GenAlEx Tutorial 2

Genetic Distance and Analysis of Molecular Variance (AMOVA)

GenAlEx

Genetic Analysis in Excel

6.5

©2006 to 2012

Based on material provided at the graduate workshop *Molecular Marker Analysis of Plant Population Structure and Process* by Rod Peakall at Copenhagen University, Denmark, August 2009.

Table of Contents

About this Tutorial Module

This GenAlEx tutorial is the second in a series of modules. It is strongly recommended that you complete the first module before attempting this second one since important background to both the underlying statistics and the use of GenAlEx is provided in that first module. Furthermore, the first module includes a glossary of terms that are used throughout all of the modules.

This document is based on material provided by Rod Peakall at the workshop *Molecular Marker Analysis of Plant Population Structure and Process* held at Copenhagen University, Denmark, August 2009. That material was in turn based on a much expanded tutorial module presented at the national graduate workshop *Genetic Analysis for Populations Studies* offered by Rod Peakall (Australian National University) and Peter Smouse (Rutgers University, USA) at the Australian National University in July 2009.

This tutorial is intended to provide an introduction to individual-by-individual genetic distances and the Analysis of Molecular Variance (AMOVA). It will also describe how to perform these analyses in the software *GenAlEx*.

This tutorial module is provided free for personal use by registered users of the software package GenAlEx. This document and associated data files must not be used for any other purpose, including teaching in any undergraduate or graduate course, without express permission of the authors. While every effort has been taken to ensure the accuracy of this document, supporting data files and the software package GenAlEx, we are unable to take responsibility for unintentional errors or software problems that may be encountered by users. We regret that we are also unable to provide individualized support. This tutorial has been updated to reflect the options provided in GenAlEx 6.5.

Rod Peakall and Peter Smouse, Aug 2012

Professor Rod Peakall
Evolution, Ecology and Genetics
Research School of Biology
The Australian National University
Canberra ACT 0200 Australia
Email: rod.peakall@anu.edu.au

Professor Peter Smouse
Department of Ecology, Evolution and Natural
Resources. School of Environmental and
Biological Sciences, Rutgers University
New Brunswick NJ 08901-8551 USA
Email: smouse@AESOP.Rutgers.edu

Goals of the Tutorial

1. To describe the procedures for the hand calculation of individual-by-individual genetic distance for haploid, codominant and binary genetic data.
2. To introduce the Analysis of Molecular Variance (AMOVA) framework and demonstrate how to calculate AMOVA by hand.
3. To illustrate how to use the software GenAlEx to calculate genetic distance and perform Analysis of Molecular Variance.
4. To explore the biological interpretation of the statistics described for some real data sets.

Individual x Individual Genetic Distance

Broadly speaking, population genetic analyses proceed along one of two pathways: frequency-based analysis or distance-based analysis. Frequency-based analyses will be familiar to all population genetic researchers (see GenAlEx Tutorial 1 for revision). In these analyses an estimate of allele frequencies is the basis for most downstream calculations. For codominant data, frequency-based analyses include *F*-statistics, Nei's genetic identity and distance measures, population assignment procedures, tests for sex-biased dispersal, estimates of genotypic probabilities, probabilities of identity, probabilities of exclusion and pairwise relatedness estimates. A subset of these frequency-based analyses is also applicable to haploid and binary data.

By contrast to frequency-based analyses, genetic distance-based analyses are relatively new. For these analyses the starting point is the conversion of genetic data into a pairwise individual-by-individual genetic distance matrix. Distance matrices can be calculated for all kinds of genetic data including codominant, haploid and binary data genetic markers, and DNA sequences. Once a genetic distance matrix is calculated, further extensive genetic analyses can be performed including: Analysis of Molecular Variance (AMOVA); Principal Coordinates Analysis (PCoA); UPGMA and Neighbor Joining Tree building; Mantel Tests; Spatial Autocorrelation analyses; and *TwoGener*. Many of these distance-based analysis options are particularly relevant to the study of spatial genetic structure.

In this module we will introduce the various genetic distance metrics and how to calculate them. The genetic distance matrices generated will then form the input for a range of distance-based statistical analyses including AMOVA and spatial genetic analysis.

Haploid Distance

Haploid genetic distance is applicable to all DNA markers and DNA sequence generated from haploid organisms. It is also applicable to DNA sequence and genetic markers from mtDNA and cpDNA in diploid organisms.

The calculation of pairwise individual-by-individual genetic distance for haploid data is as follows. Any comparison with the same state yields a value of 0 (e.g. both 1 vs 1 comparisons, 2 vs 2 comparisons, and 3 vs 3 etc.), while any comparison of different states (e.g. 1 vs 2 or 1 vs 3) yields a value of 1. Genetic distances are summed across loci, in which case for a given pair of samples the total distance is equivalent to the tally of differences between the two genetic profiles.

Within GenAlEx a pairwise, individual-by-individual (*N x N*) genetic distance matrix can be easily generated for haploid data. This genetic distance matrix can then be used in subsequent Principal Coordinate Analysis (PCoA), Mantel and all Spatial analyses involving haploid data. While usually transparent to the user, this distance option is also used when performing an Analysis of Molecular Variance (AMOVA) for haploid data. AMOVA yields an estimate of Φ_{PT} a measure of population genetic differentiation that is analogous to F_{ST}.

Ex 2.1 Calculating Haploid Distance in GenAlEx (Optional)

The small data set in this example is drawn from a much larger study of cpSSR variation in orchids (see Ebert and Peakall 2009 & Ebert et al. 2009). Here results are shown at 10 loci for 6 samples.

cpSSR variation scored across 10 loci for 6 orchid samples

Sample	Pop	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10
1	CP	12	11	7	16	9	9	9	12	11	10
2	CP	12	11	7	16	9	9	9	12	10	10
3	CP	12	11	7	16	9	9	9	12	11	10
4	CV	14	10	8	14	10	12	9	12	9	10
5	CV	14	10	8	14	10	12	9	12	9	10
6	CV	14	9	8	14	10	10	9	12	9	11

Step 1. For each pair of samples (15 pairwise contrasts in all), tally the total number of different positions across the 10 loci. Two examples are provided below.

Pairwise tally of difference for Samples 1 and 2

Sample	Pop	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10
1	CP	12	11	7	16	9	9	9	12	11	10
2	CP	12	11	7	16	9	9	9	12	10	10
Tally		0	0	0	0	0	0	0	0	1	0
Total	1										

Pairwise tally of difference for Samples 1 and 4

Sample	Pop	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10
1	CP	12	11	7	16	9	9	9	12	11	10
4	CV	14	10	8	14	10	12	9	12	9	10
Tally		1	1	1	1	1	1	0	0	1	0
Total	7										

Step 2. Record your totals in the lower triangular matrix below.

The Haploid Distance matrix

1	**2**	**3**	**4**	**5**	**6**	
0						**1**
1	0					**2**
0	1	0				**3**
7			0			**4**
				0		**5**
					0	**6**

Step 3. Check your answers using GenAlEx. The workbook containing the data is called *Ex 2.1 cpSSR Haploid Distance.xls.*

Step 4. Choose *Distance->Genetic* from the **GenAlEx** menu, then select *Haploid*, *Output Total Distance Only, Tri Matrix*, *Label Matrix by Pop* and *Labeled Opt* in the Genetic Distance Options dialog box.

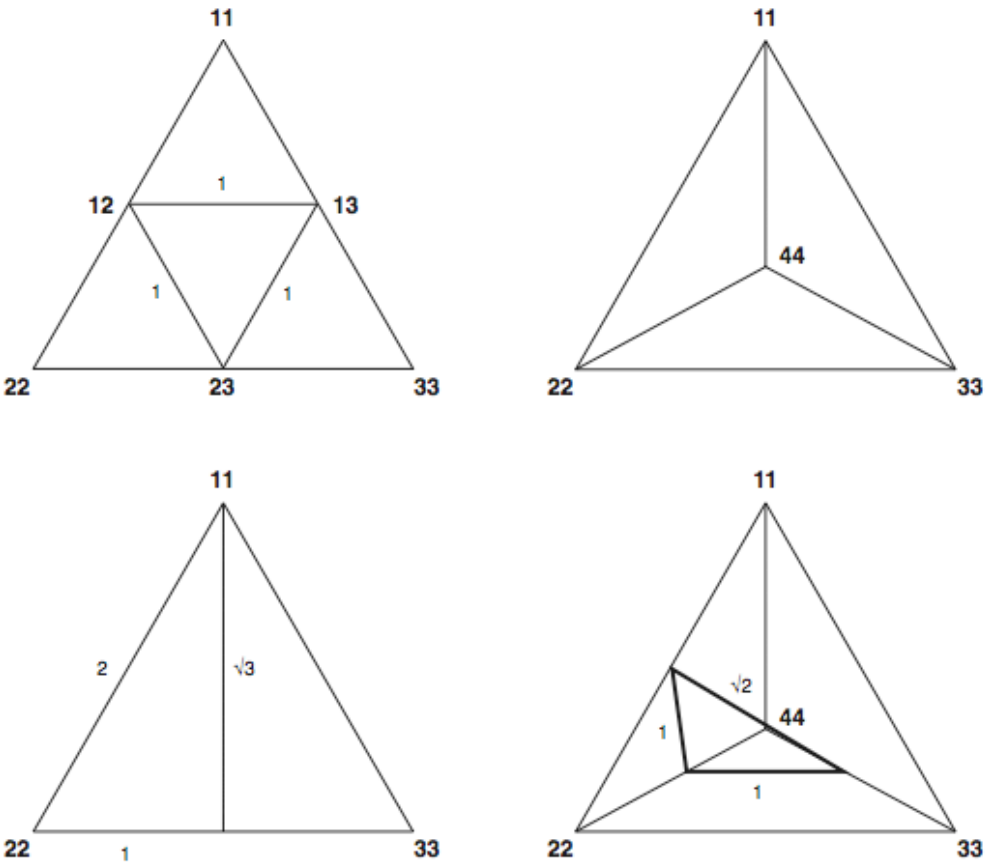
Codominant Genotypic Distance

Given that microsatellites or Simple Sequence Repeats (SSRs) are often the markers of choice for many population genetic studies, the calculation of codominant genetic distances is a key step in all studies using these markers in distance-based analyses.

There are at least 3 different ways to calculate a genetic distance for codominant data. The *codominant genotypic* distance is the one we discuss here. A graphical explanation showing the linear genotypic distances among diploid codominant genotypes with 3 and 4 alleles, respectively, is shown below (following Smouse and Peakall 1999). When squared, it is apparent that the squared genetic distances take on the values of 0, 1, 2, 3 and 4. Conveniently, for any pairwise comparison at a given locus, the maximum number of different alleles in the comparison is 4, irrespective of the total number of alleles in the population. Therefore, the schematic below is generic, all one needs to do is change the labels to match the alleles involved in any given comparison.

Within GenAlEx, a pairwise, individual-by-individual (*N x N*) genotypic distance matrix is calculated for codominant data by the *Codom-Genotypic* distance option using the following rules: For a single-locus, with *i*-th, *j*-th, *k*-th and *l*-th different alleles, a set of squared distances is defined as $d^2(ii, ii) = 0$, $d^2(ij, ij) = 0$, $d^2(ii, ij) = 1$, $d^2(ij, ik) = 1$, $d^2(ij, kl) = 2$, $d^2(ii, jk) = 3$, and $d^2(ii, jj) = 4$. Genetic distances are summed across loci under the assumption of independence.

This is the most important genetic distance option for codominant data, since the matrix generated is used in GenAlEx for subsequent PCoA, Mantel and all Spatial analyses. This distance option can also be used to calculate Φ_{PT} via AMOVA, a measure of population genetic differentiation that suppresses intra-individual variation and is therefore ideal for comparisons between codominant and haploid or binary data, where no intra-individual variation (heterozygosity) is available.



Ex 2.2 Calculating Codominant Genotypic Distance

This is a hypothetical data set consisting of 10 samples at a single locus with 4 alleles. Each sample has a unique genotype representing 1 of the 10 possible genotypes for a diploid 4 allele system.

Genotype Data

Sample	**Pop**	**Locus1**	
1	Pop1	1	1
2	Pop1	1	2
3	Pop1	1	3
4	Pop1	1	4
5	Pop1	2	2
6	Pop1	2	3
7	Pop1	2	4
8	Pop1	3	3
9	Pop1	3	4
10	Pop1	4	4

Step 1. By reference to the figures above, complete the matrix below showing the *Squared genetic distances* for each of the pairwise genotype contrasts. To help you get started some answers have been completed for you, however, it remains important to understand these completed answers.

Hint: you should see a pattern emerging that once recognized will help you to complete the matrix quickly. For convenience you might wish to use the upper triangular matrix to record the linear distances and the lower triangular matrix to record the squared distances.

Matrix of pairwise contrasts showing squared genetic distances below the diagonal

11	**12**	**13**	**14**	**22**	**23**	**24**	**33**	**34**	**44**	**Genotype**
0										**11**
1	0									**12**
1	1	0								**13**
1	1	1	0							**14**
4	1	3		0						**22**
3	1	1			0					**23**
3	1	2				0				**24**
4	3	1					0			**33**
3	2	1						0		**34**
4	3	3							0	**44**

Step 3. Check your answers using GenAlEx. The workbook containing the data is *Ex 2.2 Codom Distance.xls.* Use *Distance->Genetic*, then select *Codom-Genotypic, Output Total Distance Only, Tri Matrix* and *Label Matrix by Sample* in the Genetic Distance Options dialog box.

Binary Genetic Distance

Within GenAlEx a pairwise individual-by-individual (*N x N*) genetic distance matrix is easily generated for binary data. When calculated across multiple loci for a given pair of samples, the Binary Genetic Distance calculated by GenAlEx is equivalent to the tally of state differences among the two DNA profiles.

This genetic distance matrix can be used for subsequent PCoA, Mantel and all Spatial analysis involving binary data. This distance option is also used to calculate Φ_{PT} via AMOVA, a measure of population genetic differentiation for binary data that is analogous to F_{ST}. This is a Euclidean distance metric, unlike other binary measures such as Nei's (1 – *F*), and is therefore appropriate for AMOVA, which requires a Euclidean metric.

Ex 2.3 Calculating Binary Genetic Distance

The data set is a small subset of real AFLP data from a study of a mangrove tree species (Maguire *et al.* 2000). The data represents 14 AFLP loci for 6 samples.

Mangrove AFLP data

Sample	Pop	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	L14
SA 01	S	1	1	1	1	1	1	1	1	1	1	1	1	0	0
NW 01	N	1	1	1	0	0	1	1	0	1	1	1	1	0	0
Vic 01	V	1	1	1	1	1	1	1	1	1	0	0	1	0	0
NT 01	T	1	1	0	0	1	1	1	1	0	1	1	1	0	0
Qld 01	Q	1	1	1	1	0	1	0	1	0	1	1	0	0	1
WA 01	W	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Step 1. Return to Excel and open the workbook *Ex 2.3 Binary Distance.xls.* Activate the data worksheet. Choose *Distance->Genetic*, then select *Binary*, *Output Total Distance Only, Tri Matrix* and *Label Matrix by Sample* in the Genetic Distance Options dialog box.

Step 2. Record the pairwise individual-by-individual Binary genetic distances in the matrix below.

The Binary Distance Matrix

SA 01	**NW 01**	**Vic 01**	**NT 01**	**Qld 01**	**WA 01**	
0						**SA 01**
3	0					**NW 01**
2	5	0				**Vic 01**
			0			**NT 01**
				0		**Qld 01**
					0	**WA 01**

Analysis of Molecular Variance

The Analysis of Molecular Variance (AMOVA) is an important statistical procedure that allows the hierarchical partitioning of genetic variation among populations and regions and the estimation of the widely used *F*-statistics and/or their analogues. While not necessarily apparent to the user, AMOVA is a distance-based analysis applicable to all forms of genetic data. The data type and choice of distance calculation input into AMOVA lead to related but different analysis. A further important feature of the AMOVA framework is that it provides procedures for statistical testing.

Before one conducts a statistical test it is important to understand the null (*H0*) and alternative hypotheses (*H1*). Typically in biology the null hypothesis concerns the condition of '*No Difference*'.

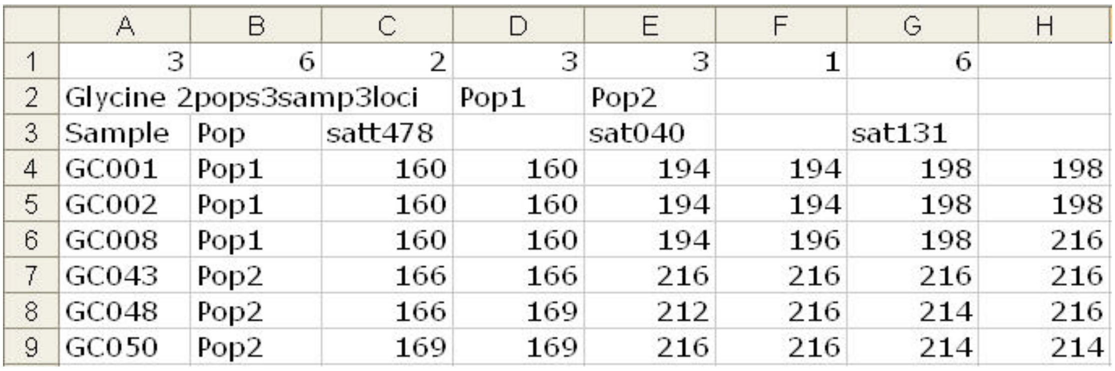
In the case of AMOVA:
H0=No genetic difference among the populations (*PhiPT*=0, or F_{ST}=0 or R_{ST}=0)
H1=Genetic difference among the populations (*PhiPT*>0, or F_{ST}>0 or R_{ST}>0)

Thus, for AMOVA our null hypothesis (*H0*) is that subpopulations can be considered part of a single large random mating genetic population. If true, any subpopulation groups we define are arbitrary and merely represent a sample from the same gene pool. Thus, we should find little difference (other than minor sampling effects) between the arbitrary subpopulations. It follows that if we shuffle (randomize) the samples in our data set, and calculate AMOVA for each shuffle, we should get values close to that expected by chance in a randomly mating population. Because of sampling effects, the results will of course vary from shuffle to shuffle. On the other hand, if we perform multiple shuffles (say 100 or 1000 times) we can obtain a good estimate of the value we would expect if the null hypothesis was true.

This is the rationale for statistical testing by random permutation that is used in an AMOVA analysis. To determine if the observed value is significantly greater than that expected by chance, we simply compare our observed value, against the outcomes of the permutations. If our observed value is greater than the permuted values 95% or more of the time, we declare the results significant at the 5% level.

Ex 2.4 Hand Calculation of AMOVA

To illustrate the hand calculation of AMOVA we will work with a small data set to make the task manageable. This data set is a small subset from the *Glycine* examples introduced in Tutorial 1. In this case there are just 6 samples, representing 2 populations, Pop1 and Pop2. Genotypes are provided for 3 microsatellite loci.



	A	B	C	D	E	F	G	H
1	3	6	2	3	3	1	6	
2	Glycine 2pops3samp3loci			Pop1	Pop2			
3	Sample	Pop	satt478		sat040		sat131	
4	GC001	Pop1	160	160	194	194	198	198
5	GC002	Pop1	160	160	194	194	198	198
6	GC008	Pop1	160	160	194	196	198	216
7	GC043	Pop2	166	166	216	216	216	216
8	GC048	Pop2	166	169	212	216	214	216
9	GC050	Pop2	169	169	216	216	214	214

Step 1. Compute pairwise codominant genetic distance matrix. (In this case, this has been completed for you).

The codominant distance matrix

Pop1	Pop1	Pop1	Pop2	Pop2	Pop2	
0	0	2	12	9	12	Pop1
0	0	2	12	9	12	Pop1
2	2	0	8	6	10	Pop1
12	12	8	0	3	8	Pop2
9	9	6	3	0	3	Pop2
12	12	10	8	3	0	Pop2

$SSTOT = sum\mathbf{D}/2N =$ _________

Within population components of the distance matrix

Pop1	Pop1	Pop1	
0	0	2	Pop1
0	0	2	Pop1
2	2	0	Pop1

Pop2	Pop2	Pop2	
0	3	8	Pop2
3	0	3	Pop2
8	3	0	Pop2

$SSWP1 = sum\mathbf{D}_{WP1}/2n_{WP1} =$ _________ $SSWP2 = sum\mathbf{D}_{WP2}/2n_{WP2} =$ _________

$SSWP = SSWP1 + SSWP2 =$ _________

Among population components of the distance matrix

Pop1	Pop1	Pop1	Pop2	Pop2	Pop2	
			12	9	12	Pop1
			12	9	12	Pop1
			8	6	10	Pop1
12	12	8				Pop2
9	9	6				Pop2
12	12	10				Pop2

$SSAP = SSTOT - (SSWP1 + SSWP2) =$ _________

Using the data in the distance matrices above, complete steps 3 to 5, recording you answers both above and in the table below.

Step 3. Calculate the sum of all genetic distances by summing all elements of the total matrix. Divide by 2x the total number of samples (*2N*) to give the Sums of Squares (*SS*) for the Total (*SSTOT*).

Step 4. For each population, calculate the sum of within population genetic distances by summing the elements of the respective within population square matrices. Divide by 2x the number of samples in each population (2*n*) to give the respective Sums of Squares (*SS*) within each population (*SSWP1*, *SSWP2*). Sum across populations to give the overall *SS* within populations.

Step 5. Compute the Sums of Squares among populations as *SSAP=SSTOT-*(*SSWP1+SSWP2*).

Step 6. Calculate Degrees of Freedom (*df*) among populations as the number of populations minus 1 (*Np-1*).

Step 7. Calculate the Mean Sums of Squares (*MS*) among populations (*MSAP*) by dividing the Sums of Squares by the degrees of freedom among pops.

Step 8. Calculate Degrees of Freedom (*df*) within populations as the number of samples minus the number of populations (*N-Np*).

Step 9. Calculate the Mean Sums of Squares within populations (*MSWP*) by dividing the Sums of Squares by the degrees of freedom within pops.

Step 10. Calculate *N0*. When all pops are of equal size *n*, then *N0=n*.

Step 11. Calculate the estimated variance among populations *VAP* as (*MSAP-MSWP*)/*N0*.

Step 12. Calculate the estimated variance within populations *VWP* as *MSWP*.

Step 13. Compute *PhiPT* as *VAP*/(*VAP+VWP*).

Step 14. Convert the estimated variances to percentages of the total variance.

AMOVA summary

N0	3				
SSTOT					
Pop	**Pop1**	**Pop2**			
n	3	3			
SSWP					
Source	**df**	**SS**	**MS**	**Est. Var.**	**%**
Among Pops					
Within Pops					
Total					
Stat	**Value**				
PhiPT					

Step 15. Check your answers using GenAlEx via the ***AMOVA*** menu option. Be sure to set the number of permutations to zero. The workbook containing this data set is called *Ex 2.4 AMOVA by Hand.xls*

Q 2.4. Questions

1. Inspect the *Glycine* data above. What do you note about the genotypic patterns?

2. Based on your inspection of the data, do you predict that the two populations will exhibit high or low genetic differentiation?

3. What do you conclude about the degree of genetic differentiation from your AMOVA analysis? Are your results statistically significant?

Box 2.1 AMOVA Formulae

Recall that an important measure of variability is provided by the Variance, calculated as:

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

while the Standard Deviation is calculated as the square root of the Variance:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

For Analysis of Variance (ANOVA) the Sums of Squares (*SS*) and Mean Sums of Squares (*MS*) are calculated as:

$$SS_{TOT} = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

$$MS_{TOT} = \frac{\sum_{i=1}^{N} (y_i - \bar{y})^2}{N-1}$$

For Analysis of Molecular Variance (AMOVA) the Sums of Squares (*SS*) and Mean Sums of Squares (*MS*) are calculated from a square genetic distance matrix as:

$$SS_{TOT} = \frac{\sum d_{ij}^2}{2N}$$

$$SS_{WP1} = \frac{\sum d_{ij}^2}{2n_1}$$

$$SS_{WP} = SS_{WP1} + SS_{WP2} ... SS_{WPn}\,,\; SS_{AP} = SS_{TOT} - SS_{WP}$$

$$MS_{WP} = \frac{SS_{WP}}{df_{WP}}\,,\; MS_{AP} = \frac{SS_{AP}}{df_{AP}}$$

Where d_{ij}^2 = the squared genetic distance between the *i*th and *j*th sample. The estimates of variances are calculated as:

$$V_{AP} = \frac{MS_{AP} - MS_{WP}}{N0}$$

$$V_{WP} = MS_{WP}$$

$$\text{Where}, N0 = \frac{1}{(Np-1)} \times \left(\sum_{k=1}^{N_p} np_k - \left(\frac{\sum_{k=1}^{Np} np_k^2}{\sum_{k=1}^{Np} np_k} \right) \right)$$

Finally, PhiPT is calculated as:

$$\phi_{PT} = \frac{V_{AP}}{(V_{AP} + V_{WP})}$$

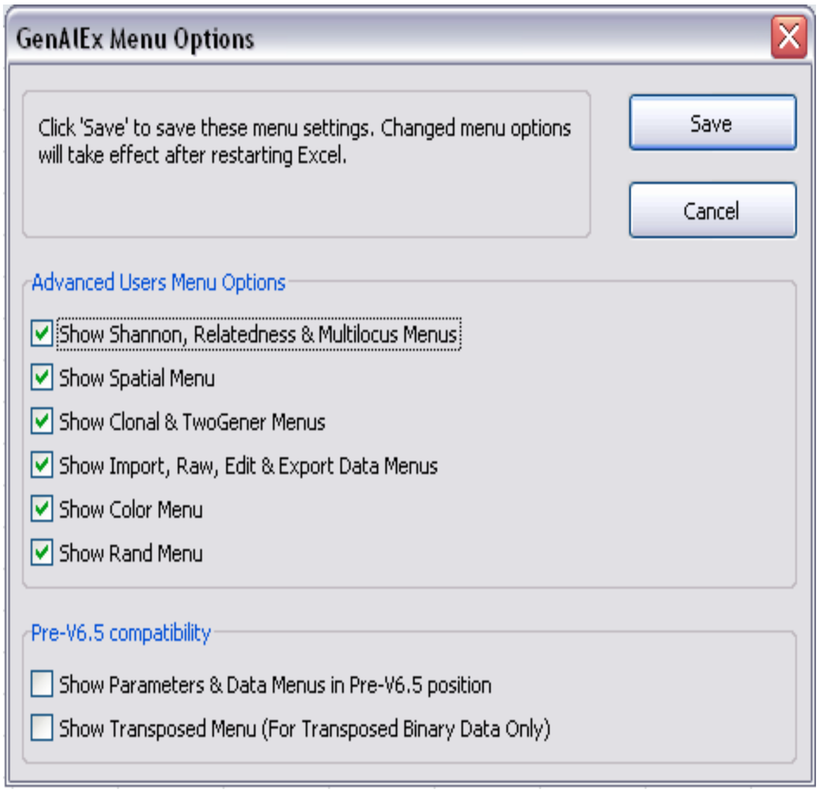
N = number of samples, Np = number of populations with *k*th populations, n_k = number of samples in the *k*th population. Degrees of freedom among populations= $df_{AP} = Np - 1$, degrees of freedom within populations= $df_{WP} = N - Np$

Ex 2.5 Understanding Statistical Testing in AMOVA

A key feature of the AMOVA framework is that it allows for statistical testing by random permutation. This exercise will introduce in more detail the process involved. The data are drawn from *Glycine clandestine* except that here we will work with a small subset of the original data.

Step 1. Return to Excel. Check that the **GenAlEx** menu shows the option ***Rand Data***. If not follow step 2 before proceeding, otherwise jump to step 3.

Step 2. From the **GenAlEx** menu choose ***Options***->***Menus.*** In the GenAlEx Menu Options dialog box check *Show Rand Menus*. Next click *Save*, then restart Excel.


GenAlEx Menu Options

Click 'Save' to save these menu settings. Changed menu options will take effect after restarting Excel.

Advanced Users Menu Options

☑ Show Shannon, Relatedness & Multilocus Menus

☑ Show Spatial Menu

☑ Show Clonal & TwoGener Menus

☑ Show Import, Raw, Edit & Export Data Menus

☑ Show Color Menu

☑ Show Rand Menu

Pre-V6.5 compatibility

□ Show Parameters & Data Menus in Pre-V6.5 position

□ Show Transposed Menu (For Transposed Binary Data Only)

Step 3. Open the workbook *Ex 2.5 Shuffle.xls. Activate* the data worksheet then select ***Rand Data***->***Color Shuffle***. This will create a new worksheet [Shuffle]. Inspect the new data and answer question 1 before proceeding.

Tip 1: This data has been colored using the GenAlEx option **Color Data->By Pop***.*

Tip 2: To enable easy comparison of the colored raw and colored shuffled data, you may wish to copy the colored raw data to the shuffled data worksheet.

Step 4. With the worksheet [Shuffle] active, use the ***AMOVA*** menu to calculate *PhiPT* for the shuffled data. Be sure to set the number of permutations to zero for this step.

Step 5. Repeat steps 1 and 2, four more times, recording the *PhiPT* values in the table below. Calculate the mean *PhiPT* value across the shuffles. Answer question 2 before proceeding.

Outcome of PhiPT for shuffles compared to original data

Run	***PhiPT***
Original Data	0.545
Shuffle 1	
Shuffle 2	
Shuffle 3	
Shuffle 4	
Shuffle 5	
Mean PhiPT of Shuffles	

Step 6. Repeat the AMOVA analysis for the original data. At the AMOVA Options dialog box, set the number of permutations to 99 and select *Pie Graph, Freq. Dist., Standard Permute* and *Pm. Values*.

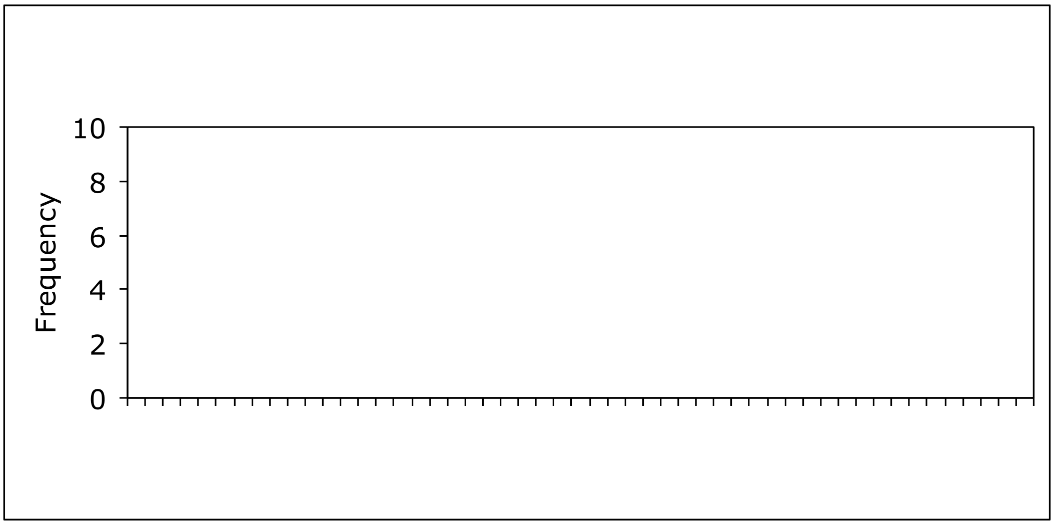
Step 7. Inspect the results of your analysis. The outcomes are presented in 3 worksheets: [*PhiPT*], [*PhiPTFD*] and [*PhiPTPV*].

Step 8. In the worksheet [*PhiPTPV*] a sorted list of the *PhiPT* values for the 99 permutations and the observed value is shown. Scroll down the list and locate the position of the observed value (*PhiPT*=0.545).

Calculate the Probability as:
Number of Random Values => Observed Value (Including Observed Value)/Number of Permutations + 1. (Where '=>' means 'greater than or equal to')

Step 9. Compare you hand calculated probability with the outcome in GenAlEx (listed in the AMOVA summary table in the worksheet [PhiPT]).

Step 10. In the worksheet [*PhiPTFD*], a graph showing the 'Frequency Distribution of Permuted PhiPT versus Observed PhiPT' is shown. Draw a diagram of this graph below, include a title and axis labels. Next, answer question 3.



Q 2.5 Questions

1. Based on your inspection of the shuffled data, describe what has happened to the data during the shuffle.

2. How does the mean *PhiPT* value across the shuffles compare with the *PhiPT* for the original data?

3. What do you conclude about the degree of genetic differentiation and its statistical significance?

Box 2.2 PhiPT(Φ_{PT}), F_{ST} and R_{ST}

$$\phi_{PT} = \frac{V_{AP}}{(V_{AP} + V_{WP})}, F_{ST} = \frac{V_{AP}}{(V_{AP} + V_{WP})},\ R_{ST} = \frac{V_{AP}}{(V_{AP} + V_{WP})}$$

Where V_{AP} is the variance among populations and V_{WP} the variance within populations. It is apparent that Φ_{PT}, F_{ST} and R_{ST} estimate the proportion of the variance among populations, relative to the total variance. The difference between the estimators is only in the genetic distance matrix used to calculate the statistic. Within GenAlEx, for haploid and binary data only estimates of Φ_{PT} are available.

For codominant data, all three estimates of differentiation can be calculated. Φ_{PT} is calculated when *Codom-Genotypic* distance is used. In this case, variation within individuals is suppressed. F_{ST} is calculated when *Codom-Allelic* distance is the input. Φ_{PT} is best for comparisons between Codominant data and Binary or Haploid data; otherwise F_{ST} is recommended, because F_{ST} is more widely used. R_{ST} is an estimator of genetic differentiation for microsatellite loci that assumes a stepwise mutation model. It is calculated when *Codom-Microsat* genetic distance is the input. Normally,Φ_{PT}, F_{ST} and R_{ST} will be greater than zero. Within GenAlEx, AMOVA procedures follow Excoffier et al. (1992), Huff et al.(1993), Peakall et al. (1995) and Michalakis and Excoffier (1996).

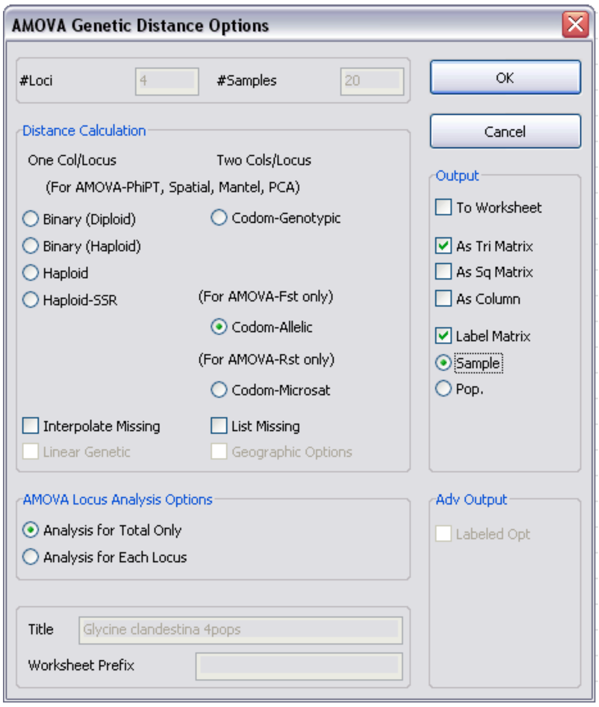
Ex 2.6 Calculating Regional F-statistics via AMOVA

In a standard AMOVA analysis a single statistic, *PhiPT* (or F_{ST}, or R_{ST}) is the outcome, along with a probability value indicating whether the statistic is statistically significant from zero or not. With multiple populations our interest likely goes beyond this single statistic. For example, we might also be interested to determine if there is significant regional genetic structure. Alternatively, we might be interested in further exploration of the patterns of genetic distances among populations. Are all pairwise contrasts significantly different, or not etc.?

In this exercise we will perform a regional F-statistics analysis and at the same time generate pairwise estimates of F_{ST} among populations for *Glycine clandestina*. We will also discover the *F-statistics* to be informative about the level of inbreeding within this species.

You have seen a version of this data set before in Tutorial 1. Just to remind you, this *Glycine* data set consists of 4 populations: 2 populations, Aranda and Taylor from lower altitude, and 2 populations, Brind and Franklin from higher altitude.

Step 1. Open the workbook *Ex 2.6 AMOVA Glycine.xls.* Inspect the *Data* worksheet, note that column 1 has been used to identify the pops and column 2 to identify the regions. The first task is to use the ***Parameters*** menu option to obtain the population and regional parameters. Which submenu will you select?



AMOVA Genetic Distance Options

#Loci 4 #Samples 20

Distance Calculation

One Col/Locus Two Cols/Locus

(For AMOVA-PhiPT, Spatial, Mantel, PCA)

○ Binary (Diploid)
○ Codom-Genotypic
○ Binary (Haploid)
○ Haploid
○ Haploid-SSR

(For AMOVA-Fst only)

● Codom-Allelic

(For AMOVA-Rst only)

○ Codom-Microsat

□ Interpolate Missing
□ List Missing
□ Linear Genetic
□ Geographic Options

Output

□ To Worksheet
☑ As Tri Matrix
□ As Sq Matrix
□ As Column
☑ Label Matrix
● Sample
○ Pop.

AMOVA Locus Analysis Options

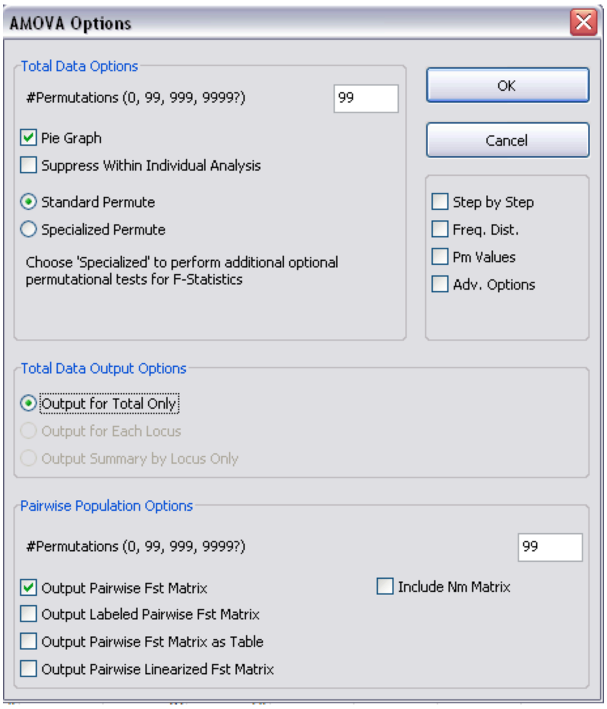
● Analysis for Total Only
○ Analysis for Each Locus

Adv Output

□ Labeled Opt

Title Glycine clandestina 4pops

Worksheet Prefix



AMOVA Options

Total Data Options

#Permutations (0, 99, 999, 9999?) 99

☑ Pie Graph
□ Suppress Within Individual Analysis

● Standard Permute
○ Specialized Permute

Choose 'Specialized' to perform additional optional permutational tests for F-Statistics

□ Step by Step
□ Freq. Dist.
□ Pm Values
□ Adv. Options

Total Data Output Options

● Output for Total Only
○ Output for Each Locus
○ Output Summary by Locus Only

Pairwise Population Options

#Permutations (0, 99, 999, 9999?) 99

☑ Output Pairwise Fst Matrix
□ Include Nm Matrix
□ Output Labeled Pairwise Fst Matrix
□ Output Pairwise Fst Matrix as Table
□ Output Pairwise Linearized Fst Matrix

Step 2. Run *AMOVA*, selecting the option *Codom-Allelic* in the Genetic Distance Options dialog box. At the AMOVA Options dialog box, select *Pie Graph*, in the upper section of the dialog box, and set the number of permutations to 99. In the lower section of the dialog box, select *Output Pairwise Fst Matrix* and set the number of permutations to 99.

Step 3. Inspect the results of your analysis presented in the two worksheets: [Fst] and [FstP] and answer the questions below.

Q2.6 Questions

1. Record the Estimated Variance and % values from the AMOVA summary table, and the F-statistics and their probabilities in the table below.

Source	**df**	**SS**	**MS**	**Est. Var.**	**%**
Among Regions	1	20.723	20.723		
Among Pops	2	6.143	3.071		
Among Indiv	52	85.357	1.641		
Within Indiv	56	20.500	0.366		
Total	111	132.723			
F-Statistics	**Value**	**P(rand >= data)**			
F_{RT}					
F_{SR}					
F_{ST}					
F_{IS}					
F_{IT}					

2. Draw a schematic of the pie chart.

3. Summarise and interpret the % of variance values and the values of F_{RT}, and F_{ST}. Did you detect significant differentiation? Explain your answer.

4. Summarise and interpret the outcomes for F_{IS}. Did you detect evidence for inbreeding?

5. Suggest a plausible biological explanation for the values of F_{RT}, F_{ST} and F_{IS}.

Based on your inspection of the results in the worksheet [FstP]*:*

6. Summarise the outcomes of the pairwise AMOVA analysis in words (include range and magnitude). Were all pairwise F_{ST} values significantly different from zero?

7. What biological explanation(s) can you offer for the patterns of pairwise F_{ST} values?

Box 2.3 The Magnitude of F_{ST}

In practice, F_{ST} is rarely larger than 0.5 and often very much less. Wright (1978) proposed for the simple 2 allelic systems that he studied that values of F_{ST} = 0.25 are taken to mean very great differentiation between subpopulations; the range 0.15 to 0.25 indicates moderate differentiation; while differentiation is not negligible if F_{ST} is 0.05 or less. The interpretation of the magnitude of F_{ST} is more complex than simple reference to this quantitative guide. Hedrick (1999) has shown that with modern hypervariable markers characterized by many alleles, F_{ST} values can be considerably lower than for genetic markers with very few alleles. In part as a consequence, there has been much recent debate about the utility of F_{ST} as a measure of population genetic structure (Jost, 2008; Ryman and Leimar, 2009; Whitlock, 2011). GenAlEx 6.5 offers the calculation of G'_{ST}, G''_{ST} and Jost's D_{est}, providing [0,1]-standardized allele frequency based estimators of population genetic structure, following Meirmans and Hedrick (2011), testing the null by random permutation, and estimating variances via jackknifing and bootstrapping over loci. New AMOVA routines now enable the estimation of standardized F'_{ST}, following Meirmans (2006).

References and Further Reading

Note that for a more extensive literature on these topics, please see Appendix 1 provided with GenAlEx: Freely available from the Australian National University, Canberra, Australia. http://biology.anu.edu.au/GenAlEx/

Brown AHD and Weir BS (1983) Measuring genetic variability in plant populations, in *Isozymes in Plant Genetics and Breeding, Part A*, (Tanksley SD, Orton TJ, Editors). Elsevier Science Publ.: Amsterdam. p. 219-239.

Conner JK and Hartl DL (2004) *A Primer of Ecological Genetics*, Sunderland, Massachusetts: Sinauer Associates, Inc.

Ebert D, and Peakall R. (2009) A new set of universal de novo sequencing primers for extensive coverage of non-coding chloroplast DNA: new opportunities for phylogenetic studies and cpSSR discovery. *Molecular Ecology Resources* 9, 777-783.

Ebert D, and Peakall R. (2009) Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. *Molecular Ecology Resources* 673-690. Invited Technical Review

Ebert D, Hayes C, and Peakall R. (2009) Chloroplast simple sequence repeat (cpSSRs) markers for evolutionary studies in the sexually deceptive orchid genus *Chiloglottis*. *Molecular Ecology Resources* 9, 784-789.

Frankham R, Ballou JD and Briscoe DA (2002) *Introduction to Conservation Genetics*, Cambridge University Press: Cambridge.

Frankham R, Ballou JD and Briscoe DA (2004) *A Primer of Conservation Genetics*, Cambridge: Cambridge University Press.

Hartl DL (2000) *A Primer of Population Genetics 3rd Ed*, Sunderland, Massachusetts: Sinauer Assoc, Inc.

Hartl DL and Clark AG (1997) *Principles of Population Genetics 3rd Ed*, Sunderland, Massachusetts: Sinauer Associates, Inc.

Hedrick PW (2000) *Genetics of Populations 2nd Ed*, Boston: Jones and Bartlett.

Jost L (2008) GST and its relatives do not measure differentiation. *Molecular Ecology* **17**(18), 4015-4026.

Meirmans PG (2006) Using the AMOVA framework to estimate a standardized genetic differentiation measure. *Evolution* **60**(11), 2399-2402.

Meirmans PG, Hedrick PW (2011) Assessing population structure: FST and related measures. *Molecular Ecology Resources* **11**(1), 5-18.

Nei M (1972) Genetic distance between populations. *American Naturalist*, **106**, 283-392.

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, **89**, 583-590.

Peakall R and Beattie AJ (1996) Ecological and genetic consequences of pollination by sexual deception in the orchid *Caladenia tentactulata*. *Evolution*, **50**, 2207-2220.

Peakall R and Lindenmayer DB (2006) Genetic insights into population recovery following experimental perturbation in a fragmented landscape. *Biological Conservation*, **132**, 520-532.

Peakall R and Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, **6**, 288-295.

Peakall R, Ruibal M and Lindenmayer DB (2003) Spatial autocorrelation analysis offers new insights into gene flow in the Australian bush rat, *Rattus fuscipes*. *Evolution*, **57**, 1182-1195.

Peakall R, Smouse PE, and Huff DR (1995) Evolutionary implications of allozyme and RAPD variation in diploid populations of dioecious buffalograss (Buchloë dactyloides (Nutt.) Engelm.). *Molecular Ecology* 4, 135-147.

Peakall, R. and Smouse P.E. (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinformatics* In press. First published online July 20, 2012 doi:10.1093/bioinformatics/bts460. Advanced print Epub available here

Ryman N, Leimar O (2009) GST is still a useful measure of genetic differentiation — a comment on Jost's D. *Molecular Ecology* **18**(10), 2084-2087.

Weir BS (1990) *Genetic Data Analysis*, Sunderland, Massachusetts: Sinauer Ass. Inc.

Whitlock MC (2011) G'ST and D do not replace FST. *Molecular Ecology* **20**(6), 1083-1091.

Wright S (1946) Isolation by distance under diverse systems of mating. *Genetics*, **31**, 39-59.

Wright S (1951) The genetical structure of populations. *Annual Eugenics*, **15**, 323-354.

Wright S (1965) The interpretation of population structure by F-Statistics with special regard to systems of mating. *Evolution*, **19**, 395-420.

Wright S (1978) *Evolution and the Genetics of Populations. Variability within and among natural populations*. Vol 4. The University of Chicago Press, Chicago.