Independent Project Introduction

Michael Hajkowski

2025-05-04

Metagenomic Insights into AMR Gene Prevalence in Municipal Wastewater: One Health Approach

Data set and other supporting materials can be found in the GitHub Repository: BIOL710_AMRWWTP

```r
# load data
dataset <- read.csv("combined_amr_results_seasons.csv")
```

Introduction

With the evolution and emergence of novel bacterial and viral pathogens, identifying and classifying the threats they pose is the first step toward understanding their impacts on environmental and public health. Wastewater Treatment Plants (WWTP) play a major role in disseminating antimicrobial resistance (AMR), further circulating their infectivity range (Nguyen et al., 2021). Traditional wastewater-based epidemiology (WBE) approaches rely on membrane filtration concentration and polymerase chain reaction (PCR)-based molecular methods to identify targeted microorganisms and functional genes (Harrington et al., 2022). However, nanotrap-based concentration methods paired with next-generation sequencing (NGS) for metagenomic analysis provide higher species resolution with the ability to detect AMR presence and abundance (Ahmed et al., 2023). While the identification of AMR-associated genes is important, the integrative "One Health" approach recommended by the World Health Organization (WHO) elucidates the risk severity of AMR-associated genes across human, animal, and environmental health (Smith et al., 2024). Although monitoring pathogenic shedding in wastewater influent provides insight into population-level infectious rates, there is limited knowledge on how wastewater treatment affects pathogens and AMR dissemination, particularly in effluent discharged into the ocean (Fahrenfeld et al., 2014; Gothwal & Thatikonda, 2020).

To address these gaps, I propose to: (1) Develop and implement a WBE workflow integrating nanotrap technology with shotgun sequencing for metagenomic analysis, validating taxa detection accuracy using a bacterial mock community as a positive control. Detection efficacy will be assessed through taxonomic classification consistency and relative abundance comparisons (Lu et al., 2024). Data will be visualized through bar plots and heatmaps in ggplot2 (Hajkowski et al., 2024). (2) Characterize AMR gene variability across wastewater influent using the AMR One Health framework, leveraging metagenomic annotation databases (e.g., CARD, ResFinder) and quantifying AMR gene abundance through gene functional profiling with Zhang et al. and CZID pipelines (Zhang et al., 2021). Statistical significance of AMR gene enrichment was determined using a negative binomial generalized linear model (GLM) with season as a predictor. Residuals were assessed to confirm model assumptions. Data were visualized using bar plots and residual diagnostic plots in base R and ggplot2. (3) Apply an integrated risk assessment by combining the quantitative risk model from Zhang et al. with the Resistance Persistence Index (RPI) from Hajkowski et al., incorporating genomic AMR determinants and pathogen virulence factors to assess public health risk using CZID and NCBI databases (Hajkowski et al., 2024). Density plots, risk heatmaps, and network graphs in ggplot2 will visualize AMR persistence and risk across samples.

For this project I will be conducting a statistcal analyslsis on my second aim:

(2) Characterize AMR gene variability across wastewater influent using the AMR One Health framework, leveraging metagenomic annotation databases (e.g., CARD, ResFinder) and quantifying AMR gene abundance through gene functional profiling with Zhang et al. and CZID pipelines (Zhang et al., 2021). Statistical significance of AMR gene enrichment was determined using a negative binomial generalized linear model (GLM) with season as a predictor. Residuals were assessed to confirm model assumptions. Data were visualized using bar plots and residual diagnostic plots in base R and ggplot2.

Hypothesis:

Null Hypothesis: AMR gene abundance per sample is the statically similar across the two seasons (winter vs summer). Alternative Hypothesis: AMR gene abundance per sample is statistically higher in one season over the other (summer vs winter).

Statistical Model:

Because this dataset is over dispersed (variance is higher than the means, an assumption that violates Poisson models), I used a Negative binomial generalized linear model (GLM). GLM is great for this project because it accounts for over dipsersion of variances.

Visualizations:

I used ggplot2 to create a simple bar plot showing avaerage AMR reads per sample per season,including standard error bars, legend, and titles. Winter season is color coated as blue and summer is red.

R code Statical Analysis and Visuals

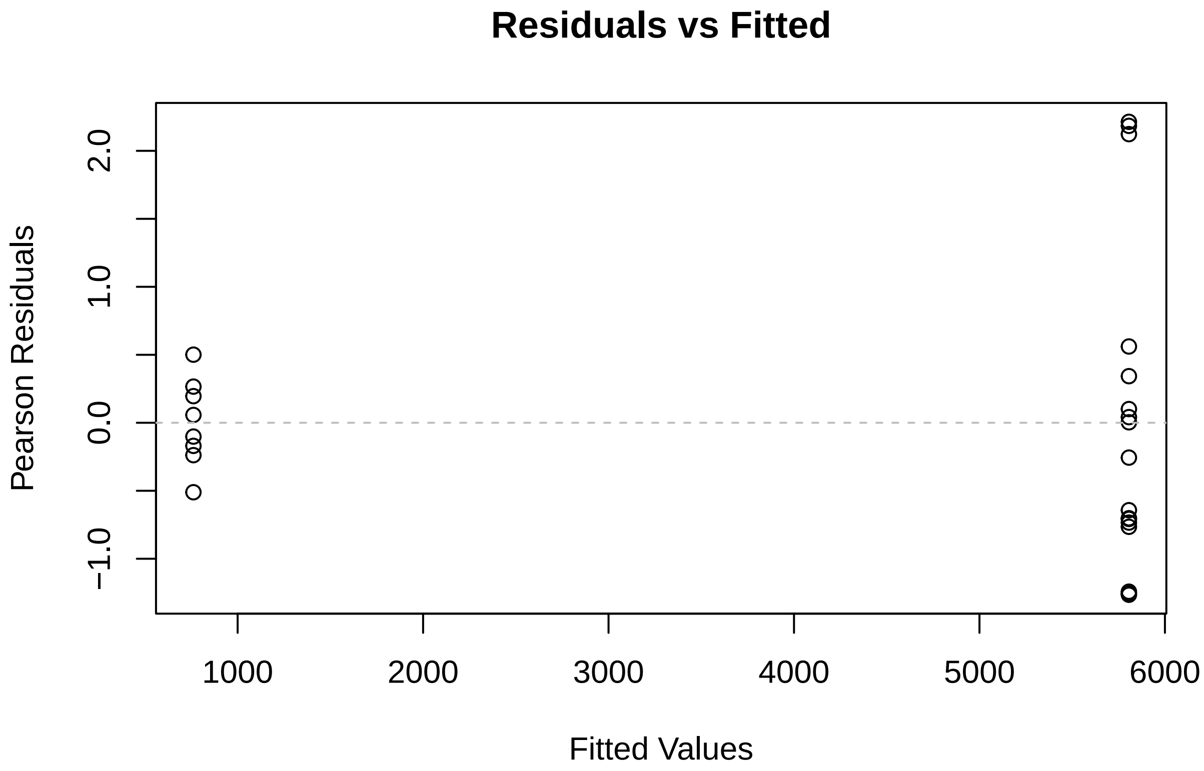
```r
# Load packages
library(MASS)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

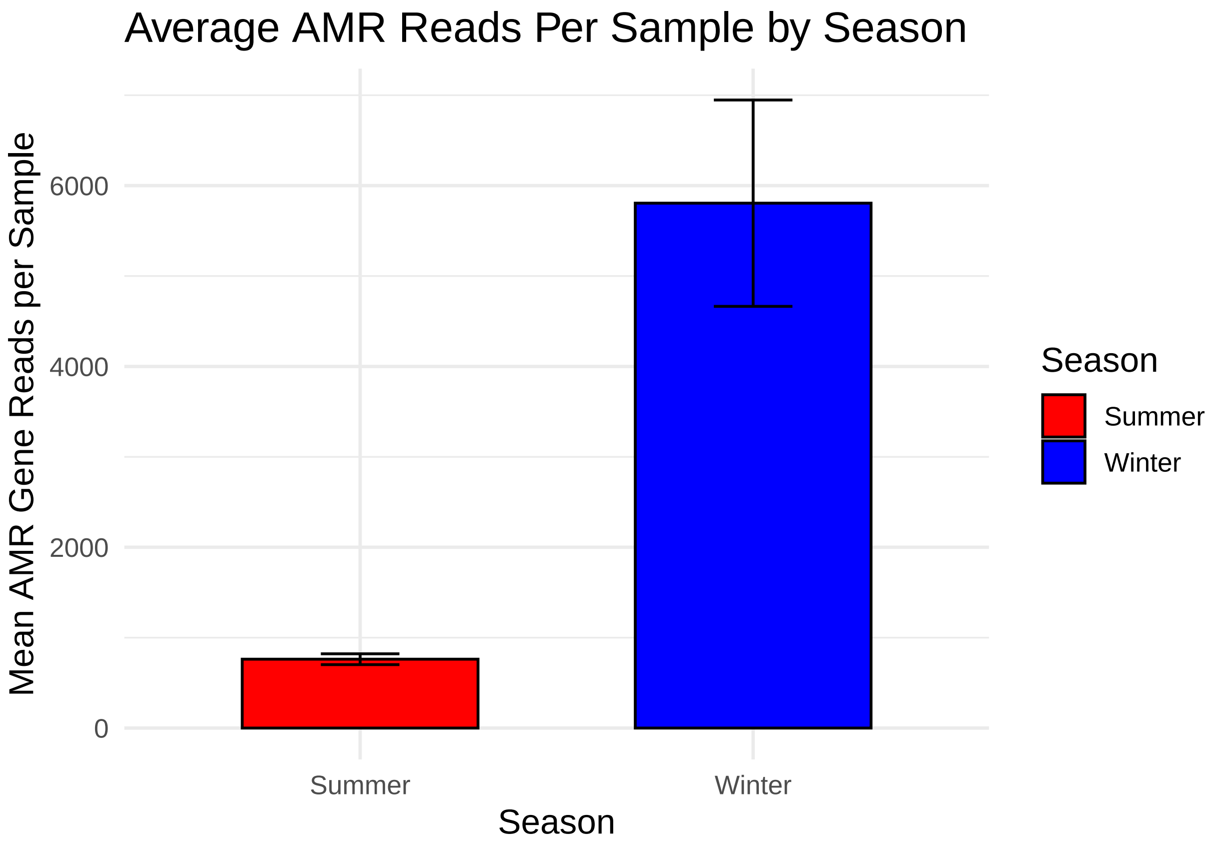
```r
# Read the data
df <- read.csv("combined_amr_results_seasons.csv")

# Clean data: remove NA values
df_clean <- df %>%
  filter(!is.na(num_reads), !is.na(Season), !is.na(sample_name))

# Summarize total reads per sample per season
df_summary <- df_clean %>%
  group_by(sample_name, Season) %>%
  summarise(total_reads = sum(num_reads), .groups = "drop")

# Fit negative binomial GLM
glm_nb <- glm.nb(total_reads ~ Season, data = df_summary)

# Show model summary
summary(glm_nb)
```

```
##
## Call:
## glm.nb(formula = total_reads ~ Season, data = df_summary, init.theta = 2.087472953,
##     link = log)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)     6.636      0.245  27.079  < 2e-16 ***
## SeasonWinter    2.031      0.297   6.838 8.03e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(2.0875) family taken to be 1)
##
##     Null deviance: 60.743  on 24  degrees of freedom
## Residual deviance: 26.936  on 23  degrees of freedom
## AIC: 450.42
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  2.087
##           Std. Err.:  0.550
##
##  2 x log-likelihood:  -444.423
```

```r
# Diagnostic plot: model assumptions check
res <- residuals(glm_nb, type = "pearson")
fitted_vals <- predict(glm_nb, type = "response")

plot(fitted_vals, res,
     main = "Residuals vs Fitted",
     xlab = "Fitted Values",
     ylab = "Pearson Residuals")
abline(h = 0, lty = 2, col = "gray")
```



```r
# Bar plot of AMR reads by season
df_plot <- df_summary %>%
  group_by(Season) %>%
  summarise(mean_reads = mean(total_reads),
            se = sd(total_reads)/sqrt(n()))

# Plot with legend and cleaner colors
ggplot(df_plot, aes(x = Season, y = mean_reads, fill = Season)) +
  geom_bar(stat = "identity", color = "black", width = 0.6) +
  geom_errorbar(aes(ymin = mean_reads - se, ymax = mean_reads + se), width = 0.2) +
  labs(
    title = "Average AMR Reads Per Sample by Season",
    y = "Mean AMR Gene Reads per Sample",
    x = "Season",
    fill = "Season"
  ) +
  scale_fill_manual(values = c("Summer" = "red", "Winter" = "blue")) +
  theme_minimal(base_size = 13)
```


Average AMR Reads Per Sample by Season
Mean AMR Gene Reads per Sample
6000
4000
2000
0
Summer
Winter
Season
Season
Summer
Winter