Analyzing the tumor microbiome to predict cancer patient survival and drug response

Leandro C. Hermida[1,2†], E. Michael Gertz[1†], Eytan Ruppin[1*].

[1] Cancer Data Science Laboratory (CDSL), National Cancer Institute (NCI), National Institutes of Health (NIH), Bethesda, MD, USA.

[2] Department of Computer Science, University of Maryland, College Park, MD, USA.

[†] Equally contributing first authors

[*] Corresponding author (eytan.ruppin@nih.gov)

Abstract

Poore et al.[1] recently published a computational approach for deriving microbial abundances from human tumor whole genome (WGS) and transcriptome sequencing (RNA-seq) data by leveraging tools commonly used to remove microbial contamination from such data. They have shown that microbial abundances could be used to predict various tumor-related phenotypes across The Cancer Genome Atlas (TCGA) cohort, including distinguishing tumor from adjacent normal tissue samples, cancer type, and tumor stage. Here, we investigated whether the microbial abundances inferred by Poore et al. in the TCGA cohort, to the best of our knowledge the most comprehensive dataset of its kind, are predictive of patient survival and drug response, two fundamentally important and clinically relevant phenotypes. We find that in four cancer types, adrenocortical carcinoma, cervical squamous cell carcinoma, brain lower grade glioma, and subcutaneous skin melanoma, microbial features are better predictors of survival than clinical covariates alone. In addition, we find seven cancer-drug pairs where microbiome features are more predictive of patients' response than clinical covariates alone. These seven pairs include chemotherapy treatments for bladder urothelial carcinoma, docetaxel treatment for breast invasive carcinoma and sarcoma, and several treatments for stomach adenocarcinoma.

Main

Our approach utilized the normalized, batch effect corrected, and decontaminated microbial abundance data generated by Poore et al.[1], which was derived from TCGA WGS and RNA-seq sequencing of 32 cancer types. We analyzed primary tumors from these data and applied additional filters to further reduce technical variation present in legacy TCGA (see Supplemental Methods). We then built survival and drug response machine learning (ML) models using these data, while including and adjusting for clinical covariates. We identified the predictive subsets of microbial genera from the ML model feature importance results. For comparison, we also built corresponding models using TCGA gene expression data, as well as models combining microbial abundance and gene expression data. We evaluated the predictive performance of our models using standard metrics, employing Harrell's concordance index (C-index) for measuring the accuracy of patient survival prediction and the area under the receiver operating characteristic (AUROC) for measuring the accuracy of drug response prediction.

We first examined whether tumor microbial abundances could predict patients' overall survival (OS) and progression-free interval (PFI) better than the classical clinical prognostic covariates: age at diagnosis, gender, and tumor stage. We built microbiome survival models using Coxnet – regularized Cox regression with elastic net penalties[2], which allowed us to include and control for these clinical prognostic covariates while jointly selecting the most predictive subset of microbial features via cross-validation (CV). For comparison, we also built standard Cox regression models based on the clinical covariates alone. Each analysis generated 100 model instances and C-index scores from different randomly shuffled train and test CV splits on the data (**Fig. 1a**). We found five microbiome models that had a mean C-index score ≥ 0.6 and significantly outperformed their corresponding clinical covariate-only models with an FDR-adjusted, two-sided Wilcoxon signed-rank test p-value ≤ 0.01 when comparing their scores (**Fig. 1b**). Going into a higher resolution, we also examined how well the models predicted along the disease progression axis by calculating time-dependent, cumulative/dynamic AUCs[3,4]. We found that in adrenocortical carcinoma (ACC), microbial features predicted OS significantly better than clinical covariates starting at approximately 6 years after diagnosis and in cervical squamous cell carcinoma (CESC), the microbial features predicted OS better than clinical covariates from approximately 6 months to 10 years after diagnosis (**Fig. 1c**).

We performed analogous survival modeling and analysis using TCGA gene expression data (see Supplemental Methods). We found that gene expression is predictive of survival in the same 13 cancer types that were recently reported by Milanez-Almeida et al.[5], as well as in four additional cancer types: CESC, thymoma (THYM), pheochromocytoma and paraganglioma (PCPG), and uterine corpus endometrial carcinoma (UCEC). Overall, we find that gene expression is a more powerful predictor of survival, across a wider range of cancer types, than microbial abundances (**Fig. 1d**, Extended Data Figs. 1-2). We also evaluated whether models built from combined microbial and gene expression data could enhance survival predictive power, but only found a modest improvement in one cancer type, THYM OS (Extended Data Fig. 3a).

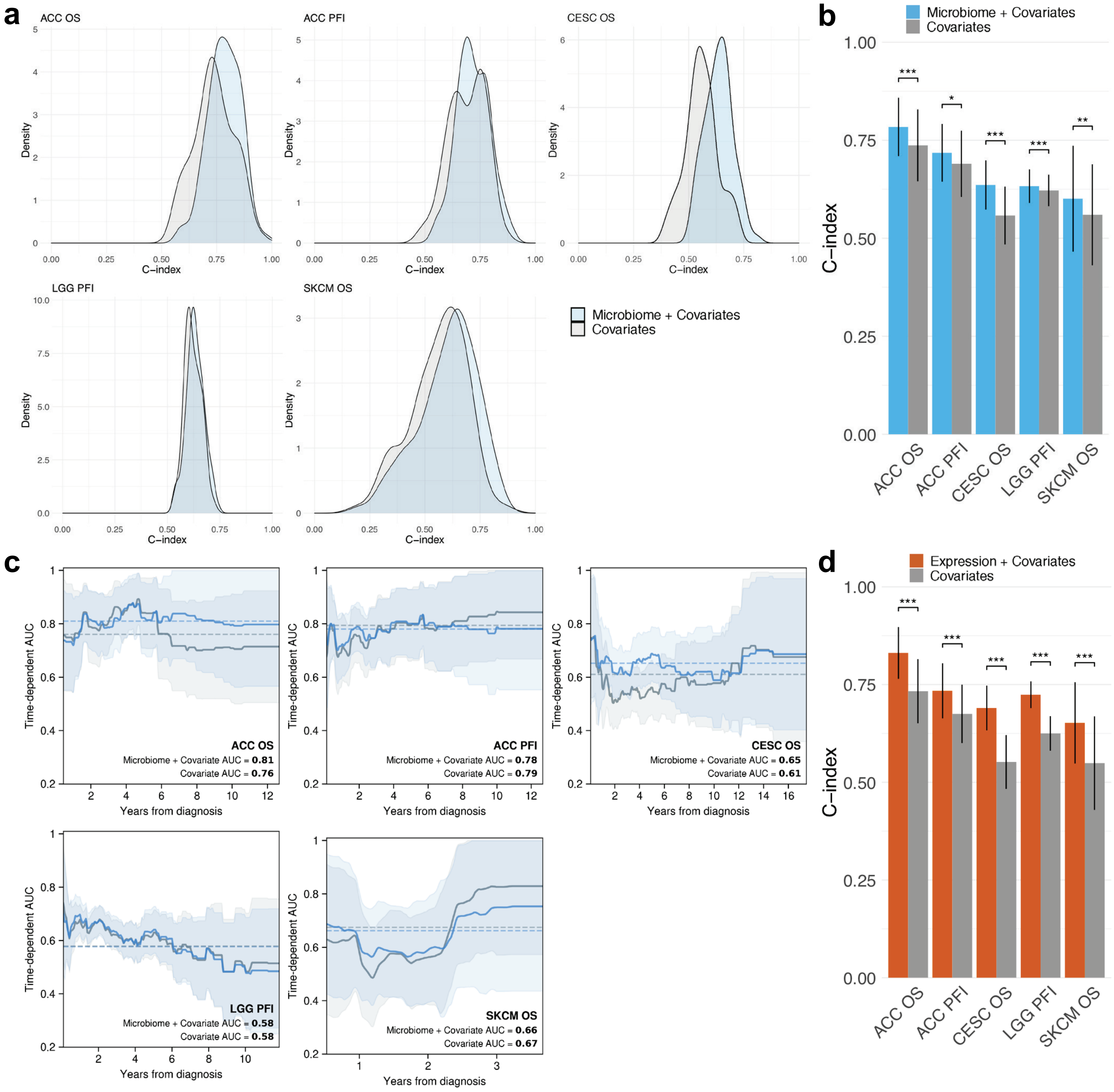


Figure 1. Performance of combined microbiome and clinical covariate survival models in the five cancer types where microbial features add predictive power. (**a**) C-index score density distributions for combined microbiome and clinical covariate Coxnet models (blue) and corresponding clinical covariate-only Cox models (grey). (**b**) Mean C-index score comparison between combined microbiome and clinical covariate models (blue bars) and corresponding clinical covariate-only models (grey bars). (**c**) Time-dependent cumulative/dynamic AUCs for combined microbiome and clinical covariate models (blue) and clinical covariate-only models (grey) following years after diagnosis. Shaded areas denote standard deviations. (**d**) Mean C-index score comparison between combined gene expression and clinical covariate models (orange bars) and corresponding clinical covariate-only models (grey bars). Error bars in (**b**) and (**d**) denote standard deviations. Significance test for comparing combined microbiome (or gene expression) and clinical covariate model scores to clinical covariate-only model scores was calculated using a two-sided Wilcoxon signed-rank test and adjusted for multiple testing using the Benjamini-Hochberg method. Significance: * ≤ 0.01, ** ≤ 0.001, *** ≤ 0.0001.

We next asked whether tumor microbial abundances could predict drug response better than clinical covariates using a similar framework as our survival analysis. All TCGA samples with drug response phenotypic data were from pre-treatment biopsies. Due to the limited cancer-drug combination cohort sizes in TCGA, we analyzed each drug individually, even if a patient received multiple drugs concurrently. We only considered the first treatment response if a patient received the same drug at different times. In total, we analyzed 30 cancer-drug combinations in the TCGA cohort that met our minimum dataset size thresholds (see Supplemental Methods). We built drug response models using a variant of the linear support vector machine recursive feature elimination (SVM-RFE) algorithm[6] that we developed, where we could include clinical covariates while jointly selecting the most predictive subset of microbial features via cross-validation. For comparison, we also created linear SVM models using the clinical covariates alone. Each analysis generated 100 model instances, AUROC, and area under the precision-recall curve (AUPRC) scores from different randomly shuffled train and test CV splits on the data (**Fig. 2a-c**). We found seven cancer-drug microbiome model types that had a mean AUROC score ≥ 0.6 and significantly outperformed corresponding covariate-only models with an FDR-adjusted, two-sided Wilcoxon signed-rank test p-value ≤ 0.01 when comparing their scores (**Fig. 2d**). Of particular interest, three of these cancer-drug combinations involve stomach adenocarcinoma (STAD).

We performed analogous drug response modeling and analysis using TCGA gene expression data as was done for survival. We found that only five cancer-drug gene expression models performed better than clinical covariates alone. Two of these, for urothelial bladder cancer (BLCA) cisplatin and gemcitabine treatments, overlapped with the microbiome model results (**Fig. 2e**, Extended Data Fig. 4). We also evaluated whether models built from combined microbial and gene expression data could enhance predictive power and found a modest improvement in four cancer-drug combinations (Extended Data Fig. 3). Overall, in contrast to survival, the tumor microbiome represents a promising new predictive tool for drug response, consistent with recent reports interrogating its potential role[7,8].

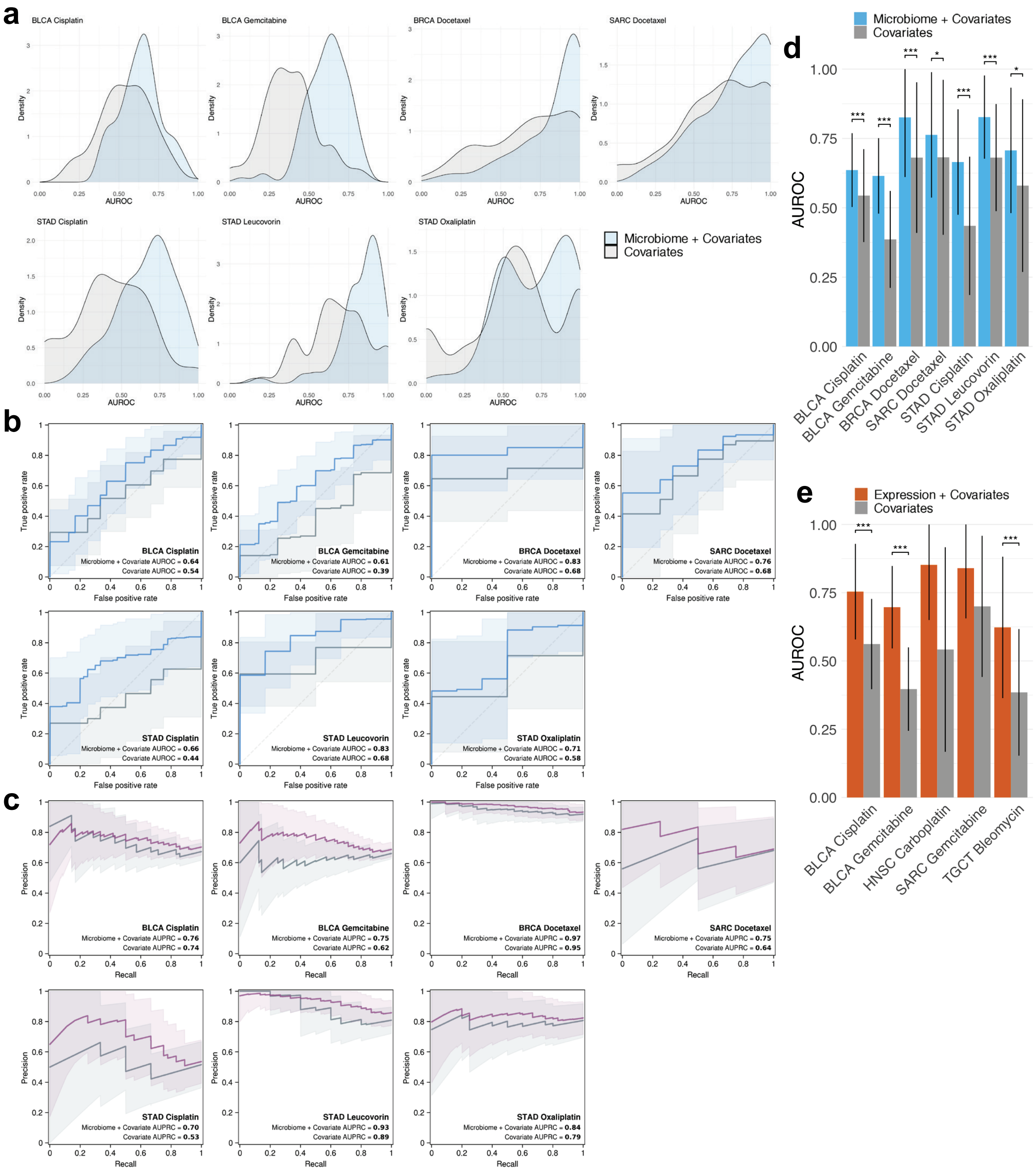


Figure 2. Performance of combined microbiome and clinical covariate drug response models in the seven cancer types where the microbiome adds predictive power. (**a**) Area under the receiver operator characteristic (AUROC) density distributions for combined microbiome and clinical covariate SVM-RFE models (blue) and corresponding clinical covariate-only SVM models (grey). (**b**) ROC curves for combined microbiome and clinical covariate models (blue) and clinical covariate-only models (grey). (**c**) Precision-recall (PR) curves for combined microbiome and clinical covariate models (purple) and clinical covariate-only models (grey). Shaded areas in (**c**) and (**d**) denote standard deviations. (**d**) Mean AUROC comparison between combined microbiome and clinical covariate models (blue bars) and corresponding clinical covariate-only models (grey bars). (**e**) Mean AUROC comparison between combined gene expression and clinical covariate models (orange bars) and corresponding clinical covariate-only models (grey bars). Error bars in (**d**) and (**e**) denote standard deviations. Significance tests for comparing combined microbiome (or gene expression) and clinical covariate model scores to clinical covariate-only model scores were calculated using a two-sided Wilcoxon signed-rank test and adjusted for multiple testing using the Benjamini-Hochberg method. Significance: * ≤ 0.01, ** ≤ 0.001, *** ≤ 0.0001.

We ranked the top microbial genera selected by each microbiome model that performed better than clinical covariates alone according to their selection frequency and weight coefficients across the 100 model instances generated for each model (see Supplemental Methods). Overall, we identified 438 distinct microbial genera that were highly ranked features in at least one drug response or survival model (Extended Data Table 1). The median number of genera selected per model was 50, with a minimum of 3 (BRCA docetaxel) and a maximum of 75 (ACC PFI). Of the 438 genera seen, only 94 were selected in more than one model and only 20 were selected in more than two models. This is consistent with the observation of Nejman et al.[9] that the tumor microbiome is tumor type specific.

All non-eukaryotic domains of life were represented in the selected features, in total encompassing 374 bacterial, 22 archaeal, and 42 viral genera (Extended Data Table 2). At the phylum level, Proteobacteria and Firmicutes were the most frequently selected features, followed by Actinobacteria and Bacteroidetes. Among viruses, Herpesvirales were the most frequently selected. There was a significant trend for selected microbial genera to be negatively predictive of drug response or survival (two-sided binomial test p-value = .000545, with 331 genera being negatively associated and 247 positively associated). Herpesvirales and the most frequently selected bacterial phyla also exhibited an overall negative-over-positive predictive trend, except for the phylum Firmicutes. Among Firmicutes, there were more genera which were positively predictive of drug response or survival, which was marginally significant with a two-sided Fisher's exact test p-value of 0.007. The cancers for which Firmicutes is associated with drug response or survival are shown in Extended Data Table 2. Notably, CESC is one such cancer. Even though CESC is known to often arise from HPV infection, the presence of other microbial species, in particular the Firmicutes species Lactobacillus, are positively associated with the risk of developing CESC[10].

Among cancer types, the involvement of the microbiome in breast cancer (BRCA)[9,11] and bladder cancer (BLCA)[12,13] has received recent attention. In BRCA, we found the genus containing Epstein-Barr virus (EBV) was negatively associated with response to docetaxel, which is similar to previous findings that EBV is associated with chemoresistance to docetaxel in gastric cancer[14]. STAD stood out among the drug response model results, with the microbiome being predictive of response to three different drugs: cisplatin, leucovorin, and oxaliplatin. Patients infected with *H.*

pylori have an increased risk of developing STAD[15]. However, *H. pylori* was not a predictive feature of drug response in STAD patients in our models. We found Cedecea and Sphingobacterium abundances were both strongly negatively predictive of leucovorin response in STAD. Both genera have been implicated in bacteremia in immunocompromised individuals in rare cases, including cancer[16,17,18,19].

In summary, we find that the microbial abundances generated by Poore et al. are predictive of patient survival and response to chemotherapy in some cancer types and treatments. Overall, their predictive capacity is quite modest. The tumor microbiome is considerably less predictive than the tumor transcriptome in predicting patient survival, but notably, better in predicting chemotherapy response. Our investigation lays the basis for future research, studying the role of the tumor microbiome (based on abundances derived from sequencing data) in predicting the response to targeted and immunotherapies.

References

1. Poore, G. D. *et al.* Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579**, 567–574 (2020).

2. Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software, Articles* **39** (5), 1-13 (2011).

3. Hung, H. & Chiang, C.T. Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics*, **38** (1), 8–26 (2010).

4. Lambert, J. & Chevret, S. Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves. *Statistical methods in medical research*, **25** (5), 2088–2102 (2016).

5. Milanez-Almeida, P., Martins, A. J., Germain, R. N. & Tsang, J. S. Cancer prognosis with shallow tumor RNA sequencing. *Nature Medicine* **26**, 188–192 (2020).

6. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* **46**, 389–422 (2002).

7. Geller, L. T. *et al.* Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. *Science* **357**, 1156–1160 (2017).

8. Pushalkar, S. *et al.* The Pancreatic Cancer Microbiome Promotes Oncogenesis by Induction of Innate and Adaptive Immune Suppression. *Cancer Discov* **8**, 403–416 (2018).

9. Nejman, D. *et al.* The human tumor microbiome is composed of tumor type–specific intracellular bacteria. *Science* **368**, 973–980 (2020).

10. Lin, D. *et al.* Microbiome factors in HPV-driven carcinogenesis and cancers. *PLoS Pathog* **16**, (2020).

11. Eslami-S, Z., Majidzadeh-A, K., Halvaei, S., Babapirali, F. & Esmaeili, R. Microbiome and Breast Cancer: New Role for an Ancient Population. *Front Oncol* **10**, (2020).

12. Bajic, P., Wolfe, A. J. & Gupta, G. N. The Urinary Microbiome: Implications in Bladder Cancer Pathogenesis and Therapeutics. *Urology* **126**, 10–15 (2019).

13. Bučević Popović, V. *et al.* The urinary microbiome associated with bladder cancer. *Scientific Reports* **8**, 12157 (2018).

14. Shin, H. J., Kim, D. N. & Lee, S. K. Association between Epstein-Barr virus infection and chemoresistance to docetaxel in gastric carcinoma. *Mol. Cells* **32**, 173–179 (2011).

15. Parsonnet, J. et al. Helicobacter pylori infection and the risk of gastric carcinoma. *N. Engl. J. Med*. **325**, 1127–1131 (1991).

16. Abate, G., Qureshi, S. & Mazumder, S. A. Cedecea davisae bacteremia in a neutropenic patient with acute myeloid leukemia. *J. Infect.* **63**, 83–85 (2011).

17. Akinosoglou, K. *et al.* Bacteraemia due to Cedecea davisae in a patient with sigmoid colon cancer: a case report and brief review of the literature. *Diagn. Microbiol. Infect. Dis.* **74**, 303–306 (2012).

18. Koh, Y. R. *et al.* The first Korean case of Sphingobacterium spiritivorum bacteremia in a patient with acute myeloid leukemia. *Ann Lab Med* **33**, 283–287 (2013).

19. Wu, P. *et al.* Profiling the Urinary Microbiota in Male Patients with Bladder Cancer in China. *Front Cell Infect Microbiol* **8**, 167 (2018).

Acknowledgements

The results shown here are in part based upon data generated by the TCGA Research Network (https://www.cancer.gov/tcga). This research was supported by the Intramural Research Program of the National Institutes of Health, National Cancer Institute. The authors would like to personally thank Christopher Buck from the NCI, Pedro Milanez-Almeida and John Tsang from NIH NIAID, and Alejandro Schäffer, Welles Robinson, Fiorella Schischlik, Sanju Sinha, and Sanna Madan from NCI CDSL for their assistance in this project. This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). The authors would like to thank Richard Lehr and Tim Miller for assistance running this analysis on the NIH HPC Biowulf cluster. The authors would also like to thank Joel Nothman, Andreas Mueller, and Adrin Jalali from the scikit-learn core development team and scikit-survival author Sebastian Pölsterl for their assistance with developing extensions to their libraries.

Author contributions

L.C.H., E.M.G., and E.R. designed the study. L.C.H. and E.M.G. performed all computational analyses and results interpretation. L.C.H., E.M.G., and E.R. wrote the paper.

Competing interests

All other authors declare that they have no competing interests.

Data and code availability

All data and code used to produce this work are available under https://github.com/ruppinlab/tcga-microbiome-prediction.

Supplemental Methods

Data retrieval and processing

Normalized and batch effect corrected microbial abundance data for 32 TCGA cancer types were downloaded from the online data repository referenced in Poore et al.[1] (ftp://ftp.microbio.me/pub/cancer_microbiome_analysis). Specifically, the "Kraken-TCGA-Voom-SNM-Plate-Center-Filtering-Data.csv" microbial abundance data file and adjoining "Metadata-TCGA-Kraken-17625-Samples.csv" metadata file were used as the starting input for further data processing.

We first filtered the data for primary tumor samples (TCGA "Primary Tumor" or "Additional - New Primary" sample types). Poore et al. generated microbial abundances from all the available WGS and RNA-seq data in legacy TCGA (after some quality filters), which frequently contained replicate WGS and RNA-seq data for each case and sample type. It was common in legacy TCGA to increase WGS sequencing coverage by performing an additional sequencing run from the same sample. Secondary runs typically had a much lower number of reads and coverage compared to their corresponding primary sequencing runs. We found that microbial abundance data which came from these lower coverage secondary runs could be substantially different from abundances derived from the larger primary sequencing runs. Therefore, we excluded microbial abundance data which came from secondary runs. In addition, legacy TCGA commonly contained data for the same samples analyzed using different computational pipeline versions. We excluded replicate microbial abundance data from older TCGA analysis pipeline versions if a replicate from a newer version existed. After the above filters, the Poore et al. data went from 17,625 samples and 10,183 unique cases to 12,111 samples and 9,812 unique cases.

Legacy TCGA curated survival phenotypic data[2] were obtained from UCSC Xena. The latest TCGA gender, age at diagnosis, and tumor stage demographic and clinical data and primary tumor RNA-seq read counts were obtained from the NCI Genomic Data Commons (GDC Data Release v24) using the R package GenomicDataCommons. TCGA GENCODE v22 gene annotations were obtained from the GDC data portal.

Drug response data were compiled from the TCGA Research Network. Our drug response models used the following binary classification targets: complete response (CR) and partial response (PR) as responders and stable disease (SD) and progressive disease (PD) as non-responders. All TCGA samples with drug response phenotypic data were from pre-treatment biopsies. Due to the limited cancer-drug combination cohort sizes in TCGA, we modeled each drug individually, even if a patient received multiple drugs concurrently. If the same drug was given at multiple timepoints to a patient, we only considered their first drug response. We considered cancer-drug combinations that contained a minimum of 18 cases and at least 4 cases per response binary class, except for STAD oxaliplatin, where we allowed a minimum of 14 cases so that the gene expression dataset could be included. In total, we analyzed 30 cancer-drug combinations which had paired microbial abundance and gene expression data that met the above thresholds. Combined microbial abundance and gene expression datasets were created by joining data from each individual dataset which had matching TCGA sample UUIDs. For some TCGA cases, data existed from multiple different aliquots per sample or multiple technical runs per aliquot, therefore in these cases all combinations were joined at the sample level. Cross-validation sampling probability weights as well as model and scoring sample weights were applied to account and adjust for any imbalance caused by the process.

ML modeling

Machine learning (ML) models were built using the scikit-learn[3] and scikit-survival libraries[4,5,6]. Custom extensions to scikit-learn and scikit-survival were developed to add new methods and functionalities required by this project. Survival models were built using Coxnet – regularized Cox regression with elastic net penalties[7]. Drug response models were built using a variant of the linear support vector machine recursive feature elimination (SVM-RFE) algorithm[8] that we developed, where we could include features in the modeling algorithm that bypassed feature elimination. Coxnet models controlled for gender, age at diagnosis, and tumor stage prognostic covariates by including them as unpenalized features in the model (i.e. penalty factor = 0). Gender was one-hot encoded and tumor stage ordinal encoded by major stage. These same covariates were included in drug response SVM-RFE models as penalized features that were excluded from elimination. All models included normalization and transformation steps integrated into the ML modeling pipeline. Training data was normalized and transformed independently from held-out test data within the ML pipeline before learning. Models built using gene expression read count data included edgeR[9,10] low count filtering, TMM normalization, and logCPM transformation steps within the ML pipeline. These were developed and integrated into our scikit-learn-based framework via R and rpy2. All models also included standardization of features within the ML pipeline before learning. During prediction, held-out test data was normalized and transformed through the ML pipeline using the parameters learned from the training data at each pipeline step before model prediction.

Each cancer, data type, and survival or drug response target type combination was modeled individually using a nested cross-validation (CV) strategy to perform model selection and evaluation on held-out test data. All cross-validation iterators kept replicate sample data per case grouped together such that data would only reside in either the train or test split during each CV iteration.

Survival models used a stratified, randomly shuffled outer CV with 75% train and 25% test split sizes that was repeated 100 times. The CV procedure stratified the splits on event status. Each training set from the outer CV was used to perform hyperparameter tuning and model selection by optimizing C-index over a stratified, randomly shuffled, 4-fold inner CV on the training set, repeated 5 times. A few cancer datasets contained fewer than four uncensored cases which required reducing the number of inner CV folds for these models such that at least one case per fold was uncensored. The data derived from Poore et al. often included more than one sample per case, and an unequal number of samples between cases, therefore requiring either ML model sample weighting or CV random sampling per case. The Coxnet implementation in scikit-survival does not currently support sample weights, therefore our outer CV iterator randomly sampled one replicate sample per case during each iteration, using a sampling procedure with probability weights that balanced the probability that a replicate WGS- or RNA-seq-based sample was selected during each CV iteration. Model selection grid search was performed on the following hyperparameters: elastic net penalty L1 ratios 0.1, 0.3, 0.5, 0.7, 0.8, 0.9, 0.95, 0.99, and 1, and for each L1 ratio a default alpha path of 100 alphas using an alpha min ratio of 10^{-2}. Alpha is the constant multiplier of the penalty terms in the Coxnet objective function. Optimal alpha and L1 ratio settings were determined via inner CV and a model with these settings was then refit on the entire outer CV train split. Model performance was evaluated in both inner and outer CV on the held-out test split by generating test predicted risk scores and using these scores to calculate a Harrell's concordance index (C-index). We also evaluated and compared model predictive performance for the test data survival time period by calculating time-dependent cumulative/dynamic AUCs[11,12].

Drug response models used a stratified, randomly shuffled, 4-fold outer CV that was repeated 25 times. Each training set from the outer CV was used to perform hyperparameter tuning and model selection by optimizing the area under receiver-operator curve (AUROC) over a stratified, randomly shuffled, 3-fold inner CV repeated 5 times. Case replicate sample weights were provided to SVM-RFE and all model selection and evaluation scoring methods. Class weights were provided to SVM-RFE to adjust for any class imbalance. Model selection grid search was performed on the following hyperparameters: SVM C regularization parameter from a range of 10^{-5} to 10^{3}, and RFE k top-ranking features to select from 1 to 100 microbial abundance or gene expression features. Clinical covariate features bypassed recursive feature elimination but were always included in each RFE recursive feature elimination model fitting step as well as final model refitting. Optimal C and k settings were determined via inner CV and a model with these settings was then refit on the entire outer CV train split. Model performance was evaluated in both inner and outer CV on the held-out test split by AUROC, area under precision-recall curve (AUPRC), average precision, and balanced accuracy.

Gender, age at diagnosis, and tumor stage clinical covariate-only survival models were built using standard unpenalized Cox regression. Clinical covariate-only drug response models were built using linear SVM. Models included standardization of features as part of the ML pipeline. Models were trained and tested using the same outer CV iterators and train/test data splits as their corresponding microbial, gene expression, or combination data type models. To test whether a Coxnet or SVM-RFE microbial or gene expression model was significantly better than its corresponding Cox and linear SVM clinical covariate-only model, a two-sided Wilcoxon signed-rank test was performed between the 100 pairs of C-index or AUROC scores between both models. All raw p-values generated from this test across survival or drug response analyses from the same data type were adjusted for multiple testing using the Benjamini-Hochberg (BH) procedure to control the false discovery rate (FDR), and a threshold $\mathrm{FDR} \leq 0.01$ was used to determine statistical significance.

Microbial model feature analysis

For each analysis, 100 Coxnet or SVM-RFE model instances were generated from the outer CV procedure. Each model instance selected a subset of features that performed best during cross-validation and the model algorithm learned coefficients (or weights) for each feature. To select microbial genera for downstream investigation from the feature results across all these model instances, we proceeded as follows. First, we applied a two-sided Wilcoxon signed-rank test that the mean feature coefficient generated by the model is shifted away from zero, and thus that the genus is identifiably positively or negatively associated with survival or drug response. Coefficients were ignored when a genus was assigned a zero coefficient or absent from a model. Second, within each model, all coefficients, ignoring the results of the Wilcoxon test, were ranked by absolute magnitude. We then kept genera that were among the top 50 features in at least 20% of the models and for which the Holm-adjusted, two-sided Wilcoxon signed-rank test p-value was ≤ 0.01. Having a Coxnet feature coefficient equal to zero or feature being absent from an SVM-RFE model was not strong enough evidence that the genus has no effect, but rather that one or more features with stronger effect were chosen. Thus, we ignored genera with a zero coefficient or absent from a model when computing mean coefficient weight and Wilcoxon statistics on the means.

We analyzed the distribution of features, selected by the rules described above, that had positive or negative signs for their mean coefficient. We used a two-sided binomial test to show that selected features had significantly more negative the positive mean coefficients. We used a two-sided Fisher's exact test to determine if selected genera belonging to Firmicutes had a statistically significant difference in the breakdown between positive and negative mean coefficients than selected features as a whole.

References

1. Poore, G. D. *et al.* Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579**, 567–574 (2020).

2. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**, 400-416.e11 (2018).

3. Pedregosa *et al*. Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**, 2825-2830 (2011).

4. Pölsterl, S., Navab, N. & Katouzian, A. Fast Training of Support Vector Machines for Survival Analysis. in *Machine Learning and Knowledge Discovery in Databases* (eds. Appice, A. et al.) 243–259 (Springer International Publishing, 2015).

5. Pölsterl, S., Navab, N., & Katouzian, A., An Efficient Training Algorithm for Kernel Survival Support Vector Machines. *4th Workshop on Machine Learning in Life Sciences*, 23 September 2016, Riva del Garda, Italy.

6. Pölsterl, S. *et al.* Heterogeneous ensembles for predicting survival of metastatic, castrate-resistant prostate cancer patients. *F1000Res* **5**, 2676 (2017).

7. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J Stat Softw* **39**, 1–13 (2011).

8. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* **46**, 389–422 (2002).

9. 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).

10. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* **40**, 4288–4297 (2012).

11. Hung, H. & Chiang, C.T. Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics*, **38** (1), 8–26 (2010).

12. Lambert, J. & Chevret, S. Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves. *Statistical methods in medical research*, **25** (5), 2088–2102 (2016).

Extended Tables

Extended Table 1: By cancer and by comparator, the genera selected as features, the number of times each genus was seen with rank at most 50 among the 100 instances of each model, the mean coefficient of the genus, the median absolute rank of the genus, and the p-value of a Holm-corrected two-sided Wilcoxon signed rank test that the coefficient was shifted away from zero. Means and medians were only taken for those instances for which the genus had rank of at most 50.

Extended Table 1 is supplied in a separate Excel file.

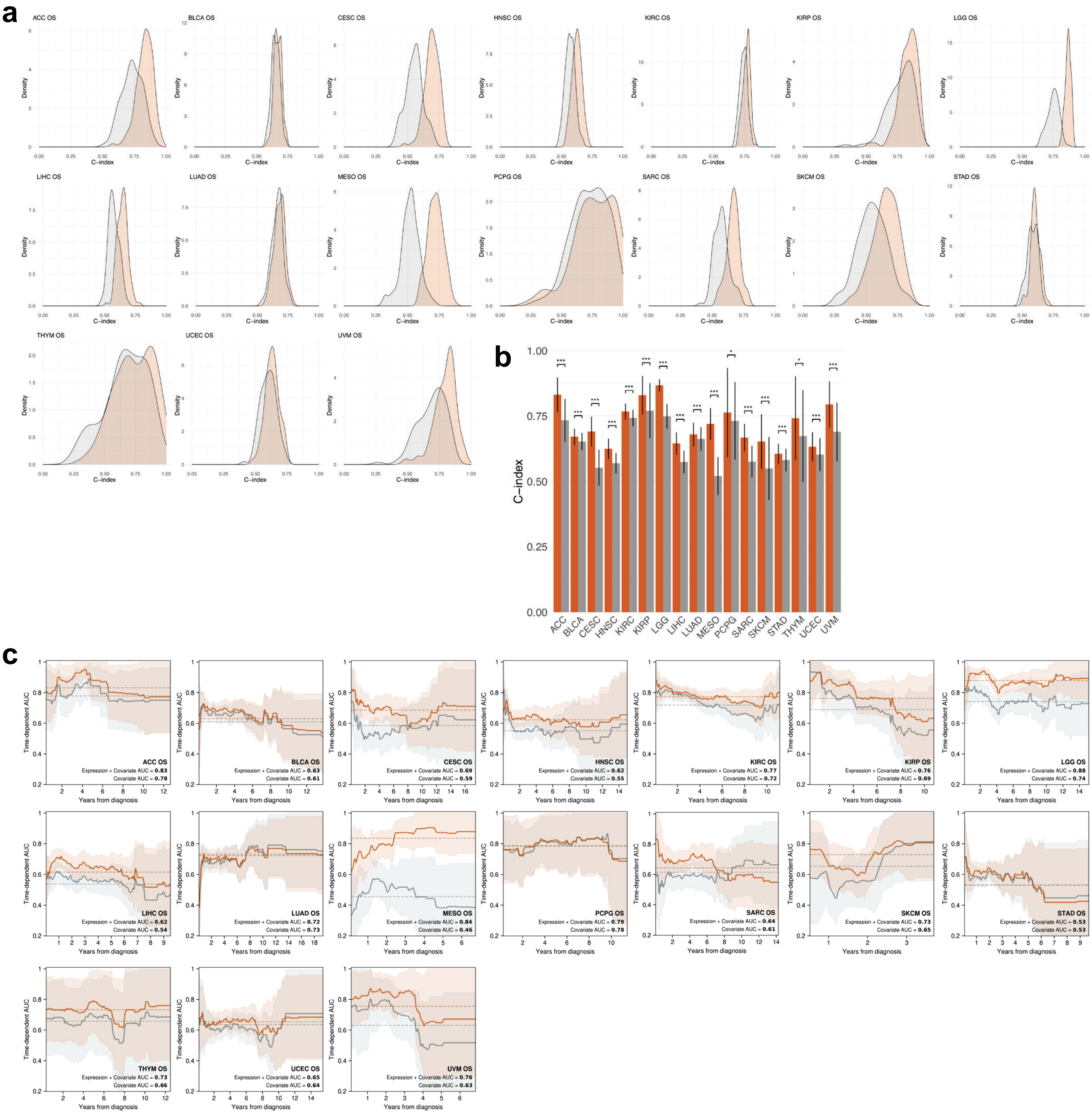
Extended Table 2: By cancer and by comparator, the number of features identified, the number of features that were positive or negative, and the median number of times the identified features were seen in a model with rank of at most 50.

Cancer	Versus	Features Selected	Positive Features	Negative Features	Median Models
ACC	OS	68	31	37	34
ACC	PFI	75	29	46	32
BLCA	Cisplatin	62	21	41	32
BLCA	Gemcitabine	41	13	28	30
BRCA	Docetaxel	3	1	2	41
CESC	OS	66	29	37	36
LGG	PFI	42	23	19	29
SARC	Docetaxel	18	11	7	37.5
SKCM	OS	52	30	22	31.5
STAD	Cisplatin	67	41	26	29
STAD	Leucovorin	48	13	35	35
STAD	Oxaliplatin	36	5	31	28

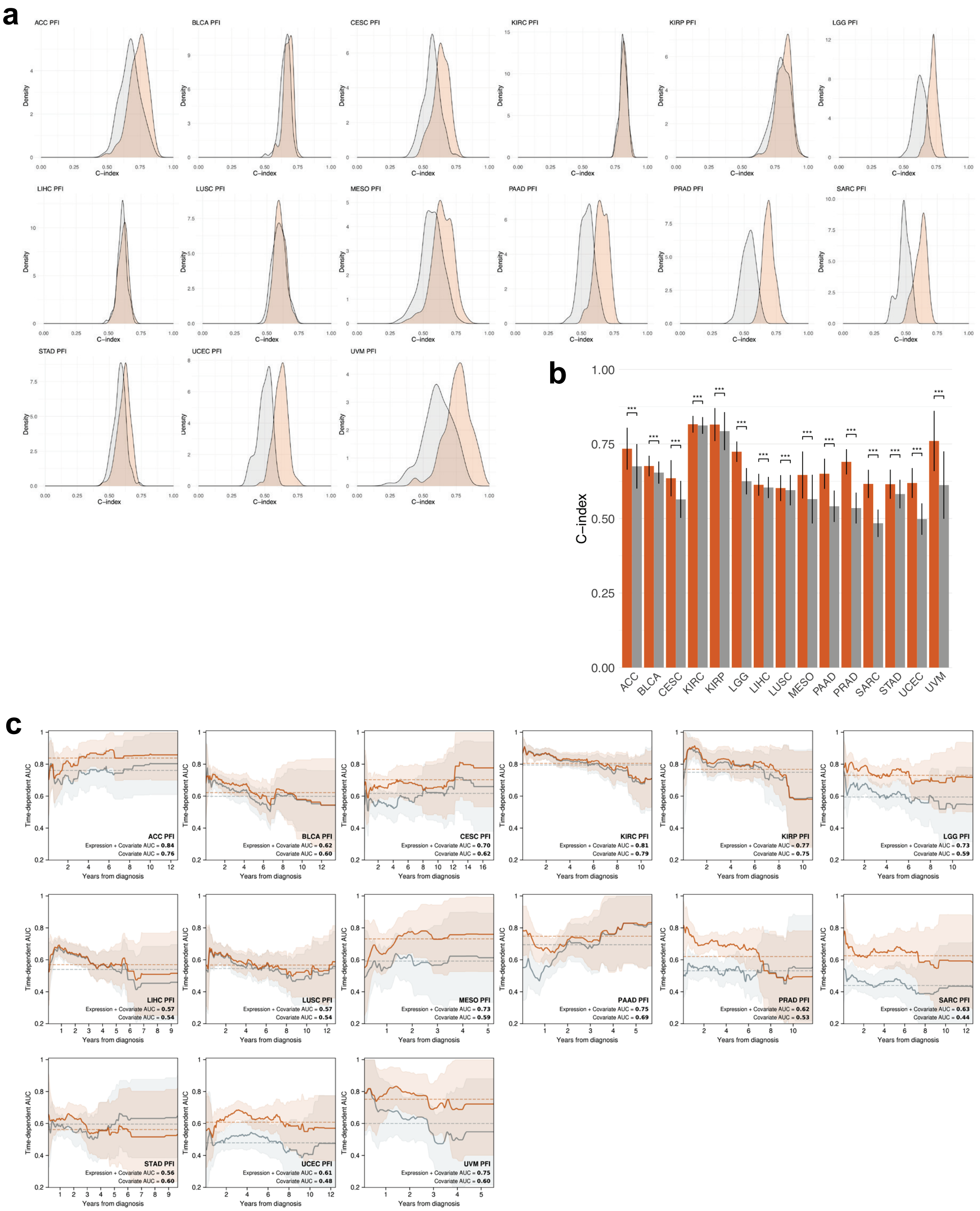
Extended Table 3. Per cancer, the number of times genera from the phylum Firmicutes were found among the selected features, whether positively or negatively associated with drug response or survival.

Cancer	Selected Positively	Selected Negatively	Total
ACC	10	11	21
BLCA	8	3	11
CESC	7	3	10
LGG	6	1	7
SARC	3	2	5
SKCM	3	3	6
STAD	8	11	19

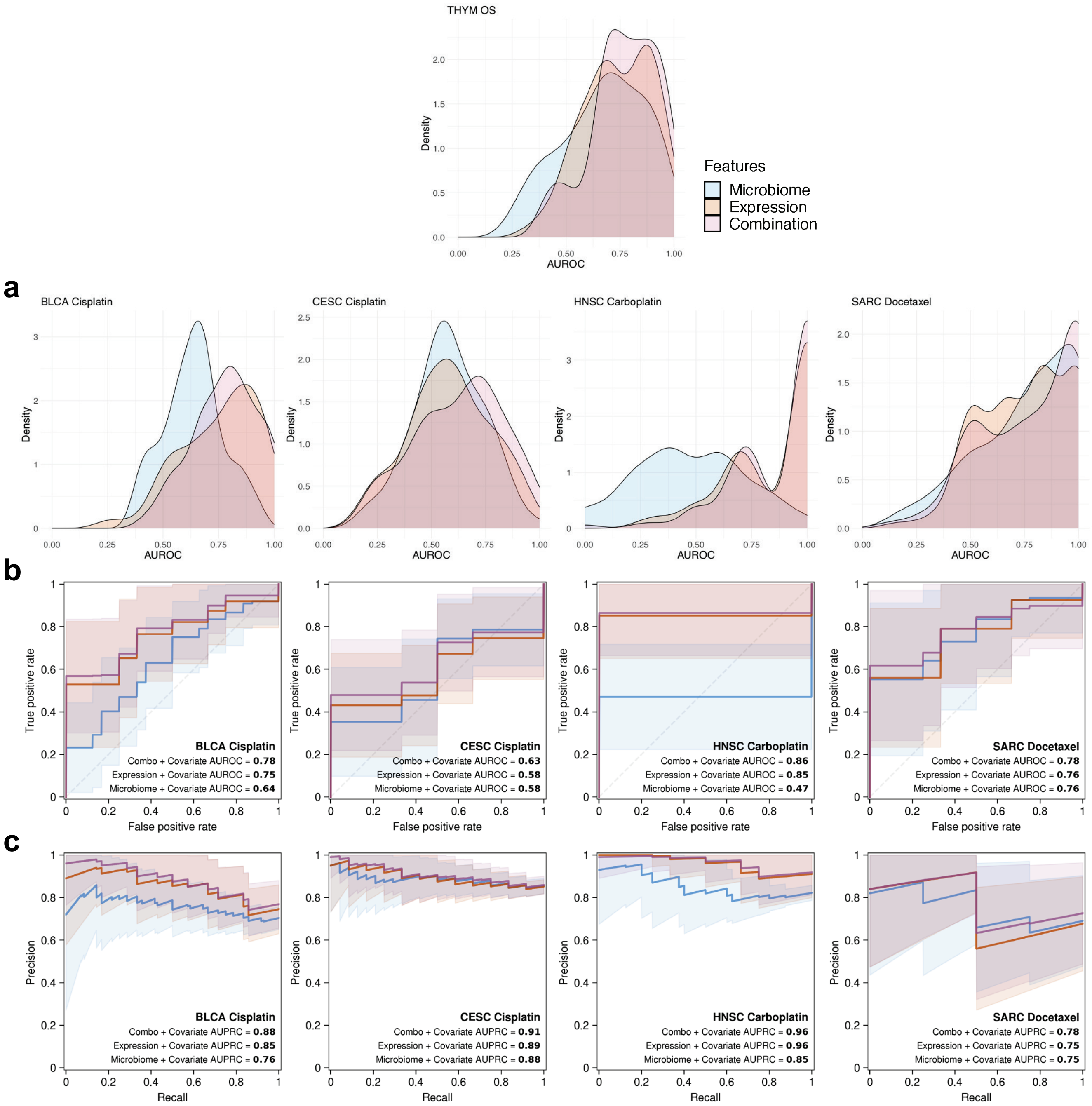
Extended Figures



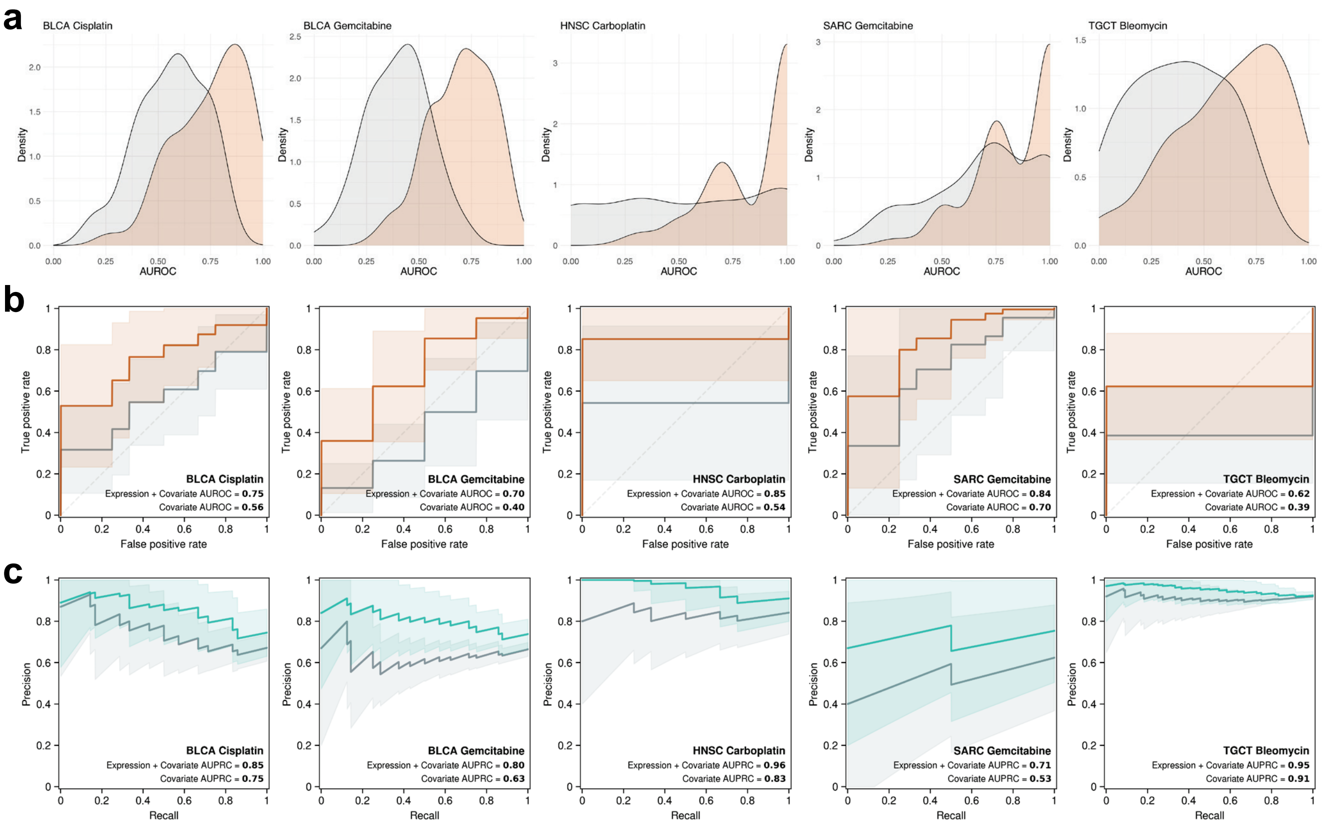
Extended Figure 1. Performance of combined gene expression and clinical covariate overall survival (OS) models in the 17 cancer types where gene expression adds to OS predictive power. (**a**) C-index score density distributions for combined gene expression and clinical covariate Coxnet models (orange) and corresponding clinical covariate-only Cox models (grey). (**b**) Mean C-index score comparison between combined gene expression and clinical covariate models (orange bars) and corresponding clinical covariate-only models (grey bars). Error bars in denote standard deviations. Significance test for comparing combined gene expression and clinical covariate model scores to clinical covariate-only model scores was calculated using a two-sided Wilcoxon signed-rank test and adjusted for multiple testing using the Benjamini-Hochberg method. Significance: * ≤ 0.01, ** ≤ 0.001, *** ≤ 0.0001. (**c**) Time-dependent cumulative/dynamic AUCs for combined gene expression and clinical covariate models (orange) and clinical covariate-only models (grey) following years after diagnosis. Shaded areas denote standard deviations.



Extended Figure 2. Performance of combined gene expression and clinical covariate progression-free interval (PFI) survival models in the 15 cancer types where gene expression adds to PFI predictive power. (**a**) C-index score density distributions for combined gene expression and clinical covariate Coxnet models (orange) and corresponding clinical covariate-only Cox models (grey). (**b**) Mean C-index score comparison between combined gene expression and clinical covariate models (orange bars) and corresponding clinical covariate-only models (grey bars). Error bars in denote standard deviations. Significance test for comparing combined gene expression and clinical covariate model scores to clinical covariate-only model scores was calculated using a two-sided Wilcoxon signed-rank test and adjusted for multiple testing using the Benjamini-Hochberg method. Significance: * ≤ 0.01, ** ≤ 0.001, *** ≤ 0.0001. (**c**) Time-dependent cumulative/dynamic AUCs for combined gene expression and clinical covariate models (orange) and clinical covariate-only models (grey) following years after diagnosis. Shaded areas denote standard deviations.



Extended Figure 3. Performance of combined microbiome, gene expression, and clinical covariate models where combining both data types add predictive power. (**a**) C-index or AUROC density distributions for combined (purple) versus microbiome (blue) and gene expression (orange) models. (**b, c**) ROC and precision-recall curves for combined (purple) versus microbiome (blue) and gene expression (orange) models.



Extended Figure 4. Performance of combined gene expression and clinical covariate drug response models in the five cancer types where gene expression adds predictive power. (**a**) Area under the receiver operator characteristic (AUROC) density distributions for combined gene expression and clinical covariate SVM-RFE models (orange) and corresponding clinical covariate-only SVM models (grey). (**b**) ROC curves for combined gene expression and clinical covariate models (orange) and clinical covariate-only models (grey). (**c**) Precision-recall (PR) curves for combined gene expression and clinical covariate models (green) and clinical covariate-only models (grey). Shaded areas in (**b**) and (**c**) denote standard deviations.