Theor Appl Genet (2011) 123:251–260
DOI 10.1007/s00122-011-1581-0

ORIGINAL PAPER

Selection strategies for marker-assisted backcrossing with high-throughput marker systems

Eva Herzog · Matthias Frisch

Received: 10 September 2010 / Accepted: 23 March 2011 / Published online: 8 April 2011

Abstract Application of marker-assisted backcrossing for gene introgression is still limited by the high costs of marker analysis. High-throughput (HT) assays promise to reduce these costs, but new selection strategies are required for their efficient implementation in breeding programs. The objectives of our study were to investigate the properties of HT marker systems compared to single-marker (SM) assays, and to develop optimal selection strategies for marker-assisted backcrossing with HT assays. We employed computer simulations with a genetic model consisting of 10 chromosomes of 160 cM length to investigate the introgression of a dominant target gene. We found that a major advantage of HT marker systems is that they can provide linkage maps with equally spaced markers, whereas the possibility to provide linkage maps with high marker densities smaller than 10 cM is only of secondary use in marker-assisted backcrossing. A three-stage selection strategy that combines selection for recombinants at markers flanking the target gene with SM assays and genome-wide background selection with HT markers in the first backcross generation was more efficient than genome-wide background selection with HT markers alone. Selection strategies that combine SM and HT assays were more efficient than genome-wide background selection with HT assays alone. This result was obtained for a broad range of cost ratios of HT and SM assays. A further considerable reduction of the costs could be achieved if the population size in the first backcross generation was twice the population size in generations BC_2 and BC_3 of a three-generation backcrossing program. We conclude that selection strategies combining SM and HT assays have the potential to greatly increase the efficiency and flexibility of marker-assisted backcrossing.

Communicated by Y. Xu.

E. Herzog · M. Frisch (✉)
Institute of Agronomy and Plant Breeding II,
Justus Liebig University, 35392 Giessen, Germany
e-mail: matthias.frisch@uni-giessen.de

Introduction

Marker-assisted backcrossing is used for transferring genes which are responsible for favorable agronomic traits from a donor line into the genome of a recipient line. Using molecular markers for selection against the genetic background of the donor can reduce the time and resources required for gene introgression. Although background selection has become a standard tool in plant breeding, the high costs of marker analysis still limit its use in practice and are the crucial factor for the experimental designs of gene introgression programs (Collard and Mackill 2008). These designs depend on the number of target genes to be transferred, the employed marker map, and the number of generations available for the gene introgression. Computer simulations are a robust tool for optimizing the design parameters of a marker-assisted backcrossing program before implementing it in practice (Prigge et al. 2008).

The design of marker-assisted backcrossing programs was studied with respect to the introgression of single dominant and recessive genes (Hospital et al. 1992; Frisch et al. 1999a, b; Frisch and Melchinger 2001a), two genes (Frisch and Melchinger 2001b), and favorable alleles at quantitative trait loci (Hospital and Charcosset 1997; Bouchez et al. 2002). More recently, marker-assisted backcrossing for developing libraries of near-isogenic lines was studied (Peleman and van der Voort 2003; Falke et al. 2009; Falke and Frisch 2011). These studies have mainly focused on optimizing the number of genotyped individuals as well as the positions and density of background selection markers with respect to the required number of marker data points. The optimizations have been carried out assuming marker systems in which each marker locus is analyzed in a separate assay (cf. Prigge et al. 2009). We refer to such systems as single-marker (SM) systems. Typical examples are the simple sequence repeat (SSR) and the restriction fragment length polymorphism (RFLP) marker systems.

Recently, high-throughput (HT) marker systems based on single nucleotide polymorphisms (SNPs) have been developed. Due to the high level of automation of systems such as DNA chips, they allow for cheap and fast analysis of hundreds of marker loci in a single analysis step (Gupta et al. 2001; Syvänen et al. 2005). HT marker systems have been developed for crops (Ragot and Lee 2007) and are becoming the marker systems of choice in commercial breeding programs of many economically important crops.

The crucial difference between HT and SM marker systems is that with SM marker systems, only those markers are analyzed in advanced backcross generations which were not already fixed for the recipient alleles in earlier generations. In contrast, with HT marker systems, the entire panel of markers used in a gene introgression program needs to be analyzed also for individuals of advanced backcross generations, even if 80 or 90% of these markers have already been fixed for the recipient alleles. To our knowledge, no study investigating the implications of this property on the efficiency of marker-assisted backcrossing is available. The combination of SM marker systems for the reduction of the chromosome segment attached to the target gene and HT markers for genome-wide background selection promises to further enhance selection efficiency in marker-assisted backcrossing and is not yet investigated.

The objectives of our simulation study were to (1) compare the relative costs of genome-wide background selection with SM and HT marker systems for different cost ratios of HT:SM markers, (2) compare the efficiency of equally spaced and randomly distributed markers with respect to the recovery of the recipient genome, (3) develop selection strategies combining SM and HT assays, which are more efficient than genome-wide background selection with SM or HT assays alone.

Simulations

A genetic model with ten equally sized chromosomes of 160 cM length was used for the simulations. Its genome length of 1,600 cM is similar to that of published linkage maps of maize (cf. Schön et al. 1994). Markers for genome-wide background selection were assumed to be (a) randomly distributed in the genome or (b) equally spaced. Average marker distances (randomly distributed markers) or marker distances (equally spaced markers) between two adjacent marker loci of $\delta_{\mathrm{GW}} = 2$, 5, 10, 20 cM were investigated. For equally spaced markers, two markers were located at the telomeres of each chromosome. One dominant target gene to be introgressed was located on Chromosome 1. It was 81, 82.5, 85, and 90 cM distant from the telomere for linkage maps with $\delta_{\mathrm{GW}} = 2$, 5, 10, 20 cM, respectively. Flanking markers for selection against the donor chromosome segment attached to the target gene were located on both sides of the target gene. The distances between target gene and each flanking marker were $\delta_{\mathrm{F}} = 5$, 10, 20, 30, 40 cM.

The investigated breeding scheme started with the cross of two homozygous parents (donor and recipient), which were polymorphic at all loci. The recipient carried the desirable alleles at all loci of the genome except for the target locus, while the donor carried the desirable allele at the target locus. The donor and recipient were crossed to create an F_1 individual, which was backcrossed to the recipient. From the BC_1 population of size n_1, one individual was selected with two- or three-stage selection, as described below, and backcrossed to the recipient. This procedure was repeated for t backcross generations.

Two-stage selection consisted of pre-selection of carriers of the target gene in the first selection step. The pre-selected individuals were subjected to genome-wide background selection in the second step. A selection index $i = \sum_m x_m$ was constructed, where summation is over markers and $x_m = 1$ if a marker is homozygous for the recipient allele. A plant with the highest value of i was selected and backcrossed to the recipient. Two-stage selection was carried out with SM and HT assays. For SM assays, only those markers were analyzed in advanced backcross generations which were not yet fixed for the recipient allele in the non-recurrent parent.

Three-stage selection combined selection for recombinants between the target gene and its two flanking markers, genotyped with SM assays, and genome-wide background selection with HT assays. It consisted of (1) selection for the target gene followed by (2) pre-selection with flanking markers and (3) genome-wide selection with background markers. For selection step (2), a selection index f was created, which took the values 0, 1, or 2, depending on whether recombination occurred between the target gene and none, one, or both flanking markers, respectively. On the basis of f, pre-selection of individuals was carried out according to one of two decision rules. Either (a) individuals with $f \geq 1$ were selected, or (b) all individuals having the maximum observed score of f ($f = \max$) were selected.

Four series of simulations were carried out with software Plabsoft (Maurer et al. 2008), assuming no interference in crossover formation. Each simulation was replicated 10,000 times in order to reduce sampling effects and to obtain results with high numerical accuracy and a small standard error. The 10% quantile (Q10) of the distribution of recipient genome (in percent) was determined in the last backcross generation to measure the success of a marker-assisted backcrossing program with respect to restoring the genome of the recipient. The number of SM and HT assays was determined as a measure for the costs of a marker-assisted backcrossing program.

In the first series of simulations, the population size n_t (constant across all backcross generations BC_t, $t = 1, \ldots, 3$) and the number of marker assays were determined which were required to reach Q10 values of 93, 94, 95, 96, 97, 98%, respectively. For 93–96%, we investigated two-generation backcrossing programs, and for 96–98% three-generation backcrossing programs. Two-stage selection with either SM or HT assay or a combination of both systems (HT in backcross generation BC_1 and SM in the following backcross generations) was carried out for linkage maps with $\delta_{\mathrm{GW}} = 5$, 10, 20 cM.

In the second series of simulations, two-stage selection with HT assays was carried out. Background selection markers were either equally spaced or randomly distributed with $\delta_{\mathrm{GW}} = 2$, 5, 10, 20 cM. We considered three backcross generations and constant values of n_t ranging from 40 to 200 individuals.

In the third series of simulations, three-stage selection was carried out either in backcross generation BC_1 or BC_3. In the remaining two generations, two-stage selection with HT assays was carried out. The flanking markers for three-stage selection had distances of $\delta_{\mathrm{F}} = 5$, 10, 20, 30, 40 cM from the target gene and individuals with $f \geq 1$ were selected for genome-wide analysis with HT assays. Distances between genome-wide background selection markers were $\delta_{\mathrm{GW}} = 5$ cM. In the generations with two-stage selection, we investigated population sizes from $n_t = 40$ to 200. In the generation with three-stage selection, these population sizes were multiplied by a factor $m = 1$, 2, 5.

In the fourth series of simulations, three-stage selection was carried out in backcross generations BC_1 and BC_2. Marker distances of $\delta_{\mathrm{GW}} = 5$ cM and $\delta_{\mathrm{F}} = 20$ cM were employed. Individuals with $f \geq 1$ were pre-selected for genome-wide analysis in backcross generation BC_1, while only individuals having the highest observed number of recombinations between target gene and flanking markers ($f = \max$) were pre-selected in backcross generation BC_2. In backcross generation BC_3, two-stage selection was carried out with HT assays. We investigated population sizes from $n_t = 40$ to 200 for generations BC_2 and BC_3. In backcross generation BC_1, these population sizes were multiplied by the factor $m = 1$, 2, 5.

For comparing the costs of marker-assisted backcrossing programs with different selection strategies, linkage maps, and population sizes, the numbers of SM and HT assays required for the entire backcrossing program were assessed. For SM analyses, only those markers not yet fixed for the recipient allele in the non-recurrent parent of a backcross population were considered. For HT analyses, the number of assays was the same as the number of individuals subjected to genome-wide background selection. Calculation of costs was based on five cost ratios of one HT assay (corresponding to all HT marker loci on the linkage map) compared to one SM assay (corresponding to one SM locus). Cost ratios of HT:SM of 200:1, 100:1, 50:1, 20:1, 10:1 were investigated. For example, a cost ratio HT:SM of 100:1 corresponds to a price of 200€ for analyzing all SNP background marker loci with a DNA chip, and 2€ for analyzing one SSR marker locus. Comparisons were carried out to compare (a) the costs of two-stage selection with HT assays to those of two-stage selection with SM assays, (b) the costs of two-stage selection with HT assays in generation BC_1 and SM assays in BC_2 and BC_3 to those of two-stage selection with HT assays in all backcross generations, (c) the costs of three-stage selection in BC_1 to those of two-stage selection with HT assays in all generations. For (a) the costs of SM assays were set 1 and the relative costs of HT assays were determined, for (b) the costs of using HT assays in all backcross generations were set 1 and the relative costs of the strategy combing HT and SM were determined, and for (c) the costs of two-stage selection were set 1 and the relative costs of three-stage selection were determined.

Results

For two-stage selection, HT assays were considerably more expensive (up to factor 4.77) than SM assays for scenarios with high relative costs of HT markers (200:1, 100:1, and 50:1) in combination with large marker distances and/or large attempted Q10 values (Table 1). For scenarios with small marker distances and/or low relative cost ratios of HT:SM assays and low attempted Q10 values, HT assays were cheaper. To reach a Q10 value of 96% in two generations, the number of required marker assays was 9–14 times greater than those required to reach the same Q10 value in three generations. The increase in the required number of marker assays, which accompanied the shortening of a backcrossing program from three to two generations, was greater for SM than for HT marker systems.

For high cost ratios of HT:SM markers (200:1, 100:1, and 50:1) and large marker distances, combining HT assays

Table 1 Relative costs of a gene introgression program using HT assays in generations BC_1 to BC_3 (HT[$BC_{1–3}$]) compared to using SM assays in BC_1 to BC_3 (SM[$BC_{1–3}$]) depending on the cost ratio of HT:SM assays

δ_{GW}	Q10 (%)	No. of BC generations	n_t	No. of assays		Cost ratio HT:SM				
						Relative costs				
				HT[$BC_{1–3}$]	SM[$BC_{1–3}$]	200:1	100:1	50:1	20:1	10:1
20 cM ($n_m = 90$)	93	2	44	44	2,643	3.33	1.66	0.83	0.33	0.17
	94	2	72	72	4,260	3.38	1.69	0.85	0.34	0.17
	95	2	133	133	7,737	3.44	1.72	0.86	0.34	0.17
	96	2	291	291	16,583	3.51	1.75	0.88	0.35	0.18
	96	3	17	26	1,158	4.40	2.20	1.10	0.44	0.22
	97	3	30	45	1,975	4.56	2.28	1.14	0.46	0.23
	98	3	70	105	4,401	4.77	2.39	1.19	0.48	0.24
10 cM ($n_m = 170$)	93	2	39	39	4,442	1.76	0.88	0.44	0.18	0.09
	94	2	62	62	6,960	1.78	0.89	0.45	0.18	0.09
	95	2	110	110	12,141	1.81	0.91	0.45	0.18	0.09
	96	2	222	222	24,050	1.85	0.92	0.46	0.18	0.09
	96	3	16	24	2,070	2.32	1.16	0.58	0.23	0.12
	97	3	26	39	3,258	2.39	1.20	0.60	0.24	0.12
	98	3	53	80	6,382	2.49	1.25	0.62	0.25	0.12
5 cM ($n_m = 330$)	93	2	38	38	8,406	0.90	0.45	0.23	0.09	0.05
	94	2	60	60	13,077	0.92	0.46	0.23	0.09	0.05
	95	2	104	104	22,292	0.93	0.47	0.23	0.09	0.05
	96	2	206	206	43,361	0.95	0.48	0.24	0.10	0.05
	96	3	15	23	3,780	1.19	0.60	0.30	0.12	0.06
	97	3	25	38	6,094	1.23	0.62	0.31	0.12	0.06
	98	3	50	75	11,719	1.28	0.64	0.32	0.13	0.06

Two-stage selection, n_m equally spaced background selection markers with distances δ_{GW}, and population sizes n_t were used to recover Q10 target values of 93–98% in two or three backcross generations

in generation BC_1 with SM assays in generations BC_2 and BC_3 for genome-wide background selection was cheaper (up to 60%) than using HT assays alone (Table 2). This cost reduction was more pronounced for three-generation than two-generation backcross programs.

To reach a given Q10 value with randomly distributed background selection markers, linkage maps with two to four times more markers are required than with equally spaced markers of marker distances $\delta_{GW} = 20$ or 10 cM (Table 3). With equally spaced markers and $\delta_{GW} = 5$ cM, approximately the same Q10 values were reached as with randomly distributed markers and $\delta_{GW} = 2$ cM. A decrease in the distance between equally distributed markers from $\delta_{GW} = 10$ to 5 cM resulted in only marginally greater Q10 values in generation BC_3. No difference in the Q10 values was observed for $\delta_{GW} = 5$ and 2 cM.

With three-stage selection combining SM and HT assays in generation BC_1, the flanking marker distance δ_F had only marginal influence on the recovered genome-wide Q10 values (Table 4). For population sizes $n_2 = n_3 < 100$ in generations BC_2 and BC_3, a substantial increase of the Q10 values was observed, if in generation BC_1 larger populations $n_1 > n_2 = n_3$ were employed. Doubling the population size in generation BC_1 ($n_1 = mn_2 = mn_3$, $m = 2$) had approximately the same effect on the Q10 values as increasing a constant population size by about 20 individuals ($n_1' = n_2' = n_3' = n_2 + 20$). The combination of doubled population sizes in generation BC_1 and small flanking marker distances δ_F resulted in less required HT assays at the expense of more required SM assays to reach a certain Q10 value, compared to backcrossing programs with constant population sizes across generations.

Three-stage selection in generation BC_3 recovered similar Q10 values as three-stage selection in generation BC_1 for all combinations of n_t and m. However, more HT assays were required (data not shown).

Three-stage selection in generations BC_1 and BC_2 required more SM assays but less HT assays compared to three-stage selection only in generation BC_1 for all combinations of n_t and m (Table 5). For population sizes smaller than 100, slightly lower Q10 values were recovered.

Three-stage selection combining SM and HT assays in generation BC_1 of a three-generation backcrossing program was cheaper than two-stage selection with HT assays for all

Table 2 Relative costs of a gene introgression program using HT assays in backcross generation BC_1 and SM assays in backcross generations BC_2 and BC_3 (HT[BC_1], SM[$BC_{2,3}$]) compared to using HT assays in all backcross generations (HT[$BC_{1–3}$], data presented in Table 1) depending on the cost ratio of HT:SM assays

δ_{GW}	Q10 (%)	No. of BC generations	n_t	No. of assays		Cost ratio HT:SM				
						Relative costs				
				HT[BC_1]	SM[$BC_{2,3}$]	200:1	100:1	50:1	20:1	10:1
20 cM ($n_m = 90$)	93	2	44	22	664	0.58	0.65	0.80	1.25	2.01
	94	2	72	36	1,019	0.57	0.64	0.78	1.21	1.92
	95	2	133	67	1,749	0.57	0.64	0.77	1.16	1.82
	96	2	291	146	3,490	0.56	0.62	0.74	1.10	1.70
	96	3	17	9	393	0.42	0.50	0.65	1.10	1.86
	97	3	30	15	624	0.40	0.47	0.61	1.03	1.72
	98	3	70	35	1,250	0.39	0.45	0.57	0.93	1.52
10 cM ($n_m = 170$)	93	2	39	20	1,130	0.66	0.80	1.09	1.96	3.41
	94	2	62	31	1,686	0.64	0.77	1.04	1.86	3.22
	95	2	110	55	2,787	0.63	0.75	1.01	1.77	3.03
	96	2	222	111	5,183	0.62	0.73	0.97	1.67	2.83
	96	3	16	8	712	0.48	0.63	0.93	1.82	3.30
	97	3	26	13	1,051	0.47	0.60	0.87	1.68	3.03
	98	3	53	27	1,880	0.46	0.57	0.81	1.51	2.69
5 cM ($n_m = 330$)	93	2	38	19	2,129	0.78	1.06	1.62	3.30	6.10
	94	2	60	30	3,194	0.77	1.03	1.56	3.16	5.82
	95	2	104	52	5,138	0.75	0.99	1.49	2.97	5.44
	96	2	206	103	9,359	0.73	0.95	1.41	2.77	5.04
	96	3	15	8	1,300	0.63	0.91	1.48	3.17	6.00
	97	3	25	13	1,969	0.60	0.86	1.38	2.93	5.52
	98	3	50	25	3,479	0.57	0.80	1.26	2.65	4.97

Two-stage selection, n_m equally spaced background selection markers with distances δ_{GW}, and population sizes n_t were used to recover Q10 target values of 93–98% in two or three backcross generations

Table 3 Q10 values recovered in generation BC_3 for constant population sizes n_t in generations BC_1 to BC_3 and equally spaced or randomly distributed markers ($\delta_{GW} = 2, 5, 10, 20$ cM) applying two-stage selection with HT assays

δ_{GW} (cM)	Generation	Equally spaced markers, n_t					Randomly distributed markers, n_t				
		40	80	120	160	200	40	80	120	160	200
20	BC_1	79.7	81.4	82.4	83.0	83.4	78.0	79.6	80.5	80.9	81.4
	BC_2	92.8	94.2	94.9	95.3	95.6	91.3	92.6	93.2	93.6	94.0
	BC_3	97.4	98.1	98.4	98.6	98.7	96.4	97.0	97.3	97.4	97.5
10	BC_1	79.9	81.7	82.7	83.3	83.8	78.8	80.5	81.3	81.9	82.3
	BC_2	93.0	94.5	95.2	95.6	95.9	91.9	93.4	94.1	94.4	94.8
	BC_3	97.6	98.4	98.7	98.9	99.0	97.0	97.8	98.1	98.3	98.4
5	BC_1	80.0	81.7	82.7	83.4	83.9	79.3	81.0	81.9	82.5	83.0
	BC_2	93.1	94.5	95.3	95.7	96.0	92.4	93.8	94.4	94.8	95.1
	BC_3	97.8	98.5	98.8	99.0	99.1	97.1	97.9	98.3	98.4	98.6
2	BC_1	80.0	81.8	82.8	83.4	83.8	79.8	81.5	82.5	83.1	83.7
	BC_2	93.2	94.6	95.3	95.7	96.0	93.0	94.4	95.1	95.5	95.9
	BC_3	97.8	98.5	98.8	99.0	99.1	97.7	98.5	98.7	98.9	99.1

investigated combinations of n_t with $m = 1$ and $m = 2$ (Fig. 1). The costs were ranging between 75.3–83.0% ($m = 1$) and 57.1–89.7% ($m = 2$) of the costs of two-stage selection. For $m = 5$, three-stage selection was only cheaper for cost ratios of HT:SM from 200:1 to 50:1. Three-stage selection with doubled population size ($m = 2$) in generation

Table 4 Q10 values recovered in generation BC_3 and number of required SM/HT assays for increased population sizes $n_1 = mn_t$ (m = 1, 2, 5; t = 2, 3) in generation BC_1 and equally spaced markers (δ_{GW} = 5 cM) applying three-stage selection (δ_F = 5, 10, 20, 30, 40 cM; $f \geq 1$) in generation BC_1 and two-stage selection in generations BC_2 and BC_3

m	δ_F (cM)	n_t								
		40	60	80	100	120	140	160	180	200
		Q10 (%) in generation BC_3								
1	40	97.8	98.2	98.5	98.7	98.8	98.9	99.0	99.0	99.1
	30	97.8	98.2	98.5	98.7	98.8	98.9	99.0	99.0	99.1
	20	97.8	98.2	98.6	98.7	98.9	99.0	99.0	99.1	99.1
	10	97.6	98.2	98.5	98.7	98.9	99.0	99.1	99.1	99.2
	5	97.4	98.0	98.3	98.6	98.8	98.9	99.0	99.1	99.1
2	40	98.0	98.4	98.6	98.8	98.9	99.0	99.0	99.1	99.2
	30	98.0	98.4	98.6	98.8	98.9	99.0	99.1	99.1	99.2
	20	98.0	98.5	98.7	98.8	99.0	99.1	99.1	99.2	99.2
	10	97.9	98.4	98.7	98.8	99.0	99.1	99.2	99.2	99.3
	5	97.7	98.2	98.6	98.8	98.9	99.1	99.1	99.2	99.3
5	40	98.2	98.6	98.8	98.9	99.0	99.1	99.1	99.2	99.2
	30	98.2	98.6	98.8	98.9	99.0	99.1	99.2	99.2	99.2
	20	98.2	98.6	98.8	99.0	99.1	99.1	99.2	99.2	99.3
	10	98.2	98.6	98.9	99.0	99.1	99.2	99.3	99.3	99.4
	5	98.1	98.6	98.8	99.0	99.1	99.2	99.3	99.3	99.4
		No. of required SM/HT assays								
1	40	40/49	60/73	80/98	100/123	120/148	140/172	160/197	180/222	200/246
	30	40/47	60/71	80/95	100/119	120/143	140/167	160/191	180/215	200/239
	20	40/45	60/68	80/91	100/114	120/137	140/160	160/183	180/206	200/229
	10	40/43	60/64	80/86	100/108	120/129	140/151	160/173	180/195	200/216
	5	40/44	60/64	80/84	100/104	120/125	140/146	160/166	180/187	200/208
2	40	80/58	120/88	160/117	200/146	240/176	280/205	320/235	360/265	400/294
	30	80/55	120/83	160/111	200/139	240/167	280/195	320/223	360/251	400/279
	20	80/51	120/77	160/103	200/129	240/155	280/181	320/207	360/233	400/259
	10	80/46	120/69	160/93	200/116	240/140	280/163	320/187	360/210	400/234
	5	80/44	120/65	160/86	200/108	240/130	280/152	320/174	360/196	400/218
5	40	200/86	300/130	400/174	500/218	600/261	700/305	800/349	900/393	1,000/437
	30	200/79	300/119	400/159	500/199	600/239	700/279	800/319	900/359	1,000/399
	20	200/69	300/104	400/140	500/175	600/210	700/245	800/280	900/315	1,000/350
	10	200/56	300/85	400/114	500/142	600/171	700/200	800/228	900/257	1,000/285
	5	200/48	300/73	400/98	500/122	600/147	700/172	800/196	900/221	1,000/245

BC_1 was the optimal selection strategy for reaching Q10 values of 98 and 99%. The only exception was the combination of a cost ratio of HT:SM assays of 10:1 and a desired Q10 value of 99%. In this case, constant population size over generations (m = 1) was optimal.

Discussion

HT marker systems

HT marker systems are expected to increase the cost-efficiency of marker-assisted backcrossing programs (Ragot and Lee 2007; Collard and Mackill 2008). However, previous studies on the efficiency of gene introgression programs have rarely taken differences between marker systems into account (Ribaut et al. 2002). In this study, we investigated the different properties of SM and HT marker systems and their effect on the efficiency of gene introgression. The simultaneous analysis of a large number of marker loci at comparatively low cost per individual marker locus is made feasible in HT assays (Syvänen et al. 2005). They, therefore, promise to be a powerful tool for marker-assisted background selection, especially when the expected number of required marker analyses is high. However, HT assays do not provide the possibility to

Table 5 Q10 values recovered in generation BC$_3$ and number of required SM/HT assays for increased population sizes $n_1 = mn_t$ (m = 1, 2, 5; t = 2, 3) in generation BC$_1$ and equally spaced markers (δ_{GW} = 5 cM) applying three-stage selection (δ_F = 5, 10, 20, 30, 40 cM) in generations BC$_1$ ($f \geq 1$) and BC$_2$ (f = max) and two-stage selection in generation BC$_3$

m	δ_F (cM)	n_t								
		40	60	80	100	120	140	160	180	200
		Q10 (%) in generation BC3								
1	20	97.3	98.0	98.4	98.7	98.8	98.9	99.0	99.1	99.1
2	20	97.6	98.3	98.6	98.8	99.1	99.0	99.1	99.2	99.2
5	20	98.0	98.5	98.8	99.0	99.1	99.1	99.2	99.3	99.3
		No. of required SM/HT assays								
1	20	58/30	86/45	115/61	143/77	172/93	200/109	228/125	256/141	285/157
2	20	97/36	146/55	194/73	242/92	291/111	338/131	387/151	436/169	484/189
5	20	217/54	325/82	433/111	542/138	650/168	758/196	866/224	974/252	1082/281

selectively analyze individual markers. In contrast to SM assays, all markers on the linkage map need to be analyzed for every backcross individual, even if a large proportion of markers has already been fixed for the recipient alleles, as is the case in advanced backcross generations.

Comparing two-generation with three-generation gene introgression programs showed that SM marker systems require relatively less assays in three-generation programs than HT assays. For example, in a two-generation gene introgression program with distances of genome-wide background selection markers of $\delta_{GW} = 20$ cM, both 44 HT and 2,643 SM assays resulted in a Q10 value of 93%, whereas in a three-generation program, 45 HT or 1,975 SM assays resulted in a Q10 value of 97% (Table 1). This effect is expected to be even more pronounced for background selection in higher backcross generations, and when background selection is carried out in selfing generations or during doubled haploid production. In line, using HT assays for genome-wide background selection in the first backcross generation, and SM assays in advanced backcross generations reduced the costs of marker analysis compared to using HT assays in all backcross generations (Table 2). Only 5–9% of all marker analyses in a three-generation backcross program fell upon backcross generation BC$_3$. The cost reduction compared to using HT assays in all backcross generations was consequently greater for three-generation than for two-generation programs. We conclude that HT assays are particularly suited for short gene introgression programs, while SM assays are efficient for marker-assisted background selection when in advanced generations already large percentages of the markers have been fixed for the recipient alleles.

Marker distance and distribution for genome-wide background selection

HT systems based on SNP markers are often analyzed with techniques employing marker numbers that are multiples of 96. We did not limit our investigations to these marker numbers for two reasons. Firstly, usually not all markers of such a set are polymorphic for a certain cross. Moreover, reduced representation sequencing approaches have recently emerged and a trend towards genotyping by sequencing can be observed. For these systems, fixed marker numbers are less relevant. Therefore, we focused in our study on marker distances δ_{GW}, but not on the fixed marker numbers employed by a certain marker technology. The results discussed below can be regarded as thresholds, which, if they are surpassed for two parental lines and a certain HT markers system, result in the presented Q10 values.

SNPs occur in abundance in plant genomes. Dense linkage maps with marker distances below 5 cM can consequently be established at reasonable costs. However, the effect of such dense markers on the recipient genome recovery has not yet been investigated. Decreasing the marker distances δ_{GW} below 10 cM had only marginal effect on the recipient genome recovery (Table 1). An explanation for this result is that on expectation one crossover per meiosis and chromatid occurs on a chromosome segment of length 1 M. In two- or three-generation backcrossing programs, the number of recombination events resulting in chromosome segments of different parental origin is therefore limited. To detect these chromosome segments and to efficiently identify the backcross individuals with the smallest percentage of donor genome, a marker distance of $\delta_{GW} = 10$ cM is sufficient. Smaller marker distances are not required, because the factor limiting selection response is not the precise estimation of the donor genome percentage, but the limited number of crossovers.

The difference in the Q10 values between equally spaced and randomly distributed markers was considerable for all marker distances δ_{GW} except 2 cM. Less than half the markers were required to reach a certain Q10 value with equally spaced markers compared with randomly

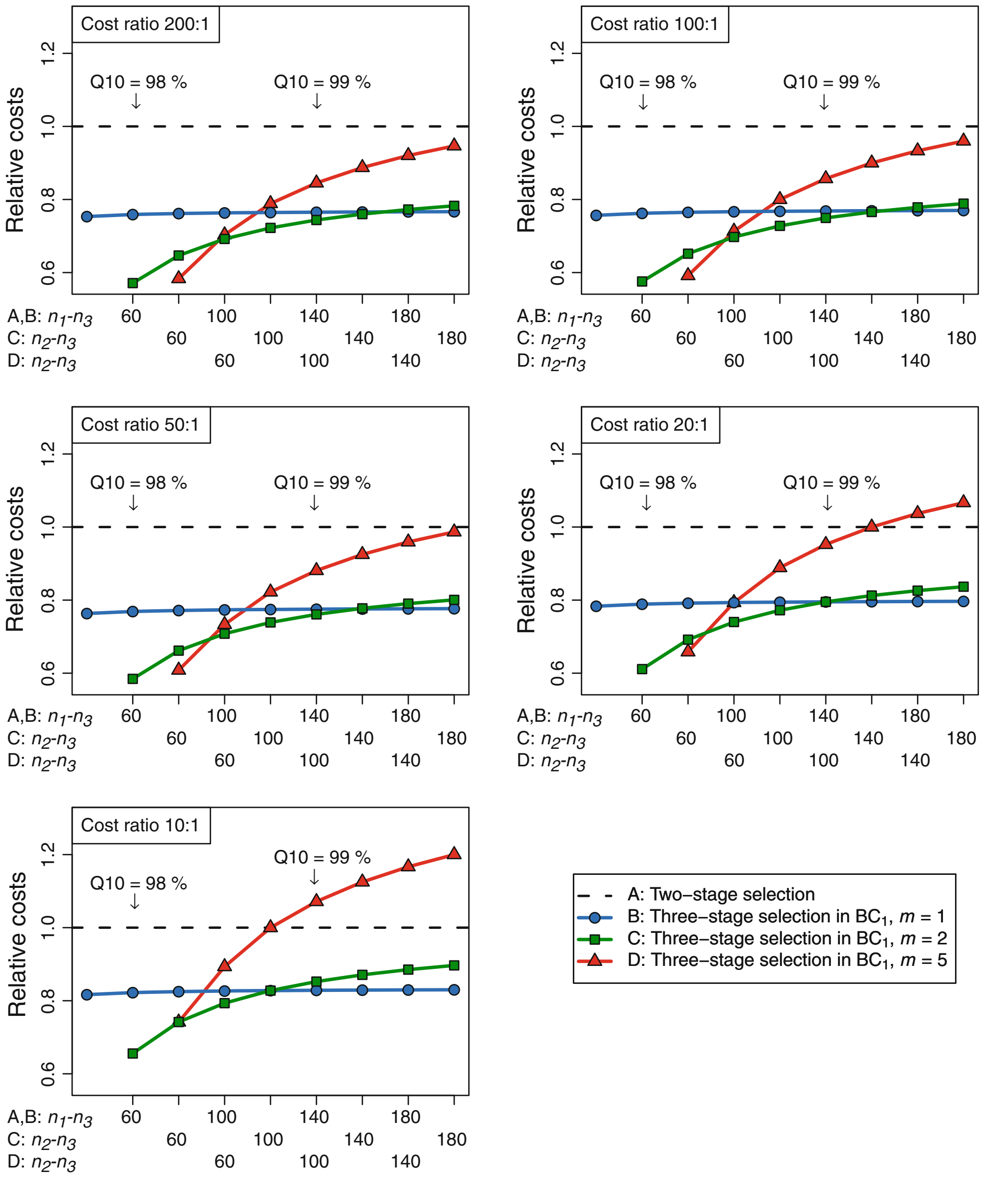

Fig. 1 Relative costs of three-stage selection with $m = 1, 2, 5$ in generation BC$_1$ and two-stage selection in generations BC$_2$ and BC$_3$ compared to two-stage selection in generations BC$_1$ to BC$_3$ for cost ratios for HT:SM assays of 200:1, 100:1, 50:1, 20:1, and 10:1

distributed markers (Table 3). This difference can be explained by the fact that, with random marker distribution, occasionally the distance between adjacent markers can get quite large, resulting in random gaps in the marker coverage. The recipient genome content of the chromosome regions in these gaps is not assessed and, therefore, the correlation of the marker estimate of the recurrent parent genome contribution and the true recurrent parent genome contribution is lower than for equally spaced markers. This results in a smaller response to marker-assisted background selection for randomly distributed compared to equally spaced markers.

We conclude that the possibility to generate linkage maps with equidistant marker distribution is a major advantage of HT marker systems, while the possibility to establish linkage maps with marker distances below 10 cM is only of secondary importance for gene introgression programs.

Pre-selection with flanking markers

In three-stage selection, the pre-selection of backcross plants showing recombination between the target gene and flanking markers allows an efficient control of the donor chromosome segment attached to the target gene. This reduces the probability of introducing negative alleles linked to the target gene into the genome of the recipient. Further, three-stage selection reduces the number of backcross plants subjected to genome-wide background selection and, therefore, reduces the number of required marker assays (Frisch et al. 1999a). To take advantage of these favorable properties of three-stage selection, a preselection for recombination between the target gene and flanking markers analyzed with SM assays can be combined with genome-wide background selection on the basis of HT assays. The design decisions required to implement such a selection strategy are discussed in the following.

Distances of flanking markers

Tightly linked flanking markers result in short donor chromosome segments attached to the target gene. However, they also result in a greater reduction of the number of individuals subjected to genome-wide background selection than loosely linked flanking markers. This reduced selection intensity can result in a decline of the genome-wide recovery of the recurrent parent genome. Therefore, the smallest δ_F that has no negative effect on the genome-wide response to selection can be regarded as an optimal flanking marker distance.

In backcrossing programs with constant ($m = 1$) population sizes ≤ 60, marker distances $\delta_F = 20$ cM between each flanking marker and the target gene resulted in high overall Q10 values while minimizing the number of HT assays required for background selection (Table 4). For larger populations, $\delta_F = 10$ was optimal. With $\delta_F = 5$ cM, controlling the donor genome segment attached to the target gene resulted in a decrease of the overall Q10 values. For such tightly linked flanking markers, only few recombinations do occur in a backcross population (see Frisch et al. 1999a, b for theoretical results) and, hence, only few plants are pre-selected and subjected to genome-wide background selection. This small number of individuals available for genome-wide background selection results in a smaller response to selection compared with less tightly linked flanking markers. We conclude that for gene introgression programs with constant population sizes, an optimum exploitation of the advantages of three-stage selection is reached with flanking marker distances of $\delta_F = 20$–10 cM, and that with smaller flanking marker distances, controlling the donor segment attached to the target gene is only possible at the cost of a lower overall Q10 value.

Generation of three-stage selection

Carrying out pre-selection for recombinants at markers flanking the target gene in only some, but not all generations of a gene introgression program can considerably reduce the logistic effort required for the marker analysis. A comparison of three-stage selection in generations BC_1 and BC_3 showed similar genome-wide Q10 values, but three-stage selection in generation BC_3 required more HT marker analyses (results not shown). Therefore, carrying out three-stage selection in generation BC_1 can be regarded as superior to three-stage selection in generation BC_3.

Three-stage selection in generations BC_1 and BC_2 required less HT assays but more SM assays than three-stage selection in generation BC_1 (Tables 4, 5). For population sizes below 100 individuals, this was accompanied by smaller genome-wide Q10 values. For population sizes greater than 100, employing three-stage selection in generations BC_1 and BC_2 provides a means to reduce the number of required genome-wide HT assays, by increasing the number of required SM analysis. Depending on the actual costs of SM and HT analysis and the work flow in the lab, this strategy can be used to shift the number of required marker analyses from HT to SM assays.

Large population sizes in the first backcross generation

As pre-selection with SM assays reduces the number of required HT assays, it provides a means to handle larger populations without necessarily increasing the cost of marker analysis. Increasing the population size in the generation where pre-selection with flanking markers is carried out increases the chance to find an individual with a small donor chromosome segment attached to the target gene, which has in addition a high proportion of recurrent parent genome (Frisch et al. 1999b). This theoretical consideration can serve as a rationale for using large population sizes in generations with three-stage selection.

We investigated backcrossing programs with three-stage selection in BC_1 populations that had $m = 1$, 2, or 5 times the size of the BC_2 and BC_3 populations in which two-stage selection was employed (Table 4). The Q10 values reached with $m = 1$ were comparable to those reached with two-stage selection for constant population sizes across generations (Table 3). Doubling the population size for three-stage selection in generation BC_1 ($m = 2$, $n_1 = mn_2 = mn_3$) resulted in Q10 values that were comparable to those reached with constant population sizes but using 20 more individuals per generation ($n_1' = n_2' = n_3' = n_2 + 20$). Using $m = 2$ required more SM but less HT assays than $m = 1$. A similar effect was observed for $m = 5$ and $n_1' = n_2' = n_3' = n_2 + 40$. However, here the increase in the number of required SM assays was considerable, while the reduction in the number of required HT assays was only small.

In conclusion, three-stage selection can be employed to put a stronger emphasis on the reduction of the donor segment attached to the target gene, and using two times larger population sizes in generation BC_1 ($m = 2$) than in BC_2 and BC_3 allows to shift the effort in the lab from HT to

SM assays compared to constant population size in all backcross generations ($m = 1$). These effects can be exploited without a reduction in the overall Q10 values. However, neither genetic advantages nor a reduction in the required marker assays supported employing five times larger populations in generation BC_1 ($m = 5$) than in generations BC_2 and BC_3.

Relative costs of three-stage selection

To compare the costs of three-stage selection in generation BC_1 with those of two-stage selection, we assumed cost ratios of 200:1 to 10:1 for the costs of one HT assay (comprising all marker loci on the linkage map) in relation to one SM assay (for one SM locus). First, the number of marker assays required to reach a given Q10 value with three-stage selection was determined from the simulations presented in Table 4, and the number of marker assays required to reach this Q10 value with two-stage selection was determined from the simulations presented in Table 3. Then the costs required with three-stage selection were determined with the above cost ratios and were set in relation to the costs that were required with two-stage selection (Fig. 1). For example, with a cost ratio of 200:1 for HT:SM assays (first diagram in Fig. 1) reaching the Q10 value of 99% with three-stage selection and $m = 5$ required 0.85 times the costs that were required to reach the Q10 value of 99% with two-stage selection. Three-stage selection with $m = 1$ required 0.77, and three-stage selection with $m = 2$ required 0.74 times the costs of two stage selection.

From the cost comparisons, we conclude that three-stage selection reaches a given Q10 value with less cost than two-stage selection, regardless of the cost ratio of HT:SM assays. If the aspired Q10 values are 99% or less, then doubling the population size in generation BC_1 provides a means to further reduce the costs required for the marker analyses.

Acknowledgements We thank the anonymous reviewers and the editor for their helpful suggestions. In particular, we gratefully acknowledge the comments of one reviewer that considerably improved the manuscript. We thank Gregory Mahone for proofreading the manuscript.

References

Bouchez A, Hospital F, Causse M, Gallais A, Charcosset A (2002) Marker-assisted introgression of favorable alleles at quantitative trait loci between maize elite lines. Genetics 162:1945–1959

Collard BCY, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Phil Trans R Soc 363:557–572

Falke KC, Frisch M (2011) Power and false positive rate in QTL detection with near-isogenic line libraries. Heredity 106:576–584

Falke KC, Miedaner T, Frisch M (2009) Selection strategies for the development of rye introgression libraries. Theor Appl Genet 119:595–603

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

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

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

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

Gupta PK, Roy JK, Prasad M (2001) Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. Curr Sci 80:524–535

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

Maurer HP, Melchinger AE, Frisch M (2008) Population genetic simulation and data analysis with Plabsoft. Euphytica 161:133–139

Peleman JD, van der Voort JR (2003) Breeding by design. Trends Plant Sci 7:330–334

Prigge V, Maurer HP, Mackill DJ, Melchinger AE, Frisch M (2008) Comparison of the observed with the simulated distributions of the parental genome contribution in two marker-assisted backcross programs in rice. Theor Appl Genet 116:739–744

Prigge V, Melchinger AE, Dhillon BS, Frisch M (2009) Efficiency gain of marker-assisted backcrossing by sequentially increasing marker densities over generations. Theor Appl Genet 119:23–32

Ragot M, Lee M (2007) Marker-assisted selection in maize: current status, potential, limitations and perspectives from the private and public sectors. In: Guimaraes EP, Ruane J, Scherf BD, Sonnino A, Dargie JD (eds) Marker-assisted selection. Current status and future perspectives in crops, livestock, forestry and fish. FAO, Rome, pp 117–150

Ribaut JM, Jiang C, Hoisington D (2002) Simulation experiments on efficiencies of gene introgression by backcrossing. Crop Sci 42:557–565

Schön CC, Melchinger AE, Boppenmaier J, Brunklaus-Jung E, Herrmann RG, Seitzer JF (1994) RFLP mapping in maize: quantitative trait loci affecting testcross performance of elite European flint lines. Crop Sci 34:378–389

Syvänen AC (2005) Toward genome-wide SNP genotyping. Nat Genet 37:S5–S10