Simulation Experiments on Efficiencies of Gene Introgression by Backcrossing

J.-M. Ribaut,* C. Jiang, and D. Hoisington

ABSTRACT

Designing a highly efficient backcross (BC) marker-assisted selection (MAS) experiment is not a straightforward exercise, efficiency being defined here as the ratio between the resources that need to be invested at each generation and the number of generations required to achieve the selection. This paper presents results of simulations conducted for different strategies, using the maize genome as a model, to compare allelic introgression with DNA markers through BCs. Simulation results indicate that the selection response in the BC_1 could be increased significantly when the selectable population size (N_{sl}) is <50, and that a diminished return is observed when this number >100. Selectable population size is defined as the number of individuals with favorable alleles at the target loci from which selection with markers can be carried out on the rest of the genome at nontarget loci, simulations considered the allelic introgression at one to five target loci, with different population sizes, changes in the recombination frequency between target loci and flanking markers, and different numbers of genotypes selected at each generation. For an introgression at one target locus in a partial line conversion, and using MAS at nontarget loci only at one generation, a selection at BC_3 would be more efficient than a selection at BC_1 or BC_2, due to the increase over generations of the ratio of the standard deviation to the mean of the donor genome contribution. With selection only for the presence of a donor allele at one locus in BC_1 and BC_2, and MAS at BC_3, lines with $<5\%$ of the donor genome can be obtained with a N_{sl} of 10 in BC_1 and BC_2, and 100 in BC_3. These results are critical in the application of molecular markers to introgress elite alleles as part of plant improvement programs.

In a BC scheme, the strategy is to transfer a specific elite allele at a target locus from a donor line to a recipient line. The use of DNA markers, which permit the genetic dissection of the progeny at each generation, increases the speed of the selection process (Tanksley et al., 1989). Although it is easy to plot a BC-MAS strategy, the design of the most appropriate and efficient strategy is generally not a straightforward task, given the number of parameters involved. Before any experiment, the number of target genes involved in the selection and the expected level of line conversion must be defined. Then, one must identify at each generation the size of the population to be screened, the number, position, and nature of molecular markers used, and the number of genotypes selected. The expected level of conversion is closely related to the number and distribution of the DNA markers at nontarget loci and the recombination frequencies between the target gene and flanking markers. All these parameters influence the number of the generations required to achieve a specific and successful BC-MAS experiment, while offering different alternatives for defining a strategy. The screened population size has been reported in simulation studies as the most important factor affecting the efficiency of MAS (Zhang and Smith, 1992; Gimelfarb and Lande, 1994; Whittaker et al., 1995; Hospital et al., 1997; Frisch et al., 1999a). Independent of the strategies considered in those papers (selection index and BC-MAS), selection for individuals with a desirable genotype at the predetermined markers usually requires a relatively large population. Theoretically, if one considers an infinite population size, any BC-MAS experiment can be achieved in three generations (one generation to cross the two parental lines and produce the F_1 seeds, one BC, and one self-pollination generation). With recent advances in polymerase chain reaction (PCR) based markers, for example, simple sequence repeats (Chin et al., 1996) and single nucleotide polymorphisms (SNPs) (Gilles et al., 1999), a substantial improvement in the capacity to efficiently screen large populations has been achieved. Today, the screening of thousands of genotypes for a few target genes no longer poses an intractable technical problem, and can be considered in MAS strategies (Ribaut and Betrán, 1999).

During the past several years, simulations to evaluate the efficiency of MAS as a breeding tool have been reported by various groups. These simulations have been quite diverse; for example, MAS has been tested combining phenotypic and genotypic data in a selection index (Lande and Thompson, 1990; Knapp, 1994; Xie and Xu, 1998a), considering different breeding generations (Edwards and Page, 1994), and for different breeding schemes (Xie and Xu, 1998b). Efficiency of MAS has also been evaluated considering the heritability of the target trait (Hospital et al., 1997; Knapp, 1998), the genetic effect at target loci (Van Berloo and Stam, 1998), and by monitoring target genomic regions simultaneously vs. one by one (Hospital and Charcosset, 1997). Most of the theoretical papers related to MAS present complex mathematical models, making it difficult to directly derive a practical MAS experiment. In addition, the implications of using different laboratory strategies that can be considered to achieve the selection, such as different DNA markers, are rarely taken into account in those theoretical papers when comparing the efficiency of different approaches.

The objective of this paper is not to present new genetic models, but rather to provide some guidelines at both the theoretical and practical levels for identifying the most appropriate BC-MAS strategy based on the objectives of different types of applied breeding experiments. To achieve this objective, the relationship be-

J.-M. Ribaut and D. Hoisington, CIMMYT, Int. Maize and Wheat Improvement Center, Lisboa 27, Apdo. Postal 6-641, 06600 Mexico D.F., Mexico; and Changjian Jiang, Monsanto Life Sciences Research Center, 700 Chesterfield Parkway North, St. Louis, MO 63198. Received 8 Dec. 2000. *Corresponding author (j.ribaut@cgiar.org).

Published in Crop Sci. 42:557–565 (2002).

Abbreviations: BC, backcross; cM, centimorgan; MAS, marker-assisted selection; PCR, polymerase chain reaction; N, screened population size; N_i, number of individuals; N_{sl}, selectable population size; SNP, single nucleotide polymorphism.

tween the size of a segregating population that is screened at each generation and the number of BC generations required to achieve the selection was evaluated through simulations that considered different selection models for complete and partial line conversion.

METHODS

Backcrossing Schemes and the Selectable Population Size

Backcross schemes are as in Hospital et al. (1992). F_1 individuals from a cross between a donor and a recipient line are crossed with the recipient to derive a BC_1 population. In each subsequent BC generation, individuals with the desired allele at the target locus and suitable genotypic composition in the rest of the genome are selected and crossed with the recipient to develop the next generation. For the simulations presented in this paper, we consider introgression of an allele at the target locus that is uniquely identified by a genetic marker.

The objectives of a BC-MAS strategy are to identify individuals heterozygous for donor and recipient alleles at target loci and homozygous for recipient alleles at nontarget loci. Given these objectives, each generation can be divided into two steps. The first step is to identify the genotypes that are heterozygous at the target loci, reducing the screened population size (N) to the N_{sl}. The second step is to identify within the N_{sl} individuals those presenting the most suitable genomic composition at the nontarget loci. Assuming no linkage between target genes, the expected N_{se} can be obtained as $N_{se} = (1/2)^t N$, where t denotes the number of target genes. For $N_{sl} = 1$, the minimal N_{sl}, no selection pressure can be applied at nontarget loci.

Simulation Experiments

Factors considered in the simulations included N, the number of target genes (one, three, and five), the distance between the flanking markers and the target gene (2–20 cM), and the number of genotypes selected in each BC generation (one to eight) used to generate the next screened population. Target genes were assigned randomly to one marker on a chromosome and no more than one gene per chromosome was considered.

Marker-locus genotypes of progeny individuals were simulated based on marker-locus genotypes of parents and rules of Mendelian segregation. Genotypes were simulated as strings of 1 (heterozygous for donor and recipient alleles) or 0 (homozygous for recipient allele). An F_1 diploid individual consists of a string of 1's and another string of 0's, and only one string was regenerated for each individual in each BC generation since gametes from the recurrent parent were all the same. Haplotypes were simulated by "random walking" (all randomness being simulated by the computer's pseudorandom number generator) along the marker linkage map. The string for an individual began with the same bit by equal chance as one of two strings in the individual selected in the previous population and crossed over to read from the other string if a random number of uniform (0, 1) exceeded the specified recombination probability. Such practices are found in Tanksley and Nelson (1996).

A genome size of 10 chromosomes, each 200 cM in length, was chosen to approximate the genome of maize. All markers were assumed to be evenly distributed with an interval size of 20 cM (11 markers per chromosome), except for the two markers flanking the target genes. Two criteria were used to compare BC-MAS strategies: the number of generations necessary to obtain at least one desired individual, and the proportion of donor genome present after a fixed number of BCs. All simulation results presented in this study represent the average of 1000 repeats for each case in order to look at the variability and the distribution of the results. When we analytically predict the selection advance, we are pointed to the tail of the distribution that is most affected if the normality is violated. The normality assumption could be satisfied fairly well in BC_1 since all loci are under segregation. But the approximation becomes poor in BC_2 and BC_3 when most loci are fixed and only a small proportion of loci are still segregating. For the simulation, we did not assume normality, and the genetic model closely resembles the practical situation except for the recombination interference. The effect of the interference on the allele introgression would make the elimination of nontarget alleles easier but the target allele more difficult. Thus, the overall effect of the interference on gene introgression would be neutral. The results of the calculation and simulation are very similar, which make the simulation results more reliable.

Complete Line Conversion

The objective of a complete line conversion is to develop a line that will have exactly the same genetic composition as the recipient line, except at target loci where the presence of homozygous alleles from a donor line is desired. By definition, such conversion requires strong selection pressure at nontarget regions linked to the target gene, due to the *genetic drag* generated by the presence of the donor allele at the target loci (Tanksley et al., 1989). Genetic drag is lowest in genotypes with homozygous recipient alleles at the two markers flanking the target gene. Because selection requires identification of recombinations between the target gene and flanking markers, the two flanking markers for each gene involved in the model must be carefully identified. A recombination rate between target gene and flanking markers of 2 to 20 cM was employed, depending on the requirement of the conversion.

We used a selection index based on the probability that an individual generates progeny with the desirable gametic type. The desirable gametic type is defined to have recipient alleles at all marker loci except the target gene. Individuals with the highest probability of giving rise to offspring with the desired genotype, the one presenting recombination in the flanking intervals of the target gene, were considered to be the most desirable recombination. For example, assume that there are two individuals with the marker genotype on the carrier chromosome (markers on other parts of the genome can be considered in the same way) as M_1 m_2 m_3 M_4 T_5 M_6 and m_1 m_2 M_3 M_4 T_5 M_6; M and m represent alleles from donor and recipient lines, respectively, and T is the target gene. The recombination is needed in each of the two flanking intervals of T in the desirable gametes. While conditional on these two markers, segregation of other markers is independent of T and can be treated the same as markers on the noncarrier chromosome. Therefore, the probability that either of two individuals would be selected based on the number of heterozygous markers is equal. However, the probability of the gamete with all markers being m except T generated by two individuals is $1/2(1 - r_{1+2+3})r_4r_5$ for Individual 1 (since m_2 and m_3 are fixed already, only r_{1+2+3} is relevant here), and $1/2(1 - r_3)r_4r_5$ for Individual 2, where r is the recombination fraction in the corresponding interval and r_{1+2+3} represents the recombination between Marker 1 and Marker 4. The probability of a desired gamete is higher for Individual 2 than for Individual 1. The extension of the algorithm to the whole genome is straightforward, since segregation of markers from different chromosomes is independent. Therefore, the probability was calculated for each chromosome and multiplied over chromosomes. In the following simulation, a logarithm of the probability is used, which changes the multiplication to simple summation. Simulations were performed to investigate the effect of the population size, the size of the marker intervals flanking the target gene, and the number of target genes simultaneously introgressed.

Partial Line Conversion

Partial line conversion means that the conversion is complete when a limited proportion of the donor genome in an individual is found scattered over the genome in addition to the desirable homozygous alleles at the target gene. The selection index in this case is based on the estimated proportion of recipient genome. This is similar to phenotypic selection for a quantitative trait, the method used for the selection index proposed by Hospital et al. (1992). In this case, the preference for individuals with recombination in the flanking regions of the target gene is not necessary because the criterion is the total proportion of the recipient genome.

We considered marker selection on nontarget loci at only one generation while the desired allele at the target locus was selected in all generations. Let μ_t denote the mean of the proportion of the donor genome of individuals in Generation *t*, and let s_t denote the mean in the selected individuals in Generation *t* ($t = 1, 2, 3$). Based on classical selection theory, assuming normality and high marker density which provides the so-called heritability h^2 a value close to 1 (as Visscher, 1996) showed that all the variance of the genetic composition can be explained by placing three or more markers per chromosome),

$$s_t = \mu_t - i\sigma_t = \mu_t(1 - i\sigma_t/\mu_t),$$

where *i* denotes the selection differential, and σ_t the standard deviation among individuals. It would be safe to assume that the relative reduction in μ_{t+1} from s_t due to one more generation of backcrossing depends mainly on the value of s_t and has a minor effect from the genomic composition of the selected individual. That is, the mean in the next generation can be approximated as $s_t\mu_{t+1}/\mu_t = \mu_{t+1}(1 - \mathrm{i}\sigma_t/\mu_t)$. Then, the response on the mean of BC_{t2} due to the selection in BC_{t1}, $t2 \geq t1$, can be approximated as

$$s_{t2|t1} = \mu_{t2}(1 - \mathrm{i}\sigma_{ti}/\mu_{t1}).$$

Therefore, the efficiency of the selections in a generation will depend on the ratio of σ_t:μ_t. We evaluate the mean and variance for BC_1, BC_2, and BC_3 using the formulas of Stam and Zeven (1981). Simulations were also performed to compare the efficiency of the selection schemes in different generations. The simulation results were compared with the above analytical results.

RESULTS

Effect of Population Size on Selection Response

The N_{sl} decreases exponentially from *N*, screened at each generation as the number of target genes increases for both complete and partial line conversion. For a single target gene, the ratio between N_{sl} and *N* is 1:2, for five genes, 1:32. The N_{sl} is directly related to the selection pressure that can be applied to reduce the donor genome contribution at nontarget loci. Considering the number of individuals (N_i = 1, 2, 4, or 8) selected, simulation results of the selection intensity as an effect of N_{sl} are presented in Fig. 1A. The relationship between N_{sl} and the selection response is nonlinear. As expected, the selection intensity increases (i.e., less donor genome contribution at nontarget loci) with an increase in N_{sl}. However, the return in response to the increase of the population size is diminished significantly when N_{sl} > ≈100.

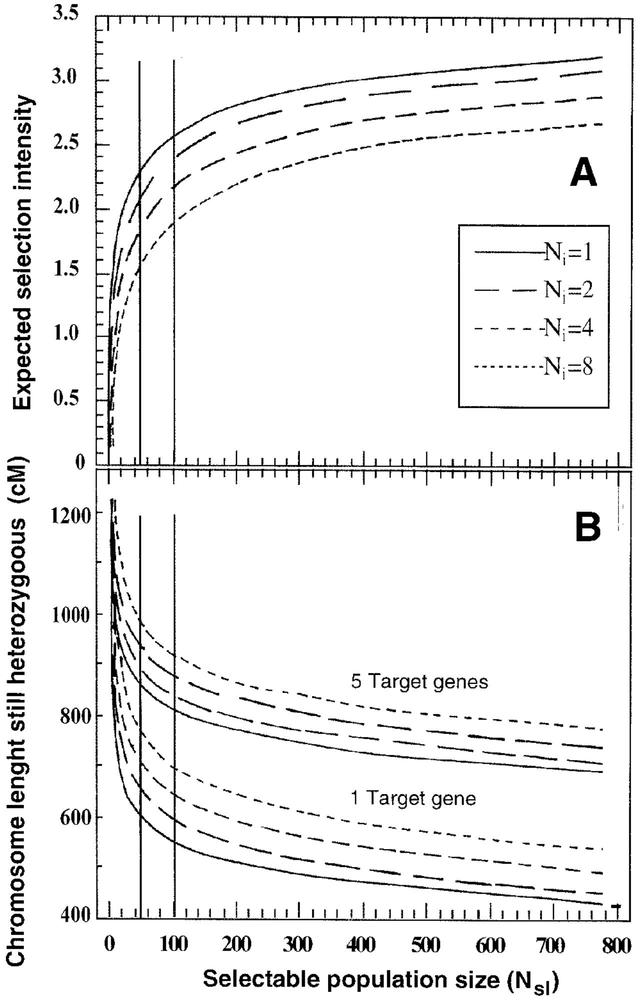

Fig. 1. Expected selection intensity as a function of population size with one, two, four, and eight individuals selected per generation, assuming an underlying normal distribution of donor genome contribution among screened genotypes. (A) Donor genome contribution at nontarget loci after BC_1 for different values of selectable population size (N_{sl}) and for the transfer of one or five target genes. (B) For this calculation, 10 chromosomes of 200 centimorgans (cM) (total genome size of 2000 cM) were considered with a DNA marker each 10 cM. The range of the sample size between vertical lines represents the most cost effective sample size.

The impact of different N_{sl}s on the genome composition at the first BC generation at nontarget loci is presented in Fig. 1B, which considers the transfer of one or five target genes with one target gene on one chromosome. The shape of the response curves is similar, suggesting that the appropriate N_{sl} is essentially independent of the number of target genes. Note that we consider the selection against the donor segments on the chromosomes with and without the target gene equivalently. However, the variation of the donor ge-

Table 1. Number of backcross generations (excluding the first cross necessary to produce the F_1 seeds) required for complete line conversion, that is, individual(s) with donor alleles only at the target gene(s) and recipient alleles at all nontarget loci every 20 centimorgans (cM) except for the flanking markers. Simulations considered one, three, or five target genes, different screened population sizes (N) corresponding to different selectable population size (N_{sl}), different recombination frequencies between a target locus and the flanking markers (2, 4, 8, 12, or 20 cM), and the selection of one or two individuals (N_i) at each cycle.

		Recombination frequencies between a target locus and the flanking markers									
		2 cM		4 cM		8 cM		12 cM		20 cM	
N	N_{sl}	$N_i = 1$	$N_i = 2$	$N_i = 1$	$N_i = 2$	$N_i = 1$	$N_i = 2$	$N_i = 1$	$N_i = 2$	$N_i = 1$	$N_i = 2$
						1 gene					
50	25	5.2	6.0	4.5	5.0	4.1	4.5	3.9	4.2	3.8	4.1
100	50	4.3	4.8	4.1	4.2	3.8	4.0	3.6	3.9	3.5	3.8
200	100	3.9	4.1	3.6	4.0	3.3	3.8	3.2	3.6	3.1	3.3
400	200	3.6	3.9	3.3	3.6	3.1	3.2	3.0	3.1	3.0	3.0
800	400	3.2	3.5	3.1	3.2	3.0	3.0	3.0	3.0	3.0	3.0
						3 genes					
200	25	9.0	10.4	6.4	7.5	5.3	5.8	4.8	5.2	4.3	4.7
400	50	6.7	7.6	5.5	6.0	4.7	5.0	4.3	4.7	4.0	4.2
800	100	5.8	6.3	5.0	5.3	4.3	4.7	4.0	4.1	3.7	4.0
1 600	200	5.2	5.5	4.5	4.8	4.0	4.1	3.8	4.0	3.4	3.7
3 200	400	4.8	5.0	4.1	4.3	3.9	4.0	3.5	3.8	3.1	3.2
						5 genes					
800	25	12.4	13.5	8.4	9.7	6.6	7.3	5.7	6.3	4.8	5.3
1 600	50	9.4	10.2	7.1	7.7	5.7	6.2	5.1	5.5	4.4	4.8
3 200	100	7.9	8.4	6.3	6.7	5.2	5.4	4.6	4.9	4.0	4.3
6 400	200	7.1	7.3	5.7	6.0	4.7	5.0	4.2	4.5	3.9	4.0
12 800	400	6.3	6.5	5.2	5.5	4.4	4.7	3.8	4.2	3.8	3.9

nome contribution in an individual from a BC_1 population is mostly from the noncarrier chromosomes because no selection for the presence of donor allele is conducted on those chromosomes, which results in the similar shape of the selection intensity curves as observed in Fig. 1A. The effect of high pressure against the genetic drag on the carrier chromosomes will be discussed in the line conversion experiment.

We defined the selection efficiency as the optimal ratio between the resources that have to be invested and the number of selection cycles required to achieve a complete selection, based on results presented in Fig. 1, the most efficient selection scheme should consider the screening of an initial population size that will result, after selection at target loci, in a N_{sl} ranging between 50 and 100 genotypes. Below these values, changes in N_{sl} still have a major impact on the number of selection generations, while above these values, changes in N_{sl} implies more resources for reduced impact on the selection process. For $N_{sl} = 100$, ≈200, 800, and 3200 individuals must be screened for one, three, and five target genes, respectively.

Complete Line Conversion

It is impossible to obtain complete line conversion, that is, the presence of only the homozygous donor alleles at the target gene. Therefore, line conversion is considered complete when, out of the selectable population, a genotype homozygous for recipient alleles at all detected nontarget loci can be identified. It implies that recombination should take place on both sides of the target gene(s), and the donor genome's contribution in the final line would be mostly around the target gene. Assuming no double recombination between two nontarget loci, and a 20-cM distance between the flanking markers and the target gene, the expected line conversion level (proportion of the genome from recipient parent) is 99% for one target gene, and 95% for five genes. For a 2-cM distance, this probability is 99.9 and 99.5% for one and five target genes, respectively. Table 1 presents the number of BC generations required to achieve the line conversion, based on simulation results. The simulations considered changes in the recombination frequency between a target gene and flanking markers, one to five target genes, different screened population sizes, and the selection of one or two individuals at each selection generation. Using these results and a manageable sample size (e.g., a selectable population of 50 to 100), the introgression can be completed in three to four generations for a single target gene, four to seven generations for three target genes, and four to nine generations for five genes, depending on the distance of flanking markers to the target gene. Dramatic effects are seen on the number of generations when $N_{sl} < 50$, and less so when $N_{sl} > 100$, independent of the number of target genes, and conforms to the results in Fig. 1B.

The most efficient selection response from a N_{sl} between 50 and 100 appears in Fig. 1 and Table 1; however, one should also consider the interval sizes flanking the target genes when defining the selection scheme. When a small interval is present in the flanking regions of the target gene, recombination in the flanking intervals becomes a rare event. Therefore, increasing the population size is preferred when the interval sizes flanking the target gene(s) are small (e.g., <5 cM).

Except for the case where a small N_{sl} (<50 genotypes) is combined with a reduced recombination frequency between the target gene and flanking markers (2 and 4 cM), the number of generations for selecting two individuals ($N_i = 2$) in each generation is almost equivalent to the results of selecting one individual ($N_i = 1$), with a population half that size for most of the cases in Table 1. Therefore, the selection fraction, which is the ratio of the N_i selected vs. the N, is a major parameter in a MAS experiment.

Partial Line Conversion with a Single Generation of Marker-Assisted Selection

The objective of a partial line conversion is to identify a line with donor alleles at target genes and a proportion of donor genome below a desired level. Usually no restriction would be enforced for the donor genome contribution outside the target loci over the genome. Mean and variation of the genome size from the donor of an individual in BC_1, BC_2, and BC_3 populations, without selection at nontarget loci, were calculated separately for carrier and noncarrier chromosomes, and summed based on our 10-chromosome genome of 2000 cM. The ratio of the standard deviation to the mean was then calculated assuming one, three, and five target genes (Table 2). As the backcrossing continues, the ratio of the standard deviation to the mean of the donor genome contribution increases. This implies that the most efficient marker-assisted selection would be in later rather than early generations if only one generation of selection at nontargeted loci is applied. Without selection, the donor genome size in an individual decreases exponentially as the backcrossing proceeds, and most of the donor genome can be reduced through the BC process, especially on the noncarrier chromosomes. However, the difference between generations decreases with the number of genes included in the model. When more genes are involved in the selection model, fewer noncarrier chromosomes are involved; and as previously mentioned, the variation in the donor genome size is mostly from the noncarrier chromosomes. A BC-MAS strategy involving MAS at nontarget loci at a single BC generation induces a larger reduction of the donor genome contribution at nonselected loci compared with BC-MAS selection conducted at an earlier generation, when the allelic introgression is conducted at one or a few target genes rather than several genes.

Simulations were performed for the introgression of one target gene, and the results were compared among different selection schemes (Table 3). Five schemes were considered, and genome sizes from the donor at BC_1, BC_2, and BC_3 were calculated. A single generation of selection at nontarget loci was performed at BC_1, BC_2, or BC_3. A population without selection at nontarget loci and continuous selection in all three generations was used as a reference.

Simulations were performed using a N_{sl} of 2, 5, 10, and 100 genotypes per generation with selection at target loci only, and 100 genotypes for the complete MAS step (selection at both target and nontarget loci). To illustrate the practical implications of the different selection schemes presented in Table 3, two schemes are presented in detail in Fig. 2. In Fig. 2A, the complete MAS step is conducted at all three BC generations, while in Fig. 2B, the complete MAS step is conducted only at the third BC. The scheme that resulted in the least amount of donor genome (1.5% after three BCs) utilized a complete MAS step at each generation (Fig. 2A). Obtaining 1.5% required screening a population of 200

Table 2. Donor genome contribution [mean (μ_1) in centimorgans (cM) and variance (σ_1^2) in cM²] at the BC_1, BC_2 and BC_3 generations after selection at target loci only, and considering an allelic introgression at one, three, and five target loci. A genome of 2000 cM (one target gene on one carrier in the middle of the chromosome along with noncarrier chromosomes each of 200 cM) was used for the calculations.†

Target	BC_1			BC_2			BC_3		
genes	μ_1	σ_1^2	σ_1/μ_1	μ_1	σ_1^2	σ_1/μ_1	μ_1	σ_1^2	σ_1/μ_1
1	1043	359	0.018	556	239	0.028	305	126	0.038
3	1130	321	0.016	667	231	0.023	415	135	0.028
5	1216	284	0.014	778	223	0.019	525	145	0.023

† Let s_1 and s_2 define the position of the target gene on a carrier chromosome and assume $s_1 = s_2 = s/2 = 100$ cM. The means and variance for the carrier (μ_c and σ_c^2) and the noncarrier (μ_{nc} and σ_{nc}^2) chromosomes in the table were calculated using the following formulas, based on the results of Stam and Zeven (1981).

$$\mu_c^{(t)} = (\tfrac{1}{2})^t\left\{(s_1 + s_2) + \sum_{k=1}^{t}\binom{t}{k}\tfrac{1}{2k}(2 - e^{-2ks_1} - e^{-2ks_2})\right\}$$

$$\mu_{nc}^{(t)} = (\tfrac{1}{2})^t\, s$$

$$\begin{aligned}\sigma_c^{2(t)} = (\tfrac{1}{2})^{2t}\Bigg\{&\sum_{k=1}^{t}\binom{t}{k}\frac{s_1 + s_2}{k} - \sum_{k=1}^{t}\binom{t}{k}\tfrac{1}{k^2}(2 - e^{-2ks_1} - e^{-2ks_2}) + \sum_{k=1}^{t}\binom{t}{k}\tfrac{1}{k}(s_1e^{-2ks_1} + s_2e^{-2ks_2})\\ &+ \left[\sum_{k=1}^{t}\binom{t}{k}\tfrac{1}{k}\right]\left[\sum_{k=1}^{t}\binom{t}{k}\tfrac{1}{2k}(2 - e^{2ks_1} - e^{-2ks_2})\right] - \sum_{k=1}^{t}\binom{t}{k}^2\tfrac{1}{k}(s_1e^{-2ks_1} + s_2e^{-2ks_2})\\ &- \sum_{\substack{k_1,\, k_2=1\\ k_1\neq k_2}}^{t}\binom{t}{k_1}\binom{t}{k_2}\frac{1}{2k_2\,(k_1 - k_2)}(e^{-2k_2s_1} + e^{-2k_2s_2} - e^{-2k_1s_1} - e^{-2k_1ks_2})\\ &- \left[\sum_{k=1}^{t}\binom{t}{k}\tfrac{1}{2k}(2 - e^{-2ks_1} - e^{-2ks_2})\right]^2\Bigg\}\end{aligned}$$

$$\sigma_{nc}^{2(t)} = (\tfrac{1}{2})^{2t}\left\{\sum_{k=1}^{t}\binom{t}{k}\frac{s}{k} - \sum_{k=1}^{t}\binom{t}{k}\tfrac{1}{2k^2}(1 - e^{-2ks})\right\},$$

where t indicates the number of backcrossing.

Table 3. Estimation of the mean and variance of donor genome contribution when the marker-assisted selection (MAS) (N_{sl} = 100) is conducted only at the BC_1, BC_2 or BC_3, at all three backcross (BC) generations, or not at all. Simulations were conducted considering one target locus, and different selectable population sizes (N_{sl}) for the BC generation at which MAS was conducted only at target locus.

		Estimate donor genome size (SD)			
MAS	N_{sl}	BC_1	BC_2	BC_3	Final %
		cM			
None	2	1043 (134)	556 (109)	305 (79)	15.3
	5	1043 (85)	556 (69)	305 (50)	15.3
	10	1043 (60)	556 (49)	305 (35)	15.3
	100	1043 (19)	556 (15)	305 (11)	15.3
Only at BC_1	2	571 (73)	311 (65)	178 (48)	8.9
	5	571 (73)	308 (52)	178 (36)	8.9
	10	571 (73)	318 (46)	183 (31)	9.2
	100	604 (73)	330 (39)	188 (23)	9.4
Only at BC_2	2	1058 (137)	248 (66)	141 (44)	7.1
	5	1039 (89)	225 (49)	133 (32)	6.7
	10	1052 (53)	225 (46)	132 (28)	6.6
	100	1044 (20)	220 (48)	128 (26)	6.4
Only at BC_3	2	1057 (126)	571 (115)	112 (40)	5.6
	5	1045 (84)	544 (68)	95 (31)	4.8
	10	1058 (53)	566 (45)	88 (27)	4.4
	100	1044 (20)	560 (15)	86 (29)	4.3
At BC_1, BC_2 and BC_3	100	604 (73)	138 (44)	30 (16)	1.5

individuals at a single locus, followed by a screening of 109 markers (11 per chromosome) at nontarget loci at each BC on the N_{sl} (N_{sl} = 100). When the single complete MAS step is conducted only at the third BC (Fig. 2B), the remnant donor genome contribution at nontarget loci is 4.4%. In this strategy, the screening of only 20 plants at the target locus in BC_1 and BC_2 is required. When complete selection is conducted only in BC_3, the proportion of donor genome in the selected individual is ≈50 and 70% of that found in a complete selection conducted only in BC_1 and BC_2, respectively (Table 3), which conforms to the results presented in Table 2. Differences in the N_{sl}, when selection is conducted only at a target locus, have the largest impact on the donor genome proportion when the complete MAS is conducted at the most advanced BC. Nevertheless, these differences become small when the N_{sl} is increased, as there is almost no difference between N_{sl} = 10 or N_{sl} = 100, even if the complete selection is conducted at the third BC. One should also consider that the smaller the N_{sl}, the larger the variation of the donor genome in the individuals in BC_3. Therefore, a N_{sl} used to maintain the population would be 10, even if the single complete MAS is performed only at BC_3.

DISCUSSION

DNA markers allow us to identify the allelic composition of genotypes in a segregating population. As previously described by Tanksley (1989), the main advantages of using DNA markers vs. conventional selection is to accelerate the fixation of recipient alleles at nontarget regions and to identify the genotypes containing crossovers close to target genes (Ribaut and Hoisington, 1998). In the case of allelic introgression at specific loci, the identification of the best strategy is really a numbers game, considering the practical implications of numerical changes of the different parameters involved in the selection model. Because several parameters are involved in the efficiency of a BC-MAS experiment, the best strategy must be adopted at the beginning of the experiment, especially taking into account time constraints and available resources. While addressing the issue of costs is not easy, particularly in view of the technical options available for conducting a BC-MAS, the options presented below should help facilitate the cost evaluation of a given selection scheme.

From the Screened Population Size to the Selectable Population Size

The ratio between N and N_{sl} depends on the number of target loci. For more than three target loci in the model, the screening of thousands of plants is required to obtain a N_{sl} of 100 genotypes. Selection at each target

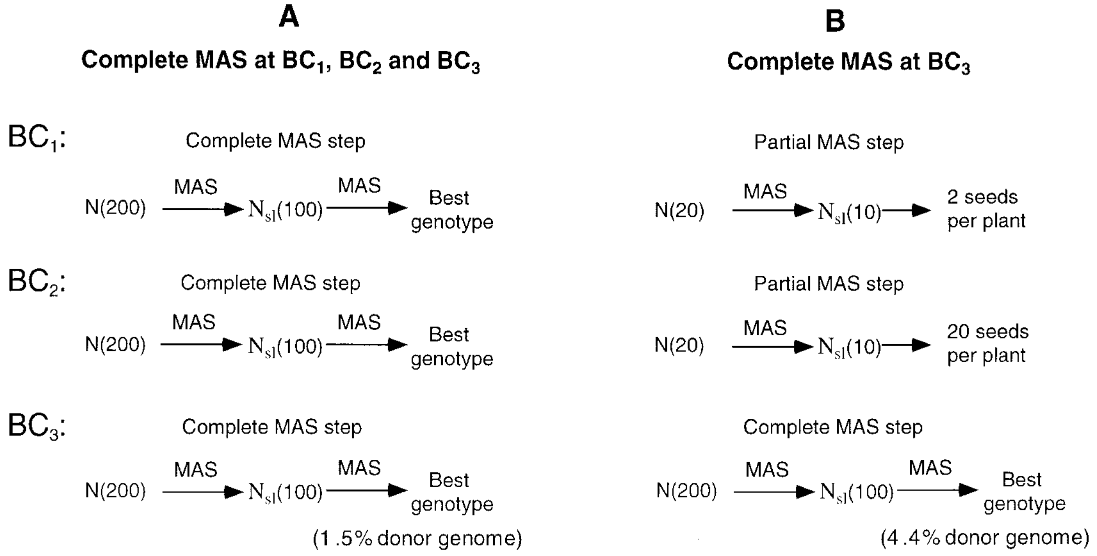

Fig. 2. Donor genome contribution for the introgression at one target locus when (A) a complete marker-assisted selection (MAS) step [selectable population size (N_{sl}) = 100] is conducted during each of the three backcross (BC) generations, and (B) partial MAS is conducted during the first two BC generations (N_{sl} = 10) and a complete MAS is conducted at the third BC generation (N_{sl} = 100).

locus reduces the N_{sl} by half. Nevertheless, the screening of the whole population has to be conducted at least once at the beginning of each BC generation. With N equal to thousands of individuals, such screening can be laborious and expensive. However, it can be optimized by using an appropriate combination of DNA markers. If markers can be amplified in the same reaction tube (Ribaut et al., 1997), a tremendous reduction in the number of PCR reactions required to conduct the selection can be achieved (e.g., in one step, duplexing reduces the population size by four, triplexing by eight). The PCR-based primers that amplify target genes could be distinguished in a single separation, because they amplify different fragment sizes. If this is not possible, other PCR-markers closely linked to the target genes might be used. Assuming the availability of fluorescent detection, the labeling of the different PCR-primers with different dyes allows direct multiplexing of the markers (Karp et al., 1997). Once the sequences of the donor and recipient alleles are known, the use of allele specific marker-like molecular beacons (Bonnet et al., 1999) and SNPs (Gilles et al., 1999) might be an efficient option. Indeed, the gel step can be eliminated by using this technique, and direct multiplexing can be obtained using different fluorescent dyes. Considering all these options, multiplexing in a BC-MAS should always be possible. Furthermore, in the context of the overall cost of an experiment, it is important to identify the most suitable set of markers at the target loci.

From the Selectable Population Size to the Selected Plants

Once the selectable population is identified, screening of N_{sl} genotypes with DNA markers should be conducted at nontarget loci in order to reduce the donor genome contribution. The selection response for this second selection step depends on the recombination frequency between the target gene and the flanking markers, and on the densities of the markers on the carrier and noncarrier chromosomes. In our simulation, we considered a fixed number of markers at each generation, therefore, the issue of different marker densities related to the BC generation is not addressed here. In regards to deciding how many unlinked nontarget loci should be screened and how they should be distributed, several strategies have already been advanced. The density of the marker coverage, for example, can be adapted to the inbreeding level of each BC generation. It has been shown that increasing the number of markers to more than three per noncarrier chromosome was not efficient at early generations (Hospital et al., 1992). At each new generation, due to additional crossover probability, an increase in the number of markers should be considered in order to optimize selection. This increase is balanced by the fact that markers that revealed fixed alleles at nontarget loci at one generation need not be screened at the next BC generation.

The Selected Genotypes at Each Generation

As presented, the N_i selected at each generation has a direct impact on the duration of the selection process. Although in BC-MAS the selection of a single individual is the fastest strategy in terms of generations required to achieve the selection, it is, nevertheless, risky from a practical point of view. A mistake at one of the selection steps or an unexpected field problem, such as low germination or poor pollen quality, will have dramatic consequences. Based on these practical considerations, the selection of more than one genotype at each generation should be considered. In practice, the number of individuals selected at each generation can be limited by the propagation ability of the studied crop, that is, the number of selected plants that are necessary to derive the suitable population size at the next selection generation. This limitation is important when several target genes are involved in the selection and the planting of thousands of plants is required. In this respect, maize, and more generally, cross-pollinated plants, offer an advantage. If the best genotype is selected before flowering, the pollen of only one selected plant is sufficient to develop the next large population, using several plants from the recipient line as females. This procedure is not general to all crops, and it may make the selection of only the best individual to create a large population at the next generation unrealistic. If the number of selected plants required at each generation to develop the next population is high, the optimal N_{sl} should be considered carefully. If this constraint is too great, other BC-MAS strategies may be considered.

Line Conversion

On the basis of several simulation studies, it is clear that BC-MAS is especially efficient when conducted on large segregating populations (Hospital et al., 1997). Nevertheless, the identification of a screened population size, which leads to the most efficient strategy for a line conversion through BC-MAS, has to consider different values for parameters that interactively influence the length of the selection process. Simulations have been widely used to evaluate and compare different strategies for the allelic introgression at single or multiple genes. Among a range of uses, simulations have been used to evaluate the optimal distribution of markers for carrier and noncarrier chromosomes (e.g., Hospital et al., 1992; Visscher, 1996), to optimize the position of the flanking markers (e.g., Frisch et al., 1999b), and to identify the minimum screened population size to obtain a given genomic composition in a given number of generations (e.g., Visscher et al., 1996; Frisch et al., 1999b). Moreover, several software programs, such as QU-GENE (Podlich and Cooper, 1998) and PLABSIM (Frisch et al., 2000), are now available to make selection predictions through simulations.

Obtaining a clear vision of the appropriate BC-MAS strategy is difficult because the implications of changing the values in the parameters involved in the selection are hard to project. As an example, it has been demonstrated in several studies that to minimize genetic drag around selected loci, emphasis should be placed on BC-MAS at an early stage of recombination on recombination events close to the target gene (Tanksley et al., 1989; Frisch et al., 1999a). If the strategy is clear, the choice of the most suitable markers to apply it must be considered carefully. Indeed, the distance between the target gene and the flanking markers has a major impact on the number of BC generations required to achieve the selection, especially when several target genes are considered. On the basis of our results, with five target genes and a selectable population of 100, having the flanking markers at 2 vs. 12 cM almost doubled the length of the complete line conversion (7.9 vs. 4.6 BC generations). In both cases, the level of conversion is different, 99.5 vs. 97% for 2 and 12 cM, respectively. Therefore, depending on the objective of the BC-MAS experiment, the position of the flanking markers can be quite different. In some cases, the most efficient strategy is less clear, especially when different theoretical approaches might serve the same purpose. For example, an increase of population size when advancing the BC generation reduces the number of required marker data points in comparison with a constant population size across all generations (Frisch et al., 1999a). The same data point reduction might be reached by increasing at each new BC generation the number of markers at nontarget loci while screening the same population size at each generation (Hospital et al., 1992). Different approaches might also be combined to increase the efficiency of the selection. Frisch proposed to identify through simulations the minimum population size that has to be screened to obtain at least one individual with a target genomic composition (Frisch et al., 1999b). This approach might be very relevant at an advanced generation of the strategy proposed in this paper. Indeed, at the end of the selection process, it is appropriate to calculate the N_{sl} that will allow the completion of the selection in one generation, thereby eliminating the need for an additional generation, even if $N_{sl} > 100$ in the last selection generation.

The nature of the germplasm considered in a BC-MAS experiment also has a major impact on the identification of the most suitable strategy. For example, the biological implication of having different levels of line conversion must be considered carefully. As already discussed, the distance between the flanking markers and the target gene has an impact on the final level of conversion, 99.5 vs. 97% for 2 and 12 cM, respectively. The biological implication on the plant phenotype of this 2.5% difference in donor genome contribution outside the target genes is difficult to predict and depends on the agronomic characteristics of the donor line (Lee, 1995). Once a target gene is introduced for the first time into an elite line, flanking markers at 2 cM should be the best option; while in the next phase, the transfer of the same target gene from elite into elite material, the use of flanking markers at 12 cM might be more effective.

Partial Line Conversion

Given the selection of only a few of the best genotypes at each generation, a single BC-MAS may be most efficient when conducted at advanced BC generations. After studying different selection schemes, Hospital et al. (1992) concluded that selection in later generations is better. The strategy of using one generation of selection in an advanced BC generation is an attractive option, especially if allelic introgression at a few target genes is considered concomitantly in a large number of recipient lines. The small population required for the first generations, in which selection is only conducted at target loci, represents a major logistical advantage. Moreover, if one target gene is linked to a phenotypic marker, or is a transgene (with a selectable gene such as herbicide resistance included in the gene construct), selection for this gene can be conducted phenotypically, reducing the cost of the selection. If this is the case, no DNA extraction is required to conduct the selection during the first generations. The "penalty" for this strategy is the retention of some donor genome contribution at nontarget loci, most of it flanking the target genes on the carrier chromosomes. Possible negative impacts from this remnant donor genome on plant performance can be minimized if the donor line is elite germplasm, because the probability of having bad agronomic characteristics dragged into the selection at nontarget loci is reduced.

CONCLUSIONS

The experimental design for line conversion through BC-MAS includes the available resources, the nature of the germplasm (e.g., agronomic quality and number of lines to be converted), and the technical options available at the marker level. Considering these parameters and the results provided through simulations for different theoretical approaches, the identification of the most efficient BC-MAS strategy for a practical experiment should be on a case-by-case basis. Several simulation results have already been reported, giving useful guidelines to identify optimal strategies. The strategies presented in this paper focus on the nonlinear relationship between a reduction of the donor genome contribution at nontarget loci for different N_{sl} and identify N_{sl} as the key parameter to be considered first in the establishment of the selection scheme. Our recommendation, once the number of target genes to be introgressed has been defined, is to determine the population size that needs to be screened at each generation, giving a target N_{sl} of 50 to 100 genotypes. Once the N_{sl} is defined, one should determine the desirable recombination frequency between the flanking markers and the target gene and the number of genotypes selected at each generation, based on the objective and the constraints of the experiment. The number of BC generations required to achieve the introgression can be easily predicted based on simulations (Table 1). When resources are limited, or introgression from a donor line into a large number of recipient lines is desired, strategies based on BC-MAS at nontarget loci solely at one advanced BC generation should be considered. Selection in later generations is more effective because the ratio of the standard deviation to the mean of the donor genome contribution increases as the backcrossing proceeds. In all cases, it is critical to put adequate effort into identifying the most convenient set of markers. From an applied breeding perspective, the analyses presented in this paper, as well as the practical points related to the nature of the germplasm and the use of the DNA markers, should help breeding programs identify the most suitable and efficient BC-MAS strategy for meeting their particular needs.

ACKNOWLEDGMENTS

The authors would like to thank Drs. M. Cooper, D. Grimanelli, and D. Podlich for helpful advice and suggestions, and D. Poland for his helpful editorial review of the manuscript.

REFERENCES

Bonnet, G., S. Tyagi, A. Libchaber, and F.R. Kramer. 1999. Thermodynamic basis of the enhanced specificity of structured DNA probes. Proc. Natl. Acad. Sci. USA 96:6171–6176.

Chin, E.C.L., M.L. Senior, H. Shu, and J.S.C. Smith. 1996. Maize simple repetitive DNA sequences: abundance and allele variation. Genome 39:866–873.

Edwards, M.D., and N.J. Page. 1994. Evaluation of marker-assisted selection through computer simulation. Theor. Appl. Genet. 88:376–382.

Frisch, M., M. Bohn, and A.E. Melchinger. 1999a. Comparison of selection strategies for marker-assisted backcrossing of a gene. Crop Sci. 39:1295–1301.

Frisch, M., M. Bohn, and A.E. Melchinger. 1999b. Minimum sample size and optimal positioning of flanking markers in marker-assisted backcrossing for transfer of a target gene. Crop Sci. 39:967–975.

Frisch, M., M. Bohn, and A.E. Melchinger. 2000. PLABSIM: Software for simulations of marker-assisted backcrossing. J. Heredity 91:86–87.

Gilles, P.N., D.J. Wu, C.B. Foster, P.J. Dillon, and S.J. Chanock. 1999. Single nucleotide polymorphic discrimination by an electronic dot blot assay on semiconductor microchips. Nat. Biotechnol. 17:365–370.

Gimelfarb, A., and R. Lande. 1994. Simulation of marker-assisted selection in hybrid populations. Genet. Res. (Cambridge) 63:39–47.

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

Hospital, F., C. Chevalet, and P. Mulsant. 1992. Using markers in gene introgression breeding programs. Genetics 132:1199–1210.

Hospital, F., L. Moreau, F. Lacoudre, A. Charcosset, and A. Gallais. 1997. More on the efficiency of marker-assisted selection. Theor. Appl. Genet. 95:1181–1189.

Karp, A., K.J. Edwards, M. Bruford, S. Funk, B. Vosman, M. Morgante, O. Seberg, A. Kremer, P. Boursot, P. Arctander, D. Tautz, and G.M. Hewitt. 1997. Molecular technologies for biodiversity evaluation: Opportunities and challenges. Nat. Biotechnol. 15:625–628.

Knapp, S.J. 1994. Selection using molecular marker indexes. p. 1–11. *In* Supplement, Joint Plant Breeding Symposia Series, American Society for Horticultural Science, Corvallis, OR. 5–6 August 1994. CSSA, Madison, WI.

Knapp, S.J. 1998. Marker-assisted selection as a strategy for increasing the probability of selecting superior genotypes. Crop. Sci. 38: 1164–1174.

Lande, R., and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124: 743–756.

Lee, M. 1995. DNA markers and plant breeding programs. Adv. Agron. 55:265–344.

Podlich, D.W., and M. Cooper. 1998. QU-GENE: a simulation platform for quantitative analysis of genetic models. Bioinformatics 14:632–653.

Ribaut, J.M., and J. Betrán. 1999. Single large-scale marker-assisted selection (SLS-MAS). Mol. Breed. 5:531–541.

Ribaut, J.M., and D. Hoisington. 1998. Marker-assisted selection: new tools and strategies. Trends Plant Sci. 3(6):236–239.

Ribaut, J.M., X. Hu, D. Hoisington, and D. Gonzlez-de-LeÛn. 1997. Use of STSs and SSRs as rapid reliable preselection tools in a marker-assisted selection-backcross scheme. Plant Mol. Biol. Rep. 15:154–162.

Stam, P., and A.C. Zeven. 1981. The theoretical proportion of the donor genome in near-isogenic lines of self-fertilizers bred by backcrossing. Euphytica 30:227–238.

Tanksley, S.D., and J.C. Nelson. 1996. Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. Theor. Appl. Genet. 92:191–203.

Tanksley, S.D., N.D. Young, A.H. Paterson, and M.W. Bonierbale. 1989. RFLP mapping in plant breeding: new tools for an old science. Biotechnology 7:257–264.

Van Berloo, R., and P. Stam. 1998. Marker-assisted selection in autogamous RIL populations: a simulation study. Theor. Appl. Genet. 96:147–154.

Visscher, P.M. 1996. Proportion of the variation in genetic composition in backcrossing programs explained by genetic markers. J. Hered. 87:136–138.

Visscher, P.M., C.S. Haley, and R. Thompson. 1996. Marker-assisted introgression in backcross breeding programs. Genetics 144:1923–1932.

Whittaker, J.C., R.N. Curnow, C.S. Haley, and R. Thompson. 1995. Using marker-maps in marker-assisted selection. Genet. Res. (Cambridge) 66:255–265.

Xie, C., and S. Xu. 1998a. Efficiency of multistage marker-assisted selection in the improvement of multiple quantitative traits. Heredity 80:489–498.

Xie, C., and S. Xu. 1998b. Strategies of marker-aided recurrent selection. Crop Sci. 38:1526–1535.

Zhang, W., and C. Smith. 1992. Computer simulation of marker-assisted selection utilizing linkage disequilibrium. Theor. Appl. Genet. 83: 813–820.