GenAlEx Tutorial 1

Introduction to Population Genetic Analysis

Based on material provided at the national graduate workshop *An Introduction to Genetic Analysis for Populations Studies* offered by Rod Peakall and Peter Smouse at the Australian National University, Canberra, Australia, July 2009.

Table of Contents

About this Tutorial Module

This GenAlEx tutorial module is based on material provided at a two-day national graduate workshop entitled *An Introduction to 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. We are also pleased to include as an appendix, an overview on Shannon Diversity analysis by Bill Sherwin (University of New South Wales) who contributed a guest lecture to our workshop.

This tutorial is intended to provide a brief refresher course in frequency-based population genetic statistics and to introduce students to the software *GenAlEx*.

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

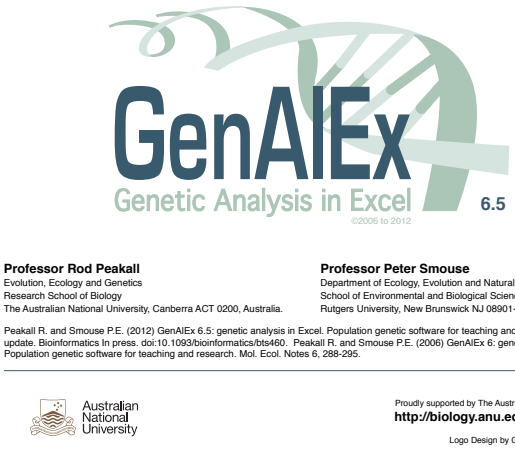
Professor Peter Smouse

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 scoring codominant genetic markers such as microsatellites.
2. To demonstrate by way of hand calculations the basic statistical procedures for frequency-based within and among population genetic analysis.
3. To introduce the software GenAlEx and outline important information about installation, data formats and operation of the software.
4. To demonstrate the basic statistical procedures for frequency-based within and among population genetic analysis including Allele Frequency, Heterozygosity, F-statistics, Nei Genetic Distance and Shannon Diversity Indices.
5. To explore the biological interpretation of the statistics described for some real data sets.

About the Software GenAlEx



GenAlEx - Genetic Analysis in Excel (Peakall and Smouse 2006, 2012) is designed as a user-friendly package with an intuitive and consistent interface that allows users to analyse a wide range of population genetic data within a software environment with which most users will have some familiarity (MS Excel). GenAlEx is now widely used by university teachers at both undergraduate and graduate levels in Australia, North America, South America, and Europe. The software also offers a wide range of analysis options for researchers, including some spatial analysis options not available elsewhere. Options for exporting data to a wide range of other population genetic packages are also provided. More than 5000 registered users, representing more than 60 countries, use the software. According to ISI, the first paper describing the software was cited more than 2200 times in the period 2006 to 2012.

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

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

Freely available from The Australian National University, Canberra, Australia at the new URL: http://biology.anu.edu.au/GenAlEx

Software Instructions

Throughout this text, instructions for using GenAlEx are provided in abbreviated form. For consistency, the same text styles as used in the *GenAlEx 6.5 Guide* have been adopted here:

Menu name (eg. GenAlEx*)*

Menu option (e.g. Distance)

Menu suboption (e.g. Genetic)

Dialog box name (e.g. Genetic Distance Options)

Dialog box option (e.g. Binary)

Tips are written in italics.

Understanding Abridged GenAlEx Instructions

Full Procedure for Calculating Genetic Distance

1. Choose the option ***Distance*** from the **GenAlEx** menu, and then select ***Genetic*** from the submenu.
2. Ensure the locus and sample parameters are correct in the Genetic Distance Options dialog box.
3. Select the appropriate Distance Calculation, and output options required (see below).
4. Enter Title and Worksheet Prefix then click *Ok*. Genetic distance is output to sheet [GD].

Abridged Procedure for Calculating Genetic Distance

Choose ***Distance***->***Genetic*** then select the appropriate *Distance Calculation* in the Genetic Distance Options dialog box.

Note the abridged options omit the prompt about entering a *Title* and *Worksheet Prefix*, however, it is strongly recommended that you take advantage of this feature which is provided to help users keep track of their data analysis. In later sections of the course instructions may be further abbreviated as students become more familiar with GenAlEx.

Genetic Marker Analysis

Broadly speaking, population genetic analyses proceed along one of two pathways: frequency-based analysis or distance-based analysis. In this introductory course we will restrict our attention to frequency-based analyses. In these analyses estimates of allele frequencies are the basis for most downstream calculations. For codominant data, frequency-based analyses include *F*-statistics, Nei's genetic distance, and Shannon diversity indices that are introduced in this first section of the course. Other allele frequency-based options, such as population assignment procedures, estimates of genotypic probabilities, probabilities of identity, probabilities of exclusion and pairwise relatedness estimates, will be covered in the main section of the workshop. 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 genetic markers, and DNA sequences. Once a genetic distance matrix is calculated, further 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*. The main section of this course will explore many of these genetic analysis options.

The first step for both frequency-based and distance-based genetic analysis is the scoring of the DNA profiles.

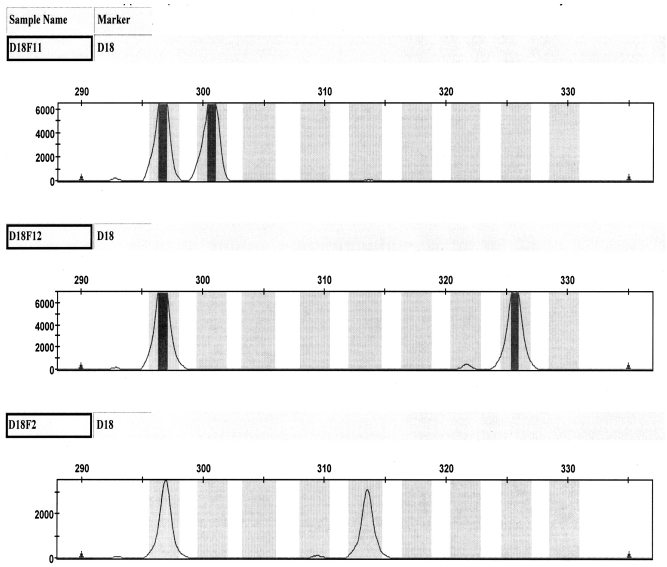
Scoring Codominant STR DNA profiles

In practice, before you begin scoring any microsatellite or STR marker it is important to inspect the profiles of a good number of samples (>20, often many more). This will allow you to get a sense of the general patterns and identify any potential artifacts that might be incorrectly scored.

1. Based on your initial inspection of the fragment sizes across multiple samples, and your knowledge of the nucleotide repeat structure, define the allele series as integers. For di-and tetra-nucleotide repeats start the allele series as either odd or even (whichever most closely matches fragment sizes) and then stick with the series chosen (unless microvariants are confirmed to disrupt the allele series).
2. For each DNA profile, identify the alleles and label the allele size(s). Note that some STR profiles show PCR artifacts such as stutter patterns that need to be identified and excluded.
3. For each DNA profile, assign the genotype score based on the allele sizes in the allele series.
4. List the genotypes in a table for downstream analysis.

The allele series and some scored genotypes are illustrated here for a tetra-nucleotide codominant microsatellite or STR locus, D18 with repeat motif [AGAA]n, that is widely used in human forensics.

Allele series = 297, 301, 305, 309, 313, 317, 321, 325, 329



Genotype

297 301

297 325

297 313

Ex 1.1 Scoring Microsatellite DNA profiles

Microsatellite genotypes at the locus *TT* for 20 samples of bush rats are shown below. Locus *TT* contains a tetra nucleotide repeat $(AAAG)_N$. Ten alleles are known at the locus with an inferred repeat range of $(AAAG)_6$ to $(AAAG)_{18}$.

Step 1. Inspect the DNA profiles of multiple samples to identify putative alleles and determine allele sizes.

Step 2. List the series of expected allele sizes as integers based on your inspection of multiple samples and knowledge of the locus sequence. Enter the list of alleles in the table below.

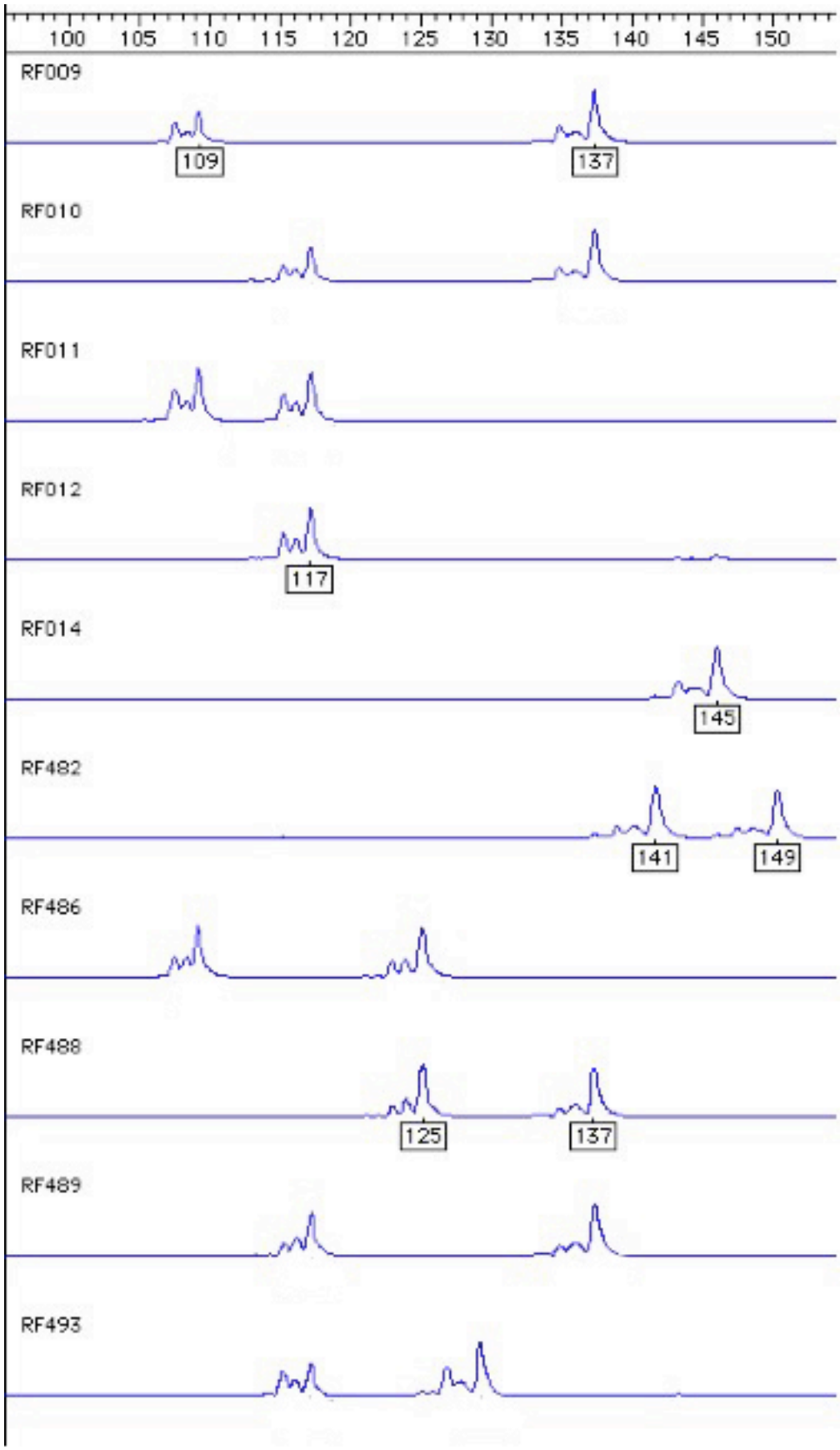
Allele Series										
109	117									

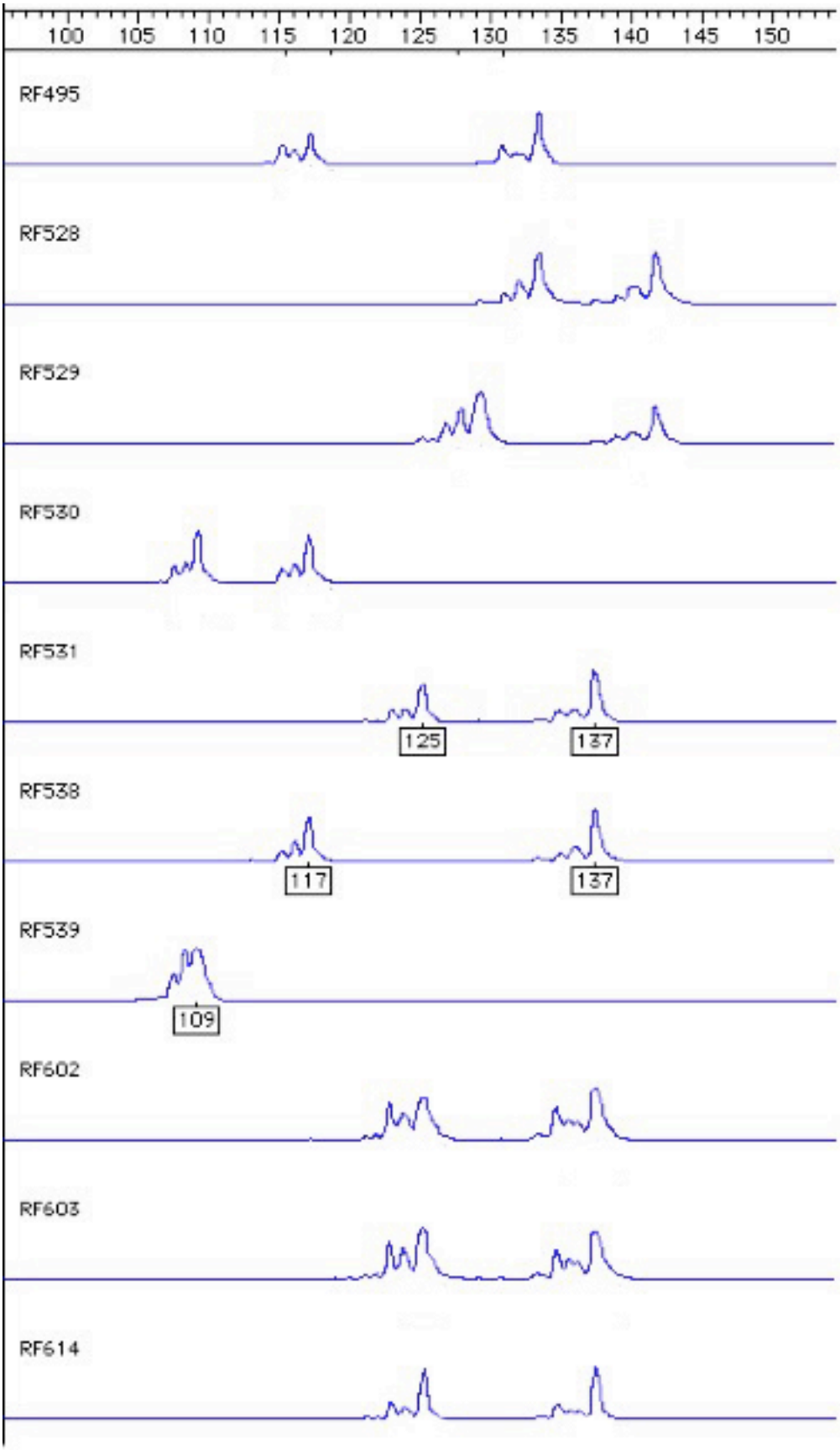
Step 3. On the DNA profiles label the alleles with their integer size in base pairs.

Step 4. Score and record the genotypes as the allele sizes in the table below.

Note that the table below is divided into 3 parts. The first 3 rows provide information about the parameters for the data set. When we begin using the computer software GenAlEx, this information will become important. Do not worry about it until then! The second block in the table is for the genotypes belonging to samples from the first population 'P1'. The third block is for samples from the second population 'P2'.

1	20	2	10	10	1	20
Scoring			P1	P2		
Sample	Pop	TT	TT			
RF009	P1	109	137			
RF010	P1	117	137			
RF011	P1					
RF012	P1					
RF014	P1					
RF482	P1					
RF486	P1					
RF488	P1					
RF489	P1					
RF493	P1					
RF495	P2					
RF528	P2					
RF529	P2					
RF530	P2					
RF531	P2					
RF538	P2					
RF539	P2					
RF602	P2					
RF603	P2					
RF614	P2					


100 105 110 115 120 125 130 135 140 145 150
RF009
109
137
RF010
RF011
RF012
117
RF014
145
RF482
141
149
RF486
RF488
125
137
RF489
RF493


100 105 110 115 120 125 130 135 140 145 150
RF495
RF528
RF529
RF530
RF531
125
137
RF538
117
137
RF539
109
RF602
RF603
RF614

Ex 1.2 Calculating Allele Frequency

This exercise continues from Ex 1.1.

Step 1. Based on your scored genotypes in the table above, calculate allele frequencies for population 1 (P1), population 2 (P2) and the total (P1 and P2). Record your answers in the table below.

Allele counts and allele frequency					
Pops./Allele	**TT**	**TT**			
P1	**Count**	**Freq**			
Allele 109	3	0.150			
Allele 117					
Allele 125					
Allele 129					
Allele 133					
Allele 137					
Allele 141					
Allele 145					
Allele 149					
P2					
Allele 109					
Allele 117					
Allele 125					
Allele 129					
Allele 133					
Allele 137					
Allele 141					
Allele 145					
Allele 149					
Total					
Allele 109					
Allele 117					
Allele 125					
Allele 129					
Allele 133					
Allele 137					
Allele 141					
Allele 145					
Allele 149					

Box 1.1 Allele Frequency for Codominant Data

$$FreqAllele\_x = \frac{2Nxx + Nxy}{2N}$$

This is calculated locus by locus. Where *Nxx* is the number of homozygotes for allele X (XX), and *Nxy* is the number of heterozygotes containing the allele X (Y can be any other allele). *N* = the number of samples. Allele Frequency can also be determined simply by direct count of the proportion of different alleles.

Ex 1.3 No. of Alleles, Heterozygosity & Fixation Index

This exercise continues from Ex 1.1 and 1.2.

Step 1. Based on the genotypes and allele frequencies for P1 and P2 calculate the number of different alleles *Na*, observed *Ho*, expected Heterozygosity *He* and the Fixation Index *F*. Show your calculations in the space below and summarise your answers in the table.

Pop		**TT**
P1	***N***	10
	Na	
	Ho	
	He	
	F	
P2	***N***	10
	Na	
	Ho	
	He	
	F	

Box 1.2 Heterozygosity and the Fixation Index

$$H_o = \frac{No.\_of\_Hets}{N}$$

Where H_o is the observed heterozygosity, i.e. the proportion of N samples that are heterozygous at a given locus.

$$H_e = 1 - \sum p_i^2$$

Where H_e is the expected heterozygosity, i.e. the proportion of heterozygosity expected under random mating and p_i is the allele frequency of the i-th allele.

$$F = \frac{H_e - H_o}{H_e}$$

The Fixation Index F (also called the Inbreeding Coefficient) exhibits values ranging from -1 to +1. Values close to zero are expected under random mating, while substantial positive values indicate inbreeding or undetected null alleles. Negative values indicate excess of heterozygosity, due to negative assortative mating, or selection for heterozygotes.

Ex 1.4 Partitioning Genetic Diversity

In Ex 1.3 you calculated the observed and expected heterozygosity for each of the populations P1 and P2 for the bush rat populations P1 and P2. Here we continue with this data set.

Step 1. For ease of calculations transcribe your data and answers from Ex 1.2 and Ex 1.3 into the reorganized table below.

Allele Frequency			
Allele	**P1**	**P2**	**Total**
109	0.150	0.150	0.150
117			
125			
129			
133			
137			
141			
145			
149			
Heterozygosity			
H_o			
H_e			
Mean H_o			
Mean H_e			
H_T			

Step 2. Calculate the mean H_o as the average of H_o across P1 and P2 and enter the value in the table.

Step 3. Calculate the mean H_e as the average of H_e across P1 and P2 and enter the value in the table.

Step 4. Calculate H_T as the expected heterozygosity of the total (using the total allele frequencies) and enter the value in the table.

Box 1.3 Genetic Diversity Within and Among Populations

For codominant genetic data at a single locus, the total genetic diversity (heterozygosity) can be divided into within and among populations as follows (based on Hartl and Clark 1989, with some modification of notation):

$\overline{H}_O$ = Observed heterozygosity averaged across subpopulations.

$\overline{H}_e$ = Expected heterozygosity averaged across subpopulations.

H_T = Total expected heterozygosity (calculated as if all the subpopulations were pooled).

$$\overline{H}_o = \sum_{i=1}^{k} H_o / k$$

Where H_O= observed heterozygosity in subpopulation i, and k is the number of subpopulations.

$$H_e = 1 - \sum_{i=1}^{h} p_{i,s}^2$$

$$\overline{H}_e = \sum_{i=1}^{k} H_e / k$$

Where H_e is the expected heterozygosity within subpopulation s, and $p_{i,s}$ is the frequency of the i-th allele in subpopulation s. The summation of the allele frequency squared is over all i-th alleles to h the max number of alleles.

$$H_T = 1 - \sum_{i=1}^{h} p_{Ti}^2$$

Where H_T is the total expected heterozygosity, and p_{Ti} is the frequency of allele i over the total population. If subpopulation sample sizes are equal then $p_{Ti} = \overline{p}_i$, where $\overline{p}_i$ is the frequency of allele i averaged over the subpopulations of equal size.

Ex 1.5 Calculating F-statistics

This exercise is a continuation of Ex 1.4. Using the values for *Mean* H_o, *Mean* H_e and H_T we can easily calculate Wrights F-statistics by the formula in Box 1.4.

Mean H_o	
Mean H_e	
H_T	
F_{IS}	
F_{IT}	
F_{ST}	

Step 1. Calculate F_{IS}, show your working out below and enter your answer in the table.

Step 2. Calculate F_{IT}, show your working out below and enter your answer in the table.

Step 3. Calculate F_{ST}, show your working out below and enter your answer in the table.

Step 4. Check that your answers fit the relationship between F_{IS}, F_{IT} and F_{ST} shown at the bottom of Box 1.4. Answer questions 1 and 2.

Q 1.5 Questions

1. Did you detect genetic differentiation between P1 and P2? Is this differentiation significant?

2. Did you expect to obtain negative values for F_{IS} and F_{IT}? If not, why not? How might you explain this outcome?

Tip: You can check your hand calculations using GenAlEx by following the instructions outlined in Ex. 1.9. If you are a new user of GenAlEx, please continue to read the essential background to GenAlEx and to complete Ex 1.6 to 1.8 first.

Box 1.4 F-Statistics

Perhaps the most widely reported statistics in population genetics are Wright's F-statistics (Wright 1946, 1951, 1965). One way to calculate these statistics is to use the partition of genetic diversity (heterozygosity) described in Box 1.3 as the starting point.

It may come as a surprise to learn that differences within versus among subpopulations can be characterised by F-statistics, since these statistics are normally associated with inbreeding. However, this is possible because population subdivision is associated with inbreeding like effects viz. excess homozygosity (reduction of heterozygosity).

F_{IS} = The inbreeding coefficient within individuals relative to the subpopulation. It measures the reduction in heterozygosity of an individual due to non random mating within its subpopulation.

$$F_{IS} = \frac{\overline{H}_e - \overline{H}_o}{\overline{H}_e}$$

F_{IT} = the inbreeding coefficient within individuals relative to the total. This statistic takes into account the effects of both non random mating within subpopulations and genetic differentiation among the subpopulations.

$$F_{IT} = \frac{H_T - \overline{H}_o}{H_T}$$

F_{ST} = the inbreeding coefficient within subpopulations relative to the total. This statistic provides a measure of the genetic differentiation between subpopulations. That is, the proportion of the total genetic diversity (heterozygosity) that is distributed among the subpopulations. F_{ST} is almost always greater than (or equal to zero). If all subpopulations are in Hardy-Weinberg equilibrium with the same allele frequencies, $F_{ST} = 0$. Note that F_{ST} as calculated in this way is equivalent to G_{ST}

$$F_{ST} = \frac{H_T - \overline{H}_e}{H_T}$$

F-statistics are related according to the following equation:

$$(1 - F_{IS})\,(1 - F_{ST}) = (1 - F_{IT}).$$

Box 1.5 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. However, 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. Therefore, in modern population genetic procedures a more important question is whether we can detect significant genetic differentiation ($F_{ST} > 0$) or not, and whether this differentiation is biologically meaningful. Procedures such as AMOVA allow for such statistical tests.

Getting Started in GenAlEx

Before you Start

Before you can work with GenAlEx you need the following:

1. Scored genetic data
2. Knowledge of whether the data are binary (haploid), binary (diploid), haploid or codominant
3. For spatial genetic analysis you also require geographic data for individuals and populations
3. Microsoft Excel installed on your computer.

Installation

GenAlEx is provided as an Excel add-in, a compiled module and the associated **GenAlEx** menu. Your download file may initially be in the zipped format. Use the extract option to unzip the download and save the files to a dedicated folder of your choice. You can work with GenAlEx directly from this folder. Please refer to the Read Me file distributed with GenAlEx for detailed installation instruction for different versions of Excel on both PC and Macintosh.

From GenAlEx 6.4 onwards only a single version of GenAlEx will be distributed which can be run in Excel 2003 through to Excel 2010 on the PC (Mac users, please see comment below). GenAlEx 6.4 now automatically assesses whether the current worksheet has 256 columns (.xls) or 16,384 columns (*.xlsx), allowing you to seamlessly use both *.xls and *.xlsx files from within Excel 2007/2010, or to use the same version of GenAlEx in both Excel 2003, or Excel 2007/2010 (but not simultaneously, unless using different copies).*

Notes for Macintosh Users

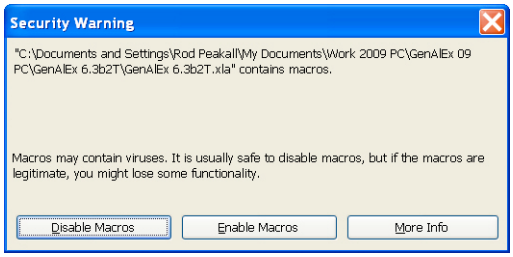
On Macintosh computers, GenAlEx can be run in Excel 2004 and Excel 2011, but not Excel 2008. This is because Microsoft removed the ability of Excel 2008 to run Visual Basic for Applications (VBA), the macro language of Microsoft Office.

Loading GenAlEx in Excel 2003

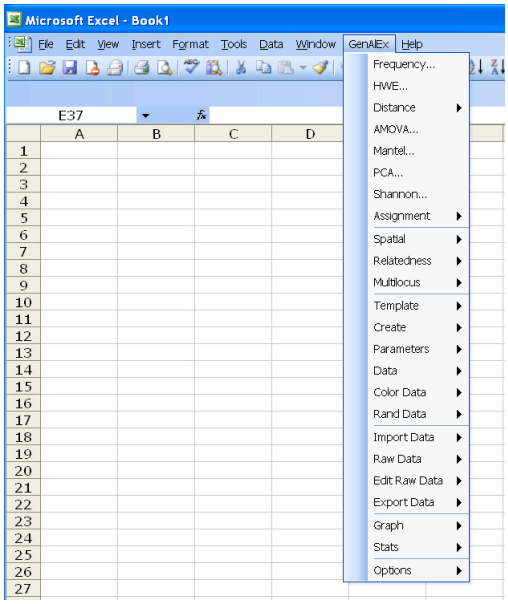
1. Copy the GenAlEx Add-in (e.g. *GenAlEx 6.5.xla*) to your choice of location on your computer. This should preferably be in a dedicated folder.
2. Launch MS Excel. Choose ***Open*** from the **File** menu, locate the GenAlEx Add-in, then click the *Ok* button.

Tip: Alternatively, you can launch GenAlEx and Excel simultaneously by clicking directly on the GenAlEx Add-in file.

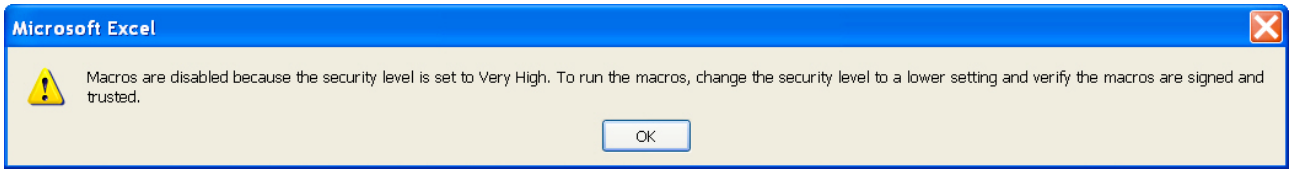
3. Depending on the settings of your Excel program, Excel may warn you that GenAlEx contains macros. Click the *Enable* button to proceed.



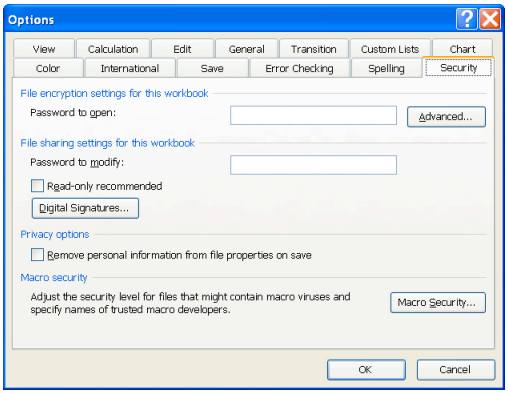
4. In a few seconds the GenAlEx splash screen will, appear click to hide, the **GenAlEx** menu will appear in the Excel menu bar, just before the **Help** menu.

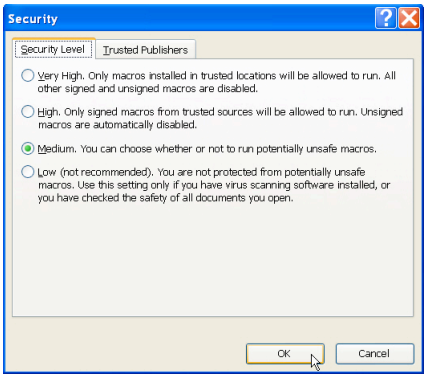


5. If you do not see the security warning dialog box and/or GenAlEx does not launch, you may see the message below instead. In this case, continue to step 6.



6. From the Excel menu **Tools**, choose ***Options***. At the options dialog box click the *Macro Security* button on the *Security* tab. In the next dialog box, choose *Medium* for the security level. Now return to step 1 to launch GenAlEx.





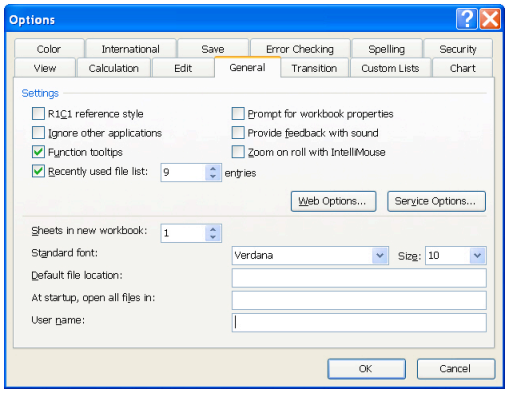
Optimizing Font Size for GenAlEx in Excel 2003

GenAlEx output is optimized for a font size of 10 pt. If not already, you should therefore set the Excel default font size to this setting.

Note that this setting does not override the font settings embedded in existing worksheets. To ensure GenAlEx output is in 10 pt, you may wish to copy data (or data sheets) to new workbook that has been created after the Excel default font is set to 10 pt. To check and change the standard font size set in Excel, select the Check Font option from the Options menu in GenAlEx. Alternatively, the font size can be changed via the excel menus, see below.

Step 1. From the Excel menu **Tools**, choose ***Options***.

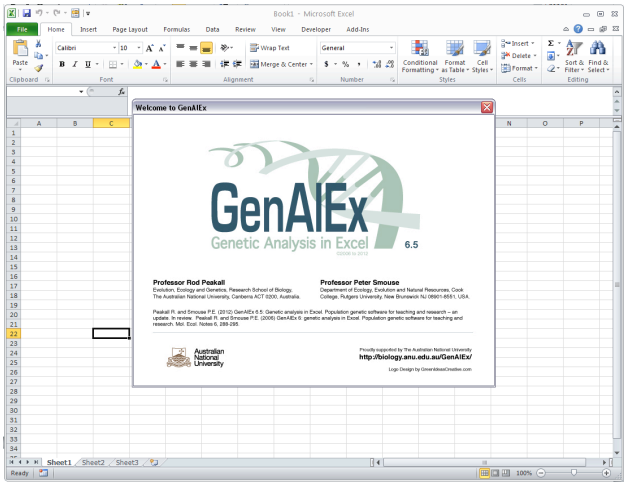
Step 2. At the options dialog box click the *General* tab. Set the standard font size to 10 pt.



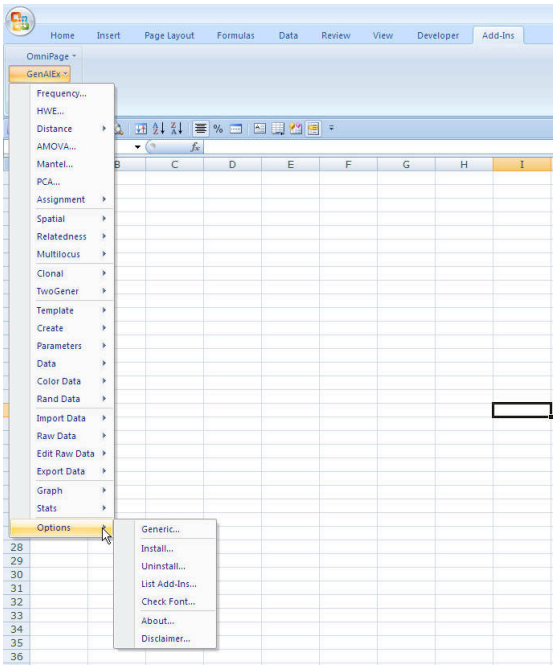
Loading GenAlEx in post-Excel 2007

1. Launch Excel.
2. Open the GenAlEx Add-in (e.g. *GenAlEx 6.5.xla)* via the Excel tab **File**.
3. When prompted, by the Security Notice, choose *Enable Macros*.

4. Shortly the GenAlEx splash screen will appear.



5. Click the Add-Ins tab to show the Add-Ins ribbon. The GenAlEx menu will appear on the right, along with any other installed Add-Ins you may be running. When using GenAlEx you may find it convenient to turn off 'Minimize Ribbon' so that the GenAlEx menu is always accessible when the Add-Ins ribbon is shown. You can access this option by right-clicking the ribbon.

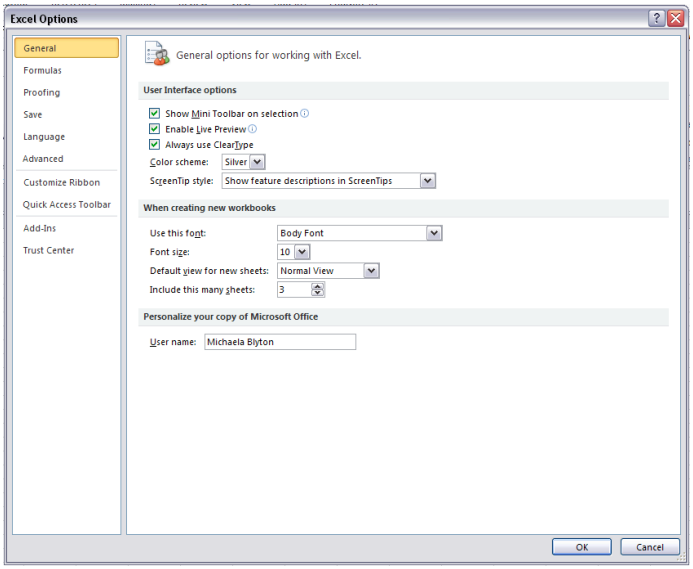


Optimizing Font Size for GenAlEx in post-Excel 2007

GenAlEx output is optimized for a font size of 10 pt. If not already, you should therefore set the Excel default font size to this setting.

Note that this setting does not override the font settings embedded in existing worksheets. To ensure GenAlEx output is in 10 pt, you may wish to copy data (or data sheets) to new workbook that has been created after the Excel default font is set to 10 pt. To check and change the standard font size set in Excel, select the Check Font option from the Options menu in GenAlEx. Alternatively, the font size can be changed via the excel menus, see below.

1. Click the ***File*** tab, then ***Options***.
2. Set the default Font Size for new workbooks to 10 pt in the *General* tab.



Understanding GenAlEx Data Formats

Input

Input consists of raw data or distance matrices in appropriate GenAlEx format (see below). In order to proceed with an analysis the worksheet containing the data must be activated (visible as the current sheet). Some analyses and procedures take several worksheets as input. Unless otherwise explained, these need to be placed starting on the left hand side (LHS) of the workbook, in the order 1 to n.

Wherever possible, GenAlEx offers two options to help users keep track of data and analysis output. In the initial Data Parameter dialog box for statistical procedures, the user may provide a worksheet prefix to help identify the output of a particular analysis, and a title for the output that can provide specific details of the analysis being performed. This title will appear at the top of each output worksheet. It is strongly recommended that both these options be used.

Output

GenAlEx can generate many worksheets in routine analysis, so the ability to create and manipulate new workbooks and new worksheets within workbooks is particularly important. Each worksheet output by GenAlEx is given a name dependent on the analysis performed. This is particularly useful in analyses that have multiple worksheet outputs. In this document (and the GenAlEx 6.5 guide) worksheet names are identified using square brackets e.g. [GD]. A user-defined prefix may be added to the worksheet name for further clarity.

Output of GenAlEx worksheets is designed so that the raw data or other input worksheet is always at the extreme left hand side (LHS) of the workbook. Thus, output worksheets for most menu options will appear to the right hand side (RHS) of the raw data worksheet. However, Genetic Distance outputs will appear to the LHS of the raw data, as the distance matrix is used as input for subsequent analyses.

Graphs are output in standard Excel format and may need to be resized in order to see all the information. All graphs can be edited using standard Excel functions.

Note: By default GenAlEx automatically saves the active workbook at the completion of each analysis. It is strongly recommend that you save a copy of your original data in a separate workbook before manipulating or analysing that data in GenAlEx.

Sample Labels

If you plan to take advantage of all the features of GenAlEx, each sample must be given a unique numerical identifier. Sample names may carry an alpha character prefix, but this must be the same (including case) for all samples in a single dataset. In this case it is important to know that when sorting on alphanumeric data, GenAlEx uses the Excel sort-order rules, sorting character by character, (e.g. A11will come after A100). For ease of sorting, we recommend that the format A001…A199 be used when using prefixes.

Note: This strict requirement for unique numerical identifiers is not essential for running most of the population genetic analyses. However, it is required for many of the useful data manipulation options.

Tip: If your samples are not in this format, it is possible to quickly create unique numerical identifiers using the ***Replace Sample code*** *option under the* ***Raw Data*** *menu in GenAlEx.*

Data Parameters and Labels

Data parameters and labels are crucial for telling GenAlEx how to read and analyze the data. GenAlEx stores all parameters and labels in rows 1, 2 and 3 of the data worksheets. For raw data, columns 1 and 2 are generally used for sample and population labels respectively; while, actual data begins in Cell C4 of a worksheet.

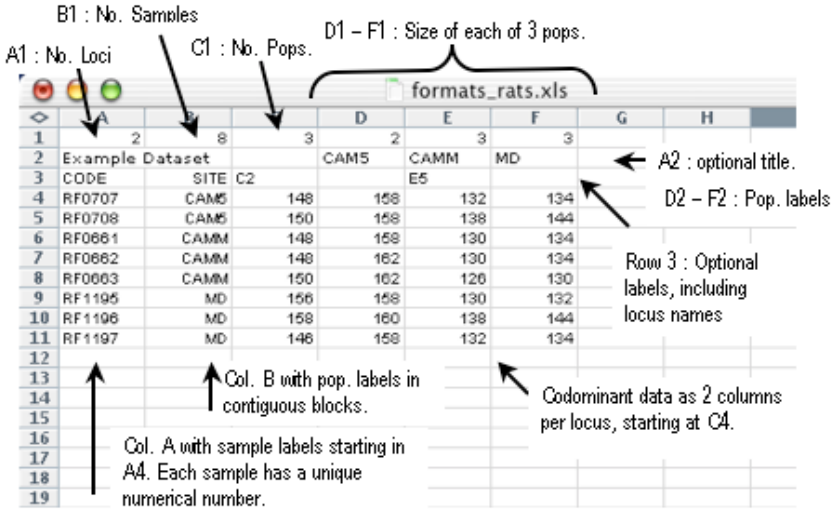
Note: When analysing your data, GenAlEx only uses the data parameters to locate and process your samples. It does not interrogate the sample or population codes in columns 1 and 2. Therefore, ensure the data parameters reflect the data format, particularly after sorting or rearranging your samples.

Data parameters and labels may be entered in GenAlEx in several ways

1. A worksheet containing data may be manually formatted to provide appropriate parameters.
2. The ***Template*** *option in the* **GenAlEx** *menu* may be used to provide parameters through a dialog box, creating a formatted worksheet into which the data are then entered (see section below for further instructions).
3. The ***Parameters*** option in the **GenAlEx** *menu may be used to obtain the relevant parameter values from an existing dataset and insert them into their appropriate location (*see section below for further instructions). This option requires that your data is bounded by blank columns and rows.
4. On initiating an analysis, GenAlEx prompts for the relevant parameters in a dialog box. Changing parameters in this box provides an easy way to select subsets of data for analysis.
5. If data is imported using GenAlEx options, essential parameters and labels will be inserted automatically, however labels for locus names may need to be entered manually.

Parameter locations

Essential parameters are inserted into Row 1. They are: No Loci (cell A1); No. Samples (cell B1); No Populations (cell C1); The size of each population (cell D1..to cell n1). If regional information is required, the parameter for the No. of Regions is inserted into the cell immediate after the last population size, and the size of each region then follows in subsequent cells (see example under format for codominant data below).



	A	B	C	D	E	F	G	H
1	2	8	3	2	3	3		
2	Example Dataset			CAM5	CAMM	MD		
3	CODE	SITE	C2		E5			
4	RF0707	CAM5	148	158	132	134		
5	RF0708	CAM5	150	158	138	144		
6	RF0661	CAMM	148	158	130	134		
7	RF0662	CAMM	148	162	130	134		
8	RF0663	CAMM	150	162	126	130		
9	RF1195	MD	156	158	130	132		
10	RF1196	MD	158	160	138	144		
11	RF1197	MD	146	158	132	134		
12								
13								
14								
15								
16								
17								
18								
19								

Data Formats

GenAlEx accepts 4 types of numerically-coded data:

1. Codominant genotypic data with 2 columns per locus.
2. Dominant (Binary), Haploid (including Haplotypes), or Sequence data coded numerically with 1 column per locus/base.
3. Codominant and Haploid raw allele frequency data.
4. Geographic data with 2 columns for X and Y coordinates.

Tip: GenAlEx also allows you to work with DNA sequences in 2 different formats, however, for most analyses the sequence needs to be coded numerically by options provided in GenAlEx. After conversion to numeric format, sequence data are treated like all other haploid data.

Format for codominant data

Codominant data are presented as two columns per locus as in the figure below. Alleles may be simply numerically-coded (1, 2, 3 etc). Alternatively, and preferably for microsatellite data, alleles may be coded as their integer size in base pairs (bp), or as the inferred number of simple sequence repeats. These last two formats are essential for calculation of the distance measure, R_{ST}. There is a limit of 999 numerically-coded alleles. Codominant alleles need not be numbered consecutively.

Example of codominant, numerically-coded data, with regional parameters.

CodominantData.xls

	A	B	C	D	E	F	G	H	I	J
1	2	8	4	2	2	2	2	2	4	4
2	Example of Codominant Data			Pop 1	Pop 2	Pop 3	Pop 4		Region 1	Region 2
3	Sample N	Pop.	Locus 1		Locus 2					
4	1	Pop 1	3	3	1	4				
5	2	Pop 1	2	3	2	4				
6	3	Pop 2	2	4	3	4				
7	4	Pop 2	1	4	1	3				
8	5	Pop 3	3	4	2	2				
9	6	Pop 3	1	2	3	3				
10	7	Pop 4	4	4	2	2				
11	8	Pop 4	2	4	4	4				
12										

In this example the 4 populations are split into 2 regions with Pops 1 & 2 in Region 1 and Pops 3 & 4 in Region 2. Note the regional parameters are only required for AMOVA.

Example of codominant microsatellite data, with genotypes by fragment size.

Formats.xls

	A	B	C	D	E	F	G
1	2	8	2	4	4		
2	Codominant data - fragment size			EC	TT		
3	Sample no.	Pop	CA2		GA8		
4	HE001	EC	294	298	274	274	
5	HE002	EC	292	300	256	258	
6	HE003	EC	296	298	258	258	
7	HE004	EC	298	300	258	258	
8	HE010	TT	298	298	256	256	
9	HE011	TT	292	296	256	260	
10	HE012	TT	296	296	254	256	
11	HE013	TT	292	296	214	248	
12							

Format for dominant, haploid or sequence data

Dominant, haploid (including haplotypes) or sequence data are presented as a single column per locus. Haploid data can be coded numerically from 1…n, or each may be represented by multiple variable sites (columns 1 … n), with multiple states. For sequence or SNP data the bases are numerically coded as follows: A=1, C=2, G=3, T=4, :=5; -=5, all other characters = 0. GenAlEx provides several options for the import of sequence data and auto conversion to numbers.

Example of dominant, or binary data.

Formats.xls

	A	B	C	D	E
1	2	6	2	3	3
2	Example of binary data			Pop 1	Pop 2
3	Sample No.	Pop.	Locus 1	Locus 2	
4	1	Pop 1	1	0	
5	2	Pop 1	0	1	
6	3	Pop 1	1	0	
7	4	Pop 2	0	0	
8	5	Pop 2	1	1	
9	6	Pop 2	1	1	
10					

Example of sequence data, coded numerically at multiple variable sites.

Formats.xls

	A	B	C	D	E
1	3	6	2	3	3
2	Example of haplotype data			Pop 1	Pop 2
3	Sample No.	Pop.	bp42	bp67	bp114
4	1	Pop 1	1	1	2
5	2	Pop 1	1	1	2
6	3	Pop 1	3	1	2
7	4	Pop 2	1	1	4
8	5	Pop 2	1	3	4
9	6	Pop 2	1	3	4
10					

Example of haplotype data, with individual haplotypes coded numerically.

Formats.xls

	A	B	C	D	E	F
1	1	6	2	3	3	
2	Example of haplotype data			Pop 1	Pop 2	
3	Sample No.	Pop.	cpDNA			
4	1	Pop 1	1			
5	2	Pop 1	1			
6	3	Pop 1	2			
7	4	Pop 2	3			
8	5	Pop 2	4			
9	6	Pop 2	4			
10						

These haplotypes correspond to the sequences shown in the previous example.

Format for geographic data

For convenience, both geographic and genetic distances can be calculated in a single analysis. Coordinates can be entered as either integer or decimal numbers.

X and Y coordinates may be read by GenAlEx from two different formats.

1. X / Y data are located in the same worksheet as the genetic data, and separated from the genetic data by a single blank column. This format is used by GenAlEx for various analyses, including Genetic Distance.

Example of geographic data after genetic data.

Formats.xls

	A	B	C	D	E	F	G	H	I	J
1	2	6	2	3	3					
2	Geographic data			CAM5	MD					
3	CODE	SITE	C2	C2	E5	E5		X	Y	
4	RF707	CAM5	148	158	132	134		670	750	
5	RF708	CAM5	150	158	138	144		150	750	
6	RF709	CAM5	156	158	116	132		510	750	
7	RF1160	MD	148	158	138	144		565	357	
8	RF1161	MD	148	158	126	132		235	537	
9	RF1162	MD	158	160	136	138		340	488	
10										

2. In a separate worksheet, in columns 3 and 4. In this case, the sample and population labels in columns 1 & 2 will correspond exactly to those for the genetic data.. This format is required for analyses such as the 2D spatial autocorrelation.

Example of geographic data in columns 3 & 4.

Formats.xls

	A	B	C	D	E	F
1	2	6	2	3	3	
2	Geographic data			CAM5	MD	
3	CODE	SITE	X	Y		
4	RF707	CAM5	670	750		
5	RF708	CAM5	150	750		
6	RF709	CAM5	510	750		
7	RF1160	MD	565	357		
8	RF1161	MD	235	537		
9	RF1162	MD	340	488		
10						

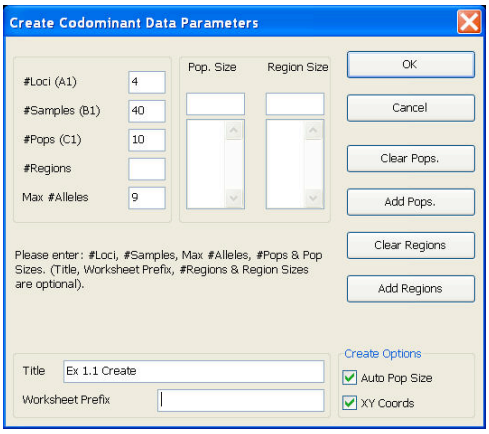
Missing Data

Virtually all GenAlEx options handle missing data. However, missing data can be particularly problematic for pairwise distance-based analyses such as AMOVA, Mantel and spatial autocorrelation. Therefore, a unique option for interpolating missing individual-by-individual pairwise distances is provided. This action will insert the average genetic distances for each population level pairwise contrast e.g. within Pop. 1, or between Pop. 1 and Pop. 2. Nonetheless, in order to avoid excessive bias, large numbers of missing data for individual-based distance calculations should be minimized.

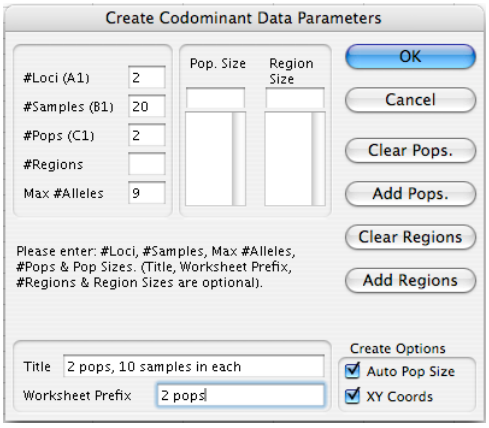
Codominant and Haploid missing data are coded as '0'. Missing Binary data are coded as '-1'.

Using *Create* to Learn about GenAlEx Data Formats

In this section you will use the ***Create*** menu option to learn about GenAlEx data formats. This menu provides options to create random examples of all GenAlEx data formats, both Genetic and Geographic. These datasets are useful for exploring the range of GenAlEx procedures.



Create dialog box on a PC



Create dialog box on a Macintosh

Ex 1.6 Using *Create* with Auto Pop Size

In this first exercise we will take advantage of the *Auto Pop Size* feature in GenAlEx that will automatically generate even pop sizes, for the number of samples and populations you specify.

Step 1. Before you proceed, randomly choose a set of numbers within the specified range as follows, and record them below:

The number of loci (suggested range 1 to 10) =

The number of samples (suggested range 10 to 40, and evenly divisible by the number of pops chosen below) =

The number of populations (suggested range 2 to 10) =

The number of alleles (suggested range 4 to 9) =

Step 2. With a workbook open, choose the option ***Create*** from the **GenAlEx** menu, and select the ***Codominant*** submenu.

Step 3. In the Create Data Parameters dialog box enter the number (#) of loci, # samples, # populations and # Alleles, as chosen above.

Step 4. Check the *Auto Pop Size* and *XY Coords* options on the dialog box.

Step 5. Inspect the data sheet generated by GenAlEx. By reference to the numbers you jotted down, identify the location of the parameters in the data sheet and study the format for the genotypes and XY coordinates.

Ex 1.7 Using *Create* with Variable Pop Sizes

In this second exercise you will be given the option to manually enter variable pop sizes.

Step 1. Choose a new set of numbers as for Ex 1.6. Also choose a set of pop sizes that add to the total number of samples you have chosen. Jot down the numbers you have chosen, then proceed.

Step 2. With a workbook open, choose the option ***Create*** from the **GenAlEx** menu, and select the ***Codominant*** submenu.

Step 3. In the Create Data Parameters dialog box enter the number (#) of loci, # samples, # populations and # Alleles required.

Step 4. Enter the size of each pop in the edit box below 'Pop. Size', and add to the population list using the *Add Pops* option.

Step 5. Uncheck the default *Auto Pop Size* and *XY Coords* options on the dialog box.

Step 6. Inspect the data sheet generated by GenAlEx. By reference to the numbers you jotted down, identify the location of the parameters in the data sheet and study the format for the genotypes.

Ex 1.8 Using *Create* with Other Data Types

Now that you are up and running, use the ***Create*** option to generate some random demonstration data for other types of GenAlEx data formats, such as haploid or binary.

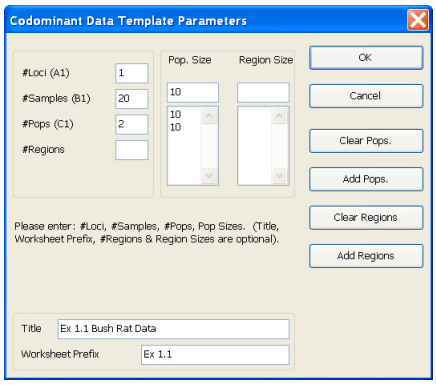
Tip: The Create option is a great way to troubleshoot the occasional data analysis problem you might encounter in GenAlEx. Suppose you have a large data set that GenAlEx is unable to analyse. Often you will be given a warning by GenAlEx to check your data and parameters. Sometime such a problem is associated with something unusual about your data, rather than the parameters. To check whether or not this is the case, simply use the **Create** *option to generate a data set of the same size (No Samples, No of Pops, Pops Size, No Loci etc). Now perform the required analysis in GenAlEx. If the created data set runs, check your own data carefully. Look for things like missing data for all samples in a population at a specific locus. Such a case might trigger an error. Check for unusual data values that might be typos etc.*

Ex 1.9 Using *Template* as a Starting Point for Data Entry

The ***Create*** option is provided in GenAlEx primarily for users to create examples of the formats of the various data types that can be analysed by the software. If you are entering small data sets by hand, and you know how many samples, and from which populations they come, you can take advantage of the ***Template*** option.

Here we return to the microsatellite data set scored in Ex 1.1. First you will use the ***Template*** option to set up the data for entry out of the completed table in Ex 1.1 Once your data has been entered you can use GenAlEx to check your answers to the hand calculations in Ex 1.2 and 1.5.

Step 1. Open a new Excel document. Now use the GenAlEx option ***Template->Codominant*** to quickly setup the table for data entry. What parameters will you use for this data set? How many samples? How many populations? How many loci?

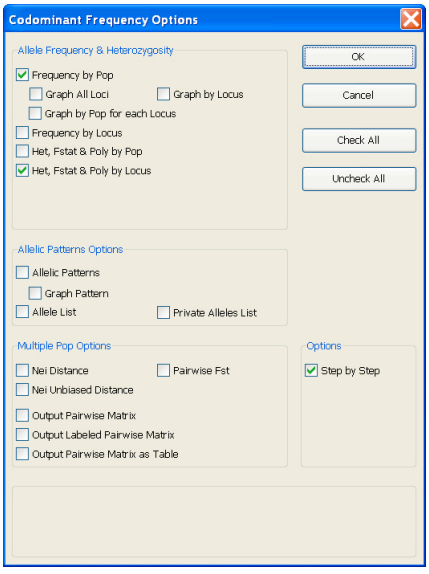


	A	B	C	D	E	F	G
1	1	20	2	10	10	1	20
2	Ex 1.1 Bush Rat Data			Pop1	Pop2		Region1
3	Sample	Pop	Locus1				
4	1	Pop1					
5	2	Pop1					
6	3	Pop1					
7	4	Pop1					
8	5	Pop1					
9	6	Pop1					
10	7	Pop1					
11	8	Pop1					
12	9	Pop1					
13	10	Pop1					
14	11	Pop2					
15	12	Pop2					
16	13	Pop2					
17	14	Pop2					
18	15	Pop2					
19	16	Pop2					
20	17	Pop2					
21	18	Pop2					
22	19	Pop2					
23	20	Pop2					

The Template dialog box and the template created shown to the left

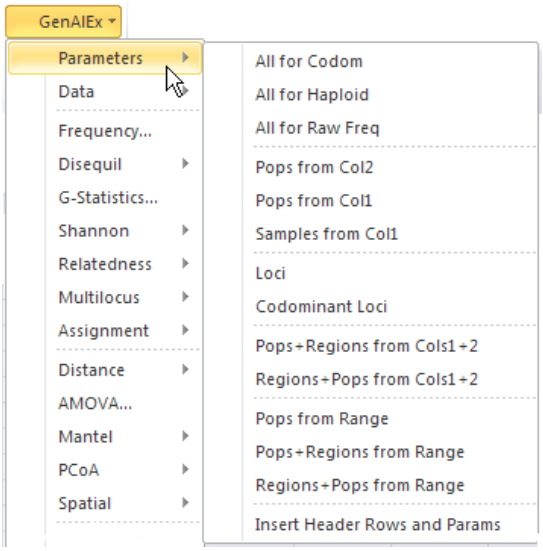
Step 2. Enter the genotypes for all samples from Ex 1.1. Name the worksheet 'Ex 1.1 Data' and save the workbook with the name *Ex 1.1 Rats.xls*.

Step 3. To check your hand calculations using GenAlEx, first activate the worksheet containing the Ex 1.1 Data. Next, choose ***Frequency*** from the **GenAlEx** menu then select *Freq by Pop, Het, Fstat & Poly by Locus* and *Step by Step* in the Codominant Frequency Options dialog box.



Step 4. Compare your hand calculations with GenAlEx. How did you go?

GenAlEx Data Parameters



Parameter Submenu on PC

The Parameters option provides a quick and easy way to obtain the necessary GenAlEx parameters from a pre-existing dataset, and insert them in their correct locations. Data must be in standard GenAlEx format with data starting in cell C4. For the ***All for Codom*** and ***All for Haploid*** options samples labels must also be in column 1 and population labels in column 2. The dataset needs to be bounded below by an empty row and to the right by an empty column, as GenAlEx uses empty cells to identify the data limits. All samples per population must have the same population label, and be in a contiguous block. For each menu sub-option, GenAlEx will interrogate the appropriate column(s) and insert the corresponding parameters in their correct locations. An option to insert the header rows into an unformatted dataset is also provided.

Always remember that the ***Parameters*** menu option requires:

1. Data to be in standard GenAlEx format, with data starting in cell C4.
2. Data to be bounded by an empty row below the last sample and an empty column at the right of the last locus entry.
3. All samples within a population must have the same population label, and be in a contiguous block.

Ex 1.10 Getting Population Parameters

Real genetic data sets may be imported into GenAlEx directly from genotyping software or other data sources. In these cases, GenAlEx can determine the parameters for you provided you follow the rules for GenAlEx formats. In this exercise you are provided with a real codominant genetic data set from a study of bush rats by Peakall and Lindenmayer (2006).

Step 1. Open the workbook called *Ex 1.10 Bush Rats Raw Data*, read the info provided then activate the worksheet containing the data.

Step 2. Inspect the data provided. Note it has not yet been formatted for GenAlEx analysis.

Step 3. Convert the data into codominant GenAlEx format.

Hint: This may require the addition and deletion of rows or columns.

Step 4. Using GenAlEx, automatically obtain parameters for the data set by selecting ***All for Codom*** from the ***Parameters*** menu and record your answers below:

The number of codominant loci =

The number of samples =

The number of populations =

The name and number of samples in the first population =

The name and number of samples in the last population =

Using *Data* to Work Efficiently

The ***Data*** menu option offers several commands for quickly manipulating your dataset. In all cases, Data must be in appropriate GenAlEx format (including parameters). Two useful options are:

Sort on Sample (Col1): Sorts the entire dataset on column 1(normally containing the sample labels), according to the Excel sort-order rules (see the 'Sample Labels' section). The sample and population parameters are automatically inserted after the data is sorted; GenAlEx assumes population codes are in column two.

Sort on Pop (Col2): Sorts the entire dataset on column 2 (normally containing the population labels), according to the Excel sort-order rules (see the 'Sample Labels' section). The sample and population parameters are automatically inserted after the data is sorted; GenAlEx assumes population codes are in column two.

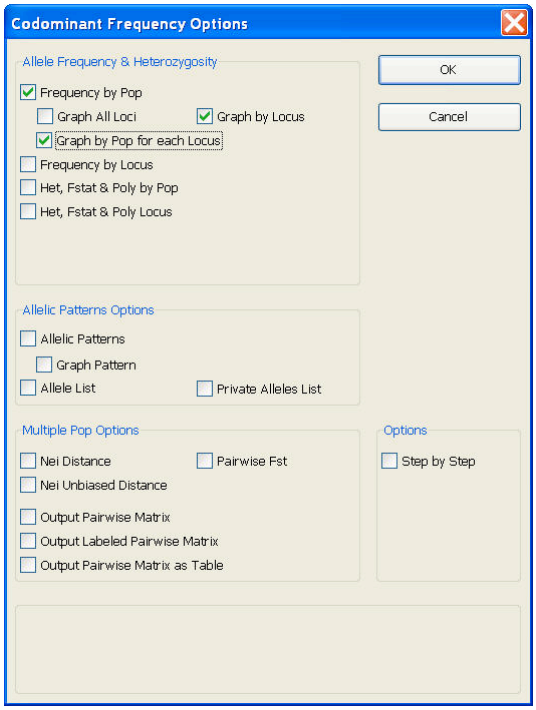
Data Exploration and Allele Frequencies

Genetic data exploration is the first step of any population genetic analysis. GenAlEx provides some powerful graphic tools to aid this important first step. The calculation of allele frequencies and various summary statistics such as the number of different alleles, observed and expected heterozygosity (or equivalent diversity estimates for haploid data) represent critical baseline statistics that should be reported in every population genetic study. However, even before further analysis, let alone publication of results, inspecting the outcomes of allele frequencies and summary statistics is important for identifying problems that might be attributable to incorrect scoring of your DNA profiles, or errors in data entry. If you find unexpected results at this stage, and you can rule out error, this data exploration step can also reveal interesting genetic patterns that might provide unexpected insights into the biology of your study species.

Ex 1.11 Plots of Allele Frequency

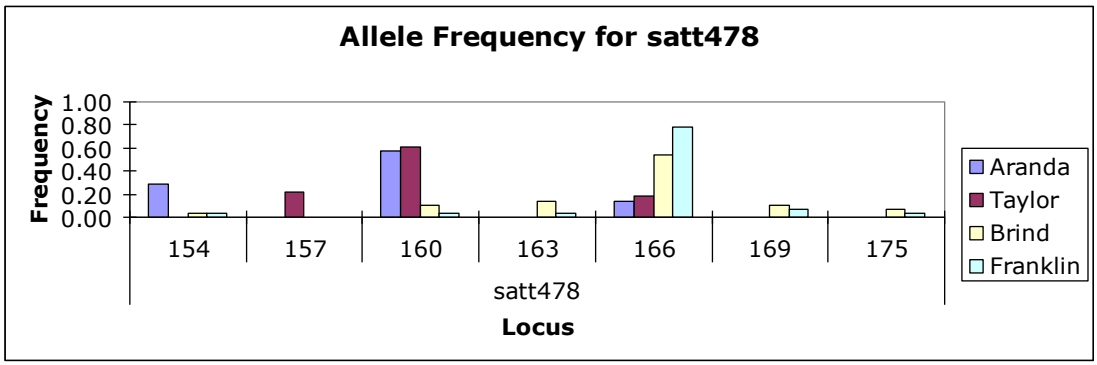
Preliminary genetic data exploration and the calculation of allele frequency are intertwined. The ***Frequency*** menu option in the **GenAlEx** menu is the entry point for this analysis. The small data set in this exercise is drawn from the plant *Glycine clandestina*, an Australian native relative of the soybean. This species has an unusual reproductive biology - it produces two kinds of flowers: Normal 'Open pollinated' flowers and 'Closed or cleistogamous' flowers. The open flowers are typical of pea flowers in general requiring insect pollinators for seed set. The 'Closed or Cleistogamous' flowers are adapted to self pollination, regularly producing seed without the aid of pollinators. The seeds of the species lack an obvious dispersal mechanism and it appears most seeds fall close to the parent plant. Data are provided for 4 populations, two from within Canberra (Aranda and Taylor) and two from the Brindabella range (Brind and Franklin) some 50 km west of Canberra.

Step 1. Return to Excel and open the workbook *Ex 1.11 Glycine,* then activate the *Data* worksheet. Next, choose ***Frequency*** from the **GenAlEx** menu. You will first be prompted with the Allele Frequency Data Parameters dialog box. Click *OK*. Next select *Freq by Pop* and check the options *Graph by Locus* and *Graph by Pop for each Locus* in the Codominant Frequency Options dialog box.

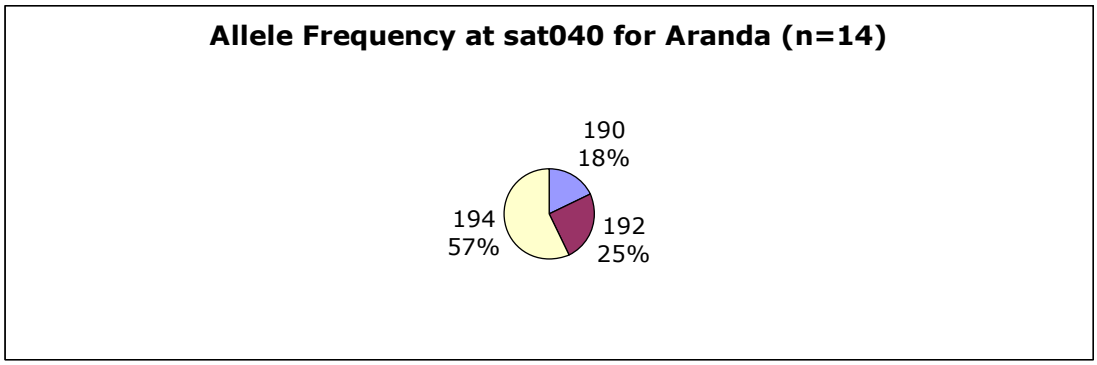


Step 2. Inspect the outcomes in the three worksheets with suffixes [*AFP*], [*AGF*] and [*AGP*] and answer the questions below. Two example graphs generated by the analysis are shown below.

Example of Allele Frequency Graph by Locus



Example of Allele Frequency Graph by Pop



Q 1.11 Questions

1. Based on the allele frequency patterns, do you predict there is genetic differentiation among the *Glycine* populations? How might you test this prediction?

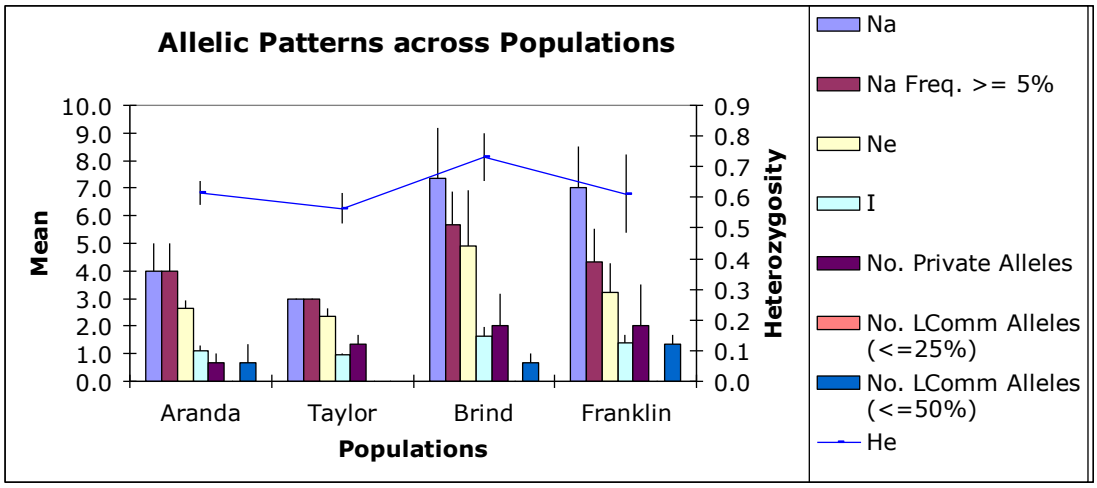
Ex 1.12 Heterozygosity, F-statistics and Allelic Patterns

Inspection of the population allele frequency graphs for the *Glycine* example reveal some interesting patterns among populations and loci. The ***Frequency*** menu option also offers other tools for data exploration and analysis including various Heterozygosity estimates and F-statistics analysis (via allele frequency).

Step 1. Return to the workbook *Ex 1.11 Glycine,* re-activate the *Data* worksheet, then choose ***Frequency*** from the **GenAlEx** menu. You will first be prompted with the Allele Frequency Data Parameters dialog box. Click *OK*. Next, uncheck the options previously run in Ex. 1.9 then check *Het, Fstat & Poly by Pop*, *Het, Fstat & Poly by Locus*, *Allelic Patterns* and *Graph Pattern* in the Codominant Frequency Options dialog box.

Step 2. Inspect the outcomes in the three worksheets with suffixes [*HFP*], [*HFL*] and [*APT*] and answer the questions below. The allelic patterns graph generated by this analysis is shown below (with minor additional modification available in Excel for all GenAlEx graphs).

Example of an Allelic Patterns Graph



Q 1.12 Questions

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

1. Summarize the findings for observed versus expected heterozygosities?

2. What do you conclude about the extent of inbreeding?

3. Did you detect genetic differentiation among the populations? Is the differentiation significant?

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

4. Briefly describe the key allelic patterns that are revealed across the four populations. What biological factors might explain the patterns observed?

Shannon Diversity Indices in Population Genetics

Shannon's diversity index for information theory (Shannon 1948) has been widely employed in ecology but has been less widely used in population genetics. In a recent series of studies, Sherwin et al. (2006) and Rossetto et al. (2008) have shown both by computer simulation and for real data sets that Shannon's Indices offer some ideal statistical properties for measuring biological information across multiple scales from genes to landscapes. In particular, the capacity to apply the indices at multiple scales is unique among the commonly employed population statistics. Furthermore, Shannon's *mutual information* index ${}^{S}H_{UA}$ not only provides a convenient measure of differentiation among populations, but it can be readily converted to the log-likelihood contingency test *G* statistic enabling a convenient chi-square based statistical test for allele frequency differences at each locus for each pairwise combination of populations. Finally, for diploid species with large estimated effective population size ${}^{S}H_{UA}$ can be converted to an estimate of *Nm* (Number of Migrants).

Note: As a test of population differentiation, the G-statistic may have a high Type I error (falsely rejecting the null hypothesis) in some circumstances. Therefore, for research purposes we recommend using the permutational tests offered in GenAlEx 6.5 via the Shannon Partition option.

Tip: For more extensive background to Shannon Diversity see Box 1.6 and Appendix 1.1.

Ex 1.13 Hand Calculation of Shannon's Indices

Shannon's indices are remarkably straightforward to calculate by hand, requiring only knowledge of the allele frequencies and the sample sizes. In this exercise we will work through the steps for calculating these indices drawing on a subset of data from a study of the plant *Glycine clandestina* introduced in earlier exercises. In this case microsatellite genotype data are provided for the locus AG48 for 10 samples each in two Canberra populations, Aranda and Taylor.

Step 1. The raw genotype data are shown below. Inspect the data and answer question 1 before proceeding.

	A	B	C	D	E
1	1	20	2	10	10
2	Glycine at Locus AG48			Aranda	Taylor
3	Sample	Pop.	AG48		
4	1	Aranda	280	280	
5	2	Aranda	272	272	
6	3	Aranda	272	280	
7	4	Aranda	280	280	
8	5	Aranda	272	280	
9	6	Aranda	272	272	
10	7	Aranda	272	272	
11	8	Aranda	280	280	
12	9	Aranda	272	272	
13	10	Aranda	266	272	
14	11	Taylor	270	270	
15	12	Taylor	270	270	
16	13	Taylor	270	270	
17	14	Taylor	270	270	
18	15	Taylor	270	270	
19	16	Taylor	270	270	
20	17	Taylor	270	270	
21	18	Taylor	270	270	
22	19	Taylor	268	270	
23	20	Taylor	272	272	

Step 2. Calculate $wt_1 = ct_1/(ct_1+ct_2)$ = _____ and $wt_2 = ct_2/(ct_1+ct_2)$ = _____ where ct_1 = 20 and ct_2 = 20 (2 x No. of Samples in each Pop).

Step 3. Allele frequencies for the two populations have been calculated for you in the table below. Calculate the weighted mean frequency for each allele as $p_{i1} \times wt_1 + p_{i2} \times wt_2$.

Tip: In this case because sample sizes are the same the weighted mean is simply the arithmetic mean.

	Aranda	**Taylor**	**Mean**	**X**	**Y**	**Z**
Allele	p_{i1}	p_{i2}	$\overline{p}_i$	$-p_{i1}Log_2 p_{i1}$	$-p_{i2}Log_2 p_{i2}$	$-\overline{p}_i Log_2 \overline{p}_i$
266	0.050	0.000				
268	0.000	0.050				
270	0.000	0.850				
272	0.550	0.100				
280	0.400	0.000				
Sum	1.000	1.000				

Step 4. To calculate ${}^{S}H_{A1}$ for the Aranda pop, first compute -1 x p_{i1} x Log_2 p_{i1} for each allele and enter the values in Col X, then sum these values across the 5 alleles.

Step 5. To calculate ${}^{S}H_{A2}$ for the Taylor pop, first compute -1 p_{i2} x Log_2 p_{i2} for each allele and enter the values in Col Y, then sum these values across the 5 alleles.

Step 6. To calculate ${}^{S}H_{U}$ compute -1 x $\overline{p}_i$ x Log_2 x $\overline{p}_i$ for each allele and enter the values in Col Z, then sum these values across the 5 alleles.

Tip: On a calculator to obtain the $\log_2$ of the value p_i you need to calculate $\log(p_i)/\log(2)$. If you are using Excel use the function LOG(number, base), in this case enter '= LOG (p_i, 2)'.

Step 7. Calculate ${}^{S}H_{UA}$ by the formula ${}^{S}H_{UA} = {}^{S}H_{U} - wt_1 \times {}^{S}H_{A1} - wt_2 \times {}^{S}H_{A2}$.

Step 8. Calculate *G* by the formula G=1.3863 x ${}^{S}H_{UA}$ x (ct_1+ct_2) where ct_1 = 20 and ct_2 = 20 (2 x No. of Samples in each Pop).

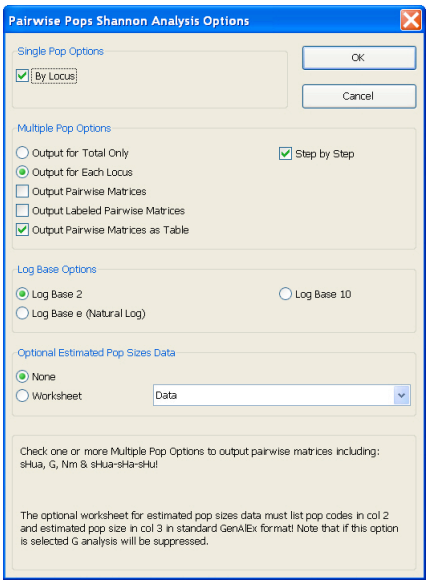
Step 9. Calculate the *DF* as the (No. of populations-1) x (No. of alleles compared – 1) = (2-1) x (5 - 1) = _____.

Step 10. Look up the Chi-Square Probability for the *G*-test given the degrees of freedom *DF* using the table provided in the section on Hardy-Weinberg Equilibrium later in this module. Record your answers from Step 7 to Step 10 in the table below.

Tip: When using Excel you can easily calculate the Chi-Square probability using the function CHIDIST(x,deg_freedom), in this case enter '=CHIDIST(G,DF).

Shannon Statistic	**Value**
${}^{S}H_{A1}$	
${}^{S}H_{A2}$	
${}^{S}H_{U}$	
${}^{S}H_{UA}$	
G	
DF	
Chi-Sq Prob	

Step 11. Check your answers using GenAlEx. The data can be found in the workbook *Ex 1.13 Glycine by Hand.xls*. Choose ***Shannon->Pairwise Pops***. When prompted by the Shannon Analysis Options dialog box choose *By Locus*, *Output for Each Locus*, *Output Pairwise Matrices as Table*, *Step by Step* and *Log Base 2*.



Q 1.13 Questions

1. Based on your inspection of raw genotypes shown above, summarise the patterns of allele frequency differences between the two populations.

2. Do you predict there will be a significant difference in allele frequencies between the two populations?

3. Summarise and interpret the outcomes of the *G*-test for allele frequency differences.

Box 1.6 Shannon's Information Indices

Based on Sherwin et al. (2006) Mol. Ecol. 15, 2857-2869, see also Appendix 1.1.

In general for a specific locus in a given population the Shannon's *Allele Information* index ${}^{S}H_{A}$ is calculated as:

$$ {}^{S}H_{A} = -\sum p_i \log_2 p_i $$

At this specific locus across multiple populations we consider each pairwise combination of populations in turn calculating ${}^{S}H_{A}$ for each of the two populations:

$$ {}^{S}H_{A1} = -\sum p_{i1} \log_2 p_{i1} \text{ and } {}^{S}H_{A2} = -\sum p_{i2} \log_2 p_{i2} $$

Where p_i is the allele frequency of the *i*th allele at the locus in question for the specified population (1 or 2).

Shannon's *Total Information* index across each pair of populations is calculated as:

$$ {}^{S}H_{U} = -\sum \overline{p}_i \log_2 \overline{p}_i $$

Where $\overline{p}_i$ is the average weighted frequency of the *i*th allele for each pair of populations:

$$ \overline{p}_i = p_{i1} \times wt_1 + p_{i2} \times wt_2 $$

$$ \text{Where } wt_1 = \frac{ct_1}{ct_1 + ct_2} \text{ and } wt_2 = \frac{ct_2}{ct_1 + ct_2} $$

and *ct* = the total allele count at the locus (2 x the number of samples for diploids) for the respective populations.

Finally, Shannon's *Mutual Information* index ${}^{S}H_{UA}$ is calculated for each pair of populations as:

$$ {}^{S}H_{UA} = {}^{S}H_{U} - wt_1 {}^{S}H_{A1} - wt_2 {}^{S}H_{A2} $$

Shannon's *Mutual Information* index can now be used to compute the log-likelihood contingency test statistic *G as:*

$$ G = 1.3863 {}^{S}H_{UA}(ct_1 + ct_2) $$

With degrees of freedom *DF* calculated as the (number of populations compared - 1) x (number of alleles compared - 1).

For diploid species with effective population sizes > 500 estimates of *Nm* among pairs of populations can be computed as:

$$ Nm = \left(\frac{0.156}{{}^{S}H_{UA}} \right)^2 $$

Nei Genetic Distance

Ex 1.14 Hand Calculation of Nei's Genetic Distance

While F_{ST} is perhaps the most widely used measure of genetic differentiation among populations, another frequently used estimate of the genetic difference among populations is Nei's Genetic Distance *D*. In this exercise we will utilise the same data set as for Ex 1.13. Record your answers in the Table below.

Step 1. Allele frequencies for the two *Glycine* populations are shown below. Calculate squared allele frequencies for each allele in population x and population y.

Step 2. Calculate *Jx* as the sum of the squared allele frequencies for population x. Calculate *Jy* as the sum of the squared allele frequencies for population y.

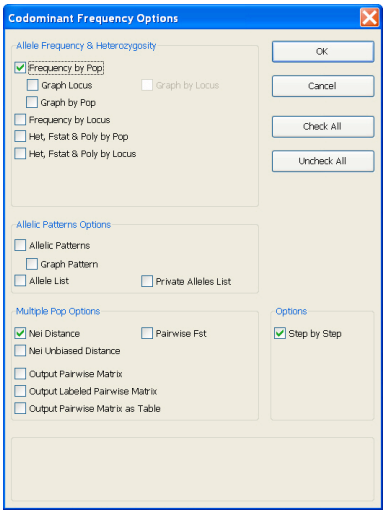
	Aranda	**Taylor**			
Allele	**x**	**y**	x^2	y^2	***xy***
266	0.050	0.000			
268	0.000	0.050			
270	0.000	0.850			
272	0.550	0.100			
280	0.400	0.000			
			Jx	***Jy***	***Jxy***
Sum	1.000	1.000			
Nei I					
Nei D					

Step 3. For each allele calculate the product of allele frequency in population x and population y. Calculate *Jxy* as the sum of the products of allele frequency.

Step 4. Calculate *Nei I* as $Jxy/(JxJy)^{0.5}$.

Step 5. Calculate *Nei D* as –Ln(*I*). Record values in the table above.

Step 6. Check your answers using the ***Frequency*** option in GenAlEx. The data can be found in the workbook *Ex 1.13 Glycine by Hand.xls*. In the Codominant Frequency Options dialog box check *Frequency by Pop*, *Nei Distance* and *Step by Step*.



Box 1.7 Nei's Genetic Identity and Distance

$$Nei\ I = \frac{J_{xy}}{\sqrt{(J_x J_y)}}$$

$$\text{Where}, J_{xy} = \sum_{i=1}^{k} p_{ix} p_{iy},\ J_x = \sum_{i=1}^{k} p_{ix}^2, \text{ and } J_y = \sum_{i=1}^{k} p_{iy}^2.$$

Where I is Nei's Genetic Identity, and p_{ix} and p_{iy} are the frequencies of the i-th allele in populations x and y. For multiple loci, Jxy, Jx and Jy are calculated by summing over all loci and alleles and dividing by the number of loci. These average values are then used to calculate I.

$$Nei\ D = -\ln(I)$$

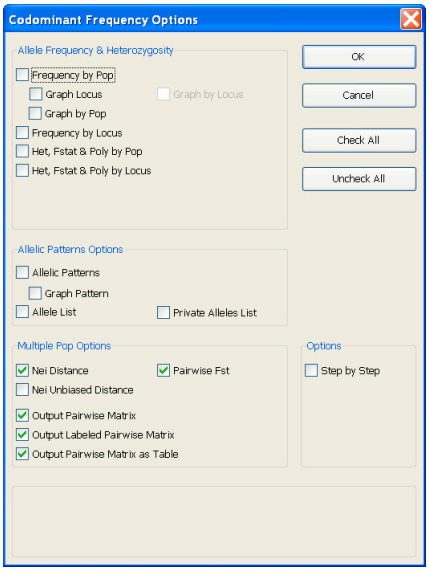
Nei's genetic identity ranges from 0 to 1. Consequently, Nei's Genetic Distance ranges from 0 to infinity (Nei 1972, 1978). Note an unbiased estimate of *Nei's I* and *Nei's D* is also available in GenAlEx. Hedrick (2000) suggests this correction may give spurious results when homozygosity is low and sample size is small. This unbiased estimator may also give slightly negative values for Nei's Unbiased Genetic distance, which should be interpreted as zero.

Pairwise Population Genetic Analysis

Ex 1.15 Pairwise Fst and Nei Genetic Distances

F_{ST} when reported as a single statistic over loci and populations provides an estimate of average differentiation. However, by exploring the patterns of differentiation among each pair of populations you can learn more about the genetic relationships than is evident from the average F_{ST} value on its own. Within GenAlEx both pairwise F_{ST} and Nei Genetic Distance can be readily computed for each pairwise combination of populations and summarized as a matrix. Both options are offered in GenAlEx via the ***Frequency*** menu.

Step 1. Return to the workbook *Ex 1.11 Glycine* and re-activate the *Data* worksheet. Choose ***Frequency*** from the **GenAlEx** menu, click *Uncheck All*, then choose *Nei Distance*, *Pairwise Fst*, *Output Pairwise Matrix*, *Output Labeled Pairwise Matrix* and *Output Pairwise Matrix as Table* from the Codominant Frequency Options dialog box.



Tip: You might find it easier to move a copy of the data sheet from Ex 1.11 into a new workbook called Ex. 1.15. Right-click on the worksheet tab at the left-hand corner of the workbook and use the Move or Copy *option to achieve this task.*

Step 2. Inspect the outcomes in the four worksheets with suffixes [*NeiP*], [*NeiL*], [*FstP*] and [*FstL*], in order to understand the nature of the output.

Step 3. Now based on the summary worksheets [*NeiT*] and [*FstT*], answer the questions below.

Q 1.15 Questions

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

1. Describe the genetic relationships among the populations indicated by the Nei Genetic distances. Which pairs of populations are genetically most similar?

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

2. Describe the genetic relationships among the populations indicated by the pairwise F_{ST} values. Which pairs of populations are genetically most similar?

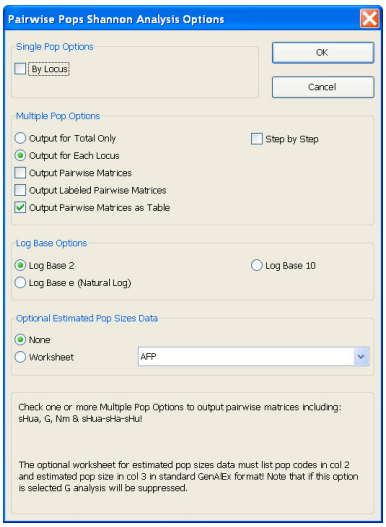
3. Compare the pairwise F_{ST} results with the pairwise *Nei D*. What do you conclude?

4. How would you test for a correlation between the pairwise F_{ST} results and the pairwise *Nei D*?

Ex 1.16 Pairwise calculation of Shannon's Indices

It will be self-evident that Shannon's diversity indices can also be computed at each locus for each pairwise combination of populations. Here we use the data from Ex 1.11 (and Ex 1.15) for Shannon analysis with a focus on the mean values over loci.

Step 1. Return to the workbook *Ex 1.11 Glycine* and re-activate the *Data* worksheet. Choose ***Shannon->Pairwise Pops***. When prompted by the Shannon Analysis Options dialog box choose *Output for Each Locus* and *Output Pairwise Matrices as Table*.



Tip: You might find it easier to move a copy of the data sheet from Ex 1.11 into a new workbook called Ex. 1.16. Right-click on the worksheet tab at the left-hand bottom of the workbook and use the Move or Copy option to achieve this task.

Q 1.16 Questions

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

1. Describe the genetic relationships among the populations indicated by the Shannon Mutual Information Index ${}^{S}H_{UA}$. Which pairs of populations are genetically similar?

2. Complete the table below drawing on the outcomes of analysis from Ex 1.15 and Ex 1.16. Do the different pairwise population statistics reveal similar genetic patterns?

Summary of pairwise population values of *F_{ST}*, *Nei Distance* and Shannon's *Mutual Information index* among two *Glycine* populations.

Pop1	Pop2	*F_{ST}*	Nei *D*	Nei *I*	*sHua*
Aranda	**Taylor**	0.061	0.208	0.813	0.337
Aranda	**Brind**				
Taylor	**Brind**				
Aranda	**Franklin**				
Taylor	**Franklin**				
Brind	**Franklin**				

Notes:

Principal Coordinate Analysis (PCoA)

Even a matrix as small as the 4x4 Nei Genetic distance matrix from our *Glycine* analysis (shown below) can be a little difficult to read and interpret. Larger matrices become impossible to interpret. Ideally, what we need is a way of visualizing the patterns of genetic relationship contained in such a matrix. Principal Coordinate Analysis (PCoA) provides such a tool.

PCoA is a multivariate technique that allows one to find and plot the major patterns within a multivariate data set (e.g. multiple loci and multiple samples). The mathematics is complex, but in essence PCoA is a process by which the major axes of variation are located within a multidimensional data set. Each successive axis explains proportionately less of the total variation, such that when there are distinct groups, the first 2 or 3 axes will typically reveal most of the separation among groups.

The Pairwise Nei Genetic Distance Matrix Among the 4 Glycine Populations

Aranda	**Taylor**	**Brind**	**Franklin**	
0.000				**Aranda**
0.208	0.000			**Taylor**
1.012	1.231	0.000		**Brind**
1.633	2.032	0.217	0.000	**Franklin**

Ex 1.17 Steps for Performing PCoA

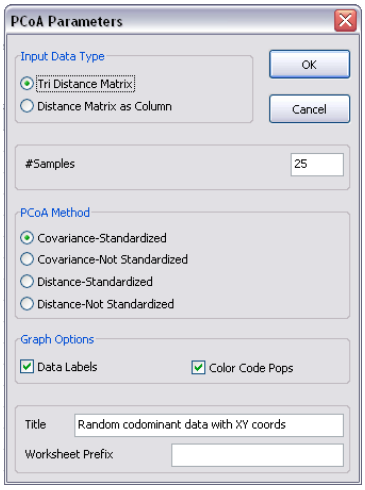
In this course we will leave the complex mathematics behind PCoA to GenAlEx (although we will return to PCoA in a later module). For now all that is needed is an appropriately formatted distance matrix. Here we will use PCoA to visualize the genetic relationships revealed by the Nei D, F_{ST} and ${}^{S}H_{UA}$ analysis of the 4 *Glycine* populations in Ex 1.15 and Ex 1.16.

Step 1. Return to the workbook *Ex 1.11 Glycine,* (or the renamed *Ex. 1. 15* & *Ex. 1.16*) and move a copy of the [*NeiP*], [*FstP*] and [*SHuaP*] worksheets to a new workbook. Name the workbook Ex 1.17 PCoA.

Step 2. Activate the worksheet [*NeiP*]. Note that for PCoA a distance matrix in GenAlEx format is required with the matrix diagonal starting in the cell A4, and parameters as shown.

	A	B	C	D	E	F	G
1	1	4					
2	Glycine cla	Data	Pairwise Population Matrix of Nei Genetic Distance				
3	Aranda	Taylor	Brind	Franklin			
4	0.000				Aranda		
5	0.208	0.000			Taylor		
6	1.012	1.231	0.000		Brind		
7	1.633	2.032	0.217	0.000	Franklin		

Step 3. Choose ***PCoA->Analysis*** from the **GenAlEx** menu, accept the default options of *Tri Distance Matrix*, *Covariance-Standardized* and *Data Labels* in the PCoA Options dialog box.



Step 4. Inspect the outcomes of the PCoA analysis and answer question 1 below.

Step 5. Repeat step 3 in order to produce a PCoA plot from the worksheets [*FstP*] and [*ShuaP*], then answer questions 2 and 3.

Q 1.17 Questions

1. Summarize the outcomes of your PCoA analysis of Nei Genetic Distance in words. How well does this PCoA plot represent the original data? (Hint: check the percentage of variation explained by the first 2 axes)

2. Compare the outcomes of the three PCoA analyses. Do they reveal a similar pattern?

3. Do your PCoA plots suggest regional genetic structure in *Glycine*? How would you test for this pattern?

Hardy-Weinberg Equilibrium

For codominant genotypes at a single locus, and for a single population, we can determine whether the observed tallies of genotypes are consistent with the expectations under random mating by performing a Chi-Square Test of Hardy-Weinberg Equilibrium.

Before one conducts any 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 tests for Hardy-Weinberg Equilibrium:
H0=No departure from random mating expectations (F=0) i.e. the population is randomly mating
H1=Departure from random mating expectations i.e. the population is not randomly mating (F<>0).

Note that the HWE option in GenAlEx is provided primarily for teaching purposes and for data exploration. Other programs such as GenePop and Arlequin provide Exact Tests which are recommended for research purposes (but note there remain some technical issues when employing Exact Tests. GenAlEx offers data export to these programs and other relevant programs. See the GenAlEx 6.5 Appendix 1 for more details.

Ex 1.18 Testing for Hardy-Weinberg Equilibrium

Small but real data sets for two plant examples with contrasting reproductive systems, *Glycine clandestina* and *Caladenia tentaculata,* are shown below. Complete the steps 1 to 8 for each species to determine whether or not they conform to Hardy-Weinberg Equilibrium then answer questions 1 to 3. For simplicity, steps 1 to 4 have already been completed for you.

Step 1. Determine the number of samples.

Step 2. Determine the number of alleles, *Na*.

Step 3. Count the numbers of each genotype.

Step 4. Calculate allele frequencies.

Step 5. Estimate the expected genotype frequencies, given the sample size of the population, either as p^2 for a homozygous genotypes or as $2pq$ for a heterozygous genotypes.

Step 6. Test for conformity with HWE expectations by calculating the Chi-squared statistic X^2.

Step 7. Determine the degrees of freedom as $DF = [Na(Na-1)]/2$

Step 8. Given the calculated Chi-squared value and the degrees of freedom, estimate the probability of the observed numbers deviating as far from the expected numbers by chance alone from the table below.

If the probability of obtaining the observed Chi-squared value (given the degrees of freedom) is greater than 0.05 (*P* in the range 0.05 to 1.0), the result is NOT statistically significant and we accept the null hypothesis *H0* = The population is mating randomly.

If the probability of obtaining the observed Chi-squared value (given the degrees of freedom) is less than 0.05 (in the range $0 < P < 0.05$), we conclude that the result is statistically significant, and we reject the null hypothesis *H0*, in favour of H1 = The population is NOT mating randomly.

Step 9. Record your answer in the tables below, and then answer the questions.

Step 10. Check your hand calculations using GenAlEx via ***Disequil->HWE***. The data are provided in *Ex 1.18 HWE Glycine.xls* and *Ex 1.18 HWE Caladenia.xls*.

Table showing critical values of X^2.

	Upper-tail Probability			
DF	**0.05**	**0.01**	**0.005**	**0.001**
1	3.841	6.635	7.879	10.828
2	5.991	9.210	10.597	13.816
3	7.815	11.345	12.838	16.266
4	9.488	13.277	14.860	18.467
5	11.070	15.086	16.750	20.515
6	12.592	16.812	18.548	22.458

You can use this table, given the degrees of freedom, to estimate the upper-tail probability for your calculated Chi-Square value. For example if DF=1 and your Chi-Square value is 5.5, the *P* value is less than 0.05, but greater than 0.01.

Tip: When using Excel you can easily calculate the Chi-Square probability using the function CHIDIST(x,deg_freedom), in this case enter '=CHIDIST(X^2,DF).

Allele Frequencies *Glycine clandestina*

Sample size = 30

Pop	**Allele**	**SATT373**
Mt Taylor	1	0.317
	2	0.683
Expected genotype frequency		
	1	**2**
1		
2		
Observed (=Observed genotype counts)		
	1	**2**
1	7	
2	5	18
Expected (=Expected genotype frequency * No. samples=30)		
	1	**2**
1		
2	12.990	
Observed - Expected		
	1	**2**
1		
2		
(Observed - Expected)^2		
	1	**2**
1		
2		
(Observed - Expected)^2/Expected		
	1	**2**
1		
2		
ChiSquare		
DF		
Prob		

Allele Frequencies *Caladenia tentaculata*

Sample size = 100

Pop	**Allele**	**MDH1**	
Pop1	**1**	0.220	
	2	0.705	
	3	0.075	
Expected genotype frequency			
	1	2	3
1			
2			
3			
Observed (=Observed genotype counts)			
	1	**2**	**3**
1	5		
2	34	48	
3	0	11	2
Expected (=Expected genotype frequency * No. samples=100)			
	1	**2**	**3**
1			
2			
3			
Observed - Expected			
	1	**2**	**3**
1			
2			
3			
(Observed - Expected)^2			
	1	**2**	**3**
1			
2			
3			
(Observed - Expected)^2/Expected			
	1	**2**	**3**
1			
2			
3			
ChiSquare			
DF			
Prob			

Q 1.18 Questions

1. Summarise your findings for *Glycine clandestina*. Which hypothesis, *H0* or *H1* is supported?

2. Summarise your findings for *Caladenia tentaculata*. Which hypothesis, *H0* or *H1* is supported?

Box 1.8 Chi-square for Hardy-Weinberg Equilibrium (HWE)

$$X^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

Where the summation from i to k genotypes is based on O_i the observed number of individuals of the i-th genotype, and E_i the expected number for the i-th genotype. E_i is calculated as either p_i^2 for a homozygous genotype or $2pq$ for a heterozygous genotype.

Degrees of freedom for the Chi-Squared test can be calculated one of two ways:

DF = (No. of genotype classes)-Na

or

$DF = [Na(Na\text{-}1)]/2$, where Na is the number of alleles at the locus.

The second formula is more convenient when there are a large number of alleles as is frequently the case with genetic markers such as microsatellites or STRs.

Putting It All Together

Ex 1.19 Revision: F-statistics in *Glycine and Caladenia*

Allele frequencies and observed heterozygosity are shown for two populations of *Glycine clandestina* and two populations of *Caladenia tentaculata* in the tables below. We will use this exercise to revise many of the formula we have learnt so far. Only minimal instructions are provided, if in doubt please refer to earlier exercises.

Step 1. Inspect the allele frequencies for both the *Glycine* and Caladenia data and answer questions 1 and 2.

Step 2. Calculate F-statistics for *Glycine* and *Caladenia* showing full hand calculations below. Record your answers in the table below, then answer the remaining questions.

Step 3. Check your hand calculations using GenAlEx. The data are provided in the workbooks *Ex 1.19 Glycine Fstats.xls and Ex 1.19 Caladenia Fstats.xls.*

Q 1.19 Questions

1. Based on your inspection of the allele frequencies in *Glycine*, how much genetic differentiation do you predict? Explain why.

2. Based on your inspection of the allele frequencies in Caladenia, how much genetic differentiation do you predict? Explain why.

3. Based on your calculations for *Glycine*, what do you conclude about the extent of genetic differentiation between the two populations Aranda and Taylor?

4. What do you conclude about the extent of inbreeding within the two populations of *Glycine*?

5. Based on your calculations for *Caladenia*, what do you conclude about the extent of genetic differentiation between the two populations W1 and W2?

6. What do you conclude about the extent of inbreeding within the two populations of *Caladenia*?

7. Are the biological conclusions you draw from the F-statistics analysis of *Glycine* and *Caladenia* the same as for the HWE tests? Explain your answer.

Glycine clandestina

Allele Frequencies			
Allele	**Aranda**	**Taylor**	**Total**
266	0.050	0.000	0.025
268	0.000	0.050	0.025
270	0.000	0.850	0.425
272	0.550	0.100	0.325
280	0.400	0.000	0.200
Heterozygosity and F statistics			
H_o	0.300	0.100	0.200
H_e			
F			
Mean F			
Mean H_o			
Mean H_e			
H_T			
F_{IS}			
F_{IT}			
F_{ST}			

Caladenia tentaculata

Allele Frequencies			
Allele	**W1**	**W2**	**Total**
1	0.021	0.146	0.083
2	0.896	0.792	0.844
3	0.083	0.063	0.073
Heterozygosity and F statistics			
H_o	0.208	0.333	0.271
H_e			
F			
Mean F			
Mean H_o			
Mean H_e			
H_T			
F_{IS}			
F_{IT}			
F_{ST}			

Ex 1.20 Bringing the Genetics and Ecology Together

Throughout these exercises we have been working with data from three different species, the bush rat, *Rattus fuscipes* and two plant species *Glycine clandestina and Caladenia tentaculata*. For the two plant species it is now time to reveal a little more about their biology. By combining ecology and genetics we can frequently discover new insights not evident from either ecology or genetic studies alone. In addition, genetic results can help us test predictions from our ecological knowledge, and vice versa. In the boxes below a brief summary of what we know about the biology of the two plant species is provided. Read these summaries before proceeding.

Box 1.9 The case of *Glycine clandestina*

Glycine clandestina is a native relative of the soybean. This species has an unusual reproductive biology - it produces two kinds of flower: Normal 'Open pollinated' flowers and 'Closed or cleistogamous' flowers. The open flowers are typical of pea flowers in general requiring insect pollinators for seed set. The 'Closed or cleistogamous' flowers are adapted to self-pollination, regularly producing seed without the aid of pollinators. The seeds of the species lack an obvious dispersal mechanism and it appears most seeds will fall close to the parent plant.

Box 1.10 The case of *Caladenia tentaculata*

Caladenia tentaculata, the green spider orchid, is exclusively pollinated by sexually attracted male thynnine wasps. The orchid, like many other Australian orchids, exploits the reproductive behavior of thynnine wasps by mimicking the sex pheromones of the female wasp. Pollination occurs when male wasps attempt copulation (pseudocopulation) with the labellum (the modified 3rd petal of orchids). After pollination, wasps immediately leave the patch, rather than visiting additional orchids. As a consequence of this behavior, pollen movements approximate a linear distribution, with a mean dispersal distance of 17 m (max = 58 m). This is among the largest mean pollen dispersal distances known for herbaceous plants (Peakall and Beattie 1996). The seeds of the species, like orchids in general, are minute and wind dispersed. However, we presently know little about the extent of seed dispersal in this and other orchids.

Step 1. Given what you now know about the biology of these two plant species, draw up a series of qualitative genetic predictions (summary in words) in the table below. Briefly justify these predictions below the table.

Step 2. Collate your answers from previous exercises in the summary table of statistical outcomes.

Step 3. To complement your statistical summary, calculate the outcrossing rate *t* as outlined in Box 1.11.

Step 4. Briefly summarize the key findings below the summary table. Then answer the questions that follow.

Box 1.11 Estimation of Outcrossing Rates in Plants

Typically the estimation of outcrossing rates in plants involves a genetic analysis of the genotypes of mother and offspring across multiple loci followed by a formal mating system analysis. In the absence of such an analysis, a simple transformation of the Fixation Index F can provide an estimate of the outcrossing rate t:

$$t = \frac{(1-F)}{(1+F)}$$

This transformation assumes no selection between fertilisation and the stage at which the samples were analysed for the estimate of F.

Q 1.20 Questions

1. Summarize in words your predictions in the table below, then justify your answer.

Statistic	***Glycine clandestina***	***Caladenia tentaculata***
HWE		
F		
F_{IS}		
F_{ST}		
t		

Justification:

2. Summarise the statistical outcomes in the table below, then list your key findings.

Statistic	***Glycine clandestina***	***Caladenia tentaculata***
HWE		
F		
F_{IS}		
F_{ST}		
t		

Key Findings:

3. Based on your findings in *Glycine clandestina*, how important is the contribution of the 'Closed flowers' to reproductive success. Explain your answer using the statistics you have calculated to back up your case.

4. What do you conclude about the extent of seed dispersal in *Glycine clandestina*? Explain your answer.

5. What do you conclude about the outcrossing rate in *Caladenia tentaculata?* Given selfing is possible in this system (i.e. the plant is self-compatible) how can you explain the result? Use the statistics you have calculated to support your case.

6. What do you conclude about the extent of seed dispersal in *Caladenia tentaculata*? Explain your answer using the statistics you have calculated to support your case.

References and Further Reading

Note that for a more extensive literature on these topics, please see the 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.

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 Associates, 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.

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, 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 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. 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

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, Ebert D, Cunningham R and Lindenmayer DB 2006. Mark-recapture by genetic tagging reveals restricted movements by bush rats, *Rattus fuscipes*, in a fragmented landscape. *Journal of Zoology,* **268**, 207-216.

Rossetto M, Kooyman R, Sherwin W and Jones R (2008) Dipersal limitation, rather than bottlenecks or habitat specificity, can restrict the distribution of rare and endangered rainforest trees. *American Journal of Botany,* **95**, 321-329.

Sherwin WB, Jobot F, Rush R and Rossetto M (2006) Measurement of biological information with applications from genes to landscapes. *Molecular Ecology,* **15**, 2857-2869.

Shannon CE (1948) A mathematical theory of communication. *The Bell System Technical Journal,* **27**, 379-423, 623-656.

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

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.

Glossary – Some Important Definitions

Allele: One or more alternative forms of a given gene or non-coding region of DNA.

Codominant: Both alleles in a diploid organism are visualized by a genetic marker system such that homozygous and heterozygous genotypes are detected. At the phenotypic level, the gene products of both alleles are expressed.

Dominant: Only one allele in a diploid organism is visualized by a genetic marker system such that only two genotypes are detected, either band presence or band absence. At the phenotypic level the gene product of only one allele is detected.

DNA: Deoxyribonucleic acid (DNA). Ribbons of sugars and phosphates held together in two opposite strands by 4 different bases or nucleotides: Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). Sequences of nucleotides make up genes.

DNA sequence: The sequence of DNA bases at a given locus.

DNA profile: Bands or genetic fingerprint produced by a genetic marker.

Electrophoresis: Migration of particles under the influence of an electric field. In the context of genetics, electrophoresis separates protein and DNA molecules of different size in a gel matrix that is subject to an electric field.

Genetic Marker: Any genetic character that can be measured and quantified. Most often genetic markers are visualized using laboratory procedures that detect variation either directly at the DNA level or indirectly via the products of DNA transcription and translation such as for allozyme or morphological characters.

Genotype: The set of alleles within an organism. In a narrower sense the alleles observed at a particular locus or loci. cf. Phenotype.

Heterozygosity: The proportion of heterozygous individuals at a locus, or heterozygous loci in an individual. Approximates genetic variance.

Heterozygote: Two different alleles at a given locus.

Homozygote: Two identical alleles at a given locus.

Locus: A specific position on the homologous chromosomes. Includes any identifiable coding (genes) and non coding region of the chromosome (pl. Loci).

PCR: Polymerase chain reaction.

Phenotype: The characteristics or appearance of an organism influenced by both the environment and genotype of the organism. In a narrower sense the characteristics displayed by a particular locus or loci cf. Genotype.

Polymorphism: The presence of one or more alternative forms at a given locus or loci = genetic variation. All genetic variation reflects variation in the sequence of nucleotides. For example, at a given locus, genetic variation can be represented by: (1) variation in the bases e.g. CGTACG vs CGAAAG, (2) variation in DNA length due to an insertion of nucleotides e.g. CGTACG vs CGTATATATATATACG, or (3) variation in DNA length due to a deletion of nucleotides e.g. CGTACG vs CGCG (TA deleted).

Restriction enzyme: An enzyme that cuts DNA at short specific sequences. Each enzyme has a unique cutting site.

Glossary - Genetic markers

AFLPs (Amplified Fragment Length Polymorphisms): A method that reveals fragment length polymorphism by PCR. First, genomic DNA is cut with two different restriction enzymes to produce short DNA fragments. Next, adapters of known DNA sequence are ligated to the ends of the cut fragments. Subsequently, selective PCR of the genomic fragments is then achieved using primers that match the known adapter sequence plus additional 'selective' nucleotides. Electrophoresis of the fragments produces a multi-locus profile or DNA fingerprint with polymorphisms apparent as either band presence or absence. Fluorescent or radioactive methods are used to visualize the fragments.

Allozymes: Alternate forms of enzymes encoded by different alleles at the same locus. Allozymes are prepared by homogenising tissue to produce a solution of proteins that is electrophoresed through a gel. Specific enzyme products are then visualized by a specific reaction. Alleles with different charges have different mobilities.

PCR-based genetic markers: Genetic markers produced via the amplification of DNA by the polymerase chain reaction.

RAPD's (Random amplified polymorphic DNA): An arbitrary-primed PCR method that uses arbitrary primers, of known sequence, usually 10 base pairs long to serve as both forward and reverse primers. Typically the amplified DNA fragments are resolved by low resolution agarose electrophoresis and staining with ethidium bromide. A multi-locus profile or DNA fingerprint with polymorphisms apparent as either band presence or absence is produced.

RFLPs (Restriction fragment length polymorphisms): Polymorphisms at specific sites in the DNA sequence revealed by the following method: DNA is cut with restriction enzymes, electrophoresed, blotted to a membrane and probed with radioactive DNA. Depending on the probe, single-locus or multi-locus profiles will be produced.

SNPs (Single Nucleotide Polymorphisms: Single base changes at a specific position in the genome, in most cases with two alleles. SNPs represent the most common form of DNA variation in the genome, and the analysis of a set of linked nuclear SNPs (haplotypes) provide an essentially inexhaustible source of stable polymorphic markers. An array of new methods are rapidly being developed for the routine screening of SNPs.

STRs (Short Tandem Repeats) or SSR (Simple Sequence Repeats) or Microsatellites: Tandem repeats of very short nucleotide motifs (1-6 bases long) eg: $[(CA)_{17}]$ or $[(AAT)_{10}]$ obtained by STS-PCR amplification using specific primers. Typically high resolution electrophoresis is required. A single-locus codominant genetic marker is produced. The standard genetic marker in human forensics and widely used in population genetics.

STS-PCR (Sequence-tagged-site PCR): A PCR method that uses two different specific primers, complementary to opposite strands of conserved DNA, to amplify the intervening sequence. A single-locus codominant genetic marker is produced.

Appendix 1.1 - Shannon Diversity

By Associate Professor WB Sherwin
Co-coordinator, Genetics Plan
Evolution & Ecology Research Centre
Deputy Head of School, Biological Earth and Environmental Science
University of New South Wales, Sydney NSW 2052 Australia
Email: W.Sherwin@unsw.edu.au
WWW: http://www.bees.unsw.edu.au/school/staff/sherwin/sherwinwilliam.html

1.0 INTRODUCTION

Biological diversity is quantified for reasons ranging from primer design, to bioprospecting, and community ecology. Therefore there is interest in merging and comparing biodiversity databases over a range of levels: ecosytems, species, genes (*Science* 2000; Hubbell 2001; Etienne 2004; Vellend 05). However, there has been little attention to providing common measures for biodiversity at different levels, thus risking comparisons of "apples with oranges". We need to answer the question "Does the index measure what we want to measure?"

Shannon's index ${}^{S}H$ (Shannon 1948) is: ${}^{S}H = - \Sigma\, p_i \log_2 p_i$
where p_i is the proportion of the "i"th allele in the population (or the "i"th species in the community, the "i"th base (A,C,T or G) at a certain position in the DNA strand, etc). ${}^{S}H$ asesses how much information could be spelt out using the entities in a collection. IE, if a stretch of DNA has three A's, two G's and one T, how many different messages could be spelt out by arranging them in various orders? This is, after all, what genes do. Groups of bases which allow more possible messages are said to have higher diversity. The same approach can be used to assess diversity at the level of the allele, haplotype, species, etc. IE, by placing the available entities (alleles, or species, etc), in particular orders, how many different messages could one spell out?

Another way of thinking of it is that the exponential of ${}^{S}H$ tells us how many EQUALLY FREQUENT alleles we would need in order to be able to spell out the same number and complexity of messages as in the actual array of alleles with unequal representation (Sherwin et al 06). This is similar to the "effective number of alleles", which tells us the number of equally frequent alleles (or bases, species, etc) would be needed in order to give the same heterozygosity as in the actual array of alleles with unequal representation.

${}^{S}H$ is the most commonly used index of diversity for ecological communities (Buddle 2004), and has seen a variety of uses in genetics, including describing human allelic diversity (Sherwin et al 2006). Does the frequent usage in ecology mean that ${}^{S}H$ is actually the best one to use over all levels of biodiversity? It has been suggested that the exponential of ${}^{S}H$ is the overall best diversity measure, without undue emphasis on rare or common alleles (or species, if operating at the community level) (Jost 2006, 2007). Also, when summing diversity over hierarchical levels, the exponential of ${}^{S}H$ shows more intuitive behaviour than many other measures of diversity (Jost 2006, 2007). This is not true for many other measures of diversity.

On the other hand, the main criticism of ${}^{S}H$ in the past has been a shortfall of theoretical expectations – what value of ${}^{S}H$ is expected in a population with a given history of size, dispersal,

etc? Such predictions are commonplace for statistics based on He and Fst (eg Halliburton, 2004), so much attention has been given to these statistics (eg Hubbell, 2001, Jost 2008). However, lack of predictive theory for ${}^{S}H$ is no longer a barrier: there are recent developments in theory of ${}^{S}H$ and related measures for both ecological communities (Dewar and Porte 2008) and one or more populations of genes under the Infinite Alleles Model of Mutation (IAM), or the Stepwise Mutation Model (SMM) (Sherwin et al 06). Like the predictive theory for He and Fst, the theories for ${}^{S}H$ are mostly neutral, so they must be taken as null hypotheses.

2.0 SINGLE POPULATIONS

In a single population, Shannon diversity is calculated as: ${}^{S}H_{AI} = -\Sigma p_i \log_2 p_i$
where p_i is the proportion of the "i"th allele in the population (Sherwin et al 06).

This can be compared with expectations based on various different assumptions about the population's history of population size, and the mutation mechanism of the loci analysed. For example, in a single population with mutation following the IAM, ${}^{S}H$ is expected to be

$$E^{S}H = \theta\sum_{j=1}^{\infty}\left\{\left[1-(2N_e)^{-(j+\theta)}\right]/\left[j(j+\theta)\right]\right\} \quad \text{(Sherwin et al 06)}$$

where $\theta = 4N_e u$, and u is the mutation rate, N_e is the effective population size ($N_e > 20$). For the stepwise mutation model, which is more appropriate for microsatellite DNA, the expected mutual information is:

$${}^{S}H \approx -\frac{\int_0^1 (Z\log p)dp}{\int_0^1 Zdp} \quad \text{where } Z = p^B(1-p)^{\theta-1} \text{ and } B = \frac{1+\theta-\sqrt{1+2\theta}}{-1+\sqrt{1+2\theta}} \quad \text{(Sherwin et al 06).}$$

Programs to calculate these expectations are available at:
http://www.bees.unsw.edu.au/school/staff/sherwin/sherwinwilliam.html

It is also possible to quantify the effect of assortative mating, such as inbreeding, on Shannon information, and relate this to statistics such as Fis (see supplement of Sherwin et al 2006).

3.0 MULTIPLE POPULATIONS – MUTUAL INFORMATION

As with other diversity measures, ${}^{S}H$ can be assessed within one locality, and between localities (Sherwin et al. 2006; Rossetto et al. 2008). Shannon's measure was very specifically set up to deal with a hierarchical situation such as multiple ecotypes within one locality, and multiple localities within one catchment, etc. Therefore, Shannon Information at each hierarchical level will always be independent of information at other levels; this removes many of the problems that occur using between-population measures based on *He*/*Fst* (Hedrick 2005, Jost 2008, 2009, Heller and Siegismund 2009; Ryman and Leimar 09).

The between-population measure for Shannon Information is called "Mutual Information", or ${}^{S}H_{UA}$. For alleles in two populations, mutual information can be seen as the ability to identify population of origin based on allele identity. If population number one had only I^A alleles, while population 2 had only I^B alleles, then knowledge of an individual's genotype would provide exact information on the individual's population membership – ie, mutual information is high between allele identity and population membership. But if each population had the two alleles at the same proportion, then

genotyping would give no information about population membership – mutual information is zero in this case.

3.1 MUTUAL INFORMATION FROM ALLELE PROPORTIONS

Calculation of mutual information from allele proportions is as follows (Sherwin et al 2006). For two localities (#1, #2) containing proportions *r* and *s* of the total individuals, within population #1 the allelic information is: ${}^{S}H_{A1} = - \Sigma\, p_i \log_2 p_i$ and similarly ${}^{S}H_{A2}$ for population #2.

Mutual information is then ${}^{S}H_{UA} = {}^{S}H_{U} - r\,{}^{S}H_{A1} - s\,{}^{S}H_{A2}$

where ${}^{S}H_{U} = -\sum \overline{p_i} \log_2 \overline{p_i}$ and $\overline{p_1} = rp_1(pop1) + sp_1(pop2)$

Computer simulation and real data sets show that mutual information is robust to a wide range of effective population sizes and dispersal rates. Relative to other measures, ${}^{S}H_{UA}$ has lower bias and/or variance over a range of sample sizes and conditions, making it a more robust estimator of dispersal (Sherwin 2006).

3.2 CONVERTING MUTUAL INFORMATION INTO DISPERSAL ESTIMATES

As with other between-population genetic diversity measures, mutual information can be converted to an estimate of dispersal rate between populations. Empirical results from simulations of microsatellite loci (Sherwin et al 2006) show that for haploids

$$ {}^{S}H_{UA} = \frac{0.22}{\sqrt{N_e m}} - \frac{0.69}{N_e \sqrt{m}} $$

Or for diploid organisms with uncorrelated alleles (eg no inbreeding)

$$ {}^{S}H_{UA} = \frac{0.22}{\sqrt{2N_e m}} - \frac{0.69}{2N_e \sqrt{m}} $$

Thus these equations can be used to convert the value of ${}^{S}H_{UA}$ estimated from allele proportions into an estimate of $N_e m$. For accurate estimates below N_e=500 for diploids (1000 for haploids), it is necessary to have at least a rough estimate of N_e. For large N_e, the second term in the equations is negligible, so that no estimate of N_e is needed.

3.3 MUTUAL INFORMATION AND STATISTICAL TESTING

Importantly, ${}^{S}H_{UA}$ also has an explicit relationship to standard statistical tests, so that ${}^{S}H_{UA}$ can be readily converted into a statistical test of partitioning of diversity between localities, habitats, etc. In fact the mutual information has a very straightforward interpretation. When multiplied by the total sample size and a constant, mutual information becomes the log-likelihood contingency test *G* statistic:

$$ G = 1.3863\ {}^{S}H_{UA}\ N_{tot} $$

where N_{tot} is the total number of alleles counted in all populations and ${}^{S}H_{UA}$ is in the log2 scale as above. To assess significance of the information, this G value can be compared with the chi-square table using appropriate degrees of freedom. For example, for a two-way table of populations by alleles such as Table 1, the degrees of freedom would be (#populations -1) x (#alleles -1) so DF = 2 in this case. Note that when calculating ${}^{S}H_{UA}$ for use as a statistical test in this way, the values of *r* and *s* should reflect the relative sizes of the samples, not the relative sizes of the actual populations. Thus in Table 1, *r* = 340/(500+340) = 0.405 and *s* = 500/(500+340)= 0.595.

The close relationship between mutual information ($^{S}H_{UA}$) and contingency tests facilitates incorporation into a statistical testing framework for various types of partitioning, such as habitat types, spatial structure, community structure and genetic structure (Smouse 1974; Sherwin et al 06)

Table 1 Partitioning biodiversity with Mutual Information via contingency testing and $^{S}H_{UA}$. Each cell entry is a count, *f* calculated from *N* the total number of individuals in the patch, and *p* the proportion of the particular allele in the population of that patch.

Locality	Allele			Column total
	I^A	I^B	I^O	
Loc #1	0	300	40	340
Loc #2	216	0	284	500
				N_{tot} = 840

4.0 INFORMATION AND APPROXIMATE BAYESIAN COMPUTATION (ABC)

It has been suggested that we should use several diversity indices simultaneously, and thus characterise the different aspects of diversity to which each index is most sensitive (Pielou 1966; Routledge 1979; Ricotta 2003). For example two communities that have identical Simpson's indices might have different Shannon indices, which would suggest that the communities have differences in their rare species composition. In this spirit, ABC methods use a variety of different measures in attempting to identify the most likely history of population size, dispersal, etc (Beaumont 2002). The Shannon indices hare been shown to perform well in ABC, and a user-friendly ABC program incorporating them is available (MS Bayes by Hickerson) http://msbayes.sourceforge.net/

GenAlEx 6.3 onwards provides options for calculating Shannon Diversity estimates among multiple populations and also provides the outcomes of locus-by-locus *G*-tests of mutual information. Further options associated with Shannon Diversity analysis are planned for future releases of GenAlEx.

REFERENCES

Beaumont M, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025.

Buddle CM, Beguin J, Bolduc E*, et al*. (2004) The importance and use of taxon sampling curves for comparative biodiversity research with forest arthropod assemblages. *Canadian Entomologist* **137**, 120-127.

Etienne RS, Olff H (2004) A novel genealogical approach to neutral biodiversity theory. *Ecology Letters* **7**, 170-175

Dewar RC, Porté A 2008. Statistical mechanics unifies different ecological patterns. Journal of Theoretical Biology 251:389–403

Halliburton R. 2004 Introduction to Population Genetics. Pearson Education, Upper Saddle river, NJ, USA

Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution* **59**, 1633–1638.

Heller R and Siegismund H 2009 Relationship between three measures of genetic differentiation Gst Dest and Gst': how wrong have we been? Molecular Ecology **18**:2080-2083.

Hubbell SP (2001) *The Unified Neutral Theory of Biodiversity and Biogeography* Princeton University Press, Princeton.

Jost L. 2006 Entropy and Diversity. *Oikos* **113**:363-375

Jost L. 2007 Partitioning Diversity into Independent Alpha and Beta Components. *Ecology* **88**:2427-2439

Jost L. 2008 Gst and its Relatives do not Measure Differentiation. *Molecular Ecology* **17**:4015-4026

Jost L. 2009. D vs Gst: Response to Heller and Sieismund (2009) and Ryman and Leimar (2009). Molecular Ecology **18**:2088-2091

Pielou EC (1966) The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology* **13**, 131-144.

Ricotta C (2003) Additive partition of parametric information and its associated β-diversity measure. **51**, 91-100.

Rossetto M, Kooyman R, Sherwin WB, Jones R. 2008. Dispersal limitations, rather than bottlenecks or habitat specificity, can restrict the distribution of rare and endemic rainforest trees. Amer. J Bot **95**: 321–329.

Routledge RD (1979) Diversity indices: which ones are admissible? *J. Theor. Biol.* **76**, 503-515.

Ryman N and Leimar O. 2009. Gst is still a useful measure of differentiation: a comment on Jost's D. *Molecular Ecology* **18**:2084-2087.

Science (2000) Special Issue Science **289**(5488).

Shannon CE (1948) A Mathematical Theory of Communication. *The Bell System Technical Journal* **27**, 379-423, 623-656.

Sherwin WB, Jabot F, Rush R and Rossetto M. 2006. Measurement of biological information with applications from genes to landscapes Molecular Ecology **15**:2857-2869.

Smouse PE. 1974. Likelihood analysis of recombinational disequilibrium in multiple-locus gametic frequencies. Genetics 76: 557-565

Vellend M (2005) Species diversity and genetic diversity: parallel processes and correlated patterns. *American Naturalist* **166**, 199-215.