Predicting cancer prognosis and drug response from the tumor microbiome

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

Tumor gene expression is predictive of patient prognosis in some cancers, although reads from expression studies also contain reads from the tumor microbiome, which can be used to infer the microbial abundances in each tumor. Here, we show that tumor microbial abundances, alone or in combination with tumor expression data, can predict cancer prognosis and drug response to some extent: microbial abundances are significantly less predictive of prognosis than gene expression, although remarkably, modestly more predictive of chemotherapy drug response.

Main

Milanez-Almeida et al.[1] recently showed that gene expression from The Cancer Genome Atlas (TCGA) RNA-seq data could predict overall survival (OS) or progression-free interval (PFI) better than classical clinical prognostic covariates – age at diagnosis, gender, and tumor stage. Poore et al.[2] recently published estimated, decontaminated, and normalized microbial abundances for the TCGA cohort, derived from either tumor whole genome sequencing (WGS) or RNA-seq data. Importantly, for RNA-seq experiments, both gene expression and microbial abundances can be inferred from raw reads and may represent non-redundant information. To compare the prognostic predictive power between tumor microbial abundances and gene expression, we began by replicating the Milanez-Almeida et al. [1] analysis using our own machine learning methodology and the TCGA cohort of primary tumor gene expression data.

We built OS and PFI models of 32 TCGA tumor types using the Coxnet[3] algorithm, which jointly selects the most predictive subset of features via cross-validation (CV) while simultaneously controlling for clinical covariates. For comparison, we also built standard Cox regression models based on the clinical covariates alone. We evaluated the predictive

performance of our models using Harrell's concordance index (C-index), which is a metric of survival model predictive accuracy. Each model analysis generated 100 model instances and C-index scores from different randomly shuffled train-test CV splits on the data (Supplemental Methods). We found 32 OS and PFI models for 20 tumor types that had a mean C-index score ≥ 0.6 and significantly outperformed their corresponding clinical covariate-only models (**Figure 1a, Extended Data Figs. 1-2**). Our models were predictive of prognosis in 11 of the same 13 tumor types that were reported by Milanez-Almeida et al.[1] (we did not analyze one tumor type that Milanez-Almeida did, acute myeloid leukemia (LAML), because Poore et al. excluded it from their analysis). Among the cancers and outcomes that Milanez-Almeida et al. analyzed, our methodology produced predictive models for four additional tumor types: cervical squamous cell carcinoma (CESC), sarcoma (SARC), stomach adenocarcinoma (STAD), and uterine corpus endometrial carcinoma (UCEC), as well as predictive models for additional cancers and outcomes that were not analyzed in their study.

We applied Coxnet[3] using the same methodology to build prognosis models using the microbial abundance estimates provided by Poore et al.[2]. We found five microbial abundance models that had a mean C-index score ≥ 0.6 and significantly outperformed their corresponding clinical covariate-only models (**Fig. 1b, c**). We also evaluated model performance by calculating the time-dependent, cumulative/dynamic area under the curve ($AUC^{C/D}(t)$)[4,5], which is an extension of the area under the receiver operating characteristic (AUROC) for continuous outcomes. We found that in only two of the five models, microbial abundances significantly outperformed clinical covariates alone in terms of $AUC^{C/D}(t)$ (**Extended Data Fig. 3a**). In adrenocortical carcinoma (ACC), microbial features predicted overall survival significantly better than clinical covariates starting at approximately 6 years after diagnosis. In CESC, microbial features predicted overall survival better than clinical covariates from approximately 6

months to 10 years after diagnosis. Overall, we found that tumor microbial abundances are only marginally predictive of prognosis across the TCGA cohort, and that gene expression is a significantly more powerful predictor of prognosis (**Fig. 1a, c, Extended Data Figs. 1-2**).

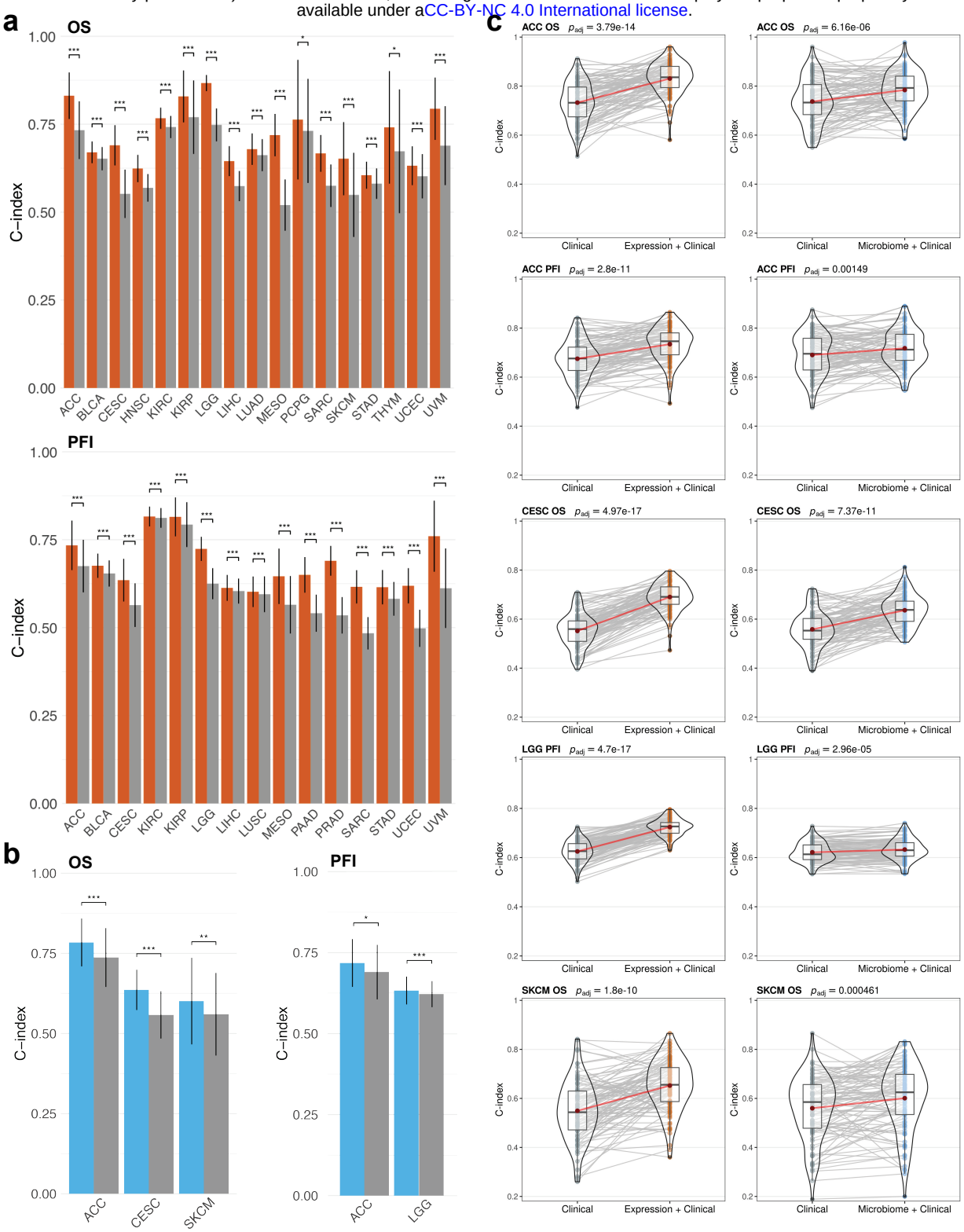


Figure 1. Performance of gene expression and microbial abundance prognosis prediction models where features add predictive power to clinical covariates. Mean C-index scores for **(a)** gene expression with clinical covariate models (orange) and **(b)** microbial abundance with clinical covariate models (blue) vs clinical covariate-only models (grey). Error bars denote standard deviations. Significance: * ≤ 0.01, ** ≤ 0.001, *** ≤ 0.0001. **(c)** C-index score density distributions for the five models where microbial abundance with clinical covariate features outperform clinical covariate-only models. Corresponding gene expression models shown for comparison. Lines connecting points (light grey) represent score pairs from same train-test split on the data. Mean C-index scores and connecting lines shown in red. Significance for the prediction improvement over clinical covariate-only models was calculated using a two-sided Wilcoxon signed-rank test and adjusted for multiple testing using the Benjamini-Hochberg method with adjusted p-values shown at top.

We next asked whether tumor gene expression from pre-treatment biopsies could predict drug response better than clinical covariates alone. Thirty TCGA cancer-drug combinations met our minimum dataset size thresholds (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 to unconditionally include clinical covariates while selecting the most predictive subset of features (Supplemental Methods). For comparison, we built linear SVM models using clinical covariates alone. We evaluated the predictive performance of drug response models using AUROC. Each analysis generated 100 model instances, AUROC, and area under the precision-recall curve (AUPRC) scores from different random train-test CV splits. Five cancer-drug gene expression models had a mean AUROC score ≥ 0.6 and performed better than clinical covariates alone. (**Figure 2c, Extended Data Fig. 3b**).

We performed the same drug response modeling using the TCGA microbial abundance estimates. Here, seven cancer-drug microbial abundance models had a mean AUROC score ≥ 0.6 and significantly outperformed their corresponding clinical covariate-only models (**Fig. 2a, b**). Three of these cancer-drug combinations involved stomach adenocarcinoma (STAD). Two of these combinations, urothelial bladder cancer (BLCA) cisplatin and gemcitabine treatments, overlapped with the gene expression model results. Overall, our results support the notion that the tumor microbiome may contain information that is predictive of drug response, consistent with recent reports[7,8].

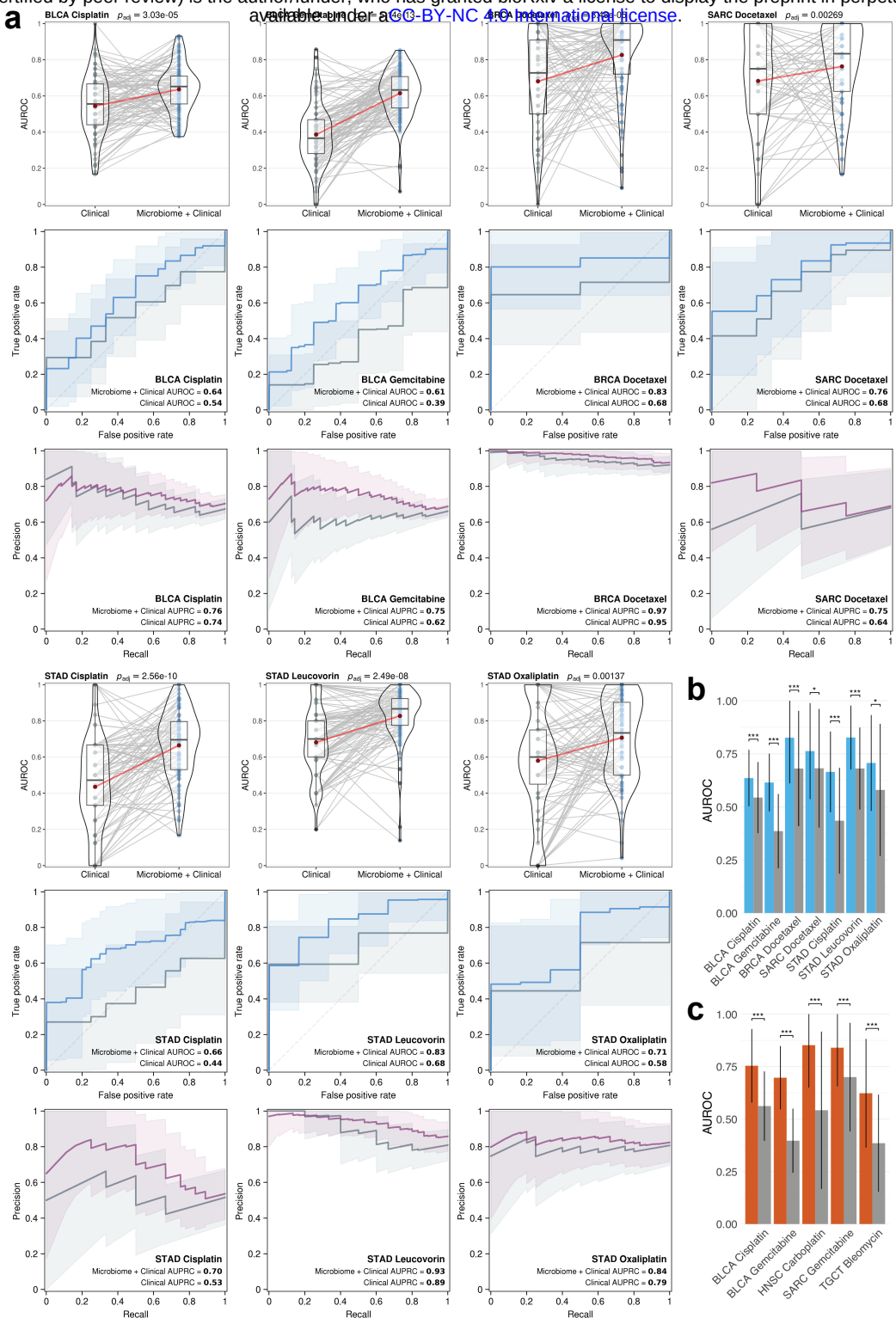


Figure 2. Performance of microbial abundance drug response prediction models in the seven cancer-drug combinations where models performed better than clinical covariates alone. (a) AUROC score density distributions for microbial abundance with clinical covariate models (top rows, blue) vs clinical covariate-only models (grey). Lines connecting points (light grey) represent score pairs from same train-test split on the data. Mean AUROC scores and connecting lines shown in red. Mean ROC (middle rows, blue) and precision-recall (PR) curves (bottom rows, purple) for microbial abundance with clinical covariate models vs clinical covariate-only models. Mean AUROC and AUPRC scores shown in legends and shaded areas denote standard deviations. Mean AUROC scores for **(b)** microbial abundance with clinical covariate (blue) and **(c)** gene expression with clinical covariate (orange) models vs clinical covariate-only models. Error bars denote standard deviations. Significance: * ≤ 0.01, ** ≤ 0.001, *** ≤ 0.0001. Significance was calculated using a two-sided Wilcoxon signed-rank test and adjusted for multiple testing using the Benjamini-Hochberg method with adjusted p-values shown at top of violin plots in **(a).**

Finally, we investigated if models built from combined microbial abundance and gene expression features would result in an improvement in predictive power over their corresponding single data type models. The combined feature prognosis models resulted in a modest predictive improvement in just three models: SARC OS, STAD PFI, and THYM OS (**Extended Data Fig. 4a**). Although this improvement was not statistically significant in terms of C-index score, the $AUC^{C/D}(t)$ metric showed a clear improvement in prognostic predictive power from the combined data type models. We also found a modest improvement in drug response models for two cancer-drug combinations (**Extended Data Fig. 4b**).

To learn more about the most predictive microbial features, we determined the top microbial genera selected by each of the significantly predictive microbial abundance models according to their selection frequency and weight coefficients across the 100 model splits. There were 444 distinct such microbial genera appearing in at least one prognosis or drug response model (**Extended Data Table 1**). Of these 444 genera, 149 were individually significantly predictive of prognosis or drug response by a Wilcoxon test, indicating that the other genera were significantly predictive in combination (Supplemental Methods). 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 444 genera, 114 were selected in more than one model and only 21 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.

The predictive genera we found span all non-eukaryotic domains of life, in total encompassing 380 bacterial, 22 archaeal, and 42 viral genera (**Extended Data Table 2**). Proteobacteria and Firmicutes were the most frequent phyla, followed by Actinobacteria and Bacteroidetes. Among viruses, Herpesvirales were the most frequent. We found that more

microbial genera, when predictive of prognosis or drug response, were negatively associated with the prediction target than positively associated (two-sided binomial test p-value = 0.0004). Herpesvirales and the most frequently selected bacterial phyla individually exhibited an overall negative-over-positive predictive trend, except for Firmicutes, where the abundance of more genera was positively associated with prognosis or drug response (two-sided Fisher's exact test p-value = 0.007; genera shown in **Extended Data Table 2**). Notably, though CESC is known to often arise from HPV infection, the presence of other microbial species, in particular the Firmicutes species Lactobacillus, has been previously associated with the risk of developing CESC[10].

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]. For STAD, the microbiome was predictive of response to three different drugs: cisplatin, leucovorin, and oxaliplatin. In STAD, though patients infected with *H. pylori* have an increased risk of developing gastric cancer[15], *H. pylori* was not a predictive feature of drug response in our models. Cedecea and Sphingobacterium were both strongly negatively predictive of leucovorin drug response in STAD. Notably, both genera have been previously implicated in bacteremia in immunocompromised individuals in rare cases, including cancer[16,17,18,19].

In summary, we find that the microbial abundance estimates generated by Poore et al.[2] are predictive of patient prognosis and response to chemotherapy in a subset of tumor types and treatments. Overall, in terms of the number of significant models, based on their cross-validated C-index or AUROC scores and improvement over clinical covariates alone, the tumor

microbiome is considerably less predictive than the tumor transcriptome at predicting patient prognosis, but notably, modestly better at predicting chemotherapy response. Our investigation motivates future studies investigating the role of the tumor microbiome in predicting the response to targeted therapies and immunotherapies.

References

1. 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).

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

3. 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).g

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

5. 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).

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 their assistance in 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 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 tumor 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.

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 abundance, gene expression, or combination data type models. To test whether a Coxnet or SVM-RFE microbial abundance 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 the signed-rank 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 FDR ≤ 0.01 was used to determine statistical significance. To test whether a combined data type model was significantly better than its corresponding microbial abundance or gene expression model, a two-sided Dunn test was performed between all three groups of scores. Each Dunn test raw p-value was adjusted for multiple testing using the Benjamini-Hochberg (BH) procedure to control the false discovery rate (FDR), and a threshold FDR ≤ 0.05 was used to determine statistical significance.

Microbial abundance 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 rank 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. For all Wilcoxon tests, we used the package coin[13], which allows exact calculation of p-values. 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.

For each selected feature, we tested whether it was a significantly univariate feature of survival or drug response. This is a strictly different question than whether the coefficient of a feature has consistent sign – sign may be consistent when used in combination with other features, but the feature may not be individually predictive. For drug response models, we divided individuals into responsive or non-responsive, and for survival data we divided individuals whose survival time was greater or less than the censored median, ignoring those who were lost to follow up before median time. For each cancer-test type pair, we applied a two-sided Wilcoxon rank-sum test. We applied a Benjamini-Hochberg multiple hypothesis correction for each cancer-test type pair and report the false discovery rate in Extended Table 1.

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).

13. Hothorn, T., Hornik, K., Wiel, M. A. van de & Zeileis, A. Implementing a Class of Permutation Tests: The coin Package. *Journal of Statistical Software* **28**, 1–23 (2008).

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, and the FDR that the feature is univariately predictive. 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.

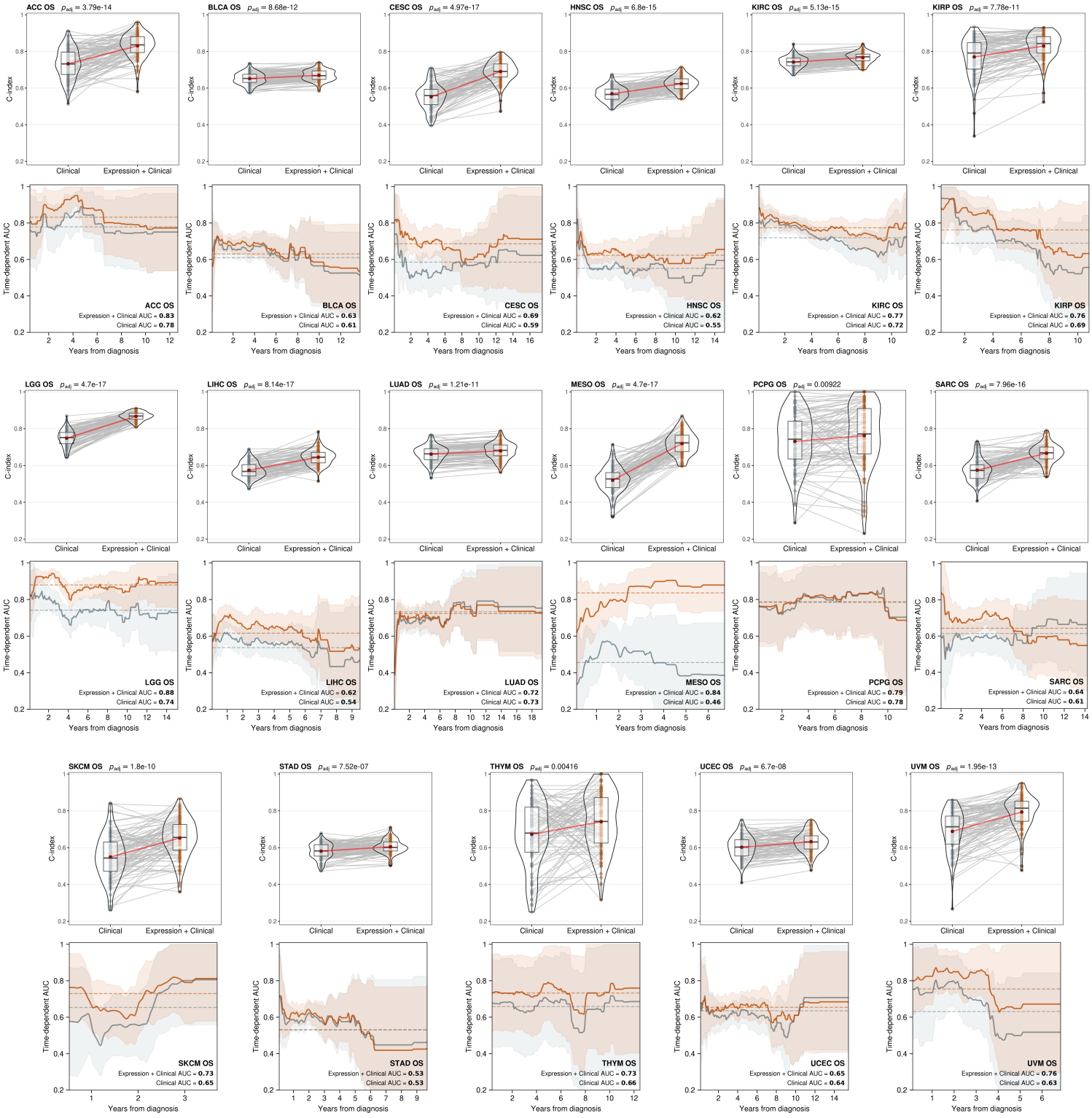
Cancer	Versus	Features Selected	Positive Features	Negative Features	Median Models
ACC	OS	68	31	37	34
ACC	PFI	75	29	46	32
BLCA	Cisplatin	62	21	41	32
BLCA	Gemcitabine	41	13	28	30
BRCA	Docetaxel	3	1	2	41
CESC	OS	66	29	37	36
LGG	PFI	42	23	19	29

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

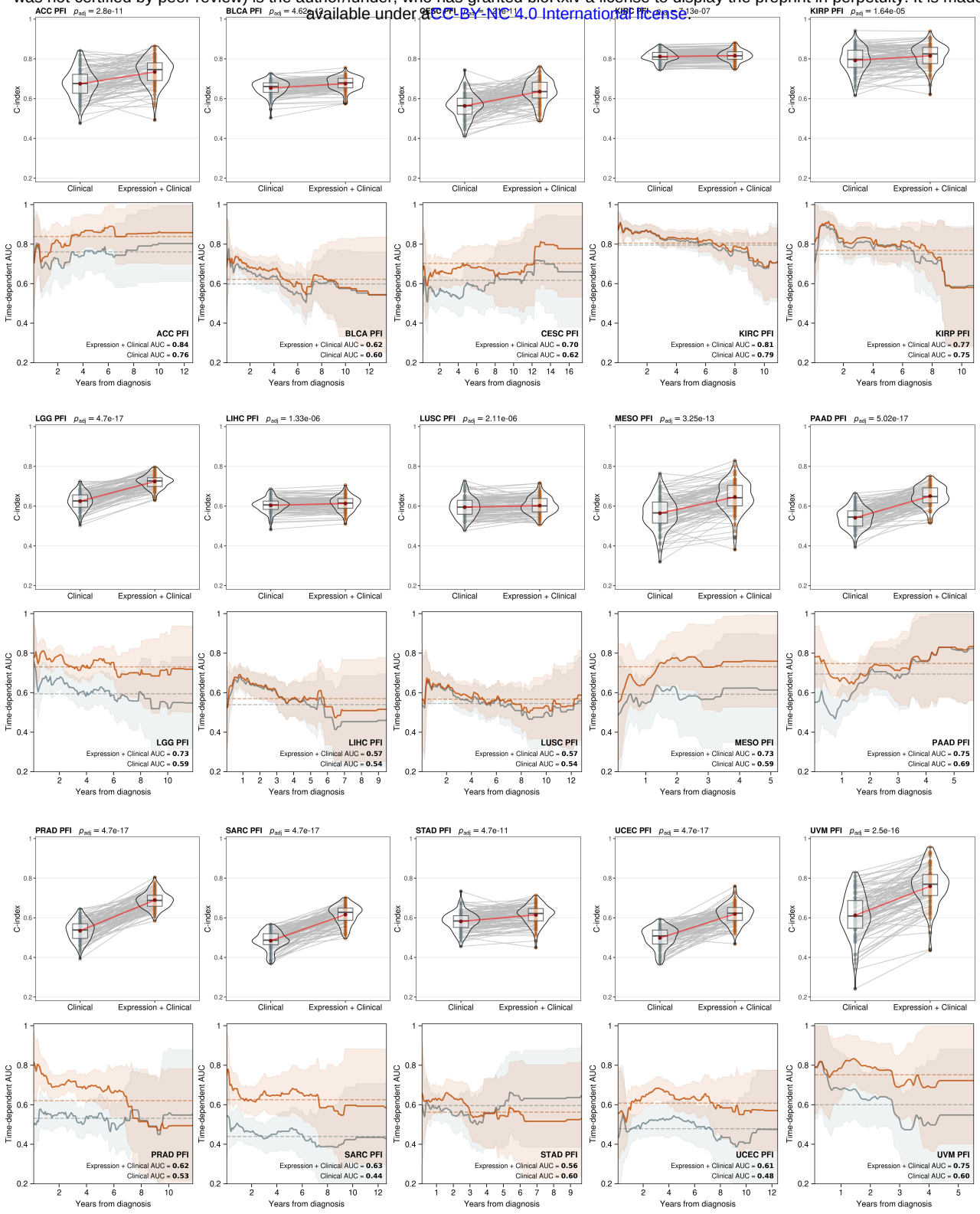
Extended Table 3. Per cancer, the number of times genera from the phylum Firmicutes were found among the selected features, whether positively or negatively associated with drug response or survival.

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

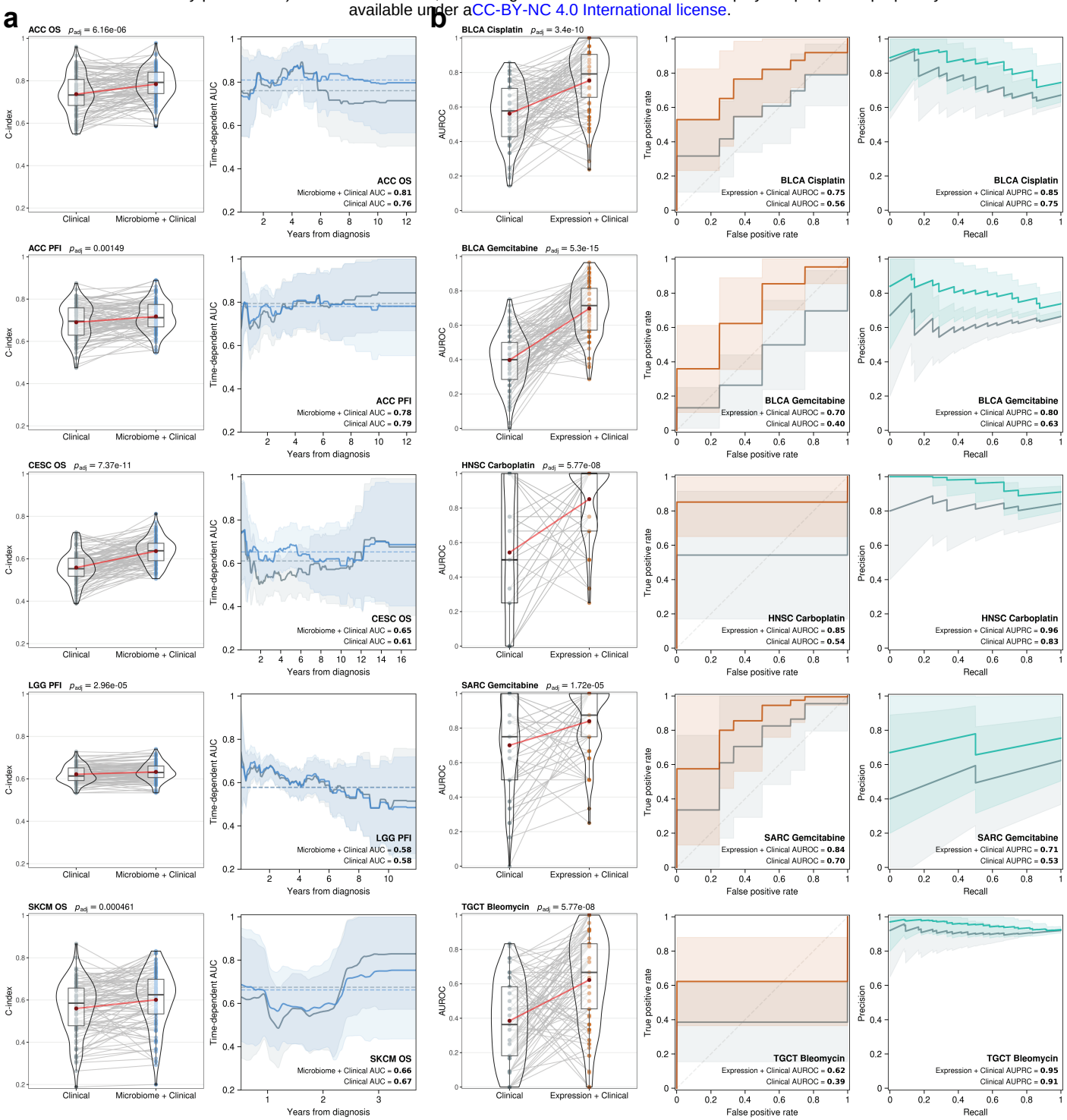
Extended Figures



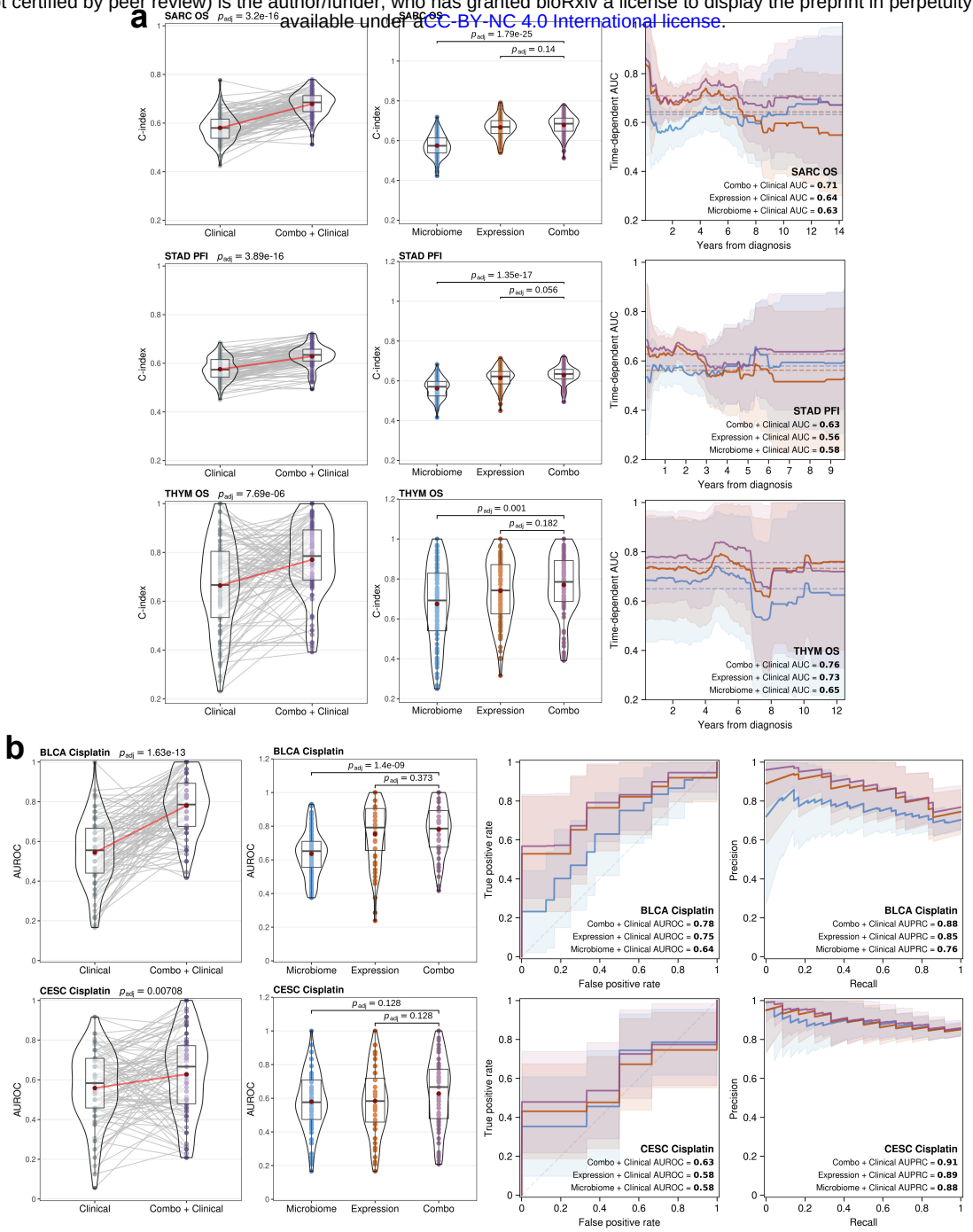
Extended Figure 1. Performance of combined gene expression and clinical covariate overall survival (OS) models in the 17 tumor types where gene expression adds to OS predictive power. C-index score density distributions (top rows) for gene expression with clinical covariate models (orange) vs clinical covariate-only models (grey). Lines connecting points (light grey) represent score pairs from same train-test split on the data. Mean C-index scores and connecting lines shown in red. Significance for the prediction improvement over clinical covariate-only models was calculated using a two-sided Wilcoxon signed-rank test and adjusted for multiple testing using the Benjamini-Hochberg method with adjusted p-values shown at top. Time-dependent, cumulative/dynamic AUCs (bottom rows) for gene expression with clinical covariate models vs clinical covariate-only models following years after diagnosis. Mean AUCs across entire time scales shown as a horizontal dotted line and in legends and shaded areas denote standard deviations.



Extended Figure 2. Performance of combined gene expression and clinical covariate progression-free interval (PFI) models in the 15 tumor types where gene expression adds to PFI predictive power. C-index score density distributions (top rows) for gene expression with clinical covariate models (orange) vs clinical covariate-only models (grey). Lines connecting points (light grey) represent score pairs from same train-test split on the data. Mean C-index scores and connecting lines shown in red. Significance for the prediction improvement over clinical covariate-only models was calculated using a two-sided Wilcoxon signed-rank test and adjusted for multiple testing using the Benjamini-Hochberg method with adjusted p-values shown at top. Time-dependent, cumulative/dynamic AUCs (bottom rows) for gene expression with clinical covariate models vs clinical covariate-only models following years after diagnosis. Mean AUCs across entire time scales shown as a horizontal dotted line and in legends and shaded areas denote standard deviations.



Extended Figure 3. Performance of microbial abundance prognosis models and gene expression drug response prediction models where features performed better than clinical covariates alone. (a) C-index score density distributions (left panels) for microbial abundance with clinical covariate models (blue) vs clinical covariate-only models (grey). Lines connecting points (light grey) represent score pairs from same train-test split on the data. Mean C-index scores and connecting lines shown in red. Time-dependent, cumulative/dynamic AUCs (right panels) for microbial abundance with clinical covariate models vs clinical covariate-only models following years after diagnosis. Mean AUCs across entire time scales shown as a horizontal dotted line and in legends and shaded areas denote standard deviations. **(b)** AUROC score density distributions (left panels) for gene expression with clinical covariate models (orange) vs clinical covariate-only models. Mean ROC (middle panels) and precision-recall (PR) curves (right panels) for gene expression with clinical covariate models (orange, green) vs clinical covariate-only models. Mean AUROC and AUPRC scores shown in legends and shaded areas denote standard deviations. Significance was calculated using a two-sided Wilcoxon signed-rank test and adjusted for multiple testing using the Benjamini-Hochberg method with adjusted p-values shown at top of violin plots.



Extended Figure 4. Performance of combined microbial abundance, gene expression, and clinical covariate models where combining both data types adds to predictive power. **(a, b)** Model score density distributions (left panels) for combined data type (microbial abundance and gene expression) with clinical covariate models (purple) vs clinical covariate-only models (grey). Lines connecting points (light grey) represent score pairs from same train-test split on the data. Significance was calculated using a two-sided Wilcoxon signed-rank test and p-values were adjusted for multiple testing using the Benjamini-Hochberg method with adjusted p-values shown at top of violin plots. Model score density distributions (middle left panels) between microbial abundance (blue), gene expression (orange), and combined data type (purple) models. Mean scores shown in red. Significance was calculated using a two-sided Dunn test and p-values were adjusted for multiple testing using the Benjamini-Hochberg method. **(a)** Results from prognosis models where combining data types add to predictive power. Time-dependent, cumulative/dynamic AUCs (right panels) for combined data type (purple), microbial abundance (blue), and gene expression (orange) models following years after diagnosis. Mean AUCs across entire time scales shown as a horizontal dotted line and in legends and shaded areas denote standard deviations. **(b)** Results from drug response models where combining data types adds to predictive power. Mean ROC and PR curves (right panels) for combined data type (purple), gene expression (orange), and microbial abundance (blue) models. Mean AUROC and AUPRC scores shown in panel legends and shaded areas denote standard deviations.