# Comparison of Selection Strategies for Marker-Assisted Backcrossing of a Gene

Matthias Frisch, Martin Bohn, and Albrecht E. Melchinger*

## ABSTRACT

**Marker-assisted selection can accelerate recovery of the recurrent parent genome (RPG) in backcross breeding. In this study, we used computer simulations to compare selection strategies with regard to (i) the proportion of the RPG recovered and (ii) the number of marker data points (MDP) required in a backcross program designed for introgression of one target allele from a donor line into a recipient line. Simulations were performed with a published maize (*Zea mays* L.) genetic map consisting of 80 markers. Selection for the target allele was based on phenotypic evaluation. In comparison to a constant population size across all generations, increasing population sizes from generation $BC_1$ to $BC_3$ reduced the number of required MDP by as much as 50% without affecting the proportion of the RPG. A four-stage selection approach, emphasizing in the first generations selection for recombinants on the carrier chromosome of the target allele, reduced the required number of MDP by as much as 75% in comparison to a selection index taking into account all markers across the genome. Adopting the above principles for the design of marker-assisted backcross programs resulted in substantial savings in the number of MDP required.**

The backcross procedure is used in plant breeding to transfer favorable alleles from a donor genotype, which has mostly poor agronomic properties, into a recipient elite genotype (Allard, 1960, p. 150). Marker assays can be of advantage in backcross breeding for foreground selection and background selection (Hospital and Charcosset, 1997). In the first approach, the presence of a target allele in an individual is diagnosed by monitoring the genotype at flanking markers for alleles of the donor parent. This is a powerful tool for manipulation of oligogenic traits under numerous situations in plant breeding (for review see Melchinger, 1990), but also for manipulation of quantitative trait loci (QTL) (Stuber, 1995). The second approach, devised by Tanksley et al. (1989), accelerates recovery of the RPG. Individuals are selected which are homozygous for the alleles of the recurrent parent at a large number of marker loci covering the entire genome. Marker-assisted background selection has meanwhile been established as a standard tool in plant breeding (see, e.g., Ragot et al., 1995).

Computer simulations have proved to be a powerful tool for investigating the design and efficiency of marker-assisted selection programs (for review see Visscher et al., 1996). These authors studied marker-assisted QTL introgression in an animal breeding context, using an infinitesimal model to explain differences among breeds. Hospital and Charcosset (1997) determined the optimal position and number of marker loci for manipulating QTL in foreground selection. Further, they investigated the combination of foreground and background selection in QTL introgression. Openshaw et al. (1994) determined the population size and marker density required in background selection. They recommended the use of four markers per chromosome (of 200-cM length) and a selection strategy for proximal recombinants of the target allele.

Although efficient PCR-based DNA markers such as simple sequence repeats and amplified fragment length polymorphisms are available (Ribaut et al., 1997), their use in background selection is restricted by the large number of required MDP. In this study, we investigate strategies for reducing the total number of MDP needed in background selection. Our research objectives were to (i) determine the number of MDP required in background selection, (ii) investigate the effects of varying population sizes from early to late backcross generations on the level of RPG and the MDP required, and (iii) compare a two-stage selection procedure, consisting of one foreground and one background selection step, with alternative selection procedures consisting of one foreground selection step and two or three background selection steps.

## METHODS

### Genetic Map

Our simulations were based on a published linkage map of maize (Schön et al., 1994) constructed from a population of 380 $F_2$ individuals derived from the cross of two flint inbred lines. The total map length was 1612 cM. On the basis of previous investigations (Openshaw et al., 1994; Visscher et al., 1996; Frisch et al., 1998), an average marker density of about 20 cM is sufficient to warrant a good coverage of the genome in marker-assisted selection programs. Hence, 80 of the 89 polymorphic restriction fragment length polymorphism markers used by Schön et al. (1994) were chosen to obtain an average marker density of 20 cM. Markers *umc128*, *umc5*, *umc175*, *bn16.06*, *umc54*, *umc51*, *umc110*, *bnl7.61*, and *bnl9.44* were tightly linked to other markers and, therefore, excluded from the present study. There were two larger gaps on this map: one 90-cM marker interval on Chromosome 3 and one 89-cM marker interval on Chromosome 9. The target locus was assumed to be located on Chromosome 5, 30 cM from the telomere. In our simulations, the entire map was additionally covered with equally spaced (1 cM) background loci to monitor the parental origin of the whole genome.

### Algorithm

Software PLABSIM (Frisch et al., 1999b), a computer program written in C++, was used to simulate the recombination process during meiosis. Crossover events were generated by

Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany. Received 24 Nov. 1998. *Corresponding author (melchinger@uni-hohenheim.de).

Published in Crop Sci. 39:1295–1301 (1999).

**Abbreviations:** $BC_t$, *t*th backcross generation; cM, centimorgan; MDP, marker data points; QTL, quantitative trait locus; RPG, recurrent parent genome.

a random-walk algorithm (Crosby, 1973, p. 237). Recombination frequencies required for the random walk were calculated from the map distance by Haldane's (1919) mapping function. This assumes that neither chiasma interference nor chromatide interference (Stam, 1979) occur. To check our simulation software, the original linkage map of Schön et al. (1994), which was based on experimental $F_2$ data, was compared with a linkage map constructed from simulated data of $F_2$ individuals by MAPMAKER software (Lander et al., 1987). Both maps were in excellent agreement, confirming that the models underlying the two software packages were similar.

### Simulation Runs

Each simulation of a backcross program started by the cross of two parents, which were assumed to be homozygous and polymorphic at all loci (target locus, marker loci, background loci). The recurrent parent was assumed to carry the desirable alleles at all loci of the genome except for the target locus. The donor parent was assumed to carry the desirable allele at the target locus in homozygous state. One heterozygous $F_1$ individual was backcrossed with the recurrent parent and $n_1$ $BC_1$ individuals were produced. The best $BC_1$ individual was selected according to the selection strategies described below and, for production of generation $BC_2$, backcrossed with the recurrent parent. This procedure was repeated for $t$ backcross generations. For the selected individual in each generation $BC_t$, the percentage of the RPG was determined by dividing the number of loci (marker and background loci) homozygous for the recurrent parent allele by the total number of loci monitored. Furthermore, each analysis of a marker locus in a backcross individual was counted as a MDP. In $BC_1$, the entire set of markers was analyzed (at least in the individual selected as parent for producing generation $BC_2$). In the following generations, only those markers not fixed for the recurrent parent allele in the nonrecurrent parent (i.e., individual selected in the previous generation) were analyzed. The number of MDP required in each generation was counted and summed over the whole backcross program. The simulation of each backcross program was repeated 10 000 times to reduce sampling effects and obtain results with sufficient numerical accuracy.

### Threshold for the RPG

The values gained from these 10 000 repetitions can be regarded as realizations of random variables that describe the proportion of RPG and the total number of MDP required after $t$ generations in a backcross program with the parameter settings considered. The 10% percentile of the empirical distribution of the RPG in the selected individual (Q10) is used as an estimator for the amount of RPG reached after selection in generation $BC_t$ with probability 0.90. Compared with arithmetic means, percentiles have two advantages.

1. The skewness of the RPG distribution increases in advanced backcross generations. Percentiles are more suitable than arithmetic means for comparison of skewed distributions.
2. Inferences about the probability to achieve a certain goal can be made. For example, a Q10 value of 98% means that "with probability 0.90 an RPG proportion greater than 98% is attained" under the considered parameter combination.

### Simulations to Determine Threshold Values

A full backcross program usually consists of six generations (Allard, 960, p. 155). Hence, the Q10 values reached in generation $BC_6$ by applying random selection among all individuals carrying the target allele was used as a termination threshold for a marker-assisted backcross program. This threshold was determined by simulations with selection only for presence of the target allele but no selection for any marker loci.

### Selection Strategies

For describing our selection strategies in general terms, we consider a chromosome carrying the target locus (carrier chromosome) of length $l_0$ and $c$ further chromosomes (non-carrier chromosomes) with length $l_c$. Positions on the chromosomes are represented by a scale in Morgan units ranging from 0 to $l_c$. The target locus is located at position $x$ on the carrier chromosome and two flanking markers at positions $y_1$ and $y_2$; $i$ additional markers on the target chromosome are located at positions $z_i$. On the non-carrier chromosomes are altogether $m$ markers positioned at positions $u_{ck}$. Let $X$, $Y_1$, $Y_2$, $Z_i$, and $U_{ck}$ be indicator variables, which take the value 1, if the corresponding locus is homozygous for the recurrent parent allele and 0 otherwise. From these random variables we obtain the count variables $Y = Y_1 + Y_2$ and $U = Y_1 + Y_2 + \Sigma_i Z_i + \Sigma_c \Sigma_k U_{ck}$. Furthermore, we define the indicator variable $Z$, which is 1 if all $i$ additional markers on the carrier chromosome are homozygous for the recurrent parent allele and 0 otherwise.

By means of the random variables $X$, $Y$, $Z$, and $U$ as selection indices, three sequential selection strategies were applied. The first step always involved selection of individuals carrying the target allele ($X = 0$). Subsequently one, two, or three steps with background selection followed (Table 1). In each selection step, only those individuals selected in the previous step are subjected to marker assays. In the selected individual

**Table 1. Description of selection steps and their sequence in the three selection strategies investigated.**

| | | Sequence of selection steps in | | |
|---|---|---|---|---|
| **Selection step** | **Condition†** | **Two-stage selection** | **Three-stage selection** | **Four-stage selection** |
| **Select individuals carrying the target allele** | $X = 0$ | **1** | **1** | **1** |
| **Select individuals homozygous for the recurrent parent allele at most flanking markers** | $\max(Y)$ | **–‡** | **2** | **2** |
| **Select individuals homozygous for the recurrent parent allele at all additional markers on the carrier chromosome** | $\max(Z)$ | **–** | **–** | **3** |
| **Select one individual which is homozygous for the recurrent parent allele at the maximum number of all markers across the genome** | $\max(U)$ | **2** | **3** | **4** |

† $X$, $Y_1$, $Y_2$, $Z_i$, and $U_{ck}$ are indicator variables, which take the value 1, if the loci at positions $x$, $y_1$, $y_2$, $z_i$, and $u_{ck}$ are homozygous for the recurrent parent allele and 0 otherwise. From these random variables the count variables $Y = Y_1 + Y_2$ and $U = Y_1 + Y_2 + \Sigma_i Z_i + \Sigma_c \Sigma_k U_{ck}$ are obtained. The indicator variable $Z$ is 1 if all $i$ additional markers on the carrier chromosome are homozygous for the recurrent parent allele and 0 otherwise.

‡ Not carried out.

for producing the next backcross generation, all markers not fixed in the previous generation(s) are assayed to determine homozygosity and, hence, which need not to be assayed in the following generation(s).

The selection strategies differ in the selection pressure applied to carrier versus non-carrier chromosomes. In two-stage selection, selection in the second step is based on the Index $U$, which takes into account all marker loci irrespective of their position in the genome. In three-stage selection, the second selection step rests on the flanking markers (Index $Y$), while the final step is again based on all markers (Index $U$) irrespective of their genomic location. Four-stage selection is similar to three-stage selection, but inserts after the second step one additional selection exclusively based on the markers located on the carrier chromosome (Index $Z$). Hence, emphasis given to RPG recovery on the carrier chromosome increases from two- to four-stage selection. A selection procedure preferring recombinants at flanking markers similar to our three-stage selection was proposed by various authors (Tanksley et al., 1989; Hospital et al., 1992; Openshaw et al., 1994; Hospital and Charcosset, 1997).

### Population Size

Backcrossing with a constant number of individuals in each generation $BC_t$ ($n_t$ = 20, 40, 60, 80, 100, 125, 150, 200) was compared with backcrossing, in which the population size $n_t$ varied from $BC_1$ to $BC_3$. The total number of individuals $\Sigma n_t$ = 300 was allocated to backcross generations $BC_1$:$BC_2$:$BC_3$ with ratios of 3:2:1, 1:1:1, 1:2:4, 1:2:3:, 1:3:5, and 1:3:9.

## RESULTS

In backcrossing, when selection is performed only for the presence of the target allele, the mean of the RPG was about 1% below the theoretical values expected without selection (Table 2). After six generations of backcrossing, a Q10 value of 96.7% was reached. This value was subsequently used as a threshold to determine the termination of a marker-assisted backcrossing program. From $BC_7$ to $BC_{10}$, Q10 increased only 2.0% with marginal gains in advanced generations.

Under two-stage selection with a constant population size, Q10 amounted to 97.8% with $n_t$ = 20 in $BC_4$ and 97.1% with $n_t$ = 60 in $BC_3$ (Table 3). The first parameter setting resulted in saving two backcross generations and required a total of 1180 MDP, while the second parameter setting saved three generations and required 3340 MDP. Even with $n_t$ = 200, the Q10 value did not exceed the threshold of 96.7% in $BC_2$. For $n_t$ = 150 and 7990 MDP, Q10 reached 97.6% in $BC_3$, which corresponds to a saving of four backcross generations.

**Table 2. Simulation results for the mean and 10% percentile (Q10) of the distribution of the recurrent parent genome in generation $BC_t$ with random selection of individuals carrying the target allele and expected values for the mean without selection.**

| | No selection | Selection | |
|---|---|---|---|
| Generation | Mean | Mean | Q10 |
| | | % | |
| $BC_1$ | 75.0 | 74.0 | 67.4 |
| $BC_2$ | 87.5 | 86.1 | 80.7 |
| $BC_3$ | 93.8 | 92.4 | 88.3 |
| $BC_4$ | 96.9 | 95.6 | 92.7 |
| $BC_5$ | 98.4 | 97.3 | 95.2 |
| *$BC_6$* | 99.2 | 98.2 | *96.7†* |
| *$BC_7$* | *99.6* | *98.7* | *97.6* |
| *$BC_8$* | *99.8* | *99.0* | *98.1* |
| *$BC_9$* | *99.9* | *99.1* | *98.5* |
| *$BC_{10}$* | *100.0* | *99.3* | *98.7* |

† Used as threshold in subsequent tables.

After generation $BC_3$, the required number of MDP increased slowly for all values of $n_t$ (Table 3). A large proportion of markers were fixed for the recurrent parent allele in the individual selected in generation $BC_3$. Increasing $n_t$ beyond 100 had little effect on the recovery of the RPG, but was consuming of MDP. For example, in a two-stage selection program with constant $n_t$, with $n_t$ = 100 resulted in Q10 = 97.4% in $BC_3$ and required 5430 MDP, while with $n_t$ = 200 resulted in Q10 = 97.8% but required 10 500 MDP. The total number of MDP required in two-stage selection with constant population size was approximately proportional to $n_t$. The greatest proportion of total MDP was consumed in generation $BC_1$: about 60% for $n_t$ = 20 and about 80% for $n_t$ = 200.

Three-stage selection with constant $n_t$ yielded lower Q10 values than two-stage selection only in $BC_1$ and $BC_2$, but in subsequent backcross generations the difference was only marginal especially for greater $n_t$ values (Table 3). Increasing $n_t$ from 20 to 60 resulted in a substantial increase of Q10 values only up to $BC_3$ but not in later backcross generations. Likewise, increasing $n_t$ beyond 60 resulted only in marginal gains in Q10. In comparison with two-stage selection, less than half the total number of MDP were required in a three-generation backcross program for all values of $n_t$. This reduction was attributable to considerable savings in $BC_1$ (Table 3).

For four-stage selection with constant $n_t$, the Q10 values followed the same trends as for three-stage selection. Corresponding Q10 values never exceeded those for the latter procedure, but differences were negligible after generation $BC_2$, irrespective of the choice of $n_t$ (Table 3). However, the total MDP number was reduced, compared with three-stage selection (about 15% for $n_t$ = 20 and 28% for $n_t$ = 200), and even more when compared with two-stage selection.

Variation in $n_t$ values for $BC_1$ to $BC_3$ with the restriction $\Sigma n_t$ = 300 hardly influenced the Q10 values reached in $BC_3$ under two-stage selection (Table 4). In contrast, the number of MDP required was strongly reduced with larger values for $n_t$ in advanced backcross generations. In comparison to the ratio 1:1:1, increasing ratios of $n_t$ reduced the required number of MDP up to 50%, while decreasing ratios of $n_t$ increased the required number of MDP up to 150%. Variation of $n_t$ in three- and four-stage selection had only marginal influence on both the RPG and the required number of MDP for ratios of 3:2:1 to 1:2:4. A reduction in RPG was observed for the ratio 1:3:9 (Table 4).

## DISCUSSION

### Recurrent Parent Genome

In analogy to response to selection for a quantitative character with a normal distribution (Falconer and Mackay, 1996, p. 185), response to selection for the RPG

**Table 3. Simulation results for the 10% percentile (Q10) of the distribution of the recurrent parent genome (RPG) and total number of marker data points (MDP) required in a backcross program to introgress one target allele, using constant population size $n_t$ in all backcross generations. Values for MDP are rounded to multiples of ten.**

| | Number of individuals $n_t$ per backcross generation | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Generation | 20 | 40 | 60 | 80 | 100 | 125 | 150 | 200 |
| | Q10 (%) of the RPG | | | | | | | |
| Two-stage selection | | | | | | | | |
| $BC_1$ | 76.7 | 78.7 | 79.7 | 80.3 | 80.7 | 81.3 | 81.7 | 82.2 |
| $BC_2$ | 90.3 | 91.9 | 92.8 | 93.3 | 93.6 | 93.9 | 94.0 | 94.6 |
| $BC_3$ | 95.8 | 96.2 | *97.1* | *97.3* | *97.4* | *97.5* | *97.6* | *97.8* |
| $BC_4$ | *97.8*† | *97.9* | 98.4 | 98.5 | 98.5 | 98.6 | 98.6 | 98.7 |
| $BC_5$ | 98.7 | 98.9 | 99.0 | 99.0 | 99.0 | 99.0 | 99.0 | 99.0 |
| Three-stage selection | | | | | | | | |
| $BC_1$ | 71.2 | 72.7 | 73.4 | 73.6 | 73.3 | 73.2 | 72.8 | 72.2 |
| $BC_2$ | 86.1 | 87.2 | 88.5 | 89.3 | 90.2 | 90.7 | 91.3 | 91.8 |
| $BC_3$ | 94.4 | 95.7 | 96.5 | *96.9* | *97.2* | *97.3* | *97.5* | *97.6* |
| $BC_4$ | *97.7* | *98.2* | *98.4* | 98.4 | 98.4 | 98.5 | 98.5 | 98.5 |
| $BC_5$ | 98.7 | 98.8 | 98.9 | 98.9 | 98.9 | 98.9 | 99.0 | 99.0 |
| Four-stage selection | | | | | | | | |
| $BC_1$ | 71.0 | 71.9 | 72.1 | 71.7 | 71.6 | 71.5 | 71.2 | 71.0 |
| $BC_2$ | 85.5 | 86.2 | 87.2 | 87.6 | 88.2 | 88.7 | 89.1 | 89.8 |
| $BC_3$ | 93.7 | 95.0 | 96.0 | 96.5 | *96.8* | *97.0* | *97.2* | *97.4* |
| $BC_4$ | *97.6* | *98.2* | *98.3* | *98.4* | 98.4 | 98.4 | 98.4 | 98.5 |
| $BC_5$ | 98.7 | 98.8 | 98.9 | 98.9 | 98.9 | 98.9 | 98.9 | 98.9 |
| | Number of MDP required in total | | | | | | | |
| Two-stage selection | | | | | | | | |
| $BC_1$ | 800 | 1560 | 2400 | 3200 | 4000 | 5000 | 5990 | 8 000 |
| $BC_2$ | 1010 | 2130 | 3150 | 4170 | 5180 | 6430 | 7670 | 10 100 |
| $BC_3$ | 1180 | 2280 | *3340* | *4390* | *5430* | *6720* | *7990* | *10 500* |
| $BC_4$ | *1210* | *2310* | 3380 | 4430 | 5470 | 6750 | 8030 | 10 600 |
| $BC_5$ | 1220 | 2320 | 3380 | 4430 | 5470 | 6760 | 8030 | 10 600 |
| Three-stage selection | | | | | | | | |
| $BC_1$ | 250 | 320 | 420 | 510 | 590 | 690 | 750 | 840 |
| $BC_2$ | 440 | 610 | 830 | 1100 | 1390 | 1780 | 2210 | 3 110 |
| $BC_3$ | 550 | 820 | 1130 | *1470* | *1810* | *2260* | *2740* | *3 740* |
| $BC_4$ | *590* | *860* | *1170* | 1500 | 1840 | 2280 | 2760 | 3 760 |
| $BC_5$ | 590 | 860 | 1170 | 1500 | 1840 | 2280 | 2760 | 3 760 |
| Four-stage selection | | | | | | | | |
| $BC_1$ | 230 | 270 | 340 | 390 | 430 | 470 | 480 | 520 |
| $BC_2$ | 370 | 460 | 590 | 750 | 910 | 1140 | 1360 | 1 900 |
| $BC_3$ | 460 | 660 | 900 | 1140 | *1390* | *1710* | *2020* | *2 690* |
| $BC_4$ | *500* | *710* | *950* | *1190* | 1430 | 1740 | 2050 | 2 720 |
| $BC_5$ | 510 | 710 | 950 | 1190 | 1430 | 1740 | 2050 | 2 720 |

† Q10 values exceeding for the first time the threshold of 96.7% and the respective total number of MDP required are printed in italics.

in background selection can be calculated as $R = i \sigma r$. Here, $i$ denotes the selection intensity, $\sigma$ the standard deviation of the RPG, and $r$ the correlation between the proportion of recurrent parent alleles at marker loci and the proportion of recurrent parent alleles across the whole genome. Values of $\sigma$ and $r$ for the three selection strategies are given in Table 5.

In addition to background selection for RPG, the backcross process itself increases the RPG values in each backcross generation. By expectation, the donor genome proportion is halved with each backcross generation, irrespective of its amount present in the nonrecurrent parent. This implies that increasing the RPG proportion by selection in a backcross generation has a carry-over rate of one half to the next backcross generation. Consequently, increasing the RPG by selection is more effective (with regard to the RPG in the end product of the breeding program), if it is realized in an advanced backcross generation. This proposition can be proved analytically and is a generalization of results of Hospital et al. (1992). They demonstrated that a single generation background selection is most efficient if selection is performed in the last backcross generation.

Marker-assisted selection is different from selection for a quantitative character, where a high selection intensity in early generations can take advantage of the large segregation variance among individuals. There is no such optimum generation for applying high selection intensities in marker-assisted background selection. If large $BC_1$ population sizes are chosen, the response to selection is high due to large values of $\sigma$ and $r$ (Table 5). However, in each of the following backcross generations this initial gain in RPG is halved. In contrast, the response to background selection achieved by large population sizes in the last backcross generation is fully recovered in the breeding product and not diluted by further backcrossing, even if due to smaller $\sigma$ and $r$ values (Table 5) the absolute values of the response to selection are smaller in advanced backcross generations. A compensation of both effects explains why in $BC_3$ the content of RPG in the selected individual is hardly influenced by the ratio of population sizes used in $BC_1$ to $BC_3$, given a constant total number of individuals.

Compared with two-stage selection, in three-stage or four-stage selection greater emphasis is given to the carrier chromosome in generation $BC_1$. This is illus-

**Table 4. Simulation results for the 10% percentile (Q10) of the distribution of the recurrent parent genome (RPG) and total number of marker data points (MDP) required in a backcross program to introgress one target allele, for increasing and decreasing population sizes $n_t$. Values for MDP are rounded to multiples of ten.**

| | Ratio $n_1 : n_2 : n_3$ | | | | | | |
|---|---|---|---|---|---|---|---|
| Generation | 3:2:1 | 1:1:1 | 2:3:4 | 1:2:3 | 1:3:5 | 1:2:4 | 1:3:9 |
| | Number of individuals $n_t$ | | | | | | |
| $BC_1$ | 150 | 100 | 66 | 50 | 33 | 43 | 23 |
| $BC_2$ | 100 | 100 | 100 | 100 | 100 | 86 | 68 |
| $BC_3$ | 50 | 100 | 133 | 150 | 166 | 171 | 209 |
| | Q10 (%) of the RPG | | | | | | |
| Two-stage selection | | | | | | | |
| $BC_1$ | 81.6 | 80.7 | 80.0 | 79.3 | 78.3 | 78.9 | 77.1 |
| $BC_2$ | 93.8 | 93.6 | 93.2 | 93.1 | 92.8 | 92.8 | 91.9 |
| $BC_3$ | 97.3 | 97.4 | 97.4 | 97.4 | 97.4 | 97.4 | 97.3 |
| Three-stage selection | | | | | | | |
| $BC_1$ | 72.8 | 73.1 | 73.7 | 73.1 | 72.3 | 72.8 | 71.4 |
| $BC_2$ | 90.5 | 90.0 | 89.5 | 88.8 | 88.1 | 88.3 | 86.9 |
| $BC_3$ | 97.0 | 97.1 | 97.1 | 97.0 | 96.9 | 97.0 | 96.7 |
| Four-stage selection | | | | | | | |
| $BC_1$ | 71.2 | 71.6 | 72.0 | 72.0 | 71.5 | 71.9 | 71.1 |
| $BC_2$ | 88.5 | 88.3 | 88.0 | 87.4 | 87.0 | 87.0 | 86.0 |
| $BC_3$ | 96.5 | 96.7 | 96.8 | 96.8 | 96.6 | 96.6 | 96.3 |
| | Number of MDP required in total | | | | | | |
| Two-stage selection | | | | | | | |
| $BC_1$ | 6010 | 4000 | 2680 | 2000 | 1370 | 1720 | 920 |
| $BC_2$ | 7120 | 5180 | 3910 | 3290 | 2720 | 2850 | 1900 |
| $BC_3$ | 7240 | 5430 | 4280 | 3720 | 3230 | 3380 | 2650 |
| Three-stage selection | | | | | | | |
| $BC_1$ | 750 | 590 | 450 | 370 | 290 | 340 | 250 |
| $BC_2$ | 1740 | 1390 | 1070 | 930 | 740 | 790 | 580 |
| $BC_3$ | 1930 | 1820 | 1690 | 1660 | 1620 | 1680 | 1760 |
| Four-stage selection | | | | | | | |
| $BC_1$ | 480 | 430 | 350 | 300 | 260 | 290 | 240 |
| $BC_2$ | 1070 | 910 | 740 | 640 | 540 | 570 | 440 |
| $BC_3$ | 1310 | 1390 | 1400 | 1400 | 1400 | 1450 | 1500 |

trated by the low value of $r = 0.38$ for the carrier-chromosome in $BC_3$ under four-stage selection (Table 5). Because of a high selection pressure in early backcross generations, almost all markers on the carrier-chromosome are homozygous for the recurrent parent allele. Hence, they describe only poorly the differences in RPG that still do exist between the individuals.

Preferential selection of individuals with high RPG content on the carrier chromosome in $BC_1$ and $BC_2$ results in a lower overall RPG content, because the non-carrier chromosomes, on which only a reduced selection pressure is applied, form the major part of the genome. In three- or four-stage selection, non-carrier chromosomes selection is less intensive in $BC_1$. Therefore the corresponding value for $r$ in $BC_3$ is distinctly higher. This results in efficient $BC_3$ selection, which compensates for the lower RPG values derived from $BC_1$ and $BC_2$.

## Number of Marker Data Points Required

The major portion of MDP required in a two-stage selection program with constant $n_t$ is required in generation $BC_1$ (Table 4). Its expectation is $mn_1/2$, where $m$ is the total number of marker loci. A reduction in $n_1$ results in a proportional reduction of the MDP required in generation $BC_1$ (Table 4). In advanced backcross generations, many marker loci are already fixed for the recurrent parent allele. This results in a substantial MDP decrease if larger population sizes are used in advanced backcross generations instead of $BC_1$ or $BC_2$.

In the second selection step of three-stage selection, only the flanking markers are analyzed in all carriers of the target allele. Hence, instead of $mn_1/2$ MDP only $n_1$ MDP are required by expectation. Subsequently, analysis of the remaining marker loci in the third selection step requires $(m - 2)a$ MDP for the $a$ preselected individuals. This smaller number of MDP in generation $BC_1$ results in the observed overall MDP reduction (up to 50%) (Table 4). In four-stage selection, a further MDP reduction is achieved by investigating only the $i$ non-flanking markers on the carrier chromosome in the third selection step. This requires $ia$ MDP instead of $(m - 2)a$. The whole marker set is only analyzed on

**Table 5. Factors determining response to marker-assisted selection for the recurrent parent genome (RPG) in backcrossing: $\sigma$ = standard deviation of the RPG and $r$ = correlation between the proportion of recurrent parent alleles at marker loci and the proportion of recurrent parent alleles across the whole genome are given for the carrier chromosome, the non-carrier chromosomes, and for all chromosomes. Only individuals carrying the target allele are considered.**

| | | Standard deviation $\sigma$ | | | Correlation $r$ | | |
|---|---|---|---|---|---|---|---|
| $n_1 : n_2 : n_3$ | Chromosomes | $BC_1$ | $BC_2$ | $BC_3$ | $BC_1$ | $BC_2$ | $BC_3$ |
| Two-stage selection | | | | | | | |
| 100:100:100 | carrier | 0.125 | 0.112 | 0.067 | 0.964 | 0.947 | 0.894 |
| | non-carrier | 0.055 | 0.029 | 0.013 | 0.911 | 0.813 | 0.642 |
| | all | 0.051 | 0.027 | 0.012 | 0.913 | 0.814 | 0.681 |
| 50:100:150 | carrier | 0.125 | 0.117 | 0.068 | 0.964 | 0.948 | 0.899 |
| | non-carrier | 0.055 | 0.031 | 0.013 | 0.911 | 0.830 | 0.669 |
| | all | 0.051 | 0.029 | 0.013 | 0.913 | 0.830 | 0.700 |
| 150:100:50 | carrier | 0.125 | 0.113 | 0.067 | 0.964 | 0.947 | 0.896 |
| | non-carrier | 0.055 | 0.028 | 0.012 | 0.911 | 0.807 | 0.642 |
| | all | 0.051 | 0.026 | 0.012 | 0.913 | 0.807 | 0.683 |
| Three stage-selection | | | | | | | |
| 100:100:100 | carrier | 0.125 | 0.096 | 0.055 | 0.964 | 0.918 | 0.698 |
| | non-carrier | 0.055 | 0.041 | 0.020 | 0.910 | 0.884 | 0.795 |
| | all | 0.051 | 0.037 | 0.019 | 0.913 | 0.877 | 0.795 |
| Four stage-selection | | | | | | | |
| 100:100:100 | carrier | 0.125 | 0.088 | 0.036 | 0.964 | 0.887 | 0.380 |
| | non-carrier | 0.055 | 0.043 | 0.024 | 0.911 | 0.896 | 0.883 |
| | all | 0.051 | 0.039 | 0.022 | 0.913 | 0.887 | 0.830 |

the *b* individuals preselected in the third step, which requires $(m - 2 - i)b$ MDP.

## Transferability to Other Situations in Breeding

Like simulations in general, the results presented in this study depend on the underlying model. In the present context, simulation results are influenced by (i) the theoretical assumptions underlying the simulation of the meiotic recombination and (ii) the choice of genetic and dimensioning parameters.

We chose the map of Schön et al. (1994) because it represents a typical linkage map used in breeding programs. To investigate the robustness of our results with regard to the target allele position, we analyzed two additional scenarios.

1. The target locus was located on Chromosome 7, with a distance of 40 cM from the telomere.
2. The target locus was assigned to a random position on the genome in each repetition of the simulation. While the absolute Q10 values under these scenarios differed slightly from the results presented here, the general trends were the same (data not shown).

Simulations with varying linkage maps demonstrated that an average marker density higher than 20 cM results only in a marginal increase of Q10 values, but requires a substantially larger number of MDP (Frisch et al., 1998). In generation $BC_1$ and $BC_2$, a chromosome only consists of several segments of different origin (for a chromosome of length *l*, the expected number of segments in $BC_1$ is $l + 1$). Hence, the bottleneck limiting marker-assisted selection in early backcross generations is the number of chromosome segments itself, not the number of markers used for monitoring the composition of the chromosomes.

With a linkage map with equally spaced markers (Frisch et al., 1998), smaller population sizes and fewer MDP were required than with the linkage map underlying this study, which has regions of 60 or 80 cM length not covered by markers. For example, with a linkage map uniformly covered by markers, a saving of four backcross generations can be achieved with population sizes that resulted in a saving of three backcross generations with the linkage map used in this study (Frisch et al., 1998). This shows that an equally covered linkage map is mandatory for obtaining maximum RPG values in $BC_2$ and $BC_3$.

The differences in Q10 and MDP values between the selection strategies are caused by a different treatment of carrier and non-carrier chromosomes. Hence, the ratio between carrier and non-carrier chromosomes determines the different outcome of the selection strategies. The amount of reduction in the required number of MDP reported here is specific for 10 chromosomes and map length of 16 Morgan. In crops with genomes consisting of less than 10 chromosomes, the differences are expected to be smaller, because the ratio between carrier and non-carrier chromosomes increases. For more than 10 chromosomes, the proportion of genome on the non-carrier chromosomes increases and, consequently, the differences between the selection strategies are expected to be greater.

The presented results should cover a wide range of gene introgression programs in crops with $2x = 20$ and also $2x = 18$ chromosomes, such as maize or sugar beet (*Beta vulgaris* L.). For different linkage maps, our simulation software PLABSIM (Frisch et al., 1999b) can be used for conducting simulations to compare the effect of selection strategies or breeding designs in marker-assisted backcrossing.

## Design of Marker-Assisted Backcross Programs

Tanksley et al. (1989) stated that a sufficiently high proportion of the RPG is recovered after three generations of marker-assisted backcrossing. Hospital et al. (1992) expected a saving of two backcross generations because of marker-assisted background selection. This is in accordance with our simulations, resulting in a saving of two to four backcross generations in the transfer of a single target allele (Table 3).

The backcross procedure can be terminated after four instead of six backcross generations even with small population sizes and a limited number of MDP (Table 2). This demonstrates that marker technology can be advantageous even when the resources in a breeding program are limited. A shortening from six to three backcross generations can be regarded as a realistic goal for practical breeders, because moderate population sizes and number of MDP are required, and the breeding program is two times faster than it is without markers. As demonstrated by our results, marker-assisted selection has the potential to reach in generation $BC_3$ the same level of RPG as reached in $BC_7$ without use of markers. However, large numbers of MDP are required to unlock this potential. With the marker systems presently available, this application is yet unrealistic or at least not economic.

In generations $BC_1$ and $BC_2$, two-stage selection is superior to three- and four-stage selection because it reaches a larger RPG proportion with a given population size (Table 3). Thus, two-stage selection seems appropriate in two-generation backcross programs with limited population size. Furthermore, it can be applied without information about the marker linkage map and, hence, is the only option for application in generation $BC_1$, if no marker linkage map is available.

An increasing population size $n_t$ is preferable over a constant population size in a two-stage selection program, because the number of marker analyses is reduced without reducing the Q10 values. Limits for varying $n_t$ are practical restrictions for handling large values of $n_3$ and the risk of loosing the target allele in $BC_1$ with low values of $n_1$. With probability $P = 1/2^{n_1}$, none of the $n_1$ backcross individuals carries the target allele. Hence, a minimum of 15 to 20 individuals per generation should be produced to obtain with almost certainty at least one carrier of the target allele.

Reduction of the linkage drag is one of the main goals in marker-assisted backcrossing (Tanskley et al., 1989). Theoretical results (Stam and Zeven, 1981) show that the donor segment attached to the target allele remains surprisingly large in backcrossing without marker-assisted selection even in advanced backcross generations. In introgression of target alleles from unadapted germplasm, linkage drag is the main cause for the differences between the recipient line and the converted line. Tightly linked flanking markers can be used for a substantial reduction of the linkage drag. Individuals with recombination between tightly linked loci have a low frequency in backcross populations, but may not be selected by applying two-stage selection. Hence, if reduction of the linkage drag has high priority, three- or four-stage selection should be applied. This avoids the necessity of additional backcross generations at the end of the breeding program to ascertain detection of a recombination event between tightly linked flanking markers and the target locus.

While three- and four-stage selection yield considerably lower RPG values in $BC_2$ than two-stage selection, the slightly lower Q10 values reached in $BC_3$ can be compensated by larger population sizes $n_3$. Thus, without restrictions on $n_3$, applying three- or four stage selection in three-generation backcross programs results in a reduction of the required number of MDP by as much as 50 or 75% (Table 3). They combine economic marker use with the possibility to efficiently reduce the linkage drag.

In a separate paper (Frisch et al., 1999a), we give equations for calculating the minimal population size for obtaining at least one carrier of the target allele homozygous for the recurrent parent allele at one or both flanking markers. The required population size depends on (i) the map distances between the flanking markers and the target allele and (ii) the chosen probability of success. These results can be used for the design of efficient three- and four-stage selection backcross programs in marker-assisted background selection.

## ACKNOWLEDGMENTS

The financial support from fellowships by KWS Kleinwanzlebener Saatzucht AG, Einbeck, Germany, and Pioneer Hi-Bred Intl. Inc., Johnston, IA, USA, to M. Frisch is gratefully acknowledged.

## REFERENCES

Allard, R.W. 1960. Principles of plant breeding. Wiley, New York.

Crosby, J.L. 1973. Computer simulation in genetics. Wiley, New York.

Falconer, D.S., and T.F. Mackay. 1996. Introduction of quantitative genetics. Longman Group Limited, Harlow, UK.

Frisch, M., M. Bohn, and A.E. Melchinger. 1998. Markerdichte und Anzahl benötigter Markeranalysen in markergestützten Rückkreuzungs-programmen. Vorträge fr Pflanzenzüchtung 42:1–3.

Frisch, M., M. Bohn, and A.E. Melchinger. 1999a. 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. 1999b. PLABSIM: Software for simulations of marker-assisted backcrossing. J. Heredity (In press).

Haldane, J.B.S. 1919. The combination of linkage values and the calculation of distance between the loci of linkage factors. J. Genet. 8:299–309.

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 132:1119–1210.

Lander, E.S., P. Green, J. Abrahamson, A. Barlow, M.J. Daly, S.E. Lincoln, and L. Newburg. 1987. MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1:174–181.

Melchinger, A.E. 1990. Use of molecular markers in breeding for oligogenic disease resistance. Plant Breeding 104:1–19.

Openshaw, S.J., S.G. Jarboe, and W.D. Beavis. 1994. Marker-assisted selection in backcross breeding. *In* Proceedings of the Symposium "Analysis of Molecular Marker Data", Corvallis, OR. 5–6 Aug. 1994. Am. Soc. Hortic. Sci. and Crop Sci. Soc. Am.

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

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

Schön, C.C., A.E. Melchinger, J. Boppenmaier, E. Brunklaus-Jung, R.G. Herrmann, and J.F. Seitzer. 1994. RFLP mapping in maize: Quantitative trait loci affecting testcross performance of elite European flint lines. Crop Sci. 34:378–389.

Stam, P. 1979. Interference in genetic crossing over and chromosome mapping. Genetics 92:873–594.

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

Stuber, C.W. 1995. Mapping and manipulating quantitative traits in maize. Trends Genetics 11:477–481.

Tanksley, S.D., N.D. Young, A.H. Patterson, and M.W. Bonierbale. 1989. RFLP mapping in plant breeding: new tools for an old science. Bio/Technology 7:257–263.

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