References

Danin-Poleg Y, Teis N, Baudracco-Arnas S, Pitrat M, Staub JE, Oliver M, Arús P, deVicente MC, and Katzir N, 2000. Simple sequence repeats in *Cucumis* mapping and map merging. Genome 43:963–974.

Devos KM, Pittaway TS, Reynolds A, and Gale MD, 2000. Comparative mapping reveals a complex relationship between the pearl millet genome and those of foxtail millet and rice. Theor Appl Genet 100:190–198.

Grant D, Cregan P, and Shoemaker RC, 2000. Genome organization in dicots: genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis.* Proc Natl Acad Sci USA 97:4168–4173.

Joobeur T, Periam N, deVicente MC, King G, and Arús P, 2000. Development of a second generation linkage map for almond using RAPD and SSR markers. Genome 43:649–655.

Maaliepaard C, Alston F, Van Arkel G, Brown LM, Chevreau E, Dunemann G, Evans KM, Gardiner S, Guilford P, van Heusden AW, Janse J, Laurens F, Lynn JR, Manganaris AG, den Nijs APM, Periam N, Rikkenrink E, Roche P, Ryder C, Sansavini S, Schmidt H, Tartarini S, Verhaegh JJ, Vrielink-van Ginkel M, and King G, 1998. Aligning male and female linkage maps of apple (*Malus pumilla* Mill.) using multi-allelic markers. Theor Appl Genet 97:60–73.

Received March 21, 2001
Accepted January 30, 2002

Corresponding Editor: Leif Andersson

MDM: A Program to Compute Fully Informative Genotype Frequencies in Complex Breeding Schemes

B. Servin, C. Dillmann, G. Decoux, and F. Hospital

In many genetics studies it is necessary to compute the expected frequencies of genotypes at marker loci and/or to infer the genotypes at chromosomal locations from the known genotypes at markers. This is the case, for example, for quantitative trait loci (QTL) detection, where likelihood ratio tests or multiple regressions are based on the probabilities of the different genotypes at a putative QTL location, given the genotypes at flanking markers. This is also the case for "graphical genotypes" (Young and Tanksley 1989), where it is wanted to estimate the genomic composition of the chromosomes (parental origin of the alleles) given the genotypes at markers.

The calculations performed by most existing programs (e.g., for QTL detection) are based solely on the genotypes at the two closest markers flanking the putative position on each side, observed at only one generation. This is sufficient only if the population considered has issued from a single generation of effective recombination (e.g., BC_1, F_2) and if marker genotypes are known without ambiguity. If there is ambiguity in marker genotypes (e.g., dominance, missing data) or if more than one effective meiosis has taken place (e.g., F_3, recombinant inbred lines (RILs), advanced backcross generation), then additional markers further from the putative position and/or marker genotypes at previous generations may also be informative and could permit more accurate prediction of the genotype at the putative position.

Also, available programs generally consider fixed and reasonably simple breeding schemes (e.g., F_2, F_3, RIL, BC_n), whereas breeding schemes of higher (and arbitrary) complexity are more and more often used in practice (e.g., random mating before selfing to produce highly recombinant inbred lines (HRILs) with higher apparent recombination rate, BC followed by selfing to fix the introgressions, etc.).

MDM is a program that computes frequencies of multilocus genotypes in populations derived from breeding schemes involving any combination of selfing, full-sib mating, random mating, backcrossing, or hybrid mating that takes into account all the genotypic information available (flanking and nonflanking markers, intermediate generations). It can be used interactively to perform the relevant calculations on experimental data, or it can be included as a function in QTL detection programs or in simulation programs aimed at optimizing breeding schemes before proceeding to the experiments. More generally, MDM was designed for fast and easy numerical computation of multilocus genotype frequencies in arbitrary breeding schemes, avoiding cumbersome analytic derivations.

Principle

The program works with a collection of loci (typically marker loci) described by their positions on a genetic map. Given a pedigree, the program computes the probabilities of the offspring genotypes at generation n. The pedigree is defined by the genotypes of the ancestors at each former generation (from generation 1 to $n - 1$), and by the mating systems used between generations. The breeding scheme (i.e., the succession of mating systems) can be any combination of backcrossing, hybrid mating, full-sib mating, or self-ing. Depending on the mating system, one or two ancestor genotypes are needed at each generation. A single ancestor (herein called the maternal ancestor) is needed for selfing or full-sib mating. A second ancestor (herein called the paternal ancestor) is needed for hybrid mating or backcrossing.

In practice, the genotyping of an individual produces an observation (i.e., "phenotype") that poorly reflects its true genotype. Indeed, usually the marker phenotypes do not provide the gametic phase of the chromosomes (which allele originates from which parent), for example, in the case of a double or multiple heterozygote. Furthermore, genotyping data may not be fully informative, because of missing or incomplete data (e.g., in the case of dominant markers). So the program distinguishes between "observed genotypes" (OGs), allowing missing or incomplete genotyping data, and "true genotypes" (TGs), where all alleles at all loci as well as the gametic phase are assumed to be known.

Individuals are described by their OGs at all loci. The coding of the OGs and the relationship between OGs and TGs is user defined, allowing the user to work with any genotype-coding system used in particular experiments. According to the coding system, OGs at each generation are converted into all possible sets of corresponding TGs. Then, the probabilities of transition between all possible sets of TGs at different generations are computed according to the recursion equations of Hospital et al. (1996). Finally, these probabilities are summed to provide the probability that each offspring genotype at generation n issued from the ancestors in previous generations given the breeding scheme.

An additional locus (typically a putative QTL position, or a point on a chromosome) can be included in the calculations. Thus the program computes two sets of genotypic frequencies: the frequencies of OGs at marker loci only, and the frequencies of OGs at marker loci plus the additional locus. This allows the user to compute the conditional probabilities of putative genotypes at the additional locus given the observed genotypes at marker loci.

The maximal number of loci that can be considered simultaneously depends on computer memory size (e.g., taking seven loci into account requires 64Mb RAM). The number of ancestor generations is unbounded, but affects computing time (in conjunction with the number of loci).

Running the Program

A text format (ASCII) file is used to set the value of the parameters used for the computation: number of generations of the breeding scheme, number of offspring, number of genotypes to allocate to the additional locus in the offspring, and name of the file containing the coding system. It also contains the mating systems used in each generation. Finally, it contains the genetic map (chromosomes, names and positions of markers) along with the genotypes of the ancestors and the offspring at the marker loci. The genetic map used by MDM is constant during the whole breeding scheme. It is therefore assumed that recombination rates between loci are evaluated once (either on the population studied or on another one, e.g., if using a consensus or a joint map) and that they are not reestimated during the breeding scheme.

It is possible to compose a large marker dataset only once, then run the program with different subsets of these markers on a given chromosome. This is particularly useful if the total number of loci on the chromosome is larger than MDM can handle.

The computations performed by MDM can be customized by using options, such as including an additional locus in the computation. MDM can be run several times for different positions of an additional locus. This can be used to perform a chromosome scan of the different offspring or to analyze a particular chromosome segment of interest (e.g., containing a QTL). In this case, it is possible to obtain the conditional probabilities of several genotypes at the additional locus, given the observed genotype at markers for each offspring. The output of the results can be either detailed, including recalling of the input parameters, or brief. This last option makes it easier for other programs to use the results provided by MDM.

Another way to make MDM interact with other programs is to use its core computation function as a subroutine. For this, the source code of the MDM is split into two files, one containing the core computation function, one containing other functions used to manage input and output.

Package

MDM is written in ANSI C and has been developed under a Linux/UNIX environment using the GNU C compiler (gcc). This compiler is included in all Linux and UNIX distributions. It is also freely available for Windows and DOS environments (www.delorie.com/djgpp). It is therefore easy to compile MDM and to use it under other environments (e.g., Windows 9x and NT or DOS).

The MDM package contains the source code and binaries of the program (including a Windows executable file) along with a user's manual including the rules to write input files and examples. The package can be obtained free of charge by sending a blank DOS-formatted floppy disk to the corresponding author. The files are also freely available for downloading at http://moulon.inra.fr/~servin/mdm.

From the *Station de Génétique Végétale, INRA/UPS/INAPG, Ferme du Moulon, 91190 Gif sur Yvette, France.* Address correspondence to Bertrand Servin at the address above or e-mail: servin@moulon.inra.fr.

References

Hospital F, Dillmann C, and Melchinger AE, 1996. A general algorithm to compute multilocus genotype frequencies under various mating systems. Comput Appl Biosci 12:455–462.

Young ND and Tanksley SD, 1989. Restriction fragment length polymorphism maps and the concept of graphical genotypes. Theor Appl Genet 77:95–101.

Received February 1, 2001
Accepted December 31, 2001

Corresponding Editor: Robert Angus