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. However, RNA-seq and whole genome sequencing data contain not only reads from host tumor and normal tissue, but also 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 gene 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, similarly as predictive of drug response, but in mostly different cancer-drug combinations. Thus, it appears possible to leverage existing sequencing technology, or develop new protocols, to obtain more non-redundant information about prognosis and drug response from RNA-seq and whole genome sequencing experiments than could be obtained from gene expression or mutation data alone.

Introduction

The Cancer Genome Atlas (TCGA), available from the NCI Genomic Data Commons[1], provides RNA-seq and whole genomic sequencing (WGS) data for thousands of cases across dozens of cancer types. RNA-seq data is typically used to measure the expression of human genes, and there is a long history linking tumor gene expression to cancer outcomes[2-8]. Milanez-Almeida et al.[9] recently showed that gene expression from 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. Importantly, Milanez-Almeida et al. used a data-driven machine learning (ML) based approach that selected features that were predictive of and correlated with prognosis, rather than an approach based on classical statistics or biological knowledge that chose features *a priori*. Though RNA-seq and WGS data not only

contain human sequencing reads, but also reads from the local microbiome that are typically filtered out from the data when analyzing human gene expression. Our central research questions then were, 1) does a data-driven ML approach reveal that tumor microbial abundances in TCGA data, quantified from these reads, are predictive of cancer prognosis or drug response, 2) what microbial genera are potentially predictive biomarkers of prognosis or drug response, 3) how do these models compare to equivalent models based on tumor gene expression data, and 4) does combining both microbial abundance and gene expression features produce models and select combinations of gene and microbial genera features that are more predictive of prognosis or drug response than models from each individual data type?

Poore et al.[10] recently published estimated, decontaminated, and normalized microbial abundances for the TCGA cohort, derived from either WGS or RNA-seq data. We used these abundances, taken directly from the Poore et al. dataset, to build predictive models of prognosis and drug response for TCGA. As a positive control, we also showed that our prognosis ML modeling methods, which differed somewhat from Milanez-Almeida et al.[9], identified a similar set cancers and outcomes for which gene expression was predictive of prognosis.

Results

Tumor microbial abundances are substantially less predictive of prognosis than gene expression

We built OS and PFI gene expression ML models of 32 TCGA tumor types (cohort information in **Supplementary Data 1**) using the Coxnet[11] algorithm, which jointly selects the most predictive subset of features via cross-validation (CV) while simultaneously being able to control for prognostic clinical covariates. In our models, we included and controlled for the clinical covariates age at diagnosis, gender, and tumor stage. For comparison, we also built

56 standard Cox regression models based on the clinical covariates alone. We evaluated the
57 predictive performance of our models using Harrell's concordance index (C-index), which is a
58 metric of survival model predictive accuracy. Each model analysis generated 100 model
59 instances and C-index scores from different randomly shuffled train-test CV splits on the data.
60 We found 33 OS and PFI models for 21 tumor types that had a mean C-index score ≥ 0.6 and
61 significantly outperformed their corresponding clinical covariate-only models (**Fig. 1a & c,**
62 **Supplementary Fig. 1a, 2a**). Our models were predictive of prognosis in 11 of the same 13
63 tumor types that were reported by Milanez-Almeida et al.[9] (**Supplementary Table 1**). We did
64 not analyze one tumor type that Milanez-Almeida did, acute myeloid leukemia (LAML), because
65 Poore et al. excluded it from their analysis. Among the cancers and outcomes that Milanez-
66 Almeida et al. analyzed, our methodology produced predictive models for four additional tumor
67 types: breast cancer (BRCA), cervical squamous cell carcinoma (CESC), sarcoma (SARC), and
68 uterine corpus endometrial carcinoma (UCEC), as well as quite a few predictive models for
69 additional cancers and outcomes that were not analyzed in their study (**Supplementary Table 1**).
70 We also evaluated prognosis model performance by calculating the time-dependent,
71 cumulative/dynamic area under the curve ($AUC^{C/D}(t)$)[12,13], which is an extension of the area
72 under the receiver operating characteristic (AUROC) for continuous outcomes and can provide a
73 more detailed resolution picture of predictive performance throughout the test outcome time
74 range compared to the C-index score. Although 33 of our OS and PFI gene expression models
75 had a statistically significant C-index score improvement compared to clinical covariates alone,
76 only 22 of these models showed an improvement in $AUC^{C/D}(t)$, where the improvement in mean
77 $AUC^{C/D}(t)$ over the entire test time range after diagnosis was ≥ 0.025 (**Supplementary Fig. 1b,**
78 **2b**).

79 We applied Coxnet[11] using the same methodology to build prognosis models using the
80 microbial abundance estimates provided by Poore et al.[10]. We found six microbial abundance
81 models that had a mean C-index score ≥ 0.6 and significantly outperformed their corresponding
82 clinical covariate-only models (**Fig. 1b & c, Supplementary Fig. 3a**). We found that in only two
83 of the six models, microbial abundances outperformed clinical covariates alone in terms of
84 $AUC^{C/D}(t)$, where the improvement in mean $AUC^{C/D}(t)$ over the entire test time range after
85 diagnosis was ≥ 0.025 (**Supplementary Fig. 3b**). In adrenocortical carcinoma (ACC), microbial
86 features predicted OS significantly better than clinical prognostic covariates starting at
87 approximately 6 years after diagnosis. In CESC, microbial abundances predicted OS better than
88 clinical covariates from approximately 6 months to 10 years after diagnosis. In contrast to a
89 recent report[14] on intratumor microbes predicting prognosis in TCGA thyroid carcinoma
90 (THCA), we did not find that Poore et al. tumor microbial abundances estimated from TCGA
91 were predictive of OS or PFI using our computational methods. A few possible reasons for this
92 are that Gnanasekar et al. analyzed the THCA cohort by tumor subtype and they applied different
93 methods, including 1) they used harmonized and normalized GDC TCGA data instead of legacy
94 TCGA followed by normalization and batch effect correction, 2) they applied different methods
95 for extraction of microbial reads and decontamination compared to Poore et al., and 3) they used
96 classical statistics on all of their data to infer correlations between features and prognosis instead
97 of a regularized ML model approach focused on fitting models on training data that have the
98 potential to generalize better and discover features that are predictive of prognosis on unseen
99 data. Overall, we found that tumor microbial abundances from Poore et al. were only marginally
100 predictive of prognosis across the TCGA cohort, and that gene expression was a significantly
101 more powerful predictor of prognosis (**Fig. 1, Supplementary Figs. 1-3**).

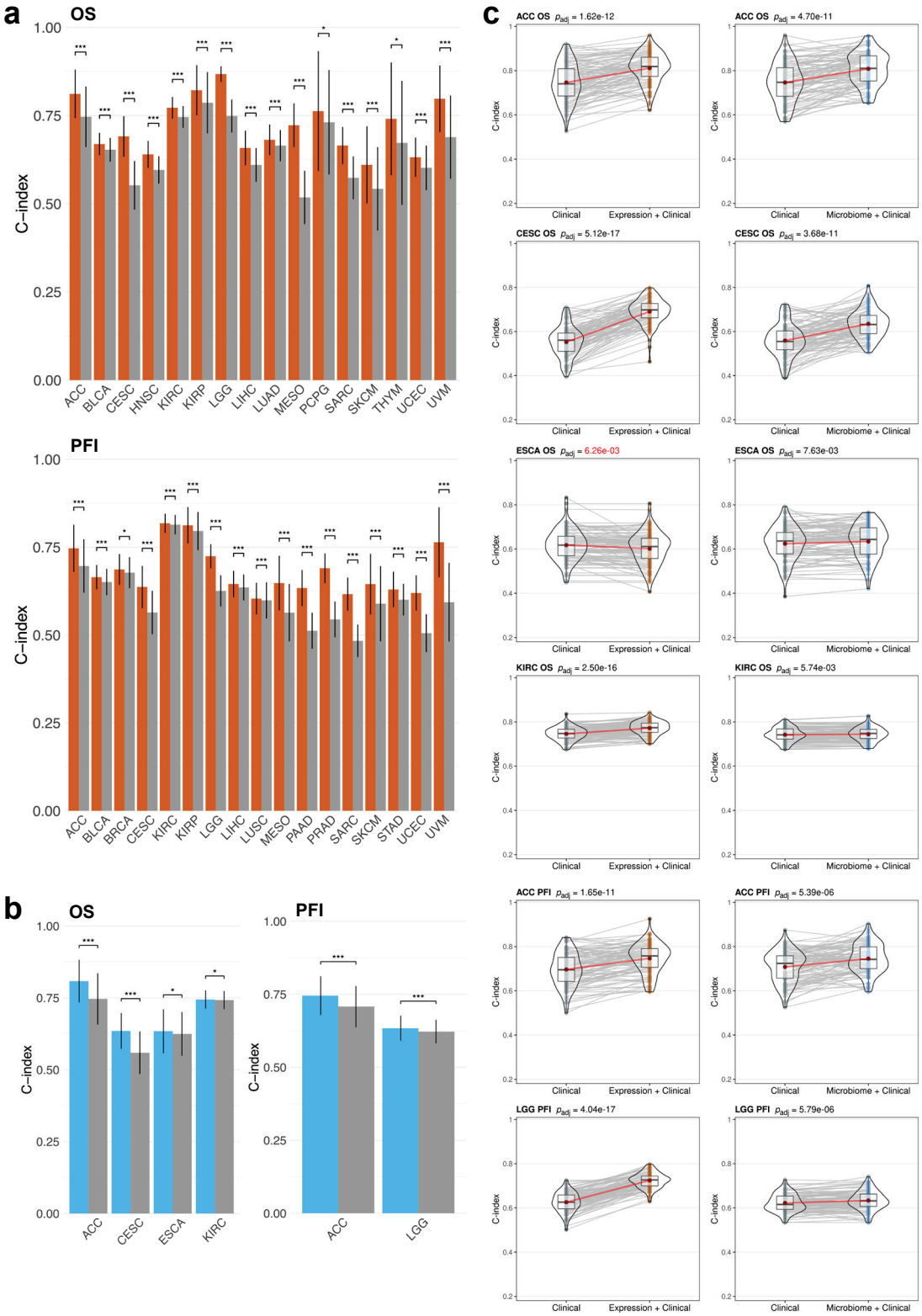


Figure 1. Performance of gene expression and microbial abundance prognosis prediction models where features add predictive power to clinical covariates. Mean C-index scores for **(a)** gene expression with clinical covariate models (orange) and **(b)** microbial abundance with clinical covariate models (blue) vs clinical covariate-only models (grey). Error bars denote standard deviations. Significance: * ≤ 0.01, ** ≤ 0.001, *** ≤ 0.0001. **(c)** C-index score violin density plots for the six 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 shown as red dots with red lines connecting the means. 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. Adjusted p-values colored in red signify difference where clinical covariate-only model is better.

113 **Tumor microbial abundances are predictive of chemotherapy drug response in some**
114 **cancers and in mostly different cancer-drug combinations than gene expression**

115 We next asked whether tumor microbial abundances from pre-treatment biopsies could
116 predict drug response better than the clinical covariates age at diagnosis, gender and tumor stage
117 alone. Drug response data were obtained from the TCGA Research Network as described in
118 Methods (cohorts shown in **Supplementary Data 1**). Cases with complete response (CR) or
119 partial response (PR) were labeled as responders and those with stable disease (SD) or
120 progressive disease (PD) as non-responders. Thirty TCGA cancer-drug combinations met our
121 minimum dataset size thresholds. We built drug response models using three different ML
122 methods: 1) a variant of the linear support vector machine recursive feature elimination (SVM-
123 RFE) algorithm[15] that we developed, 2) logistic regression (LGR) with elastic net[16] (L1 + L2)
124 penalties and embedded feature selection, and 3) logistic regression with an L2 penalty and
125 limma[17] (for tumor microbial and combination datasets) or edgeR[18,19] (for RNA-seq count
126 datasets) differential abundance/expression feature scoring and selection methods (see Methods
127 for details). All three ML methods unconditionally included the clinical covariates – age at
128 diagnosis, gender, and tumor stage – in the model (bypassing feature selection) while selecting
129 the most predictive subset of microbial abundance or gene expression features. For comparison,
130 we built standard linear SVM or LGR models using the clinical covariates alone. We evaluated
131 the predictive performance of drug response models using AUROC. Each analysis generated 100
132 model instances, AUROC, and area under the precision-recall curve (AUPRC) scores from
133 different random train-test CV splits.

134 We found five cancer-drug microbial abundance models had a mean AUROC score ≥ 0.6 and
135 performed better than clinical covariates alone in at least two out of three ML methods (**Fig. 2**).
136 Three of these cancer-drug combinations involved stomach adenocarcinoma (STAD). We

137 performed the same drug response modeling using TCGA gene expression data and here we
138 found six cancer-drug combinations had a mean AUROC score ≥ 0.6 and significantly
139 outperformed their corresponding clinical covariate-only models in at least two out of three ML
140 methods (**Fig. 3**). Only one cancer-drug combination, SARC docetaxel, overlapped between the
141 microbial abundance and gene expression drug response model results, suggesting that tumor
142 microbial abundances have independent predictive power. Overall, our results support the notion
143 that the tumor microbiome may contain information that is predictive of drug response in some
144 cancers, consistent with recent reports[20,21]. We also found there was considerable overlap in the
145 selected microbial abundance and gene expression features reported by each ML method (**Fig. 4a**
146 **& c**). We also frequently found a significant correlation between the feature importance rankings
147 reported by each ML method when comparing the two most significant methods in each cancer-
148 drug combination (**Fig. 4b & d**).

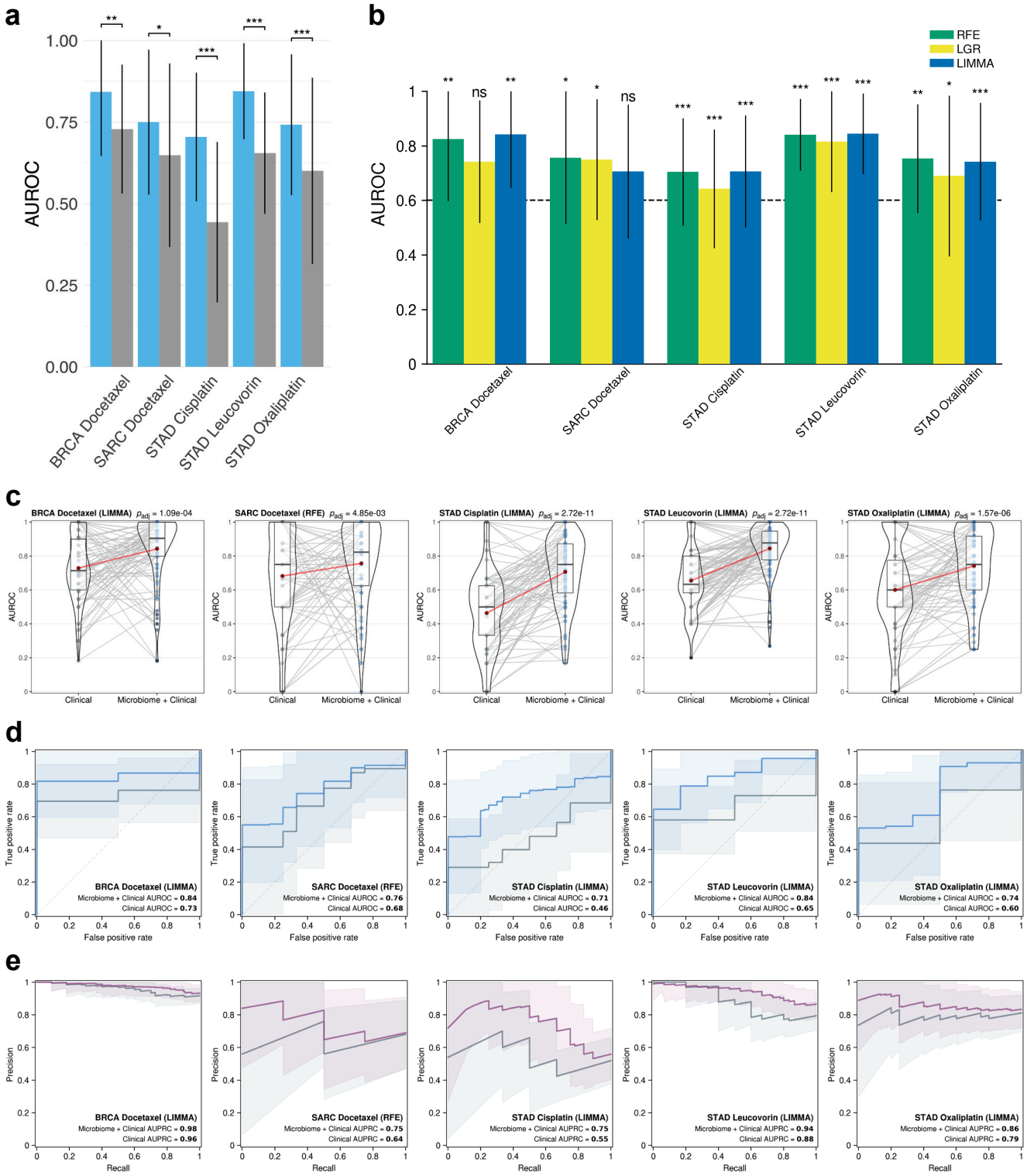


Figure 2. Performance of microbial abundance drug response prediction models in the five cancer-drug combinations where models performed better than clinical covariates alone. (a) Mean AUROC scores for microbial abundance with clinical covariate models (blue) vs clinical covariate-only models (grey) and **(b)** mean AUROC scores for each ML method. In both **(a)** and **(b)** error bars denote standard deviations. Significance: * ≤ 0.01, ** ≤ 0.001, *** ≤ 0.0001. **(c)** Violin density plots of AUROC scores (top row) for microbial abundance with clinical covariate models vs clinical covariate-only models. Lines connecting points (light grey) represent score pairs from same train-test split on the data. Mean AUROC scores are shown as red dots connected by red lines. **(d)** Mean ROC (middle row, blue) and **(e)** precision-recall (PR) curves (bottom row, purple) for microbial abundance with clinical covariate models vs clinical covariate-only models (grey). Mean AUROC and AUPRC scores shown in legends and shaded areas denote standard deviations. 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 of violin plots in **(c)**.

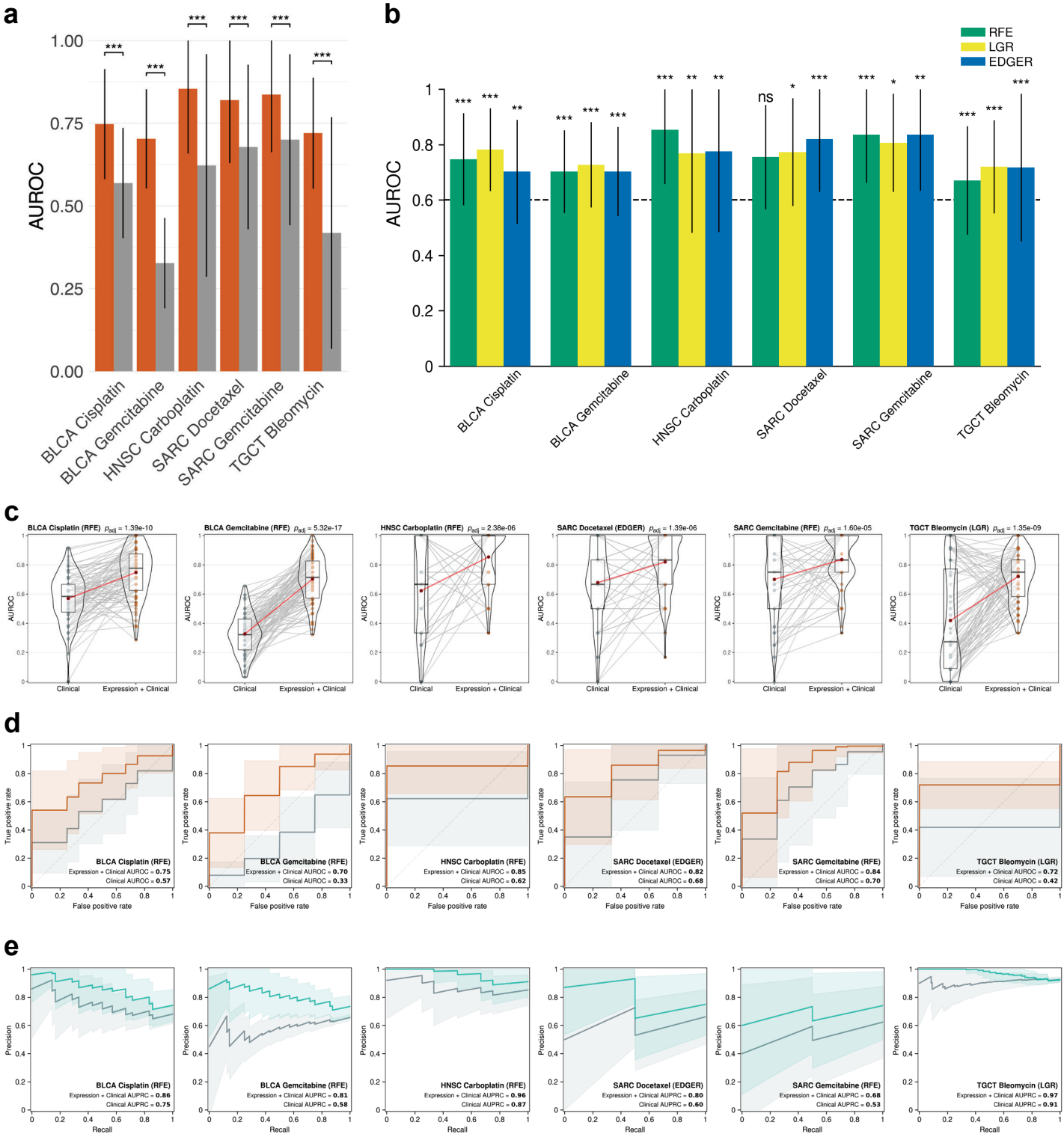


Figure 3. Performance of gene expression drug response prediction models in the six cancer-drug combinations where models performed better than clinical covariates alone. (a) Mean AUROC scores for gene expression with clinical covariate models (orange) vs clinical covariate-only models (grey) and **(b)** mean AUROC scores for each ML method. In both **(a)** and **(b)** error bars denote standard deviations. Significance: * ≤ 0.01, ** ≤ 0.001, *** ≤ 0.0001. **(c)** Violin density plots of AUROC scores (top row) for gene expression with clinical covariate models vs clinical covariate-only models. Lines connecting points (light grey) represent score pairs from same train-test split on the data. Mean AUROC scores are shown as red dots connected by red lines. **(d)** Mean ROC (middle row, orange) and **(e)** precision-recall (PR) curves (bottom row, green) for gene expression with clinical covariate models vs clinical covariate-only models (grey). Mean AUROC and AUPRC scores shown in legends and shaded areas denote standard deviations. 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 of violin plots in **(c)**.

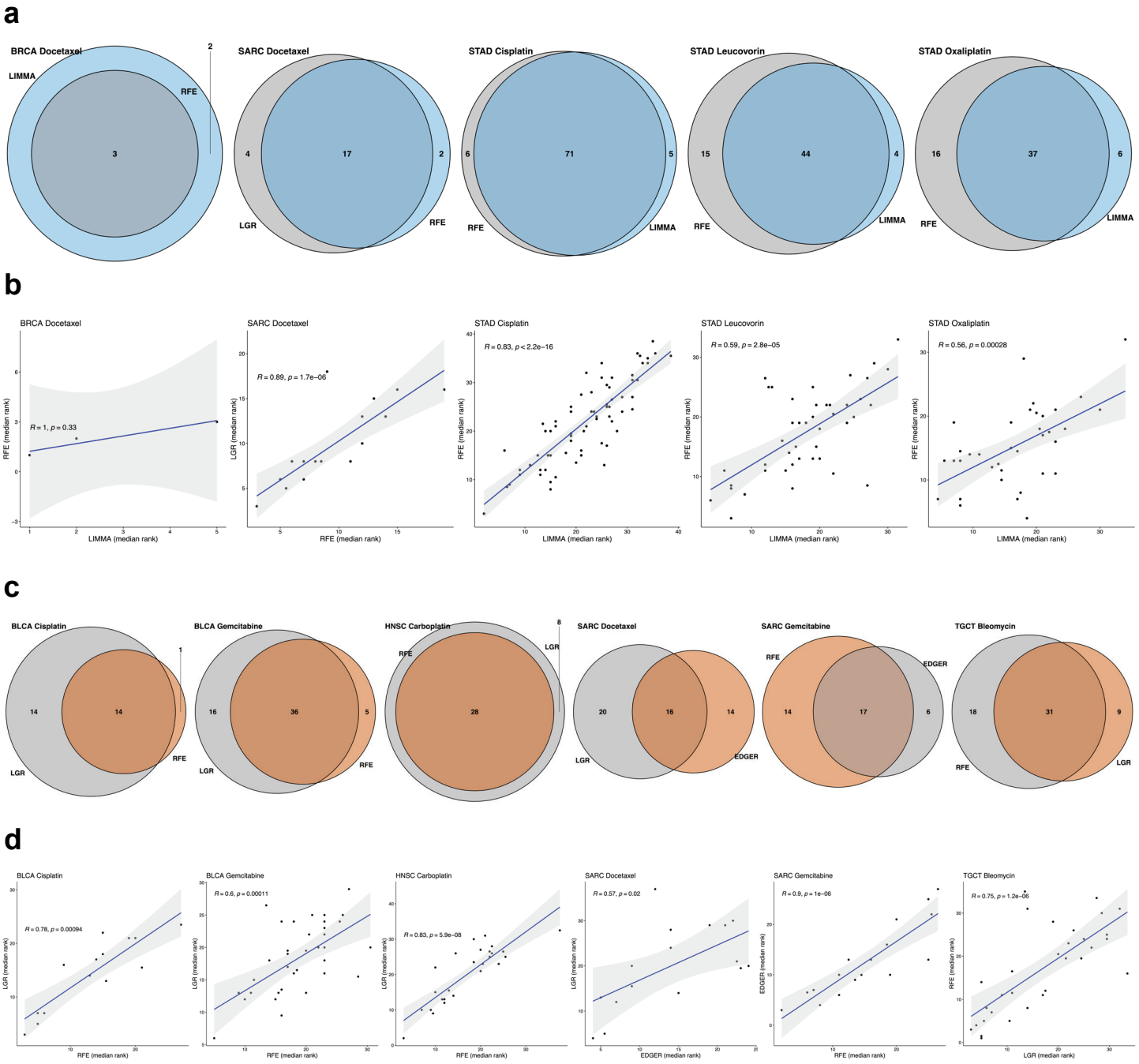


Figure 4. Comparison of top-ranked selected drug response model features by each ML method. For each drug response model, we selected the two best ML methods by significance for the prediction improvement over their respective clinical covariate-only model. **(a, c)** Venn diagrams for microbiome **(a)** or gene expression **(c)** models comparing the number of features individually selected by each ML method, and the intersection of the two ML methods. **(b, d)** Spearman rank correlation plots microbiome **(b)** or gene expression **(d)** models showing that the median rank of features (among the 100 model instances in which the feature was selected) often correlated between the two most significant ML methods. The best method is shown on the x-axis, the second best on the y-axis.

180 **Combining tumor microbial abundance and gene expression features add a modest**
181 **predictive improvement in some cancers**

182 Finally, we investigated if models built from combining microbial abundance and gene
183 expression features would result in an improvement in predictive power over their corresponding
184 single data type models. Combining data types resulted in a modest predictive improvement in
185 only three prognosis models: SARC OS, STAD PFI, and THYM OS (**Supplementary Fig. 4a**).
186 Although this improvement was not statistically significant in terms of C-index score, the
187 $AUC^{C/D}(t)$ metric showed a clear improvement in prognostic predictive power for the combined
188 data type models, where the improvement in mean $AUC^{C/D}(t)$ over the entire time range after
189 diagnosis was ≥ 0.025 compared to their respective single data type models. We also found five
190 combined data type drug response models which performed significantly better than clinical
191 covariates alone, although none of these models reached statistical significance when compared
192 to the their respective single data type model in terms of improvement in AUROC score, but one
193 of these models, for BLCA cisplatin, did show an improvement in AUROC ≥ 0.025 compared to
194 its corresponding single data type models (**Supplementary Fig. 4b-c**).

195 **Feature analysis reveals a wide range of predictive microbial genera**

196 To learn more about the most predictive features, we determined the top microbial genera
197 and top genes (**Supplementary Data 2**) selected by each of the significantly predictive microbial
198 abundance and gene expression models, respectively, according to their selection frequency and
199 model coefficients across the 100 model instances from each analysis. There were 428 distinct
200 microbial genera appearing in at least one prognosis or drug response model. Of these 428
201 genera, 160 were individually significantly predictive of prognosis or drug response by a
202 Wilcoxon test, indicating that the other genera were significantly predictive in combination. The
203 median number of genera selected per model was 52, with a minimum of 3 (BRCA docetaxel)

204 and a maximum of 78 (STAD cisplatin). Of the 428 genera, 95 were selected in more than one
205 model and only 13 were selected in more than two models. This is consistent with the
206 observation of Nejman et al.[22] that the tumor microbiome is tumor type specific. The predictive
207 genera we found span all non-eukaryotic domains of life, in total encompassing 365 bacterial, 17
208 archaeal, and 46 viral genera (**Supplementary Data 2**).

209 Discussion

210 In summary, we find that the microbial abundance estimates generated by Poore et al.[10] are
211 predictive of cancer patient prognosis and response to chemotherapy in a subset of tumor types,
212 survival outcomes, and treatments. Machine learning methods, such as those applied in this
213 study, are not able to infer causality, but only inform on the positive or negative predictive
214 associations covariates have with the response variable. The potential causal role that those may
215 play in determining patient prognosis or drug response can only be ascertained via dedicated
216 mechanistic studies. Overall, in terms of the number of significant models, based on their cross-
217 validated C-index or AUROC scores and improvement over clinical covariates alone, the tumor
218 microbiome is considerably less predictive than the tumor human transcriptome at predicting
219 patient prognosis, but notably, performs similarly to gene expression at predicting chemotherapy
220 response and in mostly different cancer-drug combinations. Our investigation motivates future
221 studies investigating the role of the tumor microbiome in predicting the response to targeted
222 therapies and immunotherapies.

223 There are also limitations to our current study. With respect to the drug response models, the
224 relatively small cohort size of most TCGA cancers where drug response data were available and
225 also the often small responder or non-responder case numbers within these cohorts does have an
226 impact on the interpretability of the results (see **Supplementary Data 1** for cohort size

227 information). We have attempted to overcome this limitation as best as possible by applying
228 robust and advanced ML techniques at every level of model building and evaluation, including
229 fitting and evaluating 100 randomly shuffled model instances, performing an exhaustive grid
230 search for hyperparameter optimization and model selection, using a nested cross-validation
231 (CV) strategy, utilizing advanced stratified and replicate grouping CV iterators, and applying
232 both sample and class weighting to the learning algorithm and scoring methods (see Methods for
233 full details).

234 There are further limitations to this study inherited from limitations in the data, originally
235 raised by Poore et al.[10] First, the study was retrospective, using existing data from the TCGA. As
236 such, it did not involve any specific protocols to capture microbial reads or to control for
237 contamination. Second, decontamination of such retrospective data is a highly involved and
238 dataset-specific process, which they made great effort to validate. Poore et al. conclude from this
239 validation that the retrospective study of TCGA was successful, and that similar retrospective
240 studies would be valuable. A third point, which they touch on briefly, is that the protocols that
241 were used have limitations with respect to capturing microbial reads and cannot distinguish if the
242 source of microbial reads is intracellular or extracellular, or alive or dead when the sample was
243 taken. Poore et al. suggest, correctly we believe, that additional protocols need to be developed
244 for prospective studies.

245 Accepting the limitations of the study, we observed certain trends. Proteobacteria and
246 Firmicutes were the most frequent phyla identified as predictive features (**Supplementary Data**
247 **2**), followed by Actinobacteria and Bacteroidetes. Among viruses, Herpesvirales were the most
248 frequent. More microbial genera were negatively predictive of drug response or prognosis than
249 those positively predictive (negative for 306/537 features; two-sided binomial test p-value =

250 0.0014). Firmicutes reversed this trend, being more often positively predictive (positive for 49/82
251 features, two-sided Fisher's exact test p-value = 0. 0.0036; **Supplementary Table 2**).

252 Further examining the predictive features of specific cancer types, we noted that several
253 genera of Firmicutes were predictive of OS in CESC, including genera of *Lactobacillales*.
254 Notably, though CESC is known to often arise from HPV infection, the presence of other
255 microbial species, in particular the Firmicutes genus *Lactobacillus* has been reported in the
256 literature to be positively associated with the risk of developing CESC[23].

257 In STAD, tumor microbial abundances were predictive of response to three different drugs:
258 cisplatin, leucovorin, and oxaliplatin. The genus *Helicobacter* was a quantified microbial
259 abundance feature in the Poore et al. dataset although notably, even though it is well established
260 that patients infected with *H. pylori* have an increased risk of developing gastric cancer[24],
261 *Helicobacter* was not identified as a predictive feature of drug response in our STAD models.
262 This findng is in line with recent research indicating reduced microbial diversity, decreased
263 abundance of *H. pylori*, and enrichment of other mostly commensal bacterial genera in gastric
264 carcinoma[25]. Instead, in STAD we found that known opportunistic bacteria *Cedecea* and
265 *Sphingobacterium* were both strongly negatively predictive of leucovorin response,
266 *Sphingobacterium* was strongly negatively predictive of cisplatin response, and the opportunistic
267 bacteria *Rouxiella* was strongly negative predictive of oxaliplatin response. *Cedecea* and
268 *Sphingobacterium* have been implicated in bacteremia in immunocompromised individuals in
269 rare cases, including cancer[26,27,28,29]. As dysbiosis is frequent in stomach cancer[30,31], and in light
270 of the mechanism of action of leucovorin, it may be of interest to study whether organisms from
271 these two genera may sequester or prevent the bacterial production of folinic acid[32].

272 We found three microbial genera whose abundances were strongly associated with breast
273 cancer response to doxetaxel. Indeed, the involvement of the microbiome in breast cancer
274 (BRCA)[22,33] has recently received considerable attention. In BRCA, we found that the genus
275 containing Epstein-Barr virus (EBV) was negatively associated with response to docetaxel,
276 which is concordant with previous findings that EBV is associated with chemoresistance to
277 docetaxel in gastric cancer[34]. Interestingly, Cyanobacteria were predictive features in several
278 cancers in our study and we identified a genus of Cyanobacteria as predictive of response to
279 docetaxel in BRCA. Notably, the presence of Cyanobacteria in BRCA was recently confirmed
280 by Nejman et al.[22] by 16S-rRNA sequencing. While the genus we identified, *Raphidiopsis*, a
281 planktonic Cyanobacteria that produces toxins harmful to human health and found in freshwater,
282 is possibly a taxonomic identification error in the original microbial abundance estimates, our
283 findings may point to a related genus under the recently discovered clade Melainabacteria of
284 Cyanobacteria[35], which is present in humans. Though Melainabacteria are difficult to culture, we
285 believe that confirmation of the relationship between BRCA to response to docetaxel and
286 Melainabacteria should be tested, and a first step would be to confirm our computationally
287 derived findings in a dedicated 16S-rRNA analysis.

288 Interestingly, in sarcoma (SARC), among the most predictive microbial features we found
289 the genus *Lactococcus* to be positively associated with response to docetaxel. *Lactococcus*
290 contains species that can sometimes cause opportunistic infections in humans, as *Lactococcus*
291 are similar to *Streptococcus* and formerly belonged to that genus. The result that this genus was
292 positively associated with response in our model initially appeared counterintuitive, although
293 while the use of therapeutic bacteria as antitumor agents has not been an extensively studied
294 field, there have been some limited findings in the literature that suggest the use of
295 bacteriotherapy as anticancer agents[36]. Historically, the intentional use of the toxins of various

296 *Streptococcus* species showing significant antitumor activity in SARC has been
297 documented[37,38,39]. One possible testable explanation for some microbes being strongly positive
298 predictive of docetaxel response in our model is that they might produce some extracellular
299 products or toxins that could work as an adjuvant to the chemotherapy.

300 In summary, while these findings and others reported in the paper are computationally
301 derived associations, we believe that they can serve as leads for further experimental studies of
302 the role of microbial species in modulating patients survival and drug response, potentially by
303 metabolizing drug levels in the tumor microenvironment as suggested above, or by altering the
304 immune response, either by changing the levels of specific immunometabolites or by having the
305 tumors present specific bacterial antigens[40].

306 Methods

307 Data retrieval and processing

308 Normalized and batch effect corrected microbial abundance data for 32 TCGA tumor types
309 were downloaded from the online data repository referenced in Poore et al.[10]
310 (ftp://ftp.microbio.me/pub/cancer_microbiome_analysis). Specifically, the "Kraken-TCGA-
311 Voom-SNM-Plate-Center-Filtering-Data.csv" microbial abundance data file and adjoining
312 "Metadata-TCGA-Kraken-17625-Samples.csv" metadata file were used as the starting input for
313 further data processing.

314 We first filtered the data for primary tumor samples (TCGA "Primary Tumor" or "Additional
315 - New Primary" sample types). Poore et al. generated microbial abundances from all the
316 available WGS and RNA-seq data in legacy TCGA (after some quality filters), which frequently
317 contained replicate WGS and RNA-seq data for the same case and sample type. It was common

318 in legacy TCGA to increase WGS sequencing coverage by performing an additional sequencing
319 run from the same sample and these secondary runs typically had a much lower number of reads
320 and coverage compared to their corresponding primary sequencing runs. When comparing the
321 normalized and batch effect corrected read counts between these WGS runs, we found that
322 microbial abundance data which came from lower coverage secondary runs could be
323 substantially different from abundances derived from the larger primary sequencing runs.
324 Therefore, we excluded microbial abundance data which came from secondary runs. In addition,
325 legacy TCGA commonly contained data for the same samples analyzed using different
326 computational pipeline versions. We excluded replicate microbial abundance data from older
327 TCGA analysis pipeline versions if a replicate from a newer version existed. After the above
328 filters, the Poore et al. data went from 17,625 samples and 10,183 unique cases to 12,111
329 samples and 9,812 unique cases (comprising of 1,944 WGS samples from 1,904 unique cases
330 and 10,167 RNA-seq samples from 9,745 unique cases).

331 TCGA gender, age at diagnosis, and tumor stage demographic and clinical data and as well
332 as primary tumor RNA-seq read count data for the 32 TCGA tumor types included in our study
333 were obtained from the NCI Genomic Data Commons (GDC Data Release v29.0) using the R
334 Bioconductor package GenomicDataCommons. TCGA GENCODE v22 gene annotations were
335 obtained from the GDC data portal and Ensembl Gene v98 using the R package rtracklayer and
336 R Bioconductor packages AnnotationHub and ensembldb. The downloaded GDC primary tumor
337 cohort with RNA-seq read count data comprised of 9,735 samples from 9,680 unique cases.
338 There were 68 cases at the GDC which had missing age of diagnosis but existing values in the
339 Poore et al. data and we chose not to exclude these data and used the Poore et al. age of diagnosis
340 values for these cases. TCGA curated survival phenotypic data[41] were obtained from UCSC

341 Xena. Cases which had both missing overall survival (OS) and progressive-free interval (PFI)
342 outcome data were excluded from survival modeling.

343 Drug response data were compiled from the TCGA Research Network. Our drug response
344 models used the following binary classification targets: complete response (CR) and partial
345 response (PR) were labeled as responders and stable disease (SD) and progressive disease (PD)
346 as non-responders. All TCGA samples with drug response phenotypic data were from pre-
347 treatment biopsies. Due to the limited cancer-drug combination cohort sizes in TCGA, we
348 modeled each drug individually, even if a patient received multiple drugs concurrently. If the
349 same drug was given at multiple timepoints to a patient, we only considered their first drug
350 response. We considered cancer-drug combinations that contained a minimum of 18 cases and at
351 least 4 cases per response binary class, except for STAD oxaliplatin, where we allowed a
352 minimum of 14 cases so that the gene expression dataset could be included. In total, we analyzed
353 30 cancer-drug combinations which had paired microbial abundance and gene expression data
354 that met the above thresholds. Combined feature microbial abundance and gene expression
355 datasets were created by joining data from each individual dataset which had matching TCGA
356 sample UUIDs. For some TCGA cases, data existed from multiple different aliquots per sample
357 or multiple technical runs per aliquot, therefore in these cases all combinations were joined at the
358 sample UUID level. Cross-validation sampling probability weights as well as model and scoring
359 sample weights were applied to account and adjust for any imbalance caused by the process.

360 **ML modeling**

361 Machine learning (ML) models were built using the scikit-learn[42] and scikit-survival
362 libraries[43,44,45]. Custom extensions to scikit-learn and scikit-survival were developed to add new
363 methods and functionalities required by this project. Survival models were built using Coxnet –

364 regularized Cox regression with elastic net penalties[11]. Coxnet models controlled for gender, age
365 at diagnosis, and tumor stage clinical prognostic covariates by including them as unpenalized
366 features in the model (Coxnet penalty factor = 0). Drug response classification models were built
367 using three different ML methods: 1) a variant of the linear support vector machine recursive
368 feature elimination (SVM-RFE) algorithm[15] that we developed with a number of additional
369 features and better performance than the scikit-learn built-in version, 2) logistic regression
370 (LGR) with elastic net[16] (L1 + L2) penalties and embedded feature selection, and 3) LGR with
371 an L2 penalty and limma[17] (for tumor microbial and combination datasets) or edgeR[18,19] (for
372 RNA-seq count datasets) differential abundance/expression feature scoring inside a k-best
373 wrapper feature selection method around the learning algorithm. limma differential abundance
374 analysis was run inside the ML pipeline with default parameters except for fitting an intensity-
375 dependent trend to the prior variances and running a robust empirical Bayes procedure (eBayes
376 function parameters trend = TRUE and robust = TRUE). edgeR differential expression analysis
377 was run inside the ML pipeline with default parameters except for enabling robust estimation of
378 the negative binomial dispersion (calcDispersions function robust = TRUE) and robust
379 estimation of the prior quasi-likelihood (QL) dispersion (glmQLFit function robust = TRUE).
380 Both limma and edgeR methods scored and ranked features by differential abundance/expression
381 p-value.

382 All three drug response ML methods unconditionally included the same three clinical
383 covariates in the model as in the prognosis models by having them bypass feature selection in the
384 ML pipeline, thought in drug response models clinical covariates where modeled as L2 penalized
385 features. In SVM-RFE, clinical covariate features bypassed recursive feature elimination but
386 were always included at each RFE recursive feature elimination model fitting step as well as
387 final model refitting. To the best of our knowledge, no available comprehensive ML library in

388 python or R currently provides an elastic net LGR algorithm with the functionality to specify
389 features that can bypass embedded feature selection and be modeled with an L2 penalty (setting
390 the R glmnet penalty factor, for example, does not provide this functionality as it isn't a penalty
391 factor per regularization term but a factor applied to the sum of both L1 and L2 terms). In order
392 to develop this functionality for our study, our elastic net LGR model pipeline was designed as a
393 two-level LGR, 1) an elastic net LGR and embedded feature selection on only microbial
394 abundance or gene expression features with clinical covariates bypassing this step, followed by
395 2) an L2 penalized LGR on features selected by the elastic net LGR step and the clinical
396 covariates. We know this design doesn't likely produce the exact same model settings and results
397 of a single-level elastic net LGR algorithm with the functionality we needed, if such a
398 implementation it existed, though we tested every drug response model through an ML pipeline
399 with elastic net LGR and no clinical feature selection bypass and found that model predictive
400 performance, feature coefficients and signs, and feature importance rankings were similar to our
401 two-level ML pipeline setup.

402 Gender was one-hot encoded and tumor stage ordinal encoded by major stage. In the final
403 cohort included in our prognosis and drug response models, 3363 out of 9708 tumor microbial
404 abundance cases (34.64%) and 3244 out of 9484 gene expression cases (34.21%) had tumor
405 stage "not reported" or BRCA stage "X". Since missing tumor stage metadata is so prevalent in
406 TCGA, we took the approach of including these in our study and modeled missing tumor stage
407 with as neutral an ordinal encoding as possible. Looking at the distribution of reported major
408 tumor stages in our cohort, we determined that encoding missing data as an ordinal between
409 tumor stage II and III was as close to the middle of the distribution of stages in TCGA as we
410 could possibly achieve with ordinal encoding.

411 All prognosis and drug response models included the previously described feature selection
412 as well as normalization and transformation steps integrated into the ML modeling pipeline using
413 an extended version of the scikit-learn Pipeline framework. Training data splits always
414 underwent feature selection, normalization, and transformation through the ML pipeline
415 independently from held-out test and validation data splits before learning. Models built using
416 gene expression read count data included edgeR low count filtering, weighted trimmed mean of
417 M-values (TMM) normalization, and log counts per million (CPM) transformation steps within
418 the ML pipeline. These were developed and integrated into our scikit-learn-based framework via
419 R and rpy2. All models also included standardization of features within the ML pipeline just
420 before learning. During prediction, held-out test or validation data were feature selected,
421 normalized and transformed through the ML pipeline using the parameters learned from the
422 training data at each pipeline step before model prediction and scoring. Each cancer, data type,
423 and survival or drug response target type combination was modeled individually using a nested
424 cross-validation (CV) strategy to perform model selection and evaluation on held-out test data.
425 All cross-validation iterators kept replicate sample data per case grouped together such that data
426 would only reside in either the train or test split during each CV iteration.

427 Survival models used a stratified, randomly shuffled outer CV with 75% train and 25% test
428 split sizes that was repeated 100 times. The CV procedure stratified the splits on event status.
429 Each training set from the outer CV was used to perform hyperparameter tuning and model
430 selection by optimizing C-index over a stratified, randomly shuffled, 4-fold inner CV on the
431 training set, repeated 5 times. A few cancer datasets contained fewer than four uncensored cases
432 which required reducing the number of inner CV folds for these models such that at least one
433 case per fold was uncensored. The data derived from Poore et al. often included more than one
434 sample per case, and an unequal number of samples between cases, therefore requiring either ML

435 model sample weighting or CV random sampling per case. The Coxnet implementation in scikit-
436 survival does not currently support sample weights, therefore our custom outer CV iterator
437 randomly sampled one replicate sample per case during each iteration, using a sampling
438 procedure with probability weights that balanced the probability that a replicate WGS- or RNA-
439 seq-based sample was selected during each CV iteration. Model selection grid search was
440 performed on the following hyperparameters: elastic net penalty L1 ratios 0.1, 0.3, 0.5, 0.7, 0.8,
441 0.9, 0.95, 0.99, and 1, and for each L1 ratio a default alpha path of 100 alphas using an alpha min
442 ratio of 10^{-2}. Alpha is the constant multiplier of the penalty terms in the Coxnet objective
443 function. Optimal alpha and L1 ratio settings were determined via inner CV and a model with
444 these settings was then refit on the entire outer CV train data split. Model performance was
445 evaluated in both inner and outer CV on the held-out validation or test data split, respectively, by
446 generating model test predicted risk scores and using these scores to directly calculate a Harrell's
447 concordance index (C-index). We also evaluated and compared model predictive performance
448 for each test data split survival time period by calculating time-dependent cumulative/dynamic
449 AUCs[12,13].

450 Drug response models used a stratified, randomly shuffled, 4-fold outer CV that was repeated
451 25 times (i.e. 100 model instances). Each training set from the outer CV was used to perform
452 hyperparameter tuning and model selection by optimizing the area under receiver-operator curve
453 (AUROC) over a stratified, randomly shuffled, 3-fold inner CV repeated 5 times. Case replicate
454 sample weights were provided to SVM-RFE and LGR learning algorithms and all model
455 selection and evaluation scoring methods. Class weights were provided to SVM-RFE and LGR
456 learning algorithms to adjust for any class imbalance. Model selection grid search was performed
457 on the following hyperparameters: SVM and L2 penalized LGR C regularization parameter from
458 a range of 10^{-5} to 10^{3}, elastic net LGR L1 ratios of 0.1, 0.3, 0.5, 0.7, 0.8, 0.9, 0.95, 0.99 and 1,

459 elastic net LGR C regularization parameter from a range of 10^{-2} to 10^{3} (microbial abundance) or
460 from 10^{-2} to 10^{1} (gene expression), and finally RFE, elastic net LGR, and limma and edgeR
461 feature scorer k-best feature selection search range from 1 to 400 top scoring microbial
462 abundance or gene expression features. Optimized hyperparameter settings were determined via
463 inner CV and a model with the optimized settings was then refit on the entire outer CV train data
464 split. Model performance was evaluated in both inner and outer CV on the held-out validation or
465 test data split, respectively, by AUROC, area under precision-recall curve (AUPRC), and
466 balanced accuracy (BCR). AUROC was used to evaluate and select the best model and
467 optimized hyperparameter settings from the grid search. Average precision and BCR scores were
468 also reported for each model instance.

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

483 performed between all three groups of data type model scores. Each Dunn test raw p-value was
484 adjusted for multiple testing using the Benjamini-Hochberg (BH) procedure to control the false
485 discovery rate (FDR), and a threshold FDR ≤ 0.05 was used to determine statistical significance.

486 **Microbial abundance model feature analysis**

487 For each analysis, 100 prognosis or drug response model instances were generated from the
488 outer CV procedure. Each model instance selected a subset of features that performed best
489 during CV and the model algorithm learned coefficients (or weights) for each feature. To select
490 microbial genera for downstream investigation from the feature results across all these model
491 instances, we proceeded as follows. First, we applied a two-sided Wilcoxon signed-rank test that
492 the mean feature coefficient rank generated by the model is shifted away from zero, and thus that
493 the genus is identifiably positively or negatively associated with survival or drug response. For
494 all Wilcoxon tests, we used the package coin[46], which allows exact calculation of p-values.
495 Coefficients were ignored when a genus was assigned a zero coefficient or absent from a model.
496 Second, within each model, all coefficients, ignoring the results of the Wilcoxon test, were
497 ranked by absolute magnitude. We then kept genera that were among the top 50 features in at
498 least 20% of the models and for which the Holm-adjusted, two-sided Wilcoxon signed-rank test
499 p-value was ≤ 0.01. Having a Coxnet feature coefficient equal to zero or feature being absent
500 from an SVM-RFE or LGR model was not strong enough evidence that the genus has no effect,
501 but rather that one or more features with stronger effect were chosen. Thus, we ignored genera
502 with a zero coefficient or absent from a model when computing mean coefficient weight and
503 Wilcoxon statistics on the means.

504 For the drug response models, where three ML methods were tested, we noted the features
505 selected by individual models and the median rank the feature attained in the instances in which

506 it appeared, but further filtered the features to account for the consensus between ML models.
507 We kept features selected in any two ML model methods that individually met our criteria for
508 inclusion, ignoring features in ML models that did not meet these criteria. We then computed the
509 Spearman correlation between the median ranks attained by the features.

510 For each selected microbial feature, we tested whether it was a significantly univariate
511 feature of survival or drug response. This is a strictly different question than whether the
512 coefficient of a feature has consistent sign – sign may be consistent when used in combination
513 with other features, but the feature may not be individually predictive. For drug response models,
514 we divided individuals into responders and non-responders, and for survival data we divided
515 individuals whose survival time was greater or less than the censored median, ignoring those
516 who were lost to follow up before median time. For each cancer-test type pair, we applied a two-
517 sided Wilcoxon rank-sum test. We applied a Benjamini-Hochberg multiple hypothesis correction
518 for each cancer-test type pair and report the false discovery rate in **Supplementary Data 2**.

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

525 Data availability

526 All results generated from this work are available under https://doi.org/10.5281/zenodo.5221525.

527 Code availability

528 All code and data used to produce this work are available under

529 https://github.com/ruppinlab/tcga-microbiome-prediction.

530 References

531 1. Grossman, R. L. et al. Toward a Shared Vision for Cancer Genomic Data. *New England*
532 *Journal of Medicine* **375**, 1109–1112 (2016).

533 2. Ahluwalia, P., Kolhe, R. & Gahlay, G. K. The clinical relevance of gene expression based
534 prognostic signatures in colorectal cancer. *Biochimica et Biophysica Acta (BBA) - Reviews*
535 *on Cancer* **1875**, 188513 (2021).

536 3. Brodsky, A. S. *et al.* Expression profiling of primary and metastatic ovarian tumors reveals
537 differences indicative of aggressive disease. *PLoS One* **9**, e94476 (2014).

538 4. Liu, Y. *et al.* Pan-cancer analysis of clinical significance and associated molecular features of
539 glycolysis. *Bioengineered* **12**, 4233–4246 (2021).

540 5. Selfors, L. M., Stover, D. G., Harris, I. S., Brugge, J. S. & Coloff, J. L. Identification of
541 cancer genes that are independent of dominant proliferation and lineage programs. *Proc Natl*
542 *Acad Sci U S A* **114**, E11276–E11284 (2017).

543 6. Shimoni, Y. Association between expression of random gene sets and survival is evident in
544 multiple cancer types and may be explained by sub-classification. *PLoS Comput Biol* **14**,
545 e1006026 (2018).

546 7. Shukla, S. *et al.* Development of a RNA-Seq Based Prognostic Signature in Lung
547 Adenocarcinoma. *J Natl Cancer Inst* **109**, (2017).

548 8. Venet, D., Dumont, J. E. & Detours, V. Most random gene expression signatures are
549 significantly associated with breast cancer outcome. *PLoS Comput Biol* **7**, e1002240 (2011).

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

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

554 11. Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. Regularization Paths for Cox's
555 Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software* **39** (5),
556 1-13 (2011).

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

559 13. Lambert, J. & Chevret, S. Summary measure of discrimination in survival models based on
560 cumulative/dynamic time-dependent ROC curves. *Statistical methods in medical research* **25**
561 (5), 2088–2102 (2016).

562 14. Gnanasekar A. The intratumor microbiome predicts prognosis across gender and subtypes in
563 papillary thyroid carcinoma. *Comput Struct Biotechnol J* **19**, 1986-1997 (2021).

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

566 16. Zou H. & Hastie T. Regularization and variable selection via the elastic net. *J R Statist Soc B*
567 **67** (2), 301–320 (2005).

568 17. Ritchie, M.E. *et al. limma* powers differential expression analyses for RNA-sequencing and
569 microarray studies. *Nucleic Acids Res* **43** (7), e47 (2015).

570 18. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for
571 differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140
572 (2010).

573 19. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor
574 RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* **40**, 4288–4297
575 (2012).

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

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

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

582 23. Lin, D. *et al.* Microbiome factors in HPV-driven carcinogenesis and cancers. *PLoS Pathog*
583 **16**, (2020).

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

586 25. Ferreira R.M. et al. Gastric microbial community profiling reveals a dysbiotic cancer-
587 associated microbiota. *Gut* **67** (2): 226-236 (2018).

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

590 27. Akinosoglou, K. *et al.* Bacteraemia due to Cedecea davisae in a patient with sigmoid colon
591 cancer: a case report and brief review of the literature. *Diagn Microbiol Infect Dis* **74**, 303–
592 306 (2012).

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

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

597 30. Coker, O. O. *et al.* Mucosal microbiome dysbiosis in gastric carcinogenesis. *Gut* **67**, 1024–
598 1032 (2018).

599 31. Castaño-Rodriguez N. *et al*. Dysbiosis of the microbiome in gastric carcinogenesis. *Sci Rep* **7**
600 (1), 15957 (2015).

601 32. Ogwang, S. et al. Bacterial Conversion of Folinic Acid Is Required for Antifolate Resistance.
602 *Journal of Biological Chemistry* **286**, 15377–15390 (2011).

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

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

607 35. Di Rienzi, S. C. *et al.* The human gut and groundwater harbor non-photosynthetic bacteria
608 belonging to a new candidate phylum sibling to Cyanobacteria. *eLife* **2**, e01102 (2013).

609 36. Sedighi M. *et al*. Therapeutic bacteria to combat cancer; current advances, challenges, and
610 opportunities. *Cancer Med* **8** (6), 3167-3181 (2019).

611 37. Coley WB. Late results of the treatment of inoperable sarcoma by the mixed toxins of
612 erysipelas and bacillus prodigiosus. *Trans Southern Surg Gynecol Ass* **18**, 197 (1906).

613 38. Fehleisen F. Ueber die Züchtung der Erysipelkokken auf künstlichem Nährboden und ihre
614 Übertragbarkeit auf den Menschen. *Dtsch Med Wochenschr* **8**, 553-554 (1882).

615 39. Busch W. Aus der Sitzung der medicinischen Section vom 13 November 1867. *Berl Klin*
616 *Wochenschr* **5**, 137 (1868).

617 40. Kalaora S. *et al*. Identification of bacteria-derived HLA-bound peptides in melanoma. *Nature*
618 **592** (7852), 138-143 (2021).

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

621 42. Pedregosa *et al*. Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**, 2825-2830
622 (2011).

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

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

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

46. 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).

Acknowledgements

The results shown here are in whole or 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 and by NIH grant 1ZIABC011803-03. 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, Tim Miller, Wolfgand Resch, and Steve Fellini 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 information

Affiliations

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

Leandro Cruz Hermida, E. Michael Gertz & Eytan Ruppin

Department of Computer Science, University of Maryland, College Park, MD, USA

Leandro Cruz Hermida

654 **Contributions**

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

657 **Corresponding authors**

658 Correspondence to Eytan Ruppin.

659 Ethics declaration

660 **Competing interests**

661 All authors declare that they have no competing interests.

662

663 # Supplementary Tables

664 **Supplementary Table 1.** Prognosis model results comparison to Milanez-Almeida et al.

	Milanez-Almeida		**Hermida & Gertz**					
Cancer	**Tested**	**Hit**	**Tested**	**Hit**	**Common Hit**	**New Hit**	**M-A Tested New Hit**	**Missing M-A Hit**
ACC	OS	OS	OS, PFI	OS, PFI	OS	PFI		
BLCA	OS	OS	OS, PFI	OS, PFI	OS	PFI		
BRCA	PFI		OS, PFI	PFI		PFI	PFI	
CESC	OS		OS, PFI	OS, PFI		OS, PFI	OS	
CHOL	*EXCL*		OS, PFI					
COAD	OS		OS, PFI					
DLBC	*EXCL*		OS, PFI					
ESCA	OS		OS, PFI					
GBM	OS		OS, PFI					
HNSC	OS	OS	OS, PFI	OS	OS			
KICH	*EXCL*		OS, PFI					
KIRC	OS	OS	OS, PFI	OS, PFI	OS	PFI		
KIRP	OS	OS	OS, PFI	OS, PFI	OS	PFI		
LAML	OS	OS	*EXCL*					OS
LGG	PFI	PFI	OS, PFI	OS, PFI	PFI	OS		
LIHC	OS	OS	OS, PFI	OS, PFI	OS	PFI		
LUAD	OS	OS	OS, PFI	OS	OS			
LUSC	OS		OS, PFI	PFI		PFI		
MESO	OS	OS	OS, PFI	OS, PFI	OS	PFI		
OV	OS		OS, PFI					
PAAD	OS	OS	OS, PFI	PFI		PFI		OS
PCPG	*EXCL*		OS, PFI	OS		OS		
PRAD	PFI	PFI	OS, PFI	PFI	PFI			
READ	PFI		OS, PFI					
SARC	OS		OS, PFI	OS, PFI		OS, PFI	OS	
SKCM	*EXCL*		OS, PFI	OS, PFI		OS, PFI		

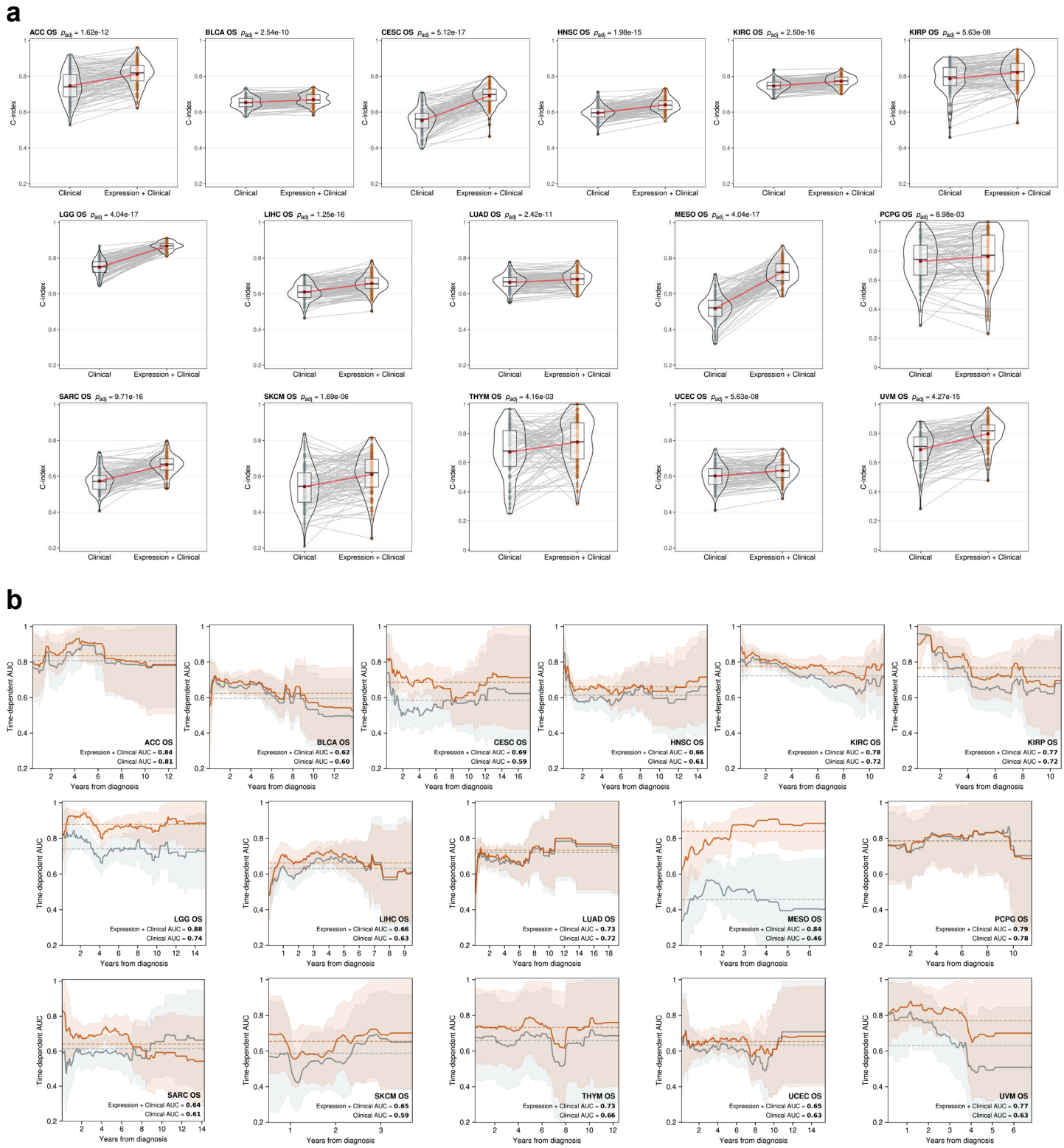
STAD	OS		OS, PFI	PFI		PFI		
TGCT	OS		OS, PFI					
THCA	PFI		OS, PFI					
THYM	PFI		OS, PFI	OS		OS		
UCEC	OS		OS, PFI	OS, PFI		OS, PFI	OS	
UCS	OS		OS, PFI					
UVM	OS	OS	OS, PFI	OS, PFI	OS	PFI		

665 **Supplementary Table 2**. Per cancer, the number of times genera from the phylum Firmicutes
666 were found among the selected features, whether positively or negatively associated with drug
667 response or prognosis.

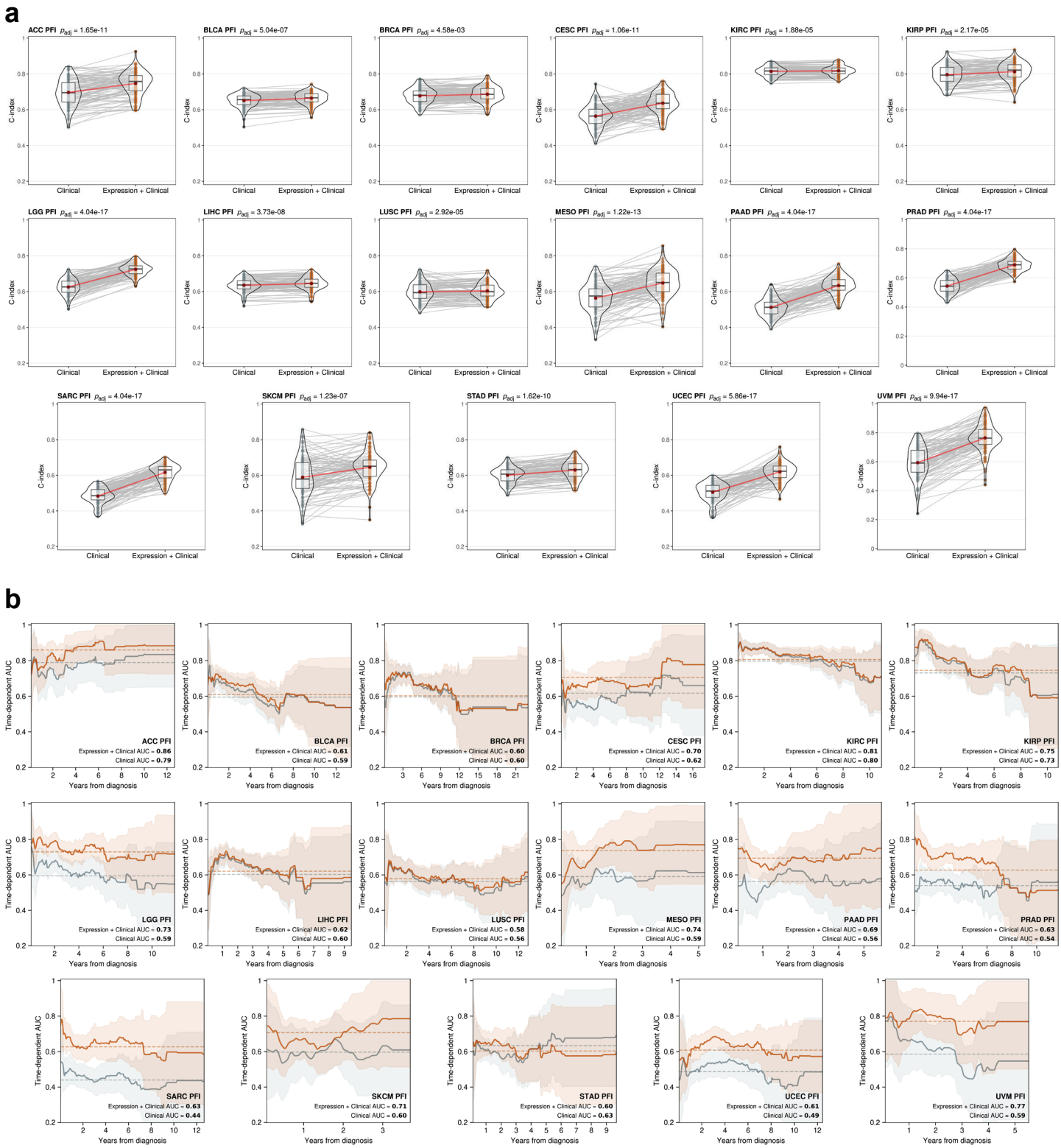
Cancer	Selected Negatively	Selected Positively	Total
ACC	8	13	21
BRCA	1	0	1
CESC	4	7	11
ESCA	0	4	4
KIRC	4	7	11
LGG	1	6	7
SARC	2	2	4
STAD	13	10	23

668

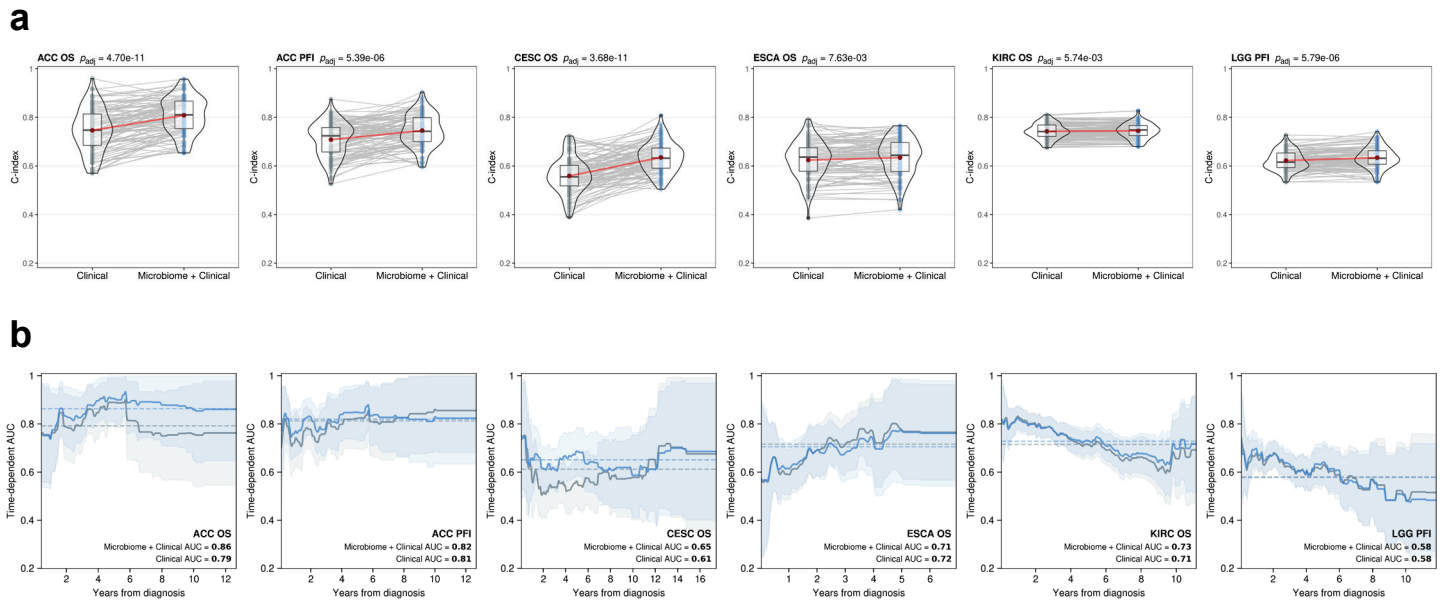
669 # Supplementary Figures



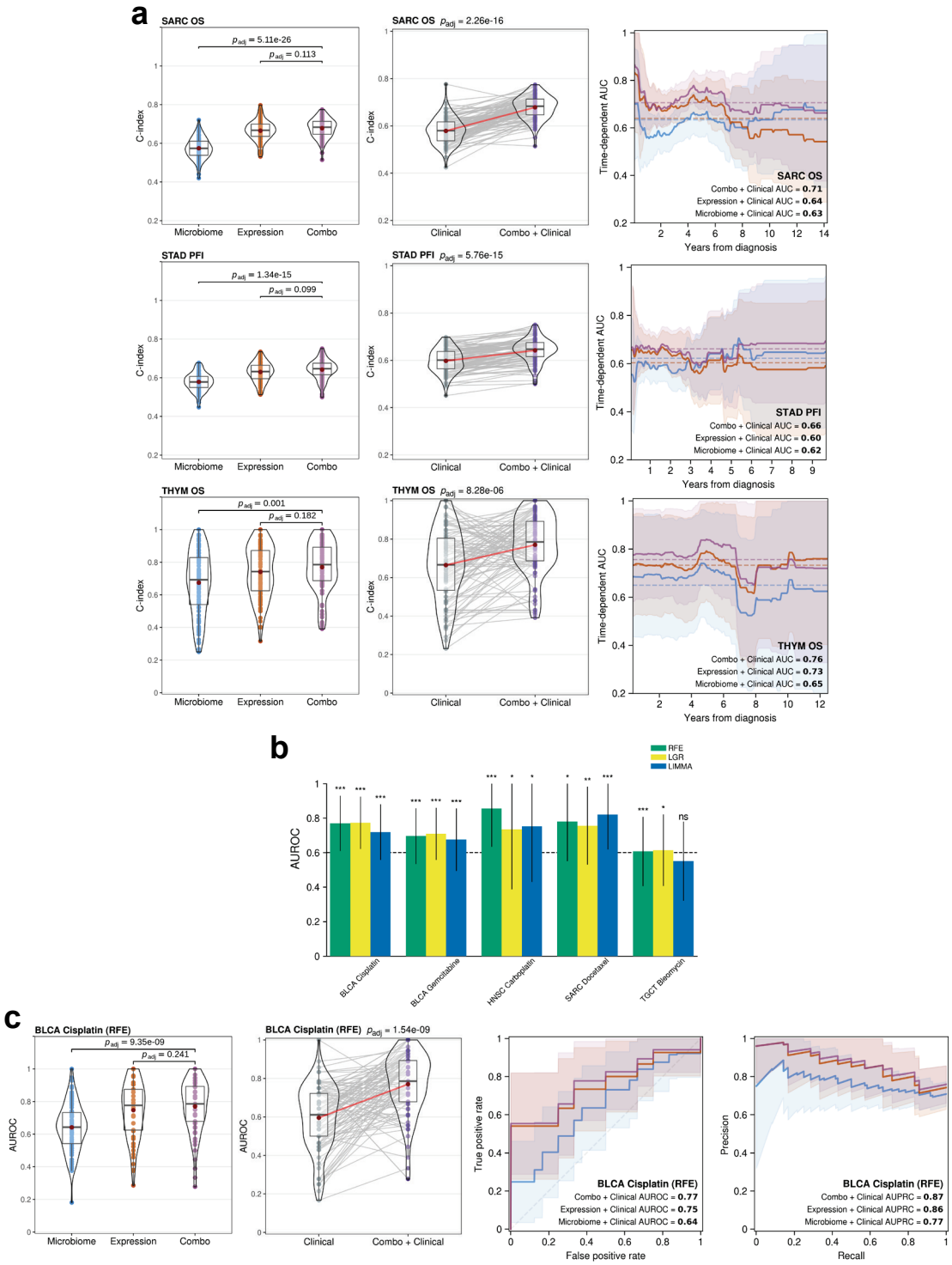
Supplementary Figure 1. Performance of gene expression overall survival (OS) models in the 16 tumor types where gene expression adds predictive power to clinical covariates. (a) C-index score density distributions (top rows) for gene expression with clinical covariate models vs clinical covariate-only models. 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. **(b)** Time-dependent, cumulative/dynamic AUCs (bottom rows) for gene expression with clinical covariate models (orange) vs clinical covariate-only models (grey) following years after diagnosis. Mean AUCs across entire test time range after diagnosis shown as a horizontal dotted line and in legends and shaded areas denote standard deviations.



Supplementary Figure 2. Performance of gene expression progression-free interval (PFI) models in the 17 tumor types where gene expression adds predictive power to clinical covariates. (a) C-index score density distributions (top rows) for gene expression with clinical covariate models vs clinical covariate-only models. 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. **(b)** Time-dependent, cumulative/dynamic AUCs (bottom rows) for gene expression with clinical covariate models (orange) vs clinical covariate-only models (grey) following years after diagnosis. Mean AUCs across entire test time range shown as a horizontal dotted line and in legends and shaded areas denote standard deviations.



Supplementary Figure 3. Performance of microbial abundance prognosis models in the five tumor types where microbial abundance features add predictive power to clinical covariates. (a) C-index score density distributions (top row) for microbial abundance with clinical covariate models vs clinical covariate-only models. 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. **(b)** Time-dependent, cumulative/dynamic AUCs (bottom row) for microbial abundance with clinical covariate models (blue) vs clinical covariate-only models (grey) following years after diagnosis. Mean AUCs across entire test time range shown as a horizontal dotted line and 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 in **(a)**.



Supplementary Figure 4. Performance of combined microbial abundance and gene expression models where combining both data types adds to predictive power. **(a)** Prognosis model C-index score violin density distributions (left) between microbial abundance, gene expression, and combined data type 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. Model C-index score violin density distributions (middle) for combined data type with clinical covariate models vs clinical covariate-only models. Time-dependent, cumulative/dynamic AUCs (right) for combined data type (purple), microbial abundance (blue), and gene expression (orange) models following years after diagnosis. **(b)** ML method drug response model mean AUROC scores where combining data types performed better than clinical covariates alone. For the combined data type drug response model closest to reach significant improvement over respective single data type models **(c)** AUROC score violin density distributions (left) comparing each data type model, AUROC score violin density distributions (middle left) comparing combined data type and clinical covariate model to clinical covariate-only model, and mean ROC and PR curves (right) for each data type model. Mean AUROC and AUPRC scores shown in panel legends and shaded areas denote standard deviations.

714 **Description of Supplementary Data Files**

715

716 **Supplementary Data 1.xlsx**

717 **Worksheet #1 Prognosis model cohort sizes:** number of cases in each prognosis model dataset.

718 **Worksheet #1 Drug response model cohort sizes:** number of cases in each drug response

719 model dataset, including the breakdown between responder (R) and non-responder (NR) cases.

720

721 **Supplementary Data 2.xlsx**

722 **Worksheet #1 Microbial genera features:** by cancer and by comparator, the genera selected as

723 features by the criteria outline in Methods, and the ML methods that selected each genus.

724 Further we show the number of times each genus was seen with rank at most 50 over the number

725 of times the genus could be seen: 100, 200 or 300 if one, two, or three ML methods chose the

726 feature, respectively. The column conditional direction shows the direction assigned to the

727 feature for each ML method; an asterisk indicates that there was more than one relevant ML

728 method, but all methods chose the same direction. The conditional direction is the typical sign of

729 the coefficient within a model, which is conditional on other features in the model. Also shown

730 is the Univariate FDR, which is the FDR that the feature is univariately predictive (n.s. is not

731 significant), and the direction in which the feature is univariately predicative, which may differ

732 from the conditional direction.

733

734 **Worksheet #2 Gene expression features**: by cancer and by comparator, the genes selected as

735 features by the criteria outline in Methods, and the ML methods that selected each gene. Further

736 we show the number of times each gene was seen with rank at most 50 over the number of times

737 the gene could be seen: 100, 200 or 300 if one, two, or three ML methods chose the feature,

738 respectively. The column conditional direction shows the direction assigned to the feature for

739 each ML method; an asterisk indicates that there was more than one relevant ML method, but all
740 methods chose the same direction. The conditional direction is the typical sign of the coefficient
741 within a model.