Molecular Ecology Notes (2004) **4**, 137–138 doi: 10.1046/j.1471-8286.2003.00566.x

PROGRAM NOTE

DISTRUCT: a program for the graphical display of population structure

NOAH A. ROSENBERG
Program in Molecular and Computational Biology, 1042 W 36th Place, DRB 289, University of Southern California, Los Angeles, CA 90089–1113, USA

Abstract

In analysis of multilocus genotypes from structured populations, individual coefficients of membership in subpopulations are often estimated using programs such as STRUCTURE. DISTRUCT provides a general method for visualizing these estimated membership coefficients. Subpopulations are represented as colours, and individuals are depicted as bars partitioned into coloured segments that correspond to membership coefficients in the subgroups. DISTRUCT, available at www.cmb.usc.edu/~noahr/distruct.html, can also be used to display subpopulation assignment probabilities when individuals are assumed to have ancestry in only one group.

Keywords: admixture, ancestry, assignment test, clustering, subdivided population

Received 12 August 2003; revision accepted 23 October 2003

A genetically structured population can frequently be viewed as a set of discrete subgroups, in each of which alleles have distinctive frequencies. Individuals have membership in one or more of the subgroups, so that the membership coefficients of an individual sum to one across subgroups. The membership coefficient of an individual for a subgroup represents the fraction of its genome that has ancestry in the subgroup.

In analysis of data from structured populations, individual multilocus genotypes are often employed to estimate the membership coefficients of individuals in subgroups. This kind of analysis can proceed in two ways: supervised and unsupervised (e.g. Hastie *et al*. 2001). In the supervised approach, the subgroups are specified in advance, with subgroup allele frequencies regarded as known, or with some individuals regarded as having known membership coefficients. Membership coefficients are then estimated for individuals of unknown origin. In the unsupervised approach, subgroups are not specified in advance. Instead, estimation of membership coefficients proceeds simultaneously with estimation of allele frequencies and other properties of a series of abstract clusters. The estimation procedure constructs these clusters, and it is with respect to the set of clusters that individual membership coefficients are estimated.

Correspondence: Noah A. Rosenberg. E-mail: noahr@usc.edu

If K subgroups exist, and $K = 2$ or $K = 3$, several methods are available for graphical illustration of estimated membership coefficients (e.g. Pritchard *et al*. 2000a, b; Beaumont *et al*. 2001). However, most graphical strategies are not easily adapted to larger values of K. For any value of K, including $K > 3$, a convenient approach to representation of membership coefficients depicts each subgroup in a different colour, and each individual as a fixed-length line segment partitioned into K coloured components. These components correspond to membership coefficients of the individual in the various subgroups (Rosenberg *et al*. 2002). The program DISTRUCT converts estimated membership coefficients into this kind of figure (Fig. 1).

DISTRUCT is designed for use with STRUCTURE (Pritchard et al. 2000a; Falush *et al*. 2003), a program that can perform either supervised or unsupervised estimation of membership coefficients, so that STRUCTURE output (using models without linkage) is input for DISTRUCT. For a given individual, DISTRUCT requires only its K estimated membership coefficients and its predefined group identifier, such as a sampling location or phenotypic classification. Thus, in principle, membership coefficients obtained by any approach (e.g. Millar 1987) could potentially be the input for DISTRUCT. The program is run from a command line either in Unix or Windows, and it produces a PostScript output file (Adobe 1986, 1999). Using DISTRUCT, details such as the colour scheme and the order in which groups and individuals are printed can be controlled.

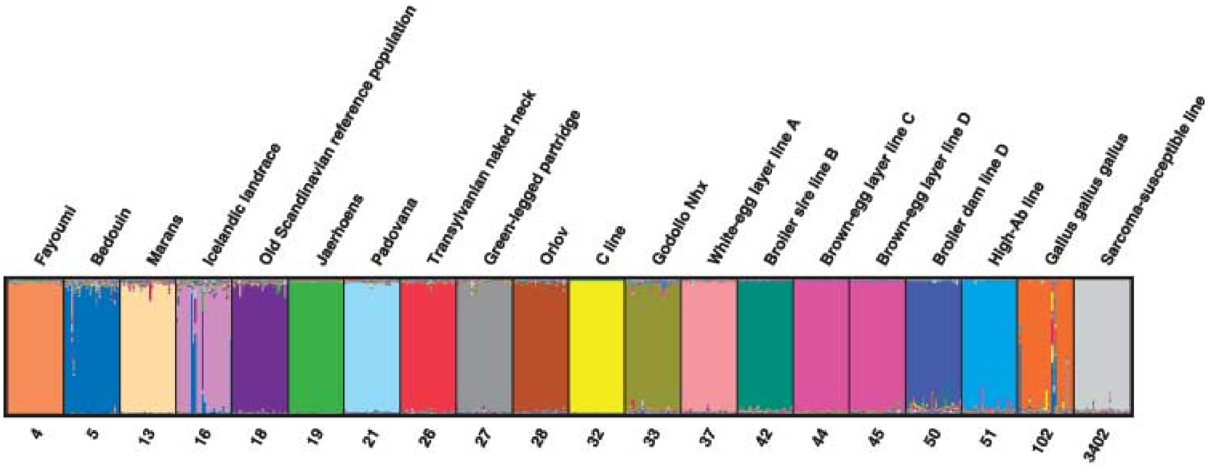

Fig. 1 DISTRUCT plot for 20 chicken breeds. The graph is based on the STRUCTURE run of highest estimated probability among those performed in the unsupervised population structure analysis in Table 2 of Rosenberg *et al.* (2001). Each individual is represented by a line partitioned into 19 segments corresponding to its membership coefficients in 19 inferred clusters. Each colour represents a different cluster, and black segments separate the individuals of different breeds. Population names are above the figure, and code numbers for populations are below it. The only two populations to have substantial sharing of a cluster are breeds 44 and 45. Clusters were permuted so that the bottom-to-top order of clusters would correspond to the left-to-right order of populations; thus, the bottom of the segment for an individual split equally among the 19 clusters would be the orange colour of breed 4 and the top of the segment would be the grey colour of breed 3402.

In addition to its use in displaying membership coefficients, DISTRUCT can depict probabilities in supervised or unsupervised assignment analyses. In these analyses, each individual of unknown origin is assumed to have ancestry in only one of the subgroups, and a probability of assignment is estimated for each unknown individual and each subgroup. Because assignment probabilities for an individual sum to one across subgroups, they can be represented graphically in the same way as membership coefficients for individuals who have ancestry in multiple subgroups. Similarly to the multiple membership case, DISTRUCT is applicable regardless of the method used to obtain assignment probabilities (e.g. Rannala & Mountain 1997; Banks & Eichert 2000; Pritchard *et al*. 2000a; Anderson & Thompson 2002). In theory, the type of graphic produced by DISTRUCT can also be used for allele frequencies or for other quantities that sum to one.

References

Adobe Systems Incorporated (1986) *Postscript Language Tutorial and Cookbook*. Addison-Wesley, Reading, MA.

Adobe Systems Incorporated (1999) *Postscript Language Reference*, 3rd edn. Addison-Wesley, Reading, MA.

Anderson EC, Thompson EA (2002) A model-based method for identifying species hybrids using multilocus genetic data. *Genetics*, **160**, 1217–1229.

Banks MA, Eichert W (2000) WHICHRUN (version 3.2): a computer program for population assignment of individuals based on multilocus genotype data. *Journal of Heredity*, **91**, 87–89.

Beaumont M, Barratt EM, Gottelli D *et al.* (2001) Genetic diversity and introgression in the Scottish wildcat. *Molecular Ecology*, **10**, 319–336.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Hastie T, Tibshirani R, Friedman J (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Millar RB (1987) Maximum likelihood estimation of mixed stock fishery composition. *Canadian Journal of Fisheries and Aquatic Sciences*, **44**, 583–590.

Pritchard JK, Stephens M, Donnelly P (2000a) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000b) Association mapping in structured populations. *American Journal of Human Genetics*, **67**, 170–181.

Rannala B, Mountain JL (1997) Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences*, **94**, 9197–9201.

Rosenberg NA, Burke T, Elo K *et al.* (2001) Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics*, **159**, 699–713.

Rosenberg NA, Pritchard JK, Weber JL *et al.* (2002) Genetic structure of human populations. *Science*, **298**, 2381–2385.