The Tumor Microbiome: Review and Research Proposal

Aashna Soni

Capstone: Microbiology

Dr. Gauravjit Singh

10 February 2025

Table of Contents

Introduction and Learning Objectives

Cancer is a set of diseases in which cells accumulate mutations to cell-cycle-regulating genes and divide uncontrollably, forming tumor masses and metastasizing to other regions of the body. The tumor microbiome consists of bacteria, viruses, and fungi inhabiting the tumor microenvironment (TME), which refers to the cells, compounds, blood vessels, and fluids surrounding tumor cells and modulating their growth [1]. The tumor microbiome is an emerging frontier in cancer research due to its potential to illuminate microbial targets for new personalized treatments and microbial biomarkers distinguishing cancer types.

This paper will accomplish the following learning objectives:

1) Holistically understand the tumor microbiome: methods of TME colonization, cancer type specificity of the tumor microbiome, microbial methods influencing cancer progression, and progress in the field of personalized treatment targetting the tumor microbiome.
2) Explore the difficulties and limitations in analyzing tumor microbiome data.
3) Provide an overview of machine learning (ML) and multi-modal modeling approaches for analyzing microbiome data.
4) Propose a method to address limitations in the field of analyzing tumor-resident microbiome data: lack of understanding of tumor-microbe dynamics, unexplored integration of tumor-cell and decontaminated tumor-localized-microbe data for cancer type diagnosis.
5) Summarize prospective results and implications.

Literature Review

Microbial Colonization of the TME

Many factors make the TME an optimal location for microbial colonization; the most notable are the TME's immunosuppressive and hypoxic conditions. Cancer cells are able to proliferate due to their ability to evade immune recognition. Thus, in the TME, microbes can divide without being as strongly targeted by the immune system compared to other regions of the body. Furthermore, most solid tumors create a hypoxic (low oxygen content) environment, which enables anaerobic bacteria like *Fusobacterium nucleatum* and *Peptoniphilus Porphyromonas* to survive and thrive [2].

Microbes colonize tumors through various methods, including mucosal barrier invasion, adjacent tissue invasion, and hematogenic invasion [3].

In mucosal barrier invasion, microbes enter tumors through damaged mucosa, the inner lining of organs [3]. This method of microbial invasion and colonization has been well-studied in colorectal cancer: Tjalsma et al. developed a model to describe the evolution of microbiome composition in the colon. In this model, “driver” bacteria colonize the intestine through the intestinal mucosa, initiating tumor development. In the process, the TME undergoes chemical compositional changes, making it attractive to other opportunistic pathogens that then invade the space. Through the accumulation of diverse microbes, CRC progression evolves over time [4]. Mucosal barrier invasion has also been validated in other tumors of mucosal organs like the esophagus, lung, and colon [3].

The next key method of microbial tumor invasion is adjacent tissue invasion, in which normal tissue-inhabiting microbes invade an adjacent cancerous tissue [3]. In these patients, great similarity can be observed between the microbial communities of these normal and cancerous tissues, making it difficult to distinguish cancerous from normal tissue based on microbial compositions. For example, Nejman et al. found that *Proteobacteria* dominate the pancreatic cancer microbiome, making it similar in microbial composition to the normal duodenum. Identifying such similarities can help us trace the path of migration of microbes to the tumor site (in this case, for pancreatic cancer, these microbiome similarities suggest adjacent tissue microbe migration from the duodenum to the pancreatic duct) [5].

The final key method of microbial tumor invasion is hematogenic invasion, in which bacteria from normal human tissues in the oral cavity and intestines enter tumor tissue through destroyed or weakened blood vessels [3]. Factors contributing to this microbial invasion include slow blood flow, blood leakage, and angiogenesis, the formation of new blood vessels [2]. Through blood vessels, microbes can reach distant regions of the body, as opposed to adjacent tissue invasion, which is a local method of invasion [3].

Cancer Type Specificity of Tumor Microbiome

The tumor microbiome is highly specific: both in differentiating cancer types and stages of cancer. For example, *Helicobacter pylori* (HP) is a significant contributor to gastric cancer (GC) and gastric mucosa-associated lymphoid tissue (MALT) lymphoma [2]. Nejman et al. found that bacteria belonging to the phyla *Firmicutes* and *Bacteroidetes* have the greatest presence in colorectal tumors [5]. Breast tumors are notable for having the greatest microbial species diversity compared to all other tumor types, with the most prevalent phyla being *Proteobacteria* and *Firmicutes* [2, 5]. As breast cancer is a very high mortality cancer, understanding its microbiome holds significant potential for advancing breast cancer treatments.

The microbiome composition can distinguish benign from malignant tumors: Heiken et al. found that malignant breast tissue is marked by a greater relative abundance of genera like *Fusobacterium, Atopobium, Gluconacetobacter, Hydrogenophaga*, *and Lactobacillus* compared

to benign tissue, and this difference arises due to metabolic changes induced by microbes, such as increased cysteine and methionine metabolism and fatty acid biosynthesis [7]. Furthermore, the *P. acnes* species is abundant in prostate tumors, and the *Shewanella* genus is enriched in malignant prostate tissues [6].

However, characterizing cancer types based on microbial community composition remains challenging due to the low biomass of microbes in tumor tissues and the difficulty in extracting and decontaminating these samples from tumor biopsies [8].

Microbial Factors Influencing Cancer Progression

The mechanisms by which the tumor microbiome contributes to cancer is a key current area of investigation. Known mechanisms include inducing genomic instability and mutations, modifying host cell epigenomes, inducing the immune response through toll-like receptors (TLRs), and initiating metastasis [3].

Inducing genomic instability is a prevalent mechanism by which microbes promote cancer onset. For example, oncoviruses integrate their genetic material into the host cell's chromosomes. This leads to mutations to cell-cycle regulating genes and triggers the production of viral oncoproteins, both of which deregulate cell-cycle-related pathways [3]. Examples of oncoviruses include human papillomavirus (HPV) in cervical and head-and-neck cancers and the hepatitis B virus (HBV) in liver cancer [9, 10, 11]. Genomic instability can also be induced by carcinogenic bacteria that secrete toxins or metabolites causing DNA damage; an example of this is pks+ *Escherichia coli* in colon cancer. The secretion of DNA-damaging toxins activates the host's DNA damage response, which then triggers the immune response, creating a pro-inflammatory microenvironment that can facilitate metastasis through mechanisms such as inflammatory factors like vascular endothelial growth factor A (VEGFA) promoting angiogenesis [3].

In addition to genomic instability, epigenetic modifications induced by microbes contribute to cancer [3]. Microbes can modulate the host epigenome directly and indirectly: by inducing the expression of long non-coding RNAs (lncRNAs), modifying histones, and impairing the functionality of histone acetylases and histone deacetylases. Liu et al. found, through bioinformatic analysis, that when *H. pylori* infects host cells, the guanine nucleotide-binding protein subunit beta-4 (GNB4) becomes demethylated, contributing to gastric carcinogenesis [12]. However, the mechanism of *H. pylori* inducing DNA methylation remains to be understood.

Furthermore, tumor microbes often interact with Toll-like receptors (TLRs), sensors that compromise a key component of the innate immune system, in the TME to induce immune responses [13, 14, 15]. Through TLRs, microbes can exert both pro-cancer and anti-cancer effects through the release of pathogen-associated molecular patterns (PAMPs), compounds

released directly by microbes, as well as damage-associated molecular patterns (DAMPs), compounds released by infected human cells in the TME [16, 17]. Lorenzo et al. discovered an example of microbial pro-cancer effects: they found that in colon cancer, microbial TLR2 activation leads to the production of reactive oxygen species (ROS) and drives cholesterol biosynthesis, which leads to cancer cell proliferation and chemoresistance [18]. On the other hand, microbes can induce anti-cancer effects: Gonzalez et al. found that flagellin, one of the proteins constituting bacterial flagella, binds TLR5 on macrophages, triggering the activation of CD8+ T-cells, which drive cancer cell destruction [19, 20]. These examples show that the immune system is modulated in a complex manner to enable tumor development.

Finally, tumor microbes can promote cancer metastasis: the invasion of tumors to other regions of the body. For example, many studies have shown that microbial activity correlates with the epithelial-mesenchymal transition (EMT) through mechanisms such as activating EMT-related pathways (NF-κB, Wnt/β-catenin, MAPK, and PI3K); increasing the expression of EMT-associated transcription factors (like Snail, Slug, and ZEB1); and improving the adhesion of tumor cells to endothelial cells in the bloodstream [21, 22, 23, 24]. The EMT describes the process by which epithelial cells develop mesenchymal features, such as reduced adhesion to other cells, resistance to apoptosis, or programmed cell death, and modified production of extracellular matrix compounds. These mesenchymal features enhance the ability of tumor cells to migrate to other regions of the body [24].

Personalized Treatments Targeting the Tumor Microbiome

Given the tumor microbiome's complex mechanisms of influencing cancer progression, a key area of current investigation is developing cancer treatments targeting the tumor microbiome. Different treatment modalities that have been explored include antibiotics, synthetically engineered bacteriophages, and synthetically engineered bacteria. While antibiotics generally affect a wide range of microbial species (including those in the gut), genetically modified bacteriophage or bacteria treatment is highly microbe-specific.

Many retrospective studies have shown that antibiotics can improve cancer patients' prognoses [3]. Using antibiotic adjuvants to target the tumor microbiome is a promising direction for a few reasons: existing FDA-approved antibiotics can be repurposed, and the tumor microbiome impairs the efficacy of traditional treatments like chemotherapy, warranting a need to weaken these resistance-driving microbes' activity. Geller et al. found that Gamma-Proteobacteria in pancreatic ductal adenocarcinoma (PDAC) produces the enzyme cytidine deaminase (CDD), which breaks down the chemotherapy compound into its ineffective form, contributing to reduced drug efficacy and the onset of drug resistance [25]. By weakening microbes that promote tumor development, chemotherapy resistance can be reduced, improving patients; overall chance of survival. An example of successful antibiotic treatment was found by Mohindroo et al:

patients treated with the chemotherapy drugs gemcitabine and 5-fluorouracil had improved progression-free survival after antibiotic treatment adjuvants [26].

However, antibiotic therapy comes with challenges, most pressing of which is that it oftentimes damages the gut microbiome, which can have a counterproductive effect–weakening the gut microbiome can reduce immune efficacy, reducing the ability of the host immune cell to target cancer cells [3]. Taking antibiotics within a short time window of immune checkpoint inhibitor (ICI) treatment has been shown to impair ICI effectiveness [27]. Additional investigation is needed into methods of using antibiotics to selectively target tumor microbes while not inducing gut microbiome dysbiosis.

Synthetic engineering techniques, which involve modifying a virus's or bacteria's DNA, RNA, or metabolic pathways, offer various advantages over antibiotics: they are highly specific in targeting microbes at species resolution (as opposed to broad-spectrum antibiotics, which often cannot distinguish between bacteria of the same genus), have a low toxicity level (leave human cells undamaged), are better at penetrating bacterial biofilms, and can deliver high drug payloads because bacteria and viruses multiply rapidly in the immunosuppressive TME [28].

Synthetic engineering can be used in one of two ways: to target the tumor microbiome or to target cancer cells themselves [28]. Bacteria have been synthetically engineered to target cancer cells: for example, researchers at the University of Massachusetts, Amherst developed BacID, a genetically engineered *Salmonella* strain that becomes activated upon salicylic acid treatment, destroying tumor cells [29].

In addition to targeting tumor cells, synthetic engineering methods have been employed to modify bacteriophages to target tumor-resident microbial species. For example, Zheng et al. genetically modified bacteriophages to both target *F. nucleatum*, a common bacterial species present in tumors, as well as bind murine colon tumor cells to optimize the delivery of irinotecan chemotherapy treatment [28]. Further investigation is needed to understand how best to leverage methods like genetically modified bacteriophage treatment to target tumor microbes, and if and how different treatments can be combined to enhance efficacy in targeting a specific set of microbes while leaving the gut microbiome intact.

Specific Area of Investigation: Data Analysis

The advent of next-generation sequencing (NGS) methods and the increase in computational power have created immense opportunities to analyze vast biomedical datasets, from sequencing to clinical to imagining, for diagnostic, treatment development, and prognostic value. The following sections will attempt to give an overview of this field: challenges, machine learning approaches, specific studies, and next steps.

Challenges in Sample Data Collection

The tumor microbiome remains underexplored compared to other human microbiomes primarily due to sample extraction and decontamination challenges [8].

Tumor biopsies, which are necessary in order to collect samples of tumor-resident microbes, are highly invasive procedures. It is much easier to non-invasively characterize gut, oral, and bloodborne microbial communities by collecting fecal samples, mucosal swabs, and blood samples, respectively [8].

Additionally, tumor samples generally contain a low microbial biomass, making external contamination a significant issue in data analysis, for even small amounts of contamination (large relative to the total microbial biomass) can introduce artifacts [30]. Contamination can occur in two main ways: acquisition of genetic material from the local environment during sample collection or introduction of contaminants during computational analysis. To prevent external contamination, aseptic techniques, in addition to sterilized DNA extraction kits and laboratory reagents (that do not contain foreign genetic material), must be used. Regarding the introduction of data preprocessing contaminants, it is important that for each diseased sample collected, a matched-normal sample be collected to serve as the control so that contaminants present in both groups can be identified and disregarded in downstream data analysis [8].

Machine Learning and the Tumor Microbiome

Supervised machine learning (ML) methods represent a set of algorithms that “learn” correlations between input data and output value by minimizing a cost function, which quantifies the discrepancy between ground truth and predicted values.

ML models have been developed and applied to detect cancer from microbiome data. Most ML applications have focused on understanding the relationship between the gut microbiome and colorectal cancer [8]. For example, Murovec et al. applied ML and linear discriminant analysis (LDA) to publicly available data of stool samples of 2,951 patients (healthy, colorectal cancer (CRC), and colorectal adenocarcinoma (CRA)). They found no statistically significant difference in microbiome diversity between these three groups but successfully identified microbial taxonomic and functional prediagnostic indicators for CRC and CRA using their Random Forest ML model. The ML model achieved a greater than 95 percent accuracy in differentiating the three groups of patients [31].

While tumor, oral, gut, and bloodborne microbiomes all differ significantly in composition between healthy and cancerous patients and thus have the potential to serve as cancer correlative indicators, understanding the tumor microbiome is of particularly great interest because it is localized to the tumor region. Analyzing tumor-resident microbiome data can illuminate

tumor-microbe dynamics, such as signaling pathway crosstalk and release of microbial metabolites that bind to tumor cell receptors. By understanding these complex cellular and molecular dynamics, we can identify pathways and proteins that can be targeted therapeutically. On the other hand, sampling the microbiome at locations in the human body distant from the tumor's primary site can help us understand how cancer development harms normal microbiota health but is not as informative for understanding the precise molecular dynamics between tumor and microbiome that contribute to tumor formation and development [8].

The Cancer Microbiome Atlas (TCMA)

Significant progress has been made in curating decontaminated microbiome databases for internal organs. One such notable project was conducted by Dohlman et al. of Duke University. They applied an unbiased statistical model to genomic sequencing data from The Cancer Genome Atlas (TCGA), an extensive repository of multi-omic data related to cancers, to produce a decontaminated database of tumor microbiome data for oropharyngeal, esophageal, gastrointestinal, and colorectal tissues [32].

The principle of their decontamination algorithm was to first compare microbiome signatures between cancer tissues of various organs and the blood to rule out species appearing indiscriminately. After this round, they compared microbiome composition between identical samples processed at different DNA sequencing company sites to identify computationally induced artifacts. Through their decontamination pipeline, they found that species equiprevalent across tissue and blood samples are predominantly contaminants that were introduced during the sample collection and data processing stages. They filtered out this contaminant data to produce a pure database containing tissue-resident microbial profiles for 3,689 unique samples from 1,772 patients collected from 21 different anatomic sites [32].

There is great potential to mine this publicly available database to extract insights into how the tumor microbiome contributes to cancer. For one, the creators of TCMA remark on its potential to illuminate pan-cancer biomarkers (cancer-related biomarkers irrespective of cancer type). Cancer-microbe relationships have largely been studied in a cancer-type-specific manner, as microbes can have different effects in different cancers: for example, *H. pylori* contributes to the onset of gastric cancer but has been shown in studies to offer a protective effect in esophageal adenocarcinoma [32]. Nevertheless, there are certain microbe-induced processes, including chronic inflammation and altering of host cell metabolism, that are known to span cancer types, and mining TCMA data can help us better understand these [8]. On the other hand, TCMA's inclusion of microbial data for four distinct cancer types also creates an opportunity to identify host-microbe dynamics that distinguish these cancers and could serve as cancer-type-specific biomarkers.

Freitas et al. developed a framework to analyze tissue-resident microbiome data from TCMA for the purpose of cancer-type differentiation. Specifically, they trained a multi-class Random Forest algorithm, a supervised ML algorithm whose output value is the aggregation of computations across easy-to-interpret decision trees. Their model achieved strong performance in predicting the following cancer classes: head and neck, stomach, and colon. However, they noted low model performance in distinguishing esophageal and rectum cancers, suggesting that cancers that are anatomically adjacent have greater similarity in microbial composition, making them difficult to identify based on microbial data alone. Thus, a key limitation of this study was that it only included microbial relative abundance data and did not integrate data from the host tumor cells [33]. A future direction of study is how microbial and tumor-cell data can be aggregated to improve the identification of cancer types, especially those that are anatomically adjacent.

Multi-Modal Modeling

An emerging area of investigation in cancer research is multimodal deep-learning modeling, i.e. integrating datasets of various structures into a single computational model. The idea behind multimodal modeling is that the complex principles governing health and disease can only be understood by bringing together different types of data (clinical, genomic, transcriptomic, epigenetic, imaging, etc.) and modeling nonlinear relationships. Multimodal approaches can provide greater accuracy and insight than unimodal approaches [34].

To study a cancer patient's health state, different data types provide different qualities of information: tumor genomic data can illuminate cancer driver genes, and whole-slide microscopic images of a biopsy can provide information about the tumor's morphology and the TME [34]. These two data samples are "complementary" because they provide information on different aspects of a single patient's tumor. Data types can also be "redundant": for example, transcriptomic and proteomic data overlap as a significant number of mRNAs (but not all) are translated into functional proteins. Data types can also be classified as "cooperative" if together they increase the complexity of the biological phenomenon under study [34].

As illustrated through the study of Freitas et al, there is a need to fuse structured and unstructured data categories via machine learning to detect trends and learn complex nonlinear correlations between input data and output value [34].

An example of such a multimodal model is MDL4Microbiome, developed by Jae Lee et al. This model combines three classes of microbial data (genome-level abundance, metabolic functional abundance, and taxonomic profiles) to predict disease status for inflammatory bowel disease (IBD), type 2 diabetes (T2D), liver cirrhosis (LC), and colorectal cancer (CRC). These three microbial data categories were chosen as they together provide essential information on the gut microbiome. Four public datasets were used to train this model: the first contained healthy individuals and patients with IBD, the second contained healthy individuals and patients with

T2D, the third contained healthy individuals and patients with LC, and the fourth contained healthy individuals and patients with CRC. The model architecture was as follows: each datatype was fed into a separate supervised deep neural network to produce a compact feature embedding. The resulting embeddings were fed into a final classifier that outputted the patient's disease status. The MDL4Microbiome classifier achieved accuracies of 98 percent, 76 percent, 84 percent, and 97 percent in distinguishing IBD, T2D, LC, and CRC from corresponding healthy patients, respectively [35]. Such models show the significant potential of fusing different data modalities to create diagnostic tools.

Goals

This research proposal has two main goals:

1. To build a multimodal supervised machine-learning model for cancer subtype prediction that can serve as a potential diagnostic assistant for oncologists.
2. To identify microbial community biomarkers that distinguish four cancer types using hierarchical clustering, an unsupervised machine-learning algorithm.

Methodology

A flowchart for the computational workflow is shown in **Figure 1**, and a link to data samples is included in the "Data Samples" section.

First, microbial relative abundance data will be downloaded from The Cancer Microbiome Atlas for oropharyngeal, esophageal, gastrointestinal, and colorectal tissues [32]. The data is a CSV file: columns contain IDs of different patient tissue samples, rows are IDs for different microbial taxonomic units, and cell values are the bacterial relative abundance in each sample. The patient tissue samples' metadata and the species names corresponding to the microbial IDs will be obtained through The Cancer Genome Atlas's GDC API.

This data will be preprocessed by normalizing data values to Reads Per Kilobase per Million (RPKM) mapped reads and removing rows with greater than 80% of values equal to 0.0 or empty.

From there, hierarchical clustering will be implemented on the microbial relative abundance data to identify groups of bacterial species with similar relative abundance levels for different cancer types. Hierarchical clustering is an unsupervised machine-learning method (i.e. no output value) optimal to employ when the number of clusters is not predetermined. The grouped microbes will represent microbial community biomarkers for distinct cancer types.

In addition to this tissue-resident microbial data analysis, tumor-cell data will be downloaded from The Cancer Genome Atlas (TCGA) and fused with the tumor-resident microbial data to build a machine-learning cancer-type prediction model. The tumor-cell data categories that will be downloaded from TCGA are DNA methylation data, copy number variation (CNV) data, clinical data, and a biopsy image of the tumor sample under a microscope. The DNA methylation data is a CSV file: the rows are CpG sites located in different genes, and the values are the methylation beta values of these sites (beta values range from 0 to 1, where < 0.5 is considered hypomethylated and > 0.5 is considered hypermethylated). The CNV data is a TSV file format: the rows are gene names, and for each gene, there is a copy number value that determines whether the gene is present in normal diploid number (copy number = 2), is possibly deleted (copy number < 2) or is possibly amplified or duplicated (copy number > 2). For the clinical data type, the rows are tissue sample identifiers and the columns are age, race, and gender. Separately, the cancer type label for each tumor tissue sample will be downloaded from TCGA by matching tumor IDs to the GDC database via API calls, and the cancer type names will be encoded as numerals (1, 2, 3, 4). Finally, the biopsy data is a microscopic slide image of the tumor. Every dataset will be split in an 80-10-10 ratio into a training, test, and cross-validation dataset: the model will be trained on the training dataset, hyperparameters will be tuned on the cross-validation dataset, and the model's generalizability to predicting cancer type on unknown data will be tested on the test dataset.

These training datasets (microbial relative abundance, tumor DNA methylation, tumor CNVs, tumor clinical phenotype, tumor microscopic slide image) will be inputted into separate computational frameworks for feature extraction. The microbial data will be fed through a feed-forward neural network (NN) to extract a feature vector representing the microbial composition of different samples. The DNA methylation and CNV data will be inputted into a single feed-forward NN to extract a feature vector summarizing the tumor's molecular biology. The clinical data will be converted into a vector representing total clinical information: [age, race, gender]. Finally, the microscopic image will be processed into a feature embedding using a pre-trained, open-source convolutional neural network (CNN) called Residual Network (ResNet). ResNet was developed by He et al as a generalizable, high-depth and high-efficiency tool for training computer vision models and will be used as a feature extractor for the biopsy images [36].

These feature embeddings will then be inputs to a final feedforward NN, which will output patient cancer type (oropharyngeal, esophageal, gastrointestinal, or colorectal). The feedforward NN will contain three categories of layers: an input (concatenation) layer, deep layers for learning nonlinear correlation between input and output, and an output layer for cancer type.

The machine learning models and hierarchical clustering algorithms will be built and tested in Python using the packages ResNet, TensorFlow, sci-kit learn, numpy, and pandas.

Results

Hierarchical clustering will produce a heatmap: the x-axis will be microbe species names and the y-axis will be cancer type. A heatmap is a visual representation of the clustering results, with colors ranging from dark red to dark green to illuminate overabundance and underabundance of microbial species in different tumor types. Using this heatmap, we will be able to identify groups of microbes that have similar relative abundance levels in a given cancer type and understand how their abundance levels differ between cancer types. For example, if a set of microbes has high relative abundance levels in colorectal cancer and low relative abundance in oropharyngeal, esophageal, and gastrointestinal cancers, that community of microbes could serve as a prediagnostic indicator for colorectal cancer.

The multimodal machine learning model's performance will be evaluated using standard metrics: accuracy, precision, recall, F1 score, and Area Under the Receiver Operating Characteristic (ROC) Curve. Furthermore, recursive feature elimination will be used to identify which features (microbial and tumor) make the most significant contribution to determining cancer type. These features will have the greatest potential to serve as cancer-type-differentiation biomarkers.

Discussion of Implications

This study's results will open vast doors in personalized cancer research.

Analyzing tumor microbiome data will elucidate intricate cellular and molecular dynamics that take place between tumor cells and tumor-resident microbes on a cancer-type-specific basis. By identifying the most impactful features using recursive feature elimination, we can understand which components of the TME (microbial or tumor cell) make the most significant contribution to determining cancer type.

We can also identify microbial communities that distinguish cancer types through hierarchical clustering. Clusters of microbes that are clearly distinct to a specific cancer type could be further investigated as potential treatment targets.

Furthermore, we can investigate whether the tumor-resident microbe biomarkers identified overlap with gut microbiome species. This is an important consideration in developing treatments targeting the tumor microbiome because it is imperative that the gut microbiome, which is key to the functioning of the immune system, not be damaged. Microbial biomarkers specific to different cancer types can become targets of future personalized treatments, such as genetically modified bacteriophages or antibiotics.

Conclusion

In this paper, we accomplished the key learning objectives. Future research should focus on developing new, more complex multimodal ML approaches to integrating tumor-resident-microbe and tumor-cell data to predict cancer type. Techniques like graph-based fusion offer greater flexibility in merging structured and unstructured datasets to train ML models. Additionally, models incorporating drug chemical structure information (of drugs in development or potentially repurposed drugs) can help identify successful pairs between tumor biology and potentially effective treatment options. Finally, the approach described in this paper introduces the challenge of it being a “black box”: machine learning algorithms learn correlations between input and output data through complex nonlinear mathematical functions which are not easily interpretable in research or clinical settings. Further investigation should be made into developing models, like Random Forests or Decision Tree Ensembles, which operate in a logical structure through traversing decision nodes. These more interpretable algorithms would be more helpful in advancing our understanding of the biological phenomena at play, as opposed to simply learning correlations between input and output data to predict cancer type.

Figures and Data Samples

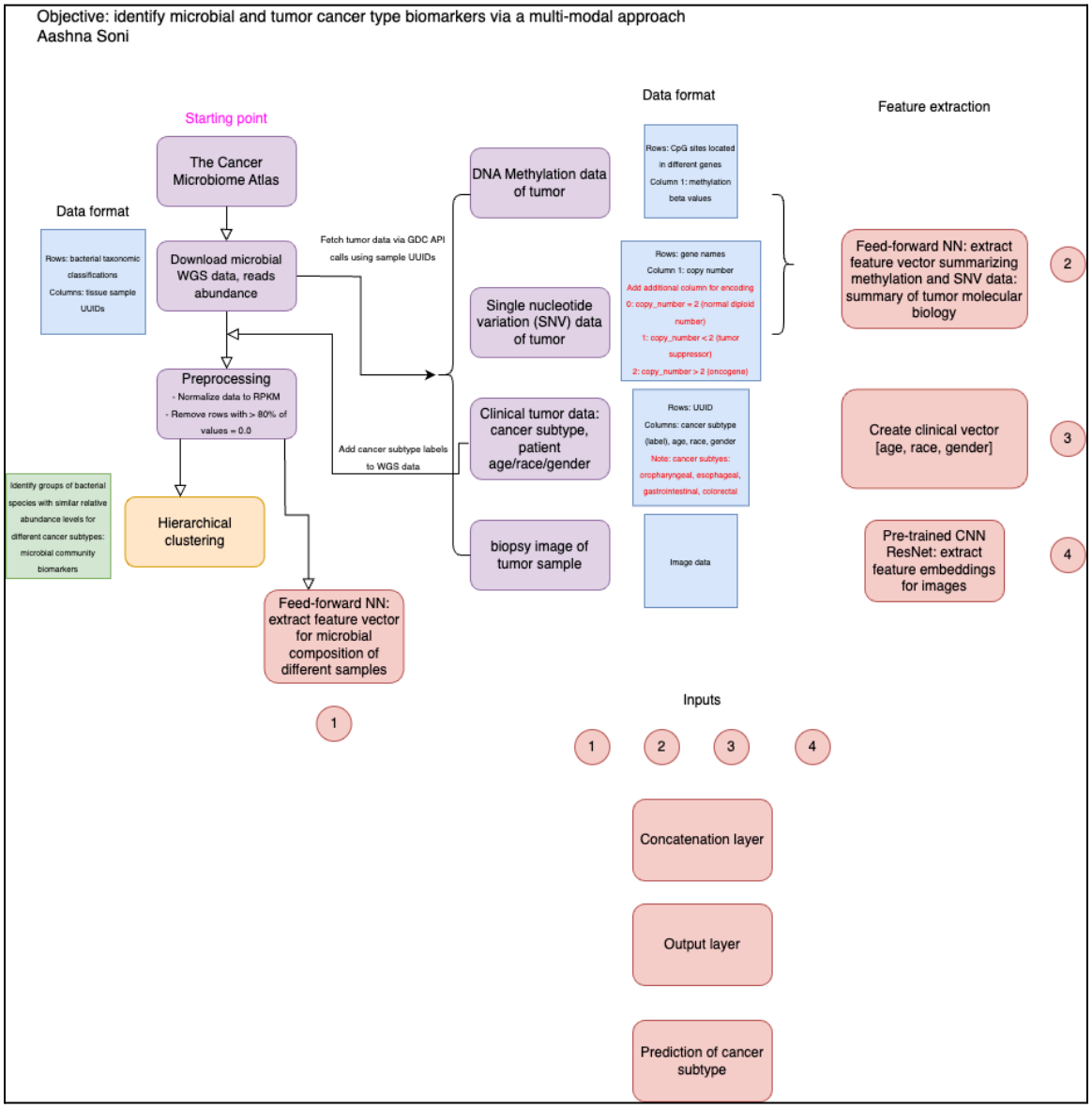

Figure 1: A diagram of the methodological workflow

Samples of data inputs (DNA methylation data, SNV data, biopsy image, microbial relative abundance data): https://drive.google.com/drive/folders/1SUXSwRxIWHXnzYg_xW4wFompekbzfpJT?usp=sharing

References

[1] *NCI Dictionary of Cancer Terms*. (n.d.-b). Cancer.gov. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/tumor-microenvironment

[2] Chen, Y., Wu, F., Wu, P., Xing, H., & Ma, T. (2022). The role of the tumor microbiome in tumor development and its treatment. *Frontiers in Immunology*, *13*. https://doi.org/10.3389/fimmu.2022.935846

[3] Cao, Y., Xia, H., Tan, X., Shi, C., Ma, Y., Meng, D., Zhou, M., Lv, Z., Wang, S., & Jin, Y. (2024). Intratumoural microbiota: a new frontier in cancer development and therapy. *Signal Transduction and Targeted Therapy*, *9*(1). https://doi.org/10.1038/s41392-023-01693-0

[4] Tjalsma, H., Boleij, A., Marchesi, J. R., & Dutilh, B. E. (2012). A bacterial driver–passenger model for colorectal cancer: beyond the usual suspects. *Nature Reviews Microbiology*, *10*(8), 575–582. https://doi.org/10.1038/nrmicro2819

[5] Nejman, D., Livyatan, I., Fuks, G., Gavert, N., Zwang, Y., Geller, L. T., Rotter-Maskowitz, A., Weiser, R., Mallel, G., Gigi, E., Meltser, A., Douglas, G. M., Kamer, I., Gopalakrishnan, V., Dadosh, T., Levin-Zaidman, S., Avnet, S., Atlan, T., Cooper, Z. A., . . . Straussman, R. (2020). The human tumor microbiome is composed of tumor type–specific intracellular bacteria. *Science*, *368*(6494), 973–980. https://doi.org/10.1126/science.aay9189

[6] Xia, B., Wang, J., Zhang, D., & Hu, X. (2023). The human microbiome links to prostate cancer risk and treatment (Review). *Oncology Reports*, *49*(6). https://doi.org/10.3892/or.2023.8560

[7] Hieken, T. J., Chen, J., Hoskin, T. L., Walther-Antonio, M., Johnson, S., Ramaker, S., Xiao, J., Radisky, D. C., Knutson, K. L., Kalari, K. R., Yao, J. Z., Baddour, L. M., Chia, N., & Degnim, A. C. (2016b). The microbiome of aseptically collected human breast tissue in benign and malignant disease. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep30751

[8] Teixeira, M., Silva, F., Ferreira, R. M., Pereira, T., Figueiredo, C., & Oliveira, H. P. (2024). A review of machine learning methods for cancer characterization from microbiome data. *Npj Precision Oncology*, *8*(1). https://doi.org/10.1038/s41698-024-00617-7

[9] Jiang, Z., Jhunjhunwala, S., Liu, J., Haverty, P. M., Kennemer, M. I., Guan, Y., Lee, W., Carnevali, P., Stinson, J., Johnson, S., Diao, J., Yeung, S., Jubb, A., Ye, W., Wu, T. D., Kapadia, S. B., De Sauvage, F. J., Gentleman, R. C., Stern, H. M., . . . Zhang, Z. (2012). The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. *Genome Research*, *22*(4), 593–601. https://doi.org/10.1101/gr.133926.111

[10] Hu, Z., Zhu, D., Wang, W., Li, W., Jia, W., Zeng, X., Ding, W., Yu, L., Wang, X., Wang, L., Shen, H., Zhang, C., Liu, H., Liu, X., Zhao, Y., Fang, X., Li, S., Chen, W., Tang, T., . . . Ma, D. (2015). Genome-wide profiling of HPV integration in cervical cancer identifies

clustered genomic hot spots and a potential microhomology-mediated integration mechanism. *Nature Genetics*, *47*(2), 158–163. https://doi.org/10.1038/ng.3178

[11] Krump, N. A., & You, J. (2018). Molecular mechanisms of viral oncogenesis in humans. *Nature Reviews Microbiology*, *16*(11), 684–698. https://doi.org/10.1038/s41579-018-0064-6

[12] Liu, D., Liu, Y., Zhu, W., Lu, Y., Zhu, J., Ma, X., Xing, Y., Yuan, M., Ning, B., Wang, Y., & Jia, Y. (2023). Helicobacter pylori-induced aberrant demethylation and expression of GNB4 promotes gastric carcinogenesis via the Hippo–YAP1 pathway. *BMC Medicine*, *21*(1). https://doi.org/10.1186/s12916-023-02842-6

[13] Zhao, H. et al. Inflammation and tumor progression: signaling pathways and targeted intervention. Signal Transduct. Target Ther. 6, 2426–2471 (2021).

[14] Huang, L., Xu, H., & Peng, G. (2018). TLR-mediated metabolic reprogramming in the tumor microenvironment: potential novel strategies for cancer immunotherapy. *Cellular and Molecular Immunology*, *15*(5), 428–437. https://doi.org/10.1038/cmi.2018.4

[15] Urban-Wojciuk, Z., Khan, M. M., Oyler, B. L., Fåhraeus, R., Marek-Trzonkowska, N., Nita-Lazar, A., Hupp, T. R., & Goodlett, D. R. (2019). The role of TLRs in anti-cancer immunity and tumor rejection. *Frontiers in Immunology*, *10*. https://doi.org/10.3389/fimmu.2019.02388

[16] Jang, G., Lee, J. W., Kim, Y. S., Lee, S. E., Han, H. D., Hong, K., Kang, T. H., & Park, Y. (2020). Interactions between tumor-derived proteins and Toll-like receptors. *Experimental & Molecular Medicine*, *52*(12), 1926–1935. https://doi.org/10.1038/s12276-020-00540-4

[17] Hernandez, C., Huebener, P., & Schwabe, R. F. (2016). Damage-associated molecular patterns in cancer: a double-edged sword. *Oncogene*, *35*(46), 5931–5941. https://doi.org/10.1038/onc.2016.104

[18] Di Lorenzo, A., Bolli, E., Tarone, L., Cavallo, F., & Conti, L. (2020). Toll-Like Receptor 2 at the Crossroad between Cancer Cells, the Immune System, and the Microbiota. *International Journal of Molecular Sciences*, *21*(24), 9418. https://doi.org/10.3390/ijms21249418

[19] Gonzalez, C., Williamson, S., Gammon, S. T., Glazer, S., Rhee, J. H., & Piwnica-Worms, D. (2023). TLR5 agonists enhance anti-tumor immunity and overcome resistance to immune checkpoint therapy. *Communications Biology*, *6*(1). https://doi.org/10.1038/s42003-022-04403-8

[20] Shalapour, S., & Karin, M. (2019). Pas de Deux: Control of Anti-tumor Immunity by Cancer-Associated Inflammation. *Immunity*, *51*(1), 15–26. https://doi.org/10.1016/j.immuni.2019.06.021

[21] Vergara, D., Simeone, P., Damato, M., Maffia, M., Lanuti, P., & Trerotola, M. (2019). The Cancer Microbiota: EMT and Inflammation as Shared Molecular Mechanisms Associated with Plasticity and Progression. *Journal of Oncology*, *2019*, 1–16. https://doi.org/10.1155/2019/1253727

[22] Gupta, I., Pedersen, S., Vranic, S., & Moustafa, A. A. (2022). Implications of Gut Microbiota in Epithelial–Mesenchymal Transition and Cancer Progression: A Concise review. *Cancers*, *14*(12), 2964. https://doi.org/10.3390/cancers14122964

[23] Shao, W., Fujiwara, N., Mouri, Y., Kisoda, S., Yoshida, K., Yoshida, K., Yumoto, H., Ozaki, K., Ishimaru, N., & Kudo, Y. (2021). Conversion from epithelial to partial-EMT phenotype by Fusobacterium nucleatum infection promotes invasion of oral cancer cells. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-94384-1

[24] Zhang, Y., Zhang, L., Zheng, S., Li, M., Xu, C., Jia, D., Qi, Y., Hou, T., Wang, L., Wang, B., Li, A., Chen, S., Si, J., & Zhuo, W. (2022). Fusobacterium nucleatum promotes colorectal cancer cells adhesion to endothelial cells and facilitates extravasation and metastasis by inducing ALPK1/NF-κB/ICAM1 axis. *Gut Microbes*, *14*(1). https://doi.org/10.1080/19490976.2022.2038852

[25] Geller, L. T., Barzily-Rokni, M., Danino, T., Jonas, O. H., Shental, N., Nejman, D., Gavert, N., Zwang, Y., Cooper, Z. A., Shee, K., Thaiss, C. A., Reuben, A., Livny, J., Avraham, R., Frederick, D. T., Ligorio, M., Chatman, K., Johnston, S. E., Mosher, C. M., . . . Straussman, R. (2017). Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. *Science*, *357*(6356), 1156–1160. https://doi.org/10.1126/science.aah5043

[26] Mohindroo, C., Hasanov, M., Rogers, J. E., Dong, W., Prakash, L. R., Baydogan, S., Mizrahi, J. D., Overman, M. J., Varadhachary, G. R., Wolff, R. A., Javle, M. M., Fogelman, D. R., Lotze, M. T., Kim, M. P., Katz, M. H., Pant, S., Tzeng, C. D., & McAllister, F. (2021). Antibiotic use influences outcomes in advanced pancreatic adenocarcinoma patients. *Cancer Medicine*, *10*(15), 5041–5050. https://doi.org/10.1002/cam4.3870

[27] Lurienne, L., Cervesi, J., Duhalde, L., De Gunzburg, J., Andremont, A., Zalcman, G., Buffet, R., & Bandinelli, P. (2020). NSCLC Immunotherapy Efficacy and Antibiotic Use: A Systematic Review and Meta-Analysis. *Journal of Thoracic Oncology*, *15*(7), 1147–1159. https://doi.org/10.1016/j.jtho.2020.03.002

[28] Kabwe, M., Dashper, S., Bachrach, G., & Tucci, J. (2021). Bacteriophage manipulation of the microbiome associated with tumour microenvironments-can this improve cancer therapeutic response? *FEMS Microbiology Reviews*, *45*(5). https://doi.org/10.1093/femsre/fuab017

[29] Raman, V., Hall, C. L., Wetherby, V. E., Witney, S. A., Van Dessel, N., & Forbes, N. S. (2024). Controlling intracellular protein delivery, tumor colonization and tissue distribution using the master regulator flhDC in a clinically relevant ΔsseJ Salmonella strain. *Molecular Therapy*. https://doi.org/10.1016/j.ymthe.2024.12.038

[30] Eisenhofer, R., Minich, J. J., Marotz, C., Cooper, A., Knight, R., & Weyrich, L. S. (2018). Contamination in low microbial biomass Microbiome Studies: issues and recommendations. *Trends in Microbiology*, *27*(2), 105–117. https://doi.org/10.1016/j.tim.2018.11.003

[31] Murovec, B., Deutsch, L., & Stres, B. (2024). Predictive modeling of colorectal cancer using exhaustive analysis of microbiome information layers available from public metagenomic data. *Frontiers in Microbiology*, *15*. https://doi.org/10.3389/fmicb.2024.1426407

[32] Dohlman, A. B., Mendoza, D. A., Ding, S., Gao, M., Dressman, H., Iliev, I. D., Lipkin, S. M., & Shen, X. (2021). The cancer microbiome atlas: a pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants. *Cell Host & Microbe*, *29*(2), 281-298.e5. https://doi.org/10.1016/j.chom.2020.12.001

[33] Freitas, P., Silva, F., Sousa, J. V., Ferreira, R. M., Figueiredo, C., Pereira, T., & Oliveira, H. P. (2023). Machine learning-based approaches for cancer prediction using microbiome data. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-38670-0

[34] Stahlschmidt, S. R., Ulfenborg, B., & Synnergren, J. (2021). Multimodal deep learning for biomedical data fusion: a review. *Briefings in Bioinformatics*, *23*(2). https://doi.org/10.1093/bib/bbab569

[35] Lee, S. J., & Rho, M. (2022). Multimodal deep learning applied to classify healthy and disease states of human microbiome. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-04773-3

[36] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1512.03385