

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

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

I. Problem Statement

Cancer ranks as the second highest cause of death both 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, the third highest incidence of cancer belongs to colorectal cancer (CRC), with approximately 1.9 million new incidence and 904 thousand deaths in 2022. The countries most impacted by CRC include China, Japan, and the United States (*Colorectal Cancer Statistics*, n.d.; *Worldwide Cancer Data*, n.d.).

While data availability lags by 2-3 years due to processing time, statistical modeling can be performed using prior data. Siegel et al. (2025) conducted such modeling to predict cancer impacts in the United States. Their analyses resulted in worrying estimates, with over 2 million new cancer cases predicted for 2025. Of these cases, >154 thousand are CRC, with ~53 thousand estimated deaths, making it 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 burden on the healthcare system means that cancer, particularly CRC, is a major target for prevention, diagnosis, and treatment.

Cancer screening remains one of the best ways for positive patient outcomes, particularly because CRC is typically asymptomatic in early stages. Catching the disease early in development will result in a better prognosis (Duan et al., 2022). This is reflected in the CRC screening guideline update in 2018, reducing the recommended lower 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 these 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. Due to a screening loss of over 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 to reduce the backlog 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.

The human microbiome is a complex system, which consists of anatomical regions for microbial communities in and on the human body (Reynoso-García et al., 2022). One region of particular interest for a variety of clinical diseases and conditions is

the gut microbiome (Madhogaria et al., 2022). Microorganisms of the human gut serve a multitude of functions including metabolism, vitamin synthesis, interactions with pharmaceutical products, and even release bioactive small molecules (Fujisaka et al., 2023; Ma et al., 2022; Tarracchini et al., 2024). However, not all microorganisms are beneficial, forming more of a commensal relationship with the human gut. It is possible that some species merely outcompete pathogens instead of providing any metabolic benefit to the human host (Coyte & Rakoff-Nahoum, 2019; Dakal et al., 2025).

Through the analysis of previously acquired metagenomic data of the human gut microbiome, our aim is to generate a machine-learning (ML) model with clinical applications for detecting the progression of CRC, as well as to help guide future research focus on the gut microbiome 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 clinical tool to aid in differential diagnoses and 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 those in different stages of CRC development.

Can a ML model be developed that accurately predicts CRC progression via a microbial “fingerprint” based on human metagenomic datasets?

One critical portion of our project will be determining what characterization of the microbiome will result in the optimal model. 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). Fingerprinting CRC would mean determining which of these profiles are the most indicative of disease progression, or if a weighted combination of the two would result in a better fit. Another consideration is the idea that the dose makes the poison, commonly used in toxicological principles and often attributed to Paracelsus in 1538 (Paracelsus, 1538/2003). A similar principle may apply here with microbial compositions, where a threshold or proportion, not the mere presence, is indicative of CRC progression.

A few challenges are expected to arise during this process. With the main contributor to an individual’s microbiome being their geographical location and consequently their diets, geography will likely have a large influence on the composition of our samples (Dakal et al., 2025; Singh et al., 2017; Zmora et al., 2019). A similar phenomena occurs within family units, bringing into question how much of a role geography and familial ties will play in disease classification, and how generalizable our final model will be to a wide range of populations. This is especially pertinent given that the NCBI Sequence Read Archive (SRA) is biased towards samples in the United States, China, and European countries (Kim et al., 2024).

Additionally, microbial dark matter (MDM) are the microorganisms that remain unculturable and thus have little to no sequences in reference databases (Kim et al., 2024; Pavlopoulos et al., 2023). With metagenomics and de novo assembly, these organisms are detectable and could be an incredibly valuable resource. As such, this project will introduce an MDM score, where the 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.

II. Introduction and Literature Review

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 (Kim et al., 2024). With microbes having a large influence on the metabolism of nutrients and synthesis of molecules in downstream systems, and others playing a role in protection from disease, humans certainly do not function optimally without them (Tarracchini et al., 2024). 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). As such, the gut microbiome exists in a dynamic equilibrium, known as homeostasis. The disruption of this balance is 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.

Initial investigations 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 removing the possibility of investigating MDM (Dakal et al., 2025; Kim et al., 2024). Some drawbacks with current methods 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 identification pipelines that are not of a diagnostic or clinical quality (Balloux et al., 2018; Wu et al., 2025; Dias et al., 2020; Pan, 2021). While it does not resolve all issues, whole metagenome sequencing (WMS) and metagenomics resolves the issues with database coverage, as the entire genome of all species present are sequenced, resulting in little reliance on reference databases with the de novo assemblies of genomes (Dakal et al., 2025; Kim et al., 2024). Despite the lack of a resolution for the other issues listed, the increasing accessibility to WMS means that precision medicine, and more specifically microbiome medicine, is fast approaching as a present day solution to 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 with health- and disease-states (Lee et al., 2023; Madhogaria et al., 2022; Manghi et al., 2025; Tegegne & Savidge, 2025; Welham et al., 2025). 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). Within these studies, model accuracy varies. For example, the area under the curve (AUC), which quantifies the accuracy of the ML models, ranges from 0.585 to 0.88 in T. Li et al. (2025), and 0.51 to 0.95 in Manghi et al. (2025), 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, all negatives are classified positive). This wide range of classification accuracy seems 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 project, which will not only contribute to this research area, but there is evidence that this results in a more accurate model (T. Li et al., 2025).

Throughout our literature review, we have yet to come across a study investigating CRC with metastases as a separate category in classification, suggesting it is incredibly rare or has not been considered previously. Further, there appears to be few studies that effectively examine early stages in cancer progression, such as early adenomas (Lee et al., 2023). We hope to address this gap by investigating all stages of CRC development as separate categories, from early stage polyps/adenomas, to CRC, to CRC with metastases. This will reveal patterns about progression and the effective microbial shifts that concur alongside it.

Another area that does not appear to be addressed in depth in most studies is the value of the MDM within samples. This could be partially due to dark matter being influenced by technical factors such as contamination or sequencing errors rather than biological relevance as noted by Wu et al. (2025). Therefore MDM results must be interpreted cautiously taking into account the noise in analysis. Many studies discuss the de novo assembly of previously uncharacterized microbes, but there is rarely a statistic stating the proportion of MDM in the sample. This could potentially guide other researchers to focus on certain homeostatic or dysbiotic states in search of novel biomarkers for the disease, filling in gaps and ultimately bettering human health. We hope to also address this in our model by providing a MDM score.

III. Methodology

The inclusionary criteria of our study subjects is based on the disease progression of CRC. As we are focusing on the gut microbiome, only stool sample types will be considered for model generation. 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). We have decided to focus on these benign and cancerous growths. Our multi-cohort study would then have the following categories: healthy control (HC), polyp/adenoma (PA), CRC, CRC with metastases (CRC-M), and CRC with comorbidities (CRC+). 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 (Maghi 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.

To determine other studies contained in cMD3 that would be suitable for our analysis, the `sampleMetadata` variable provided by the package for exploration was filtered to the following parameters using RStudio: (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. These were grouped by study name, resulting in 12 studies to be included in the ML model, totaling 1797 samples. There are 771 HC and 1026 with an indicated disease (Appendix A). None of these have indicated current antibiotic use (“no” and NULL values only).

This subset is expected to reflect many of the representation biases present in cMD3 as well as the overarching public data theme, 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; Maghi et al., 2025). If, during further investigations, it is deemed necessary to supplement cMD3 with data within these regions, the SRA will be instrumental to expansion (Kim et al., 2024), with described instructions to incorporate these seamlessly into the dataset (Maghi et al., 2025).

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 is typically 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 difference in the alpha and beta diversity, which 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 our findings in statistical analyses (Dakal et al., 2025; T. Li et al., 2025; Pita-Galeana et al., 2025). These visualizations will be important for determining the MDM score for each cohort and direct further research to relevant cohorts. It will also reveal if grouping based on confounding factors will be required, such as age or geographical location.

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). One of the most widely established algorithms with the most promising results in this field are Random Forests. Li et al. (2025) completed a benchmark comparison with seven other algorithms, in which Random Forests outcompeted all others and required the least preprocessing; 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 five disease categories (HC, AP, CRC, CRC-M, CRC+), 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 disease 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).

IV. Expected Outcomes & Impact

The human gut microbiome plays a role in immunity, metabolism, and overall human health. Only in recent years have scientists begun to fully understand the complexity of the gut microbial ecosystem. 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 project 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 project. 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 Type 1 and 2 diabetes, 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 project 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>

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>

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>

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>

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>

Paracelsus. (2003). *Septem Defensiones: Die Selbstverteidigung eines Aussenseiters* mit einem Reprint der Ausgabe Basel 1589 (G. Pörksen, Trans.). Schwabe & Co. AG Verlag. <http://www.zeno.org/Philosophie/M/Paracelsus/Septem+Defensiones/Die+dritte+Defension+wegen+des+Schreibens+der+neuen+Rezepte> (Original work published 1538)

Pavlopoulos, G. A., Baltoumas, F. A., Liu, S., Selvitopi, O., Camargo, A. P., Nayfach, S., Azad, A., Roux, S., Call, L., Ivanova, N. N., Chen, I. M., Paez-Espino, D., Karatzas, E., Novel Metagenome Protein Families Consortium, Acinas, S. G., Ahlgren, N., Attwood, G., Baldrian, P., Berry, T., ... Kyrpides, N. C. (2023). Unraveling the functional dark matter through global metagenomics. *Nature*, 622(7983), 594–602. <https://doi.org/10.1038/s41586-023-06583-7>

- 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>
- Reynoso-García, J., Miranda-Santiago, A. E., Meléndez-Vázquez, N. M., Acosta-Pagán, K., Sánchez-Rosado, M., Díaz-Rivera, J., Rosado-Quiñones, A. M., Acevedo-Márquez, L., Cruz-Roldán, L., Tosado-Rodríguez, E. L., Figueroa-Gispert, M. D. M., & Godoy-Vitorino, F. (2022). A complete guide to human microbiomes: Body niches, transmission, development, dysbiosis, and restoration. *Frontiers in Systems Biology*, 2, 951403.
<https://doi.org/10.3389/fsysb.2022.951403>
- 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>
- 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>

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: Qualifying Studies Table

Study Name	Sample Count
Total	1797
YachidaS_2019	616
ZellerG_2014	156
FengQ_2015	154
HMP_2012	147
YuJ_2015	128
WirbelJ_2018	125
VogtmannE_2016	110
HanniganGD_2017	81
ThomasAM_2018a	80
ThomasAM_2019_c	80
GuptaA_2019	60
ThomasAM_2018b	60