Genetics and Molecular Biology, 30, 4, 1101-1108 (2007)

www.sbg.org.br

Research Article

Simulation of population size and genome saturation level for genetic mapping of recombinant inbred lines (RILs)

Luciano da Costa e Silva[1], Cosme Damião Cruz[1], Maurilio Alves Moreira[2] and Everaldo Gonçalves de Barros[1]

[1]*Departamento de Biologia Geral, Universidade Federal de Viçosa, Viçosa, MG Brazil.*
[2]*Departamento de Bioquímica e Biologia Molecular, Universidade Federal de Viçosa, Viçosa, MG, Brazil.*

Abstract

Various population sizes and number of markers have been used to obtain genetic maps. However, the precise number of individuals and markers needed for obtaining reliable maps is not known. We used data simulation to determine the influence of population size, the effect of the degree of marker saturation of the genome, and the number of individuals required for mapping of recombinant inbred lines (RILs). Three genomes with 11 linkage groups were generated with saturation levels of 5, 10 and 20 cM. For each saturation level populations were generated with 50, 100, 154, 200, 300, 500 and 800 individuals with 100 replications for each population size. A total of 2100 populations was generated and mapped. Small marker numbers and small population sizes produced maps with more than 11 linkage groups. As population size and marker saturation increased, marker inversion and non-linked markers decreased, moreover, between-marker distance estimates were improved. In this study, a minimum size of 200, 300 and 500 individuals were necessary for obtaining reliable maps when they were evaluated over the saturation levels of 5, 10 and 20 cM, respectively.

Key words: recombinant inbred lines, molecular markers, number of individuals, computer simulation.

Received: January 4, 2007; Accepted: April 27, 2007.

Introduction

In plant and animal breeding, genetic maps are important tools for analyzing genomes and dissecting complex traits into their simple Mendelian determinants. They also allow for identification of genome regions harboring genes controlling qualitative and quantitative traits (Lander and Botstein, 1989). However, availability of reliable maps depends on several factors such as the type and size of population and the type and number of markers. In addition, other aspects must also be considered such as single loci segregation ratio, recombination frequency and logarithm of odds (*LOD*) thresholds used to infer linkage.

In plants, populations obtained from crossing two inbred lines are commonly used for mapping, with the F_2, backcrosses (BC), F_n (n = 3, 4,..., ∞), double haploids (DH) and recombinant inbred lines (RIL) being the most frequently used populations (Burr *et al.*, 1988). Alternatively, outbred populations such as half and full sibs can be used. The choice of the population depends on the species studied, program goals and availability of time and funds.

Send correspondence to Everaldo G. Barros. Departamento de Biologia Geral, Universidade Federal de Viçosa, 36570-000 Viçosa, MG Brazil. E-mail: ebarros@ufv.br

A recombinant inbred line can be obtained from an F_2 generation by successive self-pollinations using the single seed descent method (SSD) (Burr *et al.*, 1988). The resulting inbred lines are highly homozygous and the segregation ratio for each locus tends to 1:1 (AA:aa). Disadvantages of recombinant inbred lines are that at least six generations are required to obtain the line and the inability to estimate dominance effects of mapped quantitative trait loci (QTL) due to the absence of heterozygous genotypes. However, the advantage of recombinant inbred lines is that because they are made up of homozygotes only they are stable and can thus be used in experiments with replications in several environments allowing for more accurate estimates of genetic components and identification of QTL *vs* environment interactions. Moreover, because several cycles of meiosis occur during the development of such lines there are opportunities for recombination between tightly linked loci. In recombinant inbred lines, the recombination frequency among loci is given by $R = 2r(1+2r)^{-1}$, where r expresses the recombination frequency in the corresponding F_2 (Burr and Burr, 1991).

Several articles have been published on the mapping of recombinant inbred lines using molecular markers. In soybean (*Glycine max* L. Merr.), Burnham *et al.* (2003) used 64 lines and 75 markers and Ferreira *et al.* (2000) used

330 lines and 356 marker. In maize (*Zea mays* L.), Burr *et al.* (1988) used 48 lines and 134 markers and Cardinal *et al.* (2001) used 183 lines and 185 markers. In rice (*Oryza sativa* L.), Zheng *et al.* (2003) used 96 lines and 249 markers and Xing *et al.* (2002) used 240 lines and 213 markers and in the common bean (*Phaseolus vulgaris* L.), Miklas *et al.* (2001) used 67 lines and 245 markers and Faleiro *et al.* (2003) used 154 lines and 43 molecular markers. These examples, selected for their extremes of high and low population sizes, show that there is no consensus on the number of markers and the population size to be used for mapping, even considering the same crop.

In a recent simulation study (Ferreira *et al.*, 2006) the effects of size and type of population on the accuracy of genetic maps were estimated using a model that considered one chromosome and nine makers equidistantly separated by 10.13 cM from each other. The results showed that more accurate maps are obtained with F_2-codominant and recombinant inbred lines than with backcrosses, double haploids and F_2-dominant populations and that a sample size of 200 individuals is sufficient for the construction of reasonably accurate maps.

The study described in the present paper was a more in depth simulation study of recombinant inbred lines derived from a hypothetical diploid species (2n = 2x = 22) to determine the effect of population size and genome saturation with molecular markers on the reliability of the maps obtained.

Materials and Methods

Data of a hypothetical recombinant inbred line of a diploid species were generated using the GQMOL software, which can generate information about the genome, parental genotypes, individuals from different types of populations and quantitative trait data.

Simulated genomes and simulation of parental inbred lines

A hypothetical diploid species with a chromosome complement of 2n = 2x = 22 and a genome length of 1100 cM was used as a standard to generate three genomes with saturation levels of 5 cM for 231 markers, 10 cM for 121 markers and 20 cM for 66 markers. Each genome contained 11 linkage groups of 100 cM each and markers were equally spaced within the groups, thus each genome was 1100 cM long.

For each level of genome saturation (5 cM, 10 cM and 20 cM) we simulated two parental inbred lines, both of which were homozygous but had different marker alleles at each of the simulated marker loci, leading to an F_1 generation with all loci in the coupling phase (*cis*).

Population sizes and simulation of individuals

For each of the three genome saturation levels (5 cM, 10 cM and 20 cM) we generated seven populations with different numbers (n) of individuals (n = 50, n = 100, n = 154, n = 200, n = 300, n = 500 and n = 800), each population being replicated 100 times to form a total of 2100 populations (*i.e.* 3 saturation levels x 7 population sizes x 100 replications). For the simulation of individuals we used the approach described by Ferreira *et al.* (2006), which can be summarized as follows: for each level of genome saturation (5 cM, 10 cM and 20 cM) a set of 10 000 possible recombinant inbred line genotypes that fitted the expected segregation ratio were generated from an F_1 population and 100 replications per population size were obtained by randomly identifying 100 different sets of recombinant inbred lines among the 10 000 initial genotypes. To account for the extra recombination which occurs in recombinant inbred lines as compared to an F_2 population we corrected the recombination probabilities, *e.g.* a distance of 10 cM was represented by a recombination probability of 16.667% to account for the increased recombination as a result of multiple meiotic events that occur during the development of recombinant inbred line.

Mapped genomes

Between-marker recombination frequencies were obtained using the maximum likelihood method described by Schuster and Cruz (2004). Two markers were assumed to be linked when both recombination frequency was less than 30% and the *LOD* score was greater than three. The initial marker order was estimated based on the recombination frequency and *LOD* score and the final order was determined using the sum of adjacent recombination fractions (SARF) method (Falk, 1989) using the rapid chain delineation (RCD) algorithm (Deorge, 1993).

Comparison between simulated and mapped genomes

A total of 2100 mapped genomes obtained from simulated populations were compared to the simulated genomes using the following criteria: the number of linkage groups, the number of markers per group, the mean distance between adjacent markers, marker inversion (*i.e.* the change in the order of markers as given by the Spearman correlation) and the agreement of distances between markers in mapped versus simulated genome as given by the stress coefficient (Kruskal, 1964). In the analyses of the number of linkage groups and number of markers per group all replications (100) were used, while in all other analyses only replications (*i.e.* simulated populations) that formed 11 linkage groups were used. All statistics were as described by Ferreira *et al.* (2006).

The stress coefficient (S) has been used as a measure of goodness of fit of raw distances and graphic projection of these distances. In the context of this paper, the stress coefficient provides a measure of agreement of between-marker distances in the simulated genome and mapped genome (simulated population). If the between-marker distances in

the mapped genome are the same as in the simulated genome the stress would be zero and thus indicate no changes in distances between markers from the simulated genome to the mapped genome, implying a perfect recovery of the simulated genome. In the expression of stress given by Ferreira *et al.* (2006) a meaningful interpretation can be given to the stress coefficient if we consider the term (d_{ok} - d_k) to be constant, in which case the stress (S) would be expressed as $S = 100\ (\overline{d}/d_{ok})$ where $\overline{d}$ is the mean deviation and d_{ok} the distance between adjacent markers in the simulated genome. Thus, if S is 20% for a 5 cM genome the mean deviation would be 1 cM, indicating that the markers are, on average, 4 cM or 6 cM apart in the mapped genome. The expression also implies that stress values have different meanings depending on the degree of genome saturation, *e.g.* for a 10 cM genome an S-value of 20% would indicate a mean deviation of 2 cM.

Results

Number of linkage groups and markers per group

The number of replications that led to the formation of 11 linkage groups in the mapping of the simulated populations and the minimum and maximum number of unlinked markers obtained are shown in Table 1. However, population sizes of n = 50 for the 20 cM and 10 cM genomes or n = 100 for the 20 cM genome showed no replications with 11 linkage groups (q.v. 'Simulated genomes', above), because of which we considered these combinations of population sizes and genome saturation levels inappropriate for mapping or making further comparisons and these are not shown in Table 1. Furthermore, the n = 154 populations for the 20 cM genome produced only 48 replications with 11 linkage groups, this population size for the 20 cM genome also being omitted from the analysis. The decision to omit these populations from the analysis was also supported by the minimum (min) and maximum (max) numbers of unlinked markers, which were 0 min and 9 max for the n = 50 populations in the 10 cM genome, 24 min and 47 max for the n = 50 populations in the 20 cM genome, 2 min and 19 max for the n = 100 populations in the 20 cM genome and 0 min and 3 max for the n = 154 populations in the 20 cM genome (data not shown in Table 1). These values, especially the maximum values, are higher than those for the other population sizes (Table 1). The number of linkage groups obtained as a function of population size for the 5 cM genome saturation level is shown in Figure 1.

Spearman correlation

The n = 50, n = 100 and n = 154 populations for the 5 cM genome showed replications with marker inversion and the number of inversions was greater for the smaller populations, as shown by the fact that in the n = 50 populations all linkage groups showed replications with inverted markers whereas in the n = 100 populations only five linkage groups showed marker inversion and in the n = 154 populations only one linkage group showed one replication with inverted markers. For the 10 cM genome the n = 100 and n = 154 populations were the only ones showing replications with marker inversions, while for 20 cM genome only the n = 200 populations showed marker inversion (Table 2). As discussed above, the n = 50 populations for the 10 cM genome and the n = 50, n = 100 and n = 154 populations for the 20 cM genome were not used either for Spearman correlation analysis or any subsequent analyses.

Mean distance between adjacent markers and stress

The between-marker distances showed deviations from the expected values for the 5 cM, 10 cM and 20 cM genomes as a consequence of population size and genome saturation level.

For the 5 cM genome as the population size increased the observed and expected mean distance between adjacent markers became closer and there was a reduction in the deviation from the expected value for most linkage groups (Figure 2). This indicates the effect of population size, illustrated by the fact that for the 5 cM genome n = 50 population the mean distance in linkage group 1 was 5.26 cM with a deviation of ±0.65 cM from the expected value, while for the 5 cM genome n = 800 population the respective values were 5.15 cM ± 0.19 cM. However, there were exceptions to this trend, with linkage group 6 presenting mean distances of 5.10 cM for the 5 cM genome n = 50 population and 5.14 cM for the 5 cM genome n = 800 population. Linkage groups 1, 4, 5, 6, 7, 9 and 10 showed no statistical differences between means for the different-sized 5 cM genome populations but differences were observed for the other linkage groups, with the larger populations showing mean distances approaching the expected value for the specific 5 cM genome population concerned. For the general mean (the average over all 11 linkage groups) there were significant differences between different-sized populations, with the mean distance converging to 5 cM genome as the population size increased.

The populations generated from the 10 cM genome also showed mean distances between markers approaching the expected value of 10 cM and a reduction in the deviation from the expected value as the size of the population increased. For the general mean, smaller mean distances were significantly associated with larger population sizes. Populations generated from the 20 cM genome showed the same behavior as the populations generated from the 5 and 10 cM genomes. However, at the 20 cM marker saturation level significant differences between means were only obtained for linkage group 2. For the general mean there were no significant differences between means for the different population sizes but the deviation from the expected value decreased as population size increased, *i.e.* for the 20 cM genome n = 200 population the mean was 20.19 cM ±

Table 1 - Stress as function of genome marker saturation level and population size (n). The second column shows the maximum (max) and minimum (min) number of unlinked markers in each population. Each population ($n = 50$, $n = 100$, etc) was subjected to 100 replications but the size of the linkage groups was evaluated only for those populations that led to the formation of 11 linkage groups during the mapping of the simulated populations. The general means for the percentage stress for the 11 linkage groups are shown in the last column. Within each saturation level means followed by the same letters in the same column do not differ by the Tukey test ($p < 0.01$).

Genome saturation (cM) and number of individuals per population (n)	Unlinked markers min/(max)	Number of populations with 11 linkage groups	Linkage groups											Stress general mean (%)
			1	2	3	4	5	6	7	8	9	10	11	
			Stress (%)											
5 cM populations														
$n = 50$	0/1	70	48.0a[2]	51.0a	49.3a	50.6a	48.9a	49.8a	49.7a	49.6a	49.9a	49.3a	51.6a	49.8a
$n = 100$	0/0	100	35.8b	36.1b	35.6b	35.5b	35.6b	35.2b	34.6b	35.4b	35.7b	36.1b	34.9b	35.5b
$n = 154$	0/0	100	28.7c	28.7c	28.7c	28.0c	28.7c	28.8c	29.1c	28.8c	28.7c	28.2c	27.4c	28.5c
$n = 200$	0/0	100	24.6d	25.1d	25.1d	24.6d	24.3d	25.2d	25.0d	26.3c	25.6d	24.8d	25.0c	25.1d
$n = 300$	0/0	100	19.5e	20.1e	20.8e	20.4e	20.4e	20.4e	20.2e	20.5d	20.5e	20.5e	20.5d	20.3e
$n = 500$	0/0	100	15.6f	15.6f	15.6f	16.1f	15.7f	15.9f	15.8f	15.5e	16.0f	15.9f	15.6e	15.8f
$n = 800$	0/0	100	12.9g	12.3g	12.5g	12.6g	12.7g	12.6g	12.8g	12.9f	12.4g	13.0g	12.5f	12.7g
10 cM populations														
$n = 100$	0/1	98	27.4a	27.6a	28.3a	28.1a	27.4a	28.1a	27.3a	27.7a	27.0a	27.4a	26.5a	27.5a
$n = 154$	0/0	100	22.1b	21.2b	21.4b	21.5b	22.3b	21.5b	22.2b	22.8b	22.7b	21.5b	22.1b	21.9b
$n = 200$	0/0	100	19.2c	20.3b	19.6b	18.7c	19.1c	19.2c	18.5c	18.9c	19.3c	19.3c	19.0c	19.2c
$n = 300$	0/0	100	16.1d	15.6c	16.1c	15.5d	15.3d	16.6d	16.5c	15.6d	15.5d	15.3d	15.7d	15.8d
$n = 500$	0/0	100	12.1e	12.7d	12.2d	12.0e	12.1e	11.9e	12.4d	11.9e	12.1e	12.3e	12.1e	12.1e
$n = 800$	0/0	100	9.53f	9.61e	10.18d	9.88f	9.63f	9.87e	9.82e	9.74f	10.06e	9.41f	9.37f	9.74f
20 cM populations														
$n = 200$	0/1	86	15.2a	15.4a	15.5a	15.6a	14.8a	14.6a	14.7a	14.9a	14.7a	14.9a	14.5a	15.0a
$n = 300$	0/1	98	11.6b	12.8b	12.0b	12.1b	12.5b	12.1b	12.1b	12.3b	13.4a	12.6b	12.7b	12.4b
$n = 500$	0/0	100	9.0c	9.2c	9.2c	9.7c	9.2c	9.7c	9.0c	9.5c	9.3b	9.2c	9.2c	9.3c
$n = 800$	0/0	100	7.5c	7.1d	7.7c	7.6d	7.3d	7.6d	7.7c	7.8d	7.7c	7.7c	7.7c	7.6d

Table 2 - Linkage groups with inverted markers. The table shows the genome marker saturation level (in cM) and the number of individuals (n) in each population size in respect to the number of populations with 11 linkage groups out of a total of 100 populations for each genome marker saturation level population size (n = 50, n = 100, etc). The number of populations (n = 50, n = 100, etc) with inverted markers (between parenthesis) was only evaluated in the populations that led the formation of 11 linkage groups.

Genome saturation (cM) and number of individuals per population (n)	Number of populations with 11 linkage groups	Linkage groups										
		1	2	3	4	5	6	7	8	9	10	11
		Number of populations with specified number of linkage groups (number of populations with inverted markers between parenthesis)										
5 cM populations												
n = 50	70	58 (12)	51 (19)	60 (10)	53 (17)	52(18)	50 20)	50 (20)	45 (25)	61 (9)	48 (22)	57 (13)
n = 100	100	98 (2)	100 (0)	100 (0)	98 (2)	99 (1)	100 (0)	100 (0)	100 (0)	100 (0)	99 (1)	98 (2)
n = 154	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	99 (1)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
10 cM populations												
n = 100	98	97 (1)	97 (1)	98 (0)	98 (0)	98 (0)	98 (0)	98 (0)	97 (1)	98 (0)	97 (1)	96 (2)
n = 154	100	100 (0)	99 (1)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
20 cM populations												
n = 200	86	86 (0)	86 (0)	86 (0)	86 (0)	85 (1)	85 (1)	86 (0)	86 (0)	85 (1)	86 (0)	86 (0)

1.43 cM while for the 20 cM genome n = 800 population it was 20.15 cM ± 0.73 cM. For all population sizes and all three saturation levels the distance between markers results allowed us to conclude that the accuracy of the distance estimates improves with increasing population size.

The stress (S) means for each linkage group as a function of population size and marker saturation level are shown in Table 1 and Figure 3.

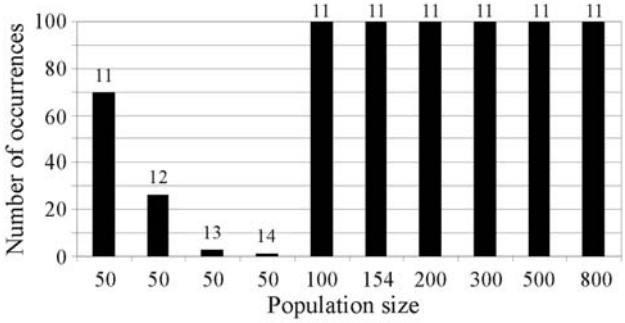

Figure 1 - Distribution of the number of linkage groups (indicated at the top of each bar) obtained in the mapping of the simulated populations as a function of population size. The evaluation used 100 replications for each population size, simulated using a genome with marker saturation level of 5 cM.

Discussion

Recombination frequency and *LOD* score are the two parameters used to infer linkage between markers.

Regarding recombination frequency, it is well-known that segregating populations consisting of only a small number of individuals do not provide a good sample of the total gametic diversity of the parents and since the distance between markers is calculated by genotyping individuals from a segregating population and counting the recombinants for each pair of loci an inadequate sample of gametes leads to a poor estimate of genetic distance. In our study the n = 50 populations for the 5 cM, 10 cM and 20 cM genomes

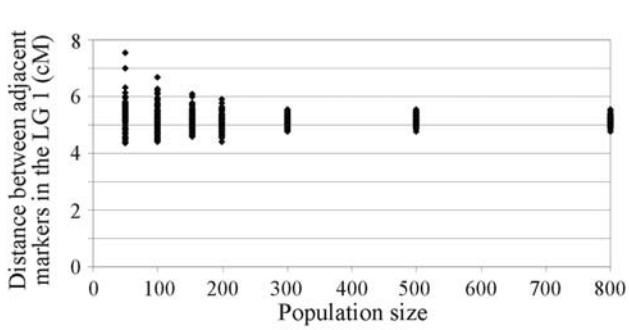

Figure 2 - Dispersion of the distance between adjacent markers in linkage group 1 (LG 1) as a function of population size, simulated using a genome with marker saturation level of 5 cM. Evaluation was done using only the replications that led to the formation of 11 linkage groups.

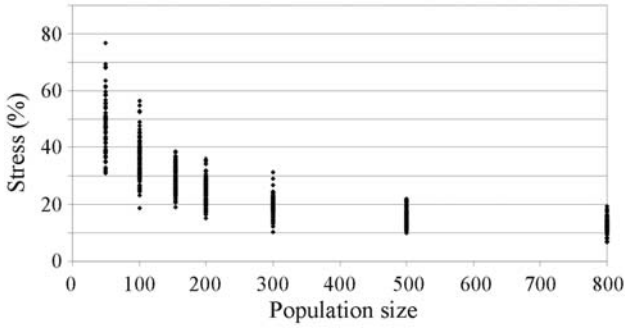

Figure 3 - Dispersion of stress coefficients for linkage group 1 as a function of population size, simulated using a genome with marker saturation level of 5 cM. Stress was evaluated only in the replications that led to the formation of 11 linkage groups.

were inadequate to allow detection of linkage where it was present. For mapping recombinant inbred lines adequate population sizes and the use of an appropriate number of markers are well-known prerequisites for producing a good estimation of recombination frequency, as illustrated by the recombination frequency variance $Var(\hat{r}) = [\hat{r}(1+2\hat{r})^2 / 2n]$ (Schuster and Cruz, 2004) in which n is the population size and $\hat{r}$ is the maximum likelihood estimate of the recombination frequency in the F_2 generation as function of the observed recombination in the recombinant inbred lines, *i.e.* $\hat{r} = R / [2(1-R)]$, where R is the recombination in the recombinant inbred lines (Burr and Burr, 1991). It is clear from this equation that a more accurate estimate of recombination frequency can be obtained either by increasing the population size or the level of marker saturation.

In the case of the *LOD* score as a factor influencing mapping, it is known that the *LOD* score is a function of sample size (n) and recombination frequency (R). The *LOD* score for recombinant inbred lines is given by

$$LOD = \log_{10}\left\{\left[0.5(1-R)\right]^{n_1+n_2}(0.5R)^{n_3+n_4}(0.25)^{-n}\right\}$$

(Schuster and Cruz, 2004) in which n_1 and n_2 are the number of individuals derived from non-recombinant gametes for a given pair of loci, n_3 and n_4 are the number of individuals derived from recombinant gametes for a given pair of loci and R is the recombination frequency in recombinant inbred lines. By replacing the values for population size and recombination frequencies in the expression above and considering $n_1 = n_2 = 0.5n(1\text{-}R)$ and $n_3 = n_4 = 0.5nR$ it follows that when R is fixed the *LOD* score values increase with n, *e.g.* if the recombination frequency in the corresponding $F_2(r)$ equals 0.05 for a 5 cM genome then for a population size of n = 50 the *LOD* score is 8.43 while for n = 800 it is 134.98. Furthermore, when n is fixed a small r value results in a large *LOD* score, *i.e.* for the n = 50 population the *LOD* score was 8.43 for r = 0.05, 5.26 for r = 0.1 and 2.06 for r = 0.20. Thus, the larger the number of individuals genotyped for mapping, the larger the *LOD* score minimum value used to infer about linkage between markers, providing more reliable linkage maps.

The *LOD* score is limited to 2.06 for a population of n = 50 with r = 0.20, so if a minimum *LOD* score is selected which is greater than the value imposed by the size of the population then markers that should be linked will be inferred to be unlinked and the number of linkage groups will be increased. For all population sizes our evaluation used a minimum *LOD* score of 3 to infer linkage between markers, and since this value was greater than the 2.06 limiting *LOD* score for a population size of n = 50 for a 20 cM genome this explains why some markers that should be linked were declared as unlinked for this population and genome saturation. The *LOD* score is limited to 8.43 for a 5 cM genome with a population size of n = 50 but in our study the minimum *LOD* score of 3 was smaller than the *LOD* score imposed by this population size, indicating that the *LOD* score was not a reason why linked markers were declared as unlinked. In the 10 cM genome linked markers were declared as unlinked not due to the *LOD* score threshold, as for population sizes of n = 50 and n = 100, the *LOD* scores were greater than the threshold of 3, *LOD* = 5.26 and *LOD* = 10.53, respectively.

By increasing population size or using populations with high linkage information it is possible to increase the probability for coupling genome segments as well as to increase genome coverage (Liu, 1998). In our study, considering that the original genome saturation levels of 5 and 10 cM were satisfactory, a low number of recombinants in small populations might be an explanation for the establishment of a higher number of linkage groups than expected. We observed changes in the order of markers within linkage groups not only in the 5 cM genome n = 50, n = 100 and n = 154 populations but also in the 10 cM genome n = 100 and n = 154 populations and in the 20 cM genome n = 200 population. Since these changes were more frequent in small populations the use of such populations can result in serious problems related to marker positioning in linkage groups and consequently generate false results in QTL mapping. The inversion of markers observed in this study has also been described by Liu (1998) as a function of population size and map saturation level.

It is known that a better sampling of gametes is achieved as population size increases, leading to a better estimate of recombination frequency. In our study, the size of linkage groups as well as the between-marker distances became closer to the true value of the simulated genome as population size was increased. Reduction in the variation (*i.e.* the standard deviation) of linkage group sizes between replications as population size increased indicated that better estimates of recombination frequency were achieved using large populations. It is possible to compare different genome saturation levels for a given population size using the mean deviation ($\bar{d}$). For example, the percentage stress and $\bar{d}$ values for the n = 200 population were 25.1% and $\bar{d}$ = 1.25 cM for the 5 cM genome population, 19.24% and $\bar{d}$ = 1.92 cM for the 10 cM genome population and 15.02% and $\bar{d}$ = 3 cM for the 20 cM genome population, the stress values being shown in the last column of Table 1. Thus, although the percentage stress value was larger for the more marker-saturated genome the mean deviation value was smaller. Within the same genome saturation level the stress values decreased as population size increased, *e.g.* the stress and $\bar{d}$ values were 49.82% and 2.49 cM for the n = 50 population but 12.71% and 0.63 cM for the n = 800 population. Depending on the objectives of a particular study the determination of the population size and number of markers to be used for mapping could be achieved by analyzing the magnitude of the mean deviation.

Accurate ordering of markers and a high-resolution linkage map are not always necessary for some applications

in genome mapping. However, high levels of accuracy are needed for QTL location when it is to serve as a basis for positional gene cloning (Van Ooijen, 1992; Liu, 1998). In this case, highly accurate QTL location estimates with a resolution of between 1 and 2 cM are needed for the application of physical mapping and QTL cloning procedures (Darvasi *et al.*, 1993) and thus fine mapping techniques are necessary for obtaining better resolution. For plant breeding, however, high accuracy of distance estimates might not be so restrictive, since processes based on marker-assisted selection can be successful if the information about markers flanking a given QLT is available and the effect of the QLT can be easily detected (Van Ooijen, 1992).

A recombinant inbred line is an suitable population for estimating recombination frequency, especially when distances between markers are relatively short. On the other hand, gene linkage above 20 cM is not frequently detected in recombinant inbred lines because of the high recombination frequency in this type of population, as already described by Burr *et al.* (1988) and confirmed in our study.

Questions concerning the size of a recombinant inbred line population and the number of markers needed to represent chromosomes in linkage groups have been addressed previously (Ferreira *et al.*, 2006) but not with such an extensive genome as used by us. It is widely accepted that population size and the number of markers used in a study are frequently defined based on the availability of funds and genetic material. The application of our results will allow breeders to define the population size and the number of markers needed for the mapping of recombinant inbred lines. By analyzing 2100 maps obtained from simulated populations we concluded that: population size and number of markers are essential factors to be considered for obtaining reliable maps; maps with severe distortions were obtained with the use of small populations even using large number of markers; maps with severe distortions were obtained with the use of a small number of markers even using large populations; the minimum population sizes necessary for obtaining maps with the same number of markers per linkage group of simulated genomes were $n = 100$ for the 5 cM population, $n = 154$ for the 10 cM population and $n = 500$ for the 20 cM population. Thus by increasing the saturation levels it is possible to substantially reduce the number of individuals to be genotyped. Alternatively, by genotyping a large number of individuals it is possible to reduce number of markers and still achieve a reliable map. Reasonable sizes of recombinant inbred lines necessary for obtaining reliable maps are $n = 200$ for a genome saturation level of 5 cM, $n = 300$ for a genome saturation level of 10 cM and $n = 500$ for a genome saturation level of 20 cM.

Acknowledgments

The authors wish to thank the Brazilian agencies CAPES, CNPq and FAPEMIG for financial support and the anonymous reviewers for their invaluable suggestions for improving this manuscript.

References

Burnham KD, Dorrance AE, Vantoai TT and Martin SKSt (2003) Quantitative trait loci for partial resistance to *Phytophtora sojae* in soybean. Crop Sci 43:1610-1617.

Burr B and Burr F (1991) Recombinant inbreds for molecular mapping in maize: Theoretical and practical considerations. Trends Genet 7:55-60.

Burr B, Burr FA, Thompson KH, Albertson M and Stuber CW (1988) Gene mapping with recombinant inbreds in maize. Genetics 118:519-526.

Cardinal AJ, Lee M, Sharopora N, Woodman-Clikeman WL and Long MJ (2001) Genetic mapping and analysis of quantitative trait loci for resistance to stalk tunneling by the European corn bores in maize. Crop Sci 41:835-845.

Darvasi A, Weinreb A, Minke V, Weller JI and Soller M (1993) Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. Genetics 134:943-951.

Deorge R (1993) Constructing genetics maps by rapid chain delineation. J Quant Trait Loci 2:121-132.

Faleiro FG, Ragagnin VA, Schuster I, Corrêa RX, Good-God PI, Brommonshenkel SH, Moreira MA and Barros EG (2003) Mapeamento de genes de resistência do feijoeiro à ferrugem, antracnose e mancha-angular usando marcadores RAPD. Fitopatol Bras 28:59-66.

Falk CT (1989) A simple scheme for preliminary ordering of multiple loci: Application to 45 CF families. In: Elston, Spence, Hodge and Cluer (eds) Multipoint Mapping and Linkage Based Upon Affected Pedigree Members (Genetic Workshop 6). Liss, New York, pp 17-22.

Ferreira A, da Silva MF, Silva LC and Cruz CD (2006) Estimating the effects of population size and type on the accuracy of genetic maps. Genet Mol Biol 29:187-192.

Ferreira AR, Foutz KR and Keim P (2000) Soybean genetic map of RAPD markers assigned to an existing scaffold RFLP map. Am Genet Assoc 91:392-396.

Kruskal JB (1964) Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. Psychometrika 29:1-27.

Lander ES and Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185-199.

Liu BH (1998) Statistical genomics, linkage, mapping and QTL analysis. CRC Press, Boca Raton, 854 pp.

Miklas PN, Johnson WC, Delorme R and Gepts P (2001) QTL conditioning physiological resistance and avoidance to white mold in dry bean. Crop Sci 41:309-315.

Schuster I and Cruz CD (2004) Estatística Genômica Aplicada a Populações Derivadas de Cruzamentos. Editora da Universidade Federal de Viçosa, Viçosa, 568 pp.

Van Ooijen JW (1992) Accuracy of mapping quantitative trait loci in autogamous species. Theor Appl Genet 84:803-811.

Xing YZ, Tan YF, Hua JP, Sun XL and Xu CG (2002) Characterization of the main effects, epistatic effects and their environmental interactions of QTLs on the genetic basis of yield traits in rice. Theor Appl Genet 105:248-257.

Zheng BS, Yang L, Zhang WP, Mao CZ, Wu YR, Yi KK, Liu FY and Wu P (2003) Mapping QTLs and candidate genes for rice root traits under different watersupply conditions and comparative analysis across three populations. Theor Appl Genet 107:1505-1515.

Internet Resource

GQMOL Software, http://www.ufv.br/dbg/gqmol/gqmol.htm.

Associate Editor: Márcio de Castro Silva Filho