Proc. Assoc. Advmt. Anim. Breed. Genet. Vol 14

A GRAPHIC DISPLAY OF GENETIC DISTANCES BETWEEN POPULATIONS

B.P. Kinghorn[1] and R.K. Shepherd[2]

[1] University of New England, Armidale, NSW 2351
[2] Central Queensland University, Rockhampton, QLD 4702

SUMMARY
The simple method describes a spatial distance diagram in which each population is plotted in an n-dimensional object, with the distances between populations being maximally correlated to their genetic distances, supplied as input data. This method has been used in two dimensions as part of a computer-aided learning module for students of conservation genetics.
Keywords: Genetic distance, genetic variance, conservation genetics.

INTRODUCTION
Methods for construction of phylogenetic trees are well established. These trees can be used to view the likely pattern of evolutionary relationships among different genetic groups, such as species, breeds, lines and even individuals within a population. Information for construction of such trees is usually genotypic in nature, such a microsatellite markers, but can be phenotypic where clear linkage to genotypes is available, such a eye or coat color.

This paper proposes a simple method to generate a spatial distance diagram in which each population is plotted in an n-dimensional object, with the distances between populations being correlated to their genetic distances, supplied as input data. This method has been deployed in module DIVERSITY of the program GENUP (http://metz.une.edu.au/~bkinghor/genup.htm), used as a training aid for students.

MATERIALS AND METHODS
DIVERSITY was used to generate genetic distance data from simulated microsatellite genotypes on six populations (Figure 1). Genetic distances were calculated using a Euclidean distance method, which gave similar results to Nei's genetic distance (Chaiwong and Kinghorn 2001). This module is used to guide students towards sensible policies on populations to conserve, and numbers of animals within populations, with the aim of balancing genetic merit and maintaining genetic diversity. Genetic distance information shown at the bottom left of figure 1 is not easy to grasp across many populations, such that a good visual aid is warranted.

The method for generating the cluster diagram employs differential evolution, adapted from Price and Storn (1997), which is a form of evolutionary algorithm (EA). In this approach, a foundation group or "population" of possible solutions to the problem is generated at random. In the prevailing case, a solution is a genetic distance diagram. Each of these possible solutions is evaluated for its "fitness" – in this case its goodness of fit to the genetic distance table. The method then propagates a new generation of possible solutions, derived from the most fit solutions in the previous generation, and so on over

generations until the results converge. Progeny solutions can inherit attributes from parent solutions, and 'crossing over' and 'mutation' can play a role (Price and Storn 1997).

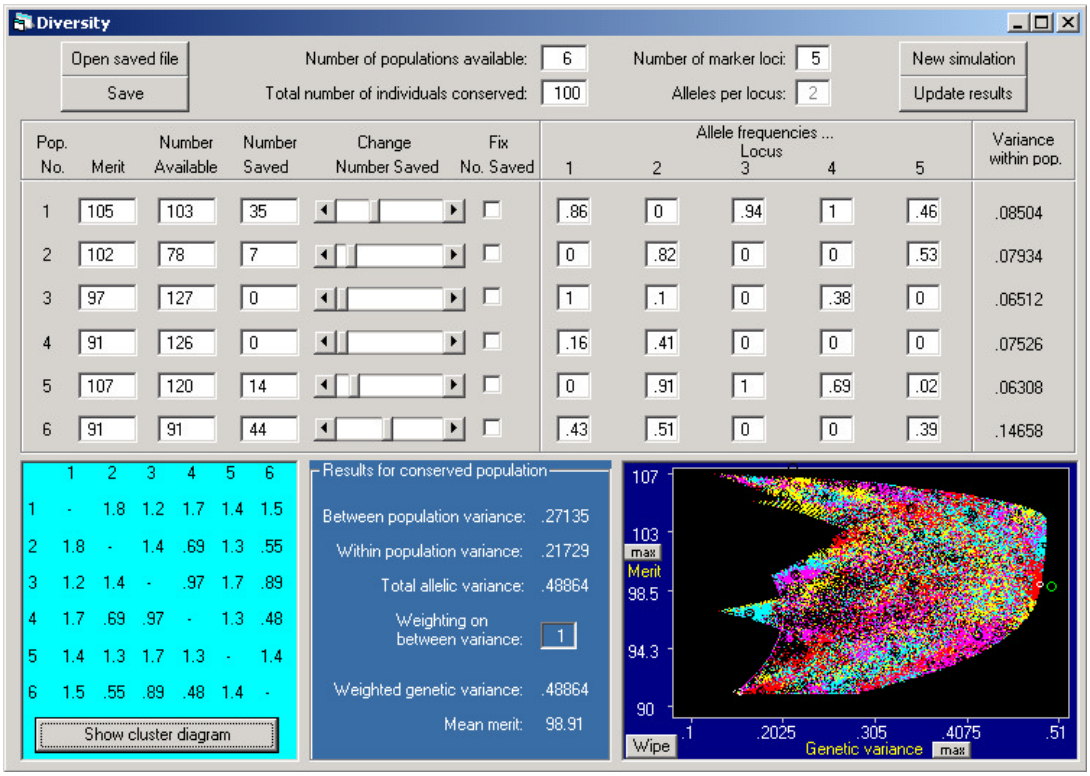

Figure 1. Screen capture from GENUP module DIVERSITY (Kinghorn 2000).

In the present case, a possible solution or "candidate solution" consists of a set of coordinates (one coordinate on a scale of 0 to 1000 for each dimension of the object to be plotted) for each population of animals. Such candidate solutions are initially created at random for the foundation generation. The distance between each set of coordinates is calculated using Pythagoras' Theorem. These coordinate distances are correlated to the genetic distance as found in the lower left part of Figure 1. The correlation is the fitness of the candidate solution under test – and this relates to the proportion of variation across genetic distances that is accounted for in the resulting diagram. Subsequent generations of solutions are created or 'bred' in a manner similar to that described by Price and Storn (1997).

This approach was used to generate a cluster diagram for the example in Figure 1. Clicking "Show cluster diagram" deploys a 2-dimensional cluster diagram, as in the middle of Figure 2.

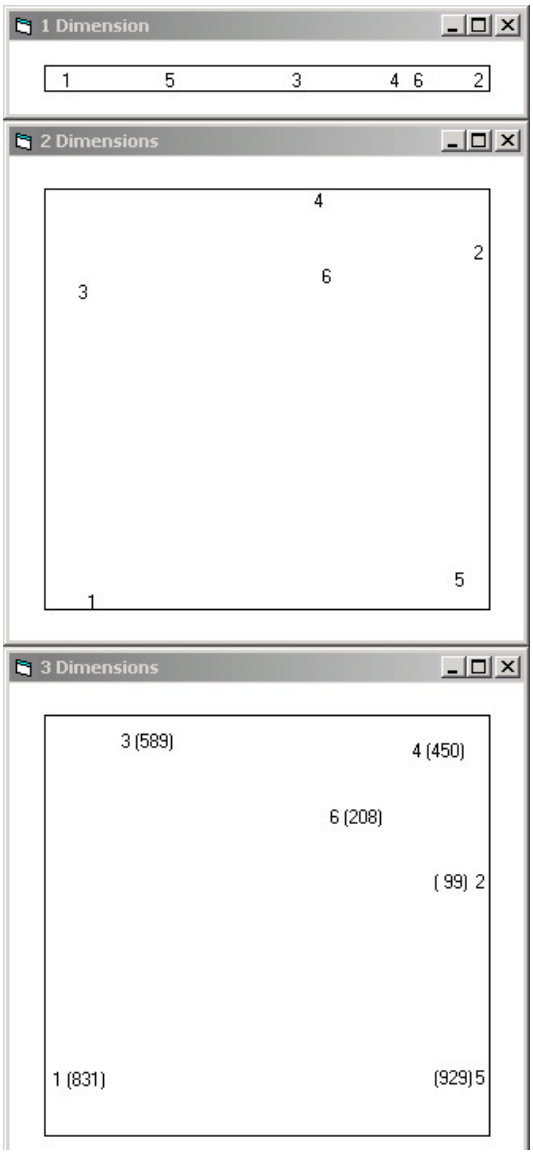

Figure 2. 1-, 2-, and 3-dimensional plots of nearest fit to the genetic distances for the six populations shown in Figure 1. In brackets are values of the third dimension (depth) on a scale of 0 to 1000.

RESULTS AND DISCUSSION

Figure 2 shows the nearest fit for plots in 1, 2 and 3-dimensions. The correlations between the distances given by these plots and the genetic distances in figure 1 are 0.90957, 0.99450 and 0.99992 respectively.

The test was repeated for examples involving 12 and 24 populations, and these results are included in Table 1. It is evident that a 2-dimensional plot provides a useful reflection of the calculated genetic distances, and that 3 dimensions gives an accurate reflection for the number of populations considered.

This method for the graphic display of genetic distances (GDGD) is related to multidimensional scaling (MDS), which is a technique designed to construct a 'map' showing the relationships between a number of objects (eg. towns), given only a table of distances between them. The map is usually in 2 or 3 dimensions as with GDGD. However GDGD differs from MDS in two key areas. Firstly, the algorithms used in MDS are usually hill-descent methods and so often get trapped in a local maximum, whereas the EA used in GDGD is better at searching the whole feasible space and finding the global maximum.

Secondly, the fitness function used in MDS, called a stress function (Schiffman *et al*. 1981), doesn't seem to have useful statistical properties except for its utility with hill-descent algorithms. As a consequence MDS has no framework for statistical inference.

The EA used in GDGD allows any appropriate fitness function to be used. The fitness function used in this application of GDGD was the correlation coefficient. Alternatively a chi-square goodness-of-fit statistic could be used as the fitness function and could be used to draw statistical inferences about the number of dimensions required assuming an appropriate statistical sampling model for the data. This is because for *p* populations there

are ½*p*(*p*–1) independent distances which can be displayed in a *d* dimensional space using the EA to find the values of the ½*d*(2*p*–1–*d*) independent coordinate parameters. Note that translations and rotations don't alter distances so only ½*d*(2*p*–1–*d*) of the *pd* population coordinates are required. Of course, the number of independent coordinate parameters equals the number of independent distances in a *p*–1 dimensional space, resulting in no lack of fit. However, in general, a 2- or 3-dimensional display of the genetic distances is sufficient.

Table 1. Correlations between predicted and actual genetic distances

# Pops.	Dimensions		
	1	2	3
6	0.90957	0.99450	0.99992
12	0.85081	0.97060	0.99362
24	0.84179	0.95280	0.98696

Use of the 2-dimensional plot in Figure 2 is quite straightforward. For example, if simple allelic variation among populations were of prime interest, then it can be seen that population 6 has little to contribute, as it is nearest to the center of the plot, and is quite close to populations 4 and 2. Indeed, if the "Weighting on between variance" in Figure 1 is set high, then population 6 does not feature in solutions with high variance, found by clicking near to the right side of the "Merit versus Genetic variance" graph (Figure 1).

REFERENCES

Chaiwong, N. and Kinghorn, B.P. (2001) *Conservation Genetics* (In press).
Kinghorn, B.P. (2000) http://metz.une.edu.au/~bkinghor/genup.htm
Price, K. and Storn, R. (1997) *Dr. Dobb's Journal*. **264**: 18.
Schiffman, S., Reynolds, M.L. and Young, F.W. (1981) Introduction to multidimensional scaling. New York, Academic Press.