About the Software GenAlEx 6.5

GenAlEx

Genetic Analysis in Excel 6.5

©2006 to 2012

Professor Rod Peakall
Evolution, Ecology and Genetics
Research School of Biology
The Australian National University, Canberra ACT 0200, Australia.

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

Peakall R. and Smouse P.E. (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research – an update. Bioinformatics 28, 2537-2539. Peakall R. and Smouse P.E. (2006) GenAlEx 6: genetic analysis in Excel. Population genetic software for teaching and research. Mol. Ecol. Notes 6, 288-295.

Australian National University

Proudly supported by The Australian National University
http://biology.anu.edu.au/GenAlEx/

Logo Design by GreenIdeasCreative.com

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 6000 registered users, representing over 60 countries, use the software. According to ISI, the first paper describing the software was cited more than 2500 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* 28, 2537-2539.

Freely available as an open access article from:
http://bioinformatics.oxfordjournals.org/content/28/19/2537

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.

The software and supporting documentation is freely available from The Australian National University, Canberra, Australia at the new URL: http://biology.anu.edu.au/GenAlEx

About this Quick Start Document

This document is provided to enable a quick start to GenAlEx 6.5 and needs to be read in conjunction with the Read Me file distributed with GenAlEx. This document is not intended as a substitute to the Tutorials offered with GenAlEx, or to replace the comprehensive guide to GenAlEx 6.5 and other supporting documentation.

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.

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 its associated GenAlEx menu. Your downloaded 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.

GenAlEx 6.5 can be run in Excel 2003 onwards. GenAlEx 6.5 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 2010, or to use the same version of GenAlEx in both Excel 2003, or Excel 2010 (but not simultaneously, unless using different copies).*

Notes for PC users: We strongly recommend that you use either Excel 2003, or Excel 2010. Excel 2010 fixed a series of major issues in Excel 2007 including issues that affect the operation of GenAlEx. Importantly, calculations can be extremely slow in Excel 2007 compared with Excel 2003/2010. It is also important to ensure that you have the most up to date release of Excel.

Notes for Macintosh Users On Macintosh computers, GenAlEx can be run in Excel 2004 and Excel 2011 (released early in 2010), 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. Note that Excel 2011 can only be run on Intel-based Macs. Be sure to read the readme file for Mac users.

For installation instructions please refer to the* Read Me *document distributed with the GenAlEx 6.5 Add-in.

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 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.

Sample Labels

To obtain maximum benefit out of GenAlEx it is ideal if each sample is given a unique numerical identifier. Sample names may carry an alpha character prefix, but this must be the same 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.

Data Parameters and Labels

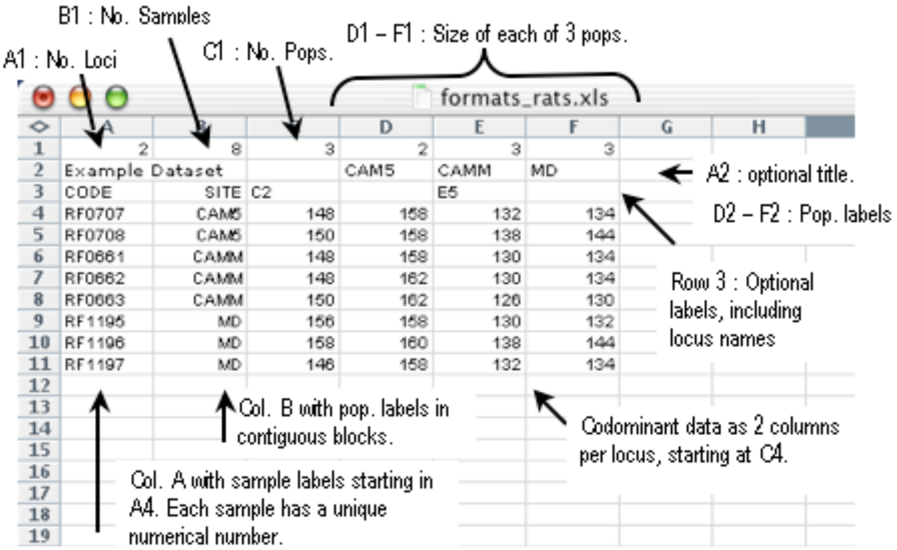
Data parameters and labels are crucial for telling GenAlEx how to read and analyse the data. GenAlEx stores all parameters and labels in rows 1, 2 and 3 of the data worksheets. Columns A and B are used for sample and population labels respectively. Actual data begins in Cell C4 of a worksheet.

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 (*s*ee section below for further instructions). This option requires that your data are 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.

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).



formats_rats.xls

	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								

B1 : No. Samples

A1 : No. Loci

C1 : No. Pops.

D1 – F1 : Size of each of 3 pops.

A2 : optional title.

D2 – F2 : Pop. labels

Row 3 : Optional labels, including locus names

Col. B with pop. labels in contiguous blocks.

Codominant data as 2 columns per locus, starting at C4.

Col. A with sample labels starting in A4. Each sample has a unique numerical number.

Data Formats

GenAlEx accepts 3 types of numerically-coded data:

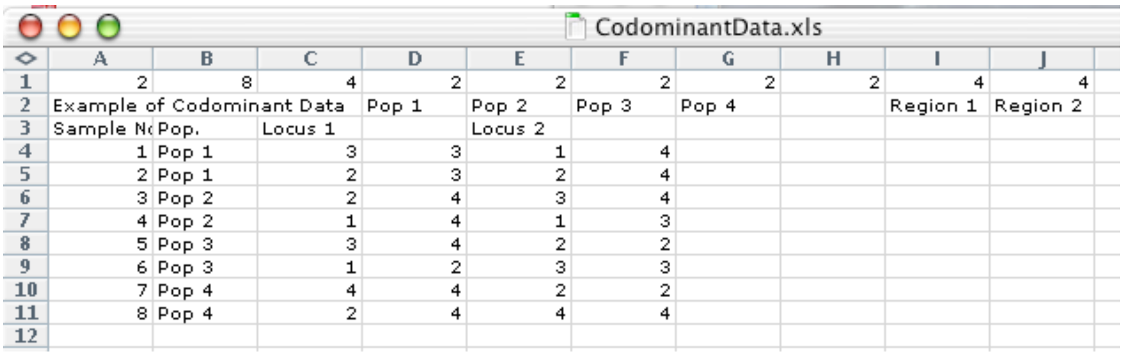
1. Codominant data with 2 columns per locus.
2. Dominant, Haploid (including Haplotypes), or Sequence data coded numerically with 1 column per locus/base.
3. Geographic data with 2 columns for X and Y coordinates.

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.

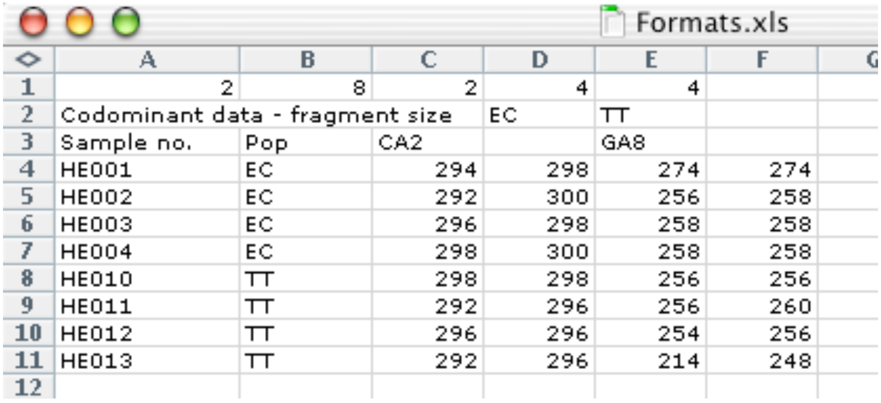


CodominantData.xls

	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.

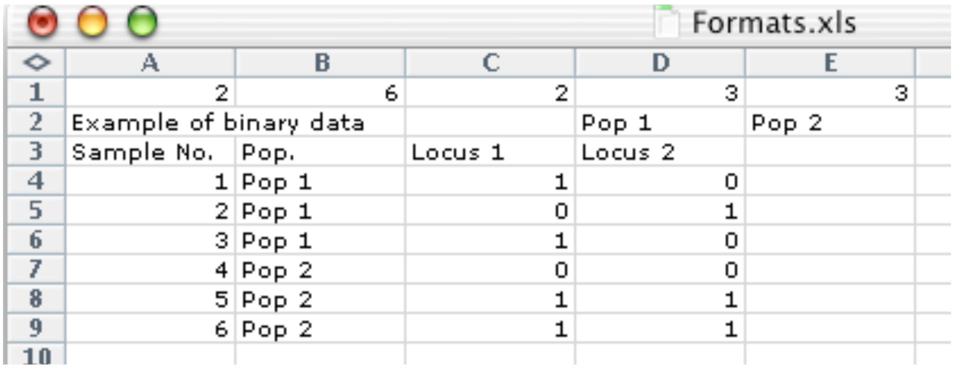


◇	A	B	C	D	E	F
1	2	8	2	4	4	
2	Codominant data - fragment size			EC	TT	
3	Sample no.	Pop	CA2		GA8	
4	HE001	EC	294	298	274	274
5	HE002	EC	292	300	256	258
6	HE003	EC	296	298	258	258
7	HE004	EC	298	300	258	258
8	HE010	TT	298	298	256	256
9	HE011	TT	292	296	256	260
10	HE012	TT	296	296	254	256
11	HE013	TT	292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

Format for dominant, haploid or sequence data

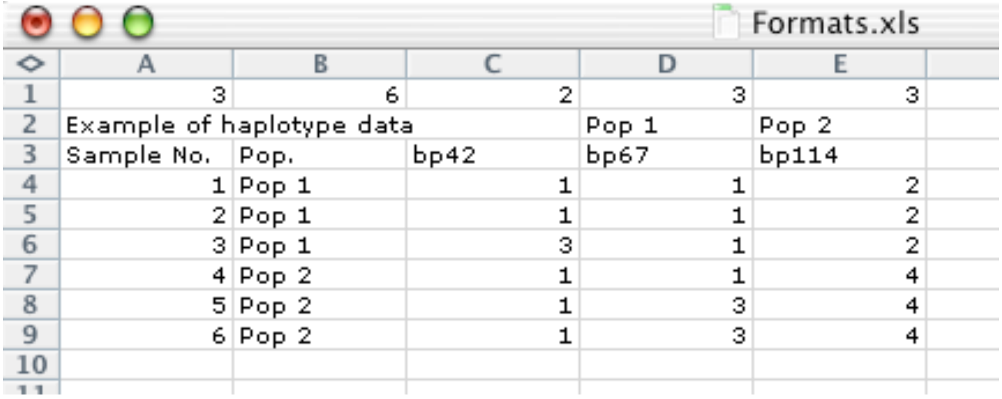
Dominant, haploid (including haplotypes) or sequence data are presented as a single column per locus. Haploid data can be coded numerically from 1…n, or each may be represented by multiple variable sites (columns 1 … n), with multiple states. For sequence or SNP data the bases are numerically coded as 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.



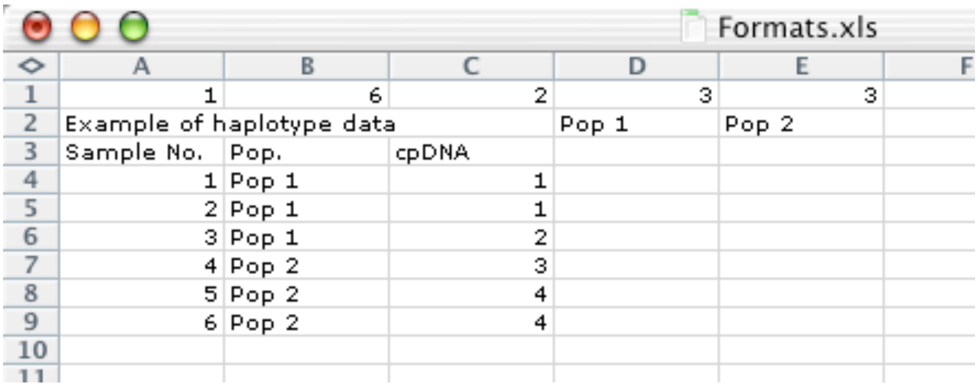
◇	A	B	C	D	E
1	2	6	2	3	3
2	Example of binary data			Pop 1	Pop 2
3	Sample No.	Pop.	Locus 1	Locus 2	
4	1	Pop 1	1	0	
5	2	Pop 1	0	1	
6	3	Pop 1	1	0	
7	4	Pop 2	0	0	
8	5	Pop 2	1	1	
9	6	Pop 2	1	1	
10					

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



◇	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.



◇	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						
11						

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

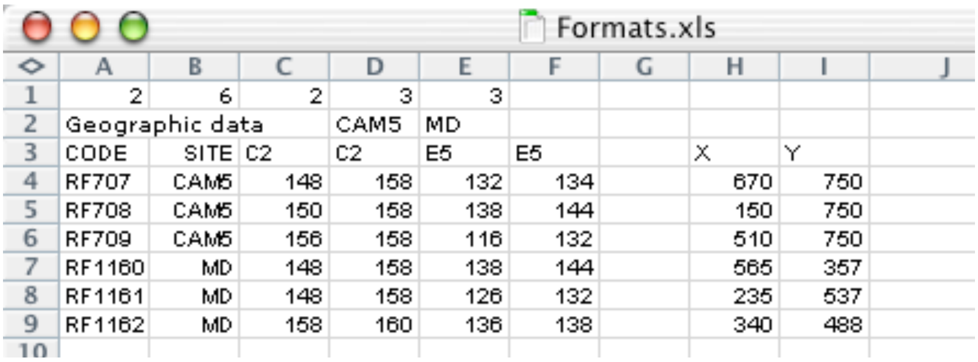
Format for geographic data

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, Clonal and *TwoGener*.

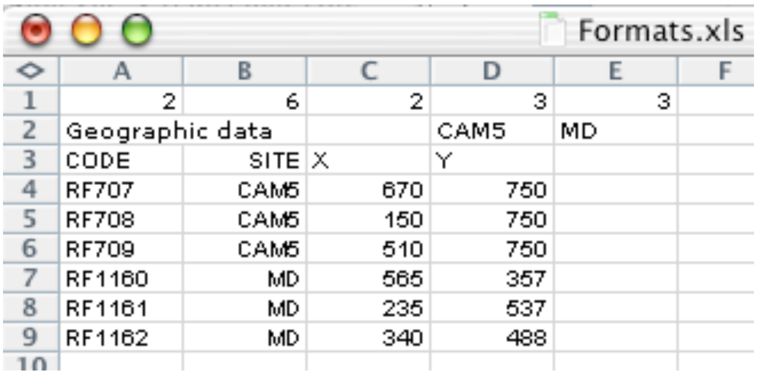
Example of geographic data after genetic data.



◇	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 C and D. In this case, the sample and population labels in columns A & B will correspond exactly to those for the genetic data. This format is also appropriate if only geographic distances are required. This format is required for analyses such as the 2D Spatial autocorrelation.

Example of geographic data in columns 3 & 4.



◇	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

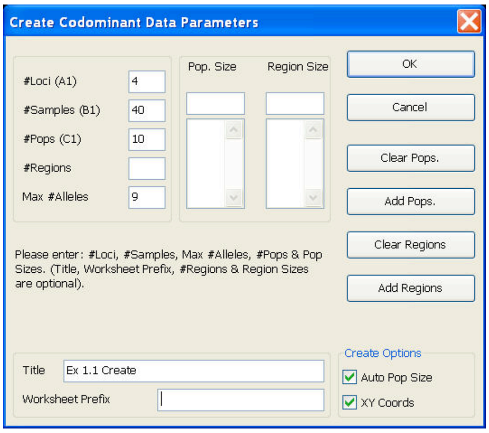
Virtually all GenAlEx options handle missing data. However, missing data can be particularly problematic for pairwise distance-based analyses such as AMOVA, Mantel and spatial autocorrelation. Therefore, a unique option for interpolating missing individual-by-individual pairwise distances is provided. 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’.

*Tip: It is important to note that missing data must be coded as either 0 (Codominant and Haploid) or -1 (Binary only). The presence of empty cells within your data, that is cells with no values, will prevent most GenAlEx analyses from running. You can use the **Data** menu option **Check Raw Data** to quickly locate any empty or non-numeric values in your data.*

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.



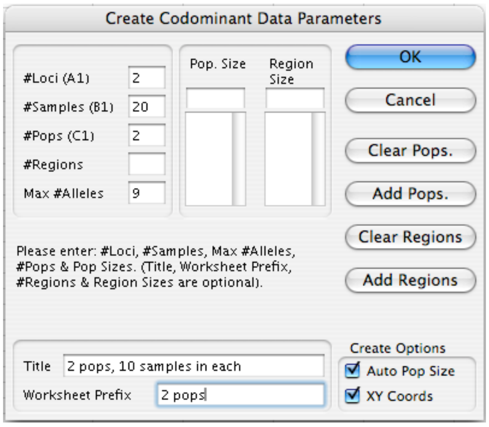
#Loci (A1) 4
#Samples (B1) 40
#Pops (C1) 10
#Regions
Max #Alleles 9

Please enter: #Loci, #Samples, Max #Alleles, #Pops & Pop Sizes. (Title, Worksheet Prefix, #Regions & Region Sizes are optional).

Title Ex 1.1 Create
Worksheet Prefix

☑ Auto Pop Size
☑ XY Coords

Create dialog box on a PC



#Loci (A1) 2
#Samples (B1) 20
#Pops (C1) 2
#Regions
Max #Alleles 9

Please enter: #Loci, #Samples, Max #Alleles, #Pops & Pop Sizes. (Title, Worksheet Prefix, #Regions & Region Sizes are optional).

Title 2 pops, 10 samples in each
Worksheet Prefix 2 pops

☑ Auto Pop Size
☑ XY Coords

Create dialog box on a Macintosh

Ex 1 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 2 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. 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.

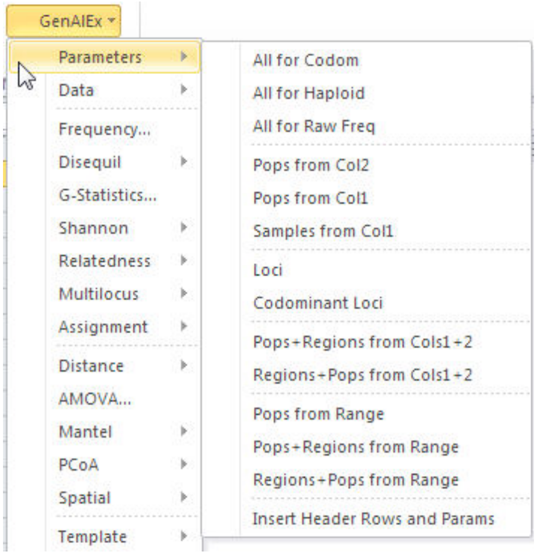
Step 6. Inspect the data sheet generated by GenAlEx. By reference to the numbers you jotted down, identify the location of the parameters in the data sheet and study the format for the genotypes and XY coordinates.

Ex 3 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, although it successfully passed the **Check Raw Data** *option confirming your data does not contain empty cells or non-numeric values. Often you will be given a warning by GenAlEx to check your data and parameters. However, 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.*

GenAlEx Data Parameters



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 location. Data must be in standard GenAlEx format, with samples in column 1, population labels in column 2 (or in col. 1), and data starting in cell C4. 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 chosen 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 sample codes in column 1, population labels in column 2, and 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.

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 the sample label (in Column 1).

Sort on Pop (Col2): Sorts the entire dataset on the population label (in Column 2).

Learning More About GenAlEx

This quick start guide is intended to merely get you started in GenAlEx. To further help GenAlEx users learn about population genetic analysis we have now made available a series of tutorial modules that provide step-by-step instructions to some of the more frequently used genetic analysis options offered in GenAlEx 6.5. In addition to these tutorials, there is a comprehensive guide to GenAlEx 6.5. The guide is supported by Appendix 1 that provides detailed background and references to the procedures used by GenAlEx.