Laboratory excercise written by Andrew J. Crawford <andrew@dna.ac>
with the support of CIES Fulbright Program and Fulbright Colombia.
Enjoy!

Filogeografía

BIOL 4211, Universidad de los Andes, Bogotá. Coord.: -074.0657, 04.6012
25 de enero a 01 de abril 2006

Lab 5
F_{ST} & *Nm* using *dnaSP* v. 4.0
and
haplotype networks by statistical parsimony using *TCS* v. 1.21

25 febrero 2006

NB: Always download the latest version of all software
and read carefully the latest accompanying documentation.

Today we have two goals. First, you will calculate pairwise F_{ST} and *Nm* for the same pairs of populations you analyzed last week using *IM*. Calculations will be made using *dnaSP*, which you used two weeks ago to calculate your summary statistics and test for population expansion. Recall, that *dnaSP* is available only for the Windows OS (though I run it on my Mac using *VirtualPC* software).

Second you will estimate one or more haplotype networks from your DNA sequence data using statistical parsimony, i.e., the network will represent a 95% probable set of the most likely networks. You will estimate these parsimony networks using *TCS* version 1.21 which is a Java program so it runs on any operating system that supports Java (e.g., Windows, Mac, and many flavors of Unix). Any single network may also be regarded as an unrooted gene genealogy in which the ancestors are allowed to persist. When these gene genealogies include multifurcations or reticulations they are called networks. Reticulations are represented as closed loops within a network and may be caused by recombination, homoplasies ("finite sites"), or ambiguities in the inferred history of the haplotype data.

Software

dnaSP version 4.10.4 (for Windows OS only)
`http://www.ub.es/dnasp/`
TCS version 1.21 (Java program, runs on Windows OS, Macintosh OSX, Linux and other types of Unix)
`http://darwin.uvigo.es/software/tcs.html`
Java Virtual Machine version 1.4.2
`http://java.sun.com/`
`http://java.sun.com/webapps/getjava/BrowserRedirect`

Citations

Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496-2497.

Clement M, Posada D, Crandall KA (2000) TCS: a computer program to estimate gene genealogies. *Molecular Ecology* **9**, 1657-1660.

Templeton AR, Crandall A, Sing CF (1992) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* **140**, 767-782. (*Explains "probability of parsimony" used to construct network that represents "95% set of plausible solutions*.")

Posada D, Crandall KA (2001) Intraspecific gene genealogies: trees grafting into networks. *Trends in Ecology and Evolution* **16**, 37-45. *(A review of networks: concepts and reconstruction methods.)*

Further information

Throughout the first part of this lab, recall that *dnaSP* has an excellent Help menu, accessed via the dropdown menu on the far right of the menu bar. The Help pages can assist you with running the software as well as background information and citations concerning each type of analysis. Further information on *dnaSP* can be found in the handout from the third lab of this course.

For help with *TCS* open and consult the documentation file called "TCS1.21.pdf" included with the program.

GOALS:

Calculate F_{ST} and *Nm* parameters for the same populations you analyzed using *IM* last week.
Compare your pairwise F_{ST} and *Nm* results with the results from *IM*.
Understand the similarities and differences behind the two approaches to population structure analysis.
Interpret the similarities and differences between the results obtained by the two methods.
Construct a statistical parsimony network.
Save this network and later we will apply Templeton's nesting procedure as our second step toward completing a Nested Clade Phylogegraphic Analysis.

I. Preparing your data for *dnaSP*.

Hopefully, you have your same NEXUS file from Lab 3, including the `TaxSet` commands (in a `SETS` block) to partition your samples into populations. Your NEXUS file might also include a `dnaSP` block:

```
begin DNASP ;
       CHROMOSOMALLOCATION= mitochondrial ;
       GENOME= Haploid ;
END;
```

If you do not have this block, be sure to format your data correctly in *dnaSP* using the Data > format... command.

If you have multiple mitochondrial loci, you should calculate F_{ST} and estimate *Nm* using the combined data set. Mitochondrial genes are presumably completely linked and therefore share the same evolutionary history. In *dnaSP* mitochondrial and nuclear genes should be analyzed independently. If you have multiple nuclear genes, these may be analyzed either independently or as a combined "supergene" sequence. To compare your results with *IM*, you might want to combine all your nuclear gene sequences.

II. Calculating population structure and estimating migration in *dnaSP*.

Open your NEXUS file in *dnaSP* as before. Recall that the first window will be the Data Information window. Review the data info to confirm that the data were read properly and the genome and ploidy are set correctly. Make any necessary corrections using the Data > format... command.

To obtain pairwise F_{ST} calculations and and *Nm* estimates, select the Analysis > Gene Flow and Genetic Differentiation option. Region to Analyze should include all your data of interest. Under "Sites with Alignment Gaps are ..." select excluded. (Version 4.10.4 of *dnaSP* has a disquieting bug that reports zero segregating sites when you select excluded only in pairwise comparisons, although the results are the same). Click (or "activate") the box that says Perform the Permutation Test and increase the number of replicates to 5000. Select which pair of populations you want to analyze by clicking on Include / Exclude Populations ... . Select all the populations in the Included Populations box and move them all to the Excluded Populations box using the << button. Now select just your desired pair of populations and click >> to move these two populations back to the Included Populations box, and select OK.

III. Reading the population structure and migration output from *dnaSP*.

After the analysis is completed, two output windows will open. In the horizontal gray and white table, you will see some familiar statistics on population divergence such as *K*xy (= π_D), G_{ST} (Nei 1973), and F_{ST} (Hudson, Slatkin, and Maddison 1992). These three statistics should be familiar to you by now. Hudson, Slatikin & Maddison (1992) defined $F_{ST} = 1 - (\pi_S / \pi_B)$, which given $\pi_B = \pi_S + \pi_D$, we can re-write as $F_{ST} = \pi_D / (\pi_S + \pi_D)$ using the notation of Charlesworth (1998). G_{ST} is defined as $(H_T - H_S)/H_T$, where $H = 1 - \Sigma p_i^2$ with p_i = frequency of the i^{th} haplotype, and H_T & H_S refer to heterozygosity of the Total combined samples & each Subpopulation, respectively.

The blue output window is divided into three blocks of text. The first provides basic summary statistics on genetic variation within each population and for the combined data. The statistics include *n*, *S*, *h*, and *K* ($\approx \pi L$), which should be familiar to you by now. The second section is labeled Genetic Differentiation Estimates and provides the results of the permutation test. These tests address the question of whether the observed F_{ST} is significantly greater than zero. Use the final line, the Z* result, to find the *p*-value of your F_{ST} estimate. The final section is labeled Gene Flow Estimates and provides us with *Nm* based on G_{ST} (Nei 1973) and *Nm* based on F_{ST} (Hudson, Slatkin, and Maddison 1992).

IV. Reporting your results of population structure and migration from *dnaSP*.

Make a table of your results with one row for each pairwise population comparison, following the format below, but of course use your own proper population names in column one.

	π_D (= *K*xy)	G_{ST}	F_{ST}	*Nm* from G_{ST}	*Nm* from F_{ST}
Pop X vs Pop Y					
Pop 2 vs Pop 3, etc.					

Note, that F_{ST} is based on genetic distances among haplotypes, while G_{ST} uses only the frequency of haplotypes.

Please answer the following questions:

(1) Explain one similarity and once difference between F_{ST} and G_{ST} (hint, they use different aspects of the data when applied to DNA sequences. Are F_{ST} and G_{ST} very different or very similiar for your data? Why? Which measure of population structure do you prefer for your DNA sequence data?

(2) Compare your estimates of migration rate obtained from *IM* with your estimates from G_{ST} and F_{ST}, above. Did the summary statistic method (above) give you very similar or different results to *IM*? If they are different, please explain why.

PART TWO: *TCS*

V. Preparing your data for *TCS* version 1.21.

You will want to analyze all mitochondrial genes together. If many individuals are missing one or more gene sequences, you may want to run two analyses, one combined and one with individual genes. (Probably it would be better to lose some DNA sequences than lose some individuals.) Independent nuclear genes should be analyzed separately. In the case of linked nuclear genes, such as multiple ITS genes, you have the option to combine data but only if your data pass the 4-gametes test (see Lab 3).

For data sets with few sequences (e.g., 20 or 40 individuals), you may simply input your whole data set into *TCS* and the program will place very divergent sequence in independent networks automatically. For larger data sets, this will be a pain in the neck, as you will soon see. If you have many individuals and you know, *a priori*, that you have very divergence clades, you should prepare separate NEXUS files for each clade (see below). For example, different species should be inputted into *TCS* as individual NEXUS files.

TCS accepts NEXUS files, but does not accept all possible options. For example, you cannot have [ ] inside of the data MATRIX. To make life easier, probably you should execute your data in *PAUP** and then export the data

(in non-interleave format) such that all extraneous text will be removed. *TCS* I believe will NOT accept data in interleave format.

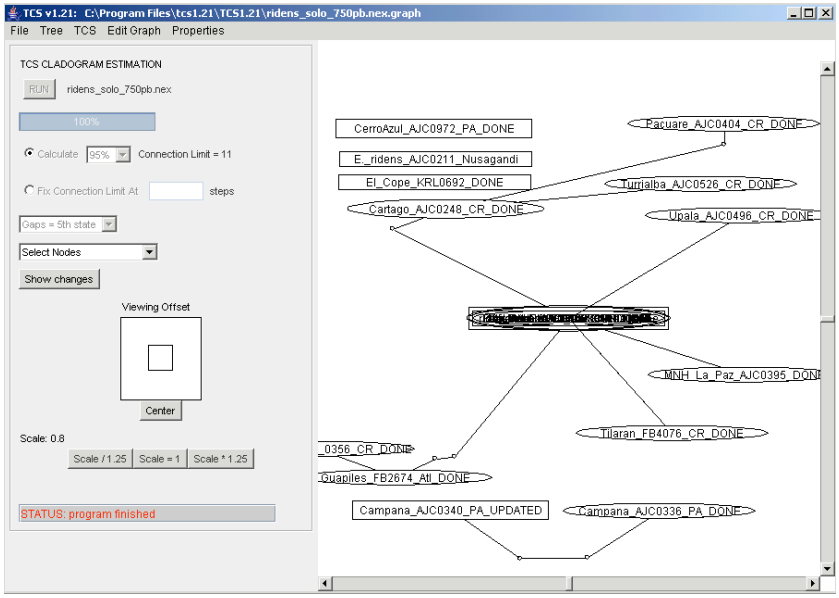
VI. Running *TCS* version 1.21.

Double click on TCS1.21.jar to launch application. Go to: File > Select NEXUS/PHYLIP Sequence file, and then select your data file. In the *TCS* window, confirm that this option is selected: Calculate 95% Connection Limit. Hit the "RUN" button!

VII. Organizing the output of *TCS* version 1.21.

The good news is that *TCS* made one or more networks for us! The bad news is that *TCS* piled nearly all the nodes into a heap. Now we have to select them one-at-a-time, and drag them off the pile and spread them out on our display screen.

1. Where it says Select Nodes and Branches change this option to just Select Nodes
2. The Scale buttons allow you to zoom in and out in the view window. You can also move around the network by grabbing with your mouse the little square inside the Viewing box.
3. Start by clicking on the middle of the "pile" and dragging of to the side whatever node you've selected.
4. Repeat this process many times, while attempting to keep all the clades spatially organized! Good luck!
   Here is what my *TCS* window looks like after 19 grabs at the pile of nodes in the center:



5. Why are some samples not connected to the network with the other samples? Note, where is says Connection Limit = there should be a number now (perhaps 11 or 16). Any samples more than 11 or 16 steps (mutations) away are NOT connected to the same network. Instead these genetically distant samples form their own independent networks. If you want more samples combined into fewer networks, you might try changing this option to Fix Connection Limit At ___ steps, and in the blank put a larger number of steps that currently displayed in Connection Limit = (*e.g*, 16? 32?).
6. If you want more information about each node, *e.g*., how many sequences each node represents, double click on the node (especially the labeled ones). In this new window that pops up, where is says Data, change this drop-down menu selection to Frequency and see what additional samples, if any, share this same haplotype.

VIII. Saving the output of *TCS* version 1.21.

Print your networks or else draw them by hand on a sheet of paper. Do not bother with the samples not included in any network. Also, save your network both as GML (Graphic Mark-up Language) and PICT (picture) file. We will need these next week when we continue with NCPA and *GeoDis*.

For your homework for this lab, you need only turn in the F_{ST} table above, along with your answers to the questions below the table.