**Title: Quantification and Characterization of *Microcystis*-Specific Cyanophages in Lake Erie**

**Background**

Harmful algal blooms (HABs) in Lake Erie, driven by *Microcystis* proliferation, present significant public health and ecological challenges because of the release of microcystins (Backer et al. 2015, Briand et al. 2003). Cyanophages are virus that attack cyanobacteria (Yoshida et al. 2008, Weirich et al, 2014). Viral lysis events associated with cyanophages regulate *Microcystis* bloom dynamics and affects the release of toxins (Steffen et al. 2017) and microbial community interactions. Understanding these viral processes is vital to determining bloom progression and potential mitigation approaches (Xia et al. 2013).

**Research Question(s)**

* What is the spatial and temporal distribution of *Microcystis*-specific cyanophages in Lake Erie?
* How does viral diversity change across different locations and seasons?

**Objective(s)**

* To quantify *Microcystis*-specific cyanophages in Lake Erie across spatial and temporal scales.
* To analyze viral diversity using high-throughput sequencing

**Approach**

* Environmental water samples collected from various locations and time points in Lake Erie were analyzed using qPCR to quantify *Microcystis*-specific cyanophages. R was used to load and clean qPCR data, handling missing values and ensuring consistency across datasets with packages like *tidyverse* and *janitor*. Exploratory data analysis was conducted using *ggplot2* for visualization of viral abundance trends over time and space.

**R codes used**

* 1. **To generate fake qPCR data**

# Load required libraries

library(readxl)

library(ggplot2)

library(dplyr)

library(lubridate)

# Read the Excel file

data <- read\_excel("~/Desktop/Fake.xlsx")

# Clean and format data

data\_clean <- data %>%

mutate(

Date = as.Date(Date), # Convert to Date format

Target = as.factor(Target),

Location = as.factor(Location)

) %>%

filter(!is.na(AVG), !is.na(STD)) # Remove rows with missing AVG/STD

* 1. To determine is the spatial and temporal distribution of *Microcystis*-specific cyanophages in Lake Erie

# Generate the facet grid plot

ggplot(data\_clean, aes(x = Date, y = AVG, color = Target, group = Target)) +

geom\_line(linewidth = 0.7) + # Add lines connecting points

geom\_point(size = 2) + # Add data points

geom\_errorbar(

aes(ymin = AVG - STD, ymax = AVG + STD), # Add error bars

width = 0.5, linewidth = 0.5

) +

facet\_wrap(~Location, scales = "free\_x") + # Split by Location, free x-axis for dates

labs(

title = "Genes by Location",

x = "Date",

y = "Average Gene Copies",

color = "Target"

) +

theme\_bw() +

theme(

axis.text.x = element\_text(angle = 45, hjust = 1), # Rotate x-axis labels

strip.background = element\_blank(), # Remove facet background

strip.text = element\_text(face = "bold") # Bold facet titles

) + scale\_y\_log10()

ggsave(

"~/Desktop/qPCR\_Facet\_Plot.png", # Saves to desktop

width = 16,

height = 12,

dpi = 300

)

* 1. To generate fake viral load sequences

library(dplyr)

library(lubridate)

set.seed(2023)

# First create base dataframe without problematic case\_when

cyanophage\_data <- tibble(

Sample\_ID = paste0("CVP-", 1001:1200),

Collection\_Date = sample(seq(ymd('2023-01-01'), ymd('2024-12-31'), by = "day"), 200),

Location = sample(c("7M", "8M", "MB18", "Buoy", "Crib", "MB20", "4P"), 200, replace = TRUE,

prob = c(0.2, 0.2, 0.15, 0.15, 0.1, 0.1, 0.1)),

Viral\_Load\_PFU\_ml = round(10^runif(200, min = 9, max = 12)),

Sequencing\_Depth = round(rnorm(200, mean = 80, sd = 15))

)

# Now add columns using mutate() where n() works properly

cyanophage\_data <- cyanophage\_data %>%

mutate(

GC\_Content = case\_when(

Location %in% c("7M", "8M") ~ round(rnorm(n(), 52, 2), 1),

Location == "MB18" ~ round(rnorm(n(), 48, 1.5), 1),

TRUE ~ round(rnorm(n(), 55, 3), 1)

),

Genome\_Length = case\_when(

Location %in% c("Buoy", "Crib") ~ sample(c(40000, 80000), n(), replace = TRUE),

Location == "4P" ~ 120000,

TRUE ~ sample(c(160000, 200000), n(), replace = TRUE)

),

Host\_Specificity = case\_when(

Location %in% c("7M", "8M") ~ sample(c("Prochlorococcus-MED4", "Prochlorococcus-MIT9312"), n(), replace = TRUE),

Location == "MB18" ~ "Synechococcus-CC9311",

TRUE ~ "Synechococcus-WH7803"

),

CRISPR\_Spacers = round(runif(n(), 2, 15)),

Structural\_Genes = round(rnorm(n(), 25, 3)),

Metabolic\_Genes = round(rnorm(n(), 15, 2)),

Unique\_ORFs = round(abs(rnorm(n(), 50, 15))),

Contaminant\_Reads = case\_when(

Location == "Buoy" ~ round(Sequencing\_Depth \* 0.15),

Location == "Crib" ~ round(Sequencing\_Depth \* 0.25),

TRUE ~ round(Sequencing\_Depth \* 0.05)

),

Data\_Quality = case\_when(

Location %in% c("7M", "8M") ~ sample(c("Excellent", "Good"), n(), replace = TRUE, prob = c(0.7, 0.3)),

Location == "MB18" ~ sample(c("Good", "Fair"), n(), replace = TRUE),

TRUE ~ "Fair"

)

)

# Save output

write.csv(cyanophage\_data, "cyanophage\_sequencing\_dataset.csv", row.names = FALSE)

* 1. Beta diversity

library(readxl)

library(vegan)

library(ggplot2)

library(dplyr)

# 1. Load and prepare data (avoid negative values)

df <- read\_excel("~/Desktop/Fake\_cyanophage\_sequencing\_dataset.xlsx") %>%

select(-Sample\_ID, -Collection\_Date) %>%

mutate(Location = as.factor(Location)) %>%

mutate(Viral\_Load\_PFU\_ml = log10(Viral\_Load\_PFU\_ml))

# 2. Scale features WITHOUT CENTERING

community\_data <- df %>%

select(GC\_Content, Genome\_Length, Structural\_Genes,

Metabolic\_Genes, Unique\_ORFs, Contaminant\_Reads) %>%

scale(center = FALSE, scale = TRUE) # Critical fix here

# 3. Verify no negative values

print(range(community\_data)) # Should show [1] 0.0000 [some positive max]

# 4. Calculate Bray-Curtis

dist\_matrix <- vegdist(community\_data, method = "bray")

# 5. Perform PCoA ordination

pcoa <- cmdscale(dist\_matrix, eig = TRUE, k = 3)

# 6. Visualize results

pcoa\_scores <- as.data.frame(pcoa$points) %>%

rename(PCoA1 = "V1", PCoA2 = "V2") %>%

bind\_cols(Location = df$Location)

ggplot(pcoa\_scores, aes(PCoA1, PCoA2, color = Location)) +

geom\_point(size = 3, alpha = 0.8) +

stat\_ellipse(level = 0.7) +

theme\_minimal() +

labs(title = "PCoA of Cyanophage Communities by Location",

x = paste0("PCoA1 (", round(pcoa$eig[1]/sum(pcoa$eig)\*100, 1), "%)"),

y = paste0("PCoA2 (", round(pcoa$eig[2]/sum(pcoa$eig)\*100, 1), "%)"))

ggsave("cyanophage\_beta\_diversity.png", width = 10, height = 8, dpi = 300)

**References**

1. Backer, L. C., Manassaram-Baptiste, D., LePrell, R., & Bolton, B. (2015). Cyanobacteria and algae blooms: Review of health and environmental data from the Harmful Algal Bloom-Related Illness Surveillance System (HABISS) 2007-2011. *Toxins, 7*(4), 1048–1064. <https://doi.org/10.3390/toxins7041048>
2. Briand, J. F., Jacquet, S., Bernard, C., & Humbert, J. F. (2003). Health hazards for terrestrial vertebrates from toxic cyanobacteria in surface water ecosystems. *Veterinary Research, 34*(4), 361–377. <https://doi.org/10.1051/vetres:2003019>
3. Steffen, M. M., Davis, T. W., McKay, R. M. L., Bullerjahn, G. S., Krausfeldt, L. E., Stough, J. M. A., ... Wilhelm, S. W. (2017). Ecophysiological examination of the Lake Erie *Microcystis* bloom in 2014: Linkages between biology and the water supply shutdown of Toledo, OH. *Environmental Science & Technology, 51*(12), 6745–6755. <https://doi.org/10.1021/acs.est.7b00856>
4. Weirich, C. A., & Miller, T. R. (2014). Freshwater harmful algal blooms: toxins and children's health. *Current Problems in Pediatric and Adolescent Health Care, 44*(1), 2–24. <https://doi.org/10.1016/j.cppeds.2013.10.007>
5. Xia, H., Li, T., Deng, F., & Hu, Z. (2013). Freshwater cyanophages. *Virologica Sinica, 28*(4), 253–259. <https://doi.org/10.1007/s12250-013-3370-1>
6. Yoshida, T., Nagasaki, K., Takashima, Y., Shirai, Y., Tomaru, Y., Takao, Y., ... Ogata, H. (2008). Ma-LMM01 infecting toxic *Microcystis aeruginosa* illuminates diverse cyanophage genome strategies. *Journal of Bacteriology, 190*(5), 1762–1772. <https://doi.org/10.1128/JB.01534-07>