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

Student	Assignment
Cory	Country & Category
Kaitlyn	Age & Sex
	Data Availability Table (Show's what cMD has from each study)
Faith	Limitations (Missing Data, Pop size, etc.)
Jacob	Summary Stats, Glossary & Introduction

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

Yada Yada Yada

Diagnostic Categories

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

Exploratory Data Analysis

Data Availability

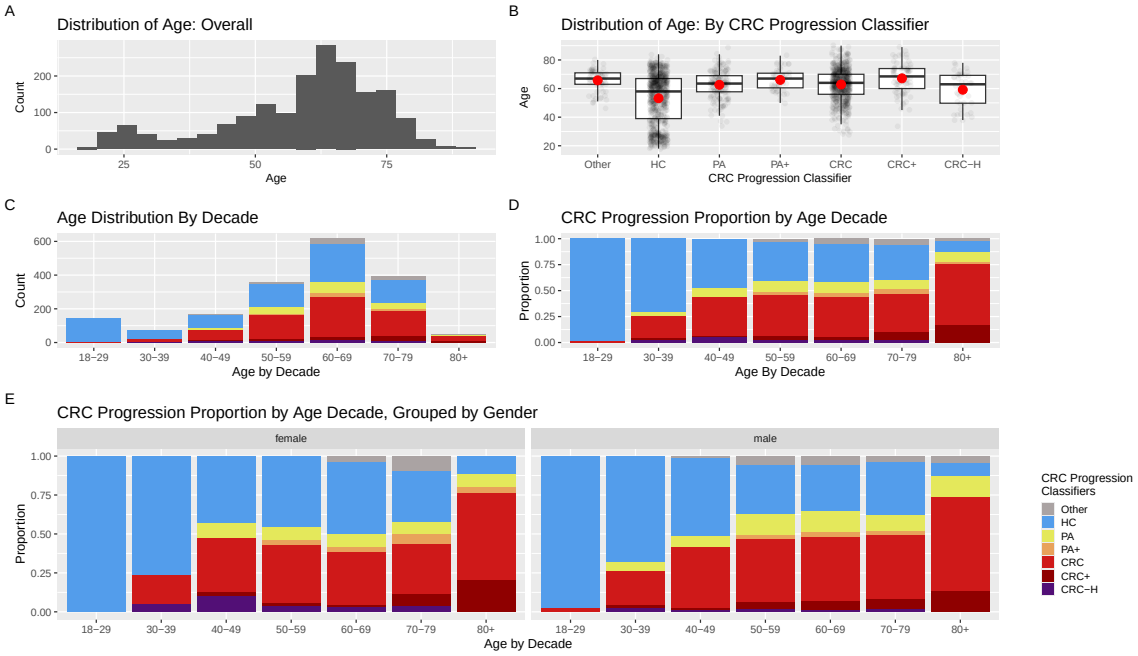
Metadata Analysis (Kaitlyn)

Of great benefit to having the cMD3 database is the standardization of the metadata, allowing for the analysis of potential confounding variables for biases and major discrepancies in the data that may influence the results and interpretations downstream. With prior evidence supporting age, gender, diet, geographical location, BMI, and drinking and smoking habits are all factors that may influence the relative risk of CRC (SOURCE), they are important variables to consider in our initial EDA, checking for bias in distribution and sources of significance.

Age, Gender, and CRC Progression Classifiers

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 (SOURCE). 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. 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 skew (SOURCE). 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-A). 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.

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

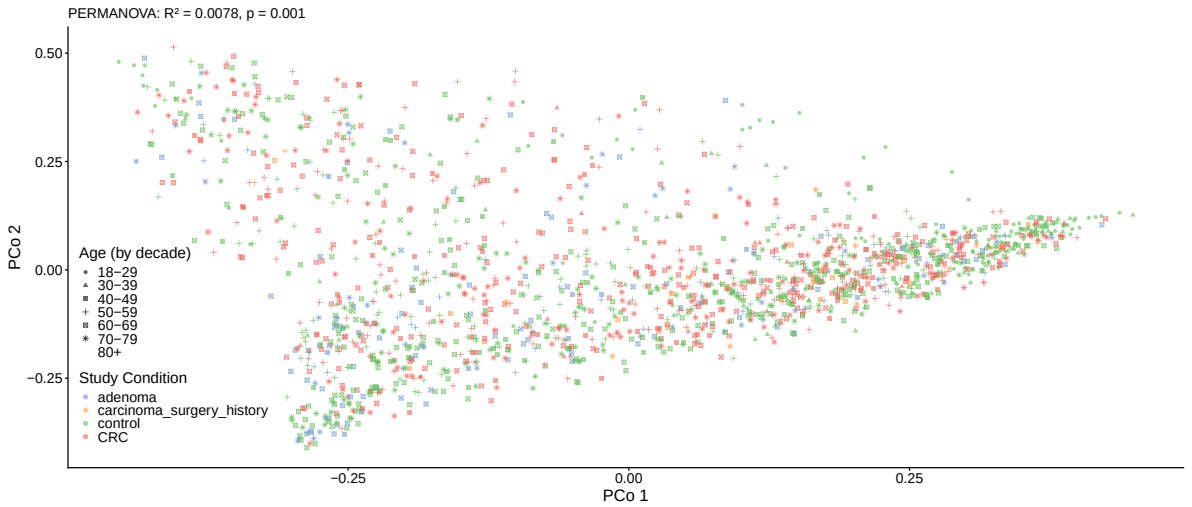


(A) Histogram demonstrating a bimodal age distribution with a secondary peak near 25 years and a primary peak near 60 years. (B) Box plot of age across CRC progression classifiers, with jittered transluscent points to demonstrate spread of the sample. This effectively displays the main contributor to the skew is coming from the HC group, with the dense cluster of points at approximately 25 years. The red dot represents the mean, the line represents the median. (C) Barplot of Age distribution by decade. Note that the first decade range includes 18−20 to avoid a bar of those only 18 and 19 years old. (D) Barplot displaying overall proportion of the CRC progression classifiers across each decade. (E) Barplot 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.

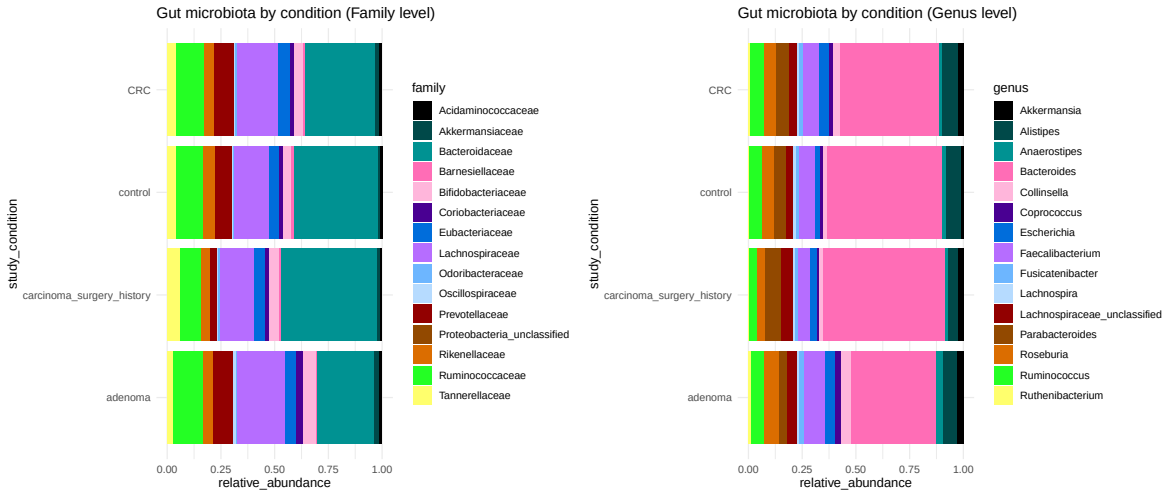
After checking normality of the data, the age was divided into decades, with those 18 and 19 years old included in the 20-29 group to avoid a decade being visualized with only two years represented, and all those 80 years or older being grouped together as there was only one individual above 89 years. The decade separations represented by Figure 1-C clearly shows 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-D and 1-E represent this data in proportions for easier comparisons across the decades, with 1-E faceted by gender. It is well established that males have a higher relative risk of CRC than females (SOURCE), which appears to be reflected across most age groups. Visually, there appears to be far less samples with recorded comorbidities, represented by the "plus", as well as relatively less "PA" compared to "CRC" and "HC" overall.

Taxa & Pathway Analysis (Cory)

The combined filtered datasets from curatedMetaGenomicData 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 ($R\hat{}2 = 0.0078$) indicates much of the disimilarity is not explained by the study conditions.

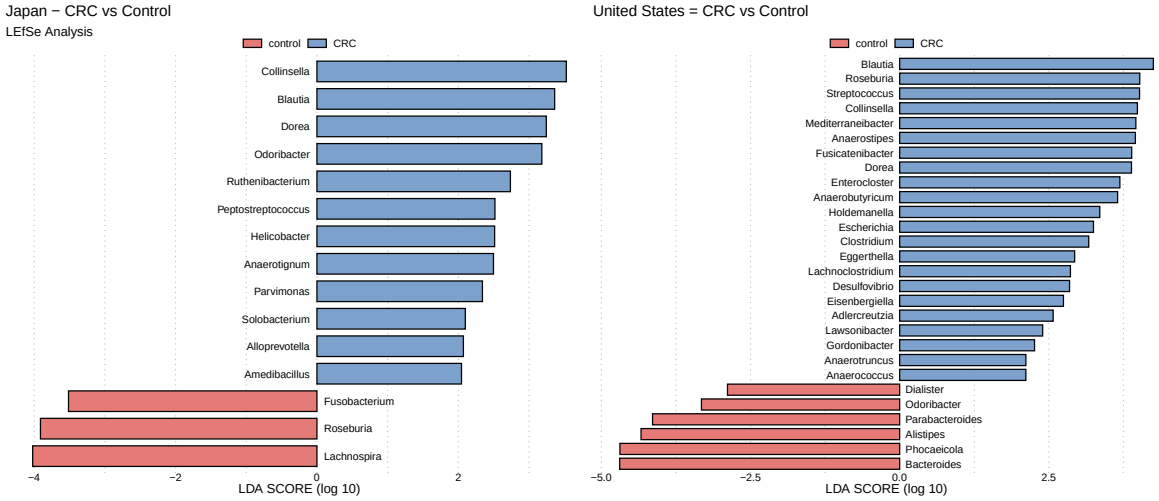


To understand how the taxa are distributed across the different study conditions, two relative abundance plots were generated at the family and genus level.



The Relative abundance stacked barplots also indicate there was very little difference at the family and genus taxonomic levels. Considering these datasets 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 (Bautia, Collinsella, and Dorea) were found to have greater abundance in CRC samples compared to controls. This warrants further investigation….

Data Limitations

Conclusion

References