Exploratory Data Analysis - Appendix

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

Kaitlyn Schisler Cory Spern Faith Shipman Jacob Anderson

Table of contents

Appendix 1 - Summary Tables

1-A: Age Summary Statistics

```
# Stats Setup
#############
age_overall <- qualifying_studies %>%
  # Summary Stats
  # -------------
  summarize(
    n = sum(!is.na(age)),
    mean = mean(age, na.rm = TRUE),
    sd = sd(age, na.rm = TRUE),
    median = median(age, na.rm = TRUE),
    Q1 = quantile(age, 0.25, na.rm = TRUE),
    Q3 = quantile(age, 0.75, na.rm = TRUE)
  )  %>%
  # Make Vertical Table
  # -------------------
  pivot_longer(
    everything(),
    names_to = "Statistic",
    values_to = "Value"
  ) %>%
  mutate(Statistic = recode(Statistic,
  n = "N",
  mean = "Mean",
  sd = "Standard Deviation",
  median = "Median",
  Q1 = "First Quartile (Q1)",
  Q3 = "Third Quartile (Q3)"
))

# Kable Table
#############
age_overall %>%
  kbl(
    col.names = c("Statistic", "Value"),
    caption = "Overall Age Summary",
    digits = 2,
    format = "latex"
```

Table 1: Overall Age Summary

Statistic	Value
N	1790.00
Mean	59.04
Standard Deviation	14.75
Median	63.00
First Quartile (Q1)	52.00
Third Quartile (Q3)	69.00

```
  ) %>%
  kable_styling(full_width = FALSE) %>%
  column_spec(1, bold = TRUE)
```

1-B: Age by Gender Summary Statistics

```
# ---

# Stats Setup
##############
age_x_gender <- qualifying_studies %>%
  group_by(gender) %>%
  summarize(
    n = n(),
    mean = mean(age, na.rm = TRUE),
    sd = sd(age, na.rm = TRUE),
    median = median(age, na.rm = TRUE),
    Q1 = quantile(age, 0.25, na.rm = TRUE),
    Q3 = quantile(age, 0.75, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  mutate(
    pct = 100 * n / sum(n),
    `N (%)` = sprintf("%d (%.1f%%)", n, pct)
  )

# Table Setup
##############
age_x_gender %>%
```

Table 2: Age Summary by Gender

Gender	N (%)	Mean	Standard Deviation	Median	Q1	Q3
female	736 (41.1%)	59.06	15.05	62	52	69
male	1054 (58.9%)	59.03	14.55	63	52	69

```
  select(gender, `N (%)`, mean, sd, median, Q1, Q3) %>%
  kbl(
    col.names = c("Gender", "N (%)", "Mean", "Standard Deviation",
                  "Median", "Q1", "Q3"),
      caption = "Age Summary by Gender", format = "latex", digits = 2) %>%
  kable_styling(full_width = F) %>%
  column_spec(1, bold = T)
```

1-C: Age by Decade Summary Statistics

```
# Stats Setup
##############
age_decade <- qualifying_studies %>%
  # Summary Stats
  # -------------
  group_by(age_decade) %>%
  summarize(
  n = n(),
  .groups = "drop") %>%
  mutate(
    pct = 100 * n / sum(n),
    `N (%)` = sprintf("%d (%.1f%%)", n, pct)
  )

# Kable Table
##############
age_decade %>%
  select(age_decade, `N (%)`) %>%
  kbl(
    col.names = c("Age (Decade)", "N (%)"),
      caption = "Age Summary by Decade", format = "latex", digits = 2) %>%
  kable_styling(full_width = F) %>%
  column_spec(1, bold = T)
```

Table 3: Age Summary by Decade

Age (Decade)	N (%)
18-29	144 (8.0%)
30-39	71 (4.0%)
40-49	164 (9.2%)
50-59	356 (19.9%)
60-69	616 (34.4%)
70-79	391 (21.8%)
80+	48 (2.7%)

1-D: Country Summary Statistics

```
# Count and Proportion
######################
country_overall <- qualifying_studies %>%
  # Summary Stats
  # -------------
  group_by(country) %>%
  summarize(
  n = n(),
  .groups = "drop") %>%
  mutate(
    pct = 100 * n / sum(n),
    `N (%)` = sprintf("%d (%.1f%%)", n, pct)
  ) %>%
  arrange('Country')

# Kable Table
#############
country_overall %>%
  select(country, `N (%)`) %>%
  kbl(
    col.names = c("Country", "N (%)"),
      caption = "Country Overall Summary (NATO country code, trigram)",
    format = "latex", digits = 2) %>%
  kable_styling(full_width = F) %>%
  column_spec(1, bold = T)
```

Table 4: Country Overall Summary (NATO country code, trigram)

Country	N (%)
AUT	154 (8.6%)
CAN	27 (1.5%)
CHN	128 (7.2%)
DEU	125 (7.0%)
FRA	156 (8.7%)
IND	60 (3.4%)
ITA	139 (7.8%)
JPN	696 (38.9%)
USA	305 (17.0%)

```
# Country Classifier Count
##########################
country_overall <- qualifying_studies %>%
  filter(!is.na(country)) %>%
  group_by(country, disease_class) %>%
  summarise(n = n(), .groups = "drop") %>%
  arrange(country, desc(n))

# Classifier Kable Table
########################
country_overall %>%
  kbl(col.names = c("Country", "CRC Classifier", "Count"),
      caption = "Country CRC Progression Summary (NATO country code, trigram)",
      format = "latex") %>%
  kable_styling(full_width = F) %>%
  column_spec(1, bold = T) %>%
  collapse_rows(columns = 1, valign = "top")
```

Appendix 2 - Supplemental Visualizations

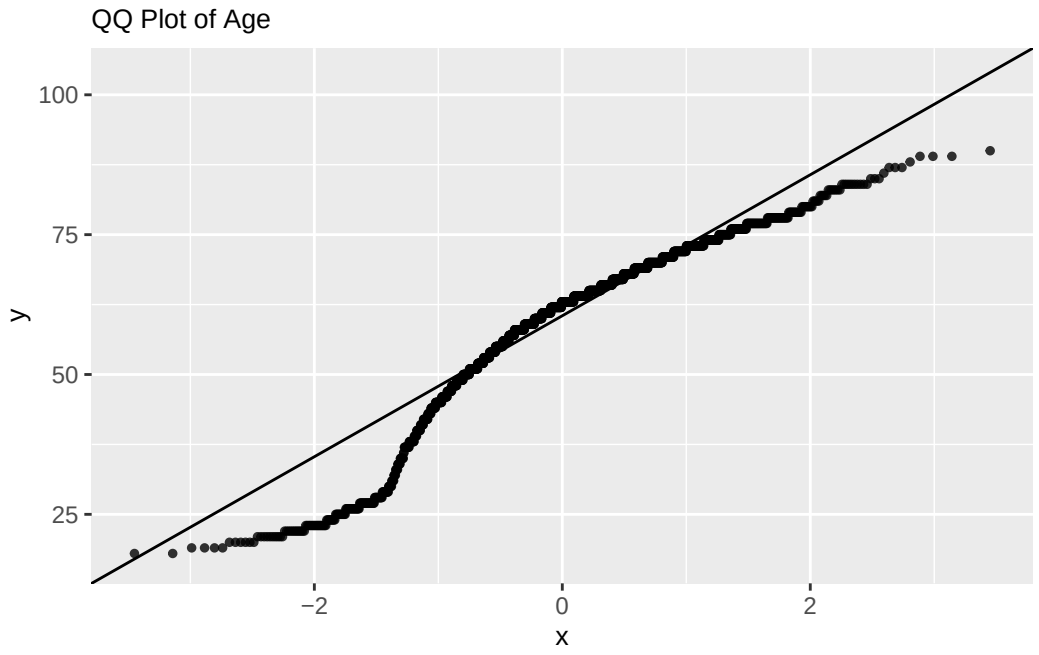
2-A: Age QQ-Plot (Supports Histogram)

```
# QQPlot
########
qualifying_studies %>%
  # Set Up
```

Table 5: Country CRC Progression Summary (NATO country code, trigram)

Country	CRC Classifier	Count
AUT	Other	45
	PA+	39
	CRC+	33
	HC	16
	CRC	13
	PA	8
CAN	PA	22
	HC	3
	CRC	2
CHN	CRC	45
	HC	38
	CRC+	29
	Other	16
DEU	HC	65
	CRC	60
FRA	HC	61
	CRC	53
	PA	42
IND	HC	30
	CRC	30
ITA	CRC	47
	HC	41
	PA	16
	CRC+	14
	PA+	11
	Other	10
JPN	CRC	298
	HC	286
	PA	72
	CRC-H	40
USA	HC	224
	CRC	77
	PA	4

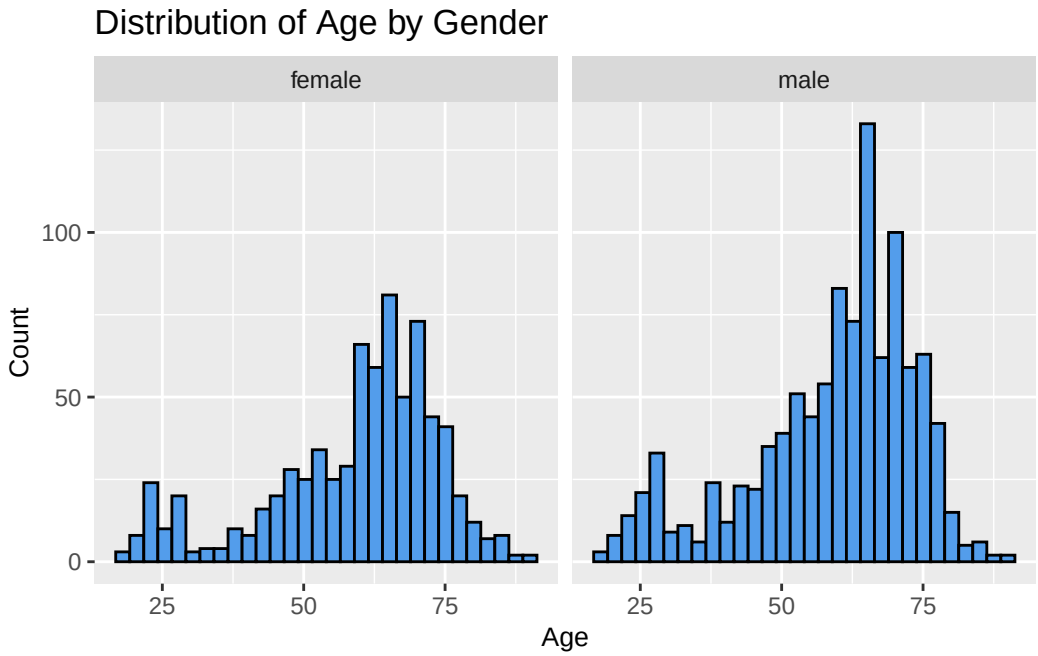
```
  # ------
filter(!is.na(age)) %>%
  # Plotting
  # --------
ggplot(aes(sample = age)) +
  geom_qq(size = 1, alpha = 0.8) +
  geom_qq_line() +
  # Labeling
  #---------
labs(title = "QQ Plot of Age") +
  theme(plot.title = element_text(hjust = 0, size = 10),
        axis.title = element_text(size = 10))
```



2-B: Age Histogram by Gender

```
# Gendered Histogram
####################
qualifying_studies %>%
  # Set Up
  # ------
```

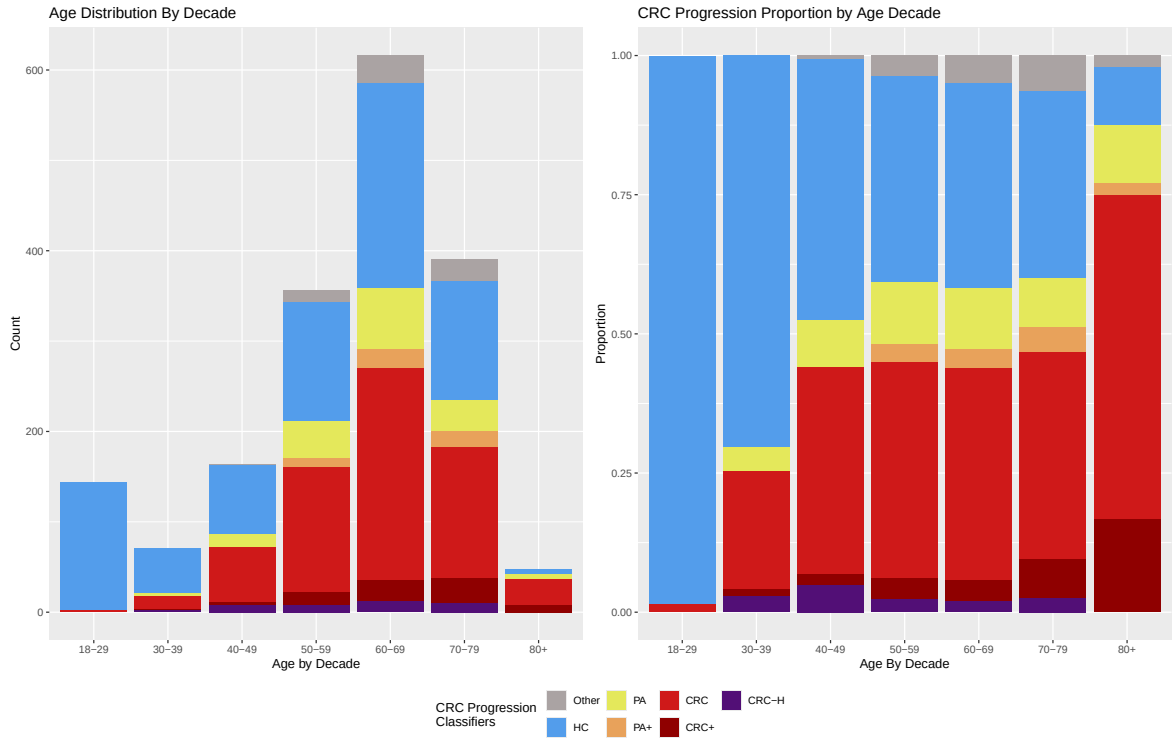
```
  filter(!is.na(age), !is.na(gender)) %>%
  # Plotting
  # --------
  ggplot(aes(x = age)) +
  geom_histogram(bins = 30, fill = "#539deb", color = "black") +
  facet_wrap(~gender) +
  # Labeling
  #---------
  labs(title = "Distribution of Age by Gender", x = "Age", y = "Count",) +
  theme(plot.title = element_text(hjust = 0),
        axis.title = element_text(size = 10))
```



2-C: Age Distribution and Proportion By Decade

```
# Barplot of age categories: Colored
####################################

decade1 <- qualifying_studies %>%
  # Set Up
  # ------
```

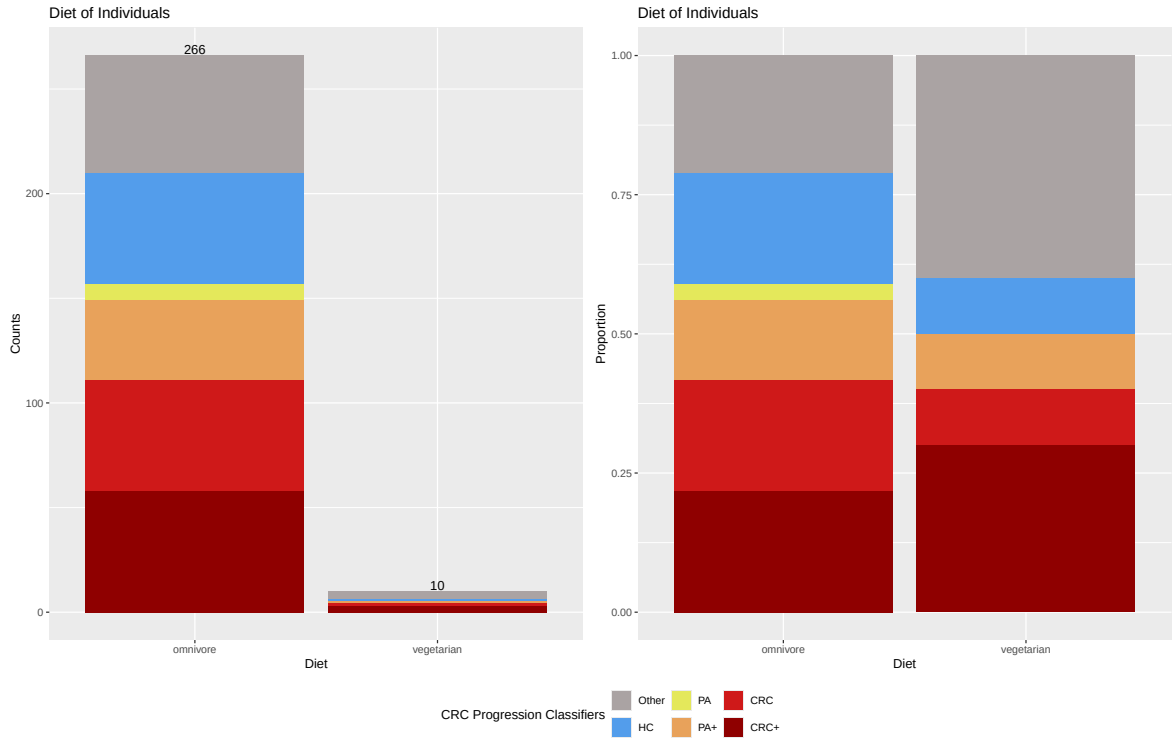
```
filter (!is.na(age_decade)) %>%
  # Plotting
  # --------
ggplot(aes(x=age_decade, fill = disease_class)) +
  geom_bar() +
  scale_fill_manual(values = progression_colors,
                    name = "CRC Progression \nClassifiers") +
  # Labeling
  #---------
  labs(
    title = "Age Distribution By Decade",
    x = "Age by Decade", y = "Count")

# Proportion: Age Decade
############################
decade2 <- qualifying_studies %>%
  # Set Up
  # ------
filter (!is.na(age_decade)) %>%
  # Plotting
  # --------
ggplot(aes(x=age_decade, fill = disease_class)) +
  geom_bar(position="fill") +
  scale_fill_manual(values = progression_colors,
                    name = "CRC Progression \nClassifiers") +
  # Labeling
  #---------
labs(title = "CRC Progression Proportion by Age Decade",
     x = "Age By Decade", y = "Proportion")

# Patchwork
(decade1 | decade2) + plot_layout(guides = "collect") &
  theme(legend.position = "bottom")
```



2-D: Diet Bar Plots

```
# Diet ~~~~~~~~~~~~~~~~~~~~
############################

# Diet Counts
#############
diet1 <- qualifying_studies %>%
  # Set Up
  # ------
  filter(!is.na(diet)) %>%
  # Plotting
  # --------
  ggplot(aes(x = reorder(diet, diet,
                         FUN = function(x) -length(x)))) +
  geom_bar(aes(fill = disease_class)) +
  scale_fill_manual(values = progression_colors,
                    name = "CRC Progression Classifiers") +
  # Labeling
```

```
  #---------
  geom_text(stat = "count", aes(label = after_stat(count)),
            vjust = -0.17, color = "black") +
  ggtitle("Diet of Individuals") +
  labs(title = "Diet of Individuals",
       x = "Diet",
       y = "Counts")

# Diet Proportion
#################
diet2 <- qualifying_studies %>%
  # Set Up
  # ------
  filter(!is.na(diet)) %>%
  # Plotting
  # --------
  ggplot(aes(x = reorder(diet, diet,
                         FUN = function(x) -length(x)))) +
  geom_bar(aes(fill = disease_class), position = "fill") +
  scale_fill_manual(values = progression_colors,
                    name = "CRC Progression Classifiers") +
  # Labeling
  #---------
  labs(
    title = "Diet of Individuals",
    x = "Diet",
    y = "Proportion")

# Patchwork
###########
(diet1 | diet2) + plot_layout(guides = "collect") &
  theme(legend.position = "bottom")
```



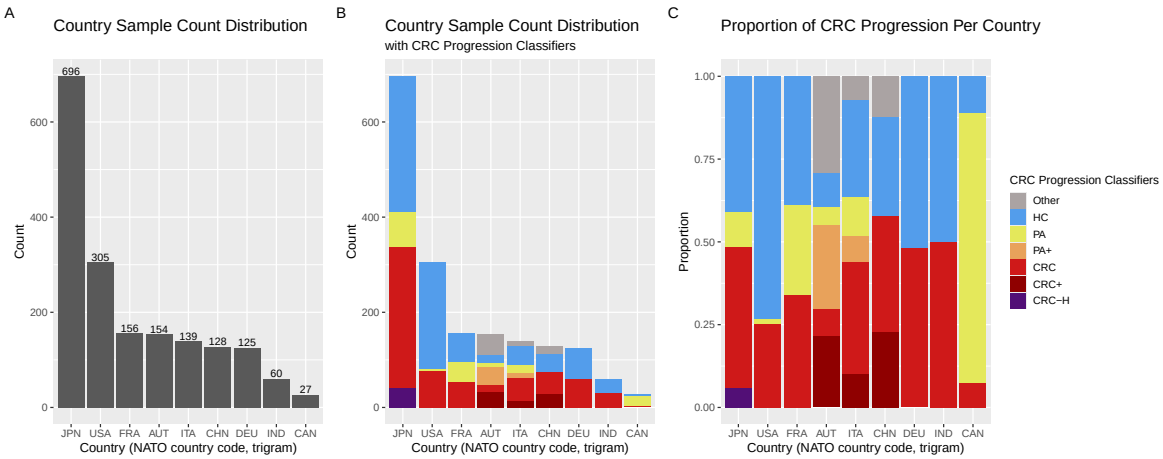
2-E: Country Distribution and Proportions

```
# ---

# Countries ~~~~~~~~~~~~~~~~
############################

# Country Count
################
country1 <- qualifying_studies  %>%
  # Set Up
  # ------
  filter(!is.na(country)) %>%
  # Plotting
  # --------
  ggplot(aes(x = reorder(country, country,
                         FUN = function(x) -length(x)))) +
  geom_bar() +
  # Labeling
```

```
  #---------
  geom_text(stat = "count", aes(label = after_stat(count)),
            vjust = -0.21, size = 3, color = "black") +
  labs(
    title = "Country Sample Count Distribution",
    x = "Country (NATO country code, trigram)",
    y = "Count") +
  theme(
        legend.position = "none",
        plot.title = element_text(size = 12))

# Country Fill
###############
country2 <- qualifying_studies  %>%
  # Set Up
  # ------
  filter(!is.na(country)) %>%
  # Plotting
  # --------
  ggplot(aes(x = reorder(country, country,
                         FUN = function(x) -length(x)),
             fill=disease_class)) +
  geom_bar() +
  scale_fill_manual(values = progression_colors,
                    name = "CRC Progression Classifiers") +
  # Labeling
  #---------
  labs(title = "Country Sample Count Distribution",
       subtitle = "with CRC Progression Classifiers",
       x = "Country (NATO country code, trigram)",
       y = "Count") +
  theme(
        legend.position = "none",
        plot.title = element_text(size = 12))

# Countries Proportions
########################
country3 <- qualifying_studies %>%
  # Set Up
  # ------
  filter(!is.na(country)) %>%
  # Plotting
```

```
  # --------
  ggplot(aes(x = reorder(country, country,
                         FUN = function(x) -length(x)), fill = disease_class)) +
  geom_bar(position = "fill") +
  scale_fill_manual(values = progression_colors,
                    name = "CRC Progression Classifiers") +
  # Labeling
  #---------
  labs(
    title = "Proportion of CRC Progression Per Country",
    x = "Country (NATO country code, trigram)",
    y = "Proportion") +
  theme(
        plot.title = element_text(size = 12))


# Display
#########
layout3 <- c( # t, l , b = t, r = l
  area(1, 1, 2, 2),
  area(1, 3, 2, 4),
  area(1, 5, 2, 6)
)

#plot(layout3)

geography_caption <- str_wrap(
  "(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
   (C) Stacked bar plot of the CRC progression classifier proportion for each country in the
  width = 200
)

country1 + country2 + country3 +
  plot_layout(design = layout3, guides = "collect", heights = c(1, 0.18)) +
  plot_annotation(
    tag_levels = "A",
    caption = geography_caption
  ) &
  theme(
        plot.title = element_text(hjust = 0, size = 14),
        plot.caption = element_text(hjust = 0, size = 11),
```

```
        legend.key.size = unit(0.5, "cm"),
        legend.title = element_text(size = 10),
        legend.direction = "vertical",
        legend.position = "right"
  )
```



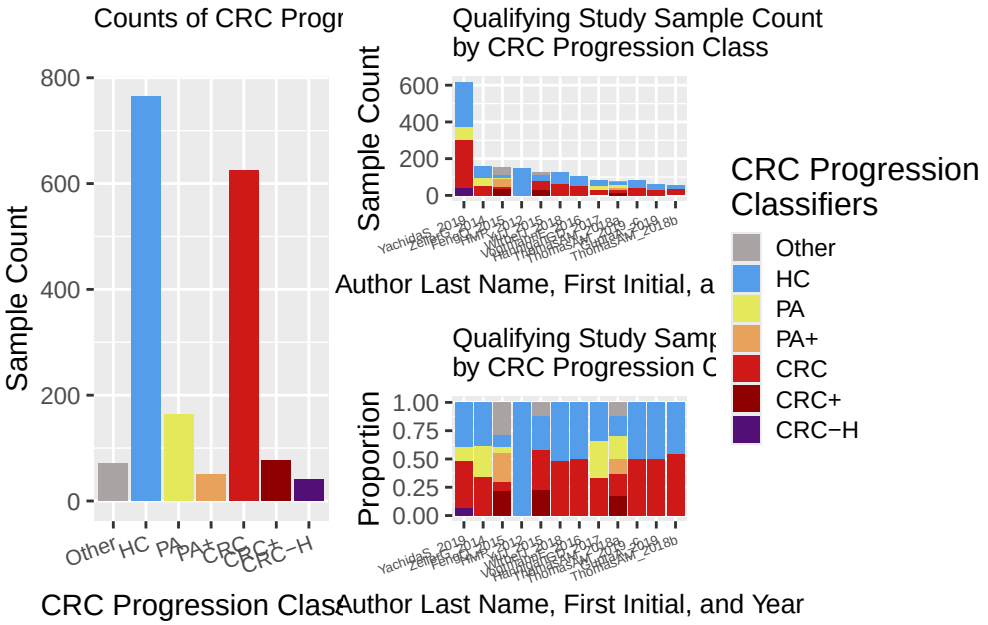
(A) Bar plot of country sample count, sorted in descending order. (B) Stacked bar plot of country sample count, sorted in descending order, filled with CRC progression classifiers. (C) Stacked bar plot of the CRC progression classifier proportion for each country in the multi−cohort study.

2-F: Contributors and CRC Progression Classifier Counts

```
# Bar Chart
###########
contributor1 <- qualifying_studies %>%
  # Plotting
  # --------
  ggplot(aes(x = reorder(study_name, study_name,
                         FUN = function(x) -length(x)),
                         fill = disease_class)) +
  geom_bar() +
  scale_fill_manual(values = progression_colors,
                    name = "CRC Progression \nClassifiers") +
  theme(axis.text.x = element_text(
    angle = 20, hjust = 1, lineheight = 1.1, size = 5),
    plot.title = element_text(size = 10),
    axis.title.x = element_text(size = 10)) +
  # Labeling
  #---------
```

```
  xlab("Author Last Name, First Initial, and Year") +
  ylab("Sample Count") +
  ggtitle("Qualifying Study Sample Count \nby CRC Progression Class")

# Bar Chart by Study
####################
contributor2 <- qualifying_studies %>%
  # Plotting
  # --------
  ggplot(aes(x = reorder(study_name, study_name,
                         FUN = function(x) -length(x)),
                         fill = disease_class)) +
  geom_bar(position = "fill") +
  scale_fill_manual(values = progression_colors,
                    name = "CRC Progression \nClassifiers") +
  theme(axis.text.x = element_text(
    angle = 20, hjust = 1, lineheight = 1.1, size = 5),
    plot.title = element_text(size = 10),
    axis.title.x = element_text(size = 10)) +
  # Labeling
  #---------
  xlab("Author Last Name, First Initial, and Year") +
  ylab("Proportion") +
  ggtitle("Qualifying Study Sample Proportions \nby CRC Progression Class")

# Bar Chart of Total CRC Progression Class Proportions
######################################################
class_count <- qualifying_studies %>%
  # Plotting
  # --------
  ggplot(aes(x=disease_class, stat = "count", fill = disease_class)) +
  geom_bar() +
  scale_fill_manual(values = progression_colors,
                    name = "CRC Progression \nClassifiers") +
  theme(axis.text.x = element_text(
    angle = 20, hjust = 1, lineheight = 1.1, size = 8),
    plot.title = element_text(size = 10))+
  # Labeling
  #---------
  labs(y = "Sample Count", x = "CRC Progression Classes") +
  ggtitle("Counts of CRC Progression Classes in Study")
```

```
class_count + (contributor1 / contributor2)  + plot_layout(guides = "collect") &
  theme(
        legend.key.size = unit(0.4, "cm"),
        legend.title = element_text(size = 12)
  )
```



Appendix 3 - Statistical Analyses

3-A: Gender Binomial Test

```
sex_counts <- table(qualifying_studies$gender)
binom.test(sex_counts["male"],
           sum(sex_counts),
           p = 0.5) # assumes 50/50 split in population
```

```
data:  sex_counts["male"] out of 1790L
number of successes = 1054, number of trials = 1790, p-value =
```

```
5.821e-14
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5656229 0.6117391
sample estimates:
probability of success
             0.5888268
```

3-B: CRC Progression Classifier and Age ANOVA

```r
age_CRC_ANOVA <- qualifying_studies %>%
  filter(!is.na(age))

age_CRC_ANOVA2 <- aov(age ~ disease_class, data = age_CRC_ANOVA)

summary(age_CRC_ANOVA2)
```

```
                Df Sum Sq Mean Sq F value Pr(>F)
disease_class    6  47581    7930   41.37 <2e-16 ***
Residuals     1783 341816     192
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
TukeyHSD(age_CRC_ANOVA2, conf.level=.95)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = age ~ disease_class, data = age_CRC_ANOVA)

$disease_class
                   diff         lwr         upr     p adj
HC-Other    -12.4863211 -17.5569314 -7.41571072 0.0000000
PA-Other     -3.0982480  -8.9042300  2.70773400 0.6983907
PA+-Other     0.3276056  -7.2176049  7.87281617 0.9999996
CRC-Other    -2.8635944  -7.9819245  2.25473572 0.6486798
CRC+-Other    1.4255004  -5.3200214  8.17102218 0.9960631
CRC-H-Other  -6.5323944 -14.6120932  1.54730448 0.2048102
PA-HC         9.3880730   5.8708635 12.90528255 0.0000000
```

```
PA+-HC       12.8139267   6.8480623 18.77979106 0.0000000
CRC-HC        9.6227267   7.4184948 11.82695863 0.0000000
CRC+-HC      13.9118214   8.9961933 18.82744958 0.0000000
CRC-H-HC      5.9539267  -0.6750151 12.58286846 0.1116511
PA+-PA        3.4258537  -3.1764093 10.02811657 0.7257036
CRC-PA        0.2346537  -3.3510090  3.82031627 0.9999957
CRC+-PA       4.5237484  -1.1473836 10.19488038 0.2189014
CRC-H-PA     -3.4341463 -10.6411712  3.77287848 0.7985262
CRC-PA+      -3.1912000  -9.1976758  2.81527577 0.7027781
CRC+-PA+      1.0978947  -6.3440481  8.53983755 0.9994830
CRC-H-PA+    -6.8600000 -15.5296010  1.80960101 0.2275496
CRC+-CRC      4.2890947  -0.6757430  9.25393246 0.1424118
CRC-H-CRC    -3.6688000 -10.3343144  2.99671442 0.6663334
CRC-H-CRC+   -7.9578947 -15.9412427  0.02545319 0.0513648
```

3-C: CRC Progression Classifier and BMI ANOVA

```r
bmi_CRC_ANOVA <- qualifying_studies %>%
  filter(!is.na(BMI))

bmi_CRC_ANOVA2 <- aov(BMI ~ disease_class, data = bmi_CRC_ANOVA)

summary(bmi_CRC_ANOVA2)
```

```
                Df Sum Sq Mean Sq F value Pr(>F)
disease_class    6   1740  290.00   18.23 <2e-16 ***
Residuals     1767  28117   15.91
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
TukeyHSD(bmi_CRC_ANOVA2, conf.level=.95)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = BMI ~ disease_class, data = bmi_CRC_ANOVA)

$disease_class
                  diff        lwr         upr     p adj
```

```
HC-Other     -3.1930673 -4.6543525 -1.73178218 0.0000000
PA-Other     -2.3568365 -4.0326947 -0.68097830 0.0006860
PA+-Other     0.7770732 -1.3967514  2.95089783 0.9407962
CRC-Other    -2.9665576 -4.4421570 -1.49095829 0.0000001
CRC+-Other   -1.1657934 -3.1154702  0.78388331 0.5720723
CRC-H-Other -4.9286078 -7.2564219 -2.60079360 0.0000000
PA-HC         0.8362308 -0.1828237  1.85528533 0.1897866
PA+-HC        3.9701406  2.2509882  5.68929290 0.0000000
CRC-HC        0.2265097 -0.4117416  0.86476105 0.9427677
CRC+-HC       2.0272739  0.6020684  3.45247942 0.0005602
CRC-H-HC     -1.7355404 -3.6456949  0.17461406 0.1034071
PA+-PA        3.1339097  1.2290131  5.03880635 0.0000268
CRC-PA       -0.6097211 -1.6491975  0.42975526 0.5949128
CRC+-PA       1.1910431 -0.4534499  2.83553604 0.3307735
CRC-H-PA     -2.5717713 -4.6506740 -0.49286854 0.0049920
CRC-PA+      -3.7436309 -5.4749667 -2.01229502 0.0000000
CRC+-PA+     -1.9428667 -4.0926038  0.20687048 0.1071576
CRC-H-PA+    -5.7056810 -8.2034499 -3.20791216 0.0000000
CRC+-CRC      1.8007642  0.3608858  3.24064262 0.0043012
CRC-H-CRC    -1.9620501 -3.8831772 -0.04092306 0.0416466
CRC-H-CRC+   -3.7628143 -6.0681506 -1.45747808 0.0000324
```

3-D: Linear Regression Model by CRC Progression Classifier

```r
qualifying_studies$age_centered <- scale(
  qualifying_studies$age,
  center = TRUE,
  scale = FALSE
)

qualifying_studies$disease_class <- relevel(
  qualifying_studies$disease_class,
  ref = "HC" # Changed to refer to healthy control rather than "Other"
)

age_bmi_scatterplot_lm <- lm(BMI ~ age_centered * disease_class,
               data = qualifying_studies)

summary(age_bmi_scatterplot_lm)
```

```
Call:
lm(formula = BMI ~ age_centered * disease_class, data = qualifying_studies)

Residuals:
    Min      1Q  Median      3Q     Max
-10.753  -2.687  -0.414   2.000  33.060

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      23.852059   0.151998 156.924  < 2e-16 ***
age_centered                     -0.013049   0.008295  -1.573 0.115863
disease_classOther                2.229690   0.655294   3.403 0.000682 ***
disease_classPA                   0.853353   0.364981   2.338 0.019495 *
disease_classPA+                  4.840180   0.785051   6.165  8.7e-10 ***
disease_classCRC                  0.422289   0.227348   1.857 0.063414 .
disease_classCRC+                 1.674100   0.594710   2.815 0.004932 **
disease_classCRC-H               -1.680342   0.645964  -2.601 0.009365 **
age_centered:disease_classOther   0.168426   0.064707   2.603 0.009321 **
age_centered:disease_classPA      0.029465   0.032615   0.903 0.366438
age_centered:disease_classPA+    -0.100111   0.075586  -1.324 0.185522
age_centered:disease_classCRC    -0.018786   0.016828  -1.116 0.264401
age_centered:disease_classCRC+    0.066627   0.044103   1.511 0.131043
age_centered:disease_classCRC-H   0.145561   0.052576   2.769 0.005689 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.97 on 1760 degrees of freedom
  (16 observations deleted due to missingness)
Multiple R-squared:  0.07076,   Adjusted R-squared:  0.0639
F-statistic: 10.31 on 13 and 1760 DF,  p-value: < 2.2e-16
```

Appendix 4 - Relevant R Scripts

4-A: cMD3 Installation

A script used to install BiocManager and the cMD3 package for the first time.

```r
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("curatedMetagenomicData")
```

4-B: Dry_Runs.R

A script used to determine what information was available on our qualifying studies and print to individual text files. Text files can be found on the GitHub for this project in the described file path in the script.

```
# Run Qualifying Studies List
source("./EDA/Qualifying_Studies_List.R")

# Studies available
####################
studies <- qualifying_studies %>%
  distinct(study_name) %>%
  pull(study_name)
print(studies)

# Information Available per Study
#################################
for (study in studies) {
  outfile <- paste0("./EDA/Kaitlyn_EDA/Info_Availability/", study, "_info_availability.txt")
  writeLines(curatedMetagenomicData(study, dryrun = TRUE), outfile)
}

# Dry Run
#########
Test <- curatedMetagenomicData("ThomasAM_2018b.+",
                               dryrun = TRUE) # Doesn't download the data, dry runs it
class(Test)
length(Test)
head(Test)
```

4-C: Group EDA Setup R Chunk

The R chunk at the beginning of the file to allow for rendering of all plots and data analysis.

```
source("Qualifying_Studies_List.R")
suppressPackageStartupMessages({
  library(tidyverse)
  library(knitr)
  library(kableExtra)
  library(patchwork)
```

```
  library(vegan)
  library(mia)
  library(scater)
  library(lefser)

  library(grid)
  library(gridExtra)
})

progression_order <- c(
  "HC",
  "PA",
  "PA+",
  "PA-M",
  "CRC",
  "CRC+",
  "CRC-H",
  "CRC-M",
  "Other"
)

progression_colors <- c(
  "HC" = "#539deb",
  "PA" = "#e4e85b",
  "PA+" = "#e8a25b",
  "CRC" = "#cf1919",
  "CRC+" = "#8f0000",
  "CRC-M" = "#4d0404",
  "CRC-H" = "#520f76",
  "Other" = "#aaa3a3"
)

qualifying_studies <- qualifying_studies %>%
  mutate(
    disease_class = factor(disease_class, levels = progression_order)
  )
```

Appendix 5 - Relevant Links

5-A: Project GitHub

Link to our project GitHub Repository: https://github.com/chejj/Therayess

5-B: cMD3 Data Set

Link to the cMD3 data set Bioconductor Page: https://bioconductor.org/packages/3.16/data/experiment/html/curatedMetagenomicData.html

Appendix 6 - Packages

```
citation("tidyverse")
```

```
To cite package 'tidyverse' in publications use:

  Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R,
  Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller
  E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V,
  Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to
  the tidyverse." _Journal of Open Source Software_, *4*(43), 1686.
  doi:10.21105/joss.01686 <https://doi.org/10.21105/joss.01686>.

A BibTeX entry for LaTeX users is

  @Article{,
    title = {Welcome to the {tidyverse}},
    author = {Hadley Wickham and Mara Averick and Jennifer Bryan and Winston Chang and Lucy
    year = {2019},
    journal = {Journal of Open Source Software},
    volume = {4},
    number = {43},
    pages = {1686},
    doi = {10.21105/joss.01686},
  }
```

```
citation("knitr")
```

```
To cite package 'knitr' in publications use:

  Xie Y (2025). _knitr: A General-Purpose Package for Dynamic Report
  Generation in R_. R package version 1.51, <https://yihui.org/knitr/>.

  Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition.
  Chapman and Hall/CRC. ISBN 978-1498716963

  Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible
  Research in R. In Victoria Stodden, Friedrich Leisch and Roger D.
  Peng, editors, Implementing Reproducible Computational Research.
  Chapman and Hall/CRC. ISBN 978-1466561595

To see these entries in BibTeX format, use 'print(<citation>,
bibtex=TRUE)', 'toBibtex(.)', or set
'options(citation.bibtex.max=999)'.
```

```
citation("kableExtra")
```

```
To cite package 'kableExtra' in publications use:

  Zhu H (2024). _kableExtra: Construct Complex Table with 'kable' and
  Pipe Syntax_. doi:10.32614/CRAN.package.kableExtra
  <https://doi.org/10.32614/CRAN.package.kableExtra>, R package version
  1.4.0, <https://CRAN.R-project.org/package=kableExtra>.

A BibTeX entry for LaTeX users is

  @Manual{,
    title = {kableExtra: Construct Complex Table with 'kable' and Pipe Syntax},
    author = {Hao Zhu},
    year = {2024},
    note = {R package version 1.4.0},
    url = {https://CRAN.R-project.org/package=kableExtra},
    doi = {10.32614/CRAN.package.kableExtra},
  }
```

```
citation("patchwork")
```

```
To cite package 'patchwork' in publications use:
```

```
  Pedersen T (2025). _patchwork: The Composer of Plots_.
  doi:10.32614/CRAN.package.patchwork
  <https://doi.org/10.32614/CRAN.package.patchwork>, R package version
  1.3.2, <https://CRAN.R-project.org/package=patchwork>.

A BibTeX entry for LaTeX users is

  @Manual{,
    title = {patchwork: The Composer of Plots},
    author = {Thomas Lin Pedersen},
    year = {2025},
    note = {R package version 1.3.2},
    url = {https://CRAN.R-project.org/package=patchwork},
    doi = {10.32614/CRAN.package.patchwork},
  }
```

```
citation("vegan")
```

```
To cite package 'vegan' in publications use:

  Oksanen J, Simpson G, Blanchet F, Kindt R, Legendre P, Minchin P,
  O'Hara R, Solymos P, Stevens M, Szoecs E, Wagner H, Barbour M,
  Bedward M, Bolker B, Borcard D, Borman T, Carvalho G, Chirico M, De
  Caceres M, Durand S, Evangelista H, FitzJohn R, Friendly M, Furneaux
  B, Hannigan G, Hill M, Lahti L, Martino C, McGlinn D, Ouellette M,
  Ribeiro Cunha E, Smith T, Stier A, Ter Braak C, Weedon J (2025).
  _vegan: Community Ecology Package_. doi:10.32614/CRAN.package.vegan
  <https://doi.org/10.32614/CRAN.package.vegan>, R package version
  2.7-2, <https://CRAN.R-project.org/package=vegan>.

A BibTeX entry for LaTeX users is

  @Manual{,
    title = {vegan: Community Ecology Package},
    author = {Jari Oksanen and Gavin L. Simpson and F. Guillaume Blanchet and Roeland Kindt a
    year = {2025},
    note = {R package version 2.7-2},
    url = {https://CRAN.R-project.org/package=vegan},
    doi = {10.32614/CRAN.package.vegan},
  }
```

```
citation("mia")
```

```
To cite package 'mia' in publications use:

  Borman T, Ernst F, Shetty S, Lahti L (2025). _mia: Microbiome
  analysis_. R package version 1.18.0,
  <https://microbiome.github.io/mia/>.

A BibTeX entry for LaTeX users is

  @Manual{,
    title = {mia: Microbiome analysis},
    author = {Tuomas Borman and Felix G.M. Ernst and Sudarshan A. Shetty and Leo Lahti},
    year = {2025},
    note = {R package version 1.18.0},
    url = {https://microbiome.github.io/mia/},
  }
```

```
citation("scater")
```

```
To cite package 'scater' in publications use:

  McCarthy DJ, Campbell KR, Lun ATL, Willis QF (2017). "Scater:
  pre-processing, quality control, normalisation and visualisation of
  single-cell RNA-seq data in R." _Bioinformatics_, *33*, 1179-1186.
  doi:10.1093/bioinformatics/btw777
  <https://doi.org/10.1093/bioinformatics/btw777>.

A BibTeX entry for LaTeX users is

  @Article{,
    author = {Davis J. McCarthy and Kieran R. Campbell and Aaron T. L. Lun and Quin F. Willis
    title = {Scater: pre-processing, quality control, normalisation and visualisation of sing
    journal = {Bioinformatics},
    year = {2017},
    volume = {33},
    issue = {8},
    pages = {1179-1186},
    doi = {10.1093/bioinformatics/btw777},
  }
```

```
citation("lefser")
```

```
To cite lefser in publications, use:

  Khleborodova A, Gamboa-Tuz S, Ramos M, Segata N, Waldron L, Oh S
  (2024). "Lefser: Implementation of metagenomic biomarker discovery
  tool, LEfSe, in R." _Bioinformatics_, btae707. ISSN 1367-4811,
  doi:10.1093/bioinformatics/btae707
  <https://doi.org/10.1093/bioinformatics/btae707>,
  <https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btae707

A BibTeX entry for LaTeX users is

  @Article{,
    title = {Lefser: Implementation of metagenomic biomarker discovery tool, LEfSe, in R},
    author = {Asya Khleborodova and Samuel D Gamboa-Tuz and Marcel Ramos and Nicola Segata a
    journal = {Bioinformatics},
    year = {2024},
    month = {11},
    pages = {btae707},
    issn = {1367-4811},
    doi = {10.1093/bioinformatics/btae707},
    url = {https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics
  }
```

```
citation("grid")
```

```
The 'grid' package is part of R.  To cite R in publications use:

  R Core Team (2026). _R: A Language and Environment for Statistical
  Computing_. R Foundation for Statistical Computing, Vienna, Austria.
  <https://www.R-project.org/>.

A BibTeX entry for LaTeX users is

  @Manual{,
    title = {R: A Language and Environment for Statistical Computing},
    author = {{R Core Team}},
    organization = {R Foundation for Statistical Computing},
    address = {Vienna, Austria},
    year = {2026},
```

```
    url = {https://www.R-project.org/},
  }

We have invested a lot of time and effort in creating R, please cite it
when using it for data analysis. See also 'citation("pkgname")' for
citing R packages.
```

```
citation("gridExtra")
```

```
To cite package 'gridExtra' in publications use:

  Auguie B (2017). _gridExtra: Miscellaneous Functions for "Grid"
  Graphics_. doi:10.32614/CRAN.package.gridExtra
  <https://doi.org/10.32614/CRAN.package.gridExtra>, R package version
  2.3, <https://CRAN.R-project.org/package=gridExtra>.

A BibTeX entry for LaTeX users is

  @Manual{,
    title = {gridExtra: Miscellaneous Functions for "Grid" Graphics},
    author = {Baptiste Auguie},
    year = {2017},
    note = {R package version 2.3},
    url = {https://CRAN.R-project.org/package=gridExtra},
    doi = {10.32614/CRAN.package.gridExtra},
  }
```