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.