

Genepop

How to estimate F_{is} , F_{st} and F_{it} ?

1. Go to Option 6: F_{st} & other correlations
2. Choose Allele identity (F-statistics)

Submit file: *Sturgeon_Genepop_input*

Results: file *Genepop_Fst_W&C*

Discussion Question: Why do we want to use both? What is the better one?

How can we test for population differentiation?

→ **Exact Tests**

1. Go to Option 3: Population differentiation
2. Choose Genotypic differentiation for all populations (when they are not in HWE)

Submit file: *Sturgeon_Genepop_input*

Results: file *Genepop_exacttest_genotypic*

Note: We should apply the Bonferroni correction on each p-value obtained for each locus

Discussion Question: Should we take out population CAN_{hat} (not in HWE) and do the genic differentiation? Is the genic more powerful than the genotypic method?

FSTAT – software for calculating F-statistics, exact tests and more...

1. Open the FSTAT program. As the program opens, it may or may not ask you to provide random numbers for initialization of statistical tests. I think it only might do this the first time you open the program on a particular computer.

1a. Before you can open your file in FSTAT, you need to convert it to the proper input format. CONVERT cannot directly create an FSTAT file for you, but it can convert your data to a GENEPOP format, and FSTAT will convert from GENEPOP to FSTAT format.

2. To convert your file from GENEPOP to FSTAT format, go to the Utilities pull-down menu and select “File Conversion” (Utilities → File Conversion → Genepop → FSTAT).

3. Select the GENEPOP file you want. FSTAT might ask you under what name you want to save the FSTAT converted file. (It may just create a converted file with the same name and a .dat extension). If it asks for a name, enter in an informative name and hit “Save.”

4. Go to the File pull-down menu and select open. Select your newly converted FSTAT file. Notice that after you open the data file, you can now manipulate the analysis options on the program interface.

3. Select the following options: Under the menu “Global Statistics” choose Weir and Cockerham F-statistics. This will allow you to estimate Fis, Fit, and Fst. To perform an exact test, look under the menu “Testing – Population Differentiation” and choose Test NOT assuming HW within samples. Set the number of permutations (1000 is suggested if you have fewer than 10 loci). Also choose a “Nominal Level for Multiple Tests” (5/100), which is selecting 0.05 as your alpha level. The program will use this as your baseline alpha when it makes a Bonferroni correction on your data.

4. Hit the “Run” button. Your output file will be located in the same directory of your input file and will have the same name (with an OUT extension). This output file will give you an estimation of Fis, Fit, and Fst for each locus and over all loci, the 95% Confidence Intervals (CI) for F (Fit), Θ (Fst) and f (Fis) which are estimated by bootstrapping over loci. Results for the exact tests are also included. If the 95% CI around your Fis, Fit, and Fst estimates includes zero, than your estimate is not significantly different from zero. If the CI does not include zero, you can say the estimate is significant.

Discussion Question: What are the differences between CI and Exact Tests? Are these ones more powerful?

Some references:

Cockerham CC and Weir BS, 1993. Estimation of gene-flow from F-statistics. *Evolution*. 47:855-863.

Excoffier L 2001. Analysis of population subdivision. In *Handbook of statistical genetics*, Balding, Bishop & Cannings (Eds) Wiley & Sons, Ltd.

Goudet J, 1999. An improved procedure for testing key innovations. *American Naturalist*. 53:549-555

Goudet J, Raymond M, Demeus T and Rousset F, 1996. Testing differentiation in diploid populations. *Genetics*. 144:1933-1940

Nei M, 1973. Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. USA*. 70:3321-3323.

Raymond M and Rousset F, 1995. An exact test for population differentiation. *Evolution*. 49:1280-1283

Slatkin M, 1995. A measure of population subdivision based on microsatellite allele frequency. *Genetics*. 139:457-462

So you want to estimate pairwise Fsts in GDA...

1. Open the data file (File → Open) in GDA.

2. Start a log file (File → Log – save under informative file name). Now all your results will be saved in an output file.

2. Go to Distance pull down menu and click on Options (Distance → Options). A dialog box will appear that will give you several choices of what type of analysis you want to

perform. The top of the box will ask you about what type of genetic distance metric you want to calculate. **Coancestry identity is equivalent to the Weir & Cockerham (1984) pairwise Fst (see Bruce Weir's Genetic Data Analysis II p 445 for details)**. This is an easy way to estimate pairwise Fst in GDA, but the P-values will not be calculated. Select coancestry identity from the pull-down menu for above diagonal. Select any other genetic distance measure you might be interested in for below the diagonal (more on genetic distance later).

3. Hit the "Estimate" button. Note that the distance matrix will be strangely truncated in the program window due to space constraints. You can open the output file in Excel and play around with it to get into the proper orientation (remember there should be a diagonal down the middle).
4. Close the log file to save your results.

Getting 95% CI for pairwise Fsts in GDA

1. Open your data file in GDA. Your input file should be in a nexus format (CONVERT can do this for you).

Hint: Double-check the number of populations, loci, and individuals that will appear in GDA to make sure there are no errors in the data file (and that you have opened the correct one).

2. Open a log file (File → Log – save under informative file name). Now all your results will be saved in an output file.
3. GDA will provide you with a description of the analyses you will be running. Go to the Misc pull-down menu and selection Preferences. Click the little box that says "verbose" next to it (top right hand corner of dialog box).
4. Time to get sneaky... There are two ways to estimate pairwise Fsts in GDA. I am going to mention an alternative method when I discuss genetic distances. The first way that I will show you now will allow you get evaluate significance of your estimate. First, open the Misc pull down menu and select "Include/Exclude populations" (Misc → Include/Exclude Populations).
5. The Include/Exclude Population menu will allow you to select which populations you want to include in your analysis. The default condition is that all populations are included. First we are going to estimate pairwise Fst among our first two populations. Highlight all populations with the exception of the first two (listed alphabetically) and hit the "Exclude" button. Now only the first two populations will be included in subsequent analyses. Select "Okay" to close the box.

Hint: GDA will tell you now that only two populations will be included in analyses.

6. Open the F-stats pull down menu and select “Bootstrap across loci” (F-stats → Bootstrap across loci). This will estimate F_{is} , F_{it} , and F_{st} , and bootstrap across loci to create a 95% confidence interval for them. Because only two populations are included in analyses, Theta-P (aka Weir and Cockerham’s F_{st}) is the pairwise F_{st} estimate for the first two populations.

Hint: If your confidence interval does not include zero, the pairwise estimate is significant (aka significantly different from zero).

7. Repeat this process for all additional combinations of populations.

Hint: The only problem with this method is that it will not allow you to perform a Bonferroni correction for multiple comparisons. Other programs (FSTAT, SPAGeDi) will provide you with an exact P-value that can be used when conducting Bonferroni corrections.

Pairwise F_{st} s, P-values, and Bonferroni Corrections in FSTAT:

1. Open the FSTAT program. As the program opens, it may or may not ask you to provide random numbers for initialization of statistical tests. I think it only might do this the first time you open the program on a particular computer.

1a. Before you can open your file in FSTAT, you need to convert it to the proper input format. CONVERT cannot directly create an FSTAT file for you, but it can convert your data to a GENEPOP format, and FSTAT will convert from GENEPOP to FSTAT format.

2. To convert your file from GENEPOP to FSTAT format, go to the Utilities pull-down menu and select “File Conversion” (Utilities → File Conversion → Genepop → FSTAT).

3. Select the GENEPOP file you want. FSTAT might ask you under what name you want to save the FSTAT converted file. (It may just create a converted file with the same name and a .dat extension). If it asks for a name, enter in an informative name and hit “Save.”

4. Go to the File pull-down menu and select open. Select your newly converted FSTAT file. Notice that after you open the data file, you can now manipulate the analysis options on the program interface.

5. Select the following options: “ F_{st} per pair of samples” (under Global statistics on F-statistics, Testing, and Disequilibrium tab) and “Pairwise Tests of Differentiation” (found at bottom left of program interface). Also choose a “Nominal Level for Multiple Tests.” This sets the initial alpha level that FSTAT will modify when it conducts a Bonferroni correction on your P-values (most people select 5/100, but this would depend on the questions being asked). To minimize the amount of data in your output, I would deselect all other options on the program interface (unless you are interested in those at this point as well).

6. Hit the “Run” button. FSTAT is unusual in that it generates multiple results files (which can be confusing). Open these in either notepad or Excel. The output file that you actually named when you conducted the initial file conversion will give you a bunch of summary statistics (F-statistics at each locus along with variance components and F-statistics bootstrapped across loci for 95% CI). The file labeled FSTAT gives you the seeds it uses for its randomization tests. This may not be interesting. It will give you one output file that will include your pairwise Fst matrix (.FST extension), and another file (will have pp in the name and .pvl as an extension) will give your P-values, Bonferroni corrected alpha, and significance matrix (* will indicate which positions on the pairwise Fst matrix are significant).

Hint: The file labeled FSTAT gives you the seeds it uses for its randomization tests. This may not be interesting. You may also get a file that will appear to your computer to be in Winamp format. I thought this might be a little cadence of fanfare to announce your results, but if you open it in notepad, it turns out to be a matrix of values I can't identify. (All values in the matrix generated from the sturgeon data were above 1 so it can't be Fst or P-values).

Quick tutorial: Pairwise Fsts in SPAGeDi – Another Way Get Some Real P-values!!

1. Download SPAGeDi: <http://www.ulb.ac.be/sciences/ecoevol/spagedi.html>
2. Put your file into the proper format. Formatting the input file from scratch will take a long time (and is extremely frustrating) but the program will convert from GENEPOP or FSTAT file formats (CONVERT can make a GENEPOP file for you). Like in GDA, to get P-values for pairwise estimates, you can only analyze to populations at a time. All input files must consist of population pairs only.

Hint: Make sure input file and SPAGeDi program are stored in the same file. The program will not run otherwise.

2. Open SPAGeDi. You will see a DOS window. This may inspire a wave of panic, but don't worry, the program isn't that difficult to manipulate.
3. SPAGeDi will prompt you to enter the name of your input file. If you are opening a GENEPOP or FSTAT formatted file, hit space and then Enter. If you are opening a file formatted by hand, type in the name and be sure to add the .txt extension. Hit Enter.
 - 3a. Select whether you have a GENEPOP or FSTAT input format. Enter “1” or “2” and hit Enter.
 - 3b. Type the name of your input file. Be sure to include the file extension, if there is one.

3c. SPAGeDi will ask you to rename the file for SPAGeDi input. You might want to change the name to avoid losing the file in its original GENEPOP or FSTAT format. Don't forget the .txt file extension. Hit enter when the program asks you if you want to proceed.

4. Now the program will ask you to name the results file. Again, be sure to add the .txt extension so the output will be saved properly. Hit Enter.

5. The program will give you a quick summary of the data. Double-check this to make sure your file is being read correctly by the program. I have had instances where mistakes have occurred at this point. Hit Enter if there are no discrepancies.

6. SPAGeDi will ask you at what level you want to conduct these analyses. Select "Population Level" by entering "2." Hit Enter.

7. SPAGeDi will now list all of the statistics available under the "Population Level" option. Note that you can calculate Fsts several ways, and you can conduct several analyses at once. We only want to estimate pairwise Fsts here, so select "1." Hit enter.

8. The next menu is for Computational Options. This allows you to incorporate spatial information if you have it (option 1). We want to determine the significance of our pairwise Fst estimates, which can be conducted with permutation tests. Select "3" and hit enter.

Hint: You may get a warning message here, informing you that you have no spatial information in your data file. At this point, you have the option to add spatial info from another file. If you don't have any of these data, just hit enter.

9. Now you will see a list of permutation options. If you select "1" your output will include only the P-value for your estimates, and not the thousands of permutations necessary to get that P-value. The second two options allow you to control how the permutations are conducted. I have always stuck with default values. Select "1" to proceed with default settings. Hit enter.

10. SPAGeDi will ask you to enter the number of permutations you would like to conduct. I have always selected the maximum number (20,000). It should only take 3-5 minutes to run through 20,000 permutations.

11. Now you can select how you would like your output to be formatted. Select "3" for pairwise genetic statistics.

12. This next menu is also about formats, particular for pairwise stats. "1" will give you an output with values in columns as opposed to matrices. Because you are only dealing with a pair of populations, columns are sufficient.

13. Your output file will be saved to the same folder as your input file. I recommend opening the output in Excel. At the top, you will see allele frequency information for each locus. Next you will see a list of F-statistics (Fis, Fit, Fst) for each locus and an estimate for ALL LOCI. The estimate for ALL LOCI is your pairwise Fst for your population pair. Below this is a list of P-values for the F-stats at each locus and ALL LOCI. The P-value for ALL LOCI is the P-value for your pairwise estimate. Finally, you get the above results in an alternate format (row).

Calculating Genetic Distance in GDA:

1. Open the data file (File → Open) in GDA.
2. Start a log file (File → Log – save under informative file name). Now all your results will be saved in an output file.
2. Go to Distance pull down menu and click on Options (Distance → Options). A box will appear that will give you several choices of what type of analysis you want to perform. The top of the box will ask you about what type of genetic distance metric you want to calculate. You can calculate two genetic distance statistics simultaneously, with the results of one show above a diagonal in the distance matrix, and one below the diagonal. Choose what measure you want above and below the diagonal from the pull-down menu. Be sure to click the box labeled “Show distance matrix.”

****Hint:** Coancestry identity is equivalent to the Weir & Cockerham (1988) pairwise Fst (see Bruce Weir’s Genetic Data Analysis II p 445 for details). This is an alternative way to estimate pairwise Fst in GDA, but the P-values will not be calculated.

3. The second portion of the Options dialog box will ask you what kind of tree you want to be generated from your distance matrix (WPGMA, UPGMA, or Neighbor-Joining). Select which option you would like from the pull-down menu. The tree will be generated from the values located below the diagonal (so make sure that the estimates from which you want to infer a tree are below the diagonal on the distance matrix. Be sure to select the “Show Phenogram” box if you want to see your tree in the program interface.

Hint: I would advise not selection the “Use Line Drawing Characters” box. This will construct the branches of your tree with letters and numbers instead of lines. It looks awful and can be especially confusing if the branch is constructed with the same letters found in the population name/abbreviation.

4. Hit “Estimate.” Note that the distance matrix will be strangely truncated in the program window due to space constraints. You can open the output file in Excel and play around with it to get into the proper orientation (remember there should be a diagonal down the middle).

5. Stop the results log by clicking on (unchecking) the Log option under the File pull-down menu (File → Log).

6. If you would like to see your tree in a more interpretable format, open it in the TreeView program. First, open the TreeView program. You need to have a printer driver installed on your computer in order to do this. Then open the Dist pull-down menu and select “Invoke TreeView” (Dist → Invoke TreeView).

Hint: TreeView needs to already be open for this function to work.