e. summarized read counts at the genus taxonomic level) as well as the metadata are available at https://drive.google.com/drive/folders/18V2ON-Go5AeEtZLe1f9EeJToWOhg81ab?usp=sharing. Raw outputs of Kraken-processed and SHOGUN-processed TCGA sequencing data comprise hundreds of terabytes of files and are not directly available unless otherwise coordinated with the corresponding author. However, all raw TCGA data and the bioinformatics pipeline necessary to generate such raw outputs from Kraken can be accessed through SevenBridge's CGC.

Normalized microbial abundances for all of TCGA and the prospective validation study, and machine learning model performances (ROC and PR curves, confusion matrices), as well as ranked taxonomy features for every model, are available on our interactive TCGA Cancer Microbiome website: https://gregpoore.shinyapps.io/KL_RShiny_App/. We also provide code to generate the results shown on the website analyses below in our GitHub repository.

All programming scripts used to access, manage, and run data on the CGC as well as development of the supervised normalization, decontamination, machine learning pipelines, and so forth can be found at our GitHub repository link: https://github.com/biocore/tcga. These can be applied directly on the summarized, genus-level count data given above. Our CGC pipeline is also publicly shareable and available upon reasonable request to the corresponding author.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

We re-examined The Cancer Genome Atlas (TCGA)'s entire compendium of whole genome sequencing (WGS, n=4,831) and whole transcriptome sequencing (RNA-Seq; n=13,285) studies for viral, bacterial, and archaeal reads using the Kraken algorithm, covering 18,116 samples across 10,481 patients and 33 cancer types. We then re-processed 13,517 of these samples through the orthogonal SHOGUN microbial taxonomy assignment algorithm and 'Web of Life' database, which was also different and much smaller than the database used for Kraken (~1/6th the size and did not include viruses). This SHOGUN subset of 13,517 samples included samples from every TCGA cancer type (n=32), sample type (n=7), sequencing platform (n=6), and sequencing center (n=8) in the Kraken-based analysis.

For the prospective validation study, 169 biobanked, frozen plasma samples were processed and deeply sequenced for cell-free microbial DNA. This included plasma samples from 69 healthy, HIV-, non-cancer individuals and from 100 cancer patients across three cancer types (prostate cancer: n=59; lung cancer: n=25; melanoma: n=16). The cancer sample sizes were empirically powered using two independent simulations on blood-derived normal samples in TCGA that came from The Broad Institute or Harvard Medical School on the same cancer types as those in the prospective validation study; these simulations were done with both Kraken- and SHOGUN-derived data, for a total of

four independent simulations (Extended Data Fig. 8a). These simulations collectively suggested that at least 15 samples per cancer type was required in order to obtain good discriminatory performance between cancer types. It was not possible to power the study for healthy, non-cancer subject since TCGA did not include healthy controls as part of their cohort.

Data exclusions

For downstream processing, 491 samples were removed due to low quality metadata, which left 17,625 samples across 10,183 patients and 32 cancer types (acute myeloid leukemia was removed) (Fig. 1b). All of these 17,625 samples were then used for subsequent supervised normalization and machine learning pipelines.

For machine learning model building, we did not evaluate comparisons in which the minority class contained less than 20 samples for both training and testing. Analyses comparing the model performances and the minority class size of one-versus-all-others machine learning models showed statistically significant linear relationships (Extended Data Figs. 2g-h) between them; the 20 sample minimum was empirically determined based on the ability of the minority class to capture enough variation in the cancer microbiome to make discrimination between classes feasible. Lastly, as part of our efforts to identify and remove decontamination, as well as to evaluate its effect on model performance, we discarded as little as 9.1% and as much as 92.3% of the microbial data in four different decontaminated datasets of varying filtering stringency (Fig. 4b).

Replication

Prospective validation study:

Analyzing the TCGA cancer microbiome suggested that blood-derived microbial DNA should be able to discriminate between cancer types (Fig. 5). These blood microbial 'signatures' were also the most contentious result of the study. To test this hypothesis, we prospectively collected, carefully processed (with 58 independent negative and positive controls), and sequenced plasma samples from 169 individuals for cell-free microbial DNA, 100 of which had a known cancer diagnosis (across three cancer types) and 69 of which were known to be HIV-seronegative and not have cancer. Notably, we powered the cancer sample sizes using simulations on blood-derived normal samples in TCGA from two different sequencing centers on the same cancer types. Then, using the same tools utilized in the TCGA analysis, plus additional ones to employ gold-standard microbiology practices, we showed that it is possible to discriminate between and among healthy, non-cancer controls and cancer types, either in pairwise or simultaneous multi-class fashion, solely using plasma-derived, cell-free microbial DNA (Fig. 6; Extended Data Fig. 9). Two out of three cancer types (prostate and lung cancer) showed strong discrimination against healthy controls; these cancer types also were strongly discriminable between each other. The cancer type that performed worse (melanoma) for both healthy vs. cancer comparisons and between cancer comparisons also performed worse in our original TCGA analysis for four out of the five tested datasets (Fig. 5a). Since melanoma represented the smallest cohort of the cancer types (n=16), we iteratively subsampled the other two cancer types to the same cohort size as melanoma and showed that they consistently performed better in a statistically significant manner when discriminating between them and subsampled healthy controls than when melanoma samples were iteratively compared against subsampled healthy controls (all comparisons done in a leave-one-out fashion for machine learning). Moreover, the two subsampled cancer types still demonstrated strong discriminatory performance at the same cohort size as the melanoma group. In other words, using results found in TCGA, which studied completely different patients, had different DNA/RNA extraction protocols, and utilized different sample types (i.e. whole blood vs. plasma), we show that discriminatory performance between cancer types appears to be conserved.

In addition to the prospective validation study, many levels of replication were internally tested within TCGA that were also successful:

1. Orthogonally validating discriminatory machine learning performances between and within all cancer types using an entirely different microbial taxonomy assignment algorithm (SHOGUN) and database (Web of Life) (Extended Data Figs. 4m-t). This also replicated the same batch effects we observed in the Kraken-derived dataset (Extended Data Figs. 4k-l), which were correct using the same Voom-SNM normalization prior to downstream machine learning.
2. Validating discriminatory machine learning performances between and within four cancer types using direct genome alignment-filtered (BWA-filtered) Kraken data (Extended Data Figs. 4a-h).
3. Splitting TCGA raw cancer microbiome data into two independent halves --> normalizing both halves separately --> building models on each half --> evaluating the performance of each model on the other half. These model performances were then compared to a model trained and tested on the full dataset with 50%-50% training and testing splits (Extended Data Fig. 3a).
4. Demonstrating that machine learning models had the same (or even improved) performance when restricting samples to a single analyte type (DNA or RNA) (Extended Data Figs. 3b-e).
5. Demonstrating that machine learning models had the same (or improved) performance when restricting samples from a given sequencing center. Additionally, we selected the two largest sequencing centers that only sequenced one kind of analyte (DNA or RNA) to further reduce the likelihood of model performance being explained by technical variation (Extended Data Figs. 3f-i).
6. Systematic decontamination of the TCGA cancer microbiome dataset produced four additional datasets of varying filtering stringency. Three of these datasets decontaminated by sequencing centers as batches (n=8), such that all microbes identified as contaminants in any one sequencing center were discarded. As a more conservative measure, the other (fourth) dataset used the highest batch resolution available in TCGA by examining which samples were on the same sequencing plate at the same sequencing center; we decontaminated in a manner such that all microbes identified as a contaminant in any one plate-center batch (n=351) were removed. All machine learning models were then replicated using these four new, decontaminated datasets and showed consistent (or improved) model performance (Figs. 4g-l).
7. We spiked pseudo-contaminants into the raw dataset to track them through decontamination, normalization, and machine learning model building. If present within a trained model, we used feature importance scores to evaluate the contribution of the pseudo-contaminant(s) to the predictive capacity of the model. This effectively labels which models are and are not trustworthy, and the degree to which they are untrustworthy. However, the number of models showing significant reliance on pseudo-contaminants was small and included none of the models trained/tested on plate-center decontaminated data (Extended Data Figs. 6a-b).
8. We additionally replicated results found and published by other groups and other bioinformatic pipelines; clinical metadata; and demonstrate ecologically sensible findings using the cancer microbiome data (Fig. 3).

Randomization

Samples were randomly selected for 70% training and hold-out 30% testing splits. A random number seed was used across all machine learning models to ensure that the models' results would be comparable with exceptions for certain permutation analyses, as described in the Methods. No randomization was used for the prospective validation study, which sequenced biobanked samples in a single sequencing run.

Blinding

Blinding is not applicable in the TCGA dataset, as all samples that had available whole genome or whole transcriptome data were processed. For the prospective study, samples were each given a code to identify their group for randomization during sample processing to avoid plate or well-specific biases.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Demographic information is presented in Fig. 6a of the paper. The relevant population characteristic differences included age (range: 20-92 years; median=62 years; mean=58.54 years) and gender (n=133 males, n=36 females). Other population characteristics that were identified in our study but are not currently known to affect the blood microbiome are as follows: All cancer patients had high-grade (stage III-IV) diagnosed cancers at time of sample donation; roughly half of prostate cancer patients (27/59) had castration-resistant disease at time of sample collection; lung cancer and melanoma patients were being trialed on immunotherapy but donated prior to immunotherapy, although they may have received other medications for gold-standard cancer care prior to enrolling in the study; all analyzed healthy controls were HIV- seronegative at time of sample donation.
Recruitment	Biobanked samples came from patients previously collected under two cancer collection studies at UC San Diego (IRB #131550 for prostate cancer studies; IRB #150348 for lung cancer and melanoma samples). IRB #131550 involved collecting samples, including plasma, from prostate cancer patients to determine molecular markers of androgen-deprivation therapy resistance (i.e. castration-resistant prostate cancer, CRPC). IRB #150348 involved collecting samples, including plasma, from lung cancer and melanoma patients in order to determine molecular markers of immunotherapy response. Patients self-selected to join these studies and were consented during the course of their clinical care at Moores Cancer Center (UC San Diego Health). There are no known biases that contributed towards their study inclusion other than they had cancer and were seeking clinical care at UC San Diego Health. Healthy, HIV-, non-cancer individuals were recruited under the following IRB protocol #s at UC San Diego: 130296, 091054, 172092, 151057, and 182064. These studies took place under the UCSD HIV Neurobehavioral Research Center (HNRC) with the goal of elucidating mechanisms behind the development and persistence of HIV-associated neurocognitive disorders (HAND). Participants self-selected, were enrolled through the HNRC ,and were longitudinally studied. Participants may be biased by the purpose of the study, thereby having a higher rate of neurocognitive disorders than the general population. However, there is no known literature linking such disorders with changes in the blood microbiome; therefore, they were used as non-cancer controls.
Ethics oversight	All biobanked samples used for the prospective validation came from studies were reviewed and approved by UC San Diego IRB.

Note that full information on the approval of the study protocol must also be provided in the manuscript.