



Which genetic loci have greater population assignment power?

Michael A. Banks^{1,*}, Will Eichert² and Jeffrey B. Olsen³

¹Coastal Oregon Marine Experiment Station, Hatfield Marine Science Center, Oregon State University, 2030 SE Marine Science Drive, Newport, OR 97365-5229, USA,
²Bodega Marine Laboratory, University of California at Davis, Post Office Box 247, Bodega Bay, CA 94923, USA and ³US Fish and Wildlife Service, Alaska Region, Conservation Genetics Laboratory, 1011 East Tudor Road, Anchorage, AK 99503, USA

Received on September 24, 2002; revised on December 20, 2002; accepted on January 17, 2003

ABSTRACT

Summary: WHICHLOCI is a program that determines the relative discriminatory power of alternate genetic loci and loci combinations for population assignment of individuals.

Availability: <http://www.oregonstate.edu/dept/comes/genetics/software.htm>

Contact: Michael.Banks@oregonstate.edu

Increased information content from highly polymorphic molecular marker types such as microsatellites has markedly improved resolving power for discrimination among closely related populations. This, together with increased automation of techniques for resolving genetic variation, results in an overall boon of new information. Methods for assigning the population origin of individuals are among the new statistical techniques emerging to take advantage of this increased amount of information (Paetkau *et al.*, 1995; Waser and Strobeck, 1998; Banks and Eichert, 2000). Theoretical studies have considered how many loci and how many alleles might be necessary to maximize assignment success, the ratio of correct assignments over all decisions (i.e. both correct and incorrect assignments plus non-assignments (failures)). The general conclusion is that assignment accuracy is greatest when using a modest number of loci with a modest number of alleles (Smouse and Chevillon, 1998; Bernatchez and Duchesne, 2000). In addition, Cornuet *et al.* (1999) find a converse relationship between the number of loci (8–30) and the number of individuals (30–12) required for 100% correct assignment. These studies, however, simulated overly simplified populations and provide only general guidelines for maximizing assignment success. They do not provide a method for evaluating and selecting loci from empirical studies.

WHICHLOCI is a computer program that selects the best combination of loci for population assignment through empiric analysis of data drawn from natural populations. Successive assignment trials using data from one locus at a time allows ranking of loci in terms of their efficiency for correct population assignment and conversely their propensity to cause false assignments. Subsequent trials with increasing numbers of loci are then performed in order to determine which combination contains the minimum number of loci required to reach a specific level of assignment success (performance) set by the program user. The method is written for screening co-dominant as well as haploid marker types. Overall, WHICHLOCI is a program that determines the relative discriminatory power of alternate genetic loci and loci combinations for population assignment in order to assess which combination contains the minimum number of loci required to reach a specific level of assignment success.

The program requires data from study populations listed either as genotypes per sample in the same format used for GENEPOP (Raymond and Rousset, 1995) or as allele frequencies per population as created in WHICHRUN (Banks and Eichert, 2000). Test data are created using computer generated random numbers to sample from an allele table created for each population. This table consists of an array of alleles observed in each population, repeating each allele in accord with its frequency. The user defines how many genotypes to generate in this manner and has the option to vary sample size among populations. Sample sizes used should be greater than 500 to avoid bias from inadvertent genetic drift that would result from working with smaller sample sizes.

Minimum number high-ranking loci combinations that will match user-defined accuracy for population assignment are determined through two basic procedures. First repeated iterations for assignment of test data using the method applied in WHICHRUN (Banks and Eichert,

*To whom correspondence should be addressed.

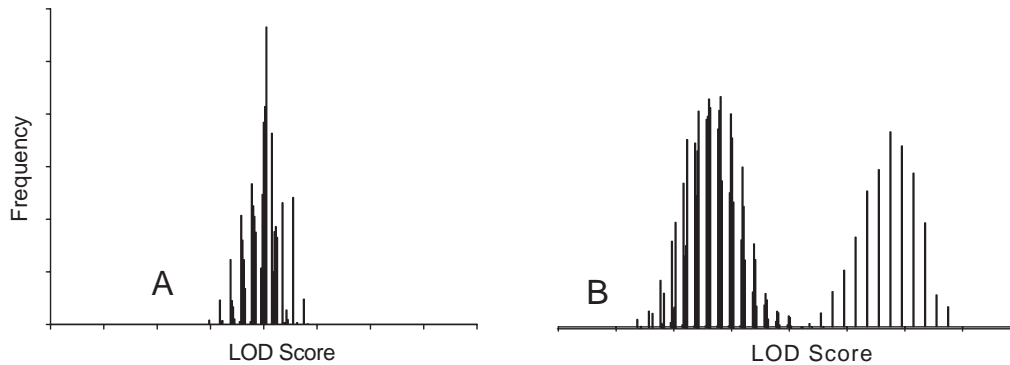


Fig. 1. Poor and good discriminatory power for identification of the endangered winter-run chinook salmon is demonstrated in (A) and (B) respectively. Results in (A) were attained using the five of 18 loci that rank lowest for discriminatory power. Here the range of LOD scores attained was common among all of four alternate chinook life history types considered. In (B), however, winter run is uniquely described using five loci that ranked highest. Here the range of LOD scores attained for winter (on extreme right) had no overlap with ranges observed for any other life history type. Data are from 17 microsatellites (Banks *et al.*, 2000), Greig and Banks, in review), five single nucleotide polymorphisms observed at an MHC class two exon (Kim *et al.*, 1999) and 5 haplotypes resolved from D-loop mtDNA sequence (Nielsen *et al.*, 1994).

2000), along with log of the odds stringency options described in WHICHRUN, are performed employing data from each locus separately, scoring the number of times genotypes might be correctly assigned to appropriate source populations for each locus. This score divided by the total possible number of correct assignments is then used to rank loci. A second round of iterations invokes loci from this rank increasing the number of loci one at a time until the assignment score matches or exceeds accuracy criteria set by the user. The above description covers procedure for accuracy considered across all populations. An alternate, critical population routine, allows focus on accuracy for assignment to a specific population set by the user. Iterations using data from each locus separately occurs as above but loci are scored only according to how many of the trial genotypes from the critical population are assigned correctly. Also the number of genotypes which might originate from other populations but are falsely assigned to the critical population are tallied. Rank order under the critical population routine is determined by applying the following formula:

$$\text{LocusScore} = \frac{\% \text{ correctly assigned}}{(\% \text{ incorrectly assigned} * \text{ScoreMultiplier})}$$

where % correctly assigned = % of members of the critical population correctly assigned, % incorrectly assigned = # of genotypes from other populations assigned to critical population/ total # of genotypes from other populations, and ScoreMultiplier = $(100 - \text{Accuracy})/\text{Inaccuracy}$. Accuracy and inaccuracy being criteria of intent for % correctly and % incorrectly assigned set by the user, respectively. One can thus weight correct assignment

or false assignment according to how important each criterion might be. An allele frequency differential method described in Shriver *et al.* (1997) can also be implemented for ranking loci. As above, a second round of iterations determines empirically how many of which loci are required to match accuracy criteria.

One area concerning individual based population assignment that has not received much attention is the issue of confidence (but see Almudevar, 2000). This parameter is obviously closely linked to the accuracy of allele frequency information for populations under consideration and is addressed through ensuring that sample sizes among baseline populations matches estimates required for polymorphic marker types (see Banks *et al.*, 2000). The issue of confidence estimation in the scenario of population assignment, however, becomes multidimensional given a comparison between alternate likelihoods that a genotype may come from each of the populations under study. The critical population method presented above provides a convenient means of summarizing these multidimensional likelihoods. WHICHLOCI provides a means for creating multiple trial data sets. Statistical parameters of variance, standard deviation and standard error are determined following typical formulae (Sokal and Rohlf, 1995) assessing the confidence that a specific combination of loci will indeed provide the assignment success estimated with the test data. It is also possible to bypass the loci ranking routine and just determine assignment success, variance, standard deviation and standard error for a user-selected bank of loci.

We thus present an empirical method for determining which combination contains the minimum number of

loci that would most likely provide defined population assignment power together with a means of determining statistical bounds on their performance. In our hands, the method provided important insight for resolving population origin among closely related chinook salmon from California's Central Valley. Two of four different life-history types present in this watershed have suffered precipitous declines in recent years. Interest in reducing loss of individuals from these threatened populations at water diversion sites, fishing and other perturbations has focused attention on resolving power for discrimination among these populations. We found that no generalization in terms of number of loci or locus characteristics (e.g. number of alleles) was a consistent predictor of performance. Rather one needs a resource such as WHICHLOCI to determine statistical power. While more polymorphic loci were typically better performers, on occasion loci with as few as four or five alleles would rank high for particular population assignments owing to unique distributions of genotypes among populations. The gain from using an empirically based method such as WHICHLOCI to determine the minimum number high-ranking loci combination is demonstrated in Figure 1.

We believe that this method will allow researchers to maximize power limits and minimize costs in focused population assignment contexts. This resource may also provide useful insight to loci linked or associated with particular fitness traits, disease resistance or other molecular traits that may have evolved to be unique among specific populations.

ACKNOWLEDGEMENTS

This research was supported by funds attained from CALFED and California's Department of Water Resources.

REFERENCES

- Almudevar, A. (2000) Exact confidence regions for species assignment based on DNA markers. *Can. J. Stat.-Rev. Can. Stat.*, **28**, 81–95.
- Banks, M.A. and Eichert, W. (2000) WHICHRUN (version 3.2): a computer program for population assignment of individuals based on multilocus genotype data. *J. Hered.*, **91**, 87–89.
- Banks, M.A., Rashbrook, V.K. et al. (2000) Analysis of microsatellite DNA resolves genetic structure and diversity of chinook salmon (*Oncorhynchus tshawytscha*) in California's Central Valley. *Can. J. Fish. Aquat. Sci.*, **57**, 915–927.
- Bernatchez, L. and Duchesne, P. (2000) Individual-based genotype analysis in studies of parentage and population assignment: how many loci, how many alleles? *Can. J. Fish. Aquat. Sci.*, **57**, 1–12.
- Cornuet, J.-M., Piry, S. et al. (1999) New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics*, **153**, 1989–2000.
- Kim, J.T., Parker, K.M. et al. (1999) Major histocompatibility complex differentiation in Sacramento river chinook salmon. *Genetics*, **151**, 1115–1122.
- Nielsen, J.L., Tupper, D. et al. (1994) Mitochondrial DNA polymorphism in unique runs of chinook salmon (*Oncorhynchus tshawytscha*) from the Sacramento–San Joaquin river basin. *Cons. Biol.*, **8**, 882–884.
- Paetkau, D., Calvert, W. et al. (1995) Microsatellite analysis of population structure in Canadian polar bears. *Mol. Ecol.*, **4**, 347–354.
- Raymond, M. and Rousset, F. (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J. Hered.*, **86**, 248–249.
- Shriver, M.D., Smith, M.W. et al. (1997) Ethnic-affiliation estimation by use of population-specific DNA markers. *Am. J. Hum. Genet.*, **60**, 957–964.
- Smouse, P.E. and Chevillon, C. (1998) Analytical aspects of population-specific DNA fingerprinting for individuals. *J. Hered.*, **89**, 143–150.
- Sokal, R.R. and Rohlf, F.J. (1995) *Biometry*. Freeman, San Francisco.
- Waser, P.M. and Strobeck, C. (1998) Genetic signatures of interpopulation dispersal. *TREE*, **13**, 43–44.