

## PLABSIM: Software for Simulation of Marker-Assisted Backcrossing

M. Frisch, M. Bohn, and  
A. E. Melchinger

In plant breeding and population genetics, computer simulations are a useful tool to investigate problems for which no analytical solutions are available. We present PLABSIM, a computer program to simulate marker-assisted selection in arbitrarily designed backcross programs. The simulated data can be evaluated with PLABSIM for gene frequencies, genotype frequencies, frequency of homozygous loci, length of chromosome segments originating from one ancestor, and the number of marker data points required for a breeding program. In addition to data analysis with PLABSIM, the simulated data can be exported for analysis with statistical software.

For investigations of the potential applications of marker-assisted selection, computer simulations are a useful tool. They have been used to investigate problems concerning the use of flanking markers in selection for one or more target genes (Hospital and Charcosset 1997; Visscher et al. 1996) or to accelerate the recovery of the recurrent parent genome in marker-assisted backcross programs (Frisch et al. 1999; Hospital et al. 1992; Openshaw et al. 1994). Following Hospital and Charcosset (1997), we refer to the first approach as foreground selection and to the second as background selection.

Until now, no software allowed one to simulate marker-assisted selection under realistic genetic models. The program GREGOR (Tinker and Mather 1993) implements the basic principles, but the interactive use and the fact that it simulates only some predefined genetic linkage

maps restricts its feasibility for simulation of breeding programs.

In this article we present PLABSIM, a tool for simulation of marker-assisted selection programs. The software can be used to investigate the effect of varying population size, marker density, marker positions, and selection strategies on the genetic composition of the breeding product and on the required number of marker data points.

PLABSIM is characterized by the following features: (1) Simulations can be made for any diploid genome with an arbitrary number of loci at arbitrary positions on an arbitrary number of chromosomes. (2) The implemented reproduction schemes include all common breeding methods. (3) An arbitrary number of selection steps can be combined to a selection strategy. Selection can be carried out for genotypes at defined loci, or for selection indices calculated from allele frequencies at several loci. (4) The simulated data can be analyzed for a broad range of genetic parameters with PLABSIM. In addition, the data can be exported for analysis with statistical software.

### Methods

The key element in genetic simulation is the algorithm used for simulation of meiosis. PLABSIM simulates the production of gametes with a random-walk algorithm (Crosby 1973). Depending on the map distance  $d$  (in Morgans) between two adjacent loci, the recombination frequency  $r$  is calculated with Haldane's (1919) mapping function  $r = (1 - e^{-2d})/2$ . This assumes that neither chiasma nor chromatid interference occurs (Stam 1979). With a pseudo-random number generator, a realization of a uniformly distributed random variable  $0 \leq Z \leq 1$  is generated for each pair of adjacent loci. A crossover occurs between the two loci under consideration if and only if  $z \leq r$ . Applying this rule to

all adjacent loci of all linkage groups generates two gametes, one of which is chosen at random.

PLABSIM was written in C++ and runs in batch mode. It requires an input file (in ASCII format) that contains information about the genetic map, the initial population(s), the design of the breeding program, and the computations to be made. The output of PLABSIM is stored as an ASCII file.

### Simulation of Breeding Programs

#### Step 1: Defining the Linkage Map

A linkage map is the basis for a simulation run. It consists of a list of loci, identified by their names. Optionally, a map position may be given. It is either 0, if the locus is at the start of the chromosome, or it represents the distance of the locus from the start of the chromosome in centiMorgans (cM). If no linkage values are given, the loci are assumed to be unlinked. A number of loci may be grouped together by assigning a common class name to them, for example, marker or target. In addition to loci of which the genotype can be determined, the genetic background can be simulated by filling up the interval between two linked loci with background loci. Background loci have map distances of 1 cM. With the evaluation of background loci, it is possible to draw conclusions about the portion of the genotype that is not explicitly defined in the linkage map. They are useful to estimate the proportion of the recurrent parent genome of an individual or a population in backcrossing, or to estimate the length of chromosome segments originating from one ancestor.

#### Step 2: Defining the Initial Populations

After defining a map, at least one population has to be defined that consists of one or more individuals. There is no differentiation between individuals and populations in PLABSIM, an individual is repre-

sented by a population of size one. For each individual of a population, the allelic composition at all loci defined in the map must be given by two arbitrarily chosen characters that represent the two alleles, for example, aa, bb, CC, 11, or 12. Background loci carry at the start of the simulation the same allele as the tightest linked locus whose genotype is known.

### Step 3: Description of the Breeding Program

Subsequently the breeding program has to be described by a list of commands for reproduction and selection. Commands for crossing, selfing, and outbreeding are available. Furthermore, varying modes of reproduction and various breeding designs are implemented, such as single-seed descent, double haploids, factorial, and top-cross mating designs.

Selection can be carried out for a certain genotype by specifying a list of loci and the respective allelic composition, or with respect to a selection index. A selection index is constructed by summing the number of loci of a certain class that have a specified allelic composition. For example, in a backcross program the following indices are possible. Selection of individuals (1) heterozygous for the donor allele at all target loci; (2) heterozygous for the donor allele at the maximum number of foreground selection markers; (3) homozygous for the recurrent parent allele at the background selection markers flanking a target gene; or (4) homozygous for the recurrent parent allele at a maximum number of background selection markers. An arbitrary number of selection steps can be carried out in sequence, this allows flexibility in the design and investigation of various selection strategies. In addition to reproduction and selection commands, population manipulations such as random sampling, taking subsets of populations, and merging of populations can be performed.

Because of the stochastic nature of a simulation study, it must be repeated to obtain reproducible results. PLABSIM implements the option to execute the input file in total or in part repeatedly for a number of repetitions specified by the user. The populations that are generated during the repetitions of a simulation can be stored in order to evaluate the results together after finishing all repetitions. Storage of the data can be done in an internal format for evaluation with PLABSIM or in an export format for analysis with statistical software.

### Step 4: Analysis of the Simulated Data

The simulated data can be evaluated with PLABSIM for gene frequencies in three ways: (1) The frequency of an allele that occurs at a defined locus can be calculated. (2) The frequency of the alleles originating from one ancestor can be calculated for a class of loci. This can be used, for example, to estimate the proportion of the recurrent parent alleles at marker loci in backcrossing. (3) The frequency of the alleles originating from one ancestor can be calculated for the whole genome. A possible application is the estimation of the recurrent parent genome proportion in backcrossing.

Genotype frequencies can be calculated either by specifying for each locus on the map an allelic composition, or by specifying for an arbitrary subset of the map an allelic composition. The frequency of the combination of two alleles can be determined for a class of loci. This is useful for estimation of homozygosity or heterozygosity. Furthermore, the distribution of the length of chromosome segments that originate from one ancestor can be calculated.

In addition to estimation of genetic parameters, the number of marker data points required in the selection steps of the breeding program can be estimated.

### Availability

Compiled versions for varying operating systems such as AIX, LINUX, UNIX, or WindowsNT are available. Noncommercial users can obtain the PC version of PLABSIM via e-mail for a nominal charge. Please contact the corresponding author.

From the Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany. The financial support from fellowships by Pioneer Hi-Bred International Inc., Johnston, IA, USA, and KWS Kleinwanzlebener Saat-zucht AG, Einbeck, Germany, to M. Frisch is gratefully acknowledged. We thank Drs. D. Borchart, M. Ouzunova, G. Seitz, and H. F. Utz for comments and suggestions on earlier versions of the simulation software. Address correspondence to A. E. Melchinger at the address above or e-mail: melchinger@uni-hohenheim.de.

© 2000 The American Genetic Association

### References

- Crosby JL, 1973. Computer simulation in genetics. New York: John Wiley & Sons.
- Frisch M, Bohn M, and Melchinger AE, 1999. Comparison of selection strategies for marker-assisted backcrossing of a gene. *Crop Sci* 39:
- Haldane JBS, 1919. The combination of linkage values and the calculation of distance between the loci of linkage factors. *J Genet* 8:299-309.
- Hospital F, Chevalet C, and Mulsant P, 1992. Using

markers in gene introgression breeding programs. *Genetics* 132:1119-1210.

Hospital F and Charcosset A, 1997. Marker-assisted introgression of quantitative trait loci. *Genetics* 147: 1469-1485.

Openshaw SJ, Jarboe SG, and Beavis WD, 1994. Marker-assisted selection in backcross breeding. In: Proceedings of the Symposium "Analysis of Molecular Marker Data," Corvallis, Oregon, 5-6 August 1994. Madison, WI: American Society of Horticultural Science.

Stam P, 1979. Interference in genetic crossing over and chromosome mapping. *Genetics* 92:573-594.

Tinker NA and Mather DE, 1993. GREGOR: software for genetic simulation. *J Hered* 84:237.

Visscher PM, Haley CS, and Thompson R, 1996. Marker-assisted introgression in backcross breeding programs. *Genetics* 144:1923-1932.

Received July 27, 1998  
Accepted August 18, 1999

Corresponding Editor: Robert Angus

## 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 ( $\epsilon$ ) of baseline population  $B_1$  is  $p_1^2$  [the square of its allele frequency ( $p_1$ ) in population  $B_1$ ]. For heterozygotes,  $s_2 \in B_1 = 2p_1q_1$  ( $q_1$  being the frequency of an alternate allele in population  $B_1$ ), and the likelihood that  $s_i \in B_k = p_k^2$  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/which-run.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](mailto: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