Group 6:

Tiffany Cook

Drashti Mehta

Joshua Mikombo

BINF 6201 Lab Report 5 [Part1]

Question 1: How many SNPs (variants) are in each dataset? How many samples are in each dataset? [3 pts total]

1A. Aripo? (1 pt) **83,777 SNPs & 38 samples**

1B. Guanapo? (1 pt) **57,931 SNPs & 37 samples**

1C. Tacarigua? (1 pt) **26,874 SNPs & 25 samples**

Question 2: How many SNPs significantly deviate from Hardy-Weinberg Equilibrium in each dataset [3 pts total]

2A. Aripo? (1 pt) **12751 SNPs**

2B. Guanapo? (1 pt) **14522 SNPs**

2C. Tacarigua? (1 pt) **8884 SNPs**

Question 3: Of the sites that are not in HWE, how many of them show Excess Heterozygosity (versus a heterozygosity deficit)? Get a list of all of the non-HWE, HET_EXCESS positions for each dataset. [6 points total].

3A. Aripo? (2 pts) **882**

3B. Guanapo? (2 pts) **740**

3C. Tacarigua (2 pts) **1179**

Question 4: How many SNPs are present in all 3 datasets? [3 pts]

- 82895 SNPs in ARIoutName.recode.vcf (aripo)
- 57191 SNPs in GUAoutName.recode.vcf (guanapo)
- 25874 SNPs in TACoutName.recode.vcf (tacarigua)
- **17231 overlapping SNPs in ALL 3 datasets**

BINF 6201 Lab Report 5 [Part2]

Question 5: What are the dimensions (number rows/number columns) of your final VCF file, and do they match up with what you were expecting based on your earlier calculations? (2 pts)

- **110 columns & 17231 rows**

Question 6: How much variation in the data can be explained by Principal Components? [3 pts total]

Question 6A: How much variation is explained by the first 2 Principal Components? (1 pt)

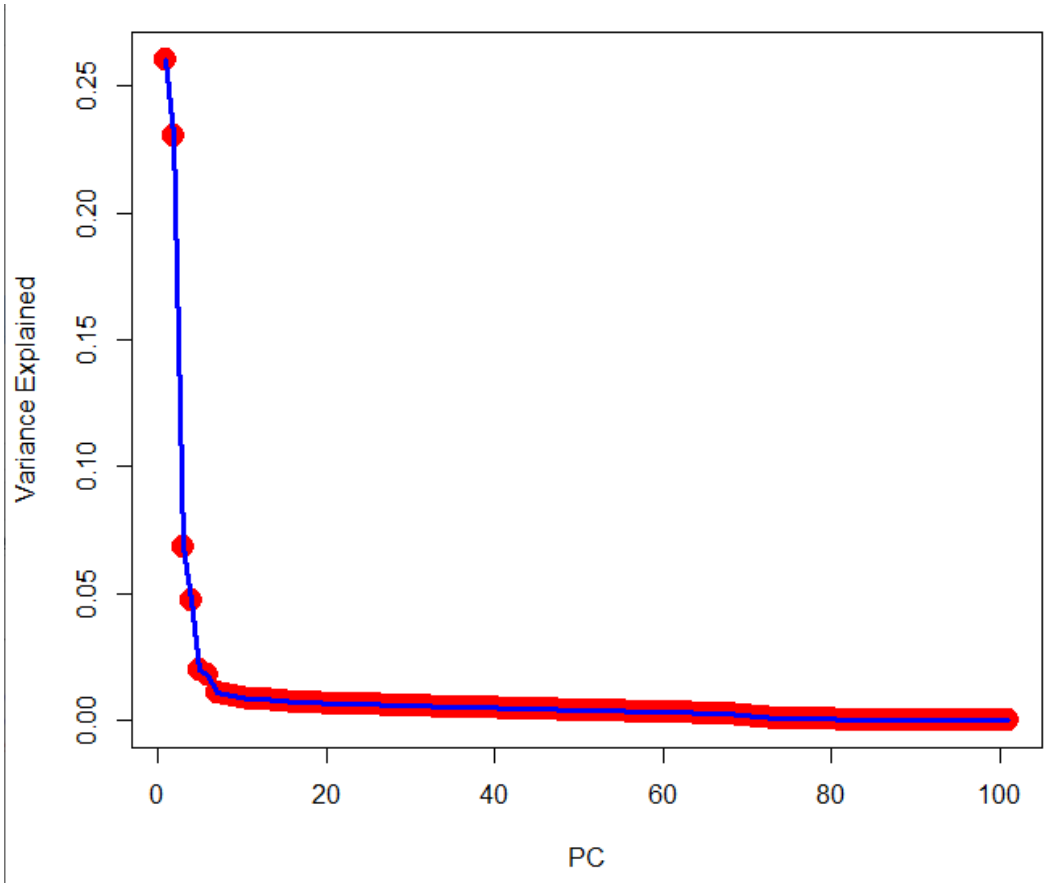
- **The first PC explains 26% of the variance and the second PC explains 23% of the variance**

Question 6B: How many Principal Components do you need to keep in order to explain 90% of the variance? (1 pt)

- **We will consider the first 45 PCs to explain 90% of the variance. It adds up to around 89.9% of the variance.**

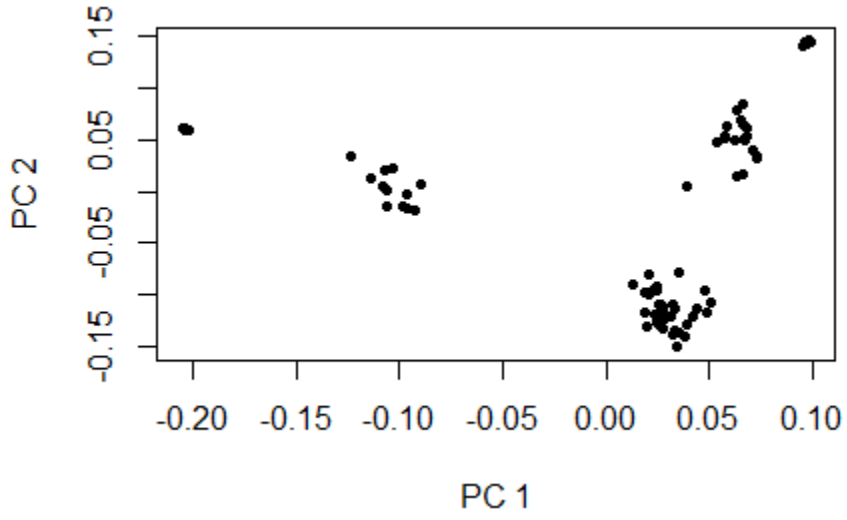
Question 6C: Make a scree plot. How many PCs would you keep based on this plot? (1 pt)

- **Based on the plot, we should keep 6 PCs to explain 90% of variance**



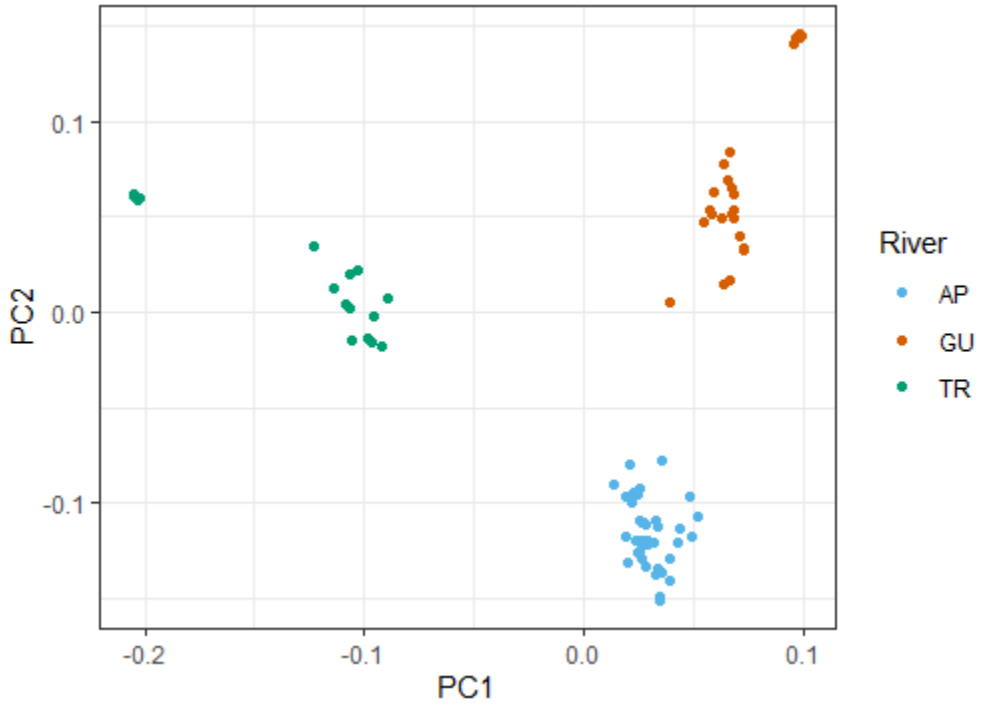
Question 7: What does the PCA clustering look like? [3 pts total]

Question 7A: How many clusters do there appear to be based on your PCA plot (without any color coding)? (1 pt)



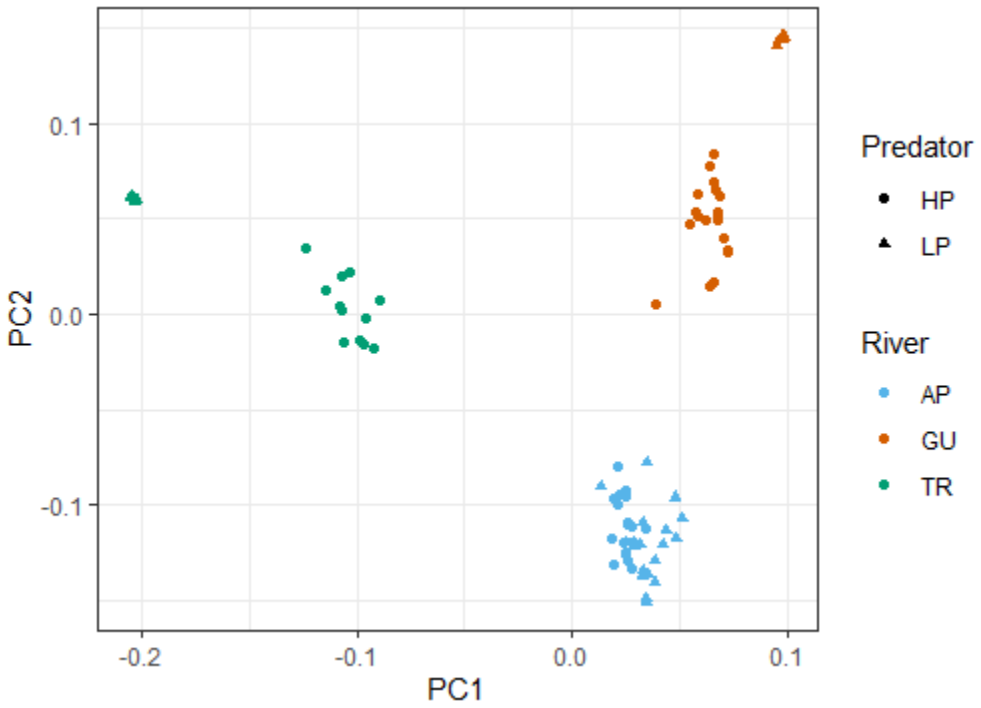
- **Approximately 5 clusters**

Question 7B: Do the clusters appear to correspond to the different rivers? Which river appears the most distinct? (1 pt)



- **Yes, each the clusters correspond to the different rivers with the Aripo river appearing to be the most distinct**

Question 7C: Do the High Predation (HP) and Low Predation (LP) fish form distinct clusters? Are they distinct in all of the rivers, none of them, or just some of them? (1 pt)



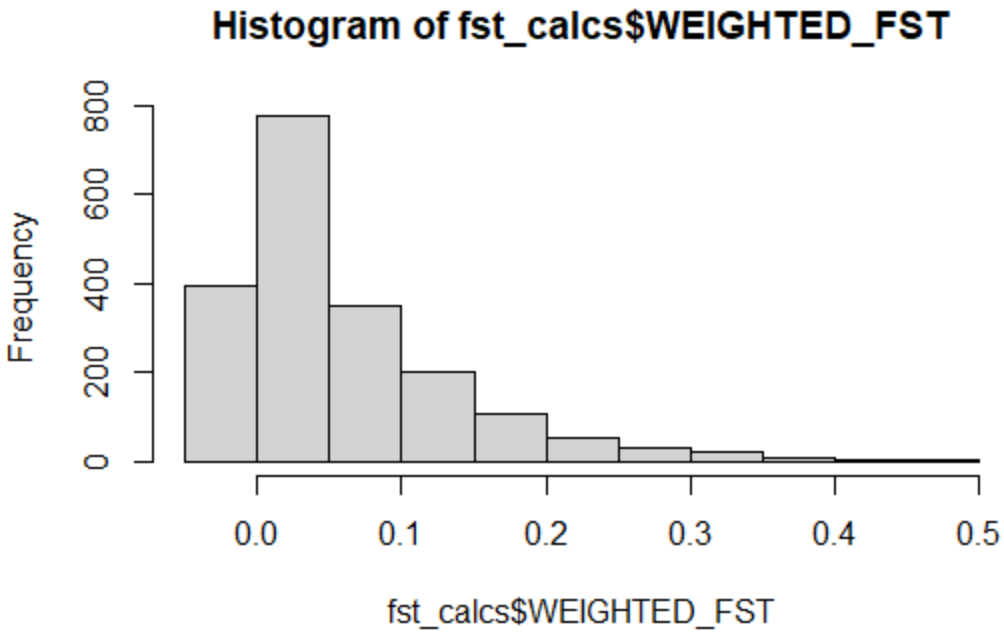
- **The HP and LP form distinct clusters in the Tacarigua and Guanapo rivers. The are not distinct in the Aripo river**

Question 8: Calculate Fst between High and Low Predation Populations [3 pts total]

Question 8A: What is the overall mean weighted Fst score for the whole dataset? Does this score suggest a high, low, or intermediate level of differentiation? (1 pt)

- **The overall mean weighted FST score is 0.0589, or 0.06. This indicates a low level of differentiation. The FST is close to 0. This means that individuals vary just as much within a group as they do between groups- they are close to being identical.**

Question 8B: What does the distribution of Fst scores look like? (1 pt)



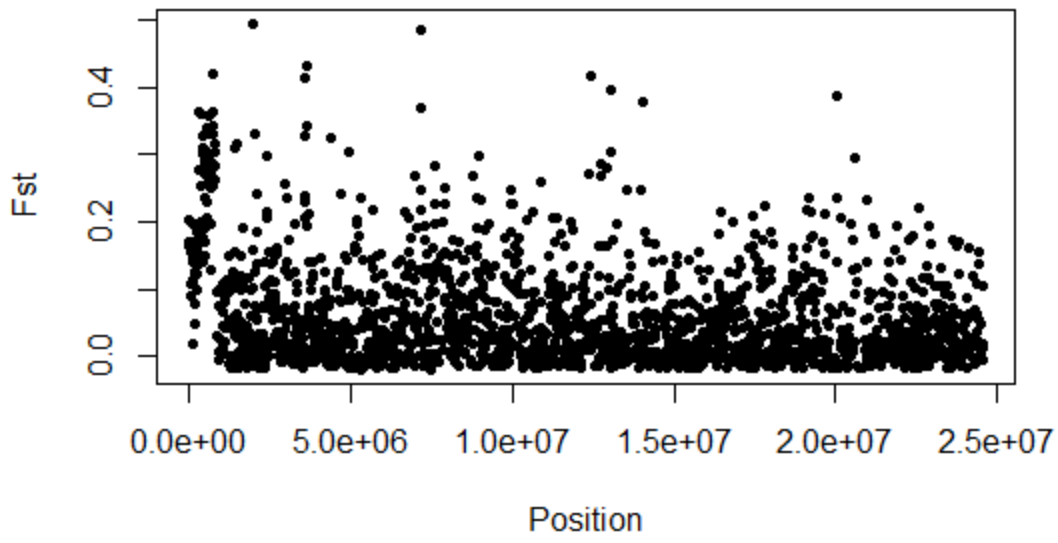
- **Distribution is positively skewed**

Question 8C: On average, which value is higher, mean or weighted Fst? (1 pt)

- **The mean FST score is higher than the weighted FST score by 0.008. The mean FST score is 0.0581 while the weighted FST score is 0.0589.**

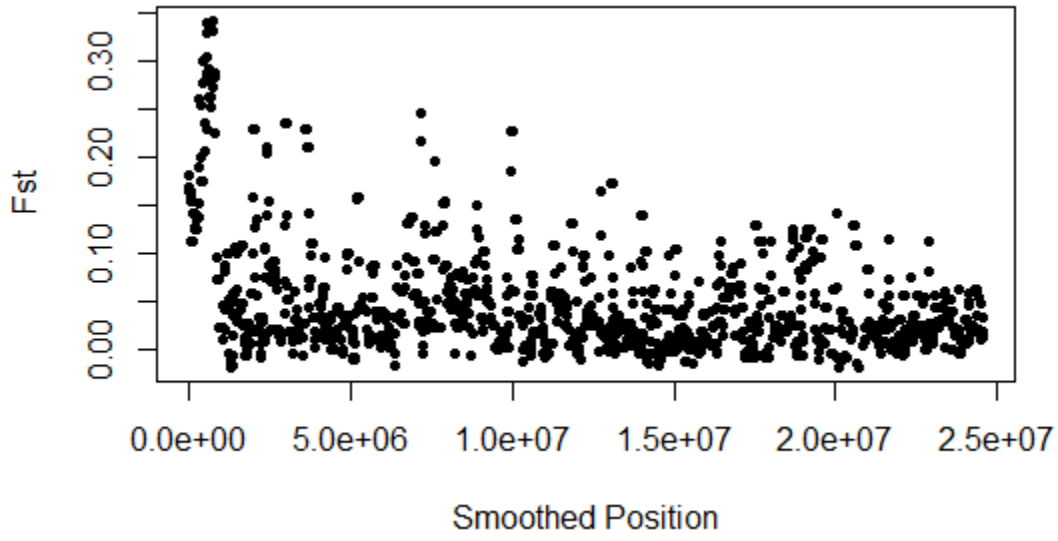
Question 9: Identify highly differentiated genomic regions using Fst [4 pts total]

Question 9A: Plot weighted Fst by position along the chromosome (you can just use the BIN START positions). Is there any region that stands out, or are there high Fst sites all along the chromosome? (1 pt)



- **The positions around 0.0e+0 are the most apparent in this chromosome with high fst sites occurring in that area. Intermediate fst sites occur in later positions, but the fst sites closer to 0e0 are more easily noted.**

Question 9B: Use a smoothing function to try and remove some of the noise, then re-plot Fst along the chromosome. Does any region stand out now? Roughly where is it? (1 pt)



- **The 0.0e+00 region is the one that stands out having the highest Fst values ranging from approximately 0.26 through Fst values over 0.30**

Question 9C: What are the beginning and end positions for the windows with the maximum smoothed Fst values? (1 pt)

- **The maximum smoothed Fst value is 0.341**

- **Beginning: 760001 & End: 770000**

Question 9D: What gene(s) are located in this region, and what is the function of their human orthologs? (1 pt).

- **The cnn3 (Poecilia reticulata calponin 3) gene. The gene is a thin filament-associated protein that is implicated in the regulation and modulation of smooth muscle contraction. It is capable of binding to actin, calmodulin and tropomyosin. The interaction of calponin with actin inhibits the actomyosin Mg-ATPase activity.**