Software Tools (BINF6210) – Assignment #1

Veedhi Solanki

1301912

October 6, 2023

INTRODUCTION:

Biodiversity is huge and each different population is connected in some way such that they display some type of relationship meaning that they are either affected, not affected from the presence of each other or they could show beneficial or mutualistic relationships. In terms of aquatic biodiversity, there are a wide variety of families. This includes different types of fishes, reptiles, invertebrates, mollusks and many more. In this project, the interest is to see if the population of family Dreissenidae influences the population of family Cyprininae in any way over the world. This is a very exploratory approach in a way that it is not just to focus on the relationship between these two populations but also, if there is any specific unusuality seen from the BOLD data then it would be interesting even to see the trend of the population in one or both population over the world in terms of the number of species and bin richness. Family of Dreissenidae are seen in freshwaters and have a wide range of habitat such as in Asia, Europe as well as North America which includes various types of mussels including the invasive zebra mussels species.

As shows in the study by (Wisniewski et al., 2013) there are many studies done in terms of the negative relationship between the mussels and surrounding species however, less is discovered in terms of the benefit that mussels can give or have mutualistic relationship with other species. So, the family Cyprininae has a wide habitat range all over the world which includes fishes (Yang et al., 2021). Therefore, interest is to see if the population of the family Dreissenidae affects in any way to the population of family Cyprininae or vise-versa in different countries of the world where both families are found co-residing. The relationship that exists between them can be predicted and classified by the number of bin richness in different countries for these two families. It can be further looked in by looking at the species richness as well. Moreover, how the bin richness is between different countries for both families can also be used to see how dissimilar regions meaning by country can have different relationships between the population of two families. Overall, this study seeks to analyze unique trends seen in the population of the family Dreissenidae all over the world. Then it is looking to compare the relationship between the families of Dreissenidae and Cyprininae by looking at the bin dissimilarity between different countries where the population of these two families co-exist.

RESULTS AND DISCUSSION:

Analysis done in this study utilized R Studio of which specifically tidyverse (Whickham et al., 2019) and vegan packages provided efficient functions and ways to perform tests on data.

1. **Analysis of Dreissenidae family Bin counts and Species counts:**

The data frame of the Dreissenidae obtained from the BOLD was filtered by the countries with data present and then it was further filtered such that there countries with not available data for bin count and species count were removed by the following R code.

```
filtered_country_Dreissenidae <-  dfBOLD_Dreissenidae %>%
 filter(!is.na(country)) %>%
 filter(!(is.na(bin_uri) & !is.na(species_name))) %>%
 group_by(country) %>%
 summarise(SpeciesCount.D = n_distinct(species_name),
        BINCount.D = n_distinct(bin_uri)) %>%
 arrange(desc(SpeciesCount.D))
```

- **SpeciesCount.D:** The number of distinct species within Dreissenidae in each country.
- **BINCount.D:** The number of distinct Barcode Index Numbers within Dreissenidae in each country.

Interest was to discover if there were enough bins discovered as the number of species present in different countries. So, by plotting the number of Bins and number of species in specific countries the data were visualized as a bar plot as shown in the **figure 1** below.

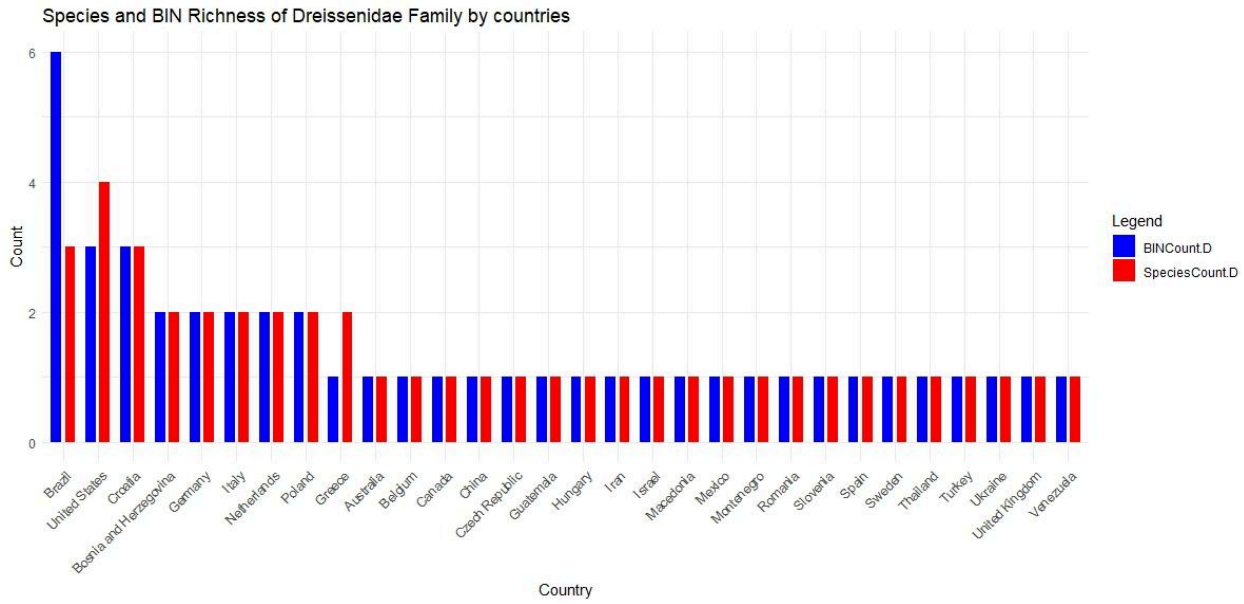

Figure 1: The grouped bar graph shows the differences in the Bin count(in blue) and species count(in red) between the different countries for the Dreissenidae family.

Code used to generate Figure 1:

```
long_data.D <- pivot_longer(filtered_country_Dreissenidae, cols = c(SpeciesCount.D, BINCount.D), names_to = "Variable")

ggplot(long_data.D, aes(x = reorder(country, -value), y = value, fill = Variable)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8), width = 0.6) +
  labs(x = "Country", y = "Count", fill = "Variable") +
  ggtitle("Species and BIN Richness of Dreissenidae Family by countries") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_manual(values = c("SpeciesCount.D" = "red", "BINCount.D" = "blue")) +
  guides(fill = guide_legend(title = "Legend"))
```

As shown in **figure 1**, there is a significantly higher number of bin count than species count in Brazil for the Dreissenidae family. This means that it is possible that some species have multiple bins while it is also possible that some species are yet to identify properly and it can be done through new bins identified. Another significance is that the United States and Greece have higher numbers of species than the bin counts, this can be due to the fact that there are not enough bins discovered for the species in these two countries. Additionally, all other countries have the same number of bins registered as the number of species. It is also very interesting to notice that there are many countries that have bin and species about three. This

figure 1, also helps to see the ranges within the Dreissenidae family reside all over the world, but it is possible that many countries do not have either bin or species data as the countries with unavailable data for either were removed at the beginning. Overall, the plot provides insight to the diversity in the bin and species richness of the Dreissenidae family.

2. **Analysis of Cyprininae family Bin counts and Species counts:**

Again, similar to the Dreissenidae family, data of the Cyprininae family were grouped based on countries and then the countries with not available data for species count or bin counts were removed. The code for it in R is shown here:

```
filtered_country_Cyprininae <-  dfBOLD_Cyprininae %>%
 filter(!is.na(country)) %>%
 filter(!(is.na(bin_uri) & !is.na(species_name))) %>%
 group_by(country) %>%
 summarise(SpeciesCount.C = n_distinct(species_name),
        BINCount.D = n_distinct(bin_uri)) %>%
 arrange(desc(SpeciesCount.C))
```

- **SpeciesCount.C:** The number of distinct species within Cyprininae in each country.
- **BINCount.D:** The number of distinct Barcode Index Numbers within Cyprininae in each country.
  (These names were automatically generated in data frames, but they mean different things in the filtered_country_Cyprininae and filtered_country_Dreissenidae dataframes)

3. **Comparison and Correlation between Dreissenidae and Cyprininae**

Next, to compare if the population of one family affects the other family, the population trends of Dreissenidae and Cyprininae families in terms of bin count were analyzed in the countries where both families are present. So, for that a new dataframe was created where it was merged based on the common countries from the previously created datta frames for individual families. The code for how the data frames were merged based on common countries in both families is as below.

```
common_countries_data <- inner_join(filtered_country_Dreissenidae,
filtered_country_Cyprininae, by = "country")
```

The merged data frame common_countries_data includes the following columns:

- Country: List of countries that includes both the population of Dreissenidae and Cyprininae
- SpeciesCount.D: Species count for Dreissenidae
- SpeciesCount.C: Species count for
- BinCount.D.x: Bin count for Dreissenidae
- BinCount.D.y: Bin count for Cyprininae
- Family.D: Name of the family Dreissenidae
- Family.C: Name of the family Cyprininae

To visualize the common_dountries_data, dot plot was created. This was to compare the bin counts of two families in the same countries. It was done as per the following code:

```
ggplot(common_countries_data, aes(x = country, y = BINCount.D.x, color = "Dreissenidae")) +
  geom_point(size = 3) +
  geom_point(aes(x = country, y = BINCount.D.y, color = "Cyprininae"), size = 3) +
  labs(title = "Dot Plot of BIN Count by Country for Dreissenidae and Cyprininae",
       x = "Country",
       y = "Count") +
  scale_color_manual(values = c("Dreissenidae" = "blue", "Cyprininae" = "red"),
                     name = "Family") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_flip() +
  scale_y_continuous(breaks = seq(min(common_countries_data$BINCount.D.x,
common_countries_data$BINCount.D.y),
                                  max(common_countries_data$BINCount.D.x,
common_countries_data$BINCount.D.y),
                                  by = 2))
```

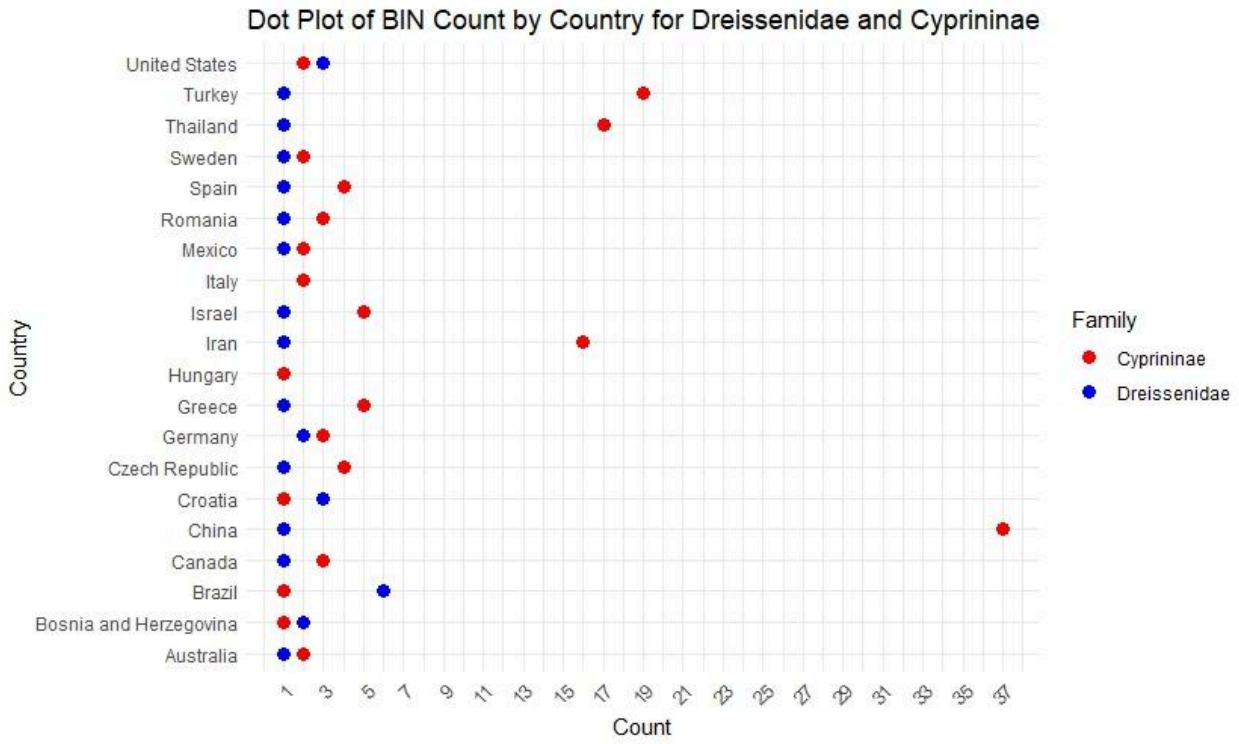

Figure 2: The plot highlights the differences in BIN counts between the two families in different countries where both families co-exist.

As from **figure 2,** it is observed that Turkey, Thailand, Iran and China have significantly higher bin counts for the Cyprininae family than the Dreissenidae family. In these countries, it is significant that the population of Cyprininae affects the population of Dreissenidae. Also, a common theme in these countries is that they are in or around Asia.

The United States, Sweden, Mexico, Germany, Bosnia and Herzegovina and Australia have about the same number of bin counts for both families meaning that there is a very minor difference in bin count for both families in these countries. In these countries there is no effect of the families on each other while they co-exist. This could be because they live in totally different regions in these countries. Here, most of them are coastline countries.

While there are few countries where Dreissenidae are higher than Cyprininae at moderate rate which includes Croatia and Brazil. They are located in different continents like one in Europe and another in South America. Here, it can be said that the Dreissenidae population moderately affects the population of Cyprininae. It is also significant because these are the many of few, only two countries with this trend.

Furthermore, bin count Dissimilarity between the countries for both families together was calculated using the Bray-Curtis dissimilarity matrix. The vegdist feature of the vegan package was used to calculate the bin dissimilarity. This matrix provides a measure of

dissimilarity between countries based on their BIN count for both families quantitatively. The matrix was created using the following code:

```
BIN_dissimilarity <- common_countries_data[, c("BINCount.D.x", "BINCount.D.y")]
dissimilarity_matrix <- vegdist(BIN_dissimilarity, method = "bray")
print(dissimilarity_matrix)
```

Then, to visualize the matrix better, a heat map was created which allows to compare the bin dissimilarity between different countries with each other. The following code was used to generate the heat map shown in **figure 3**.

```
heatmap.2(as.matrix(dissimilarity_matrix),
      dendrogram = "row",
      Rowv = TRUE, Colv = TRUE,
      col = colorRampPalette(c("white", "blue"))(100),
      main = "Bray-Curtis Dissimilarity Heatmap",
      labRow = common_countries_data$country,
      labCol = common_countries_data$country)
```

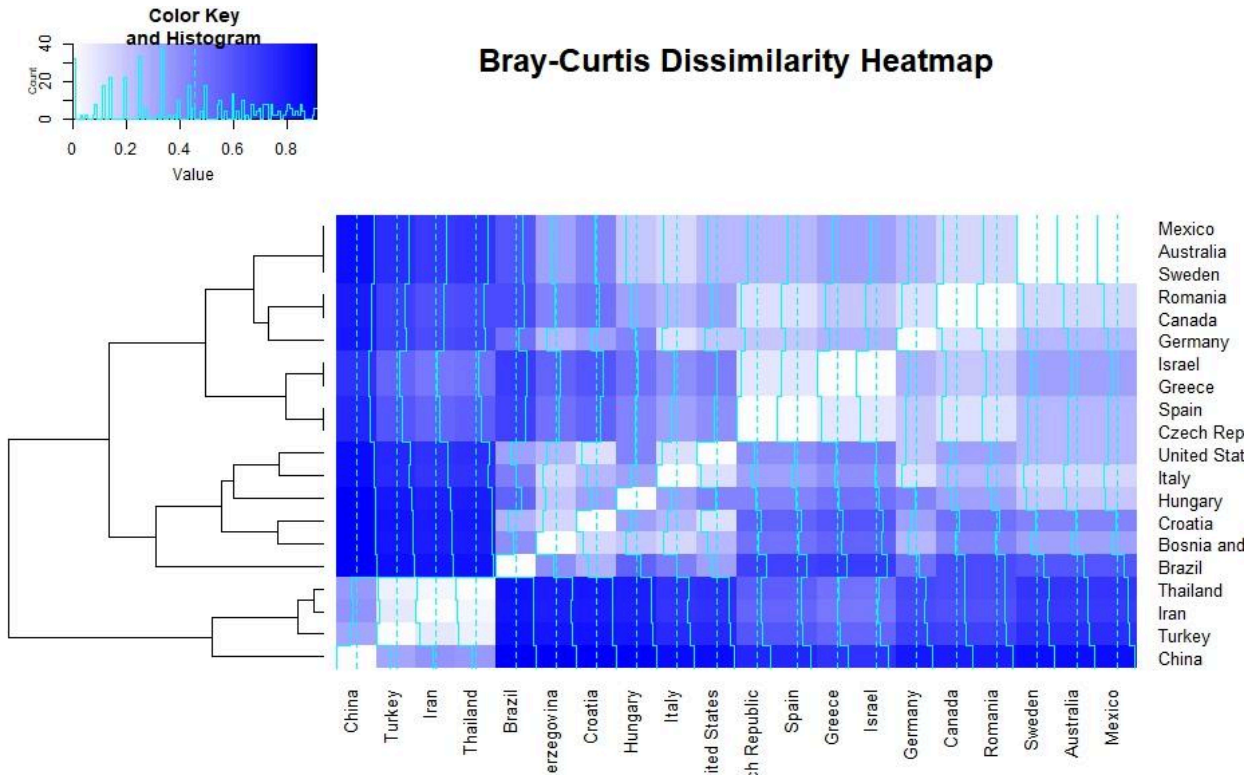

Figure 3: This heat map illustrates the Bray-Curtis dissimilarity between bin counts of countries for the Dreissenidae and Cyprininae families.

As in **figure 3,** each cell in the heatmap represents the dissimilarity in BIN counts between two countries, darker color indicating higher dissimilarity in terms of the bin counts and lighter color indicating less dissimilarity in terms of the bin count. The dendrograms in **figure 3**, help cluster countries with similar BIN count profiles, revealing patterns in the genetic diversity of aquatic organisms across geographic regions. For example, there is a diagonal pattern going through the heat map of white color, those countries have less similarity in terms of the bin count as here country is compared to the country to itself. Another example is that China has high dissimilarity in bin counts with all countries except Turkey, Thailand and Iran.

In conclusion, this study provides valuable insights into the population dynamics of the Dreissenidae and Cyprininae families worldwide. In terms of the Dreissenidae, it has countries with bin-species dissimilarity. There are many countries with about the same numbers of bin counts for both families and a good number of countries where the population of Cyprinidae outnumbers the population of Dreissenidae by a significant number. Study also shows that in some areas two families are unaffected from each other which could be due to the fact that they do not reside in the same site in that country but more research is needed there. It highlights the influence of geographical factors and mutualistic relationships between these families, shedding light on aquatic biodiversity. Further research can study deep into the specific mechanisms underlying various relationships that these two families have and their ecological implications. Moreover, specific sites in each country at which these two families co-exist can be looked closer because here, the country was a very broad term to use but narrowing down to the site will help further to discover the type of relationship these two families have and if there are any environmental factors that specifically are involved in this. Overall, the study showed mixed trends for different countries in terms of the dissimilarity in bin counts and trends in how they coexist with each other. Most significant was that most countries have the same number of bin counts for both families showing that they could be mutualistic but many have one outnumbering others showing in some cases due to influence from something like environment, Cyprininae can outcompete Dreissenidae.

REFERENCES:

Bockrath, Katherine & Wisniewski, Jason & Wares, John & Fritts, Andrea & Hill, Matthew. (2013). The Mussel–Fish Relationship: A Potential New Twist in North America?. Transactions of the American Fisheries Society. 142. 642-648. 10.1080/00028487.2013.763856.

Dreissenidae. Dreissenidae - an overview | ScienceDirect Topics. (n.d.). https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/dreissenidae

Yang, L., Naylor, G. J. P., & Mayden, R. L. (2021, October 8). *Deciphering reticulate evolution of the largest group of polyploid vertebrates, the subfamily cyprininae (Teleostei: Cypriniformes)*. ScienceDirect. https://doi.org/10.1016/j.ympev.2021.107323

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019, November 21). *Welcome to the Tidyverse*. Journal of Open Source Software. https://joss.theoj.org/papers/10.21105/joss.01686