Host- Microbiome Omics Integration for Cancer Analysis and Diagnostics

Investigating the added value of integrating microbial and host omics information for cancer diagnostics using prediction models

Gedeon d' Abreu de Paulo

4723686

MSc Computer Science, Artificial Intelligence Track

(Bioinformatics specialization)

Thesis Committee

Dr. T.E.P.M.F. Abeel, TU Delft (supervisor)

Dr. A. Lukina, TU Delft

Delft University of Technology

Preface

This report describes my master’s thesis relating to the exploration of the added value when combining data on the human microbiome with host biological information. I first became curious about the effect that the microbiome can have on human health after reading about the Human Microbiome Project. Combining this with host biological information seemed like a very interesting prospect and a practical way to contribute research to what seems to be an increasingly important field. Due to special circumstances, I decided to do this project over a longer period of time. While it has been a challenging process for various technical, personal and data-deficiency reasons, it was always fun. I am glad I got the opportunity to combine everything I have learned in my bachelor and master programs into such an interesting project.

To start with, I would like to thank my supervisor and lab professor Thomas Abeel for having me in his lab, supervising me and always giving me valuable feedback. I learned a lot in the Abeel lab and had a lot of fun while doing so. Additionally, I would like to give a special thanks to Akash Singh for being my daily supervisor for more than half of my project. He was an indispensable second brain that helped me along, especially during the most difficult initial parts of the project. I would also like to thank my parents for supporting me throughout my academic journey, and Savanne for being there and providing her unwavering support. Finally, I would like to thank Dr. Anna Lukina for agreeing to be part of my thesis defense committee.

Gedeon d’ Abreu de Paulo

Den Haag, April 2023

Abstract

Cancer is one of the leading causes of death in the world. While there have been many studies investigating the development and progression of cancer in human tissues using host omics data or microbial data, there is a lack of research combining both types of data, even though both modalities have been shown to affect cancer morphology and aetiology. Studies which do combine these modalities often use simple methods or do not consider the relation between the two modalities and disease phenotypes. Such an integrated approach could offer additional insights and lead to the discovery of new disease biomarkers and better treatment strategies and therapies.

In this paper, we investigated whether such a holo-genomic approach offers additional information compared to using the modalities separately, by comparing the performances of prediction models built using the individual and integrated modalities for various prediction endpoints. To do this, we used TCGA gene expression data for the host omics modality and bacterial genus abundance data from the TCGA-mined Cancer Microbiome Atlas (TCMA) for the microbiome modality.

We found no improvement when integrating host gene expression with microbial abundance information compared to using the gene expression data individually, and the microbial data provided the least amount of diagnostic information. This is likely due to the information density of gene expression data, high variation of the microbiome, and the quantity, specificity and validation of the TCMA data. These results suggest that the holo-omics approach might not provide additional utility in certain contexts, that additional considerations have to be made when choosing microbial and host omic datasets for holo-omic integration, and provide an insight into the usability of the TCMA data set.

Introduction

Cancer is among the leading causes of death worldwide, being responsible for millions of deaths every year. The aetiology, morphology and progression of different cancers depend on a complex interplay of various biological and environmental factors. Recently, it is becoming increasingly easy to investigate this complex interplay thanks to the development of more modern sequencing technologies and the availability of biological data. Such data availability has made it more accessible for researchers to use various omics data to perform various tasks for cancer diagnostics which relate to the analysis and integration of both host and microbial omics data.

Host multi-omics integration

The availability of host omics data of various data layers, such as gene expression (i.e. genomics), DNA methylation (i.e. epigenomics) or copy number variation has enabled researchers to derive useful insights into the aetiology and morphology of different cancers. While there are many such data sources that have been made available, one of the most impactful sources has been The Cancer Genome Atlas (TCGA), a repository of genomic profiles of over 30 types of cancer that can be used for cancer diagnostics [1]. Using host omics data from this database, researchers have been able to derive insights such as finding biomarkers, determining differences in biological composition of tumor samples versus normal samples and examining genetic features related to survival. While using individual omics types has led to useful insights, an important development has been the usage of so-called multi-omics analysis methods, where data from multiple omics layers are integrated. This has been shown to lead to additional insights and model performance over single omics methods for cancer diagnostics [2,3], however, these methods deal with additional challenges owing to the heterogeneity, noise, high dimensionality and sparsity of the multi-omics data [4]. To derive insights within these conditions, authors have combined multi-omics features with prediction models for varying prediction tasks such as cancer subtype detection [2,3], the classification of tissues as tumor or normal samples, and survival prediction [2] to diagnose cancers and allow for personalized treatments, and identify disease biomarkers that are predictive of the cancer state.

Microbiome-based analysis

Besides host omics data, a promising field of research relates to the analysis of microbial omics data. There are many microorganisms that live in communities on different human tissues, called the human microbiome. Namely, an ecosystem of 10 to 100 trillion microorganisms encompassing 500 to 1000 unique species for each individual [5]. Due to the aforementioned advances in sequencing technology, it is becoming increasingly easy to measure the identity, metabolic potential and expression of this microbiota. This is leading to various data sets on the human microbiome which can be exploited.

Such data sets include MetaHIT [6] and iHMP [7], which contain microbial data from healthy and diseased patients. However, a problem with these data sets is that they often contain data from tissue swabs and stool samples, which are not necessarily representative of the microbiome of internal organs [5]. Next-generation sequencing data sets, such as TCGA, contain, next to host sequencing data, microbial sequencing data. This aspect of TCGA is mostly unexplored yet can be mined to obtain data on, for example, viromes and bacteriomes of different cancers using different tissues, such as tumor tissues and blood. However, the microbial reads in this data set are often a result of contamination [5,8] and thus extensive care needs to be taken to properly decontaminate the data. There have been a number of studies that have managed to do this and obtain useful insights on the relation between the tumor microbiome and certain cancers [9,10], motivating the use of TCGA for microbial analyses. However, these data sets are often not readily available. A data set which attempts to combat this, is the Cancer Microbiome Atlas (TCMA), which contains batch-corrected and decontaminated microbial data mined from TCGA whole-genome sequencing (WGS) and whole-exome sequencing (WXS) experiments [5].

Using microbial datasets, various aspects of the human microbiome and its association with diseases have been studied through various means. Taxonomical data has frequently been used to investigate bacterial abundance differences between cancer and healthy samples using hypotheses tests [5,9,11,12] or prediction models, such as linear regression [9,12]. It has also been used to examine co-abundance of microbiota in certain tissues [5] and the association of microbiota with patient survival [5,13] or clinical factors such as gender and age [11–13]. Besides taxonomical data, It is also possible to examine microbial differences between tissue types on the functional level through the use of metaproteomics or metagenomics [14]. Both taxonomical and functional investigations have shown that the microbiome can be affected by many factors such as diet and environmental exposure [15] and that it exhibits significant variation across individuals [16], cancer types [9,17,18], cancer subtypes [18], healthy and unhealthy individuals [18], tumor versus normal samples [19] and cancer patients with different survival rates [20].

Furthermore, it has become clear that the microbiome is not only associated with human phenotypes but that composition and changes in the microbiome have a direct influence on oncogenesis [21–23] and tumor immunotherapy response [22,23], whether positive or negative. As an example of a mechanism through which the microbiome can affect a patient, certain bacteria can bind to and alter the function of immune system cells which infiltrate tumors, thereby affecting carcinogenesis and resistance to chemotherapy [15].

The need for a holistic view

It is clear that both host omics and microbial omics data can be used to obtain useful biological insights into the aetiology of different cancers. As shown, many studies use one or the other to understand different biological processes without considering their interplay [24]. However, it has become clear that the host can alter the human microbiota and vice versa [25]. Thus, the integration of host and microbial data could help to better understand the aetiology and physiology of different cancers and provide new insights [21]. This field, where a holistic approach is taken to biological data, is known as hologenomics. It is based on the assumption underlying the hologenome theory, which posits that the host and microbial genome are biologically dependent and must be analyzed together in order to investigate the phenotype of an organism [26].

As TCGA has seen much success in host (multi) omics analyses, the possibility of using TCGA-mined microbial data offers a valuable opportunity for matched host-microbe analyses. While limited in number, prior studies have exploited this by combining host omics features of tissues with the microbial information contained within these tissues. Chakladar et al. examined microbe-host interactions in pancreatic adenocarcinoma by investigating intra-pancreatic metastasis- and survival-linked microbe abundance data mined from TCGA and correlating it to host gene expression patterns [13]. Meanwhile, Greathouse et al. examined the interaction between the microbiome and TP53 in lung cancer by investigating the association between TP53 mutations and microbial abundance and diversity using statistical tests [11]. Finally, Dohlman et al. used the TCMA dataset to correlate tumor-normal-linked co-abundant bacterial groups with gene expression patterns of certain genes for colorectal cancer [5].

It is clear that TCGA-mined microbial datasets offer a unique opportunity and valuable data for a hologenomic approach. However, while current hologenomic studies examine host and microbial data, this is done using simple approaches such as statistical tests, which might not consider the interplay between the two types of features. Additionally, such studies often examine the relation between each modality and the disease phenotype separately, rather than investigating how the combined modalities interact to affect disease phenotype. Combining host omics and matched microbial data using powerful machine learning models provides a richer avenue for discovering how these modalities differ across diagnostic endpoints, such as tumor versus normal tissue classification, and affect tumor development and morphology.

A hologenomic approach that properly considers the interactions between host and microbial data can help derive insights that could further elucidate how cancers develop, identify targets for vaccines or microbiome-altering therapies, improve immunotherapies, identify biomarkers and identify cancer stages for prognostic assessment and specialized treatments.

Towards a holistic view

In this paper, we aimed to investigate whether a holistic view provides additional insights for cancer diagnostics when compared to simply using each modality individually by integrating host genomic and microbiome data for cancer patients. Specifically, we aimed to investigate the question:

Does integrating host and microbial omics data provide additional power over using the modalities individually?

In this case, power refers to prediction performance, as prediction models have shown much effectiveness in dealing with the challenges present in such an integrated data set and identifying important features and relationships for both microbial [9] and gene expression [27] features.

To this end, we leveraged the TCGA and TCMA data sets and integrated gene expression and microbial genus abundance data for tumor tissues and tumor-adjacent normal (NAT) tissues for colon adenocarcinoma (COAD), esophageal carcinoma (ESCA), head and neck squamous carcinoma (HNSC) and stomach adenocarcinoma (STAD). We then investigated whether the integration of each modality provided additional prediction performance for the tumor versus normal prediction and tumor stage prediction endpoints compared to simply using the modalities separately. While we discuss our main findings on STAD due to the relative simplicity of the cancer and high quantity and class balance in the dataset used, much of the experiments have also been performed for the other three cancers and can reasonably be generalized to these cancers.

We found that using the integrated modality did not provide additional performance over using the gene expression (GE) modality separately and that the prediction models that used the genus modality performed the worst. This is likely due to the high information density of the gene expression layer and the low amount of information conveyed by the genus layer. The low information of the genus layer is likely due to the high biological variability of microbial data, but also technical aspects related to the creation of the TCMA dataset. Thus, these results cannot reasonably be generalized to other microbial data sets.

Materials and methods

Data

In order to explore the performance of a holo-omic approach for cancer diagnostics, we used gene expression features from TCGA, denoted as GE, and microbial genus relative abundance features from TCMA, which we denote as genus. We only kept samples in these data sets for which there was both microbial, as well as host omics data. The set containing both modalities is denoted as GE ∩ genus. Furthermore, while we validated certain aspects of the findings in this study using other cancers, we conducted the main experiments in the study using only stomach adenocarcinoma

(STAD) data, as it had the most amount of samples and best class balance for the prediction endpoints among the cancers available.

Microbial data

For the microbial data, we used data from the Cancer Microbiome Atlas (TCMA). This is a microbial dataset, created by Dohlman et al., that was obtained by using a statistical model to isolate tissue-embedded microbial species present in TCGA samples from contaminants, and was subsequently validated using 16S rRNA amplicon sequencing on the original TCGA tissue samples. The resulting TCMA database, accessible through the TCMA portal [1], contained tissue resident microbial relative abundance data for 3689 unique samples and 1772 patients across 21 anatomic sites and 5 TCGA projects (HNSC, ESCA, STAD, COAD and READ). The authors only released data for these 5 cancers, as they were the ones with the most microbial reads among the cancers investigated in the original paper. Thus, every tissue sample contained the relative amount of each genera in that sample, where this amount is a number between 0 and 1, and the total relative amounts of all genera in a sample add up to 1. Additionally, we omitted READ due to a lack of data on that cancer.

The highest taxa specification contained within this data was the genus level. As the phylum level and other taxa levels above genus are less specific than the genus level, we decided to continue with only the genus taxonomical abundance data, which contained 221 taxa. For STAD, there were 52 genera that did not have a relative abundance of 0 across all samples (Figure 1). In total, there were 119 nonzero features among the overlapped samples for all cancers (Figure S1, Table S1).

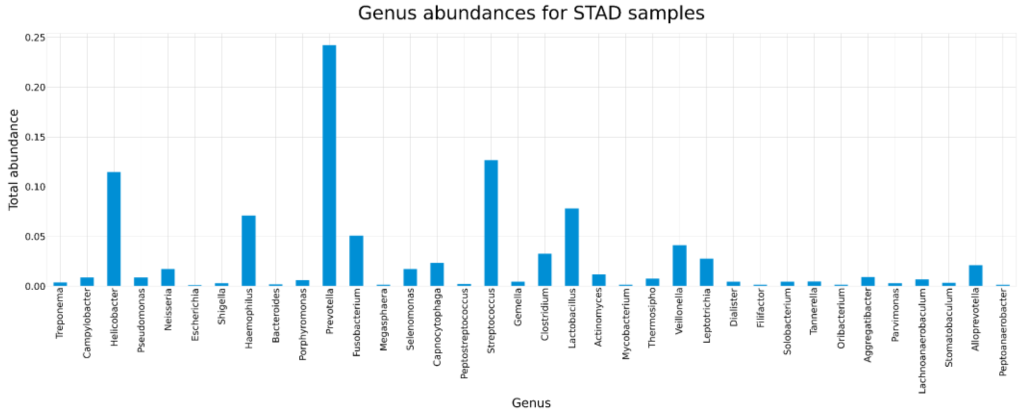

Figure 1: Relative abundances of the 52 nonzero genera in the TCMA data set for STAD samples. Note that this only includes TCMA samples for which there was also paired TCGA GE data available.

Host omics data

The host omics data consisted of a TCGA data set which was extracted and processed by a prior study [28]. This data set consisted of level 3 RNA-seq gene expression data for 9732 tumors and 727 tumor-adjacent normal samples encompassing 33 total different cancer types. The TCGA RNA-seq data was obtained using the UCSC Xena data browser on March 8, 2016. The expression values of these genes consisted of a pre-processed and batch-corrected GE x sample matrix with RSEM values normalized using a log2(FPKM + 1) transformation. The gene expression values of the 5000 genes with the highest variability were subsequently selected, as evaluated using Median Absolute Deviation (MAD). Finally, the expression values in the data set were min – max scaled. In further experiments involving the gene expression-only modality, we only used those samples for which

[1] https://tcma.pratt.duke.edu/

both host omics, as well as microbial data, was available. Thus, we also only considered the cancers STAD, HNSC, ESCA and COAD.

Clinical data

The clinical data was accessed using the Snaptron web server by the same authors [28]. We matched the clinical data with the corresponding patient samples in order to obtain details for the tumor and stage endpoints. To determine whether a sample was a tumor or NAT tissue, the sample type code [2] was used, where codes in the range 01 – 09 are tumors and those in the range 10 – 19 are normal samples [3]. Furthermore, the stage clinical data was used to determine the tumor stage of each sample. We grouped every substage together to obtain 5 total final stages. For example, samples that were classified as stage IIA and IIB were grouped together under the bin of stage II. Finally, we modeled the tumor-adjacent normal samples as stage 0.

Overlapped data

In order to investigate the effects of integrating microbial and host omics data on cancer diagnostics, we created an overlapped set of samples for which there was both host- and microbial omics data in the above-described data sets. In the first step, the TCGA gene expression data was joined with the clinical data. To do this, we used the TCGA barcode field of each row in the clinical data set, which contained the code for the project, the tissue source site (TSS), participant ID, sample type, vial, and portion id. This field was then stripped of the portion ID and vial information and the duplicate rows were dropped to remove samples with the same ID but different measuring technologies. Each row with clinical data was then joined with the samples in the GE data set by matching rows for which the TCGA barcode (excluding vial and portion id information) was equal. To join the data set with GE + clinical data to the genus abundance dataset, we matched rows for each table based on the TCGA barcode, excluding vial and portion ID information. This led to a data set containing samples with GE, genus abundance and clinical information for STAD, HNSC, ESCA and COAD.

Thus, this dataset contained information on whether each sample was a tumor or NAT tissue (Table 1) and the stage of the sample (Table 2). As can be seen, there are less total samples per cancer with the stage endpoint than with the tumor versus normal categorization. This is due to the stage information for certain samples being absent. In the end, when building models for each modality, we only used those samples that were contained in the overlap set to allow for a fair comparison.

Table 1: Number of tumor and normal adjacent tissue (NAT) samples in the overlap set for each cancer.

Cancer	Normal (NAT)	Tumor	Total
STAD	9	113	122
COAD	3	45	48
ESCA	7	59	66
HNSC	7	154	161

[2] https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes

[3] https://docs.gdc.cancer.gov/Encyclopedia/pages/images/TCGA-TCGAbarcode-080518-1750-4378.pdf

Table 2: Number of samples for each tumor stage in the overlap set for each cancer. Stage 0 corresponds to normal adjacent tissue (NAT) samples.

Cancer	Stage 0	Stage 1	Stage 2	Stage 3	Stage 4	Total
STAD	9	19	36	27	16	107
COAD	3	12	15	11	6	47
ESCA	7	8	30	15	2	62
HNSC	7	10	23	23	71	134

Complex modality integration

For the complex modality integration, we concatenated the GE and genus data set and integrated the features using an autoencoder and nonnegative matrix factorization. We also extracted features from the GE and genus data sets separately by using these integration methods on the GE data and the genus data individually. In the end, we extracted 100 features from the 5221 features in the GE + genus concatenated set, 100 features from the 5000 features in the GE set, and 50 features from the 221 in the genus set.

Autoencoder

For the integration using an autoencoder, we used *PyTorch* to define the autoencoder architecture and train the model, and used MSE as the loss function and Adam as the optimizer.

For hyperparameter tuning, we used the *Skorch* package to wrap the Autoencoder module into a *Scikit-learn* compatible model. The *scikit-learn* GridSearchCV package was then used to perform hyperparameter tuning and determine the optimal architecture for the autoencoder. Each set of parameters was evaluated using 5-fold cross-validation. We used a grid search space that we defined based on the architecture of an autoencoder built by Chaudhary et al. [2], which was previously used to successfully integrate host multi-omics features. Specifically, we explored an architecture with 3 hidden layers, with the options outer_hidden_layers_size : {100, 200} and extracted_features_amount (i.e. middle hidden layer) : {30, 50, 100}. We also evaluated an architecture with 5 hidden layers, with the same parameter space for the outer and middle hidden layers, and the added grid space of second_to_last_outer_layers_size : {50, 100}. For the Adam optimizer, we explored learning_rate : {1e-1, 1e-2, 1e-3, 1e-4}, and for the model training max_epochs : {10, 20, 30, 40, 80}. We ran the hyperparameter tuning pipeline on the GE, genus and GE ∩ genus datasets separately. The optimal architecture for the GE integration model and the GE ∩ genus model was an architecture with 3 hidden layers with an outer hidden layer size of 200, extracted features amount of 100 and trained for 80 epochs. For the genus model, the optimal architecture had 5 hidden layers, an outer and second to last outer hidden layer size of 100, extracted features amount of 50 and was trained using 80 epochs. After obtaining the optimal hyperparameters, we used each dataset to train a separate AE model per modality and then used these models to extract features from the datasets of each modality to create their AE-integrated counterpart.

Nonnegative Matrix Factorization

For the integration using nonnegative matrix factorization, the NMF function from the *scikit-learn* decomposition package was used. To compare results, we extracted the same amount of features from each modality as we did with the autoencoder. Additionally we selected the initialization to be random, and supplied a static random seed to allow for reproducibility between experiments. Finally, we created NMF-integrated counterparts of the data sets for each modality.

Prediction pipeline

We evaluated the usefulness of each data modality by evaluating the performance of a prediction model when using the modality to train the prediction model for a prediction task (Figure 2). For the stage prediction endpoint, the prediction of the different stages 0-4 was modeled as a regression problem and was performed using the *scikit-learn* random forest regressor model and the elastic net model, while the tumor versus normal prediction endpoint was modeled as a binary classification problem and was performed using the *scikit-learn* support vector machine model. All the prediction models were initialized with a random seed of 0. Each experiment was performed for each combination of cancers (i.e. STAD, COAD, ESCA and HNSC) and for each modality (i.e. gene expression, genus and the combination of gene expression and genus) separately. To train and evaluate the model, we used a random-sampling based approach (Figure 3).

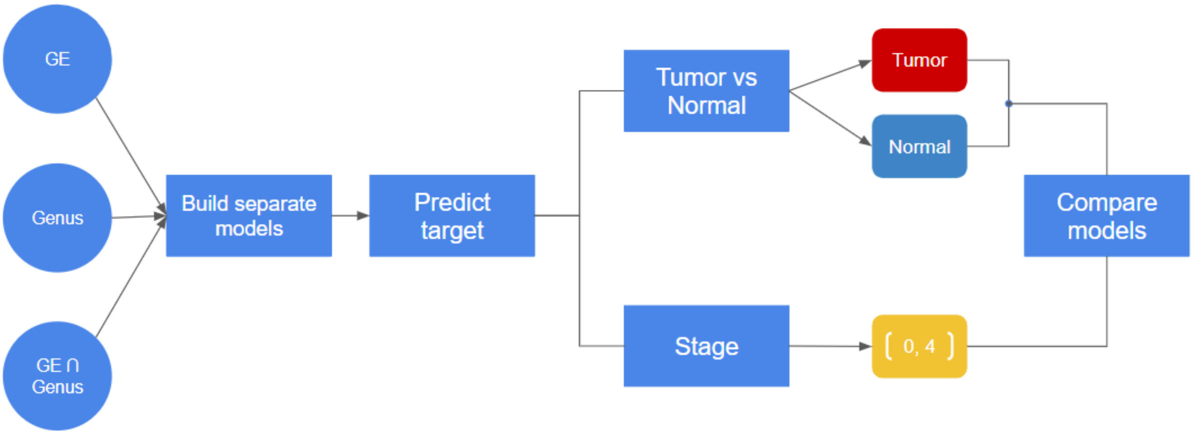

Figure 2: The general prediction pipeline. For each modality GE, genus and GE ∩ genus, we built separate prediction models for the targets tumor versus normal prediction, and tumor tissue stage prediction. We then evaluated the performance of the prediction models on each modality and compared the models in order to determine the utility of each modality.

Baseline model

To provide a reference to help interpret the performance of each prediction model, we also determined what the predictive performance would be for a model which always predicts the majority target value for each prediction task. For example, for STAD and the tumor versus normal classification endpoint, the baseline model is one which always predicts that a sample is a tumor sample and would have an F1 score of 0.96. For stage, the baseline model would be one which always predicts that a sample has a stage of II (i.e. a target value of 2), and would have an RMSE of 1.168.

Data splitting

We used a random sampling approach to obtain an estimation of how a prediction model performed when used on each data modality separately, compared to how the model performed when using a combination of the modalities using various integration methods (Figure 3). This was done by performing a random stratified split of the relevant dataset into 80% training and 20% testing using the *scikit-learn* train_test_split function. This split was performed 200 times to obtain a reliable estimate of the model performance and each split iteration was assigned a custom seed based on the iteration count to ensure consistency between experiment runs.

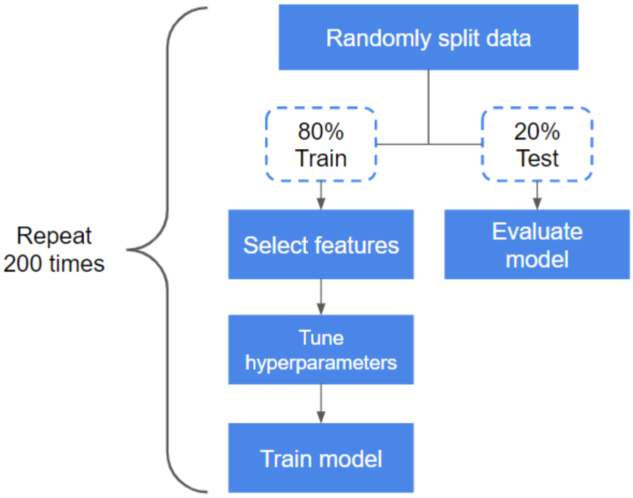

Figure 3: Data spitting and prediction pipeline. To evaluate the predictive power of each modality, we randomly split the data into a training set, consisting of 80% of the data, and a testing set, consisting of 20% of the data. On the training set, we performed feature selection using varying feature selection methods, performed hyperparameter tuning to determine the best set of parameters for the given prediction model, and then used the entire 80% of the data to train the prediction model. We then used the 20% testing data to evaluate the model performance using various prediction metrics. Ultimately, we repeated this 200 times to obtain 200 evaluation metrics for each experimental setup tested and evaluated the models based on these metrics.

Feature selection

Prior to generating a prediction model, we performed feature selection in order to reduce the effect of noisy variables and only include the most powerful predictors.

Using 80% of the data, this data set portion was used to perform feature selection for various feature amounts (i.e. 6, 10, 26, 50, 100, 200), up to the maximum amount of features present within the modality. To do this, we first performed feature selection to create an ordered list of the top 200 features, ranked according to the weights assigned by the relevant feature selection method. From this pool, we then ran the experiments using only the top 100, top 50, top 25, top 10 and top 6 features. Thus, the sets of selected features for each feature selection amount are subsets of each other. For example, the top 6 features are contained within the top 10 features, which are in turn contained within the top 26 features etc.

To perform the modality parity enforcement experiments, we performed feature selection using the GE and genus modality separately for each feature amount. For example, when selecting 200 features, we first selected 100 features using only the GE modality separately, then selected 100 from the genus modality, and then combined these together when training and evaluating the model.

For the stage prediction endpoint, the feature selection was performed using the Pearson correlation coefficient, an elastic net model, and a random forest regressor model. The Pearson correlation coefficient-based feature selection was performed by using the *scikit-learn* SelectKBest function with the r_regression method as the scoring function, while the latter two methods were performed by using the ElasticNet and RandomForestRegressor packages, respectively. Both were initialized with a random seed of zero. For the latter two model-based feature selection methods, we also first performed hyperparameter tuning to find the best parameters for the model, and then trained the model on the 80% training set. For the elastic net-based feature selection, we obtained a feature ranking using the magnitude of the coefficients for each feature value after training the model while for the random forest regressor-based model, we used a feature ranking based on the mean decrease in variance among the target values after using a certain feature as a tree splitting node. Finally, for the tumor vs normal prediction endpoint, feature selection was performed using the ANOVA F-test, implemented with the *scikit-learn* SelectKBest function with the f_classif method as the scoring function.

Model training

For each feature selection amount, after performing feature selection, we used the selected features to perform hyperparameter tuning and then trained a prediction model using these hyper parameters with the 80% training portion of the current random sampling iteration. We also trained prediction models using all of the features for the relevant modality to examine model performance when there is no feature selection.

Hyperparameter tuning

When we used a prediction model for either feature selection or to evaluate the predictive performance of a modality, we first performed hyperparameter tuning to find the optimal parameters. This was done using randomized search with the *scikit-learn* RandomizedSearchCV package. We used a stratified 5-fold cross validation split to evaluate the performance of each hyperparameter set and evaluated 100 different randomly sampled hyperparameter sets.

In terms of the searched hyperparameter space, for the elastic net we explored alpha : {1e-5, 1e-4 … 1e, 1e2} ∪ {0} and l1_ratio : {0, 0.1, 0.2 … 0.9}. For the random forest regressor, we explored n_estimators : {5, 20, 50, 100, 200, 400}, max_features : {'auto', 'sqrt'}, max_depth : {10, 30, 60, 100}, min_samples_split : {2, 5, 10}, min_samples_leaf : {1, 2, 4} and bootstrap : {True, False}. For the support vector machine, we explored C : {1e-10, 1e-9 … 1e9, 1e10}, class_weight : { 'balanced', None}, kernel : {'linear', 'rbf'} and gamma : {1e-10, 1e-9 … 1e9, 1e10} ∪ {'scale', 'auto'}.

The performance of each sampled hyperparameter set was evaluated using the root mean squared error as a scoring function for the stage prediction endpoint and balanced accuracy for the tumor versus normal endpoint. Finally, to ensure the reproducibility of the results, we also used a static random seed which only differed across random sampling iterations.

Testing and evaluation

After training a model, we then used the 20% testing set of the current random sampling iteration to evaluate the model. For the binary tumor versus normal prediction endpoint, we used the f1-score, while for the continuous stage prediction endpoint, we used the RMSE. For the latter, since the range of stage targets only spans the interval [0,4], we clamped the prediction values of the prediction model to always be within this range.

As there are 200 random sampling iterations, the prediction pipeline thus generates 200 different sets of values for these evaluation metrics. For the evaluation, we considered the average of each metric across the 200 random sampling iterations, and the standard deviation. To analyze the statistical significance of the difference between model performances, we used the Mann-Whitney U test. This was implemented using the *SciPy* mannwhitneyu package.

Results

Characterization of data

To investigate the effects of the host-omics approach, we used pre-processed gene expression (GE) samples from the Cancer Genome Atlas (TCGA), as gene expression data alone has been found to contain valuable information for multiple cancers [3]. This set consisted of 5000 normalized gene expression features with the highest variability that were selected from a larger set of TCGA gene expression data.

To investigate the effects of the microbial approach, we used bacterial genus relative abundance data from the Cancer Microbiome Atlas (TCMA), which contained 221 genus features. This is an openly available microbial database containing batch-corrected and decontaminated genus-level relative abundance data mined from TCGA for 5 cancers [5], which allows for microbial analyses using a common source of data.

To evaluate the holo-omic approach, we used a dataset with the concatenated genus and gene expression features for each sample, which contained 5221 total features. To allow for a fair comparison of this integrated set with the separate gene expression and genus modalities, we ensured each modality had the same amount of tissue samples and thus only considered tissue samples in the gene expression and genus set for which there was both gene expression and genus abundance data available. Ultimately, there remained data for stomach adenocarcinoma (STAD), colorectal adenocarcinoma (COAD), esophageal squamous carcinoma (ESCA), head and neck squamous carcinoma (HNSC) and rectal adenocarcinoma (READ).

Furthermore, we combined the patient samples with matched clinical information to determine whether the sample originated from a tumor or tumor-adjacent normal (NAT) tissue and determine the stage of the tumor tissue. For the tumor stage, the possible stages ranged from stage 0 to stage 4, as we considered the stage of NAT tissues to be 0.

Finally, we focused our analysis on the STAD data, as STAD had the most simple aetiology among the cancers in the data set and the highest amount of samples and balance between classes. However, the main experiments have also been run for COAD, ESCA and HNSC, with READ being omitted due to a lack of samples. The STAD dataset contained 122 samples for the tumor versus normal endpoint, of which 113 were from tumor tissues and 9 from NAT tissues (Table 1), and 107 samples that contained tumor stage information (Table 2).

Holo-omic approach does not lead to improved prediction performance for either prediction target

To investigate the possible benefits of a holistic view of omics integration, we used gene expression data and microbial abundance data of various cancers in a predictive model for the binary classification task of tumor versus normal prediction and the regression task of tumor stage prediction. To establish a baseline, we built prediction models on the gene expression data set (denoted as GE) and the microbial abundance data (denoted as genus) separately and evaluated the prediction performance. We then built prediction models on the concatenation of both of these data sets (denoted as GE ∩ genus) and compared it to the established baseline. An overview of the prediction pipeline can be found in Figure 2. Finally, we also compared the performance of all models to that of a model which always predicts the majority class.

We used random sampling to perform 200 random stratified splits of each modality data set into 80% training and 20% testing (Figure 3). In each random sampling iteration, we also used a feature selection method on the training set beforehand to select the most important features, as feature selection can have a significant impact on the performance and results of prediction models [29]. These features were then used to train a model on this same training set and the performance of the model was then tested on the testing set for the available prediction endpoints. This was repeated for each cancer, modality and feature selection amount, including when performing no feature selection. After this procedure, we obtained 200 scores for each evaluation metric and plotted the mean and standard deviation of these scores. For tumor vs normal prediction, we used the f1-score, as it can handle imbalanced data sets, while for tumor stage prediction, we used the root-mean-squared error (RMSE). As the distribution of these scores were nonnormal, we compared the collection of these scores for different models by using the nonparametric Mann-Whitney U test and a p-threshold of 0.01.

For tumor versus normal prediction, we used a support vector machine (SVM) due to its ability to capture nonlinear relationships, deal with imbalanced data sets and previously demonstrated performance with biological datasets [2,30,31], and the ANOVA f-test for feature selection, which has seen some success in selecting genetic features [2]. For stage prediction, we used an elastic net model as the predictor model due to its interpretability and previously demonstrated success in gene-based models [30,32] and also as the feature selector due to the demonstrated ability of penalized linear regression-based methods to select genetic features [20,29].

The results indicate that integrating STAD gene expression with bacterial genus abundance data does not lead to a significant improvement in prediction performance ($p = 0.0252$) over using gene expression data alone for tumor versus normal prediction (Figure 4A). Furthermore, these results were consistent across all cancers investigated (Figure S6). This is partially due to the GE model alone already performing optimally, which does not leave much opportunity for a performance improvement with the integrated modality. This high performance of gene expression is likely because gene expression is among the most causative host-omic modalities for tumor development and thus contains the most informative features for the discrimination between tumor and normal samples. This discriminative power can also be seen after applying dimensionality reduction techniques on the GE data and visualizing the relatively large separation between tumor and normal samples in the lower dimensional space (Figure S2, Figure S3). Additionally, gene expression can be correlated with the presence of certain microbes, as it affects the tumor environment, which in turn cultivates different microbiota. Thus, GE might already contain much of the discriminatory information that the genus data contains.

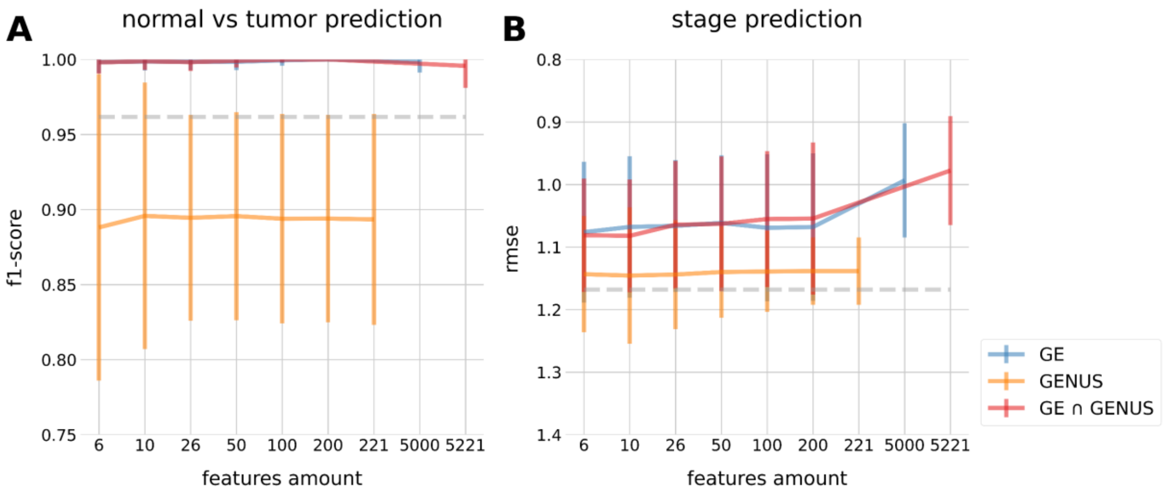

Figure 4: Predictive performance of prediction endpoints for STAD (stomach adenocarcinoma). The grey dotted line denotes the performance of a baseline model which simply predicts the majority class for the given prediction endpoint. ***A****, f1-score for the tumor versus normal prediction endpoint using ANOVA feature selection and an SVM classifier. Each line contains the f1-score for a support vector machine trained and tested on a different modality, namely on the genus abundance data (genus), the gene expression data (GE) and the concatenated genus + gene expression data (GE ∩ genus). The endpoints of each horizontal line segment indicate the average f1-score across 200 random sampling iterations, while the vertical line segments indicates the standard deviation of the f1-score across these iterations. The right-most point of each line indicates the prediction performance when all the features of the relevant modality are included (i.e. when there is no feature selection).* ***B,*** *Root-mean-squared error (RMSE) for the stage prediction endpoint using an elastic net model for the stage prediction and feature selection.*

The genus layer seems to provide the worst performance and is even outperformed by the baseline model. This is partly due to the baseline model (f1-score = 0.96) already being able to perform quite well due to the heavy class imbalance of 113 tumor samples against 9 NAT tissue samples. The taxonomic genus data alone does not seem to allow the model to discriminate well between the sample types, indicating that it does not contain enough discriminatory information. Combined with a high class imbalance, this also leads to a higher standard deviation in performance for the genus modality. We believe that the lack of discriminatory ability of the genus abundance features might be due to the high microbial variation between individuals and similarity between sample types due to microbial transfer between the tumor and NAT tissues.

We investigated whether a prediction endpoint with a better class balance might alter the results by performing the experiments for the stage prediction endpoint. Again, we observed no statistically significant difference (p = 0.1246) between the GE and GE ∩ genus model, and the genus layer performed the worst (Figure 4B). Furthermore, these results were consistent across all cancers (Figure S7). For stage prediction, the GE and GE ∩ genus models seem to perform worse than for the tumor vs normal prediction endpoint model, yet still perform better than the baseline. This is partly expected, as the prediction task is harder, with a larger range of target values. This, combined with the stage imbalance and lack of samples might also make it harder for the model to learn to discriminate between stages. This difficulty of the stage prediction task can be seen by the small amount of separation between samples with different stages when visualizing them in a lower dimensional space (Figure S4, Figure S5).

Due to the more difficult prediction task, the baseline is also easier to beat, as the model that always predicts the majority class will deviate from the actual class more often than if there are only two classes. Because of this, the genus modality seems to perform more similarly to the baseline than for the tumor vs normal classification task, as it also defaults to simply predicting the majority class. The poor performance of the genus model is similar to other prediction models which have used genus data for stage prediction [9]. Again, the high interpersonal variation of the microbiome and movement between tumor and NAT tissues might also play a role in the performance the genus model. Due to the higher class balance, we continued conducting the following experiments using only the stage prediction endpoint.

To investigate whether the lack of improvement with the holo-omics approach for the stage prediction endpoint was due to the linear elastic net model not being able to properly capture the information contained within the individual and integrated modalities and the interaction between these modalities, we ran the prediction pipeline using a random forest regressor, which is able to capture nonlinear relationships between features and has seen some success in prediction models with gene-based features (Figure S9). Again, the holo-omic approach did not offer additional improvement over the individual gene expression layer, indicating that the lack of performance improvement with the holo-omics approach was not due to the model not being able to properly capture complex nonlinear relationships.

Lack of holo-omic improvement is independent of feature selection method

To investigate whether the lack of performance improvement with the holo-omic approach was due to the feature selection method, we ran the stage prediction pipeline with the Pearson correlation coefficient and random forest as feature selection methods. The Pearson correlation coefficient has previously successfully been used to find important genetic features and is prediction model agnostic, while random forest-based feature selection has also seen some success in gene-based models and can capture relationships between features.

As can be seen, for neither the Pearson correlation coefficient (Figure 5A) nor the random forest feature selection (Figure 5B) is there any improvement when integrating the modalities for any of the feature selection amounts. These results were also consistent for the Pearson correlation coefficient selection both across different cancers (Figure S8) and when using a random forest regressor as the prediction model instead of the elastic net (Figure S10). It is worth noting that the performance of the genus model no longer changes beyond the features amount of 50, as there are only 52 nonzero features in the genus dataset for STAD, which could be informative.

Furthermore, these results show that the lack of performance improvement with the holo-omic approach is likely not due to the feature selection method used being unable to deal with the different scaling methods used for the preprocessing of the GE and genus abundance data. Both the Pearson correlation coefficient and random forest regressor are scale invariant under linear transformations and thus can select features regardless of the different (linear) scaling methods of the GE and genus modality. Additionally, while the Pearson correlation coefficient does not consider interactions between features and only captures the linear dependence between a feature and the target, the random forest regressor model is able to model interdependencies between features and capture nonlinear relationships. This implies that the lack of improvement with the holo-omic approach is also not due to the inability of the feature selection method to capture nonlinear relationships between the modality features. Thus, we conclude that the lack of performance improvement with the holo-omics approach is likely not due to a deficient feature selection technique.

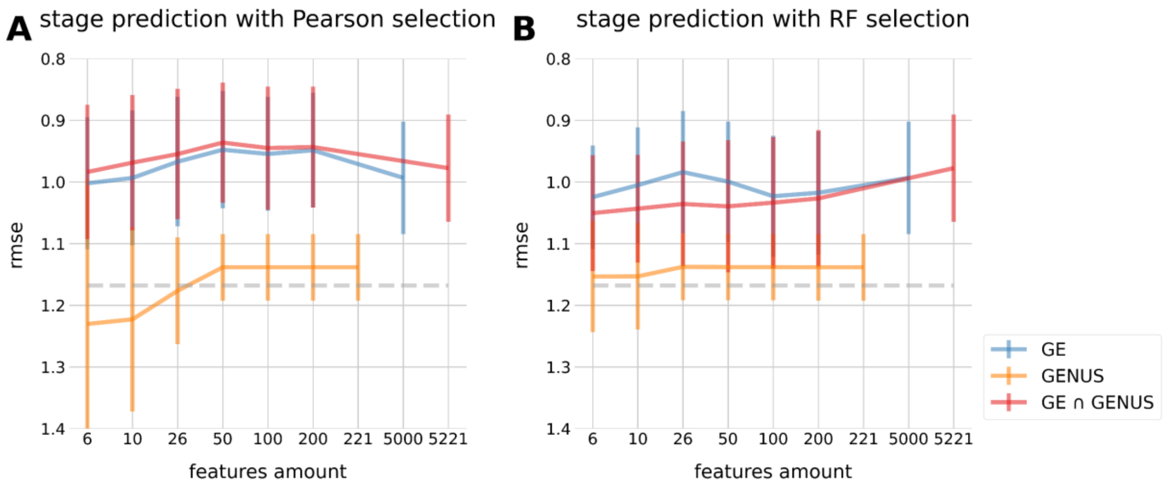

Figure 5: Stage prediction root-mean-squared error (RMSE) for an elastic net model trained and tested on the GE, genus and GE ∩ genus modalities using different feature selection methods. The last point of each line indicates the prediction performance when all the features of the relevant modality are included (i.e. when there is no feature selection) ***A,*** *Prediction performance when using the Pearson correlation coefficient to select features.* ***B,*** *Prediction performance when using random forest regressor-based feature importances to select features.*

Feature selection is dominated by gene expression features

To investigate to what degree the prediction models trained on the GE ∩ genus set were making use of both GE and genus features, we investigated what fraction of the features selected for different feature selection amounts consisted of GE features. To this end, we plotted the proportion of selected GE features for stage prediction for the three different feature selection methods used (elastic net, random forest and Pearson correlation-based feature selection) and displayed the fraction of GE features selected at different feature selection amounts (Figure 6A).

It seems that when performing feature selection on the integrated dataset, almost all of the features selected originate from the GE set. Furthermore, when investigating the absolute amount of genus features selected, this corresponds to approximately 1 genus feature being selected across the random sampling iterations and selected feature amounts (Figure 6B). These results are consistent across all cancers for elastic net-based feature selection (Figure S17) and the Pearson correlation coefficient (Figure S18), as well as for the tumor versus normal prediction endpoint using ANOVA feature selection (Figure S16).

This result is partly expected as there are roughly 25 times more GE features than genus features, with GE having 5000 features compared to 221 genus abundance features. If one assumes that both modalities are as predictive of the target, we would naturally expect 25x more GE features to be selected. Thus, for the top 6 and top 10 features, the amount of genus features selected is within expectations. However, for higher feature selection amounts, the genus features represent a disproportionately low fraction of the total selected feature set under this assumption. Furthermore, as we previously argued, the GE modality contains a higher amount of biologically relevant information than the genus modality, which is reflected in the disproportionately high number of GE features selected during the feature selection process.

Finally, the difference in number of features selected is likely also influenced by the preprocessing of the GE data, and the lack of informative features in the genus data. As previously described, the GE features used were a collection of features selected from a larger pool of GE features from TCGA that had the largest variability, as measured by the mean absolute deviation. Thus, it is more likely that features from this data set would also contain more features which vary with the target class.

Additionally, the genus data only contains 52 non-zero features for the shown STAD cancer, leading to the other 169 genus features not having any variability. This, combined with the previously seen lack of informative information in the genus features for the prediction endpoints likely leads to almost no genus features being selected when compared to the GE features.

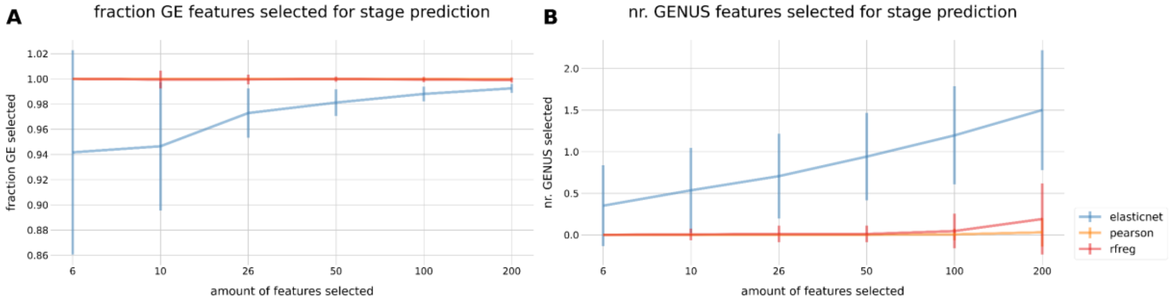

Figure 6: Features selected from each modality for STAD when performing stage prediction with the integrated genus and gene expression modality (GE ∩ genus). These features were selected out of 5221 total GE ∩ genus features using an elastic net model, the Pearson correlation coefficient and feature importances of a random forest regression model. ***A****, The fraction of GE features selected (vertical axis) from the total amount of features for each feature amount (horizontal axis). The endpoints of each horizontal line segment indicate the average fraction of GE features selected across the 200 random sampling iterations, while the error bars indicate the standard deviation across these iterations.* ***B****, The absolute amount of genus features (vertical axis) selected from the total feature set for each feature selection amount (horizontal axis).*

Enforcing modality parity during feature selection does not improve holo-omic model performance

As the domination of gene expression features during the feature selection process could prevent the prediction model from properly capturing the information of both the gene expression and the genus data, we attempted to mitigate this by repeating the prediction experiments while enforcing parity in the number of features selected from each modality. To do this, we performed feature selection prior to integrating the modalities and ensured that for each feature selection amount, half of the features were from the GE modality, while the other half were from the genus set.

As can be seen, the model trained with the modality parity-enforced integrated data, denoted as GE ∩ genus (parity), has a similar performance to the regular GE ∩ genus model (Figure 7). Comparing the scores for each modality with a feature selection amount of 200 indicates that there is no statistically significant difference between the model trained with modality parity enforcement and the one trained without the modality parity enforcement (p-value = 0.4985). This was consistent for all cancers (Figure S12), when using Pearson correlation-based feature selection (Figure S13) and for the tumor versus normal prediction endpoint (Figure S11). This result is likely because the genus modality is not offering additional information over the GE features already in the feature selection set. Thus, even when enforcing parity and ensuring that there are genus features during the model training, it achieves similar performance to the GE model as the prediction model still assigns the most importance to the GE features and uses them to discriminate between the target values. This can be seen by the nearly identical performance between the two models.

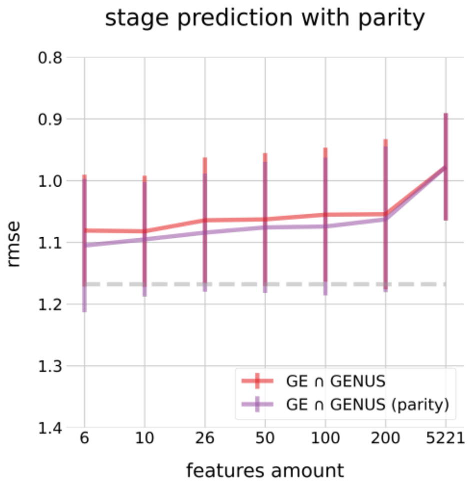

Figure 7: root-mean-squared error (RMSE) for the stage prediction endpoint for STAD with the use of an elastic net model for prediction and feature selection. Here, we compare the performance of the model when trained using concatenated genus and gene expression features with enforced parity during feature selection, denoted as GE ∩ genus (parity), with the performance of the model when trained using the regular (concatenated) integrated data (GE ∩ genus). The performance of the model with a features amount of 5221 denotes the model performance when using all features (i.e. when there is no feature selection) and is thus identical for the two models.

Selected features are supported by research

To investigate whether the lack of performance was due to the features selected having no biological basis, we investigated the top features selected by the elastic net model when using the GE ∩ genus modality for the STAD stage prediction endpoint. To do this, we determined the top features selected from each modality based on how frequently they appeared within the top 10 features set across the 200 random sampling iterations. We then analyzed the top 5 features selected for each modality, showing only 2 genus features as only 2 distinct genera were contained among the top 10 features across all random sampling iterations.

It seems that both the selected genus and gene expression features are sensible and are supported by previous research linking them to stomach cancer. The most commonly selected genus taxa was *Helicobacter*. This seems to be validated by previous studies linking STAD to *Helicobacter pylori* [33–35], which can induce gastritis, which can then lead to stomach adenocarcinoma. Furthermore, the genus *Lactobacillus* is also linked to stomach cancer and has a possible interaction with *H. pylori* [36,37]. Namely, patients with stomach cancer have been found to have a higher abundance of *Lactobacilli* in their gastric microbiota. For gene expression features, the second most frequently selected genus feature was the HOXC10 gene, which has been found to be differentially expressed in stomach cancer tissues versus normal tissues and significantly promote tumor development [38]. The third most selected feature was the PRSS21 gene, which was previously found to be among the most important biomarkers in a gene signature set for detecting metastasis in stomach cancer patients [27]. We did not find evidence of a link between TDRD9 and stomach cancer, suggesting that this gene might require further research.

These results are also consistent for the features selected when performing feature selection on the individual GE and genus modalities. For the GE modality, the selected features are similar to the GE features selected within the GE ∩ genus modality (Table S2), while the same holds for the features within the genus modality (Table S3). For the latter, the additional genera *Haemophilus*, *Fusobacterium* and *Streptococcus* were also selected. An abundance of *Streptococcus* has been linked to patients with cancer [37,39], *Haemophilus* has been linked to an increase in gastric cancer through the accumulation of nitrates [40], while *Fusobacterium* has been linked to worse prognosis in certain subtypes of gastric cancer [41].

These results indicate that the feature selection step is selecting biologically relevant features and that the lack of additional performance of the GE ∩ genus model when compared to the GE model and the lack of performance of the genus model is likely not due to the feature selection process leading to biologically irrelevant features being selected.

Table 3: Top GE and genus features selected when performing feature selection on the GE ∩ genus data set with an elastic net model and a feature selection number of 10. The table rows denote the name of the top selected GE and genus features, while the Frequency Selected column denotes the percentage of times the feature made it into the set with the top 10 most important features across 200 random sampling iterations.

Rank	Feature name	Frequency selected	Feature type
1	TDRD9	84%	GE
2	HOXC10	83%	GE
3	PRSS21	78.5%	GE
4	HOXA13	67%	GE
5	HOXC9	62%	GE
6	*Helicobacter*	53%	Genus
136	*Lactobacillus*	0.5%	Genus

Using complex integration method does not improve performance of integrated set

To determine whether the lack of performance improvement when integrating the two modalities was due to the simple concatenation-based integration method, we attempted to integrate the two modalities using a more advanced and proven integration method. Namely, an autoencoder (AE), which has successfully been used to integrate multi-omics host features [2,30], partially due to its ability to capture nonlinear relationships between features, and nonnegative matrix factorization (NMF), which has also successfully been used to integrate biological features [42]. Both models can perform feature extraction by capturing important information within the feature set and condensing it to a smaller feature space.

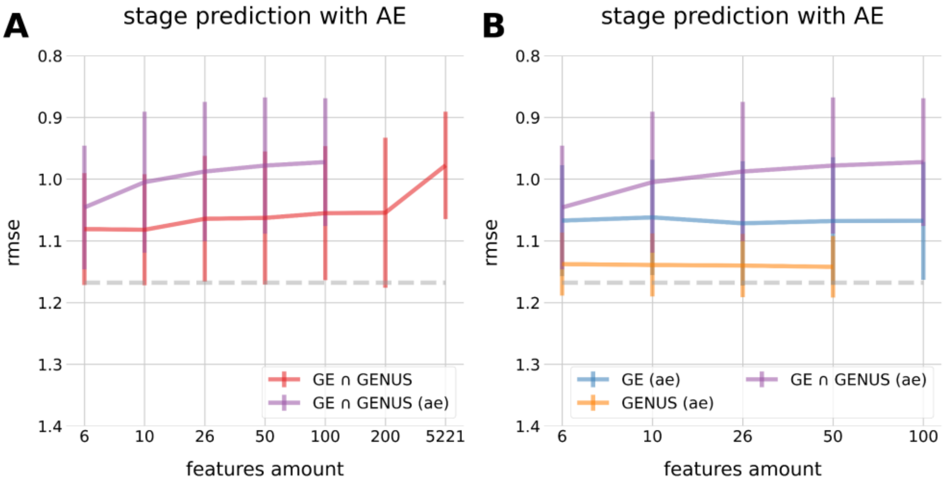

Figure 8: RMSE for the STAD stage prediction endpoint with the use of an elastic net model for the stage prediction and feature selection. ***A****, Comparison of the prediction model built using the genus + gene expression features integrated with an autoencoder, denoted as GE ∩ genus (AE), with the performance of the model built using the regular GE ∩ genus dataset. As the GE ∩ genus (AE) set contained 100 features, the performance shown at the features amount of 100 for the model built with this modality indicates the prediction performance when all the extracted features were used.* ***B,*** *Comparison of the prediction models built using the GE ∩ genus (AE) data, the AE-integrated GE data, denoted as GE (AE), and the AE-integrated genus data, denoted as genus (AE). The line for the genus (AE) model ends at a features amount of 50, as there were only 50 total features extracted for this modality.*

To evaluate the performance of the complex integration methods, we integrated the 5221 features in the GE ∩ genus set and obtained a new integrated feature set of 100 features. We then trained an elastic net prediction model on the integrated data. For the autoencoder, we based our model on a deep autoencoder architecture successfully used by Chaudhary et al. to integrate host multi-omics data for liver cancer [2], while for NMF, we selected the amount of extracted components to be equal to the AE model to allow for comparison between the methods.

We found that there was no additional improvement with the holo-omics approach when using complex integration over the simple concatenation-based approach for either the AE (p-value = 0.5057) model (Figure 8A) or the NMF (p-value = 0.3482) model (Figure 9A). Both models do converge to the optimal performance with far fewer features than the simple concatenation-based approach, indicating that these models can effectively capture a latent representation of the integrated features.

To investigate whether the ability of these complex integration methods to obtain the same prediction performance as the GE ∩ genus model with far fewer features was dependent on integrating both the GE and genus modalities, or whether running the complex integration method on the modalities separately would achieve the same results, we built prediction models using extracted features by running the AE and NMF on each modality separately. We denote the AE and NMF-integrated GE modalities as GE (AE) and GE (NMF), respectively, and the integrated genus modalities as genus (AE) and genus (NMF).

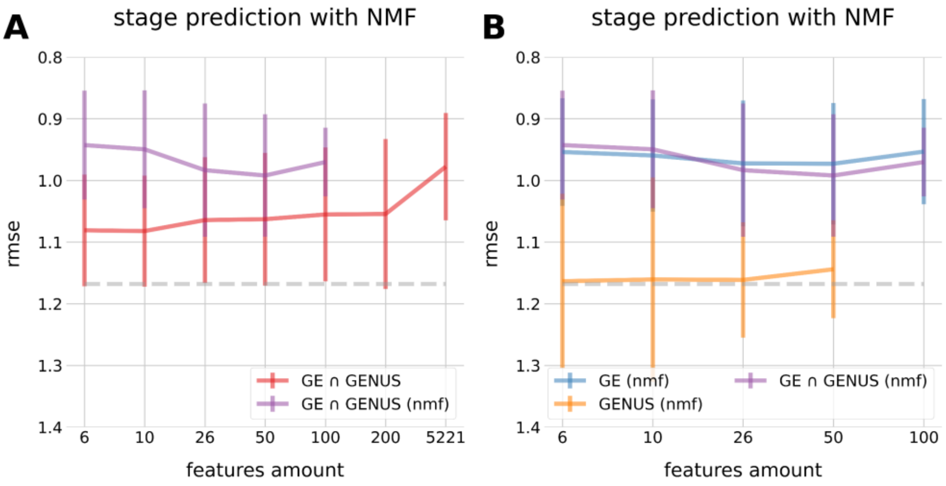

*Figure 9: RMSE for the STAD stage prediction endpoint with the use of an elastic net model for the stage prediction and feature selection. **A**, Comparison of the prediction model built using the genus + gene expression features integrated using nonnegative matrix factorization, denoted as GE ∩ genus (NMF), with the performance of the model built using the regular GE ∩ genus dataset. **B**, Comparison of the prediction models built using the GE ∩ genus (NMF) data, the NMF integrated GE data, denoted as GE (NMF), and the NMF integrated genus data, denoted as genus (NMF).*

It appears that using complex integration on the GE modality alone is enough to obtain the same prediction performance as the GE ∩ genus model with far fewer features, which indicates that yet again, the GE data contains enough information for the prediction task and that the genus data does not add additional information. The performance of the prediction model trained using the AE-integrated GE ∩ genus data performed significantly better than the one using AE-integrated GE data (p-value = $1.6*10^{-18}$), which could have indicated that both the GE and genus modalities are necessary to obtain the optimal prediction performance with fewer features (Figure 8B). However, the GE ∩ genus (NMF) model performs as optimally as the GE ∩ genus (AE) one, while the GE (NMF) model does perform similarly to its GE ∩ genus (NMF) counterpart (Figure 9B). Thus, this suggests that the lower prediction performance of the GE (AE) model in comparison to the GE ∩ genus (AE) model is due to the autoencoder architecture not being able to capture the full information of the GE data, rather than the genus data inherently being necessary to achieve a competitive prediction performance with a smaller extracted features set. Ultimately, we conclude that the genus data is not necessary to obtain an informative extracted feature set and that the lack of improvement with the holo-omic approach is unlikely to be due to the integration method used.

Discussion

It seems that across prediction targets, prediction models, feature selection methods, integration methods and cancers, integrating the gene expression modality with the genus abundance modality does not offer additional predictive power over using the GE modality individually, and that the genus modality provides the least predictive power when used alone. This does not necessarily mean that the holo-omics approach never leads to an improvement in performance, but that it is not the case with the specific TCMA genus abundance and TCGA gene expression data set used.

One of the major reasons for the lack of performance improvement with the holo-omics approach used when compared to the host omics approach is likely that gene expression data is information-dense, being able to recapitulate much of the information of other host omics layers and possibly even intra-tumoral genus abundance data. Namely, gene expression data is known to recapitulate the information of more ‘upstream’ datatypes, such as gene mutation and methylation data [3,32]. In models that integrate GE with other host omics features, GE features can end up dominating prediction models and providing most of the discriminatory information [32]. Thus, using a less informative host omics datatype might have shown additional improvement when using the genus data. Additionally, gene expression can be correlated with and be affected by microbial abundance groups [5], which follows from the general fact that the interactions between cancer and the microbiome can be bidirectional, because cancer can lead to an environment that fosters certain microbiota, which in turn can affect the cancer [43,44]. This can lead to gene expression data possibly recapitulating microbial abundance data information.

Additionally, the microbiome might occasionally contain limited information on the development and properties of cancer tissues, as the interaction between cancer and the microbiome can also be unidirectional, with the cancer affecting the microbiome but not the other way around. While there is evidence of a causal link between the human microbiome and certain cancers (e.g. *Helicobacter pylori* causing gastric cancer and hepatitis B or C causing liver cancer [43]), this causal link has only been proven for a limited amount of microbes and cancers. Especially for intratumoral bacteria, such as the TCGA-mined genus data used in this study, the presence of microbial communities might simply be due to an infection of existing tumors, rather than the tumor development actively being influenced by the microbes [11,43]. This lack of (causal) link between the microbiome and cancer can make it difficult to determine whether cancer tissues are consistently associated with a specific microbial composition, especially as tumor tissue-resident microbial composition could be sporadic and defined by the transient and random movement of microbes [45].

For the TCMA genus abundance data, the lack of informative value of the genus data in the integrated model might have been due to the removal of valuable information during the data collection process. As mentioned, this data was previously mined from existing TCGA whole genome sequencing data. This data had to be cleared of contaminants, especially for low biomass samples such as the human tumor microbiome, as contamination can arise during sample collection and due to laboratory environment [5,17]. During this decontamination step, the creators of TCMA used a statistical technique that analyzed microbial abundance data within and across tissues and eliminated those that it found likely to result from contamination. While the authors validated their approach by comparing the mined microbial abundance distribution with that of the original matched TCGA samples, this was only done with 8 samples, and only with colorectal cancer samples. Even for these samples, differences remained between the mined tumor microbiome data and the microbiome of the original samples. In the end, the decontamination process could have removed valid and informative microbial information, which is a risk when mining TCGA data [8]. Additionally, it is worth noting that we only considered TCMA genus samples for which we also found matching TCGA GE samples in this study. Thus, this excluded some TCMA samples, which could have further led to valuable information loss.

Furthermore, the lack of samples combined with the high amount of variation in the microbiome can make it harder for a prediction model to properly capture the relevant variation between individuals that leads to differing disease states. The microbiome exhibits significant person-to-person variation [16,46], and is variable across multiple axes such as age, geography, diet [46], gender [19] and time [44]. It might be necessary to examine microbes on the functional level, as some of this variation, such as the person-to-person variation, might disappear when examining the function of expressed microbial genes [46]. Additionally, while this variation can complicate model learning as it is not directly relevant to the disease state, a lack of variation and specificity in certain aspects of the TCMA microbial data used could have also complicated the model learning. For example, examining microbes on the genus level might, in certain instances, not provide enough information for discriminating between disease states, as there is less variation between individuals on the genus level than the species and strain level [47]. For tumor versus normal tissue differentiation, one genus could contain species that are correlated with tumor tissues, but also those correlated with normal tissues [5]. Thus, species-level abundance data might be needed to capture the relevant abundance differences between disease states. Additionally, it might be necessary to also look at the active expression of microbial genes, as microbial data mined from TCGA cannot determine whether microbial reads were intra- or extracellular or from alive or dead bacteria [10].

Thus, to separate microbial variation which is not relevant to the disease state from the truly discriminatory information, a higher quantity, but also more specific data might be needed. This includes clinical variables such as gender, geographical information, species- or strain level microbial data, microbial gene expression (functional) information, and possibly also repeated measurements of the microbiome at different times. Additionally, normal tissues from healthy patients rather than NAT tissues might be needed, as there can be similarities between tumor and NAT tissues [11], possibly due to the movement of microbes across these tissue types [17].

To obtain more meaningful results and truly capture the predictive power of microbial abundance data and its relation to host omics data, it might be necessary to have more specific microbial information, such as species-level information, metagenomic or transcriptonomic (functional) data, which could give a more direct measure of microbial activity and function [45], data on the virome and mycobiome, which are also associated with cancer [45] and the location and organization in the tissue of the microbial data, which could also impact tumorigenesis [43]. Finally, we would also require a validated decontamination pipeline that removes less valid microbial information, although it is unclear whether this is possible by mining TCGA data.

Conclusion

We investigated whether combining host omics and microbial information offered additional power compared to using the modalities individually. To this end, we combined TCMA tissue-resident genus abundance data with TCGA gene expression data and examined their utilities by comparing the performances of prediction models built using the individual and integrated modalities. This was done using data from STAD samples for the stage prediction and tumor versus normal prediction endpoints, and validated using ESCA, HNSC and COAD data. To our knowledge, this is the first study that used modern TCGA-mined (TCMA) microbial abundance data to examine the utility of integrating host omics and microbial information through the use of machine learning algorithms for multiple cancer diagnostic endpoints.

It appears that the holo-omics approach does not outperform the single host-omic approach and that the microbiome-only approach performs the worst. Furthermore, it is unlikely that these results are due to the specific feature selection method and process used, the prediction model or the method of integration. In general, it is clear that the human microbiome has an effect on cancer aetiology, however, for certain data sets, the prediction performance of just gene expression alone might be enough to capture the underlying patterns relevant for cancer diagnostics. This is most likely due to gene expression data being information-dense and recapitulating microbial information, and the high variability and unestablished causality of the microbiome with respect to cancer.

Finally, these results likely cannot be generalized to other holo-omic datasets, as the TCMA microbial dataset used was not well validated and contained a limited amount of data. The lower disease phenotype-related information density combined with high variation of the microbiome means that care should be taken to use an appropriate quantity and quality of microbial data to properly investigate a holo-omic approach. Thus, more specific data might be needed, such as of expressed microbial genes, species-level microbial information and other types of non-bacterial microbes. Additionally, it might be necessary to use non TCGA-mined microbial datasets, or TCGA-mined microbial datasets that use better validated decontamination methods. Future work could explore other diagnostic endpoints as well as performance differences when integrating microbial information with other host omics types.

Supplementary material

Supplementary tables

Table S1: Non-zero and zero features for each cancer in the TCMA genus relative abundance data set. Zero features are defined as features for which the relative abundance was 0 across all samples. In other words, these are the genera which were not detected in any patients with the relevant cancer type.

Cancer	Nonzero	Zero	Total
STAD	52	169	221
COAD	82	139	221
ESCA	61	160	221
HNSC	69	152	221

Table S2: top gene expression features selected when performing feature selection on the individual GE data set with elastic net and a feature selection number of 10. The table rows denote the name of the top selected GE features, while the Frequency Selected column denotes the percentage of times the feature was selected across 200 random sampling iterations.

Rank	Feature name	Frequency selected	Feature type
1	TDRD9	94%	GE
2	PRSS21	83.5%	GE
3	HOXC10	81.5%	GE
4	HOXA13	72.5%	GE
5	HOXC9	70.5%	GE

Table S3: top genus features selected when performing feature selection on the individual genus data set with elastic net and a feature selection number of 10. The table rows denote the name of the top selected genus features, while the Frequency Selected column denotes the percentage of times the feature was selected across 200 random sampling iterations.

Rank	Feature name	Frequency selected	Feature type
1	*Helicobacter*	100%	Genus
2	*Lactobacillus*	94.5%	Genus
3	*Haemophilus*	79%	Genus
4	*Fusobacterium*	67%	Genus
5	*Streptococcus*	58.5%	Genus

Supplementary figures

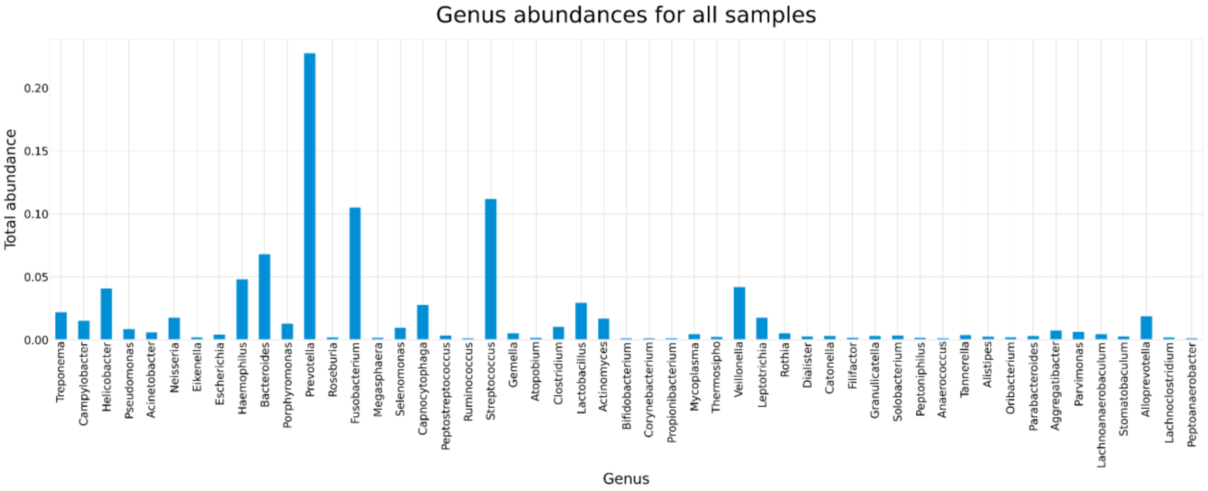

Figure S1: Summed relative abundances across all cancer samples in the TCMA dataset for the 52 genera for which the total relative abundance is larger than 0.01. Note that this only includes TCMA samples for which there were also paired TCGA GE data available.

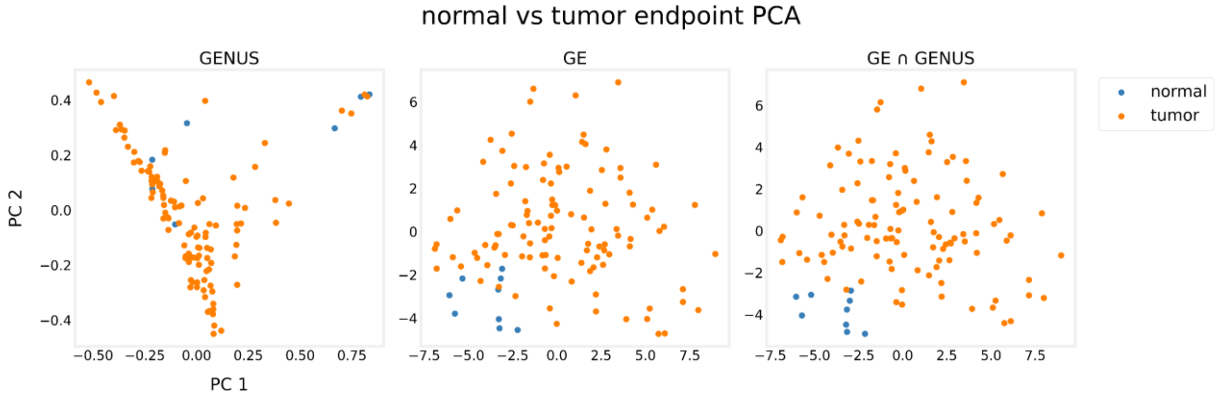

Figure S2: PCA of STAD (stomach adenocarcinoma) for all modalities for the tumor versus normal endpoint. The first graph contains the PCA for the genus abundance data (genus), the second graph for the gene expression data (GE) and the third graph for the concatenated genus + gene expression features (GE ∩ genus). The horizontal axis displays the first principal component, while the vertical axis displays the second principal component of the PCA. Finally, samples in red denote tumor samples while those in blue denote normal samples.

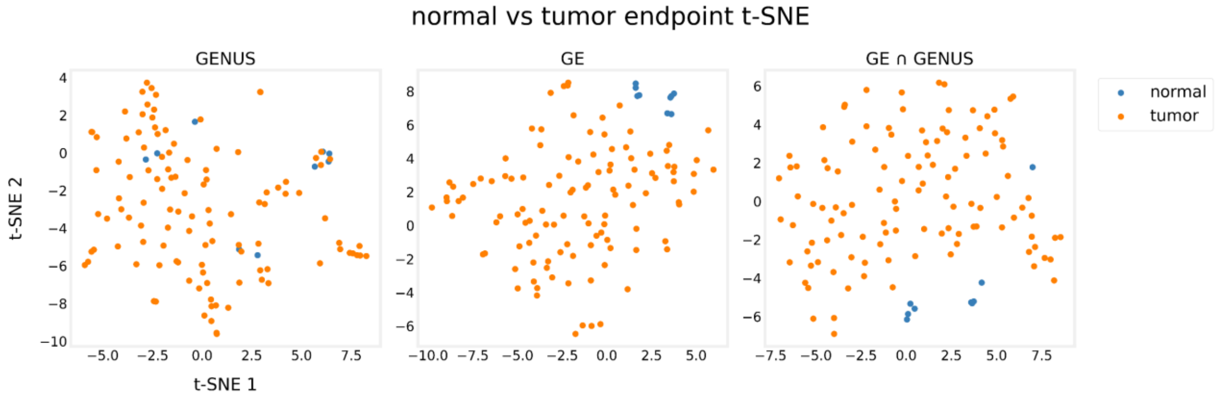

Figure S3: t-SNE of STAD (stomach adenocarcinoma) for all modalities for the tumor versus normal endpoint. The first graph contains the PCA for the genus abundance data (genus), the second graph for the gene expression data (GE) and the third graph for the concatenated genus + gene expression features (GE ∩ genus). The horizontal axis displays the first t-SNE

component, while the vertical axis displays the second t-SNE component. Finally, samples in red denote tumor samples while those in blue denote normal samples.

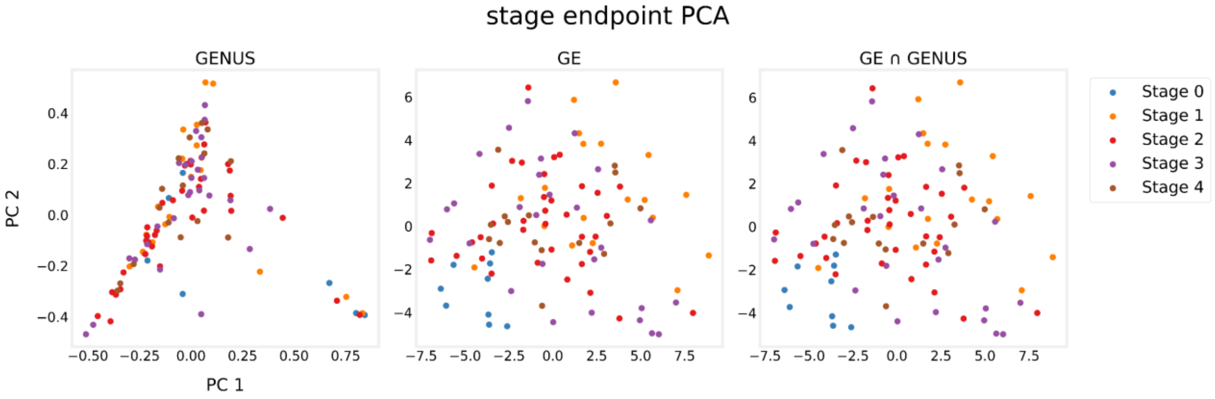

Figure S4: PCA of STAD (stomach adenocarcinoma) for all modalities for the stage endpoint. The first graph contains the PCA for the genus abundance data (genus), the second graph for the gene expression data (GE) and the third graph for the concatenated genus + gene expression features (GE ∩ genus). The horizontal axis displays the first principal component, while the vertical axis displays the second principal component of the PCA. Finally, the different colored points represent the cancer stage of the different samples, with stage one being a normal non-tumor sample.

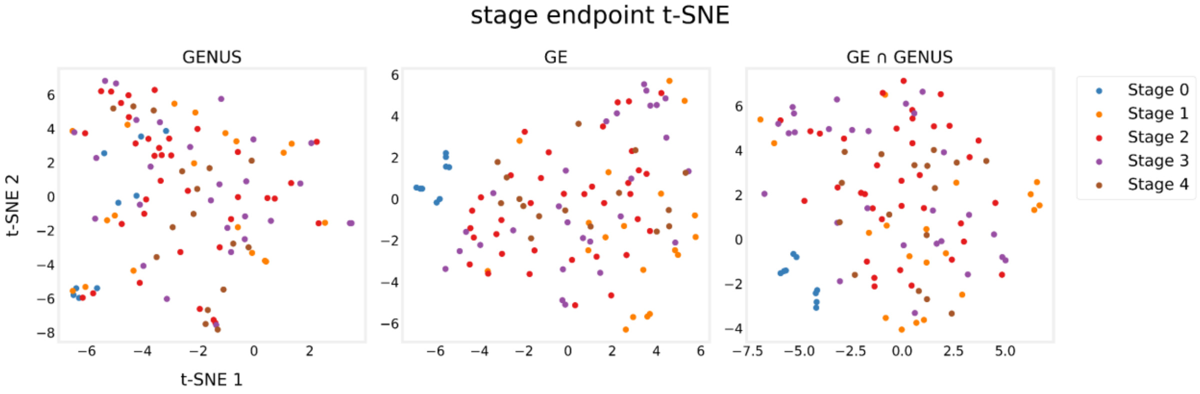

Figure S5: t-SNE of STAD (stomach adenocarcinoma) for all modalities for the stage endpoint. The first graph contains the PCA for the genus abundance data (genus), the second graph for the gene expression data (GE) and the third graph for the concatenated genus + gene expression features (GE ∩ genus). The horizontal axis displays the first t-SNE component, while the vertical axis displays the second t-SNE component. Finally, samples in red denote tumor samples while those in blue denote normal samples.

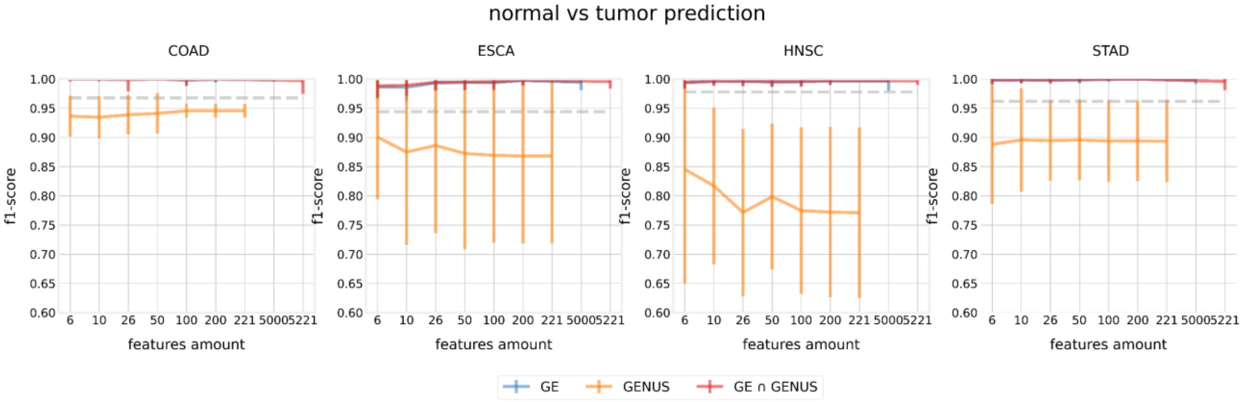

Figure S6: f1-score for the tumor versus normal prediction endpoint for COAD (colon adenocarcinoma), ESCA (esophageal carcinoma), HNSC (head and neck squamous carcinoma) and STAD (stomach adenocarcinoma), respectively. Each graph contains the f1-score for each modality (i.e. genus abundance data (genus), gene expression data (GE) and the concatenated genus + gene expression features (GE ∩ genus)). The endpoints of each horizontal line segment indicate the

average f1-score across 200 random sampling iteration, while the vertical line segments indicate the standard deviation of the f1-score across these iterations.

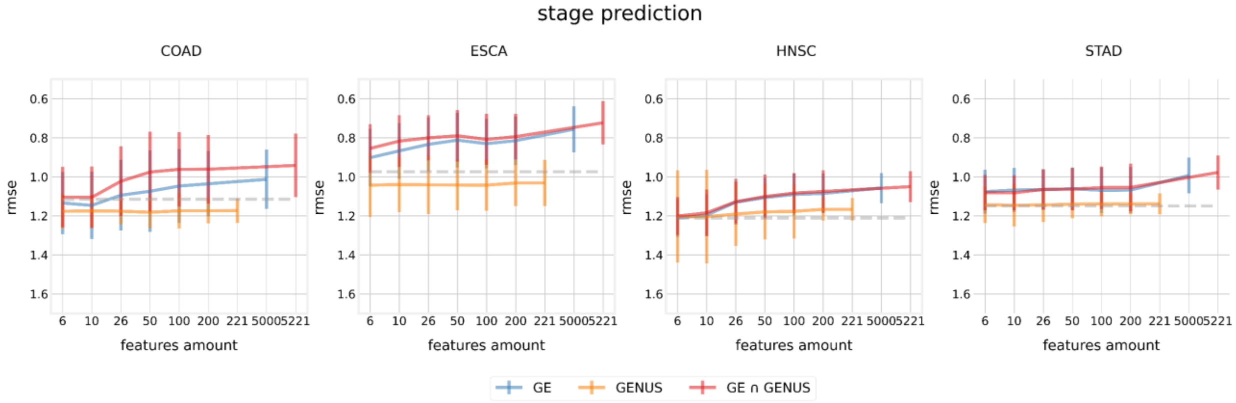

Figure S7: root-mean-squared-error (RMSE) for the stage prediction endpoint for COAD (colon adenocarcinoma), ESCA (esophageal carcinoma), HNSC (head and neck squamous carcinoma) and STAD (stomach adenocarcinoma), respectively. Each graph contains the root-mean-squared-error for each modality (i.e. genus abundance data (genus), gene expression data (GE) and the concatenated genus + gene expression features (GE ∩ genus)). The endpoints of each horizontal line segment indicate the average RMSE across 200 random sampling iterations, while the vertical line segments indicate the standard deviation of the RMSE across these iterations.

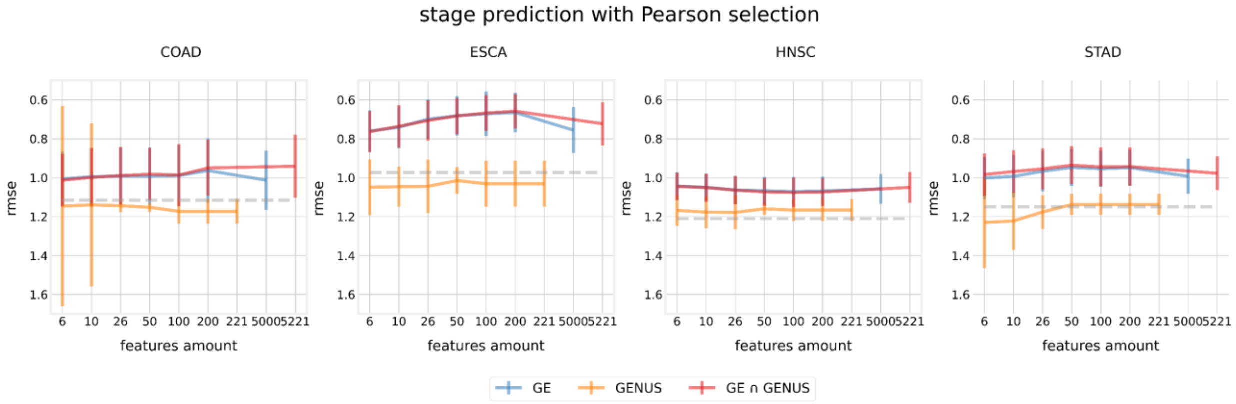

Figure S8: root-mean-squared-error (RMSE) for the stage prediction endpoint for COAD, ESCA, HNSC and STAD, respectively. Each graph contains the root-mean-squared-error for the genus, GE and GE ∩ genus. The results are shown for each feature selection amount (horizontal axis) when using the Pearson correlation coefficient for feature selection and an elastic net model for the prediction.

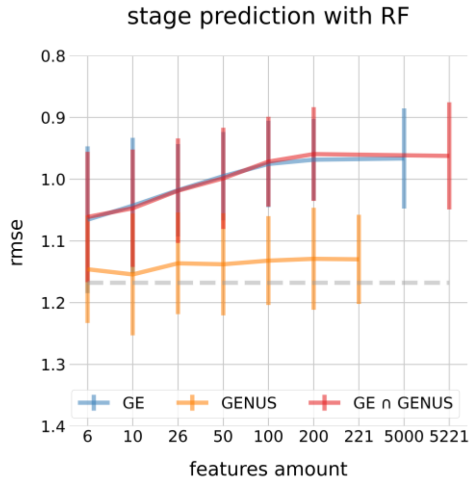

Figure S9: Predictive performance of stage prediction for STAD. The lines contain the RMSE for the random forest regressor models trained and tested on the genus, GE and GE ∩ genus sets. The results are shown for each feature selection amount (horizontal axis) using features selected by an elastic net model.

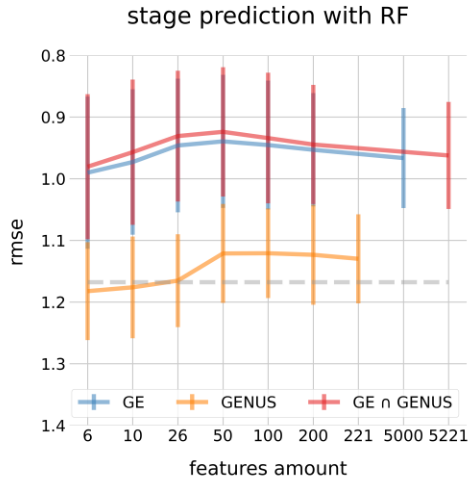

Figure S10: Predictive performance of stage prediction for STAD. The lines contain the RMSE for the random forest regressor models trained and tested on the genus, GE and GE ∩ genus sets. The results are shown for each feature selection amount (horizontal axis) using features selected with the Pearson correlation coefficient.

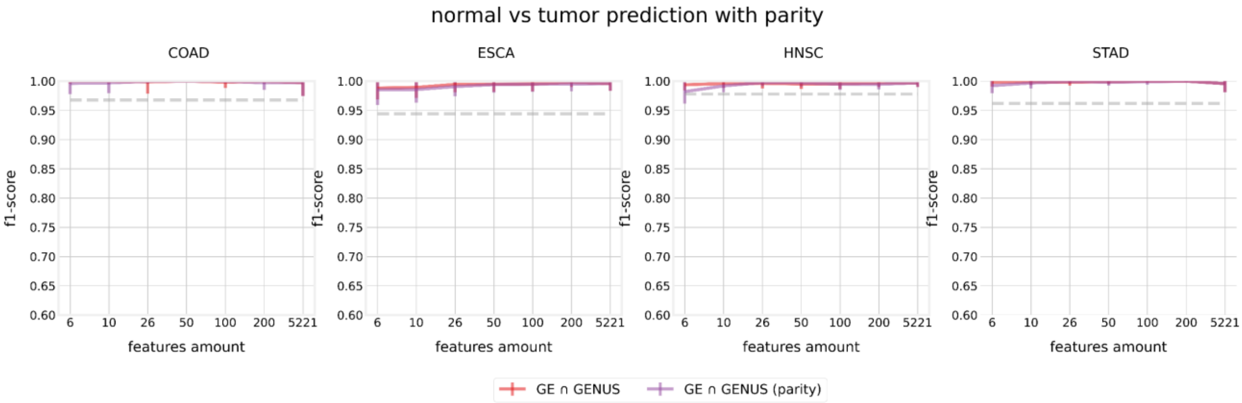

Figure S11: f1-score for the tumor versus normal prediction endpoint for COAD, ESCA, HNSC and STAD with the use of an support vector machine model for prediction and ANOVA for feature selection. Here, we compare the performance of the model when trained using concatenated genus and gene expression features with enforced parity during feature selection,

denoted as GE ∩ genus (parity), with the performance of the model when trained using the regular (concatenated) integrated data (GE ∩ genus). The performance of the model with a features amount of 5221 denotes the model performance when using all features (i.e. when there is no feature selection) and is thus identical for the two models.

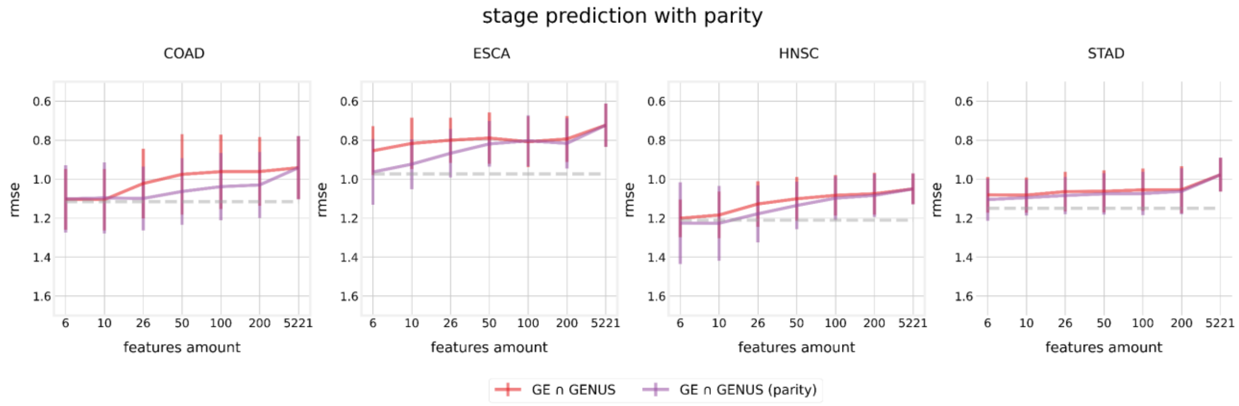

Figure S12: RMSE for the stage prediction endpoint for COAD, ESCA, HNSC and STAD with the use of an elastic net model for prediction and feature selection. Here, we compare the performance of the model when trained using concatenated genus and gene expression features with enforced parity during feature selection, denoted as GE ∩ genus (parity), with the performance of the model when trained using the regular (concatenated) integrated data (GE ∩ genus).

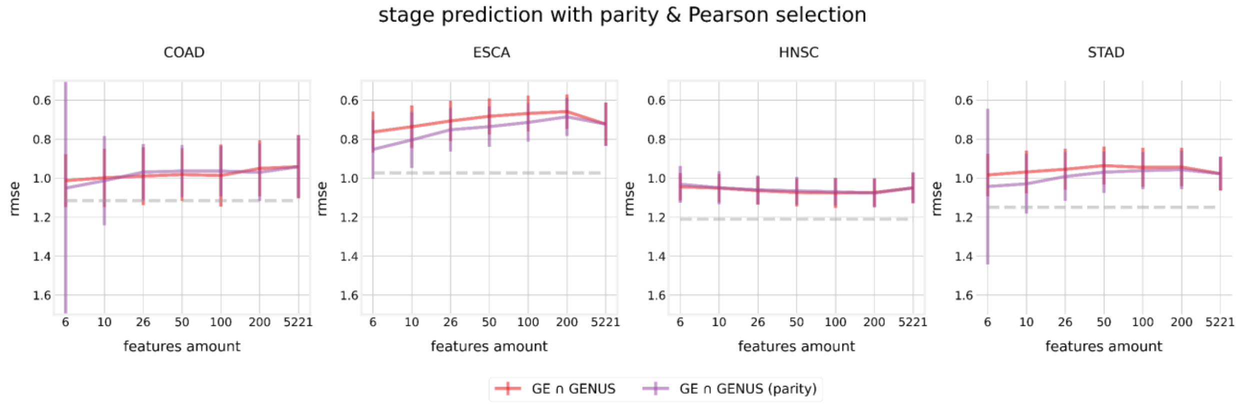

Figure S13: RMSE for the stage prediction endpoint for COAD, ESCA, HNSC and STAD with the use of an elastic net model for prediction and the Pearson correlation coefficient for feature selection. Here, we compare the performance of the model when trained using concatenated genus and gene expression features with enforced parity during feature selection, denoted as GE ∩ genus (parity), with the performance of the model when trained using the regular (concatenated) integrated data (GE ∩ genus).

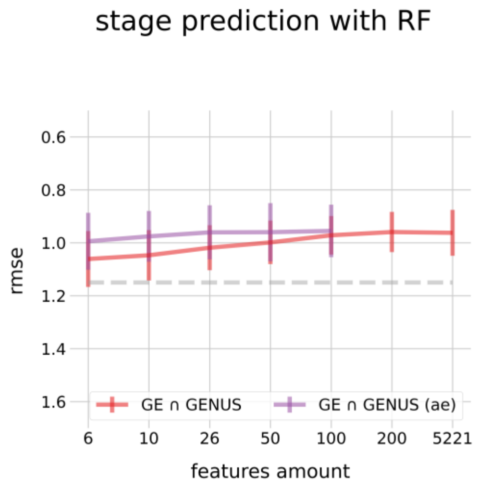

Figure S14: Predictive performance of stage prediction for STAD. The lines contain the RMSE for the random forest regressor models trained and tested on the GE ∩ genus and the feature set consisting of the AE-integrated GE ∩ genus features, denoted as GE ∩ genus (AE). The results are shown for each feature selection amount (horizontal axis) using features selected using an elastic net model.

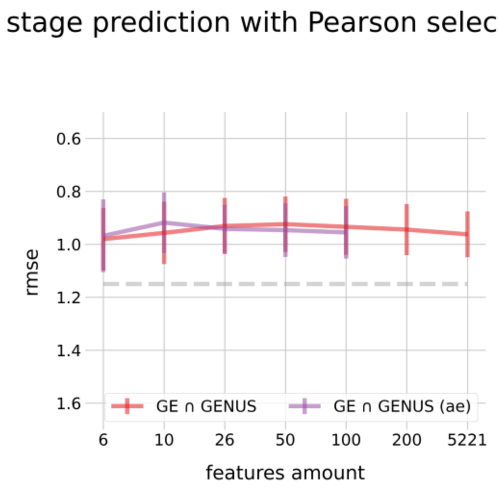

Figure S15: Predictive performance of stage prediction for STAD. The lines contain the RMSE for the random forest regressor models trained and tested on the GE ∩ genus and the feature set consisting of the AE-integrated GE ∩ genus features, denoted as GE ∩ genus (AE). The results are shown for each feature selection amount (horizontal axis) using features selected using the Pearson correlation coefficient.

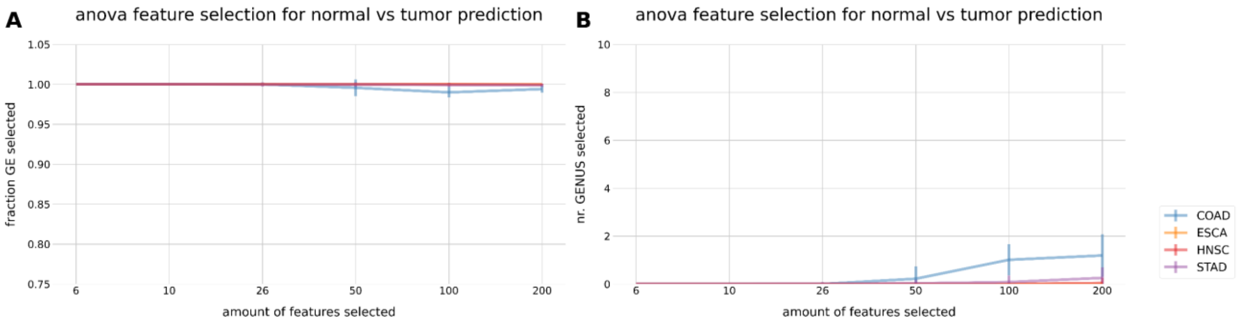

Figure S16: the fraction of GE features (vertical axis) selected with ANOVA from the total amount of features for each feature amount (horizontal axis) when performing tumor versus normal prediction with the genus + gene expression modality (GE ∩ genus). Each line displays the results for a different cancer, namely COAD (colon adenocarcinoma), ESCA (esophageal carcinoma), HNSC (head and neck squamous carcinoma) and STAD (stomach adenocarcinoma).

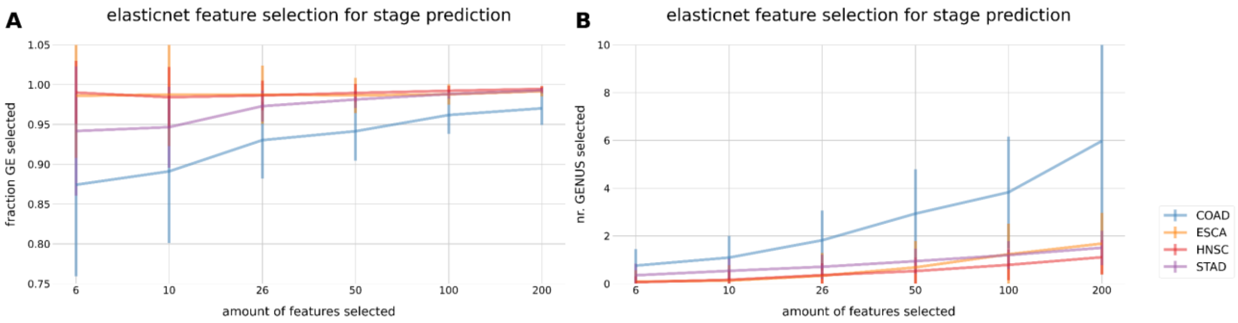

Figure S17: the fraction of GE features (vertical axis) selected using an elastic net from the total amount of features for each feature amount (horizontal axis) when performing stage prediction with the genus + gene expression modality (GE ∩ genus).

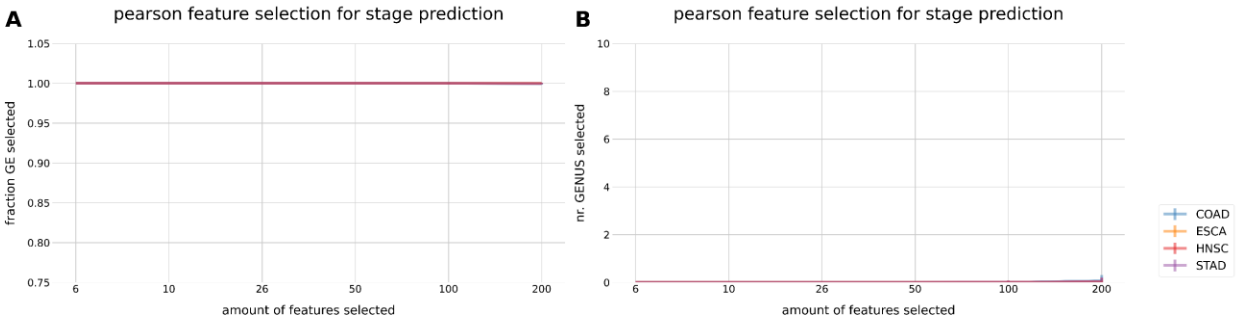

Figure S18: the fraction of GE features (vertical axis) selected using the Pearson correlation coefficient from the total amount of features for each feature amount (horizontal axis) when performing stage prediction with the genus + gene expression modality (GE ∩ genus).

References

1. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. Review<br>The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol. Onkol.* **2015**, 68–77 (2015).
2. Chaudhary, K., Poirion, O. B., Lu, L. & Garmire, L. X. Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clin. Cancer Res.* **24**, 1248–1259 (2018).
3. Duan, R. *et al.* Evaluation and comparison of multi-omics data integration methods for cancer subtyping. *PLOS Comput. Biol.* **17**, e1009224 (2021).
4. Machiraju, G., Amar, D. & Ashley, E. Multi-omics factorization illustrates the added value of deep learning approaches. 7.
5. Dohlman, A. B. *et al.* The cancer microbiome atlas: a pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants. *Cell Host Microbe* **29**, 281-298.e5 (2021).
6. Ehrlich, S. D. MetaHIT: The European Union Project on Metagenomics of the Human Intestinal Tract. in *Metagenomics of the Human Body* (ed. Nelson, K. E.) 307–316 (Springer, 2011). doi:10.1007/978-1-4419-7089-3_15.
7. Proctor, L. M. *et al.* The Integrative Human Microbiome Project. *Nature* **569**, 641–648 (2019).

8. Robinson, K. M., Crabtree, J., Mattick, J. S. A., Anderson, K. E. & Dunning Hotopp, J. C. Distinguishing potential bacteria-tumor associations from contamination in a secondary data analysis of public cancer genome sequence data. *Microbiome* **5**, 9 (2017).

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

10. Hermida, L. C., Gertz, E. M. & Ruppin, E. Predicting cancer prognosis and drug response from the tumor microbiome. *Nat. Commun.* **13**, 2896 (2022).

11. Greathouse, K. L. *et al.* Interaction between the microbiome and TP53 in human lung cancer. *Genome Biol.* **19**, 123 (2018).

12. Wang, Y., Wang, Y. & Wang, J. A comprehensive analysis of intratumor microbiome in head and neck squamous cell carcinoma. *Eur. Arch. Otorhinolaryngol.* (2022) doi:10.1007/s00405-022-07284-z.

13. Chakladar, J. *et al.* The Pancreatic Microbiome is Associated with Carcinogenesis and Worse Prognosis in Males and Smokers. *Cancers* **12**, 2672 (2020).

14. Erickson, A. R. *et al.* Integrated Metagenomics/Metaproteomics Reveals Human Host-Microbiota Signatures of Crohn's Disease. *PLoS ONE* **7**, e49138 (2012).

15. Garrett, W. S. The gut microbiota and colon cancer. *Science* **364**, 1133–1135 (2019).

16. Knippel, R. J., Drewes, J. L. & Sears, C. L. The Cancer Microbiome: Recent Highlights and Knowledge Gaps. *Cancer Discov.* **11**, 2378–2395 (2021).

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

18. Kwon, M., Seo, S.-S., Kim, M. K., Lee, D. O. & Lim, M. C. Compositional and Functional Differences between Microbiota and Cervical Carcinogenesis as Identified by Shotgun Metagenomic Sequencing. *Cancers* **11**, 309 (2019).

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

20. Riquelme, E. *et al.* Tumor Microbiome Diversity and Composition Influence Pancreatic Cancer Outcomes. *Cell* **178**, 795-806.e12 (2019).

21. Elinav, E., Garrett, W. S., Trinchieri, G. & Wargo, J. The cancer microbiome. *Nat. Rev. Cancer* **19**, 371–376 (2019).

22. Sivan, A. *et al.* Commensal Bifidobacterium promotes antitumor immunity and facilitates anti–PD-L1 efficacy. *Science* **350**, 1084–1089 (2015).

23. Gopalakrishnan, V. *et al.* Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. *Science* **359**, 97–103 (2018).

24. Nyholm, L. *et al.* Holo-Omics: Integrated Host-Microbiota Multi-omics for Basic and Applied Biological Research. *iScience* **23**, 101414 (2020).

25. Alberdi, A., Andersen, S. B., Limborg, M. T., Dunn, R. R. & Gilbert, M. T. P. Disentangling host–microbiota complexity through hologenomics. *Nat. Rev. Genet.* 1–17 (2021) doi:10.1038/s41576-021-00421-0.

26. Limborg, M. T. *et al.* Applied Hologenomics: Feasibility and Potential in Aquaculture. *Trends Biotechnol.* **36**, 252–264 (2018).

27. Izumi, D. *et al.* A genomewide transcriptomic approach identifies a novel gene expression signature for the detection of lymph node metastasis in patients with early stage gastric cancer. *EBioMedicine* **41**, 268–275 (2019).

28. Way, G. P. & Greene, C. S. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* **23**, 80–91 (2018).

29. Sen Puliparambil, B., Tomal, J. H. & Yan, Y. A Novel Algorithm for Feature Selection Using Penalized Regression with Applications to Single-Cell RNA Sequencing Data. *Biology* **11**, 1495 (2022).

30. Ding, M. Q., Chen, L., Cooper, G. F., Young, J. D. & Lu, X. Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics. *Mol. Cancer Res.* **16**, 269–278 (2018).

31. Koul, N. & Manvi, S. S. Feature Selection From Gene Expression Data Using Simulated Annealing and Partial Least Squares Regression Coefficients. *Glob. Transit. Proc.* **3**, 251–256 (2022).

32. Aben, N., Vis, D. J., Michaut, M. & Wessels, L. F. TANDEM: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. *Bioinformatics* **32**, i413–i420 (2016).

33. Peek, R. M. & Blaser, M. J. Helicobacter pylori and gastrointestinal tract adenocarcinomas. *Nat. Rev. Cancer* **2**, 28–37 (2002).

34. Blaser, M. J. *et al.* Infection with Helicobacter pylori Strains Possessing cagA Is Associated with an Increased Risk of Developing Adenocarcinoma of the Stomach1. *Cancer Res.* **55**, 2111–2115 (1995).

35. Parsonnet, J. *et al.* Helicobacter pylori Infection and the Risk of Gastric Carcinoma. *N. Engl. J. Med.* **325**, 1127–1131 (1991).

36. Noto, J. M. & Jr, R. M. P. The gastric microbiome, its interaction with Helicobacter pylori, and its potential role in the progression to stomach cancer. *PLOS Pathog.* **13**, e1006573 (2017).

37. Nardone, G. & Compare, D. The human gastric microbiota: Is it time to rethink the pathogenesis of stomach diseases? *United Eur. Gastroenterol. J.* **3**, 255–260 (2015).

38. Guo, C., Hou, J., Ao, S., Deng, X. & Lyu, G. HOXC10 up-regulation promotes gastric cancer cell proliferation and metastasis through MAPK pathway. *Chin. J. Cancer Res.* **29**, 572–580 (2017).

39. Gantuya, B. *et al.* Gastric Microbiota in Helicobacter pylori-Negative and -Positive Gastritis Among High Incidence of Gastric Cancer Area. *Cancers* **11**, 504 (2019).

40. Forsythe, S. J. & Cole, J. A. Nitrite Accumulatin during Anaerobic Nitrate Reduction by Binary Suspensions of Bacteria Isolated from the Achlorhydric Stomach. *Microbiology* **133**, 1845–1849 (1987).

41. Ellen, T. B. *et al.* Fusobacterium nucleatum is associated with worse prognosis in Lauren's difuse type gastric cancer patients. *Sci. Rep.* **10**, 1–12 (2020).

42. Hamamoto, R. *et al.* Application of non-negative matrix factorization in oncology: one approach for establishing precision medicine. *Brief. Bioinform.* **23**, bbac246 (2022).

43. Pope, J. L., Tomkovich, S., Yang, Y. & Jobin, C. Microbiota as a mediator of cancer progression and therapy. *Transl. Res.* **179**, 139–154 (2017).

44. Costello, E. K., Stagaman, K., Dethlefsen, L., Bohannan, B. J. M. & Relman, D. A. The Application of Ecological Theory Toward an Understanding of the Human Microbiome. *Science* **336**, 1255–1262 (2012).

45. Cullin, N., Azevedo Antunes, C., Straussman, R., Stein-Thoeringer, C. K. & Elinav, E. Microbiome and cancer. *Cancer Cell* **39**, 1317–1341 (2021).

46. Gosalbes, M. J. *et al.* Metagenomics of human microbiome: beyond 16s rDNA. *Clin. Microbiol. Infect.* **18**, 47–49 (2012).

47. Holmes, E., Li, J. V., Marchesi, J. R. & Nicholson, J. K. Gut Microbiota Composition and Activity in Relation to Host Metabolic Phenotype and Disease Risk. *Cell Metab.* **16**, 559–564 (2012).