MultiLocus 1.3

Paul-Michael Agapow and Austin Burt

Dept. of Biology, University College London, WC1 6BT
Dept of Biology, Imperial College, Silwood Park, Ascot, Berks, SL5 7PY

multilocus@agapow.net; a.burt@ic.ac.uk
http://www.agapow.net/software/multilocus/

2003-03-18

Documentation

Overview

Purpose. This program has been written to facilitate analysis of multi-locus population genetic data. In particular, it allows calculation of various genotypic diversity indices, various linkage disequilibrium indices, and a measure of population differentiation, and allows one to search for subpopulations which do not share polymorphisms (and thus might be reproductively isolated). In addition, there are randomization routines which allow one to test various null hypotheses.

Platforms. Versions are available to run on Macintosh (68K and PowerPC) and PC computers.

Reading in a datafile

Overview

On opening the program, a title box will appear, followed by a small menu with two options:

```
          ************************************************************
          *                    MultiLocus v1.3                      *
          *                                                         *
          *          Paul-Michael Agapow & Austin Burt             *
          *      Dept. Biology, Imperial College at Silwood Park    *
          *         <mailto://p.agapow@ic.ac.uk, a.burt@ic.ac.uk>   *
          * <http://www.bio.ic.ac.uk/evolve/software/multilocus/>   *
          *                     March 2003                          *
          ************************************************************



Command Menu :
   o> Open new datafile
   q> Quit
What is your command?                                        :
```

If you enter o, then the program will ask for the name of the input file, and then try to read it. If all goes well, the program will load the data and print out what it believes the format to be: haploid or diploid, number of loci, and number of individuals or isolates. The main menu will then appear:

```
Command Menu :
   o> Open new datafile
   1> Define linkage groups
   2> Define population groups
   3> Set preferences for handling missing data
   g> Calc. genotypic diversity & linkage disequilibrium
```

```
   l> Plot genotypic diversity vs number of loci
   d> Population differentiation analysis
   r> Test for partitions (haploids only)
   p> Print dataset to screen
   q> Quit
What is your command?                                  :
```

The options here include opening a new datafile, printing the extant datafile to the screen (to check it has been loaded properly), and quiting, all of which are self-explanatory. The other options are described later.

Further details

Datafile format

The file must be an ASCI text file in the same folder as the program. The following two examples illustrate the format for haploid and diploid data:

1	8	a	a		1/1	8/8	a/a	a/a
1	9	a	a		1/0	9/?	a/a	a/A
0	9	b	a		0/1	9/9	b/a	a/A
1	8	b	a		1/1	8/8	b/b	a/a
0	?	c	A		0/0	?/?	c/c	A/a
1	9	c	A		1/1	9/9	c/c	A/A
0	8	d	?		0/1	8/8	d/d	?/?
0	9	d	A		0/0	9/9	d/c	A/A
Example haploid dataset					Example diploid dataset			

The data should be in a table with every row a different individual or isolate and every column a different locus. There must be a single tab between each column. There cannot be any labels identifying individuals or loci – only the data matrix should be present. There should not be any blank lines between rows of data. Alleles are coded as alphanumeric strings (letters & numbers of any length), and the input is case sensitive (i.e., distinguishes small from capital letters). Missing data must be encoded as '?'. Gaps can be encoded as '-' (hyphen), and are treated in all calculations as missing data ('?'); if you want to treat them as alternative alleles, then they must be recoded as a letter or number. For diploid data the two alleles at a locus are to be separated by a '/', and the order does not matter (i.e., a/b is treated identically with b/a). If linkage groups are to be defined (see below), then loci in the same linkage group must be contiguous in the datafile; similarly, if populations are to be defined, then individuals from the same population must be contiguous.

In reading the datafile the program makes some tests and will give error messages if the file cannot be found or opened. **The first time a file is opened it should be printed to the screen and checked to confirm it has been loaded properly.**

Defining linkage groups

Choosing this option allows one to define groups of loci which are randomized together (see below for more on randomizations). The most common situation for invoking this option will be when the data consist of sequences for multiple regions of the genome (e.g., Koufopanou et al. (1997) sequenced fragments of 5 genes in each individual). Defining linkage groups has no effect on any of the statistics calculated for a particular dataset, but does affect the randomization procedure, and so the p-values associated with those statistics (details below). Linkage groups are defined by answering the following questions:

1. The number of linkage groups
2. The number of loci in group #1
3. The number of loci in group #2
   (etc)

The number in the last linkage group is calculated automatically, by subtraction.

As noted above, if linkage groups are to be defined, then loci in the same linkage group must be contiguous in the datafile. If loci with missing data have been excluded (see below), then the numbering used is that of the new, reduced dataset. This numbering is shown when the dataset is printed to the screen.

Defining populations

Choosing this option allows one to group individuals into populations. Defining populations has no effect on the genotypic diversity and linkage disequilibrium statistics, but does affect the randomizations, and so the p-values associated with those statistics (details below). Populations must be defined before testing for population differentiation. Populations are defined by answering the following questions:

1. The number of populations
2. The number of individuals in population #1
3. The number of individuals in population #2
   (etc)

The number in the last population is calculated automatically, by subtraction.

Again, if populations are to be defined, then individuals from the same population must be contiguous in the datafile. If individuals with missing data have been excluded (see below), then the numbering used is that of the new, reduced dataset.

Handling missing data

Choosing this option allows one to exclude individuals or loci with missing data (or include them, if they have previously been excluded). Note that changing

include/exclude parameters automatically sets linkage groups and populations back to the default state: every locus in its own linkage group and all individuals in the same population. Choosing this option also allows one to specify that missing data be fixed in position during randomizations for the genotypic diversity and linkage disequilibrium analyses, so that the real data is shuffled around the missing data. This ensures that any structure that exists in the missing data (e.g., particularly recalcitrant isolates) is also present in all randomizations. **If the dataset includes missing data, it will often be wise to perform the analyses with various options, to confirm the results.**

Genotypic diversity and linkage disequilibrium analyses

Overview

Choosing this option from the main menu will cause the program to calculate the following 7 statistics:

1. The number of different genotypes
2. The frequency of the most frequent genotype
3. The genotypic diversity.
4. The proportion of pairs of loci which are compatible
5. The index of association, I_A
6. A new measure of multilocus linkage disequilibrium, $\bar{r}_d$
7. Another measure of multilocus linkage disequilibrium, $\bar{r}_s$

The results are saved to a file called *filename*.stats (where *filename* is the name of the input datafile). As for all other output files, it is saved in the same folder as the application and input datafile. Before calculation can begin, the following information must first be given:

1. Whether you also want to calculate linkage disequilibrium values for all pairs of loci (saved to a file called *filename*.pairs).
2. Whether you also want to analyse randomized datasets, and if so, how many randomizations.
3. Whether you want to save the observed dataset and any randomized datasets to a file called *filename*.paup for execution in PAUP*.

Further details

This program calculates 3 measures of genotypic diversity and 4 measures of linkage disequilibrium (association between loci).

Number of different genotypes

This is obvious from the name. Only complete genotypes are included; individuals with missing data are automatically excluded in calculating this statistic.

The frequency of the most frequent genotype

This is obvious from the name. Again, individuals with missing data are automatically excluded in calculating this statistic.

Genotypic diversity

This is defined as the probability that two individuals taken at random have different genotypes. This is equivalent to

$$\frac{n}{n-1}\left(1 - \sum_i p_i^2\right)$$

where p_i is the frequency of the ith genotype and n is the number of individuals sampled. This value is 0 if every individual is the same, and 1 if every individual is different. In practice, the program takes all possible pairs of individuals and tests whether they are different, and then calculates the proportion of pairs that are different. Only pairs that are positively different are counted as different: missing data do not contribute to differences. Thus, an individual that has missing data may not count as different from two other isolates with more complete data, even though those 2 individuals are different from each other and resolving the missing data would make the original individual different from at least one of them.

Proportion of compatible pairs of loci

Two loci are compatible if it is possible to account for all the observed genotypes by mutations without having to infer homoplasy (reversals, parallelisms, or convergences), or recombination. Put another way, the two loci define a phylogeny with a consistency index of 1. For example, if there are two alleles at each of two loci, then there are 4 possible haploid genotypes, and the loci are compatible if no more than 3 of them are observed. Estabrook & Landrum (1975) outline an algorithm for multiple alleles, which is followed here. Briefly, list all the alleles that occur at locus 1 across the top of a rectangular matrix, and all the alleles that occur at locus 2 down the side, and then mark every box in the matrix for which that combination of alleles is observed. The two loci are incompatible iff it is possible to start at a marked box and then return to it by a series of horizontal and vertical moves (i.e., rook moves) to other marked boxes (though one is not allowed to go back to a box from which one has just come). Otherwise, the loci are compatible.

If the dataset consists of diploid genotypes, then they are decomposed into their constituent haploid genotypes before testing for incompatibility. This is unambiguous unless the diploid is a double heterozygote, in which case the individual is ignored in the test for that pair of loci. For diploids only, the proportion of compatible pairs of loci may therefore be an overestimate, though under the null hypothesis of no linkage disequilibrium this will be equally true of all randomizations, and so the p-value should be valid.

Loci are assumed to be compatible unless there is positive evidence to the contrary. Thus, if an individual has missing data ('?') at one or both of the loci under test, then that individual is ignored for that pair of loci.

This statistic can be used as a measure of multilocus linkage disequilibrium; note, though, that it will tend to vary with sample size, decreasing as more individuals are sampled.

The index of association (I_A)

The index of association (I_A) is the traditional measure of multilocus linkage disequilibrium (Brown et al. 1980; Maynard Smith et al. 1993; Haubold et al. 1998). In brief, the "distance" (number of loci at which they differ) between all pairs of individuals is calculated, and the variance of these distances compared to that expected if there is no linkage disequilibrium. In essence, one is asking whether two

individuals being the same at one locus makes them more likely to be the same at another. It is derived as follows. Consider first a single locus j that has been analysed in n haploid isolates. Let n_i be the number of isolates that have the ith allele. The probability that 2 isolates have a different allele at the locus will be

$$ h_j = 1 - \frac{\sum n_i (n_i - 1)}{n(n-1)} = \frac{n}{n-1}\left(1 - \sum p_i^2\right). \quad [1] $$

This is also the mean distance (either 0 or 1) between all $n(n-1)/2$ possible pairs of isolates, and the variance of those distances will be $var_j=h_j(1-h_j)$. Now suppose there are data for m loci. Let D be the distance between two isolates over all loci (i.e., the number of loci at which they differ). Then, the average distance over all pairwise comparisons will be $\overline{D} = \sum h_j$, and the variance of the distances will be

$$ V_D = \sum var_j + 2 \sum \sum cov_{j,k} \quad , \quad [2] $$

where $cov_{j,k}$ is the covariance between distance at locus j and distance at locus k and the double summation is over all $m(m-1)/2$ possible pairs of loci. If there are no associations between loci, then all these covariances are expected to be zero. Thus,

$$ I_A = \frac{V_D}{\sum var_j} - 1 \quad [3] $$

will equal zero if there is no linkage disequilibrium; this is the index of association.

In practice, the program calculates I_A using [3] where

$$ V_D = \frac{\sum D^2 - \frac{\left(\sum D\right)^2}{n_p}}{n_p} \quad [4] $$

and

$$ var_j = \frac{\sum d^2 - \frac{\left(\sum d\right)^2}{n_p}}{n_p} \quad [5] $$

In both [4] and [5] the summations are over all $n_p = n(n-1)/2$ possible pairs of individuals; D in [4] is the total distance between two isolates, over all loci, and d in [5] is the distance (0 or 1) at locus j. Thus missing data is handled in a consistent way: in calculating the distances between two genotypes, only positive differences count. That is, if there is missing data, then they do not contribute to the distance measure: a '?' is not different from either an 'A' or a 'B'.

Note that Maynard Smith et al. (1993) and previous programs by AB calculated $var_j=h_j(1-h_j)$ using [1], but without the correction factor of $n/(n-1)$; previous programs by AB also used n_p-1 in the denominator of [4] rather than n_p. The current implementation seems the most appropriate. These changes have only small effects on the I_A calculated and have no effect on the p-value (as old and new values will be monotonically related across randomizations of a dataset).

Diploids are handled in an analogous way to haploids, noting that the distance between two individuals at a particular locus may then take values of 0, 1, or 2.

$\bar{r}_d$

Substitution of [2] into [3] gives

$$I_A = \frac{2\sum\sum cov_{j,k}}{\sum var_j} \quad , \qquad [6]$$

which highlights a weakness of the statistic. Note that the number of terms in the numerator will increase as m^2, while the number in the denominator will increase only with m. Thus, if the covariances are not zero, I_A will increase with m (the number of loci). That is, if there is linkage disequilibrium, then the value will usually depend upon the number of loci included in the analysis, which makes comparisons among studies difficult (Brown et al. 1980; Maynard Smith et al. 1993). To avoid this problem, consider a slightly modified statistic which should largely remove this dependency on number of loci. An alternative standardization for the covariances is:

$$\bar{r}_d = \frac{\sum\sum cov_{j,k}}{\sum\sum \sqrt{var_j \cdot var_k}} \qquad [7]$$

$$= \frac{V_D - \sum var_j}{2\sum\sum \sqrt{var_j \cdot var_k}} \qquad [8]$$

Equation [7] has a form similar to a correlation coefficient (hence the symbol r, with the subscript d referring to distances), and will have a maximum value of 1. It is calculated using [8], [4], and [5]. Diploid data and missing data are handled as for I_A. In addition, I_A and $\bar{r}_d$ will be monotonically related across randomizations of any particular dataset, and so p-values for the two statistics will be identical.

$\bar{r}_s$

I_A and $\bar{r}_d$ are based on unordered alleles and there is no opportunity to put a sign to the linkage disequilibrium (significant negative values would usually be very difficult to interpret). Sometimes alleles may be ordered (e.g., wildtype vs mutant; common vs rare; increasing some phenotype vs decreasing it; etc), and one might then want to test whether alleles tending in the same direction are positively or negatively associated. Linkage disequilibrium would then have a sign. For this

analysis, the alleles must be integers (if there are any non-integer alleles in the dataset, this statistic is not calculated). In essence, one adds up the numerical values of all the alleles at all loci in an individual, and then calculates the variance of those sums among individuals, V_S. This can then be compared to the variance one would expect in the absence of linkage disequilibrium, which is the sum of the variances for each locus. In particular, $\bar{r}_s$ is defined as:

$$\bar{r}_s = \frac{\sum\sum cov_{j,k}}{\sum\sum\sqrt{var_j \cdot var_k}} \qquad [9]$$

$$= \frac{V_S - \sum var_j}{2\sum\sum\sqrt{var_j \cdot var_k}} \qquad [10]$$

The difference between this statistic and the previous one is that there, one is analysing a distance matrix and sums are over all pairs of individuals, whereas here one is analysing the allelic values directly and sums are over all individuals. Again, equation [9] has a form similar to a correlation coefficient (hence the symbol r, with the subscript s referring to the fact it is signed), and will have a maximum value of 1. Equation [10] is the one used for calculation.

The analysis is similar to a 2-way ANOVA on allelic values which tests for a significant between-individual effect. Because p-values are from randomizations, there is not the assumption of normality usually associated with ANOVA.

For diploid data, the value for a locus is the sum of the two allelic values. Missing data do not contribute to sums, and so are considered the same as 0. Therefore, one must be very cautious in using this statistic if there is missing data, as a significant result could simply mean that individuals differ in the amount of missing data they have (see below).

One interesting use for this statistic in diploids is to test whether heterozygosity is correlated across loci. This is done by coding the data as, say, 0 for homozygotes and 1 for heterozygotes, then treating the data as haploid, and calculating $\bar{r}_s$ for observed and randomized datasets. Correlations between heterozygosity at different loci would be expected if, for example, some individuals are more inbred than others.

Pairwise linkage disequilibrium

For this, $\bar{r}_d$ (defined above) is calculated for all possible pairs of loci.

PAUP execution file*

Burt et al. (1996) introduced a phylogenetic test for linkage disequilibrium in which the length of the most parsimonious tree fitted to the observed dataset is compared to that fitted to randomized datasets, with the expectation that deviations from panmixis would lead to significantly shorter tree lengths. To implement this test, the program can save the observed and randomized datasets to a file in a format ready

for execution in PAUP* (Swofford 1999). This file includes all the necessary commands for PAUP* to search for trees and save lengths to its own output file called *filename*.lengths.The first dataset in *filename*.paup is the observed dataset and subsequent ones are the randomized datasets. Having executed *filename*.paup in PAUP* and generated *filename*.lengths, one would then sort through the latter in a spreadsheet program to get a distribution for the randomized datasets.

Diploid data are recoded so there is one symbol per diploid genotype, and the appropriate step matrix (number of steps needed to change one genotype into another) is generated so that it can be used in the phylogenetic analysis. Because of this recoding, there is a limit of 52 distinct single locus diplotypes allowed in the entire dataset. If there is more than this, it is probably because in the input file, different symbols are used for different loci; recoding the data so all loci used the same symbols would solve this problem. A translation table is printed at the top of *filename*.paup to allow checking, but this is ignored by PAUP*.

This randomization test can be done directly in recent versions of PAUP*, but MultiLocus allows automated recoding of diploid genotypes and control over the randomization procedure (by defining linkage groups and/or populations – see below).

Randomizations

For these analyses a null hypothesis which is interesting and relatively easy to test is that of complete panmixia. To do this, one compares the observed dataset to datasets in which an infinite amount of sex and recombination has been imposed on the data by randomly shuffling the alleles amongst individuals, independently for each locus (e.g., Burt et al. 1996).

This randomization procedure can be modified in two ways. First, if linkage groups have been defined, then loci in the same group are shuffled together, and so associations within groups are maintained. If the observed dataset still has more linkage disequilibrium than randomized datasets, this must be due to associations between linkage groups (e.g., Koufopanou et al. 1997). Second, if populations have been defined, then alleles (or linkage groups) are only shuffled amongst individuals of the same population. This option can be useful because linkage disequilibrium can arise from combining samples from genetically distinct populations. With this option the linkage disequilibrium due to population differentiation is maintained in all the randomized datasets, so if the observed dataset still shows significant associations, they cannot be due solely to population differentiation (at least at the scale defined) (e.g., Koufopanou et al. 1997).

For diploid data, the two alleles at a locus are shuffled together in this program and so associations between alleles within a locus are maintained in the randomized datasets. The test is then purely for associations between loci, and is independent of tests for associations within loci (i.e., deviations from Hardy-Weinberg proportions).

If there are missing data or gaps in the dataset, then one can choose to fix them in place in all randomizations, and have the true data shuffled around them. This ensures that any structure there may be in the distribution of missing data also exists in all randomized datasets. For diploid data, if either allele is missing or a gap, then it is fixed in place with this option.

Statistics calculated for the randomized datasets are saved to *filename*.stats and, if appropriate, *filename*.pairs. Thus, for each of these output files, the first line corresponds to the observed data, and the subsequent lines correspond to the randomized datasets. The very last line gives the p-value for each statistic. These are all 1-tailed results, testing whether the genotypic diversity values are smaller than expected in a randomized population and the linkage disequilibrium values larger than expected, except $\bar{r}_s$, which is 2-tailed. The distribution of $\bar{r}_s$ will not be symmetrical around 0, so the program calculates the proportion of randomizations that are as or more extreme in the same direction as the observed value, and then doubles that.

Plot genotypic diversity vs number of loci

Overview

Choosing this option will cause the program to randomly sample from 1 to m-1 loci from the dataset and calculate the number of different genotypes and the genotypic diversity. This allows one to see whether scoring more loci is likely to increase the genotypic diversity, or whether one has reached a plateau. The results are outputted to *filename*.plot, and can be used to make a graph in any standard data analysis package.

Population differentiation analysis

Overview

Choosing this option will cause the program to calculate θ for a specified set of populations and save it to a file called *filename*.theta. This option can only be invoked after first defining the populations (above). On choosing this option, you are asked whether θ is to be calculated for randomized datasets, and if so, the number of randomizations.

Further details

θ

The program calculates Weir's (1996:180-182) formulation of Wright's F_{ST} for haploids, which for a single locus is:

$$\theta = \frac{Q_2 - Q_3}{1 - Q_3}$$

where Q_2 is the probability that two alleles from the same population are the same, and Q_3 is the probability that two alleles from different populations are the same. These are estimated as:

$$Q_2 = \frac{X - r}{r(\overline{n} - 1)}$$

$$Q_3 = \frac{1}{r(r-1)\overline{n}n_c}\left(Y - \frac{\overline{n}(n_c - 1)}{\overline{n} - 1}X\right) + \frac{\overline{n} - n_c}{n_c(\overline{n} - 1)}\left(1 - \frac{1}{r-1}X\right)$$

where

$$X = \sum_u \sum_{i=1}^{r} n_i \tilde{p}_{ui}^2$$

$$Y = \sum_u \left(\sum_{i=1}^{r} n_i \tilde{p}_{ui}\right)^2$$

$$n_c = \frac{1}{r-1}\left(\sum_{i=1}^{r} n_i - \frac{\sum_i n_i^2}{\sum_i n_i}\right)$$

r = the number of populations

n_i = the number of individuals sampled from population i

$\overline{n}$ = mean number of individuals sampled per population

and

$\tilde{p}_{ui}$ = the frequency of the uth allele in the ith population

For multiple loci, Q_2 and Q_3 are summed across loci before calculating θ:

$$\theta = \frac{\sum Q_2 - \sum Q_3}{\sum(1 - Q_3)}$$

For diploid data, the program calculates θ using the above equations, as if the data was haploid but for twice as many individuals. It is not clear to us if this is exactly equivalent to the diploid θ of Weir (1996). Nevertheless, it can be used as a measure of population differentiation, and because the randomization will be of diploid individuals across populations (see below), the p-value will be fully valid. If there is missing data, then sample sizes will differ among loci, but θ is still calculated as indicated above using (unweighted) sums of Q_2 and Q_3 across loci.

Randomizations

The null hypothesis of interest here is no population differentiation, and this is tested by comparing θ for the observed population to that for datasets in which individuals have been randomized across populations. This randomization procedure maintains any linkage disequilibrium and (for diploids) deviations from Hardy-Weinberg proportions that may exist in the observed dataset; independent randomization of individual loci or alleles across populations would not do this. If such disequilibria do exist, tests which did not take them into account could give spuriously low p-values. Note that θ is one of many statistics that could be compared between observed and randomized datasets to test for population differentiation (Goudet et al. 1996).

θ values for the randomized datasets are saved to *filename*.theta, and so the first line of this file is for the observed dataset and subsequent lines are for randomized datasets. The final line gives the p-value (1-tailed).

Test for partitions (haploids only)

Overview

Choosing this option allows one to test whether the collection of genotypes can be partitioned into two groups that do not share any polymorphisms (i.e., have no more than 1 allele in common per locus). Such groups may be good candidates for reproductively isolated species (e.g., Koufopanou et al. 1997) . If such a partition is found, then the following are saved to a file called *filename*.part: (i) the replicate number (0 for the observed data); (ii) the group sizes; and (iii) the individuals found in the smaller of the two groups (where individuals are numbered in the order they appear in the input datafile).

Further details

For a sample of n genotypes, the program tests all possible partitions into 2 and n-2 genotypes, and then all possible partitions into 3 and n-3 genotypes, and so on, up to all possible partitions of n/2 and n/2 genotypes (or (n-1)/2 and (n+1)/2, if n is odd). Each partition is tested to see if the two groups defined by it share any polymorphisms.

The analysis is done on a matrix of positively distinct genotypes: if an individual with missing data could be identical to any other, it is not included in the matrix.

Bibliography

Brown, A. H. D., M. W. Feldman, and E. Nevo. 1980. Multilocus structure of natural populations of *Hordeum spontaneum*. Genetics 96:523-536.

Burt, A., D. A. Carter, G. L. Koenig, T. J. White, and J. W. Taylor. 1996. Molecular markers reveal cryptic sex in the human pathogen *Coccidioides immitis*. Proc. Nat. Acad. Sci. USA 93:770-773.

Estabrook, G. F., and L. Landrum. 1975. A simple test for the possible simultaneous evolutionary divergence of two amino acid positions. Taxon 24:609-613.

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

Haubold, B., M. Travisano, P. B. Rainey, and R. R. Hudson. 1998. Detecting linkage disequilibrium in bacterial populations. Genetics 150:1341-1348.

Koufopanou, V., A. Burt, and J. W. Taylor. 1997. Concordance of gene genealogies reveals reproductive isolation in the pathogenic fungus *Coccidioides immitis*. Proc. Nat. Acad. Sci. USA 94:5478-5482.

Maynard Smith, J., N. H. Smith, M. O'Rourke, and B. G. Spratt. 1993. How clonal are bacteria? Proc. Nat. Acad. Sci. USA 90:4384-4388.

Swofford, D. L. 1999. PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sinauer, Sunderland.

Weir, B. S. 1996. Genetic data analysis II. Sinauer, Sunderland.