HHMI: DNA Polymorphism

6 tasks you need to complete

1: Examine DNA Sequences and to verify high quality sequences: use *FINCH* TV
2: Make a **FASTA** file of good DNA sequences
3: *BLAST* the sequences to verify species identity.
4: Align the data using *CLUSTAL W*
5: Use the alignment and make an evolutionary tree.
6: Use the aligment and use *DNASP* to quantify the variation within and among populations.

The most important data for your **Extended Abstract** is the sequence variation and populations affect this variation.

1 QC: DNA SEQUENCE

FINCH TV is used to inspect and choose good sequences.

Each sequence has a chromatogram that show the intensity of each fluorescently labeled nucleotide: A = Green, C= Blue, G = Black, T = Red.
The peak height and separation are good indications of higher quality sequence

- We need to choose the same "good sequence" from each sample. We will use FINCH TV.
  Open each sequence and inspect the quality.
  Do most or all the sequences have "good sequence" starting near the same nucleotides?
  Copy and paste the similar sequences in to "simple text" file.
  -- each file

  >Name1
  ATGAACTGACTGATGA………………………………………………………………
  ……………………………………………………………………… ACTGACTGAATGCA
  >Name2
  ATGAACTGACTGATGA………………………………………………………………
  ……………………………………………………………………… ACTGACTGAATGCA

  THUS, YOU SHOULD HAVE A SIMPLE TEXT FILE WITH ALL OF YOUR SEQUENCES, EACH NAMED WITH "> name…" Keep name simple and less than nine characters.
  THIS IS YOUR FASTA FILE. (**YOU MUST HAVE A FASTA FILE).**
- **Save this file as a simple text document: Group#_FASTA.txt**

2: DIVIDE WORKLOAD:

We need to divide the work into two groups: 1) Species identification and 2) Evolutionary analysis.

1) Species identities: Are your sequences really for your organisms?
Three questions: Are they human?
Are they your group or fish, insect?
Are they your species (this only works some of the time).
2) Evolutionary analysis: You need to align all your sequences. Make an evolutionary tree and define the sequence variation within and among populations.

3: SPECIES IDENTIFICATION BLAST AGAINST NCBI

- ➢ BLAST all of your sequences using NCBI Blastn

  http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Nucleotides&PROGRAM=blastn&BLAST_PROGRAMS=megaBlast&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on
  … In *ENTER QUERY SEQUENCE:* **CHOOSE** your simple text file with all you sequences.
  … In *CHOOSE SEARCH SET*: **CHOOSE** Nucleotide Collection
  BLAST
  Output has each sequence listed separately. What sequence is the best match?
  RECORD the best hit for each sequence.

THE "SPECIES IDENTIFICATION GROUP" SHOULD HAVE A RECORD OF EACH SEQUENCE AND HAVE IDENTIFIED SPECIES FOR EACH SEQUENCES.

4: EVOUTIONARY ANALYSIS,

What is the relationship among sequences? How much sequence polymorphism is there within populations? How different are the populations?

THREE (3) steps:

1) align -> **Get two output** phylip output (online software), and Clustal Alignment.
2) Make a phylogenetic Tree (online software)
3) Quantify the DNA sequence variation within and between populations (DNaSP on portable)

1. The first thing is to align all of the sequences. We are using CLUSTAL W
   http://www.ch.embnet.org/software/ClustalW.html

   From your simple text FASTA file copy of all your sequence and paste into "**INPUT Sequence"** box. Click "**Run Clustal W".**

   A new window (output window) should appear.
   You want to save "Clustal Alignment" for your paper and poster. Click **"clustalw.aln"** and save output.
   Return go back, to output window. Save "phylip". When new window save output as "**X.PHY"**.
2. Phylogenetic Analysis. The goal is to get a simple tree that describes the relationship among sequence. You can you use the tree in your poster and final abstract.
   By going to
   http://mobyle.pasteur.fr/cgi-bin/portal.py?form=phyml
   PASTE your PHY file into program, Run the program and Copy TREE file

   You can Plot your tree file at
   http://www.bioinformatics.nl/tools/plottree.html
   This is the tree file for your poster or abstract.

3. DNaSP

DNaSP, software to define DNA sequence within and among populations

To define the number of polymorphic loci, Open DnaSP.
File open your Phylip file. If it works you should see a list of information. Some of it is incorrect (e.g., chromosome location).
In DATA, **Assign Genetic Code** (click assign genetic code). In new window choose mitochondrial genome.
IN DATA, **Define Sequence set** (click define sequence sets). In new window,
-- choose population one sequences, >> (move sequence to other box),
--- **Add new sequence set**.
-- repeat for other populations.
-- **UPDATE ALL Entries.**

To get information:
--- in **Analysis** choose, "**Gene Flow and Genetic Differentiation**"
-- either "**Save Current Output"** or record for each populations
Number of haplotypes (for each population) this is the number of unique sequences
Nucleotide diversity Pi = X? (for both populations), where Pi is the average number of nucleotide differences per site between two sequences, or nucleotide diversity
Fst, which is the Fixation index (See below), 0= no difference, 0.1→ 0.45, meaningful difference among population, 0.45→0.7 large differences (different species??), > 0.7 are these different species

AT the END of the Day You should have:

- Species Identifications
- A Chromatogram of your sequences (good for poster)
- A Tree that provides a visual indication of the sequence similarity among individuals
- The sequence variation, Π (Pi), for both populations
- Fst the measure of genetic distance between populations.

Fst, Wikipedia, http://en.wikipedia.org/wiki/Fixation_index

FSTis a measure of population differentiation based on genetic polymorphism data.

FST is simply the correlation of randomly chosen alleles within the same sub-population relative to that found in the entire population. It is often expressed as the proportion of genetic diversity due to allele frequency differences among populations.[1]

$$Fst = (\Pi between - \Pi within)/ \Pi between$$

That is Fst is the difference Π for between and within the population divided by the Πbetween , where Π is the average number of pairwise differences between populations. Or the relative differences in sequence polymorphism, where 0.2 would be meaningful.