**Coding to Save the Frogs: Python Script to Organize and Process data**

**Introduction**

Chytridiomycosis is a wildlife disease that is lethal to many amphibian species and has caused global declines in susceptible species (Pounds et al., 2006; Skerratt et al., 2007; Voyles et al., 2009). The disease is caused by a fungal pathogen, *Batrachochytrium dendrobatidis* (*Bd*). *Bd* has a two-stage life cycle that consists of a substrate-dependent immobile sporangium and a free-living uniflagellated, motile zoospore (Berger et al. 2005; Longcore et al. 1999). Infection occurs during the motile zoospore stage of the pathogen’s life cycle (Berger, et al., 2005; Longcore et al., 1999). Zoosporangia produce motile zoospores and then release the new motile zoospores into the environment to re-infect the same host or transmit to another individual host (Van Rooij et al., 2012; Fig. 4). Once infection is established within a host, increases in infection intensity (or pathogen load) in the skin is a key feature of pathogenesis and disease development (Voyles et al. 2009). The increases in densities of zoosporangia and zoospores disrupt the epidermis, leading to dysregulation of normal skin functioning, impaired osmoregulation, and ultimately death (Voyles et al., 2009). Therefore, understanding how thermal performance of zoospore and sporangium is influenced across temperatures may be essential to making predictions about performance of isolates in the face of climate change.

*Bd* growth and reproductive traits (e.g., zoospore production and whole culture growth) are temperature sensitive (Berger et al. 2005; Muletz-Wolz et al. 2019). Generally, it is thought that *Bd* has an thermal range of 2-28℃ (Berger et al., 2005; Voyles et al., 2017). Notably, temperatures in a range 25-28°C are considered the thermal maximum for most isolates due to decreased zoospore production and motility and zoosporangium death (Piotrowski et al., 2004; Stevenson et al., 2013; Voyles et al., 2017). Lower temperatures (2-5℃) can reduce the total reproductive output and cause protracted periods of zoospore activity, which may be important for transmission and disease development (Voyles et al., 2012; Woodhams et al., 2008).

Different isolates of *Bd* may vary in their thermal performance across the range of temperatures *Bd* is known to survive at (Rosenblum et al., 2013; Stevenson et al., 2013). Investigators have collected isolates of *Bd* from infected amphibians around the globe and throughout a wide range of thermal environments (Rosenblum et al., 2013) Additionally, a single isolate can demonstrate variation in performance traits when translocated to a novel thermal environment (Muletz-Wolz et al., 2019). Therefore, variation in performance across thermal environments may occur even in isolates that are genetically similar if the local environments of origin were distinct (Byrne et al., 2019).

Our current understanding of variation among *Bd* isolates in temperature responses stems from comparisons of isolates collected from large-scale geographic locations (e.g., tropics compared to temperate regions; Muletz-Wolz et al. 2019; Voyles et al. 2017). However, several studies have noted that additional investigations should focus on variation among isolates that are found in comparatively smaller geographic regions. Examining thermal responses of multiple *Bd* isolates to from fine-scale source populations will help to characterize the functional traits that allow *Bd* to be so adaptive and lethal for amphibians across the globe (Daskin et al., 2011). I have collected multiple isolates of *Bd* from across regions of the United States to investigate fine-scale differences in thermal responses of both pathogen life stages among isolates. The geographic origins of the isolates span multiple climactic regions and were collected from various host species. I predict that these isolates will vary significantly in their zoospore production and whole population growth rate.

For my project in this course, I am focusing on coding that will enable me to be more efficient and accurate in my data organization and analyses. I have conducted seven temperature experiments over the past year and have multiple files consisting of data from 1 or more of the seven temperature experiments. My files are not all formatted the same due to my first experiments being less efficient in organizing the data. For analyses and plotting of these data I require all the data files to be combined into one file that is formatted correctly. Additionally, I need to generate new columns using operations to subtract my negative control samples from my experimental samples to give me a normalized value. Lastly, I need to ensure that there are checks in place to catch any errors where incorrect values or empty values may occur to ensure accuracy across experiments. This project will allow me to move forward in drawing conclusions from my research and continuing to try and save amphibians that face global declines from Chytridiomycosis.

**Methods**

I used R studio with a suit of packages including ggplot, readxl, and tidyverse. To accomplish this project, I made a script that joined my files together. This was done using rbind to join the cleaned data files into one new data frame. I also renamed the headers for each column so that each file has a uniform structure after cleaning. Within the script for cleaning and re-structuring the data, I created a clean data function to manipulate each imported data frame in a desirable fashion while saving the final product as a new, small data frame (df). To do this, I created a matrix and saved it as “dat” df to have the number of rows and columns that each imported file will have. Inside the new “dat” df I saved vectors that I created in the matrix and instructed R to repeat these vectors according to the number of times I need and pull values from the range of cells that exist.

Next, I created a section of script to name the new columns and pair each pulled value with the appropriate placement in the new, small df for this file. In this code, each row of the raw file was named according to the isolate that the cell values belong to. Half of each row was assigned as positive sample values, while the other half of each row was assigned as negative control sample values. Each cell value was also officially assigned to its data type, temp treatment, and day of collection in the experiment. I used the mutate function to generate a new column that used an order of operations to subtract my negative samples from my experimental sample within the Pos\_Neg column.

For the second portion of my project, I used ggplot to create a plot of a different data set from the second chapter of my thesis. I made a set of boxplots of my two evolved lineages based on temperature treatments for each of the 5 isolates that I tested. I additionally jittered the raw data points over the boxplot to show the outliers and distribution of the data used. I place significance symbols over the pairs of lineage boxplots for each isolate that are statistically different from each other using the geom\_signif package.

**Results**

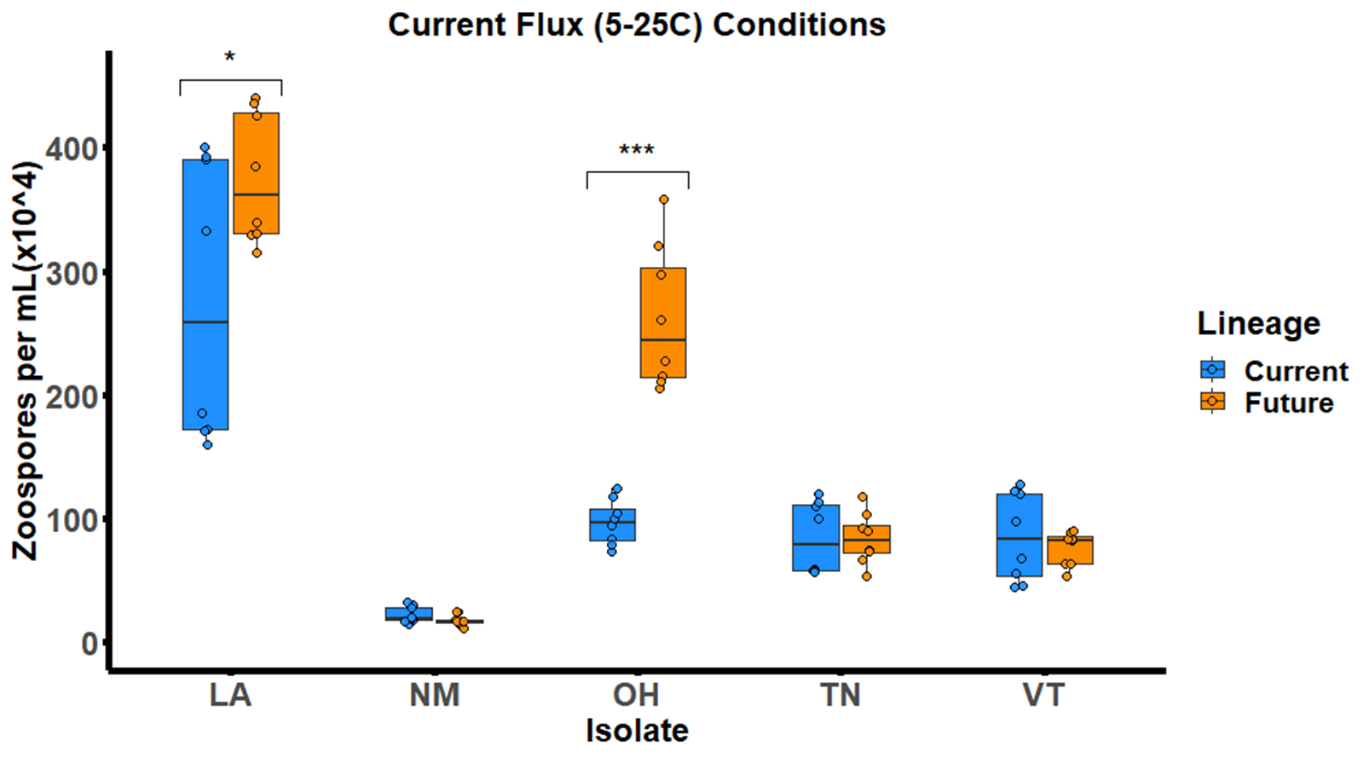
For this project, I was utilizing R studio and several common packages to take a raw excel file that my OD reader generates to manipulate and clean the data cells I desire into a new data frame. The function that I created can easily be used for every day of sample collection that I will ever use for my thesis. I can tweak the code to add in more rows or remove them according to how many isolates and lineages I run at a time. I can run this code after each sampling day and at the end of the experiment, combine all mini cleaned data frames into one large data frame to run statistics and make plots from.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 1. The final data frame produced displays the values that the reader produced and assigned their appropriate data type, day of collection in the experiment, temperature treatment, positive or negative sample labels, and the isolate the readings belong to. Additionally, the row and column that the data was pulled from the raw excel file is included in the data frame for easy QAQC.

The second portion of this project produced new figures that not only allow visualizations of the two evolved lineages for each isolate as boxplots that show the means and outliers. But also, this second portion allows for the jittering of the raw data points so that the real distribution of the data that plays into statistical analyses can easily be understood. These plots produced parameters which I can determine based on the new cleaned data frame script that I composed for the first portion of this project.

Figure 2. The two evolved lineages from each of the five isolates are shown on the x-axis in their boxplots. The evolved lineage that grew in current climate conditions is represented in blue, while the evolved lineage that grew in the future climate conditions is represented in orange. Zoospore densities are shown on the y-axis and the data used represents two sampling days that make up the peak period of zoospore production during this experiment. Significant symbols represent the level of statistical differences between the two lineages of each isolate (\* = >0.05; \*\* = >0.01; \*\*\* = >0.001).

**Conclusions**

This first R script is specific to the data collected for my first chapter of my thesis. In order to publish a manuscript on this data, I need to explore many figures and analyses which all require the data to be in a single file with the same formatting and information. In my research, I often have to copy and paste my data into an excel file from a raw excel file of optical density readings that my spectroscopy machine outputs per experimental sampling day. This leaves a fair amount of room for error and being able to check where my NAs occur is extremely valuable. Additionally, every experiment I conduct has slightly different sampling quantities and schedules so exact repeated formatting from one experiment to the other is not always possible. Being able to write a script that will allow me to reformat and check my data after collection will be crucial to more efficient turn around periods for data analyses and publications. I have written a script that will automate the copy and paste process from my raw data files of the spectroscopy machine to a single data frame across experiments, while recording the cell location from the raw data file in which the values belong for quality checking.

I also wrote coding for improved figures and data visualizations that allow me to better represent the interpretations of my data for publications. These visualizations not only allow me to clearly compare the two lineages of an isolate for zoospore production, but also allow me to compare zoospore production of lineages across all isolates measured. This semester was an opportunity to learn a improve on my coding skills to attempt a new script that I require for future experiments and hone in on my skills at being more coding literate. While my project is simple, this is beneficial to my research and can be expanded upon in the future.