Bronson, C., C-S. Chang and T-H. Tzeng

HAPMAP: a computer program for the

linkage analysis of haploids.

The development of technology for the detection of variations in DNA sequence is permitting the rapid mapping of the genomes of many organisms. Restriction fragment length polymorphism (RFLP) maps promise to be particularly useful for economically important fungi for which limited numbers of phenotypic markers are usually available. RFLP markers are virtually unlimited and may be used as the starting points for genomic walks to genes of interest. Our current efforts to create a saturated RFLP map of the maize pathogen Cochliobolus heterostrophus have made us aware of the need for computer assistance to handle the otherwise unwieldy number of marker comparisons. In this note, we describe a fast, menu-driven linkage analysis program suitable for the RFLP mapping of haploid organisms.

Program description:

HAPMAP is a computer program that calculates linkage distances based on the phenotypes of random progeny of a cross between haploid organisms. Data for different crosses are entered and analyzed independently. No assumptions are made about the number of linkage groups present in the organism, and no attempts are made to adjust linkage distances for double crossovers or interference.

Input: For each cross, data on progeny phenotype for each marker are entered into a preformed table (data file). The table accepts up to 112 progeny and up to 200 markers. Markers, progeny and their phenotypes may be added to or changed at any time. The notation for phenotype is one that was convenient for our crosses: F = phenotype of one parent, L = phenotype of the other parent. Marker-progeny combinations for which no phenotypic data are available are indicated with a dash. The table may be displayed and printed (Table 1).

Table 1. Sample data file generated by HAPMAP

Marker	4	8	12	16	20	24	28	32	36	40	44	48	52	56
GI27	:FLFF	FLFL	FLLF	LLLL	LLFF	FFLF	FFFF	FLFF	LLFL	LFLL	LLFL	LFFL	LLFF	LLFL
	:FLFL	FLLL	LFFL	FLFL	FFLF	FLLF	FFLF	LFFL	LFF					
B71	:FFFF	FLFF	FLLF	LLLF	FFFF	FFFF	FFFF	FLFF	LLFF	LLLF	LLFL	LLFF	LLFF	LFLL
	:F ----	------	------	------	------	------	-------	------	-------	------	------	------	-------	------
G214	:FLFF	FLLL	FLFF	LLLL	LLFF	FFLF	FFFF	FLFF	FLLL	LFLL	LLFL	LFFL	LLLF	LLFL
	:F-----	------	------	------	------	------	-------	------	------	------	------	------	------	------

60	64	68	72	76	80	84	88	92	96	100	104	108	II2

Output: The program provides information on both marker segregation and marker recombination.

a. Marker Segregation - This subroutine calculates, displays and prints for each marker requested the number of progeny with each parental phenotype, and the number of scored (Scd) and unscored (UScd) progeny. This permits a quick visual check for the random recovery of parental phenotypes (Table 2).

Table 2. Sample printout from marker segregation subroutine

Marker	L	F	Scd	UScd	Marker	L	F	Scd	UScd
G127	45	46	91	0	B71	22	35	57	34
G214	30	27	57	34					

b. Marker Recombination - Linkage analyses may be requested for all of any combination of markers. For each pair of markers analyzed, the program will calculate, display and print the following:

1. recombination frequency x 100 (MU = map units)

2. Chi-square (X^2) for the null hypothesis that the two markers are unlinked. The chi-square value is calculated from a 2 x 2 contingency table (1 degree of freedom) according to the equation:

$$X^2 = \sum_{i=1}^{4} \frac{([obs - exp] - 0.5)^2}{exp}$$

where i = the four possible phenotypic classes of progeny (see 6).

3. 95% confidence interval (95% CI) for the recombination frequency calculated according to the equations:

$$PL = \frac{(2np + c^2 - 1) - c\,[c^2 - (2 + 1/n) + 4p(nq + 1)]^{\frac{1}{2}}}{2(n + c2)}$$

$$PU = \frac{(2np + c^2 + 1) + c\,[c^2 + (2 - 1/n) + 4p(nq - 1)]^{\frac{1}{2}}}{2(n + c^2)}$$

where PL = lower limit, PU = upper limit, n = number of progeny scored for the markers being analyzed, p = proportion of recombinants, q = 1-p, c = 1.96 and when p = 0, the PL = 0. (J.L. Fleiss, 1981, Statistical Methods for Rates and Proportions, 2nd Ed., Wiley, NY pp 14-15).

4. the number of progeny scored (Scd)

5. the number of recombinant progeny (MM = mismatches)

6. the number of progeny in each of the four possible phenotypic classes (the two parental type, LL and FF, and the two recombinant types, LF and FL).

This subroutine can calculate and display statistics for over 100 marker comparisons in less than 18 seconds. To reduce the volume of output when large numbers of markers are analyzed, the user may request that analyses for only those marker pairs showing deviation from random association be printed (P = 0.05, $X^2 > 3.84$) (Table 3).

Table 3. Sample printout from marker recombination subroutine

M	M2	MU.	X2	95%	C.I.	Scd	MM	LL	LF	FL	FF
G127	G214	8.8	35.55	3.3	20.1	57	5	27	2	3	25
G127	B71	22.8	15.81	13.2	36.4	57	13	19	10	3	25
G214	B71	31.6	7.19	20.3	45.6	57	18	17	13	5	22

Program Requirements:

The program is written in 8086/8088 assembly language and will run on IBM PC/XT (not AT) compatible microcomputers equipped with MS/DOS or PC-DOS operating systems (versions 2.0 or higher) and a minimum of 48 K RAM. Print commands are those provided by MS-DOS.

Availability:

The length of the program precludes publication in this newsletter. The program may be obtained free of charge on a 5 ¼ " diskette from the senior author. Source codes will be included to permit the modification or addition of subroutines as desired by the user.