Metagenomic Analysis of Fecal Samples as a Differential Diagnostic Tool for Colorectal Cancer

Authors: Kaitlyn Schisler, Cory Spern, Faith Shipman, Jacob Anderson

I. Problem Statement

Cancer is the second highest cause of death worldwide and in the United States, with over 10 million deaths worldwide in 2020, and over 18 million cases in 2022 (Duan et al., 2022; Siegel et al., 2025; *The Global Cancer Burden*, n.d.; *Worldwide Cancer Data*, n.d.). Among these cases, colorectal cancer (CRC) has the third highest incidence at approximately 1.9 million new cases and 904 thousand deaths in 2022. The countries most impacted include China, Japan, and the United States (*Colorectal Cancer Statistics*, n.d.; *Worldwide Cancer Data*, n.d.). A bottle neck exists with data collection and processing, leading to a 2–3 year lag in the reporting of cancer incidence and mortality rates across the world. Subsequently, Siegel et al. (2025) statistically modeled 2025 cancer impacts in the United States using prior years rates as predictors, with their analyses resulting in worrying estimates. With over 2 million new cancer cases predicted; >154 thousand alone belonging to CRC, an estimated ~53 thousand resulting in death meaning it would be the second leading cause of cancer related deaths in the United States. CRC incidence rate is increasing as those born after 1950 approach higher risk ages (Siegel et al., 2025). This is a significant burden on the healthcare system; cancer, and particularly CRC, continues to be a fundamental target for prevention, diagnosis, and treatment.

Cancer screening remains one of the best procedures for positive patient outcomes, particularly because CRC is typically asymptomatic in early stages. Early detection often results in better prognosis (Duan et al., 2022). This is reflected in the CRC screening guideline update in 2018, lowering the recommended age limit to 45 from 55 years. As a result of this guideline change, there was a three-fold increase in health insurance claims data between January 2021 and December 2022 related to CRC screenings (Siegel et al., 2025). However, the COVID-19 pandemic had the unfortunate side effect of reduced cancer screenings as hospitals had limited resources, shifting priorities, and guidelines for reduced virus exposure, as well as patient loss of employment and health insurance. Overall, cancer screenings, CRC diagnosis, and CRC-related surgeries remain reduced even after the lifting of the quarantine. With the loss of >15 million preventative procedures during the pandemic, there are predictions of 4000–7000 more deaths by CRC in the United States than original projections, particularly affecting minority groups (van den Puttelaar et al., 2023; Siegel et al., 2025).

A top priority to reduce the impact of these losses is to dissipate the backlog of preventative care. Additional colonoscopy screenings per month would be helpful (van den Puttelaar et al., 2023), however, resource availability suggests a need for tools to determine priority level in a rapid, non-invasive way. Alongside other methods like fecal occult blood tests (FOBT) and tumor markers (Duan et al., 2022), a new, accurate screening method to reduce the burden and bolster differential diagnosis for triage purposes would be highly beneficial. The human microbiome is a prime candidate for such a purpose, as samples used for current non-invasive screenings like FOBT could also be sequenced for microorganisms.

The human gut microbiome has a well known effect on host health, as organisms inhabiting the gastrointestinal (GI) tract can be harmful, helpful, or inert within the system (Coyte & Rakoff-Nahoum, 2019; Dakal et al., 2025; Kim et al., 2024). With

microbes having a large influence on the metabolism of nutrients, synthesis of or interaction with molecules, and playing a role in protection from disease, humans certainly do not function optimally without them (Fujisaka et al., 2023; Ma et al., 2022; Tarracchini et al., 2024). Prior epidemiological and scientific evidence has also suggested that the health of the colon and cancer risks are affected by the colonic microbial metabolism (O'Keefe, 2016). For example, *Lachnospiraceae* is suspected to inhibit the oral microbes from transitioning to the gut, as oral-microbe enrichment in the GI tract is known to be associated with several diseases, including CRC (Manghi et al., 2025; Zhang et al., 2022). On the other side of the spectrum, several microbial species have a pathogenic effect on the system, such as *B. fragilis, F. nucleatum, and E. coli,* which can release toxins and cause inflammation, potentially leading to tumorigenesis (Lee et al., 2023; T. Li et al., 2025; Madhogaria et al., 2022; Zhang et al., 2022). The gut microbiome exists in a dynamic equilibrium, known as homeostasis, with its disruption known as dysbiosis (Madhogaria et al., 2022; Zhang et al., 2022). Since the gut microbiome can shift from a healthy, homeostatic state and a diseased, dysbiotic one, there is interest in identifying and characterizing the organisms present in these states and their relative abundances, often called microbial fingerprinting. The microbiome can be characterized in two main ways: the taxonomic profile (what species are present) and the functional profile (what the species can do). Determining which of these profiles are the most indicative of disease progression, or if a weighted combination of the two, is key in tracking CRC progression.

Can a supervised ML model be developed to accurately predict CRC progression via a microbial fingerprint based on human metagenomic datasets?

Through a retrospective, multicohort study of metagenomic human gut microbiome data, our aim is to generate a supervised machine-learning (ML) model with clinical applications for detecting the progression of CRC, as well as guiding future research focus on the gut microbiome's role in disease. By investigating the predictive value of a microbial "fingerprint" for the stages of CRC progression, we will generate the foundation of a proof-of-concept diagnostic tool to aid in triage for easing the burden of colonoscopy suites. Based on the foundational work of others in this field, our hypothesis is that there will be a detectable difference in the microbial fingerprint of different stages of CRC development.

II. Introduction and Literature Review

While the seemingly obvious sample for gut microbiomes is stool, there appears to be debate on whether these provide a complete representation of the colon's microbiota (Levitan et al., 2023; Rode et al., 2024). There are other methods for inter-colonic microbiome sampling that capture a highly representative sample, however, these remain invasive and are often reliant on a colonoscopy suite (Levitan et al., 2023). Alternative methods, such as lavages, are being investigated by researchers such as Levitan et al. (2023), who provided evidence that it captures the less abundant taxa, though this method is not yet widely adopted. Contrary to this, Rode et al. (2024) determined that the non-invasive and low cost sampling method of the gut microbiome of stool collection by the patient remains an accurate proxy to a rectal swab sample

collected by a gastroenterologist. Regardless, the majority of gut microbiome studies use stool samples as the method of choice, as it is a readily available specimen.

Initial investigations for characterizing the gut microbiome began with the well known use of 16S rRNA gene sequencing; however, this method has major limitations, including limited resolution of taxonomic classification, no functional information on the taxa present, lacking information on other microorganisms like fungi and protists, and the reliance on references, which removes the possibility of investigating microbial dark matter (MDM) (Dakal et al., 2025; Kim et al., 2024). While it does not resolve all issues, whole metagenome sequencing (WMS) and metagenomics with the de novo assemblies of genomes results in little reliance on reference databases, resolving issues with database coverage (Dakal et al., 2025; Kim et al., 2024). The value of MDM and de novo assembly of previously uncharacterized microbes is typically only briefly discussed, likely due to the influence of technical factors such as contamination or sequencing errors rather than biological relevance, and thus should be interpreted cautiously (Wu et al., 2025). Some drawbacks with WMS include database limitations and lack of information regarding uncharacterized microorganisms, reference database bias to western populations, data sparsity in samples (where many rows in a dataset may be "zero"), or even insufficient microbial profilers that are not of diagnostic or clinical quality (Balloux et al., 2018; Dias et al., 2020; Pan, 2021; Wu et al., 2025). Despite the lack of a solution for the other issues listed, the increasing accessibility of WMS means that precision medicine, and more specifically microbiome medicine, is fast approaching as a present day answer for many human diseases, where the human microbiome is leveraged for disease prevention, diagnostics, and treatment (Kim et al., 2024).

There is precedence for this work, as there are studies linking microbial composition with health and disease states. As more research determines the importance of the human gut microbiome, the interactions associated with microorganisms and their host have been proposed as its own “organ” in the body. Ultimately, the human gut and associated microbiome exist in a delicate balance; the threat of dysbiosis must be considered (Madhogaria et al., 2022; Manghi et al., 2025; Tegegne & Savidge, 2025). Unsurprisingly, disturbances in the gut microbiome have been linked to bowel diseases, such as irritable bowel syndrome (IBS), inflammatory bowel disease (IBD), Crohn’s disease, and ulcerative colitis via mechanisms like the gut–brain axis (Elmassry et al., 2025; Madhogaria et al., 2022; Manghi et al., 2025; Postler & Ghosh, 2017; Shaikh et al., 2023). Adjacent to this is metabolic diseases; subjects diagnosed with obesity, liver disease (cirrhosis), and both Type 1 and Type 2 Diabetes (T1D and T2D, respectively) have been linked to differing gut microbiota compositions from healthy controls (Madhogaria et al., 2022; Manghi et al., 2025; Postler & Ghosh, 2017). The influential reach of the gut microbes expands beyond the walls of the digestive tract and metabolic functions, piquing the interest of various researchers as there is evidence of some authority over other systems. Cardiovascular Disease (CVD), hypertension, and asthma are examples in the cardiovascular and pulmonary systems, as well as neurological/ mental disorders and conditions such as autism, and schizophrenia, showing both relationship to the disease and influences on treatment with microbial shifts between cases and controls (Madhogaria et al., 2022; Manghi et al., 2025; Postler & Ghosh, 2017; Vasileva et al., 2024).

Of major interest to our study is the linkages to different types of cancer. There are debates that up to 20% of cancers worldwide are linked to the gut microbiota, whether in a promotional or inhibitory manner. While some are related to the GI tract and adjacent organs such as the stomach, esophageal, and oral cancers, there is also evidence that the extended reach with other diseases also applies to cancer, with linkages to cancers such as prostate, liver (hepatocellular), cervix, and even skin (Madhogaria et al., 2022; Manghi et al., 2025; Sun et al., 2023). Specifically, in the realm of CRC diagnosis and prevention, there have been great strides made towards categorizing disease states; a few notable studies that have done this include Manghi et al. (2025), T. Li et al. (2025), Lee et al. (Lee et al., 2023), and Kiran et al. (Kiran et al., 2025).The overwhelming majority of studies in the field focus on a binary classification system, with healthy and diseased states representing a binary outcome without progression from one cohort to the next. This gap is one that our study aims to fill, investigating a progression of CRC development as separate categories, revealing patterns about the effective microbial shifts that concur alongside progression.

Modeling within the field is often done through supervised learning, as the primary goal for these algorithms is to predict and classify a specific outcome based on collected data. While the majority of studies used Random Forest algorithms for their supervised learning technique (T. Li et al., 2025; Manghi et al., 2025; Piccinno et al., 2025) for their data, some used a Naive Bayes model instead (Welham et al., 2023, 2025). P. Li et al. (2025) completed a benchmark study analyzing in sequence the four major processing steps of model building, which resulted in the evaluation of over 156 combinations of tools and model adjustments, identifying Ridge and Random Forest as the top two candidates. They mention applicability to a wide range of diseases for microbiome investigations, and that their results are based on area under the curve (AUC) comparisons (P. Li et al., 2025). AUC quantifies the accuracy of the ML models, where 1 is a perfect algorithm, 0.5 is no better than random chance, and 0 gives perfectly inverted results (i.e. all positives are classified negative, and vice versa).

Within these studies investigating CRC detection using microbiomes, model accuracy varies fairly widely. For example, the AUC ranges from 0.585 to 0.88 in T. Li et al. (2025) and 0.51 to 0.95 in Manghi et al. (2025). This wide range of classification accuracy at first glance is concerning, though this may be because there is a lack of studies that investigate the contributions of archaea and fungi to these states (Kiran et al., 2025; T. Li et al., 2025). This is partially due to the relatively low abundance in the system. Despite this challenge, we hope to include as many kingdoms as is feasible in our study, which will not only contribute to this area of research, but there is evidence that this results in a more accurate model (T. Li et al., 2025).

III. Methodology

Given the ease of access to the data, we chose only stool samples to consider for model generation. The inclusionary criteria of our study cohorts are based on the disease progression of CRC. Colorectal polyps and adenomas contribute to the occurrence of CRC; these are benign growths that have a high chance of becoming cancerous (15-40% and 2-40%, respectively), indicating relevance for CRC progression tracking (Duan et al., 2022). Our study focuses on these benign and cancerous growths. Our multi-cohort study will be divided into the following categories: healthy control

(“HC”), polyp/adenoma (“PA”), polyp/adenoma with comorbidities (“PA+”), CRC (“CRC”), history of CRC with resection (“CRC-H”), CRC with comorbidities (“CRC+”), and any diagnosed disease without current or history of CRC-related growths (“Other”). Furthermore, with microbiome shifts in aging being a major factor of variation, as well as CRC risk increasing with age, we will only be including samples of individuals over 18 years. Keeping individuals ≥18 years old rather than ≥45 years old, the recommended age to start screening, is that (i) there is evidence that incidence rate is increasing in younger populations year over year (*Cancer Over Time*, n.d.), and (ii) symptomatic individuals 20 years or older are suggested to be screened using other methods (Duan et al., 2022). Ideally, we hope to make our model generalizable to all ages that CRC effects, especially as risks are increasing.

A few papers were influential in guiding our research approaches. P. Li et al. (2025) has demonstrated that the performance of microbiome-based disease diagnostic classifiers depends heavily on data processing and model selection. Maghi et al. (2025) has generated a manually curated set of metagenomic data with accompanying metadata, in which the information is available in a standard form through Bioconductor in R. Much of the processing is completed, such as determining marker abundance/presence, and pathway abundance/coverage through HUMAnN3 and MetapPhlAn via bioBakery3, which is known to reduce ambiguity and improve functional mapping (Pita-Galeana et al., 2025). This manually curated dataset, version 3 of the curatedMetagenomicData (cMD3), will be quintessential to jumpstarting our analysis, as it contains 94 datasets from 42 countries, totaling 22,710 samples (Manghi et al., 2025).

An important portion of the dataset is several Human Microbiome Project (HMP) studies, which pioneered the development of a human microbiome reference database in two phases. The first phase (HMP1) focused on the creation of a reference catalogue in healthy human hosts, while the second phase (HMP2 or iHMP) focused on time series data and microbiome interactions with human metabolism and immunity (Kim et al., 2024; NIH Human Microbiome Portfolio Analysis Team, 2019; The Integrative HMP (iHMP) Research Network Consortium et al., 2019). Since HMP1 focused on human subjects without any clinical diagnoses, this will be an important “HC” in our study sample. Within cMD3, this has been identified as the study named HMP_2012, containing 748 samples after filtering for our age inclusion criteria, though this is not filtering for other inclusion criteria. The following parameters were applied to the full cMD3 data set using RStudio, resulting in a curated subset of 12 studies to be included in the model, totaling 1790 samples; 764 “HC” and 1026 with an indicated disease (Appendix A and B): (i) sample site is “stool”, (ii) age is ≥18 years, (iii) the disease was not blank, and (iv) the disease contained some text regarding CRC, polyp/adenoma, or (v) the study was HMP_2012.

This subset is expected to reflect many of the representation biases present in cMD3 as well as overarching public data trends, though this is yet to be determined with exploratory data analysis. It will likely be dominated by samples from the United States, China, and European countries, with little representation of those in regions such as Africa, South America, the Middle East, and much of Asia (Kim et al., 2024; Manghi et al., 2025). With the main contributor to an individual’s microbiome being their geographical location and therefore diet, geography will likely have a large influence on our subset of data (Dakal et al., 2025; Singh et al., 2017; Zmora et al., 2019).

Initial exploration of the quantitative and qualitative data through descriptive and inferential statistical analysis will be conducive to detecting any microbiome shifts prior to attempting a ML model. Analyses we plan to conduct are: (i) Alpha diversity analysis for the determination of within group variation. This will be done using Shannon, Chao1, and Faith indices (Dakal et al., 2025; Pita-Galeana et al., 2025). (ii) Beta diversity analysis for between group variation. This utilizes the Bray-Curtis dissimilarity index and Unifrac distances (Dakal et al., 2025; Pita-Galeana et al., 2025). (iii) Statistical tests to determine significant differences in the alpha and beta diversity, as well as between categorical groups such as the CRC progression classes. These tests may include the Wilcoxon rank sum test, permutational multivariate analysis of variance (PERMANOVA), and Linear Discriminant Analysis Effect Size (LEfSE) (Dakal et al., 2025; T. Li et al., 2025). (iv) Various visualization techniques, such as Principal Component Analysis (PCA) and Principal Coordinates Analysis (PCoA), box plots, bar plots, and heat maps, which will reveal hidden patterns and effectively present and support our findings from the statistical analyses (Dakal et al., 2025; T. Li et al., 2025; Pita-Galeana et al., 2025). It will also reveal if grouping based on confounding factors will be required, such as age or geographical location. (v) MDM score, which is a proportion of the MDM, will be investigated within and between samples to determine contributions to disease state, as well as offer guidance for further research on the microbial influences on disease.

It is well established that there is success using ML algorithms for diagnostics and detecting disease biomarkers for a wide range of diseases, including CRC (Kim et al., 2024; P. Li et al., 2025; Manghi et al., 2025). As previously mentioned, P. Li et al. (2025) determined Random Forests were the prime choice in supervised ML algorithm for disease classification with minimal preprocessing, which is optimal for our use case; the only suggested preprocessing step was to remove the low abundance taxa with a 0.001% threshold. This supervised learning technique allows us to indicate the seven disease categories (“HC”, “PA”, “PA+”, “CRC”, “CRC+”, “CRC-H”, “Other”), is robust to overfitting, effectively captures feature relationships, and is suitable for use with large data sets, making it the prime ML model for CRC progression identification with microbial fingerprinting (Dakal et al., 2025; P. Li et al., 2025; Pita-Galeana et al., 2025). There is also evidence that multi-kingdom analysis, such as including bacteria and archaea, results in a more accurate model, according to an AUC statistic (T. Li et al., 2025). However, with these species being in relatively low abundance in the human gut, it is possible these would be removed during preprocessing. Regardless, we will attempt to include any remaining taxa in a multi-kingdom model to increase accuracy. Model accuracy can be quantified using a per-class or macro-averaged F1 score (Haldar et al., 2024; Hicks et al., 2022; Pita-Galeana et al., 2025), performance will also be assessed using leave-one-dataset-out (LODO) cross-validation method (Kubinski et al., 2022).

IV. Expected Outcomes & Impact

The human gut microbiome plays a role in immunity, metabolism, and overall human health. Scientists have begun to scratch the surface of the interplay between the human microbiome and the intricate role it plays on human health and development. Every individual has a unique microbiome fingerprint that changes over time with aging, dietary shifts, and environmental or geographic influences. Studies have also shown that microbiomes perform biochemical functions that can impact disease diagnosis and

treatment strategies (Hajjo et al., 2022). Accurate clinical decision making depends on an integration of all the available patient information, meaning a microbiome based insight would provide an additional layer of evidence. A tool like this would support clinicians' differential diagnosis and treatment in a timely, non-invasive manner.

This study is expected to demonstrate that gut microbiome fingerprints contain measurable and clinically significant patterns applicable to early detection and treatment strategies of various diseases. Specifically, CRC is explored in comparison to patients with and without polyps. It is anticipated that microbial community features and MDM scores will identify and distinguish CRC samples from that of the healthy control and patients with polyps. These findings will contribute to the growing evidence that microbiome-based biomarkers may provide a non-invasive layer of diagnostic insight alongside traditional clinical screening methods already in place.

Primary outcomes of this work will be the development of a proof-of-concept tool capable of ranking microbiome fingerprints in comparison to diseased states based on the metagenomic microbiome database. Integration of taxonomic and functional profiles allows the model to generate interpretable outputs such as disease similarity scores in addition to MDM scores for guiding future diagnostic and biomarker focus. This approach may clarify how microbial dysbiosis emerges in CRC and how these patterns can be represented in an accessible diagnostic framework.

Utilizing CRC as a focus for a proof-of-concept disease-state tool allows for a controlled evaluation while maintaining feasibility within the scope of this study. If further developed beyond a prototypic concept piece, it could add another tool to the proverbial belt that healthcare providers have at their disposal. However, the framework developed here is expected to be broadly extensible beyond CRC alone. Future research could apply this approach to additional disease systems and conditions such as T1D and T2D, various other cancers, or even hormonal and microbial shifts with the use of birth control. Similarly, application to medication-driven microbiome shifts for visualization and quantification in cases of antibiotics or even SSRIs would be possible. All of which are important areas of study and interest to the group but are currently excluded from this study due to limitations in time and availability of curated metadata. Establishing a scalable microbiome proof-of-concept tool will lay the groundwork for future disease diagnostic mapping and expanded clinical applications as more data becomes available.

V. References

Balloux, F., Brønstad Brynildsrud, O., Van Dorp, L., Shaw, L. P., Chen, H., Harris, K. A., Wang, H., & Eldholm, V. (2018). From Theory to Practice: Translating Whole-Genome Sequencing (WGS) into the Clinic. *Trends in Microbiology*, *26*(12), 1035–1048. https://doi.org/10.1016/j.tim.2018.08.004

Cancer Over Time. (n.d.). Retrieved January 31, 2026, from https://gco.iarc.fr/overtime

Colorectal cancer statistics. (n.d.). World Cancer Research Fund. Retrieved January 31, 2026, from https://www.wcrf.org/preventing-cancer/cancer-statistics/colorectal-cancer-statistics/

Coyte, K. Z., & Rakoff-Nahoum, S. (2019). Understanding Competition and Cooperation within the Mammalian Gut Microbiome. *Current Biology*, *29*(11), R538–R544. https://doi.org/10.1016/j.cub.2019.04.017

Dakal, T. C., Xu, C., & Kumar, A. (2025). Advanced computational tools, artificial intelligence and machine-learning approaches in gut microbiota and biomarker identification. *Frontiers in Medical Technology*, *6*, 1434799. https://doi.org/10.3389/fmedt.2024.1434799

Dias, C. K., Starke, R., Pylro, V. S., & Morais, D. K. (2020). Database limitations for studying the human gut microbiome. *PeerJ Computer Science*, *6*, e289. https://doi.org/10.7717/peerj-cs.289

Duan, B., Zhao, Y., Bai, J., Wang, J., Duan, X., Luo, X., Zhang, R., Pu, Y., Kou, M., Lei, J., & Yang, S. (2022). Colorectal Cancer: An Overview. In Cellular and Molecular Oncobiology Program, Cellular Dynamic and Structure Group, National Cancer

Institute-INCA, Rio de Janeiro, Brazil & J. Andres Morgado-Diaz (Eds.), *Gastrointestinal Cancers* (pp. 1–12). Exon Publications. https://doi.org/10.36255/exon-publications-gastrointestinal-cancers-colorectal-cancer

Elmassry, M. M., Sugihara, K., Chankhamjon, P., Kim, Y., Camacho, F. R., Wang, S., Sugimoto, Y., Chatterjee, S., Chen, L. A., Kamada, N., & Donia, M. S. (2025). A meta-analysis of the gut microbiome in inflammatory bowel disease patients identifies disease-associated small molecules. *Cell Host & Microbe*, *33*(2), 218-234.e12. https://doi.org/10.1016/j.chom.2025.01.002

Fujisaka, S., Watanabe, Y., & Tobe, K. (2023). The gut microbiome: A core regulator of metabolism. *Journal of Endocrinology*, *256*(3), e220111. https://doi.org/10.1530/JOE-22-0111

Hajjo, R., Sabbah, D. A., & Al Bawab, A. Q. (2022). Unlocking the Potential of the Human Microbiome for Identifying Disease Diagnostic Biomarkers. *Diagnostics*, *12*(7), 1742. https://doi.org/10.3390/diagnostics12071742

Haldar, S., Stein-Thoeringer, C., & Borisov, V. (2024). *Interpreting Microbiome Relative Abundance Data Using Symbolic Regression* (arXiv:2410.16109). arXiv. https://doi.org/10.48550/arXiv.2410.16109

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, *12*(1), 5979. https://doi.org/10.1038/s41598-022-09954-8

Kim, N., Ma, J., Kim, W., Kim, J., Belenky, P., & Lee, I. (2024). Genome-resolved metagenomics: A game changer for microbiome medicine. *Experimental & Molecular Medicine*, *56*(7), 1501–1512. https://doi.org/10.1038/s12276-024-01262-7

Kiran, N. S., Chatterjee, A., Yashaswini, C., Deshmukh, R., Alsaidan, O. A., Bhattacharya, S., & Prajapati, B. G. (2025). The gastrointestinal mycobiome in inflammation and cancer: Unraveling fungal dysbiosis, pathogenesis, and therapeutic potential. *Medical Oncology*, *42*(6), 195. https://doi.org/10.1007/s12032-025-02761-x

Kubinski, R., Djamen-Kepaou, J.-Y., Zhanabaev, T., Hernandez-Garcia, A., Bauer, S., Hildebrand, F., Korcsmaros, T., Karam, S., Jantchou, P., Kafi, K., & Martin, R. D. (2022). Benchmark of Data Processing Methods and Machine Learning Models for Gut Microbiome-Based Diagnosis of Inflammatory Bowel Disease. *Frontiers in Genetics*, *13*, 784397. https://doi.org/10.3389/fgene.2022.784397

Lee, J. W. J., Plichta, D. R., Asher, S., Delsignore, M., Jeong, T., McGoldrick, J., Staller, K., Khalili, H., Xavier, R. J., & Chung, D. C. (2023). Association of distinct microbial signatures with premalignant colorectal adenomas. *Cell Host & Microbe*, *31*(5), 827-838.e3. https://doi.org/10.1016/j.chom.2023.04.007

Levitan, O., Ma, L., Giovannelli, D., Burleson, D. B., McCaffrey, P., Vala, A., & Johnson, D. A. (2023). The gut microbiome-Does stool represent right? *Heliyon*, *9*(3), e13602. https://doi.org/10.1016/j.heliyon.2023.e13602

Li, P., Li, M., & Chen, W.-H. (2025). Best practices for developing microbiome-based disease diagnostic classifiers through machine learning. *Gut Microbes*, *17*(1), 2489074. https://doi.org/10.1080/19490976.2025.2489074

Li, T., Coker, O. O., Sun, Y., Li, S., Liu, C., Lin, Y., Wong, S. H., Miao, Y., Sung, J. J. Y., & Yu, J. (2025). Multi-Cohort Analysis Reveals Altered Archaea in Colorectal Cancer Fecal Samples Across Populations. *Gastroenterology*, *168*(3), 525-538.e2. https://doi.org/10.1053/j.gastro.2024.10.023

Ma, Y., Liu, X., & Wang, J. (2022). Small molecules in the big picture of gut microbiome-host cross-talk. *eBioMedicine*, *81*, 104085. https://doi.org/10.1016/j.ebiom.2022.104085

Madhogaria, B., Bhowmik, P., & Kundu, A. (2022). Correlation between human gut microbiome and diseases. *Infectious Medicine*, *1*(3), 180–191. https://doi.org/10.1016/j.imj.2022.08.004

Manghi, P., Antonello, G., Schiffer, L., Golzato, D., Wokaty, A., Beghini, F., Mirzayi, C., Long, K., Gravel-Pucillo, K., Piccinno, G., Gamboa-Tuz, S. D., Bonetti, A., D'Amato, G., Azhar, R., Eckenrode, K., Zohra, F., Giunchiglia, V., Keller, M., Pedrotti, A., … Waldron, L. (2025). Meta-analysis of 22,710 human microbiome metagenomes defines an oral-to-gut microbial enrichment score and associations with host health and disease. *Nature Communications*, *17*(1), 196. https://doi.org/10.1038/s41467-025-66888-1

NIH Human Microbiome Portfolio Analysis Team. (2019). A review of 10 years of human microbiome research activities at the US National Institutes of Health, Fiscal

Years 2007-2016. *Microbiome*, *7*(1), 31. https://doi.org/10.1186/s40168-019-0620-y

O'Keefe, S. J. D. (2016). Diet, microorganisms and their metabolites, and colon cancer. *Nature Reviews Gastroenterology & Hepatology*, *13*(12), 691–706. https://doi.org/10.1038/nrgastro.2016.165

Pan, A. Y. (2021). Statistical analysis of microbiome data: The challenge of sparsity. *Current Opinion in Endocrine and Metabolic Research*, *19*, 35–40. https://doi.org/10.1016/j.coemr.2021.05.005

Piccinno, G., Thompson, K. N., Manghi, P., Ghazi, A. R., Thomas, A. M., Blanco-Míguez, A., Asnicar, F., Mladenovic, K., Pinto, F., Armanini, F., Punčochář, M., Piperni, E., Heidrich, V., Fackelmann, G., Ferrero, G., Tarallo, S., Nguyen, L. H., Yan, Y., Keles, N. A., … Segata, N. (2025). Pooled analysis of 3,741 stool metagenomes from 18 cohorts for cross-stage and strain-level reproducible microbial biomarkers of colorectal cancer. *Nature Medicine*, *31*(7), 2416–2429. https://doi.org/10.1038/s41591-025-03693-9

Pita-Galeana, M. A., Ruhle, M., López-Vázquez, L., De Anda-Jáuregui, G., & Hernández-Lemus, E. (2025). Computational Metagenomics: State of the Art. *International Journal of Molecular Sciences*, *26*(18), 9206. https://doi.org/10.3390/ijms26189206

Postler, T. S., & Ghosh, S. (2017). Understanding the Holobiont: How Microbial Metabolites Affect Human Health and Shape the Immune System. *Cell Metabolism*, *26*(1), 110–130. https://doi.org/10.1016/j.cmet.2017.05.008

Rode, J., Brengesjö Johnson, L., König, J., Rangel, I., Engstrand, L., Repsilber, D., & Brummer, R. J. (2024). Fecal samples and rectal swabs adequately reflect the human colonic luminal microbiota. *Gut Microbes*, *16*(1), 2416912. https://doi.org/10.1080/19490976.2024.2416912

Shaikh, S. D., Sun, N., Canakis, A., Park, W. Y., & Weber, H. C. (2023). Irritable Bowel Syndrome and the Gut Microbiome: A Comprehensive Review. *Journal of Clinical Medicine*, *12*(7), 2558. https://doi.org/10.3390/jcm12072558

Siegel, R. L., Kratzer, T. B., Giaquinto, A. N., Sung, H., & Jemal, A. (2025). Cancer statistics, 2025. *CA: A Cancer Journal for Clinicians*, *75*(1), 10–45. https://doi.org/10.3322/caac.21871

Singh, R. K., Chang, H.-W., Yan, D., Lee, K. M., Ucmak, D., Wong, K., Abrouk, M., Farahnik, B., Nakamura, M., Zhu, T. H., Bhutani, T., & Liao, W. (2017). Influence of diet on the gut microbiome and implications for human health. *Journal of Translational Medicine*, *15*(1), 73. https://doi.org/10.1186/s12967-017-1175-y

Sun, J., Chen, F., & Wu, G. (2023). Potential effects of gut microbiota on host cancers: Focus on immunity, DNA damage, cellular pathways, and anticancer therapy. *The ISME Journal*, *17*(10), 1535–1551. https://doi.org/10.1038/s41396-023-01483-0

Tarracchini, C., Lugli, G. A., Mancabelli, L., Van Sinderen, D., Turroni, F., Ventura, M., & Milani, C. (2024). Exploring the vitamin biosynthesis landscape of the human gut microbiota. *mSystems*, *9*(10), e00929-24. https://doi.org/10.1128/msystems.00929-24

Tegegne, H. A., & Savidge, T. C. (2025). Leveraging human microbiomes for disease prediction and treatment. *Trends in Pharmacological Sciences*, *46*(1), 32–44. https://doi.org/10.1016/j.tips.2024.11.007

The Global Cancer Burden. (n.d.). Retrieved January 31, 2026, from https://www.cancer.org/about-us/our-global-health-work/global-cancer-burden.html

The Integrative HMP (iHMP) Research Network Consortium, Proctor, L. M., Creasy, H. H., Fettweis, J. M., Lloyd-Price, J., Mahurkar, A., Zhou, W., Buck, G. A., Snyder, M. P., Strauss, J. F., Weinstock, G. M., White, O., & Huttenhower, C. (2019). The Integrative Human Microbiome Project. *Nature*, *569*(7758), 641–648. https://doi.org/10.1038/s41586-019-1238-8

van den Puttelaar, R., Lansdorp-Vogelaar, I., Hahn, A. I., Rutter, C. M., Levin, T. R., Zauber, A. G., & Meester, R. G. S. (2023). Impact and Recovery from COVID-19–Related Disruptions in Colorectal Cancer Screening and Care in the US: A Scenario Analysis. *Cancer Epidemiology, Biomarkers & Prevention*, *32*(1), 22–29. https://doi.org/10.1158/1055-9965.EPI-22-0544

Vasileva, S. S., Yang, Y., Baker, A., Siskind, D., Gratten, J., & Eyles, D. (2024). Associations of the Gut Microbiome With Treatment Resistance in Schizophrenia. *JAMA Psychiatry*, *81*(3), 292. https://doi.org/10.1001/jamapsychiatry.2023.5371

Welham, Z., Li, J., Engel, A. F., & Molloy, M. P. (2023). Mucosal Microbiome in Patients with Early Bowel Polyps: Inferences from Short-Read and Long-Read 16S rRNA Sequencing. *Cancers*, *15*(20), 5045. https://doi.org/10.3390/cancers15205045

Welham, Z., Li, J., Tse, B., Engel, A., & Molloy, M. P. (2025). Gut Mucosal Microbiome of Patients With Low-Grade Adenomatous Bowel Polyps. *Gastro Hep Advances*, *4*(8), 100687. https://doi.org/10.1016/j.gastha.2025.100687

Worldwide cancer data. (n.d.). World Cancer Research Fund. Retrieved January 31, 2026, from https://www.wcrf.org/preventing-cancer/cancer-statistics/worldwide-cancer-data/

Wu, Q., Lu, S., Wang, L., Liao, X., & Wei, D. (2025). Gut microbiota and intestinal polyps: A systematic review and meta-analysis based on 16S rRNA gene sequencing. *Gut Pathogens*, *17*, 104. https://doi.org/10.1186/s13099-025-00784-3

Zhang, Y., Zhou, L., Xia, J., Dong, C., & Luo, X. (2022). Human Microbiome and Its Medical Applications. *Frontiers in Molecular Biosciences*, *8*, 703585. https://doi.org/10.3389/fmolb.2021.703585

Zmora, N., Suez, J., & Elinav, E. (2019). You are what you eat: Diet, health and the gut microbiota. *Nature Reviews Gastroenterology & Hepatology*, *16*(1), 35–56. https://doi.org/10.1038/s41575-018-0061-2

VI. Appendices

Appendix A: Contributing Studies and Samples Per Study, Descending Order

Study Name	Frequency
YachidaS_2019	616
ZellerG_2014	156
FengQ_2015	154
HMP_2012	147
YuJ_2015	128
WirbelJ_2018	125
VogtmannE_2016	104
HanniganGD_2017	81
ThomasAM_2018a	80
ThomasAM_2019_c	80
GuptaA_2019	60
ThomasAM_2018b	59

Note:

Study name is the last name, first initial, and publication year of the study.
If there are multiple with the same name, then a lowercase alphabetical letter is also appended.

Appendix B: Sample Count per CRC Progression Category

CRC Disease Progression Category	Frequency
HC	764
PA	164
PA+	50
CRC	625
CRC+	76
CRC-H	40
Other	71