

WHICHRUN (version 3.2): A Computer Program for Population Assignment of Individuals Based on Multilocus Genotype Data

M. A. Banks and W. Eichert

Microsatellite DNA provides essentially limitless, highly varied information within species. That this provides a means for distinguishing not only among populations but also individuals has not escaped current theoretic interest (Smouse and Chevillon 1998; Waser and Strobeck 1998). Here we present a C++ computer program, WHICHRUN, that uses multilocus genotypic data to allocate individuals to their most likely source population. This program runs on Windows95, 98, or NT (including Macintosh emulations of these operating systems) and has no specific hardware requirements. WHICHRUN differs from a similar individual-based population assignment program "the assignment test" (Paetkau et al. 1995; Waser and Strobeck 1998) in that it provides a variety of methods for evaluating population assignments including maximum likelihood, jackknife, and critical population routines. WHICHRUN also provides resources for converting data into formats required for the population-based Statistical Package for Analysis of Mixtures (SPAM) available from L. Seeb, Alaska Department of Fish and Game.

Input File

WHICHRUN requires baseline genotype data for all potential source populations, as well as genotype data for candidate individuals for which population origin is to be determined. Data should be provided in simple ASCII format as required for GENPOPOP (Raymond and Rousset 1995). The download available at the site described below includes sample input files.

Theory and Program Outline

It is assumed that each baseline population ($B_1 \dots B_k$) has Hardy-Weinberg-Castle (HWC) genotype frequencies and that genetic loci employed are independent. The likelihood that an individual sample ($s_{1..n}$) may come from each of the source populations ($B_{1..k}$) is presumed to be equal to the HWC frequency of its specific genotype at each locus in each respective source population. Thus for homozygotes the likelihood that a sample (s_i) is an element (ϵ) of baseline population B_1 is p_1^2 [the square of its allele frequency (p_1) in population B_1]. For heterozygotes, $s_2 \epsilon B_1 = 2p_1q_1$ (q_1 being the frequency of an alternate allele in population B_1), and the likelihood that $s_i \epsilon B_k = p_k^2$ or $2p_kq_k$. Likelihood values for each locus are multiplied to give a series of multilocus likelihood functions for assignment to each of the source populations. Alternate hypotheses that individual samples in question may come from each source population are considered in three ways:

(1) Multilocus likelihood functions may be grouped to form ratios considering all possible pairs of baseline populations under consideration. If the ratio of the most likely allocation grouped with the second most likely allocation approaches one, there is ambiguity in the assignment of the particular sample under study. Conversely, samples for which this ratio yields a large result in comparison to all other ratios can be assigned to a single population with more confidence. For the two populations considered in the ratio, the chance of error is equal to the inverse of this ratio. Stringency for population allocation can be applied by defining a selection criterion for the \log_{10} of this ratio. For example, by selecting only assignments that have a log of odds (LOD) ratio of at least 2, all results will have a 1/100 chance of error or less.

(2) Multilocus likelihood functions may be grouped in a maximum likelihood format according to the equation $L(n)$

$L(\max)$. This yields a series of ratios between 1 (most likely) and close to 0 (least likely). Analysis of variance of log transformed data followed by a Tukey's multiple comparison enables evaluation of statistical significance in the classical sense.

(3) Jackknife iterations provide an empirical means for evaluating baseline data and the chances of correct allocation. Iterations sample individuals from the baseline one at a time, recalculating allele frequencies in the absence of each individual genotype sampled before determining the most likely population origin for that individual. Experimenting with alternate loci and populations enables one to determine which population comparisons and loci combinations enable reliable population reallocation.

Reporting Options and Special Cases

Sample ID, genotypic data, and multilocus likelihoods for population allocation can be displayed for verification. A critical population routine allows one to select a target population for calculation of LOD scores. All scores are then calculated with the critical population as the numerator in the ratio. A special case where test samples may have an allele or pair of alleles not observed in one or all of the baseline populations is treated as follows. For source populations in which the allele is not observed, an estimated allele frequency of $1/(2N + 1)$ is applied. This hypothesizes that the nonobservance of the allele in question is due to sampling error and that the allele in question would have been observed in the baseline population if one more allele had been sampled. Note that this estimation may introduce substantial bias if baseline population size (N) is small, as would be likely for any allele frequency estimation given small N , particularly when dealing with highly polymorphic marker types. The program implements a warning describing this consideration when small baseline population sizes ($N < 30$) are encountered. Alternatively, if sampling error is low, an unknown sample allele not observed in a baseline population may constitute strong evidence that the sample in question may indeed not originate from the particular baseline population under consideration. Any alleles for which the $1/(2N + 1)$ estimation is necessary are noted on the genotype output.

It is obvious that a technique such as

WHICHRUN will only be effective if there is reasonable reproductive isolation among populations under study. Three other considerations are also important. First, the rate of accumulation of variance for molecular loci employed should be closely matched with estimated divergence times among populations under study. For example, highly polymorphic microsatellites prone to homoplasy would not be suitable for diagnosis among populations that have diverged over substantial evolutionary time. However, highly polymorphic microsatellites are likely one of a few molecular marker types that have sufficient information to resolve diagnosis among recently diverged populations such as the global radiation of *Drosophila melanogaster*, which is estimated to have occurred within the last 10,000–15,000 years (Bénassi and Veuille 1995; David and Capy 1988). Second, the accuracy of determination is crucially dependent upon the lack of differential sampling error among baseline allele frequencies. While this problem is partially addressed through ensuring that sample size is equal for all populations, highly polymorphic marker types such as microsatellites require substantial sampling. Third, for population origin diagnoses where source populations are recently diverged, there will be a number of loci that have not accumulated differences in the time since divergence. As a result, simply increasing the number of loci employed may not necessarily increase the power of diagnosis. For closely related populations, additional loci that have marked differences in allele frequency profiles among populations will be necessary to achieve increased power.

WHICHRUN may be downloaded from <http://www-bml.ucdavis.edu/whichrun.htm>.

From the Bodega Marine Laboratory, University of California, Davis, P.O. Box 247, Bodega Bay, CA 94923-0247. We thank V. K. Rashbrook, H. A. Fitzgerald, and J. Olsen for a number of useful suggestions and improvements that resulted from beta testing various versions of this program. We are also grateful to F. J. Saminiego for discussions on statistical aspects during the development of WHICHRUN. Research and development of WHICHRUN was supported by funds from the California Department of Water Resources and the U.S. Fish and Wildlife Service. Address correspondence to Michael A. Banks at the address above or email: mabanks@ucdavis.edu.

© 2000 The American Genetic Association

References

- Bénassi V and Veuille M, 1995. Comparative population structuring of molecular and allozyme variation of *Drosophila melanogaster* Adh between Europe, West Africa and East Africa. *Genet Res* 65:95–103.

David JR and Capy P, 1988. Genetic variation of *Drosophila melanogaster* natural populations. Trends Genet 4:106-111.

Paetkau D, Calvert W, Stirling I, and Strobeck C, 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 86:248-250.

Smouse PE and Chevillon C, 1998. Analytical aspects of population-specific DNA fingerprinting for individuals. J Hered 89:143-150.

Waser PM and Strobeck C, 1998. Genetic signatures of interpopulation dispersal. Trends Ecol Evol 13:43-44.

Received March 31, 1999

Accepted August 18, 1999

Corresponding Editor: Robert Angus