

Computer Note

MSTRAT: An Algorithm for Building Germ Plasm Core Collections by Maximizing Allelic or Phenotypic Richness

B. Gouesnard, T. M. Bataillon, G. Decoux, C. Rozale, D. J. Schoen, and J. L. David

A core collection is a subsample of a larger germ plasm collection that contains, with a minimum of repetitiveness, the maximum possible genetic diversity of the species in question (Frankel 1984; Frankel and Brown 1984). Brown (1989a) argued that before creating the core collection, the larger collection should first be hierarchically stratified into groups of accessions that share common characters or that originate from similar ecological and geographic regions. Such a stratification could be based on passport data, knowledge of the structure of the gene pool, or both. Accessions are then drawn from each group. Several sampling strategies are used to determine how to allocate sampling effort across groups (Brown 1989b). In the absence of detailed genetic data about the individuals within the groups, all groups can be represented evenly, or in proportion to their group size, or in proportion to the logarithm of their group size (referred as the C-, P-, and L- strategies, respectively) (Brown 1989b).

An increasing number of germ plasm collections are being genotyped for marker loci such as allozymes, restriction fragment length polymorphisms (RFLPs), and random amplified polymorphic DNA (RAPD). Schoen and Brown (1993) proposed two strategies that can use marker diversity to allocate sampling effort for the construction of the core collection. The H strategy seeks to maximize the total number of alleles in the core collection by sampling accessions from groups in pro-

portion to their within-group genetic diversity. Such an approach assumes that the sampled alleles follow Ewens (1972) sampling theory for neutral alleles, though the approach is robust to several types of departures from this assumption (Brown and Schoen 1994). Schoen and Brown (1993) formulated an alternative strategy, the so-called M (or maximization) strategy which does not necessarily rely upon stratified sampling. The M strategy examines all possible core collections and singles out those that maximize the number of observed alleles at the marker loci. These can then be chosen as final candidates for the core. The expected superiority of this marker-based method is based on the correlation between observed allelic richness at the marker loci and allelic richness on other loci (hereafter referred as to as "target loci"). Such a correlation (or linkage disequilibrium between marker and target alleles) is expected on theoretical grounds either because of (1) shared coancestry of populations, (2) the mating system of the species considered, or (3) episodes of selection whereby selected (target) and neutral (marker) alleles become associated through hitchhiking. Monte Carlo simulations of germ plasm collection and sampling using several marker based sampling strategies have shown that the M strategy performs well when the accessions come from populations with restricted gene flow or when the accessions are predominantly selfing (Bataillon et al. 1996).

While it was initially based on variation at marker loci, the M strategy can be extended to the qualitative and quantitative variables. For quantitative variables, the continuous distribution can be broken into a series of discrete classes. Each accession then belongs to one or several classes for this quantitative variable, depending on the value of the individuals comprising the accession. For each qualitative variable, the number of classes is

determined by the possible values taken by the variable in question. For example, if the variable of disease resistance is coded as either resistant or susceptible, there would be two classes. Richness of a collection of accessions for such a qualitative variable is defined as the number of classes represented among the accessions. Then when considering several variables corresponding to several traits and/or marker loci, the total richness is defined as the sum richness values across variables. The independent contributions of each variable to the sum may be weighted by the importance of the variable; for instance, a given variable may be an important trait or a locus for which allelic variation is desired. The MSTRAT software implements a generalized version of the M strategy (as discussed above), and helps the user to define the size of the core collection to be sampled, as well as the choice of the type of genetic richness to be maximized in the core collection. The software also allows the user to investigate how much genetic richness has been retained for variables that were not used in the sampling of the core collection.

Sampling Procedure

Required Data

The MSTRAT software requires ASCII input files. The data file contains the description of accessions for each variable and allows for several individuals per accession. Different types of variables are defined according to their use in the MSTRAT program. Variables that are used in the optimization of the richness are called active variables. Other variables are called target variables. The diversity score for a given core subset is given for the set of active variables and also for any other series of variables.

Optimization Criterion

The method maximizes the richness of samples. In cases where several putative

core collections have the same richness, a second criterion is used that is near the generalized variance criterion used by Noirot et al. (1995). The core collection with the highest sum of squares for the set of active quantitative variable (nonorthogonal variables) is chosen. This allows one to obtain an optimized distribution of the trait in the core subset.

Algorithm for the Constitution of a Core Collection of Size r

Searching throughout all possible core collections is impractical when the collection to be sampled is large (the number of possible combinations grows factorially with the size of the core and the base collection). Instead, we implemented an algorithm based on an iterative maximization procedure. Briefly, a subset of r accessions is first chosen at random from the N accessions comprising the base collection. In step 1, all the core collections of size $(r - 1)$ are tested. The subset having the highest level of richness is retained. In step 2, among the remnant accessions, the accession bringing the greatest increase in the diversity criterion is retained. Subsequently, steps 2 and 3 are repeated until a convergence criteria is met (e.g., the richness of the core collection is no longer improved or 30 iterations have been performed).

Kernel Core

The sampling procedure allows the user to specify a compulsory set of accessions (the "kernel core") that will always be included in the core collection. For example, the kernel core may represent pivotal accessions that one wishes to retain for historical reasons (e.g., for instance key accessions for breeding programs).

Redundancy in the Collection

Redundancy occurs when accessions overlap in their contribution to the core collection. Redundancy is pivotal in deciding how large the final core collection should be. The extent of potential redun-

dancy can be visualized by graphing the richness of a randomly chosen sample of increasing size r ($1 < r < N$; N being the size of the base collection). For a collection with zero redundancy, one expects a linear increase of richness with r . A convex relationship indicates redundancy in the collection. The choice of the variables defining richness and the number of classes in the variables obviously influence the expected amount of redundancy. By graphing a second curve which shows how richness increases with r under the M strategy, one can assess the gain in diversity achieved by the M method. The inflection point of the M curve provides the optimal size for a core collection. The program also allows external validation of the M strategy by examining the diversity captured for an "independent" set of variables not used in making sampling decisions. Graphs of redundancy can be obtained jointly for both the active and target variables. The MSTRAT software allows one to interactively explore the consequences of the sampling procedure for different sets of traits, and to define a robust minimum size for the core collection based on the actual patterns of redundancy in the collection.

MSTRAT Software Availability

The MSTRAT software uses the Tcl environment, which is freely available for the following platforms: Linux, PC-Windows, and Unix. The MSTRAT.tcl program calls a compiled C program that performs the redundancy study and the core collection sampling. The MSTRAT software and the compiled C program can be obtained free of charge by sending a blank DOS-formatted floppy disk to the authors. The files can also be downloaded at <http://www.ensam.inra.fr/gap/resgen88>. The distribution of the package includes executable programs, a help file, and sample data files. If Tcl is not installed on your system, it can be downloaded at <http://www.scriptics.com/products/tcltk/8.0.html>.

From INRA-SGAP, Centre de Montpellier, Domaine de Melgueil, 34130 Mauguio, France (Gouesnard, Bataillon, Rozale, and David), INRA, Station de Génétique Végétale, Gif sur Yvette, France (Decoux), and Department of Biology, McGill University, Montreal, Quebec, Canada (Schoen). B. Gouesnard acknowledges support from the French Ministère de l'Agriculture and Bureau des Ressources Génétiques. D. J. Schoen acknowledges support from the National Science and Engineering Research Council of Canada and from INRA. Address correspondence to Brigitte Gouesnard at the address above, or e-mail: gouesnard@ensam.inra.fr.

© 2001 The American Genetic Association

References

- Bataillon TM, David JL, and Schoen DJ, 1996. Neutral genetic markers and conservation genetics: simulated germplasm collections. *Genetics* 144:409-417.
- Brown AHD, 1989a. A case for core collections. In: *The use of plant genetic resources* (Brown AHD, Frankel OH, Marshall DR, and Williams JT, eds). Cambridge: Cambridge University Press; 136-156.
- Brown AHD, 1989b. Core collections: a practical approach to genetic resources management. *Genome* 31: 818-824.
- Brown AHD and Schoen DJ, 1994. Optimal sampling strategies for core collections of genetic resources. In: *Conservation genetics* (Loeschcke V, Tomiuk J, and Jain SK, eds). Basel, Switzerland: Birkhäuser Verlag; 354-370.
- Ewens WJ, 1972. The sampling theory of selectively neutral alleles. *Theor Popul Biol* 3:87-112.
- Frankel OH, 1984. Genetic perspectives of germplasm conservation. In: *Genetic manipulation: impact on man and society* (Arber W, Llimensee K, Peacock WJ, and Starlinger P, eds). Cambridge: Cambridge University Press; 161-170.
- Frankel OH and Brown AHD, 1984. Plant genetic resources today: a critical appraisal. In: *Crop genetic resources: conservation & evaluation* (Holden JHW and Williams JT, eds). London: George Allen & Unwin; 249-257.
- Noirot M, Hamon S, and Anthony F, 1995. The principal component scoring: a new method of constituting a core collection using quantitative data. *Genet Resource Crop Evol* 41:1-6.
- Schoen DJ and Brown AHD, 1993. Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. *Proc Natl Acad Sci USA* 22: 10623-10627.
- Schoen DJ and Brown AHD, 1995. Maximising genetic diversity in core collections of wild relatives of crop species. In: *Core collections of plant genetic resources* (Hodgkin T, Brown AHD, van Hintum ThJL, and Morales AEV, eds). Chichester, UK: John Wiley & Sons; 55-76.

Received September 28, 1999

Accepted August 30, 2000

Corresponding Editor: Bruce S. Weir