Exploratory Data Analysis

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

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

Introduction

The standardization of data collection and processing is the foundation of a strong machine learning algorithm (P. Li et al., 2025). As of December 2025, Manghi et al. (2025) released version 3, their newest version, of their curatedMetagenomicData set (cMD3), which is a manually curated metagenomic data set from 94 studies in 42 countries, totaling 22,710 samples. This was standardized and made available through Bioconductor in R, well known for its package use in genomic and metagenomic studies (Drnevich et al., 2025), and from which previous versions of the cMD have been used in similar studies (Pasolli et al., 2017). Since much of the processing is completed for taxonomic and pathway identification through HUMAnN3 and MetapPhlAn via bioBakery3, it is an ideal data source for jumpstarting our analysis. With there being precedence for using the cMD for this type of work, it remains a fine choice for This data set, however, includes various sample sites, diseases, and variables irrelevant to our study. Thus, the following inclusionary criteria were applied to the full cMD3 data set using RStudio to determine which studies were to be included in our final subset: (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, PA, or (v) the study was HMP_2012, which is an important study for healthy controls from the Human Microbiome Project.

Our goal is to gather retrospective, multicohort study of metagenomic human gut microbiome data to generate a supervised machine learning (ML) model for the detection of CRC progression. This machine learning model will act as a proof-of-concept precursor to a diagnostic tool that can ease the burden of colonoscopy suites for preventative screening, giving a way for doctors to triage patients for effective backlog reduction. It will also guide future research on the influence of microbiomes on disease states.

For our exploratory data analysis, we plan to focus on the following:

I. **Data availability** - cMD3 returns a TreeSummarizedExperiment object with relative_abundance, gene_families, marker_abundance and presence, and pathway_abundance and coverage data. How much data is available from each study, and how does that affect our ability to perform our analyses and modeling?

II. **Basic metadata analysis** - All included samples contain information regarding age and gender, which is important considering they are some of the factors in CRC screening and preventative care (Cancer Over Time, n.d.; Duan et al., 2022; Siegel et al., 2025). Body mass index (BMI) and age are also known to influence each other, thus it was decided to check if there would be any clustering or relationships. Since BMI has been previously associated with microbial shifts (Li et al., 2025), it should be heavily considered when modelling. Geographical location and consequently diet are also well established to be linked to CRC development, and it should be noted that the three most affected countries by CRC are Japan, the United States, and China (*Colorectal Cancer Statistics*, n.d.; *Worldwide Cancer Data*, n.d.; Dakal et al., 2025; T. Li et al., 2025; Singh et al., 2017). The following analyses were conducted:

    A. Normality of age and BMI are analyzed with summary statistics, histograms, and box plots. Analysis of Variance (ANOVA) is used to compare the mean across

Abbreviations: 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").

CRC progression classifiers. These will be performed at a 95% confidence level, and with a large sample size, the central limit theorem allows for an ANOVA to be performed when there is slight skew. When significance arises, Tukey's Honest Significant Difference (HSD) test is used to determine which pairings are significantly different from each other.

B. Gender proportions are compared across age by decade via a faceted stacked bar chart, with a binomial test to determine if the proportion of gender is significantly different in our sample.

C. Linear regression models were conducted for each CRC progression classifier against age and BMI to check for potential relationships, which are also represented by scatter plots. These are age-centered to place the results into biological relevance and compared to "HC".

III. **Taxonomic and pathway analysis** - using a combination of PERMANOVA and Linear Discriminant Analyses (LDA) for initial identification of any taxonomic or pathway trends in the data.

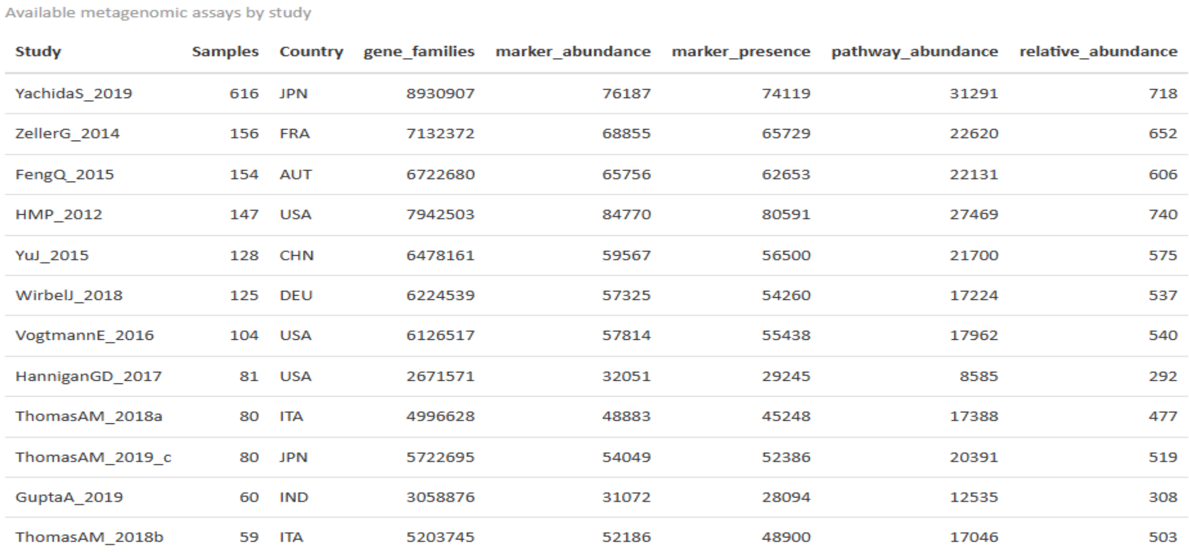
IV. **Data Limitations** - For investigation into the data limitations that will be encountered during next steps in our model creation allowing further narrowing of scope.

Our Exploratory Data Analysis provides the foundation for understanding the cMD3 dataset and our filtered data. Through our EDA we can assess the quality of the data and key characteristics to refine the scope of our project and future modeling.

Exploratory Data Analysis

Data Availability

Table 1 summarizes the data available across the CRC-relevant studies included in this analysis after filtering. Each study contains between 59 and 616 samples, with the largest cohort being YachidaS_2019 and the smallest being ThomasAM_2018b.



Available metagenomic assays by study

Study	Samples	Country	gene_families	marker_abundance	marker_presence	pathway_abundance	relative_abundance
YachidaS_2019	616	JPN	8930907	76187	74119	31291	718
ZellerG_2014	156	FRA	7132372	68855	65729	22620	652
FengQ_2015	154	AUT	6722680	65756	62653	22131	606
HMP_2012	147	USA	7942503	84770	80591	27469	740
YuJ_2015	128	CHN	6478161	59567	56500	21700	575
WirbelJ_2018	125	DEU	6224539	57325	54260	17224	537
VogtmannE_2016	104	USA	6126517	57814	55438	17962	540
HanniganGD_2017	81	USA	2671571	32051	29245	8585	292
ThomasAM_2018a	80	ITA	4996628	48883	45248	17388	477
ThomasAM_2019_c	80	JPN	5722695	54049	52386	20391	519
GuptaA_2019	60	IND	3058876	31072	28094	12535	308
ThomasAM_2018b	59	ITA	5203745	52186	48900	17046	503

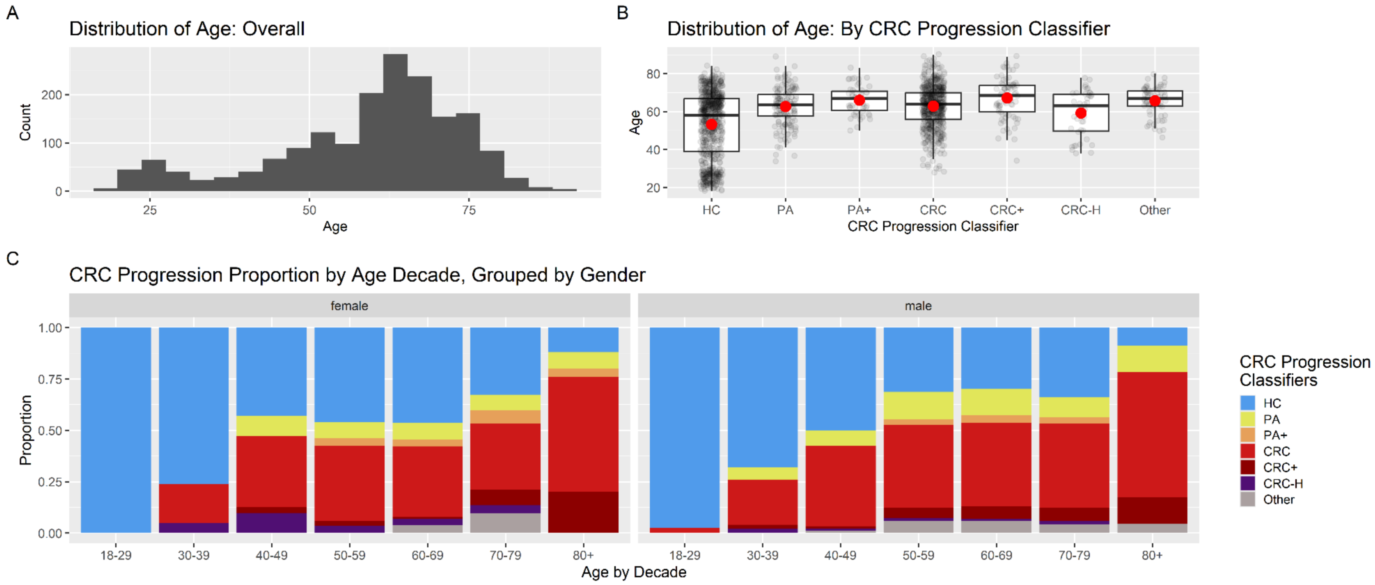
The studies span multiple geographic regions across Asia, Europe and North America, providing a diverse representation of populations relevant to colorectal cancer research.

All studies contain taxonomic relative abundance profiles, with between 292 and 740 taxa per study, allowing for cross-study comparisons of microbial composition. Functional metagenomic information is also available in the form of gene family, marker gene, and pathway abundance tables, although these datasets vary substantially in size with each value in these columns representing the available amount of features per study, reflecting the high dimensionality of functional genomic profiling from shotgun metagenomic sequencing. This high dimensionality also requires additional exploration and filtering before modelling, and increases computational demands. Variation in geographic representation also introduces heterogeneity that must also be considered in later analyses.

Metadata Analysis

The age summary statistics in Appendix 1-A show a mean of 59, standard deviation of 14.75, and median of 63, suggesting a skew. When faceted by gender, aside from a binomial test showing statistical difference in proportion (41.1% female, p = 5.821e-14, Appendix 1-B and 3-A), the median, mean, and quartiles are not remarkably different. A histogram and box plot (Figure 1-A and 1-B) were generated to corroborate this information (Q-Q plot available, Appendix 2-A). While age has a slight left-tailed skew with a bimodal shape, seemingly caused by the healthy controls visible in the box plot, it is not to an extreme that would harm our final model detrimentally. Skew and variance of age was checked between the sexes via a faceted histogram, which can be found in Appendix 2-B. An ANOVA found a statistical difference in the mean ages of the groups complementary to the box plot (95% CI, p = <2e-16, Appendix 3-B), requiring a follow-up Tukey's HSD test to determine which pairings were significantly different from each other. The following pairings with "HC" were found to be statistically different: "Other", "PA", "PA+", "CRC", and "CRC+" (all p-values <0.000, Appendix 3-B). All other differences in age between the CRC progression classifiers showed no statistical significance at a 95% confidence interval.

Figure 1: Age Normality & CRC Variation across Age and Gender



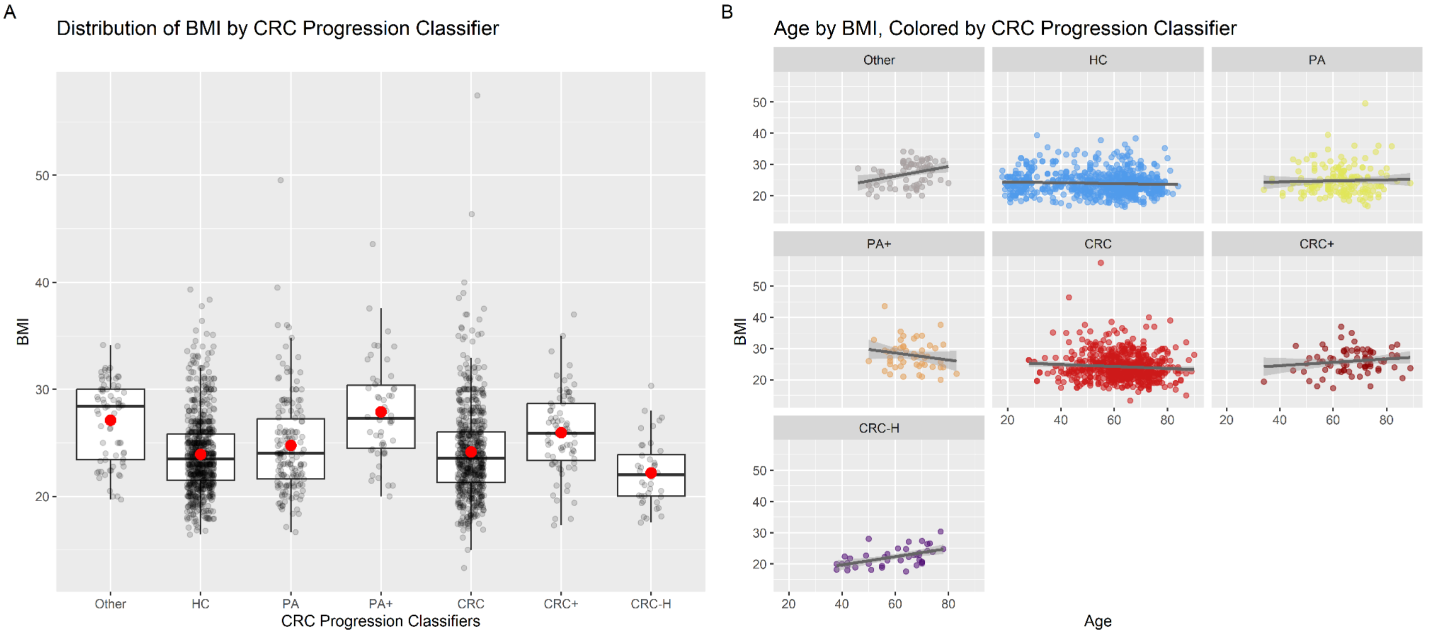
(A) Histogram demonstrating a bimodal age distribution with a secondary peak near 25 years and a primary peak near 60 years. (B) Box plot of age across CRC progression classifiers, with jittered transluscent points to demonstrate spread of the sample. The red dot represents the mean, the line represents the median. (C) Bar plot displaying the gender faceted proportion of CRC progression classifiers across each decade. Notable differences include and visual increase in HC from 18-69 for females, and increased PA and CRC/CRC+ across most decades in males.

After checking normality of the data, the age was divided into decades, with those 18 and 19 years old included in the 20-29 group to avoid a decade being visualized with only two years represented, and all those 80 years or older being grouped together as there was only one individual above 89 years. The two least represented decades in this study will be the 80+ and 30-39 decades (2.7% and 4.0%, respectively) and the three most represented being 60-69, 70-79, and 50-59 (34.4%, 21.8%, 19.9%, respectively; Appendix 1-C). Figure 1-C represents this data in proportions for easier comparisons across the decades, faceted by gender, as it is well established that males have a higher relative risk of CRC than females (Colorectal Cancer Statistics, n.d.; Duan et al., 2022). This pattern is reflected across most age groups in our data, particularly from 30-79. Furthermore, visually there are less samples with recorded comorbidities, as well as less "PA" compared to "CRC" and "HC", overall.

The three most affected countries by CRC are well represented in this multi-cohort study (Appendix 1-D and 2-E), and within these three target countries, there is representation from 6 of our classifiers: "Other", "HC", "PA", "CRC", "CRC+", and "CRC-H". The only unrepresented classifier is "PA+".

Figure 2-A shows the distribution of BMI among the CRC progression classifiers; while there were a few outliers, particularly in the upper BMI bounds of the "PA", "PA+", and "CRC", there are visually striking differences among the means. An ANOVA (Appendix 3-C) showed statistical significance (95% confidence, p = <2e-16), warranting a Tukey's HSD for further investigation. Of 21 pairings, 13 were found to be statistically significant to a 95% confidence level, meaning BMI could have an impact. It appears to have the most impact on "CRC-H" (5 pairings), "Other" (4 pairings), and "CRC" (4 pairings). BMI and age were also plotted against one another, shown in Figure 2-B with the associated linear models for each CRC progression classifier represented on the scatter plots (Appendix 3-D). Results of the age-centered, linear regression modelling shows a very low R-squared value of 0.0707, meaning that BMI and age only explains about 7% of the variance, though there are a few relationships that show

Figure 2: BMI Associations with CRC Progression Classifiers



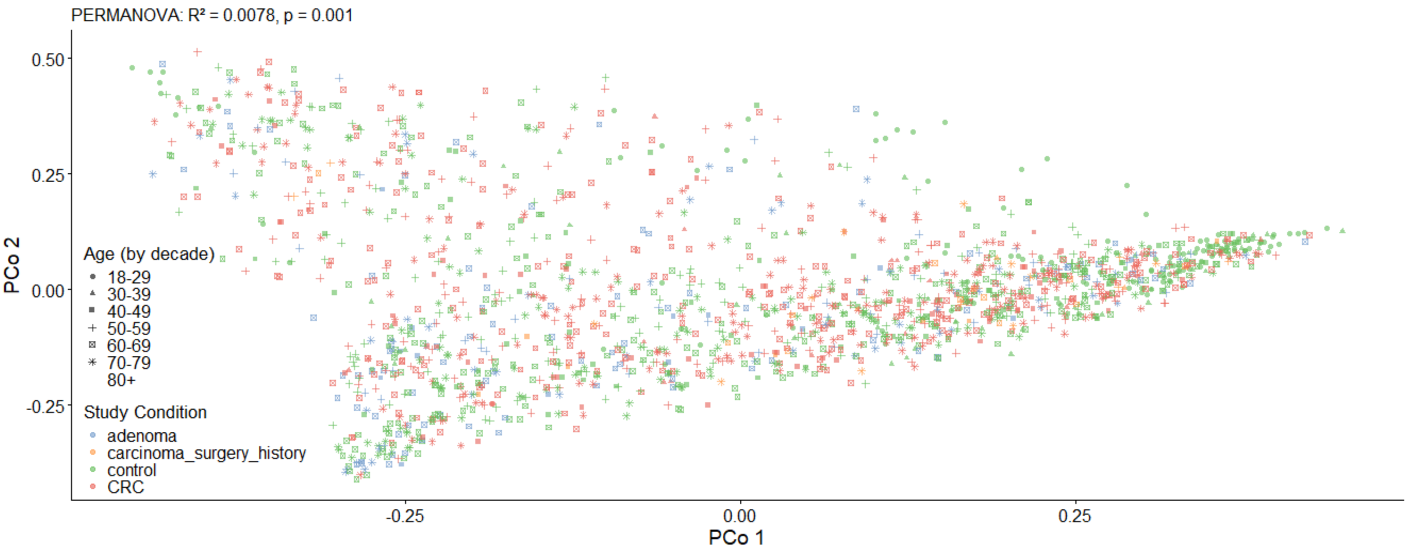
(A) Box plot of BMI by CRC progression classifier. The red dot marks the mean, while the black line marks the median. All classifiers generally have a normal distribution. (B) Scatter plots of BMI, with a linear model regression line of y ~ x, faceted by CRC progression classifier. Notice that there is minimally notable clustering.

statistical significance to a 95% confidence level, such as in "CRC-H" and "Other". With such a low explanation of the variance, it would be interesting to see how this may or may not extend towards the taxa and pathway relationships in the gut microbiome.

Taxonomic & Pathway Analysis

The combined filtered data sets from cMD3 contain over 1000 taxa. To identify any trends in the data, we looked at beta-diversity, which is a test for dissimilarity between the study conditions. Although the p-value was highly significant (P = 0.001), the R-squared value (R^2 = 0.0078) indicates much of the dissimilarity is not explained by the study conditions.

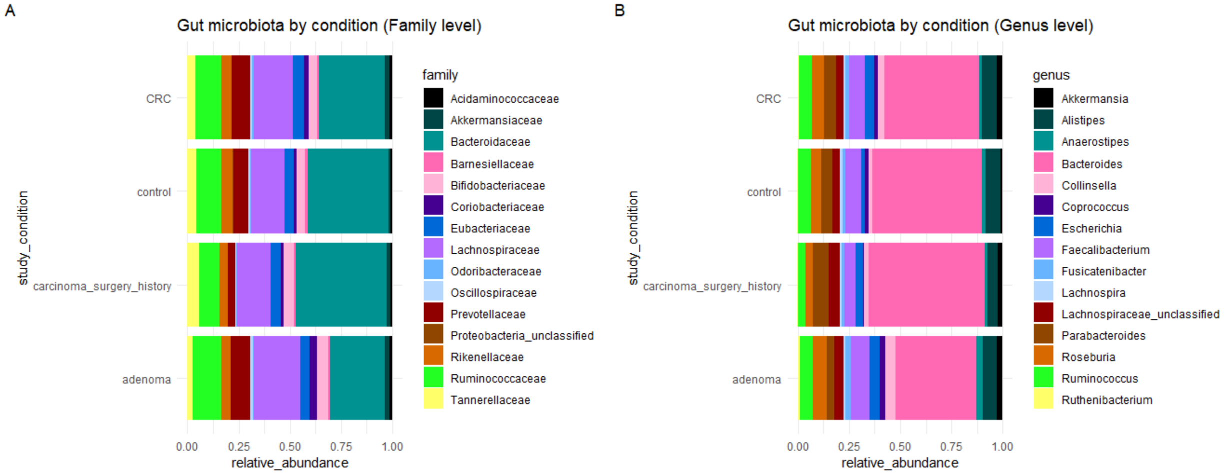
Figure 3: Beta Diversity of Samples by Age and Study Condition



Principle Coordinate Analysis illustrating Bray-Curtis dissimilarity of metagenomic data from cMD3 study. Comparisons were made between study conditions and age groups.

To understand how the taxa are distributed across the different study conditions,

Figure 4: Relative Abundance of cMD3 Study Metagenomic Data at Family and Genus taxonomic levels

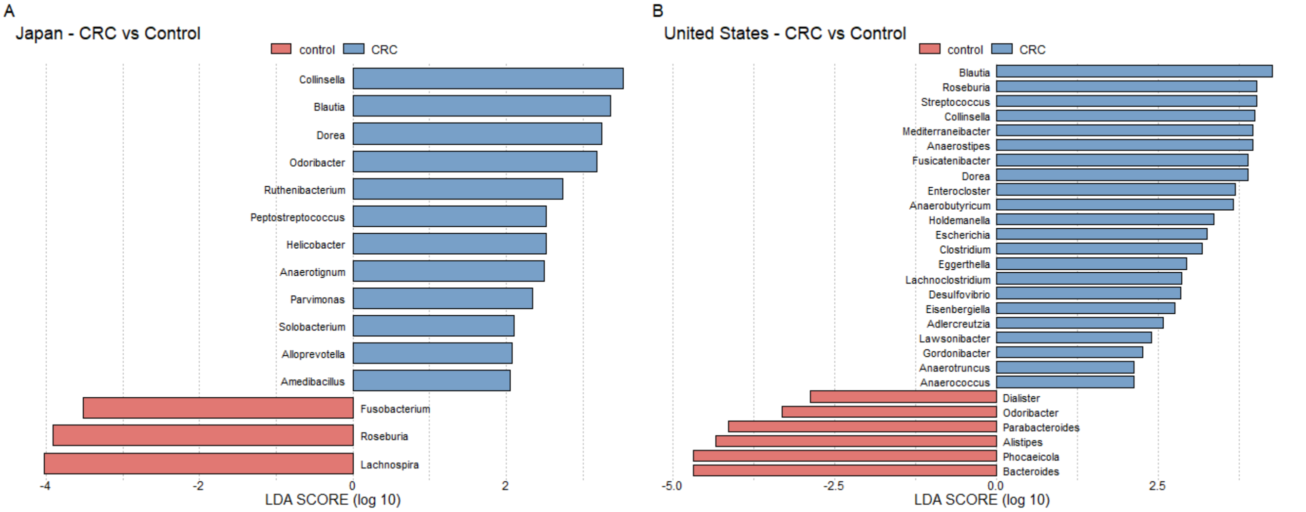


Stacked bar plots of cMD3 metagenomic data based on study conditions of filtered studies (A) Family taxonomic level (B) Genus taxonomic level.

two relative abundance plots were generated at the family and genus level.

The relative abundance stacked bar plots also indicate there was very little difference at the family and genus taxonomic levels. Considering these data sets cover several countries and age categories, it is possible that further separation into more groups may provide better comparisons. As Japan and the United States contain the greatest number of participants, these countries were separated out. Then, for the same reason, we compared only the control and CRC groups, since these health conditions also have the highest sample sizes. To identify key differences in how taxa may be enriched between groups, differential abundance was looked at using LDA Effect Size (LEfSe) analysis. Using a reference (control) and a treatment (CRC), LEfSe performs Kruskal-Wallis test to identify taxa that were statistically different from zero between the control and CRCgroups. Taxa found to be statistically significant were run with a Wilconox Rank-Sum Test to perform pairwise comparisons. Finally, the effect size is calculated using LDA.

Figure 5: Differential Abundance of cMD3 Study Metagenomic Data at Genus taxonomic level for Japan and the United States



Differential abundance of cMD3 Study metagenomic data using LDA Effect Size (LEfSe) comparing healthy controls and subjects with colorectal cancer (CRC) in (A) Japan and the (B) United States.

Three genera (*Blautia*, *Collinsella*, and *Dorea*) were found to have greater abundance in CRCsamples compared to controls. This warrants further investigation and may help guide the taxonomic portion of the machine learning model.

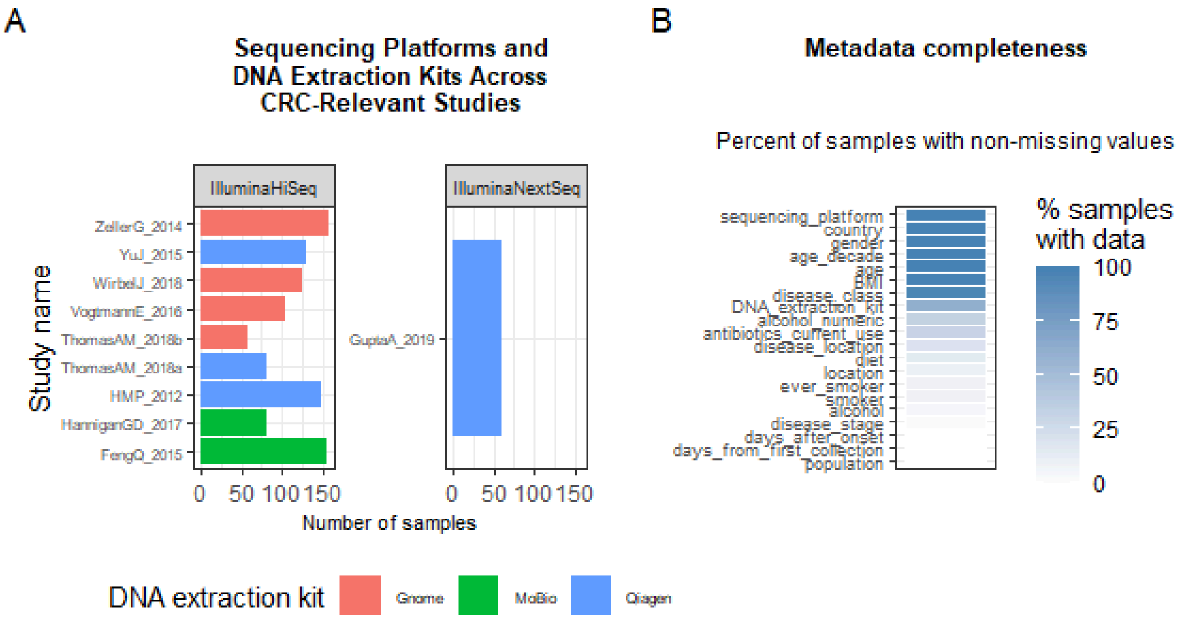
Limitations

While this work assesses the feasibility of leveraging human gut microbiome “fingerprints” for disease-state characterization, limitations do still exist from data availability, metadata completeness, and necessary inclusion criteria. All of which are not unique to this study but rather showcase the current challenges amongst large scale microbiome meta-analyses.

A rather major limitation is the incomplete availability of clinically and biologically relevant metadata across the various studies in our analysis. The metadata completeness analysis shown below visualises this. There is only a subset of variables that are consistently available throughout the data such as the sequencing platform use, participant country, gender, age and age by decade, BMI, and disease class. This is unfortunate because it is known that the microbiome is drastically influenced by variables that are not consistent such as antibiotic use, smoking status, alcohol consumption, and disease stages (Tegegne & Savidge, 2025). The original use case was intended to be applied to antibiotic usage which will unfortunately need to be pushed until more complete data sets are available.

While this study focuses on CRC metagenomic analysis/"fingerprinting", the available data has either empty or missing columns for things like tumor stage, location, and metastasis. Original plans had the separation of CRC patients with and without metastasis which is currently not feasible until more complete data is available. This will result in a computational framework that can only compare broad disease categories rather than detailed cancer stages.

Figure 6: Study variations and Data limitations



Other limitations came about after the filtering of the dataset to keep data consistent and comparable. Filters included only adult patients, stool samples only, disease labels, and CRC relevant studies. While the data quality is better using these filters, it does shrink the dataset. The filtering created a working database with mostly CRC samples where polyp-only and adenoma cases are few and far between. Due to this, disease progression from polyp to cancer modeling will not be reliable. Again this leads to a focus on broad comparisons such as healthy participants vs early lesions vs CRC.

As this is an aggregated data set, there are other considerations to factor in, specifically the difference in laboratory techniques, sequencing machines, extraction kits, and protocols used throughout all of the studies. These variables are documented but they cannot be fully corrected for so they must be mentioned as they create biological noise in an exploratory modeling approach such as this.

Summary

Exploration that occurred during this phase of our study revealed that after filtering for our inclusionary criteria, the 12 studies that would remain in our multi-cohort retrospective study contains taxonomic and pathway data required for modelling purposes. We successfully audited the taxonomic and metadata, revealing both anticipated and surprising patterns. While biometric and demographic data such as age, BMI, and gender appear to have some influence over the CRC progression classifiers, the predictive value is not as substantial as we may have assumed given prior research. Diet has also been shown to influence the human microbiome, however it is not often collected as a data point in studies, which is reflected in this data set, with only 276 data points containing information regarding diet. Approximately 98.4% classified as omnivores and the remainder (10 samples) representing vegetarians, which may not be enough data points to investigate variability in diet as a factor in CRC progression. This may over inflate proportions of different stages of the disease, like in various visualizations attempted (Appendix 2-C).

These initial analyses also revealed several imbalances. For example, countries are highly represented in our data (Figure 6-B), though 38.9% of our study belongs to Japan. The three most affected countries by CRC are well represented in this multi-cohort study (Figure 2-A, Appendix 1-D), which is important to ensure that our results and model will be applicable to the most vulnerable target populations. However, three continents are missing: Africa, Oceania, and South America. This will need to be considered in future evaluations of the model, with particular attention when searching for data to include in training sets.

Proposal of this analytical framework is intentionally framed as a proof-of-concept tool rather than a ready to use clinical diagnostic model to show the idea works in principle. Biggest limitations currently include missing or inconsistent metadata, not from the modeling itself. This is to say that when future datasets with more consistent data become available during further version releases of the cMD or more additions via the pipelines and framework provided by Manghi et al. (2025) the proof-of-concept tool can handle expansions.

Due to the bimodality in age caused by the healthy controls, if there is no real difference in microbial fingerprints in the healthy controls when the inclusion criteria is changed from 18+ to 30+, proposed computational modeling may benefit from selection of only participants 30+. In conjunction with various limitations in the selected data which also became apparent leading to the need to pivot with select variables such as exclusion of diet focus, disease stage, antibiotic use, and CRC with individual comorbidities analyses. Future analysis will focus on using Random Forest models as they are robust to this to unbalanced data (Bradter et al., 2022).

Unfortunately, due to the computing power of our personal computers, we were unable to dig into the pathway data as initial implementation failed. RAM requirements will necessitate a high performance computing cluster like that of Sol at Arizona State University (Jennewein et al., 2023), and thus gaining joint access as a team will be a high priority for completing this analysis and for further modelling needs.

References

Bradter, U., Altringham, J. D., Kunin, W. E., Thom, T. J., O'Connell, J., & Benton, T. G. (2022). Variable ranking and selection with random forest for unbalanced data. *Environmental Data Science*, *1*, e30. https://doi.org/10.1017/eds.2022.34

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/

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

Drnevich, J., Tan, F. J., Almeida-Silva, F., Castelo, R., Culhane, A. C., Davis, S., Doyle, M. A., Geistlinger, L., Ghazi, A. R., Holmes, S., Lahti, L., Mahmoud, A., Nishida, K., Ramos, M., Rue-Albrecht, K., Shih, D. J. H., Gatto, L., & Soneson, C. (2025). Learning and teaching biological data science in the Bioconductor community. *PLOS Computational Biology*, *21*(4), e1012925. https://doi.org/10.1371/journal.pcbi.1012925

Jennewein, D. M., Lee, J., Kurtz, C., Dizon, W., Shaeffer, I., Chapman, A., Chiquete, A., Burks, J., Carlson, A., Mason, N., Kobawala, A., Jagadeesan, T., Basani, P. B., Battelle, T., Belshe, R., McCaffrey, D., Brazil, M., Inumella, C., Kuznia, K., … Yalim, J. (2023). The Sol Supercomputer at Arizona State

University. *Practice and Experience in Advanced Research Computing*, 296–301. https://doi.org/10.1145/3569951.3597573

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

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

Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., Beghini, F., Malik, F., Ramos, M., Dowd, J. B., Huttenhower, C., Morgan, M., Segata, N., & Waldron, L. (2017). Accessible, curated metagenomic data through ExperimentHub. *Nature Methods*, *14*(11), 1023–1024. https://doi.org/10.1038/nmeth.4468

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

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

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/