Group 6:

Tiffany Cook

Drashti Mehta

Joshua Mikombo

BINF 6201 - Lab 6

Question 1: Alignment and SNP identification [2 *pts total*]

Perform a Multiple Sequence Alignment using Clustal-Omega. Save the resulting alignment in Clustal format (without character counts). To identify SNPs from this file, upload it to the cluster, and then run the following command:

```
java -jar /projects/class/binf6201_001/jvarkit/dist/jvarkit.jar msa2vcf -o YOUR-OUTPUT-NAME.vcf YOUR-ALIGNMENT-FILE
```

1A. Based on these initial results, would you say that these samples look very distinct or very similar? *There's no right or wrong answer, just state what you think and why you think so.(1 pt)*

- **From the initial results the samples look fairly similar to one another. This conclusion comes from the observation that the chromosome positions in each sample hardly even differ from one another**

1B. What are the dimensions of your VCF file (number of individuals, number of SNPs)?*(1 pt)*

- **66 SNPs (rows) by 57 individuals (columns)**

Question 2: What substitution model is the best fit for the data set? [4 *pts total*]

To perform the model testing and the neighbor-joining analysis, I recommend using R. (*If you run into any issues with R or the R packages, then please let me or Varnika know and we can help you get it going on the cluster*).

```
### Install and load the phangorn library
install.packages("phangorn")
library(phangorn)

### Read in the alignment file
my.aln = read.phyDat(file="YOUR-ALIGNMENT-FILE", format="clustal")

### Perform a model test on all possible models
modelTest(my.aln)
```

```
> model <- modelTest(my.aln)
Model        df  logLik   AIC      BIC
JC          111 -3106.771 6435.541 7039.853
JC+I        112 -3081.237 6386.473 6996.229
JC+G(4)     112 -3087.607 6399.215 7008.97
JC+G(4)+I   113 -3080.591 6387.183 7002.383
F81         114 -2812.33 5852.661 6473.305
F81+I       115 -2786.056 5802.113 6428.201
F81+G(4)    115 -2792.917 5815.834 6441.922
F81+G(4)+I  116 -2785.481 5802.961 6434.494
K80         112 -3083.715 6391.43 7001.186
K80+I       113 -3055.798 6337.597 6952.797
K80+G(4)    113 -3063.897 6353.795 6968.995
K80+G(4)+I  114 -3055.209 6338.418 6959.062
HKY         115 -2775.16 5780.32 6406.408
HKY+I       116 -2739.797 5711.595 6343.128
HKY+G(4)    116 -2753.506 5739.012 6370.545
TPM3        113 -3070.716 6367.433 6982.633
TPM3+I      114 -3044.278 6316.556 6937.2
TPM3+G(4)   114 -3051.222 6330.443 6951.088
TPM3+G(4)+I 115 -3043.585 6317.171 6943.259
TPM3u       116 -2768.811 5769.623 6401.156
TPM3u+I     117 -2737.421 5708.843 6345.82
TPM3u+G(4)  117 -2748.045 5730.091 6367.068
TPM3u+G(4)+I 118 -2737.643 5711.286 6353.707
TIM1e       114 -3079.501 6387.002 7007.646
TIM1e+I     115 -3051.024 6332.047 6958.136
TIM1e+G(4)  115 -3059.528 6349.056 6975.144
TIM1e+G(4)+I 116 -3050.688 6333.375 6964.908
TIM1        117 -2774.387 5782.774 6419.751
TIM1+I      118 -2738.967 5713.935 6356.356
TIM1+G(4)   118 -2752.829 5741.659 6384.08
```

2A. According to the AIC, which model is the best? *(1 pt)*

> which(model$AIC==min(model$AIC))
> model$Model[50]

- **TPM3u+I**

2B. According to the BIC, which model is the best? *(1 pt)*

> which(model$BIC==min(model$BIC))
> model$Model[14]

- **HKY+I**

2C. Briefly describe the assumptions and parameters of the best model according to the BIC (e.g., what base frequencies, how many substitution rates, do rates vary across the sequence, etc.) (*1 pt)*

https://evomics.org/resources/substitution-models/nucleotide-substitution-models/

- **The Hasegawa-Kishino-Yano (HKY) 1985 Substitution Model. The model assumes that all substitution rates are equal. It has unequal base frequencies.**
- **The "I" here mentions that there are some sites which do not change at all.**

- **It has 2 substitution rates, one transition rate and one transversion rate. These rates vary.**

2D. Which model available in the dist.dna() function is most similar to your best model according to the BIC? *(1 pt)*

- **The F84 model**

Question 3: Construct a Neighbor-Joining Tree for your data set [5 *points total]*.

You can use the R code below to get a neighbor-joining tree for your data:

```
### Calculate distance under your best fit model
my.dist = dist.dna(as.DNAbin(my.aln), model="MODEL YOU WANT TO USE", gamma=TRUE)

### Create a neighbor-joining tree
my.nj = nj(my.dist)

### Get a basic plot in R
plot(my.nj)

### Write tree to an output file
write.tree(my.nj, file="bees-NJtree.nwk")
```

3A. Provide the plot of your basic NJ tree from R. *(1 pts*)

3B. Based on your tree, are there any distinct groups that might be cryptic species? Do they correspond to anything we know (based on sample names) about color patterns or geographic locations? *(4 pts)*

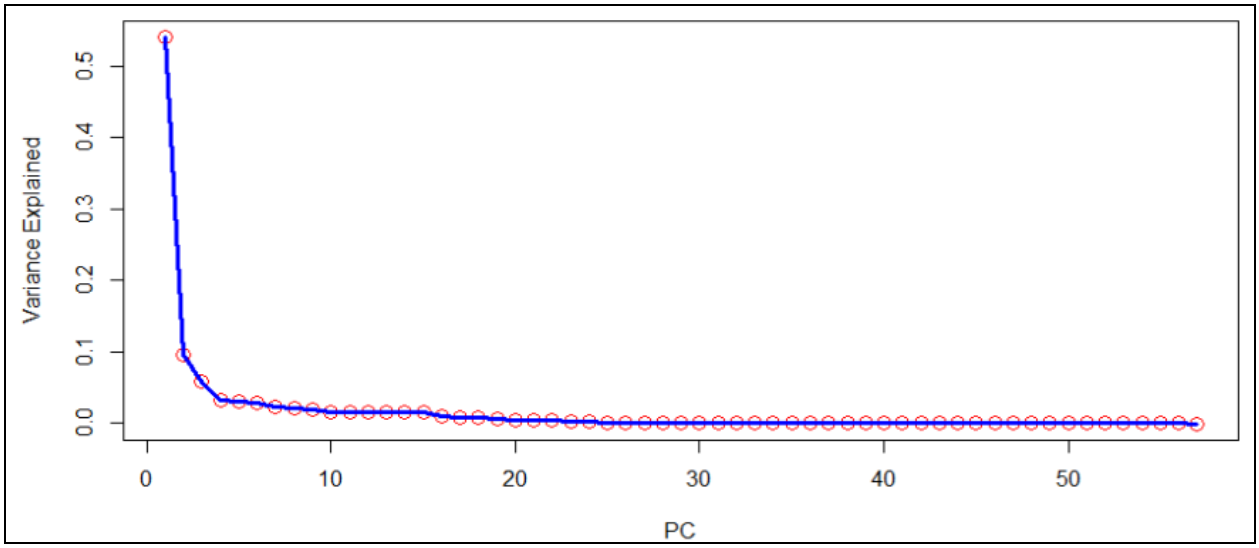
- **Yes, there groups of that are distinct which are the bees from Utah, Montana, New Mexico, California, and South Dakota.**
- **They are all color morph 4 and are the most distant from all other species as displayed on the tree**

Question 4: PCA Analysis [4 *pts*]

Using the VCF file you created in Question 1, perform a PCA analysis on the data (use your notes from the last lab to remember how to do this).

4A. How many eigenvectors did you estimate for the data set? Why did you choose this number? *(1 pt)*

- **We estimate 4 eigenvectors for the dataset and this estimate came from the scree plot we produced**



4B. Does the PCA show a clustering pattern that matches up with the pattern you see in Question 3? Again, can you determine if there is anything related to phenotype or location that might correspond to any pattern of clustering? *(3 pts)*

- **No, the clustering pattern of the PCA doesn't match up with the pattern we see in question 3. The reason being is because the PCA clustering pattern is based-off phenotypes (4), whereas question 3's clustering pattern was based on the the genotypes**