Exploratory Data Analysis

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

Kaitlyn Schisler Cory Spern Faith Shipman Jacob Anderson

Delete Before Submitting

Assignment Guidelines from Canvas

Step 1: EDA Objectives & Preparing data

- define at least 3 EDA objectives to investigate (necessary variables? distributions conducive to modelling? data quality issues? is data suitable? representative of population?)
- There's a document available about what's good to do for EDA on canvas that we can look at if needed

Step 2: Conduct EDA

- Plots/visualizations, summary stats, tables
- document findings clearly
- Code, additional plots & tables that aren't the main talking points can be included in an appendix section.

Step 3: EDA Report (Single spaced, 12 pt. font)

- **Introduction to Data (1-2 page)**
  Provide an overview of your dataset, including its source (with relevant links), purpose, and your rationale for selecting it. Clearly state your EDA objectives and briefly describe the EDA methods used. Explain the importance of EDA in the context of your overall project.
  - Pull from methods section from proposal/revised proposal and the project statement, good depth there on selection of the data and why we chose the data/methods.

- **Exploratory Data Analysis (4-5 pages)**
  This is the core of your report. Present your findings using relevant summary statistics, tables, and visualizations. Ensure each EDA objective is addressed clearly and supported by appropriate evidence from your analysis.
  - Flow of this section will be important. We can play with the flow once everyone's section is complete.
- **Summary (1-2 pages)**
  Summarize your key findings and highlight the main takeaways from your EDA. Describe the next steps for your project, including how the insights gained from EDA will inform your choice of data modeling techniques for the main analysis. Briefly outline which techniques you plan to implement next and why.

General Notes

- Potential flow of our EDA section: Data Availability & Selection → Metadata Analysis → Taxa & Pathway Analysis → Data Limitations
  - Data Availability
    - What does the whole cMD3 offer in its standardized data, including how Bioconductor data works (`summarizedExperiments` vs `TreeSummarizedExperiments`)?
    - What data remains, what could be answered, what can still be explored?
    - ***Our original plan was to include metastases, which our data set has none so that is no longer possible***, so that could be mentioned here in this EDA section.
  - Metadata EDA
    - The factors we mentioned in our proposals (ie age, sex, geography/ethnicity, diet, antibiotics, etc)
      - What has too little to give real meaning behind trends we see?
    - For geography, the top 3 affected countries by CRC are represented in our data set, which is good. We should note that we are missing representations from certain continents
  - Taxa vs pathways
    - what we can do now (personal machines) vs what we need to do later on the HPC cluster (Sol)
  - Data Limitations

- ∗ What gets lost once we filter it with our exclusionary criteria?
- ∗ What would be unable to answer with what we have, and what gaps would remain to be filled?

Introduction

Glossary

Diagnostic Categories

Abbreviation	Description
HC	Healthy control
PA	Polyp / adenoma
PA+	Polyp / adenoma with comorbidities
CRC	Colorectal cancer
CRC+	Colorectal cancer with comorbidities
CRC-H	History of colorectal cancer with resection
Other	Any diagnosed disease without current or history of CRC-related growths

Exploratory Data Analysis

Data Availability (Jake)

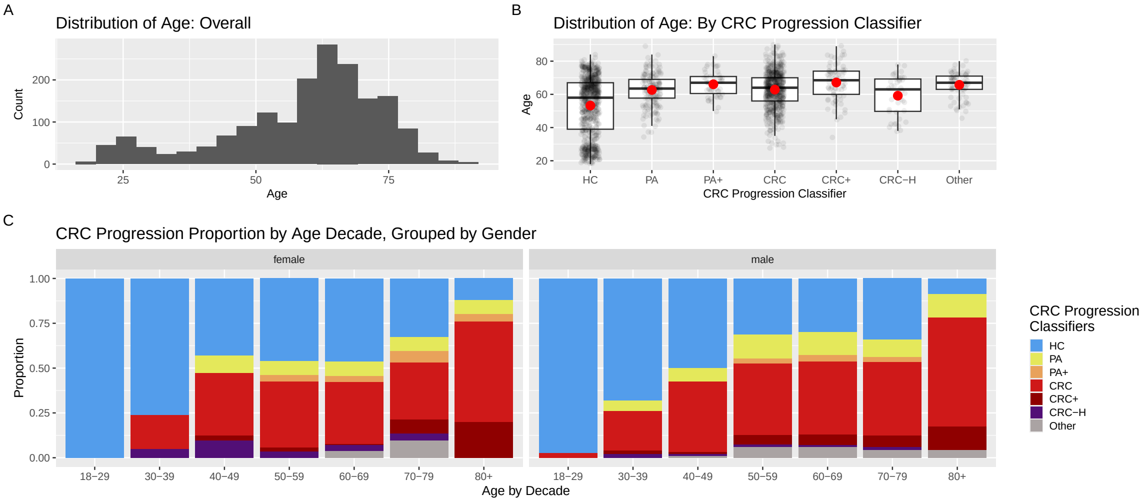
Metadata Analysis

All samples included contain information regarding the 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). Summary statistics were conducted (Appendix 1-A), showing a mean age of 59, standard deviation of 14.75, and median of 63, suggesting a skew in this continuous variable. When faceted by gender (Appendix 1-B), aside from the proportion of males and females being statistically different from each other (41.1% female, p = 5.821e-14, Appendix 1-B and 3-A), the median, mean, and quartiles are not remarkably different. A histogram (Figure 1-A) and box plot by CRC progression classifier

(Figure 1-B) were generated to corroborate this information, checking for extreme skewedness (supporting Q-Q plot available in Appendix 2-A). These show that while there is a slight left tailed skew in age overall with a bimodal pattern, seemingly caused by the healthy controls visible in the box plot, it is not to an extreme that would harm our final model detrimentally. Random Forest models are robust to unbalanced data (Bradter et al., 2022). Skew and variance of age was also checked between the sexes via a faceted histogram, which can be found in the supplementary visualizations (Appendix 2-B). Since the sample size is large, the central limit theorem allows for an ANOVA to be performed despite the slight skew, which was done to a 95% confidence level. This was to assess if there is a statistical difference in the mean age of the groups as a compliment to the box plot (Appendix 3-B). The p-value was <2e-16, and a follow-up Tukey's Honest Significant Difference (HSD) test was conducted to determine which pairings were significantly different from each other. The following pairings with "HC" were 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.

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.

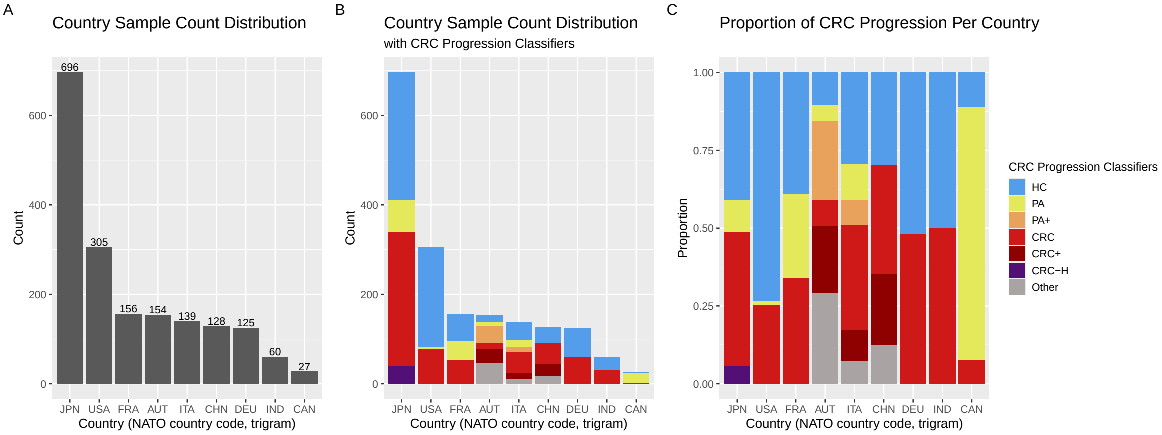
Figure 1: Age Normality & CRC Variation across Age and Gender



(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 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 pro 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.

Geographical location and consequently diet are also well established to be linked to CRC development (Dakal et al., 2025; Li et al., 2025; Singh et al., 2017). The three most affected countries by CRC—Japan, the United States, and China (Colorectal Cancer Statistics, n.d.; Worldwide Cancer Data, n.d.)—are well represented in this multi-cohort study at 38.9%, 17.0%, and 7.2% respectively (Figure 2-A, Appendix 1-D). This is important for our study to ensure that our results and model will be applicable to the most vulnerable target populations. Within these three target countries, there is representation from 6 of our classifiers, visible in Figures 2-B and 2-C: “Other”, “HC”, “PA”, “CRC”, “CRC+”, and “CRC-H”. The only unrepresented classifier is “PA+”. Country is often closely tied to diet, however it is often not 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, as it may over inflate proportions of different stages of the disease, like in various visualizations attempted (Appendix 2-C).

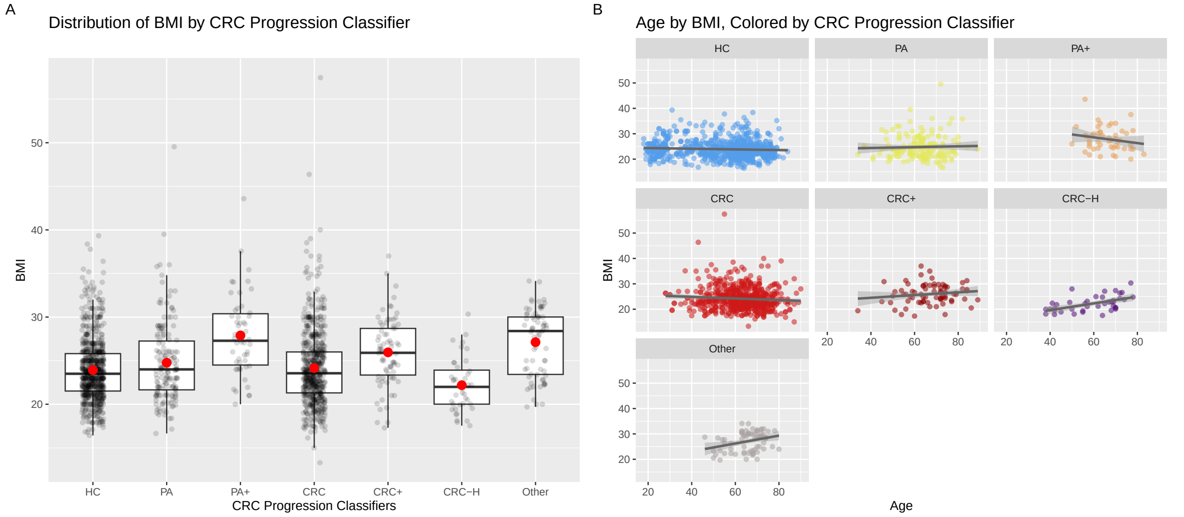
Figure 2: Country Associations with CRC Progression Classifiers



(A) Bar plot of country sample count, sorted in descending order. (B) Stacked bar plot of country sample count, sorted in descending order, filled with CRC progression classifiers. (C) Stacked bar plot of the CRC progression classifier proportion for each country in the multi−cohort study.

BMI has been previously associated with microbial shifts (Li et al., 2025) and should be heavily considered when modelling. We examined the distribution of the BMI among the CRC progression classifiers (Figure 3-A) and while there were a few outliers, particularly in the upper BMI bounds of the “PA”, “PA+”, and “CRC” classifiers, there are visually striking differences among the means of the groupings, particularly those with comorbidities. To confirm, an ANOVA was run (Appendix 3-C) which was found to be statistically significant (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 are also known to influence each other, thus it was decided to check if there would be any clustering or relationship visible when plotted against one another, shown in Figure 3-B. Associated linear models were conducted for each CRC progression classifier to check for potential relationships, which are also 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 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 taxonomic and pathway relationships in the gut microbiome.

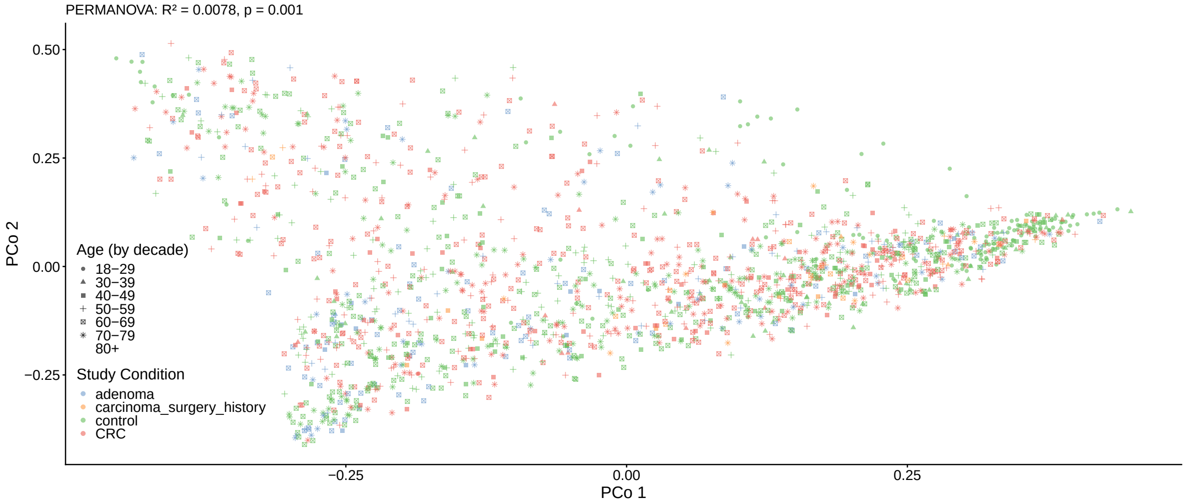
Figure 3: BMI Associations with CRC Progression Classifiers



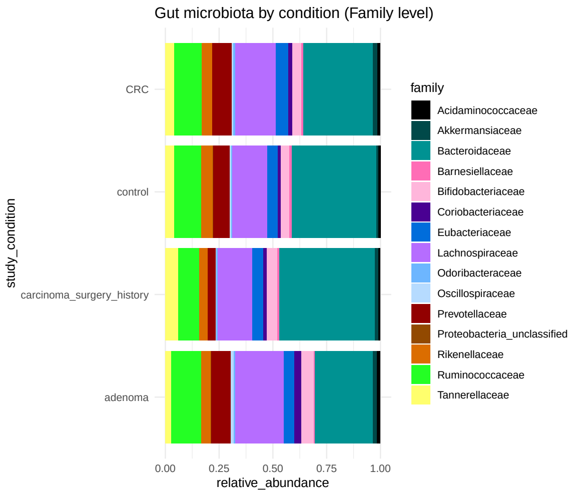
(A) Box plot of BMI by CRC progression classifier. The red dot marks the mean, while the black line marks the median. All classifiers generally have a normal distribution. (B) Scatter plots o BMI, with a linear model regression line of y ~ x, faceted by CRC progression classifier. Notice that there is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

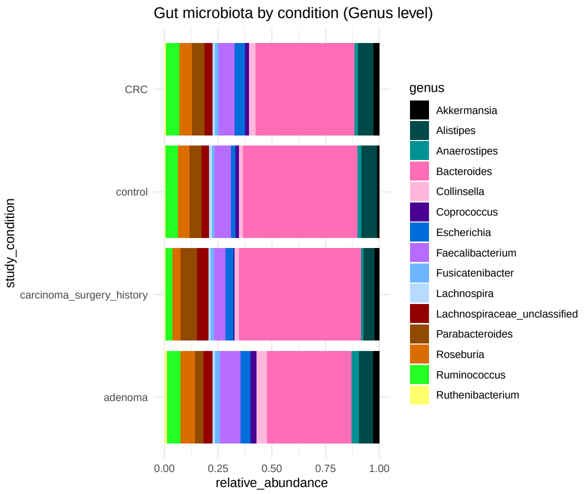
Taxonomic & Pathway Analysis

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.

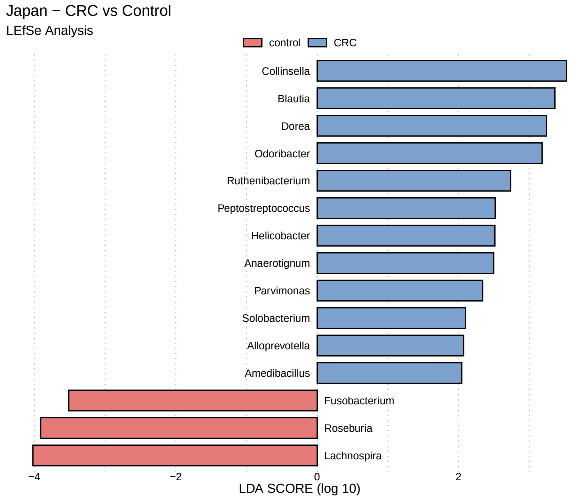


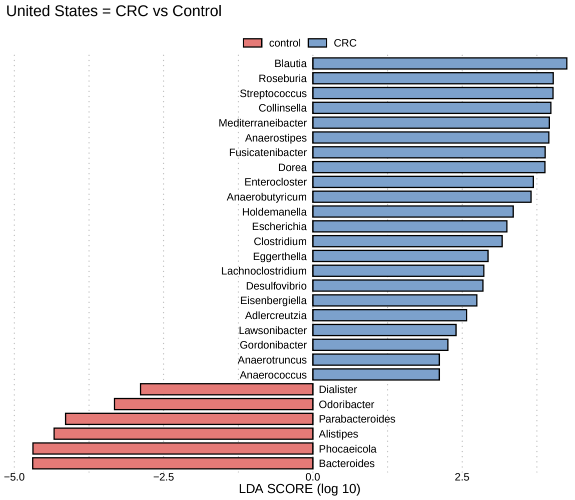
To understand how the taxa are distributed across the different study conditions, two relative abundance plots were generated at the family and genus level.





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 LEfSe analysis. Using a reference and a treatment, LEfSe performs Kruskal-Wallis test to identify taxa that were statistically different from zero between the control and CRC groups. 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 linear distriminant analysis (LDA).





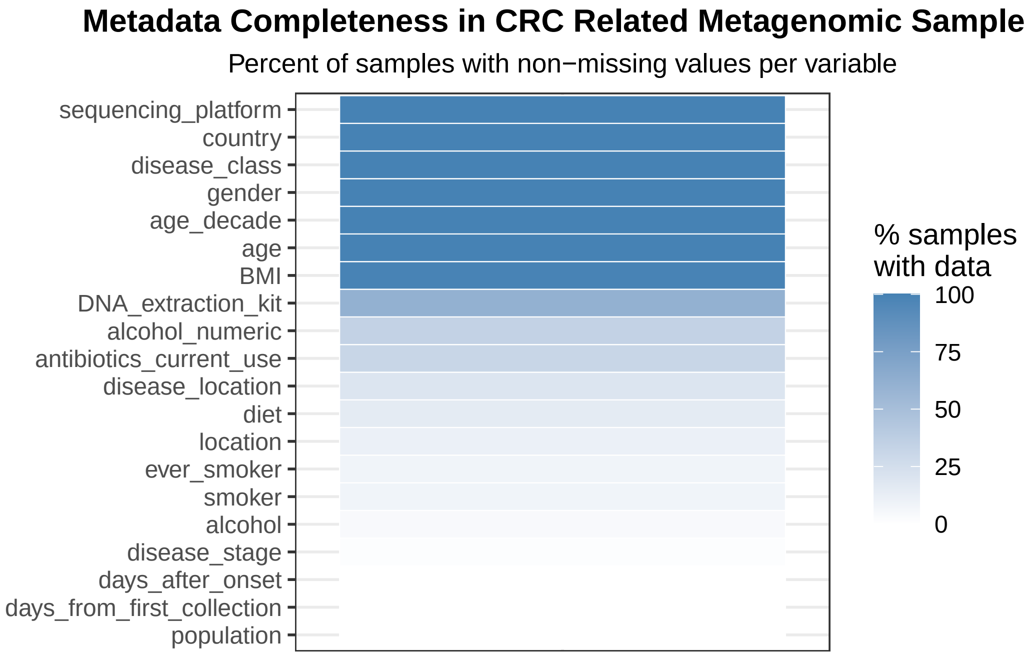
```
[1] "Dorea"       "Blautia"     "Collinsella"
```

Three genera (Blautia, Collinsella, and Dorea) were found to have greater abundance in CRC samples compared to controls. This warrants further investigation….

Data 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.

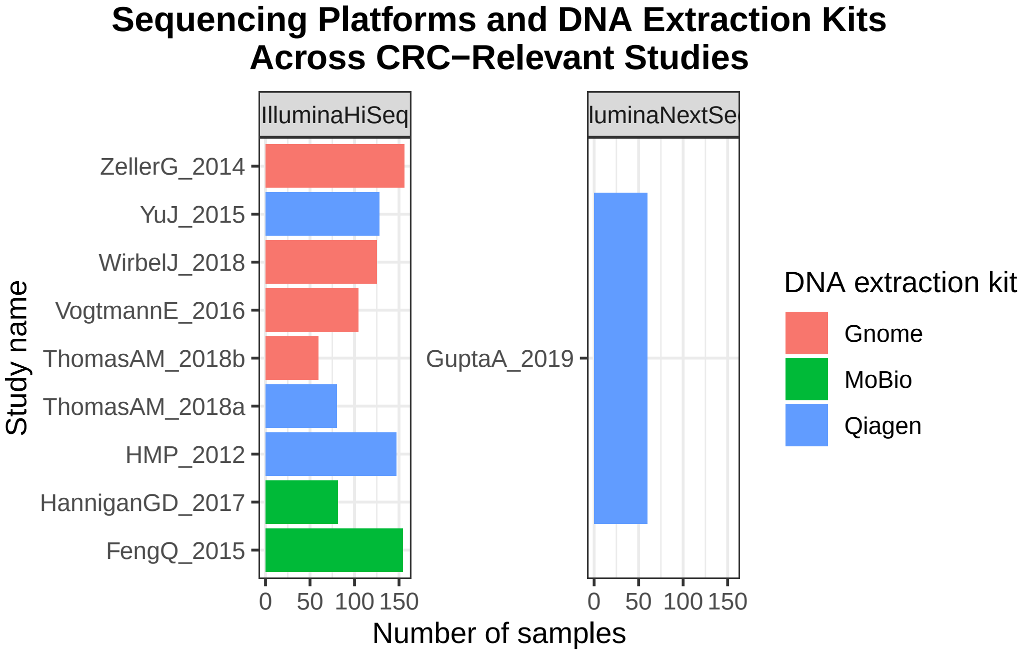
A rather major limitation is the incomplete availability of clinically and biologically relevant metadata across the various studies in our analysis. The metadata completeness analysis shown below visualises this. There is only a subset of variables that are consistently available throughout the data 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.

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 dataset is a compilation, 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



Proposal of this analytical framework is intentionally framed as a proof-of-concept tool rather than a ready to use clinically 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 the framework can handle expansions.

Conclusion

References