## Questions

Start by downloading the FASTA file with the sequence of each individual:

[bee-input.fa](https://uncc.instructure.com/courses/192992/files/20749743?wrap=1)

### 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)*

They look relatively similar as each individual looks generally similar across the samples.

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

#### Number of individuals - 57 Number of SNP’s - 67

### 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)

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

TPM3u+I is the best model according to the AIC.

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

HKY+I is the best model according to the BIC.

#### 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)*

HKY allows base frequencies to vary and are unequal to each other. There are two substitution rates: transitions and transversions and they have different rates. The parameter +I describes invariable sites where the model expects some unchanging proportion of the sites in the sequence.

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

F84 is the most similar model to our best model according to the BIC.

A lot of times, the exact model that you want based on your test won't be available in the distance calculation or tree-building program, or it may be there but go by a different name. To see what models are available in R (and their descriptions), type ?dist.dna() into the R command prompt. Then see which might be most similar based on your answer to 2C.

### 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)

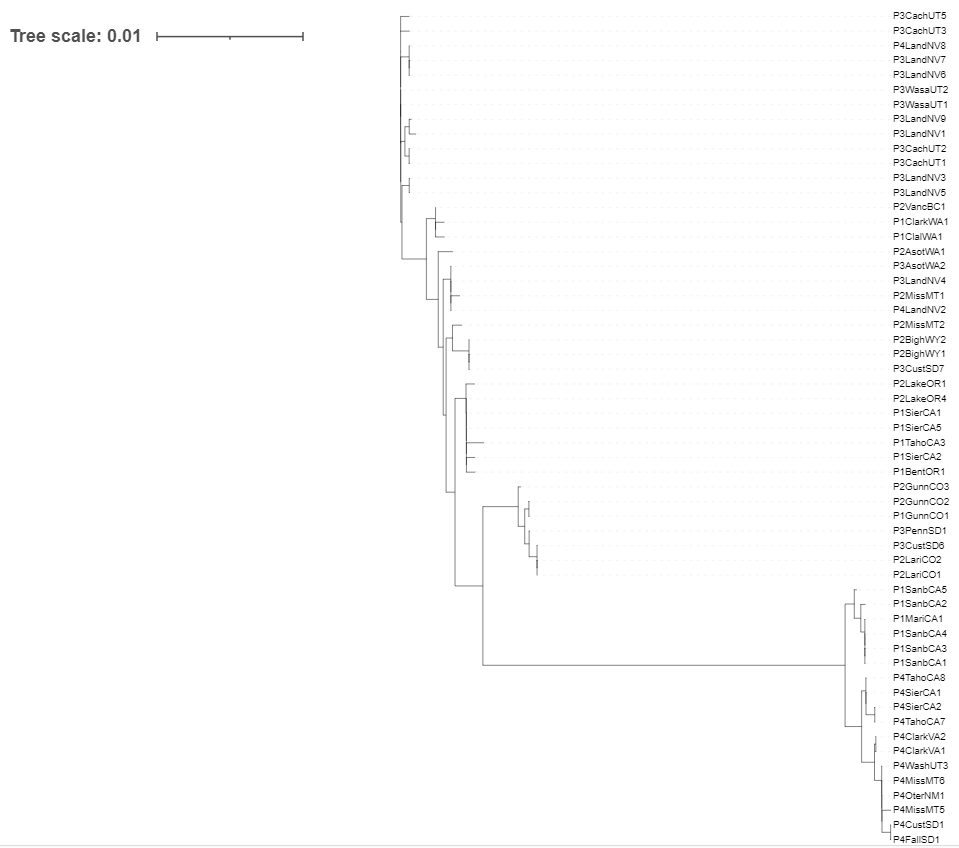
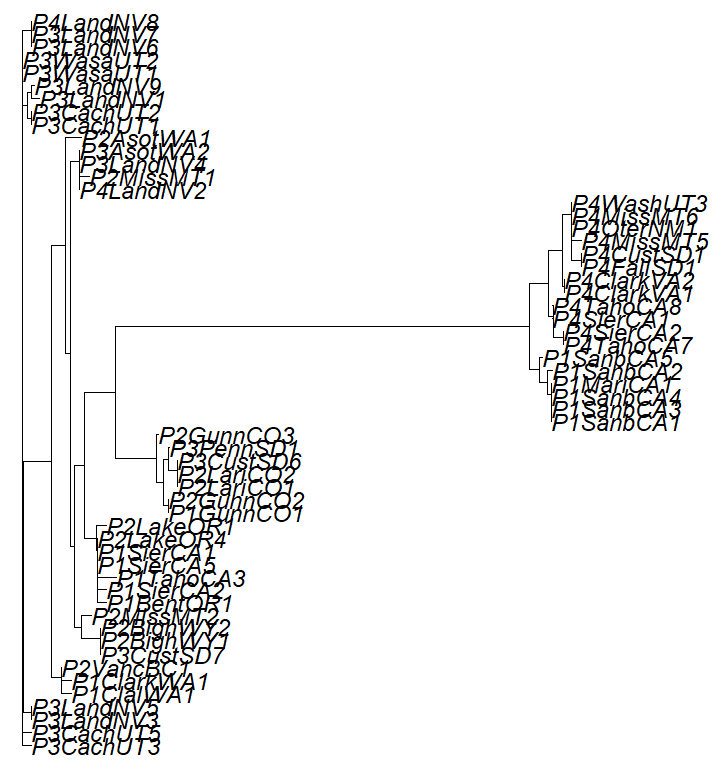
### Get a basic plot in R

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)*

P1 and P4 seem to be the most cryptic groups based on their respective locations in the tree, being the farthest clades from the root. Non-random mating in phenotype P4 appears to show a preference for that color type. In P1, it appears that there are more geographical implications in the divergence.

For this question, I recommend looking into some alternative plotting programs where you can manipulate the basic R plot and annotate it (this might make it easier to see if there are any patterns in the data). Using the ".nwk" tree file you made from R, you can use almost any existing phylogenetic tree plotting program. One option is the interaction tree of life: <https://itol.embl.de/>

Another option is the program FigTree: <https://github.com/rambaut/figtree/releases>

If you're more comfortable working in R, you can try the ggtree package: <https://4va.github.io/biodatasci/r-ggtree.html>

Spend a bit of time trying to get a tree that best demonstrates how you want to answer this question (and note that the answer may not be clean or obvious, as this is real data!).

### 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)*

There should be 57 eigenvectors for each principal component created for each individual.

#### 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)*

In our PCA plot, PC1 is the axis that shows the most variation among the bumblebees. This reflects the data that we observed in our NJ tree. The variation in (P1) appears to be due to geographic isolation, as most of P1’s are clustered in the same region of California. P4’s seem to be a subpopulation created by assortative mating (non-random), as they appear to have a preference for a specific phenotype and they are located in different regions in the U.S.