III.2 Breeding Strategies: Optimum Design of Marker-Assisted Backcross Programs

M. FRISCH[1]

1 Introduction

Recurrent backcrossing is used to transfer the genes underlying agronomically important traits from a donor into the genetic background of a recipient genotype, usually an inbred line (Allard 1960). In a backcross program, molecular markers can be used for indirect selection for the presence of a favorable allele (Tanksey 1983) and for selection against the undesired genetic background of the donor genotype (Tanksley et al. 1989). Selection against the genetic background of the donor ('background selection') allows us to reduce the number of backcross generations required for gene introgression from six to three (Frisch et al. 1999a). Due to this time saving and the possibility to monitor the donor genome content of the converted line, background selection has become a standard tool in plant breeding, as demonstrated by the example of the introgression of a gene coding for the *Bacillus thuringiensis* toxin into a maize inbred line (Ragot et al. 1995). However, the cost of a breeding program applying background selection is determined by an optimum allocation of resources, because the price of marker analyses is still high. In this chapter, principles for the optimum design of backcross programs for introgression of qualitatively inherited traits with marker-assisted background selection are described. Considered topics are (1) introgression of a dominant gene, (2) introgression of a recessive gene, and (3) simultaneous introgression of two genes.

2 Introgression of One Dominant Gene

2.1 Minimum Population Size Required for Finding Recombinant Plants

Theory (Stam and Zeven 1981) and experimental results (Young and Tanksley 1989) show that the intact donor chromosome segment around a target gene in backcrossing remains large, even in advanced backcross generations. It can

[1] Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany

Biotechnology in Agriculture and Forestry, Vol. 55
Molecular Marker Systems (ed. by H. Lörz and G. Wenzel)

form the major part of the carrier chromosome of the target gene in a backcross product, which is responsible for the transfer of undesired traits from the donor into the recipient parent (Zeven et al. 1983). By monitoring markers flanking the target locus and selecting individuals carrying the donor allele at the target locus and the recipient alleles at the flanking markers, the length of the intact donor chromosome segment around the target gene can be reduced efficiently (Tanksley et al. 1989). This rationale can be used to determine the population size in a backcross program such that recombinants between the target gene and flanking markers can be found with a high probability.

The population size required to generate in one backcross generation with a high probability at least one plant recombinant between the target gene and both flanking markers is greater than the multiplication rate of most crop species. For example, for a flanking marker distance of 5 cM on each side of the target gene, about 4000 individuals are required to find a double recombinant with a probability of 0.99; even for the rather large flanking marker distance of 25 cM at least 300 individuals are required (Frisch et al. 1999b). Therefore, we recommend a sequential strategy to find an individual with recombination between the target gene and one flanking marker in generation BC_1, and a recombinant between the target gene and the second flanking marker in generation BC_2 (Frisch et al. 1999b).

To calculate the minimum population size required in a backcross program applying this approach, we consider a chromosome on which positions are denoted in map distance from the telomere. The target locus is located at position x and two flanking markers at positions y_l and y_r such that $y_l < x < y_r$. Let $d_1 = x - y_l$ and $d_2 = y_r - x$ denote the lengths of the chromosome intervals between the target locus and its flanking markers. Without loss of generality, we assume $d_1 \leq d_2$. We denote z^- as the genotype of an individual homozygous for the recipient allele and z^+ as the genotype of a heterozygous individual at the locus at position $z \in \{y_l, x, y_r\}$.

If the probability that a plant has a desired genotype is p, then the minimum population size n required to find with probability q at least one plant, which has a desired genotype, can be obtained from the probability function of the binomial distribution as

$$n \geq \text{In } (1 - q)/\text{l}(1 - p). \tag{1}$$

Probabilities p of obtaining single or double recombinant plants between the target gene and the flanking markers are listed in Table 1. Probabilities p of obtaining plants, which are not only defined by conditions concerning the target gene and its flanking markers, but also by the condition that the complete chromosome region between a flanking marker and the nearest telomere consists entirely of the recurrent parent genome were given by Frisch et al. (1999b).

Applying these results, a simple method to carry out a two-generation backcross program designed to find with probability q_2 at least one BC_2 plant of genotype $y_l^- x^+ y_r^-$ can be conducted as follows:

Table 1. Transition probabilities p between genotypes in backcrossing

Genotype in generation BC_s	Genotype in generation BC_{s+1}	Transition probability
$y_l^+x^+y_r^+$ [a]	$y_l^-x^+y_r^-$	$p = (1 - e^{-2d_1})\,(1 - e^{-2d_2})\,/8$ [b]
$y_l^+x^+y_r^+$	$y_l^-x^+y_r^+$ or $y_l^+x^+y_r^-$	$p = (1 - e^{-2(d_1+d_2)})\,/4$
$y_l^-x^+y_r^+$	$y_l^-x^+y_r^-$	$p = (1 - e^{-2d_2})\,/4$
$y_l^+x^+y_r^-$	$y_l^-x^+y_r^-$	$p = (1 - e^{-2d_1})\,/4$

[a] The symbols y_l and y_r denote the background selection markers and x the target locus. A superscript + or − indicates that the locus is heterozygous or homozygous for the recurrent parent allele, respectively

[b] d_1 and d_2 denote the map distances between the target gene and two flanking markers

1. Choose the desired probability of success q_2. Set the probability of finding at least one BC_1 individual of type $y_l^-x^+y_r^+$ or $y_l^+x^+y_r^-$ to $q_1 = q_2$
2. Carry out BC_1 with population size n_1 such that at least one individual of genotype $y_l^-x^+y_r^+$ or $y_l^+x^+y_r^-$ is generated with probability q_1.
3. Select a BC_1 individual according to $(d_1 \leq d_2)$

   $$y_l^-x^+y_r^- \succ y_l^-x^+y_r^+ \succ y_l^+x^+y_r^- \succ y_l^+x^+y_r^+.$$

   (The symbol $\succ$ denotes that the genotype on the left-hand side is preferred over the genotype on the right-hand side.)
4. Carry out generation BC_2 with n_2 such that at least one individual with genotype $y_l^-x^+y_r^-$ is generated with probability q_2.

An optimization of this scheme is possible by choosing $q_1 \neq q_2$ in step 1. A certain probability of success q_2 can be reached irrespective of the chosen probability q_1 because the population size n_2 can be chosen in step 4 such that a desired level of q_2 is reached irrespective of the genotype of the plant selected in step 3. In consequence, the choice $q_1 = q_2$ is arbitrary, and an optimum criterion for q_1 can be defined such that q_1 is optimal if the expected total number of individuals required for the backcross program is minimized: $E(n) = n_1 + E(n_2) \rightarrow$ min. A mathematical description of this optimum criterion is given by Eq. (35) of Frisch et al. (1999b), a graphic illustration of the effect of q_1 on the expected total number of individuals required is shown in Fig. 1. For example, for two flanking markers 5 cM distant from the target gene ($d_1 = d_2 = 0.05$) and a desired probability $q_2 = 0.99$ of finding in generation BC_2 at least one double recombinant individual, choosing $q_1 = 0.90$ results in an expected total number of individuals required of $E(n) > 400$ (Fig. 1). In contrast, choosing $q_1 \approx 0.995$ results in a minimum of the expected total number of individuals of $E(n) = 222$.

Calculation of the optimum values q_1 for different flanking marker distances and probabilities of success is numerically demanding, therefore, we tabulated the population size n_1 corresponding to the optimum value of q_1 for flanking marker distances of 4, 6, 8, 12, and 16 cM in Table 2.

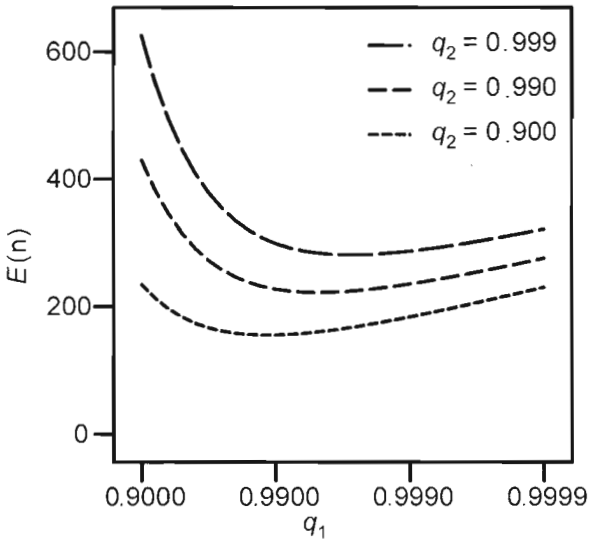

Fig. 1. Expected total number $E(n)$ of individuals required to reach probabilities of success q_2=0.900, 0.990 or 0.999 depending on the value chosen for q_1. The flanking marker distances are d_1=d_2=0.05 M

Table 2. Optimum population size n_1 in generation BC_1 and corresponding expected population size $E(n_2)$ in generation BC_2 such that the expected total number of individuals $E(n)$=n_1+$E(n_2)$ required to introgress one gene with a minimum number of individuals in a two-generation backcross program is minimized. The values depend on the map distances d_1 and d_2 between the target gene and two flanking markers

	d_2 (M)				
d_1 (M)	0.04	0.06	0.08	0.12	0.16
	n_1/E(n_2)				
0.04	143/252	136/186	130/155	123/128	117/117
0.06		91/167	88/135	83/105	79/93
0.08			66/125	63/94	60/80
0.12				43/83	41/68
0.16					32/62

In the above approach, the population size of generation BC_2 is determined after learning the result of generation BC_1 (a posteriori) such that the overall success is reached independent of the outcome of generation BC_1. Hospital and Charcosset (1997) presented an approach to determine the population sizes for all generations of a backcross program before starting the breeding program (a priori). Their approach applies constant population sizes over generations and, furthermore, does not favor individuals of type $y_i^- x^+ y_r^+$ over those of type $y_i^- x^+ y_r^-$ even if d_1<d_2. In comparison with the a priori approach,

determining the population size n_2 a posteriori has the advantages that (1) only the number of individuals actually required to reach q_2 are generated, and (2) the probability q_2 can be reached for all outcomes of generation BC_1.

2.2 Reducing the Number of Backcross Generations

The above approach for calculating the population size focuses on reduction of the length of the intact donor chromosome segment around the target gene. Alternatively, the population size of a backcross program can be targeted to save a defined number of backcross generations in the breeding program. The population size required for this approach depends on the number and positions of the markers and the selection intensity. There are no analytical solutions to determine population sizes, but simulations can be used to solve the problem.

In such a simulation, first a reference breeding plan is simulated which describes the breeding program as it would be carried out without using markers. Usually the reference plan consists of six (Allard 1960) to eight (Fehr 1987) generations of backcrossing. With this simulation, the reference value reached for the recurrent parent genome is determined. In subsequent simulations of alternative scenarios with marker-assisted backcrossing, the parameters of the backcross program are varied until the reference value for the recurrent parent genome content is reached in the desired number of generations.

For a linkage map of maize, Frisch et al. (1999a) used the simulation software Plabsim (Frisch et al. 2000). We found that the recurrent parent genome content of 96.8 %, which was reached after six backcross generations without marker-assisted background selection, was reached after three backcross generations when using 80 markers and a population size of 100 individuals in each of generations BC_1 to BC_3.

2.3 Marker Positions

If only two background selection markers on the target chromosome are used (assuming direct selection for the target gene), the distances d_1 and d_2 between target gene and markers can be chosen such that the expected donor chromosome content on the target chromosome is minimized if both markers are fixed for the recipient allele (Hospital et al. 1992) by applying

$$d_1 = d_2 = \frac{1}{2} \ln \left(1 + 2\sqrt{s}\right), \tag{2}$$

where s is the proportion of selected BC_1 individuals. This approach is based upon the assumption of an infinite population size and the optimum properties only hold true if exactly two markers on the carrier chromosome of the target gene are used.

An alternative method of calculating of d_1 and d_2 is based on the rationale that in a population with given size n, at least one single or double recombinant individual is found with a given probability q (Frisch et al. 1999b). To determine n, the probabilities p given in Table 1 are inserted in

$$p = 1 - (1 - q)^{1/n} \quad (3)$$

and the resulting equation is solved for d_1 and d_2. Tabulated results are given by Frisch et al. (1999b).

The positions of markers used for background selection on noncarrier chromosomes can be determined such that the correlation between the molecular marker estimate of the donor genome content and the true donor genome content is maximized (Visscher 1996). In this approach, the distance between the telomere and the first marker is determined by numerical comparisons of alternative map positions in order to optimize the correlation. For the remaining markers, Visscher (1996) showed that the maximum correlation is reached if these are evenly spaced. The optimum distance between telomere and the first marker is different for each backcross generation, which makes it difficult to choose marker distances for a backcross program with subsequent generations of background selection.

An approach which takes selection over several generations into account was presented by Servin and Hospital (2002). They suggest choosing the positions of markers such that after all markers have been fixed for the recurrent parent allele, the expected donor genome content on the chromosome is maximized. As with Visscher's approach, the distance between the telomere and the first marker is determined with numerical evaluations and the remaining markers are evenly spaced.

2.4 Selection Strategies

We consider here a marker-assisted backcross program consisting of $s=1...\ t$ generations, where in total $n=n_1+...+n_t$ plants are employed and the population size n_s per generation is considerably greater than the minimum population size required to find at least one recombinant between target gene and flanking markers. The goal of marker-assisted background selection is to reduce the recurrent parent genome across all chromosomes. A straightforward design to accomplish this goal is to generate in each generation $n_s=n/t$ plants and to apply a two-stage selection strategy, consisting of selection for the target gene and one marker-assisted background selection step. For the background selection step, in generation BC_1 m markers with a good coverage of the entire genome are analyzed and an individual which carries the target gene and the recurrent parent alleles at most of the m markers is selected as parent for producing the next backcross generation. In subsequent backcross generations, selection is carried out according to the same scheme, but only those markers are analyzed which have not been fixed for the recurrent parent allele in the preceding generation.

When at least three generations of marker-assisted backcrossing have been carried out, the efficiency can be enhanced considerably by: (1) employing a small population size in generation BC_1 and increasing the population size in subsequent backcross generations, or (2) employing three- or four-stage selection strategies, emphasizing selection for recombinants on the carrier chromosome of the target gene during the first generations (Frisch et al. 1999a).

Employing increasing, constant, or decreasing population sizes from generations BC_1 to BC_3 in a simulation study had little effect on the recurrent parent genome values of the selected BC_3 plants (Frisch et al. 1999a). For example, allocating a total of n=300 plants such that 100 plants are generated in each of generations BC_1 to BC_3 (ratio n_1:n_2:n_3=1:1:1) resulted in a lower 10% percentile of the recurrent parent genome (Q10) of 97.4%, while various ratios from 3:2:1 on the one extreme to 1:3:9 on the other resulted in Q10 values of 97.3 or 97.4%. In contrast, employing a large population size in generation BC_1 multiplied the number of marker data points required for the marker-assisted backcrossing program. For example, only 2650 marker data points were required for n_1:n_2:n_3=1:3:9, while 5000 or even 7250 marker data points were required for ratios of 1:1:1 and 3:2:1, respectively.

The constant recurrent parent genome values for different ratios n_1:n_2:n_3 are in contrast to what is expected in multi-stage selection for a quantitative character. There, large populations in early generations are advantageous, because when high selection intensity is applied, a large selection gain is expected due to the large segregation variance. However, in marker-assisted backcrossing the increase in recurrent parent genome is not only driven by selection, but also by the backcross process itself. It is to be expected that backcrossing reduces the donor genome content by one half in each generation, irrespective of the amount present in the nonrecurrent parent. This implies that the selection gain attained in a certain backcross generation is halved by each additional backcross. Only the selection gain attained in the last backcross generation is fully recovered in the final product of the backcross program. Consequently, if high selection pressure (i.e., selection of one individual from a large population) is applied at the beginning of a marker-assisted backcrossing program, then a high absolute value for the selection gain is reached, but it is halved with each additional backcross generation. In contrast, if high selection pressure is applied in advanced backcross generations the selection gain is smaller, but the rate of recovering it in the final product of the breeding program is greater. A compensation between both effects explains why the ratio of dividing a constant number of individuals amongst the backcross generations hardly influences the recurrent parent genome in the plants selected at the end of a marker-assisted backcrossing program.

In generation BC_1, all markers are analyzed at the plants carrying the target gene; an approximation of the required number of marker data points is $mn_1/2$. With no marker-assisted selection, in each subsequent backcross gen-

eration the number of heterozygous markers is expected to be halved, therefore, a rough approximation of the portion of markers which are still heterozygous in generation BC_s, is $mn_s/2^s$. With marker-assisted background selection, the actual number of heterozygous markers which need to be analyzed in generation BC_s is below this approximation, because in addition to the effect of the backcrossing per se, homozygosity is increased by marker-assisted selection. These approximations illustrate that the expected portion of markers which need to be analyzed is greater in early than in advanced backcross generations. Therefore, large populations in early generations of a marker-assisted backcross program require more marker data points than large populations in late backcross generations. In conclusion, increasing the population size reduces the number of marker data points in a two-stage selection program compared to applying constant population sizes, but reaches comparable percentages of recurrent parent genome in the final breeding product.

Saving marker data points by using small BC_1 populations can be successfully applied without a linkage map of the markers. If linkage information is available, a sequential three- or four-stage selection strategy is another option to increase efficiency of a marker-assisted background selection program.

A three-stage selection strategy, consisting of one foreground selection step and two background selection steps, can be conducted as follows (Frisch et al. 1999a): after preselecting all individuals carrying the target gene, these are analyzed for the two markers flanking the target gene. On the basis of the result of this analysis, a selection index is constructed for each individual, taking the value 2, if both flanking markers are fixed for the recurrent parent allele and the value 1, if one out of the two flanking markers is fixed for the recurrent parents' allele. If both flanking markers are still heterozygous, the index takes the value 0. Subsequently, all individuals for which this selection index takes the largest observed value are analyzed for the remaining $m-2$ markers. Out of these individuals, the one carrying the recurrent parent allele at the largest number of markers is selected as parent for the next backcross generation.

An additional selection step extends three-stage selection to four-stage selection. After preselecting the individuals having the best selection index with respect to the flanking markers, these individuals are analyzed for all markers on the carrier chromosome of the target gene. Then a selection index is constructed, reflecting the number of markers on the carrier chromosome which were fixed for the recurrent parent allele. The plants for which this selection index takes the largest observed value are analyzed for the markers on the remaining chromosomes and the one carrying the recurrent parent allele at most markers is selected as parent for the next backcross generation.

In a simulated backcross experiment with maize, using 100 plants per backcross generation, two-stage selection reached a Q10 value of 97.4% in

generation BC_3 while three- and four-stage selection reached Q10 values 97.2 and 96.8%, respectively (Frisch et al.1999a). However, while for two-stage selection 5430 marker data points were required, three-stage and four-stage selection required only 1810 and 1390 marker data points, respectively. These results show that three- and four-stage selection provide an option which significantly reduces the number of marker data points required compared to two-stage selection, with only a minor reduction of the recurrent parent genome percentage reached.

3 Introgression of a Recessive Gene

Introgression of a recessive gene by recurrent backcrossing without the aid of molecular markers requires progeny tests in each backcross generation in order to determine whether a plant is a heterozygous carrier of the recessive gene or not. In addition to background selection, molecular markers can be used to indirectly select for the target gene such that progeny tests are not required in each backcross generation, but only at the end of the backcross program. In this section, we focus on the carrier chromosome of the target gene and outline a strategy to answer the following questions which describe a two-generation backcross program for introgression of a recessive gene by applying combined foreground and background selection (Frisch and Melchinger 2001a): (1) What is the necessary population size n_1 in BC_1? (2) Suppose the marker genotypes of the n_1 BC_1 individuals are known. Which marker genotypes g and how many individuals i_g of each should be selected as parents for further backcrossing? (3) What should be the size f_g of a BC_2 family produced from a selected BC_1 individual of genotype g?

We consider a chromosome with a sequence of loci at positions y_l, m_l, x, m_r, y_r. The target locus is located at position x and two flanking markers, used for foreground selection, are located at positions m_l and m_r. Two markers located at positions y_l and y_r are used for background selection. We denote z^- as the genotype of an individual homozygous for the recipient allele and z^+ as the genotype of a heterozygous individual at the locus at position $z \in \{y_l, m_l, x, m_r, y_r\}$.

To calculate the population size for generation BC_1, we define a set G of multi-locus genotypes with respect to map positions y_l, m_l, m_r, y_r, which comprises the marker genotypes of all plants considered as possible parents for generation BC_2. It contains all multi-locus marker genotypes with at least one background selection marker homozygous for the recurrent parent allele and at least one heterozygous foreground selection marker. Marker genotypes with no foreground selection marker carrying the donor allele are not included in G because with high probability they do not carry the target gene. Likewise, genotypes with only heterozygous background selection markers are excluded from G because, with respect to the goal of reducing the donor

genome around the target gene, they show no improvement compared with F_1 individuals.

To calculate the selection parameters for generation BC_2, we define a set T of multi-locus genotypes with respect to map positions y_l, m_l, m_r, y_r, comprising the marker genotypes of all plants, which are considered as a successful outcome of generation BC_2. It contains all multi-locus marker genotypes with both background selection markers homozygous for the recurrent parent allele and at least one heterozygous foreground selection marker. Marker genotypes with no foreground selection marker carrying the donor allele are not included in T because with high probability they do not carry the target gene. Likewise, genotypes with only one homozygous background selection marker are excluded because the goal of reducing the donor genome around the target gene has not been reached.

Using these definitions, the population size for generation BC_1 can be determined by inserting into Eq. (1) the probability

$$p = \sum_{g \in G} p_{0,g+} \tag{4}$$

that a BC_1 individual belongs to the set of genotypes G and carries the target gene. For each marker genotype $g \in G$, the probability $p_{0,g+}$ that a BC_1 individual has marker genotype g and carries the target gene is given in Table 3.

After producing the BC_1 generation and analyzing the plants with markers at positions y_l, m_l, m_r, y_r, the selection parameters for producing generation BC_2 need to be determined. We denote o_g as the number of individuals with genotype g observed in BC_1, i_g as the number of BC_1 individuals with genotype g used for further backcrossing, and f_g as the size of a BC_2 family produced from a BC_1 individual with genotype g. A certain parameter setting for generating the BC_2 generation, consisting of the number of individuals i_g to be backcrossed and the respective family size f_g for each marker genotype g, is denoted by S. The set of all admissible parameter settings, denoted by A, is determined by the following three conditions: (1) $0 \leq i_g \leq o_g$ for all $g \in G$, i.e., the number of selected individuals of genotype g cannot exceed the number of observed individuals, (2) $0 \leq f_g \leq m$ for all $g \in G$, i.e., the number of progenies generated from one plant cannot exceed the maximum possible family size m (which can be determined either by the multiplication rate of the species or the resources of the breeder), and (3) $q(S) \leq q_2$, i.e., the desired probability of success q_2 must be reached by the parameter combination S.

The probability $q(S)$ of recovering at least one BC_2 plant of marker genotype $t \in T$ carrying the target gene when using the parameter setting S is calculated as

$$q(S) = 1 - \prod_{g \in G} [1 - q_g(i_g, f_g)], \tag{5}$$

where $q_g(i_g, f_g)$ is the probability of finding among the i_g backcross families of size f_g at least one carrier of the target gene with genotype $t \in T$

Table 3. Formulas to calculate the probabilities $p_{0,g+}$ (probability that a BC_1 individual has marker genotype g and carries the target gene), $p_{g+|0,g}$ (probability that a BC_1 individual with marker genotype g carries the target gene), and $p_{g+,T+}$ (probability that a BC_1 individual with marker genotype g, which carries the target gene, generates a BC_2 individual with marker genotype $t \in T$, which carries the target gene.)

Marker genotype $g \in G$	$p_{0,g+}$	$p_{g+\|0,g}$	$p_{g+,T+}$
$y_l^- m_l^+ m_r^+ y_r^-$ [a]	$p_b(1-p_c)(1-p_d)p_e/2$ [b]	$(1-p_c)(1-p_d)/(1-p_h)$	$(1-p_c p_d)/2$
$y_l^- m_l^- m_r^+ y_r^-$	$(1-p_b)p_c(1-p_d)p_e/2$	$p_c(1-p_d)/p_h$	$(1-p_d)/2$
$y_l^- m_l^+ m_r^- y_r^-$	$p_b(1-p_c)p_d(1-p_e)/2$	$(1-p_c)p_d/p_h$	$(1-p_c)/2$
$y_l^+ m_l^+ m_r^+ y_r^-$	$(1-p_b)(1-p_c)(1-p_d)p_e/2$	$(1-p_c)(1-p_d)/(1-p_h)$	$[p_b(1-p_c)+(1-p_b)p_c(1-p_d)]/2$
$y_l^- m_l^+ m_r^+ y_r^+$	$p_b(1-p_c)(1-p_d)(1-p_e)/2$	$(1-p_c)(1-p_d)/(1-p_h)$	$[(1-p_d)p_e+(1-p_c)p_d(1-p_e)]/2$
$y_l^+ m_l^- m_r^+ y_r^-$	$p_b p_c(1-p_d)p_e/2$	$p_c(1-p_d)/p_h$	$p_a(1-p_d)/2$
$y_l^+ m_l^+ m_r^- y_r^-$	$(1-p_b)(1-p_c)p_d(1-p_e)/2$	$(1-p_c)p_d/p_h$	$p_b(1-p_c)/2$
$y_l^- m_l^- m_r^+ y_r^+$	$(1-p_b)p_c(1-p_d)(1-p_e)/2$	$p_c(1-p_d)/p_h$	$(1-p_d)p_a/2$
$y_l^- m_l^+ m_r^- y_r^+$	$p_b(1-p_c)p_d p_e/2$	$(1-p_c)p_d/p_h$	$(1-p_c)p_f/2$

[a] The symbols y_l and y_r denote the background selection markers, m_l and m_r the foreground selection markers, and x the target locus. A superscript + or – indicates that the locus is heterozygous or homozygous for the recurrent parent allele, respectively

[b] The probabilities p_a to p_h are the recombination frequencies between the loci delimiting the intervals $[y_l, x]$, $[y_l, m_l]$, $[m_l, x]$, $[x, m_r]$, $[m_l, y_r]$, $[x, y_r]$, $[y_l, m_r]$, and $[m_l, m_r]$, respectively. They can be obtained from $p=(1-e^{-2d})/2$ by inserting the corresponding map distance d between the loci delimiting the interval

$$q_g\,(i_g, f_g) = \sum_{s=1}^{i_g} [B(i_g, s, p_{g+|0,g})\{1 - B(sf_g, 0, p_{g+T,+})\}]. \tag{6}$$

For each marker genotype $g \in G$, the probabilities $p_{g+|0,g}$ and $p_{g+,T+}$ are given in Table 3. $p_{g+|0,g}$ denotes the probability that a BC_1 individual with marker genotype g carries the target gene. $p_{g+,T+}$ denotes the probability that a BC_1 individual with marker genotype g, which carries the target gene, generates a BC_2 individual with marker genotype $t \in T$, which carries the target gene.

$B\,(n, m, p) = \binom{n}{m} p^m\,(1 - p)^{n-m}$ is the probability function of the binomial distribution. If a particular genotype occurs with probability p, the number m of individuals of this type in a sample of size n is binomially distributed with probability $B(n,m,p)$.

The number of individuals required for the parameter setting S is

$$n_2\,(S) = \sum_{g \in G} i_g f_g \tag{7}$$

and the optimum parameter setting S^* is the one requiring the smallest number of individuals among all elements in A.

$$n_2\,(S^*) = \min_{S \in A} n_2\,(S). \tag{8}$$

There is no closed analytical solution for the minimization problem in Eq. (8). To find a suitable parameter setting, we propose to calculate the probability of success $q(S)$ for various parameter settings S and choose the one which is an element of A and requires the smallest number of individuals.

4 Introgression of Two Dominant Genes

Alternative breeding schemes exist for the simultaneous introgression of two genes into the genetic background of an inbred line (Frisch and Melchinger 2001a). They differ in the generation in which a plant carrying both target genes is generated for the first time. The two genes can be merged into one individual before starting the backcross program by crossing the donors of the target genes and using the resulting F_1 as the nonrecurrent parent for backcrossing. Alternatively, the two genes can be introgressed in two separate branches of the breeding program into the recipient and only when the introgression is finished after t generations of backcrossing, the two converted BC_t individuals are crossed in order to merge the target genes. Between these two extremes, breeding plans for a t-generation backcross program can be applied, in which the target genes are merged into one individual in generation BC_s ($s<t$). These alternative breeding plans differ with respect to: (1) the minimum population size required for finding, with a given probability of

success, carriers of both target genes in different types of populations and (2) the selection intensity which has an effect on the percentage of the recurrent parent genome reached and the number of marker data points required in the backcross program.

The minimum population size required to recover carriers of both target genes depends on the degree of linkage between them and whether they are in coupling or repulsion phase linkage in the crossing or selfing parent. The required population size can be calculated by inserting the respective probabilities given by Frisch and Melchinger (2001b) into Eq. (1). Special attention is required for breeding programs in which linked target genes are merged into one individual by crossing two BC_t plants, followed by a selfing generation to generate homozygous carriers of the target genes. In the selfing parents, the target genes occur in repulsion phase, i.e., they are located on different homologous chromosomes, one originating from the male, the second from the female BC_t plant. To generate a plant which carries both target genes homozygous, it is therefore required that recombination between the target genes occurs during the formation of both parental gametes. The probability $p=(1-r)^2/4$ that such a plant occurs is lower, the tighter linkage is. This results in large populations being required for tightly linked target genes.

When two unlinked target genes are merged into one plant before the first backcross generation and a total of n individuals are generated per backcross generation, then about $n/4$ plants are expected to be subjected to marker-assisted background selection. In contrast, when each target genes is introgressed in a separate branch of the breeding program with a population size of $n/2$, then out of the n plants employed in total for a certain generation, about $n/2$ are expected to be subjected to marker-assisted background selection. In consequence, the intensity of selection for the recurrent parent genome in a breeding plan in which the target genes are merged in a later generation is greater than in a breeding plan with early merging of the target genes. The greater selection intensity is accompanied by greater values of the recurrent parent genome reached, but also by larger numbers of marker data points required. This was demonstrated numerically in a simulation study based on a model of the maize genome (Frisch and Melchinger 2001b).

5 Length of the Intact Donor Chromosome Segment Around the Target Gene

For recurrent backcrossing with selection for the presence of a target gene, the expected length of the intact donor chromosome segment attached on one side of the target gene was derived by Hanson (1959) as $(1-e^{-tl})/t$, where t is the number of backcrosses carried out and l the map distance between the target gene and the end of the chromosome. Stam and Zeven (1981) extended

Hanson's approach and derived the expected donor genome content on the carrier chromosome. Their approach includes chromosome segments not directly attached to the target gene and averages over all possible map positions of the target gene on the chromosome. For background selection with exactly two markers on the carrier chromosome of the target gene, Hospital et al. (1992) extended the approach of Stam and Zeven (1981) and determined numerically the expected donor genome content on the target chromosome.

The probability distribution of the intact donor chromosome segment around the target gene in backcrossing with selection for the presence of the target gene and selection for the recipient alleles at flanking markers was investigated by Hospital (2001) and Frisch and Melchinger (2001c). For various situations relevant for practical backcross programs in plant breeding, they derived density functions, expectations, and the variances of the lengths of the attached donor chromosome segment.

A numerical illustration of their results presented in Table 4 shows the expected length $E(X)$ of the intact chromosome segment attached to one side of the target gene in generations BC_t (t=1...5, 6, 8, 10, 15) for backcross programs with and without background selection at flanking markers in generation BC_1. The target locus is positioned at distance l=1.0 M from the chromosome end and the flanking marker is located at distance 0.1, 0.2, 0.3, 0.4, 0.5 M from the target locus. In generation BC_1, the expected length of the intact chromosome segment is 0.25 M, when selecting for a flanking marker at 0.5 M distance. Without marker-assisted selection, a value of 0.24 M is reached only in generation BC_4. With an increasing number of backcrosses, the differences between applying background selection and not applying

Table 4. Expected length $E(X)$ of the intact chromosome segment attached on one side of the target gene in generations BC_s (s=1...5, 6, 8, 10, 15) in the absence of marker-assisted selection (none) and with selection for recombinants at a flanking marker at varying distances (d=0.1,...0.5 M) in generation BC_1. The map distance between the target gene and the telomere is 1 M

	Flanking marker distance (M)					
s	None	0.5	0.4	0.3	0.2	0.1
	$E(X)$[M]					
1	0.63	0.24	0.20	0.15	0.01	0.05
2	0.43	0.21	0.17	0.14	0.09	0.05
3	0.32	0.18	0.15	0.12	0.09	0.05
4	0.25	0.16	0.14	0.11	0.08	0.05
5	0.20	0.14	0.12	0.10	0.08	0.04
6	0.17	0.12	0.11	0.10	0.07	0.04
8	0.12	0.10	0.09	0.08	0.07	0.04
10	0.10	0.09	0.08	0.07	0.06	0.04
15	0.07	0.06	0.06	0.05	0.05	0.03

background selection becomes smaller. However, an expected length of the attached chromosome segment of 0.05 M, as reached in generation BC_1 with a flanking marker distance of 0.1 M, is not reached even after 15 backcross generations without background selection.

These results show that selection for recombinants between the target gene and a flanking marker is highly effective even when the marker is fairly distant from the target gene. For example, a saving of three backcross generations concerning the expected length of the linked chromosome segment is realized with a marker distance of 0.5 M. Because recombinants between the target gene and fairly distant flanking markers occur with a high probability even in small backcross generations (Frisch et al. 1999b), marker-assisted background selection can be used to avoid large intact donor chromosome segments around the target gene, even with limited resources for the population size and marker analyses.

Acknowledgment The author thanks A.E. Melchinger for helpful discussions and comments on the manuscript.

References

Allard RW (1960) Principles of plant breeding. Wiley, New York

Fehr WR (1987) Principles of cultivar development, vol 1. Theory and technique. Macmillan, New York

Frisch M, Melchinger AE (2001a) Marker-assisted backcrossing for introgression of a recessive gene. Crop Sci 41:1485–1494

Frisch M, Melchinger AE (2001b) Marker-assisted backcrossing for simultaneous introgression of two genes. Crop Sci 41:1716–1725

Frisch M, Melchinger AE (2001c) The length of the intact chromosome segment around a target gene in marker-assisted backcrossing. Genetics 157:1343–1356

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

Frisch M, Bohn M, Melchinger AE (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, Erratum: Crop Sci 39:1903

Frisch M, Bohn M, Melchinger AE (2000) Plabsim: software for simulation of marker-assisted backcrossing. J Hered 91:86–87

Hanson WD (1959) Early generation analysis of lengths of heterozygous chromosome segments around a locus held heterozygous with backcrossing or selfing. Genetics 44:833–837

Hospital F (2001) Size of donor chromosome segments around introgressed loci and reduction of linkage drag in marker-assisted backcross programs. Genetics 158:1363–1379

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

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

Ragot M, Biasiolli M, Delbut MF, Dell'Orco A, Malgarini L, Thevenin P, Vernoy J, Vivant J, Zimmermann R, Gay G (1995) Marker-assisted backcrossing: a practical example. In: INRA (ed) Techniques et utilisations des marqueurs moléculaires. Montepellier, France, 29–31 March 1994

Servin B, Hospital F (2002) Optimal positioning of markers to control genetic background in marker-assisted backcrossing. J Hered 93:214–217

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

Tanksley SD (1983) Molecular markers in plant breeding. Plant Mol Biol Rep 1:1–3

Tanksley SD, Young ND, Patterson AH, Bonierbale MW (1989) RFLP mapping in plant breeding: New tools for an old science. Bio/Technology 7:257–263

Visscher PM (1996) Proportion of the variation in genomic composition in backcrossing programs explained by molecular markers. J Hered 87:136–138

Young ND, Tanksley SD (1989) RFLP analysis of the size of chromosomal segments retained around the *tm*-2 locus of tomato during backcross breeding. Theor Appl Genet 77:353–359

Zeven AC, Knott DR, Johnson R (1983) Investigation of linkage drag in near isogenic lines of wheat by testing for seedling reaction to races of stem rust, leaf rust and yellow rust. Euphytica 32:319–327