Reducing Bias in Mixture Estimates: a Computer Program to Bin Alleles

Jeffrey Bromaghin[1] and **Penelope Crane**[2]
[1]U.S. Fish and Wildlife Service, Div. of Fisheries and Habitat Conservation,
1011 E. Tudor Road, Anchorage, AK 99503, USA
[2]U.S. Fish and Wildlife Service, Conservation Genetics Laboratory,
1011 E. Tudor Road, Anchorage, AK 99503, USA

Keywords: DNA, microsatellites, data reduction, homogeneity, exact test, mixture analysis

Mixed-stock analysis (MSA) using genetic characters is an integral part of research programs estimating stock composition of catches of anadromous salmon and describing migration patterns of anadromous salmon in the high-seas. Widespread adoption of DNA techniques has lead to increased use of highly polymorphic loci in MSA. Greater polymorphism can enhance the power of MSA, but is not always beneficial. Researchers often bin alleles to reduce the effect of sampling error in baseline allele frequencies in studies using conditional maximum likelihood. Alleles are typically binned based on allele size or frequency, but these methods may result in a loss of information. We present a program for binning alleles to reduce the number of dimensions in a baseline while simultaneously maintaining the ability of the data to differentiate populations.

Exact tests of homogeneity can be used to test if alleles are similarly distributed across populations, with Monte Carlo simulation to estimate significance, to determine binning strategy. For any two alleles, the hypothesis of homogeneity is tested using either a likelihood ratio or Pearson test statistic. The P (number of populations) by A (number of alleles) matrix of allele frequencies is permuted such that the marginal allele and population frequencies of the entire P by A matrix remain fixed. For each permutation of the matrix, the test statistic is computed for all allele pairs and its value (Ψ_p) is compared to the test statistic from the full model (Ψ_o) providing an estimate of the probability of the distribution of the test statistic. The number of times Ψ_p exceeds Ψ_o, denoted k, is recorded for each pair of alleles. After the matrix has been permuted K times, the pair of alleles having the larges value of k is identified. The ratio p = k/K is an estimate of the significance of an exact test of the hypothesis that the allele proportions are equal across all populations, and large values of p indicate the allele proportions are not statistically different among the populations. If p exceeds a specified threshold p_{max}, the two alleles are binned to form a new allele, A is reduced by 1, and the process is repeated with the new P by A data matrix, otherwise the process terminates.

The program OptiBin uses baseline files (*.bse) for SPAM (Debevec et al. 2000) as input. The program options include choice of test statistic, threshold p value, number of permutations for Monte Carlo tests of significance, random seed, and whether to test all possible pairs of alleles or only alleles adjacent in size. The program outputs a *.bse file readable by SPAM and a log file of which alleles were binned and p-value for homogeneity test. The log file can be used by OptiBin to bin aleleles of mixture files for estimation. The program will be available at http://www.r7.fws.gov/fish/genelab/home.html.

The binning algorithm was tested on two data sets, allele frequencies for six microsatellite loci for five populations of Dolly Varden in western Alaska and allele frequencies for 11 microsatellite loci for ten populations of chum salmon from Yukon River. The binning program reduced the average number of alleles per locus by 50% and greatly reduced the number of sampling zeros. Bias was reduced in conditional maximum likelihood estimates of simulated mixtures. Further, a slight reduction in bias was also apparent when a Bayesian estimator of baseline allele frequency distributions was employed.

REFERENCES

Debevec, E.M., R.B. Gates, M. Masuda, J. Pella, J. Reynolds, and L. W. Seeb. 2000. SPAM (Version 3.2): Statistics program for analyzing mixtures. J. Hered. 91: 509–511.

All correspondence should be addressed to P. Crane.
e-mail: Penelope_crane@fws.gov