Host- Microbiome Omics Integration for Cancer Analysis And Diagnostics

# Abstract

# Introduction

Cancer is one of the leading causes of death and is responsible for millions of deaths every year. The aetiology, morphology and progression of different cancers depends 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 availability of biological data. This data availability has made it more accessible for researchers to use various omics data to perform various tasks related to cancer diagnostics. This relates to the analysis and integration of both host and microbial omics data.

## Host multi-omics integration

The availability of omics data from multiple layers has spearheaded the field of multi-omics integration, where data from multiple biological omics layers are integrated for cancer diagnostics. These methods can deliver additional insights over single omics methods for cancer diagnosis. However, they also deal with additional challenges owing to the heterogeneity of the data, noise, high dimensionality and sparsity 1.

## Microbiome multi-omics integration

Next to host omics data, a promising field of research relates to the analysis of microbial omics data. There are many microorganisms which live in communities on different human tissues, called the human microbiota. Namely, an ecosystem of 10 to 100 trillion microorganisms encompassing 500 to 1000 unique species for each individual 2. 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 microbiota which can be exploited.

Other sources, such as the him project and MetaHIT, are also available, although they often contain data from tissue swabs and stool samples, which are not necessarily representative of the microbiome of internal organs 2. The integrative human microbiome project 16 is the second stage of the human microbiome project and contains data on both host and microbial omics layers.

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 blood, which is most likely to contain useful microbial contamination 5. However, the microbial reads in this data set are often a result of contamination 2.

Using this data, various aspects of the human microbiome are being investigated. It has become clear that the microbiome exhibits significant person-to-person variation 3 and can be affected by many factors such as diet, environmental exposure and lifestyle choices 4. Additionally, microbial communities are unique to each cancer type 5. This has provided a motivation for the analyses of the impact of microbial composition on disease phenotypes and progression.

In light of this, recent research has been elucidating the effect that the microbiota can have on a host organism 6. It has become clear that the microbiota can have a clear effect on the phenotype of humans. Many microorganisms are enriched in certain cancers or differentially expressed among healthy and unhealthy individuals 7. Even further, there are indications that composition and changes in the microbiota has a direct influence on oncogenesis 8. As an example, patients with Parkinson’s have a relatively lower incidence of multiple cancers, possibly through mechanisms involving the microbiota 9. It has also been revealed that there are significantly different abundances of microbes, which are also associated with different functional categories, among different disease groups for cervical cancer, even when controlling for multiple other clinical variables 7. As an example of a mechanism through which the microbiome can affect a patient, it has been discovered that certain bacteria can bind to and alter the function of immune system cells which infiltrate tumors, thereby affecting carcinogenesis and resistance to chemotherapy 4. However, they can also positively affect patient health, for example, by positively influencing immune system cells to promote antitumor immunity.

## Towards 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. Many studies use one or the other to understand different biological processes without considering their interplay 6. However, it has become clear that the host can alter the human microbiota and vice versa 10. Thus, the integration of host and microbiota omics data could help to better understand the aetiology and physiology of different cancers and provide new insights 8. 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 11.

There are a lack of specialized methods which are able to perform this kind of integrated analysis 6 even though the development of such tools could provide helpful new insights. Such a method would have to deal with multiple challenges, namely the nonuniformity or linearity of this interaction, the high dimensionality compared to low amount of samples and missing values 6. Next to this, it would also have to deal with problems typical to multi-omics integration such as high heterogeneity of data and noise. Finally, care would have to be taken as microbial samples can be plagued with batch effects and contaminants 2.

The development of such methods could help power multiple diagnostic goals, such as predicting cancer response to therapy by elucidating why certain immunotherapies work or fail in patients, providing insights into how cancers develop, aid in the development of microbial consortia to push out disease associated microorganisms from a gut or tumor, identify targets for vaccines or targets for therapies which reduce the amount of certain microorganisms, such as antibiotics or phage-based therapy 4. Finally, such methods could also be used to identify biomarkers and predict cancer versus normal samples.

## Research question

This paper aims to integrate host and metagenomics data for cancer patients in order to investigate whether a holistic view provides additional insights for cancer diagnostics versus simply using the layers individually. Specifically, it aims to investigate the question:

Does integrating host and microbial omics data provide additional power over using the individual layers?

In this case, power refers to prediction performance.

## Previous research

The first step for the analysis of microbial omics data is choosing which type of data to use. Microbial data is often obtained through either amplicon sequencing or shotgun sequencing. Amplicon sequencing amplifies specific regions of the microbial 16S rRNA gene while shotgun sequencing reads all genomic DNA in a sample. There are numerous crucial differences between these types of sequencing methods, which can affect the type of downstream analyses which can be performed.

Shotgun sequencing makes it easier to identify species, and occasionally even strains. It can identify all taxa, including fungi and viruses, instead of just bacteria and archaea. However, it might lead to sequencing of host DNA which can obscure results, contains more complex data and might have less reference genomes available to match to a certain taxa. Either way, it has provided valuable insight into the phylogeny, biodiversity, metabolic abilities and functional diversity of many organisms.

After obtaining the data, analysis of microbial data is often done through taxonomic analysis, which can use 16S rRNA sequencing data to analyze the types of microbes present in a tissue, or functional analysis, which studies the function of present microbes by identifying and characterizing exons or analyzing metabolites. Here as well, analyses has to deal with multiple challenges, one of the biggest being batch effects, as the same tools can give different results.

Taxonomical analysis often investigates whether certain taxa or species of bacteria are overexpressed in different cancer or healthy samples using hypotheses tests 2,5,12,13, possibly also by using predictor models (e.g. regression) 5,13. It is also possible to analyze whether certain subgroups of microbiota are more likely to be found together in one tissue versus another or analyze if they are associated with certain symptoms of a certain cancer 2. Additionally, hypothesis tests can be used to see if microbial abundance is predictive of overall survival 2,14 or progression 14. Studies often also investigate the relation with clinical factors such as gender or age 12–14, which can often be confounders. This can be done by using predictor models which use these clinical factors as features 13. Finally, it is also possible to combine metaproteomics with metagenomics to investigate functional, as well as taxonomical differences of microbes between healthy and diseased samples 15.

In terms of hologenomics, previous studies have attempted to combine microbial and host omics data to investigate correlations between bacterial co-abundance groups and host gene expression patterns 2,14, host mutations 12 or proteins 2. As stated, there are not a lot of studies which directly identify the relation between host and microbial omics data, and certainly not how both datatypes relate to patient phenotype. As host multi-omics integration for cancer diagnostics is already a thriving field, one promising direction could be to use host multi-omics integration methods for holo- omics data. As holo- omics data generates highly complex data sets which require feature reduction 10, a possible method which can be investigated is the use of autoencoders to reduce and extract features which could then possibly be used in a predictor model.

How well does it perform without the microbial data? For different methods?

# Materials and methods

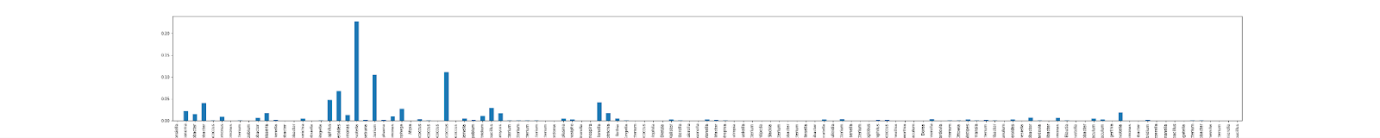
## Data

In order to explore the performance of cancer diagnostics in a holistic context, we integrated microbial data mined from patient tumor samples and host omics data mined from these same samples. We then only keep samples for which there is both microbial, as well as host omics data.

### Microbial data

This study used a statistical model to isolate tissue embedded microbial species present in TCGA samples from contaminants to obtain this data set, which was then subsequently validated using 16S rRNA amplicon sequencing on the original tissue samples. The resulting TCMA database accessible through the TCMA portal [[1]](#footnote-1), contains 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 highest specification that this data set reaches is the genus level. As the phylum level and everything above genus is less specific than the genus level, we decided to continue with only the genus taxonomical abundance data, which contains 221 taxa. The following is a visual exploration of the abundance data. There are 119 nonzero features among the overlapped samples, which can be seen below.



### Host omics data

The host omics data consisted of a TCGA data set which was extracted and processed by a prior study 18. This data set consists 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 was obtained using the UCSC Xena data browser on March 8, 2016. The expression values of these genes consists of a pre-processed and batch-corrected gene abundance 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 used, as evaluated using Median Absolute Deviation (MAD). Finally, the expression values are also min – max scaled.

### Clinical data

The clinical data was accessed using the Snaptron web server. We match the clinical data with the corresponding patient samples in order to obtain details for the tumor and stage endpoints. To determine whether a sample is tumor or normal, the sample type code [[2]](#footnote-2) is used, where codes in the range 01 – 09 are tumors and those in the range 10 – 19 are normal samples [[3]](#footnote-3). The stage clinical data is used to determine the tumor stage of each sample. Normal samples are considered as Stage 0.

### Overlapped data

In order to investigate the effects of microbial data on cancer diagnostics, we create an overlapped set samples for which there is both host- and microbial omics data in the above-described data sets.

In the first step, the TCGA gene expression data is joined with the clinical data. To do this, the “portion\_id” field of each row in the clinical data set, which contains the code for the project, the tissue source site (TSS), participant ID, sample type, vial, and portion id is used. This field is stripped of the portion ID and the vial, the duplicates are dropped (to remove samples with the same ID but different measuring technologies), and each row is then joined with the samples in the GE data set, which contains the same attributes up to and including the sample type.

To join the data set with GE + clinical data to the microbial dataset, the sample barcode for the microbial samples is stripped of the all samples with a vial type of “B”, and then the vial type attribute is removed altogether. This data set is then joined with the GE + clinical data set.

The table below displays the class balance of tumor versus normal samples for each modality and the integrated set.

Table

Description automatically generated

The next table displays the balance for the stage samples. Normal samples are classified as stage 0.

Table

Description automatically generated

Due to the lack of samples for READ, we use data for the cancers COAD, ESCA, HNSC and STAD.

## Methods

### Exploration of holistic view

Feature exploration was performed using PCA and t-SNE. PCA was performed using the sklearn PCA function with 2 components. t-SNE was performed using the sklearn t-SNE function using the default settings, including 1000 maximum iterations. A static random seed was used for all experiments.

Feature reduction was performed using the chi2 test Anthony regression

### Predictive performance

For the stage endpoint, the prediction of the different stages 0-4 it is modeled as a regression, rather than classification problem.

### Feature selection

# Results

## Data

To investigate the effects of the host-omics approach, we used preprocessed samples from the Cancer Genome Atlas (TCGA). It turns out that gene expression and DNA methylation data works well with multiple cancers 17. Thus, in order to examine the benefits of a holistic view on omics data for cancer diagnostics, gene expression features from TCGA are used. This set contains 5000 features for each sample.

To investigate the effects of the microbial omics approach, we used the Cancer Microbiome Atlas (TCMA). This is a microbial database which is based on data that is mined and processed from reads contained in TCGA, which attempts to identify microbial species in tissue and blood samples while dealing with contaminants 2. TCMA contains microbial data available from multiple cancers allowing for cross cancer analyses using the same source with common methodologies.

Finally, for the holo-omic approach, we overlapped samples for which there is both host omics and microbial omics data available.

## There is little qualitative difference in performance with holo-omic approach

We performed a preliminary exploration of the available data sets in order to examine whether a holistic view could have benefits for cancer diagnostics. To this end, we used numerous dimensionality reduction techniques in order to examine the separation between classes for multiple diagnostic endpoints as an indicator of the possible predictive value of integrating omics and microbial data. The diagnostic endpoints were tumor versus normal prediction, and tumor stage prediction. We investigated the class separation using the PCA and t-SNE dimensionality reduction techniques.

For each dimensionality reduction technique and for all 4 cancers, we performed dimensionality reduction on the gene expression data set separately, the genus taxonomical abundance data, and then on the concatenation of these datasets. Additionally, this was done for increasing amounts of selected features, and when using all features. Feature selection was performed using the chi-square test and using all data.

### Tumor PCA does not show additional class separation

The PCA for the tumor classification endpoint shows that there is no meaningful additional separation between the classes when integrating genus taxonomic data with gene expression versus using only gene expression data (Figure 1). Additionally, the genus layer does not seem to provide much by itself separating power. This is possibly because the microbial information is not directly related to the phenotype and thus does not provide enough explaining power by itself.

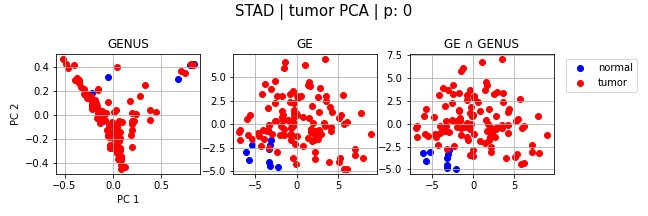


Figure : tumor PCA of STAD for all modalities with no features selected

This result is consistent across cancer types (i.e. for COAD, ESCA and HNSC). The results for these additional cancer types can be found in the appendix.

### Results are consistent across feature selection amounts

For PCA, performing feature selection leads to increased separation.

For tumor versus normal prediction, features (0,5,10) exhibit about the same behavior. GE tends to for show much better separation than genus, and GE+Genus is almost exactly the same as just GE, possibly because GE has much more features, and when features are selected, the selected features are probably GE features.

### Using different dimensionality reduction techniques does not offer additional separation

To determine whether the lack of separation of the holo-omic approach was due to the dimensionality reduction method chosen, we repeated the above experiment using t-SNE.

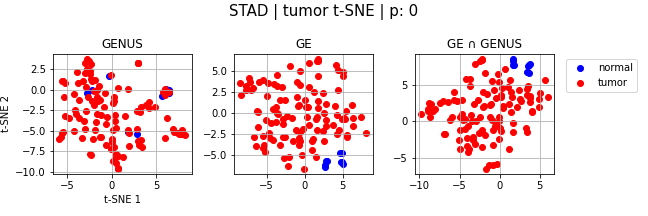


Figure : tumor t-SNE of STAD for all modalities with no features selected.

With t-SNE, using the overlapping gene expression with genus data again provides a nearly identical amount of separation as when only using gene expression data (Figure 2). Similar to PCA, using only the genus modality provides the worst separation between classes .

### There is unclear class separation for the stage class

In order to determine whether the above results were due to the classification endpoint chosen, we repeated the results using a different endpoint, the tumor stage. For the stage class endpoint, there is not much separation to be seen across any of the modalities (Figure 3). This is likely because of the class imbalance.

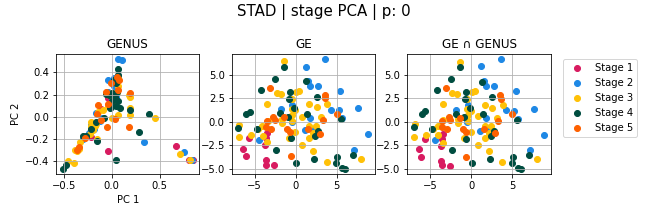


Figure : stage PCA of STAD for all modalities with no features selected

Again, there is not much difference between the GE and the overlapped set. These results are similar when using t-SNE.

### Exploring the modalities balance of integrated features

One problem with the above is that the integration occurs via simple concatenation, and if the feature selection process simply selects features from only one data set, then the integration performance will approximate that of the individual data set.

## There is little quantitative performance improvement with the holo-omic approach

To investigate the possible benefits of a holistic view of omics integration, we used the different omics layers and microbial data in a predictive model for numerous cancer diagnostic endpoints. The performance of the predictive model in different settings was used as a measure of the usefulness of each modality.

We use random sampling to randomly split the data set with a stratified split into 80% training and 20% testing n times. In each iteration, a feature selection method is first performed on the training set. These features are then used to train a model on this same training set. The performance of the model is then tested on the testing set for the available endpoints. This is repeated for each cancer, layer and feature amount. The table below displays the combinatorial options of the pipeline.

|  |  |
| --- | --- |
| Type | Values |
| Cancers | COAD, ESCA, HNSC and STAD |
| Endpoint | Tumor vs Normal, Stage |
| Layers | GE, Genus, GE ∩ Genus |
| Feature Selection Method | Chi2 and Linreg |
| **Prediction model**  Tumor vs Normal  Stage | SVM  Linear regression |
| Feature amounts | 6, 10, 26, 50, 100, 200, all |

The prediction models are then evaluated using the precision, recall and f1-score. This pipeline thus generates n different sets of values for these evaluation metrics. For the evaluation, we consider the average of the metric across each class value. For each combination of the value types above, the average of across the n random sampling iterations are then plotted along with their standard deviation across these iterations.

Again, this was done for stage prediction and tumor versus normal prediction. The predictive performance of the GE model separately, and the microbial data model were used as a baseline. The integration of both of these data sets were then compared to this baseline.

### Holo-omic approach does not lead to improved tumor prediction performance

The preliminary results indicate that integrating gene expression with microbial taxonomic data does not lead to a significant improvement in prediction performance over using gene expression data alone (Figure 4).

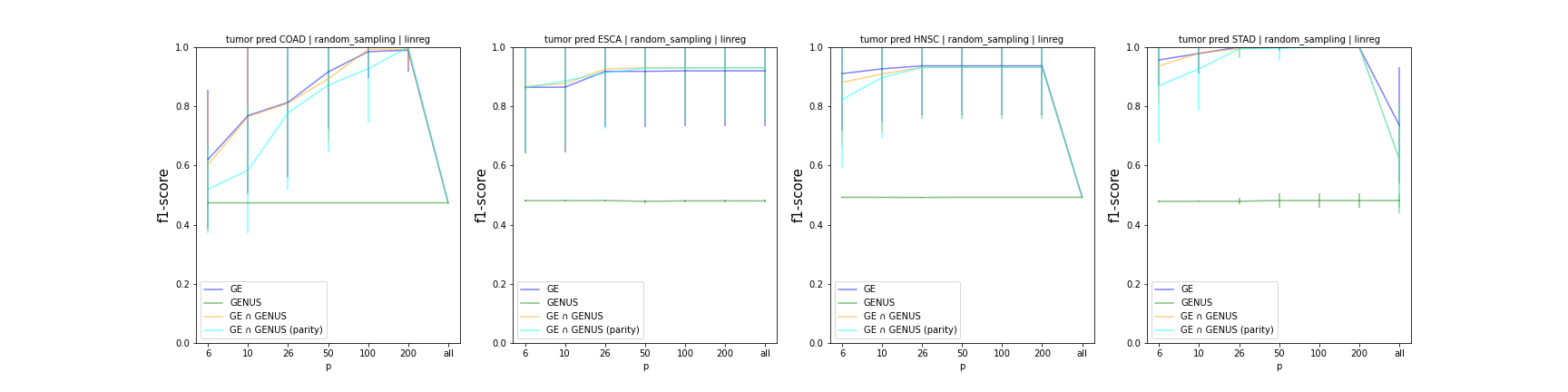


Figure : f1-score of tumor vs normal prediction for each cancer

This result is consistent for tumor versus normal prediction, but also stage prediction Figure 5.

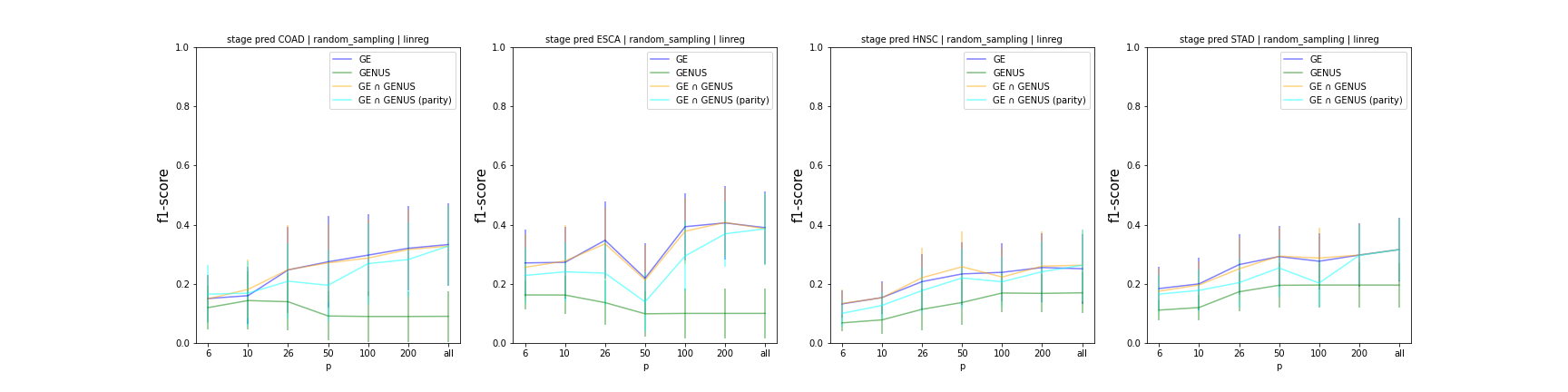


Figure : f1-score of stage prediction for each cancer

These results also seem to be consistent across the different amount of selected features.

### Genus layer provides worse performance

The genus layer seems to provide the worst performance. This is possibly because the taxonomic genus data is mined from the tumor samples and might not be directly related to the aetiology of the disease. Thus, it might not contain enough predictive power by itself. It is worth noting that this does not necessarily mean that microbial data might not be enough to predict these endpoints, but rather that the specific microbial data collected might need additional non-microbial features to be powerful.

Interestingly, this difference is smaller for stage prediction. This is possibly because of the extreme imbalance of classes in the data set. The more classes there are, the harder it is to maintain a balance. Thus, for stage prediction, all the models achieve relatively low performance.

This is relatively consistent with previous results of prediction models using TCGA microbial data 5.

### Are results independent of prediction model?

Attempt different prediction models and hyper parameter tuning.

### Feature selection is dominated by one modality

To investigate the distribution of the types of features selected for each feature selection method and feature amounts, the proportion for which each layer is represented in the final feature sets is investigated by saving the selected features for each combinatorial state of the pipeline in each iteration. The proportion is then plotted for all the different cancers for each feature amount.

It seems that when performing feature selection on the integrated data set, most of the features selected originate from the gene expression set (Figure 6).

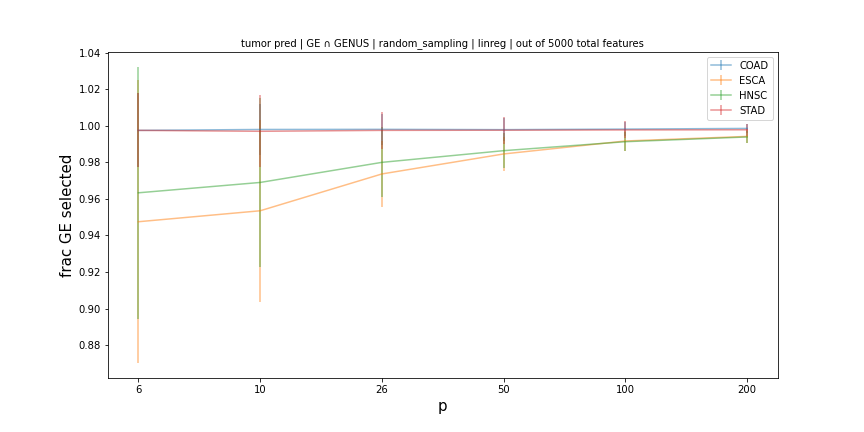


Figure : fraction of GE features selected for each feature amount and cancer during tumor versus normal prediction

Furthermore, out of the 221 total GENUS features, less than 1% is selected in the feature selection process (Figure 7).

Chart, line chart

Description automatically generated

Figure : absolute amount of GENUS features selected for each feature amount and cancer during tumor versus normal prediction

One possible reason for the domination of gene expression features is that the original data set contains much more gene expression features than taxonomic genus features (more than 20 times as much). Thus, there are simply more gene expression features to select from. It is clear that the microbial features are still useful, given that they are still reasonably well represented when selecting the top 5 or 10 features.

#### GE modality dominance is smaller in stage prediction

The dominance of the gene expression modality seems to be smaller for stage prediction, possibly because there is a wider variety of endpoint values and thus relationships between input and output to consider during the feature selection process.

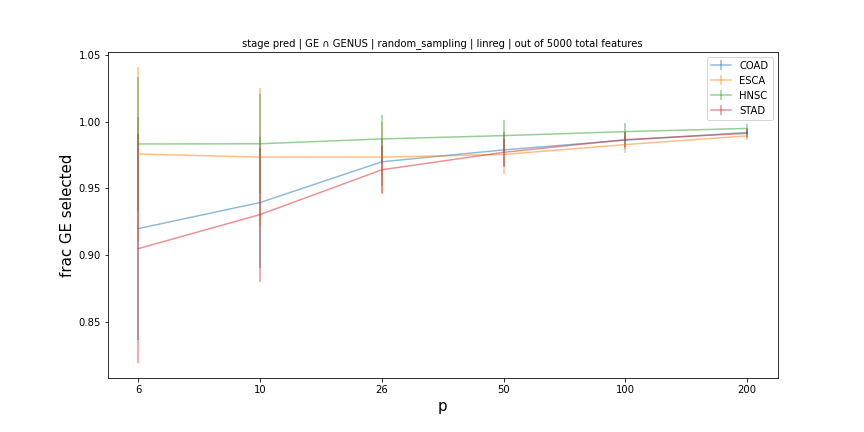


Figure : fraction of GE features selected for each feature amount and cancer during stage prediction

Chart, line chart

Description automatically generated

Figure : absolute amount of GENUS features selected for each feature amount and cancer during stage prediction

One possible way to deal with this is to correct for the amount of base features in each modality. Additionally, it is possible to use feature selection on the original modalities separately prior to concatenating.

#### Feature selection robustness

Devise experiment to examine feature selection robustness across iterations.

[Display bar plot showing the amount of features selected for different percentages of something iterations e.g. 10 feature selected 40-50% of the time]

### Results consistent for additional prediction targets

results of survival prediction

[Display prediction performance of survival]

### Enforcing class balance does not improve performance

results when over sampling

[Display prediction performance when over sampling]

### Enforcing modality parity during feature selection does not improve performance

To determine whether the domination of one modality in the features selected has an effect on the added performance of the holo-omic approach, we repeated the above experiments while enforcing parity in the amount of features selected of each modality. To do this, we performed feature selection prior to integrating the modalities to ensure that the same amount of features are selected of each modality.

As can be seen, this leads to similar or worse performance than only using GE or the non-enforced-parity feature selection approach with the overlapped layer (Figure 4).

[Display performance of integrated modality versus individual modalities when enforcing modality parity]

# Conclusion

1. Machiraju, G., Amar, D. & Ashley, E. Multi-omics factorization illustrates the added value of deep learning approaches. 7.

2. 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).

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

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

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

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

7. 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).

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

9. Fang, H., Du, Y., Pan, S., Zhong, M. & Tang, J. Patients with Parkinson’s disease predict a lower incidence of colorectal cancer. *BMC Geriatr.* **21**, 564 (2021).

10. 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.

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

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

13. 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.

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

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

16. Proctor, L. M. *et al.* The Integrative Human Microbiome Project. *Nature* **569**, 641–648 (2019).

17. Duan, R. *et al.* Evaluation and comparison of multi-omics data integration methods for cancer subtyping. *PLOS Comput. Biol.* **17**, e1009224 (2021).

18. 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).

1. <https://tcma.pratt.duke.edu/> [↑](#footnote-ref-1)
2. https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes [↑](#footnote-ref-2)
3. https://docs.gdc.cancer.gov/Encyclopedia/pages/images/TCGA-TCGAbarcode-080518-1750-4378.pdf [↑](#footnote-ref-3)