Genetics: Published Articles Ahead of Print, published on March 31, 2005 as 10.1534/genetics.104.035451

Selection Theory for Marker-assisted Backcrossing

Matthias Frisch and Albrecht E. Melchinger

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

This article is dedicated to Professor Dr. H. F. Utz on the occasion of 65th birthday. His teaching of selection theory was most instrumental on the authors.

Running head: Selection theory for marker-assisted backcrossing

Keywords: Selection theory, gene introgression, gene pyramiding, marker-assisted backcrossing

Corresponding author:

Albrecht E. Melchinger,

Institute of Plant Breeding, Seed Science, and Population Genetics,

University of Hohenheim,

70593 Stuttgart,

Germany.

E-mail: melchinger@uni-hohenheim.de

ABSTRACT

Marker-assisted backcrossing is routinely applied in breeding programs for gene introgression. While selection theory is the most important tool for the design of breeding programs for improvement of quantitative characters, no general selection theory is available for marker-assisted backcrossing. In this treatise, we develop a theory for marker-assisted selection for the proportion of genome originating from the recurrent parent in a backcross program, carried out after preselection for the target gene(s). Our objectives were to (i) predict response to selection and (ii) give criteria for selecting the most promising backcross individuals for further backcrossing or selfing. Prediction of response to selection is based on the marker linkage map and the marker genotype of the parent(s) of the backcross population. In comparison to standard normal distribution selection theory, the main advantage our approach is that it considers the reduction of the variance in the donor genome proportion due to selection. The developed selection criteria take into account the marker genotype of the candidates and consider whether these will be used for selfing or backcrossing. Prediction of response to selection is illustrated for model genomes of maize and sugar beet. Selection of promising individuals is illustrated with experimental data from sugar beet. The presented approach can assist geneticists and breeders in the efficient design of gene introgression programs.

Marker-assisted backcrossing is routinely applied for gene introgression in plant and animal breeding. Its efficiency depends on the experimental design, most notably on the marker density and position, population size, and selection strategy. Gene introgression programs are commonly designed using guidelines taken from studies focusing on only one of these factors (e.g., Hospital *et al.* 1992; Visscher 1996; Hospital and Charcosset 1997; Frisch *et al.* 1999a,b). In breeding for quantitative traits, prediction of response to selection with classical selection theory is by far the most important tool for the design and optimization of breeding programs (Bernardo 2002). Adopting a selection theory approach to predict response to marker-assisted selection for the genetic background of the recurrent parent promises to combine several of the factors determining the efficiency of a gene introgression program into one criterion.

In classical selection theory, the expectation, genetic variance, and heritability of the target trait are required, as well as the covariance between the target trait and the selection criterion in the case of indirect selection (Bernardo 2002). In backcrossing without selection, the expected donor genome proportion in generation BC_n is $1/2^{n+1}$. In backcrossing with selection for the presence of a target gene, Stam and Zeven (1981) derived the expected donor genome proportion on the carrier chromosome of the target gene, extending earlier results of Bartlett and Haldane (1935), Fisher (1949) and Hanson (1959) on the expected length of the donor chromosome segement attached to the target gene. Their results were extended to a chromosome carrying the target gene and the recurrent parent alleles at two flanking markers (Hospital *et al.* 1992) and to a chromosome carrying several target genes (Ribaut *et al.* 2002).

Hill (1993) derived the variance of the donor genome proportion in an unselected

backcross population, whereas RIBAUT *et al.* (2002) deduced this variance for chromosomes carrying one or several target genes. The covariance of the donor genome proportion across a chromosome and the proportion of donor alleles at markers in backcrossing was given by VISSCHER (1996). In their derivations, these authors assumed that the donor genome proportion of different individuals in a backcross generation is stochastically independent. This applies to large BC_n populations only (a) in the absence of selection in all generations BC_s ($1 \leq s \leq n$) and (b) if each BC_{n-1} ($n > 1$) individual has maximally one BC_n progeny (comparable to the single seed descent method in recurrent selfing). VISSCHER (1999) showed with simulations that the variance of the donor genome proportion in backcross populations under marker-assisted selection is significantly smaller than in unselected populations of stochastically independent individuals.

HILLEL *et al.* (1990) and MARKEL *et al.* (1997) employed the binomial distribution to describe the number of homozygous chromosome segments in backcrossing. However, VISSCHER (1999) demonstrated with simulations that the assumption of binomially distributed chromosome segments results in an unrealistic prediction of the number of generations required for a marker-assisted backcross program. Hence, the expectations, variances, and covariances are known for backcrossing without selection, but these approximations are of limited use as a foundation of a general selection theory for marker-assisted backcrossing.

The objective of this study was to develop a theoretical famework for marker-assisted selection for the genetic background of the recurrent parent in a backcross program to (i) predict response to selection and (ii) give criteria for selecting the most promising backcross individuals for further backcrossing or selfing. Our approach deals with selection in generation n of the backcross program taking into account

(a) preselection for presence of one or several target genes, (b) the linkage map of the target gene(s) and markers, and (c) the marker genotype of the individuals used as non-recurrent parent for generating backcross generations BC_s ($s \leq n$).

THEORY

For all derivations we assume absence of interference in crossover formation such that the recombination frequency r and map distance d are related by HALDANE's (1919) mapping function $r(d) = \left(1 - e^{-2d}\right)/2$. An overview of the notation used throughout this treatise is given in Table 1. Table 1

In the following we derive (1) the expected donor genome proportion of a backcross individual conditional on its multilocus genotype g_n at marker and target loci, (2) the expected donor genome proportion of a backcross population generated by backcrossing an individual with multilocus genotype g_n to the recurrent parent, and (3) the expected donor genome proportion of the wth-best individual of a backcross population of size u generated by backcrossing an individual with multilocus genotype g_n to the recurrent parent.

Probability of multilocus genotypes: We derive the probability that a BC_n individual has multilocus genotype g_n under the condition that its non-recurrent parent has multilocus genotype g_{n-1}. Let

$$I = \{(i,j) \,|\, (i,j) \in L,\, g_{n-1,i,j} = 1\} \tag{1}$$

denote the set of indices, for which the locus at position $x_{i,j}$ was heterozygous in the non-recurrent parent in generation BC_{n-1} ($\mathrm{F}_1 = \mathrm{BC}_0$). The elements of I are ordered according to

$$(i',j') \prec (i,j) \text{ iff } (i' < i) \text{ or } \left[(i' = i) \text{ and } (j' < j)\right]. \tag{2}$$

The conditional probability that the BC_n individual has the multilocus marker genotype g_n is

$$P(G_n = g_n|g_{n-1}) = \prod_{(i,j)\in I} \Big\{\delta_{i,j} r^*_{i,j} + (1-\delta_{i,j})(1-r^*_{i,j})\Big\}, \tag{3}$$

where

$$\delta_{i,j} = \begin{cases} g_{n,i,j} & \text{for } j = 1 \\ |g_{n,i,j} - g_{n,k,l}| & \text{otherwise} \end{cases} \tag{4}$$

with

$$(k,l) = \max\big\{(i',j') \,|\, (i',j') \in I, (i',j') \prec (i,j)\big\} \tag{5}$$

and

$$r^*_{i,j} = \begin{cases} 1/2 & \text{for } j = 1 \\ r(x_{i,j} - x_{k,l}) & \text{otherwise.} \end{cases} \tag{6}$$

Distribution of donor alleles at markers: Consider a BC_n family of size u, generated by backcrossing one BC_{n-1} individual to the recurrent parent. Let

$$B = \sum_{(i,j)\in M} G_{n,i,j} \tag{7}$$

denote the number of donor alleles at the marker loci of a BC_n individual. The probability that an individual that carries all target genes is heterozygous at exactly b loci is

$$f(b) = P_t(B = b) = \frac{\sum_{g_n \in \mathscr{G}_{n,t,b}} P(G_n = g_n|g_{n-1})}{\sum_{g_n \in \mathscr{G}_{n,t}} P(G_n = g_n|g_{n-1})}, \tag{8}$$

where

$$\mathscr{G}_{n,t} = \Big\{g_n \,|\, t = \sum_{(i,j)\in T} g_{n,i,j}\Big\} \tag{9}$$

denotes the set of all multilocus marker genotypes carrying all target genes and

$$\mathscr{G}_{n,t,b} = \Big\{g_n \,|\, g_n \in \mathscr{G}_{n,t}, b = \sum_{(i,j)\in M} g_{n,i,j}\Big\} \tag{10}$$

denotes the set of all multilocus marker genotypes carrying all target genes and the donor allele at exactly b marker loci. The respective distribution function is $F(b) = P_t(B \leq b)$.

Selection of individuals with a low number of donor alleles: We determine the distribution of donor alleles in the individual carrying (1) all target genes and (2) the w smallest number of donor alleles among all carriers of the target genes (subsequently referred to as the w-th best individual).

Assume that v out of u individuals of a backcross family carry all target genes. Then, the distribution of donor alleles in the w-th best individual among the v carriers of the target gene is described by the w-th order statistic of v independent random variables with distribution function $F(b)$. Its distribution function is (David 1981)

$$F_{w:v}(b) = \sum_{i=w}^{v} \binom{v}{i} F(b)^w [1 - F(b)]^{v-w}. \tag{11}$$

Weighing with the probability that exactly v individuals carry the target gene yields the distribution function of donor alleles in the w-th best carrier of all target genes in a BC_n family of size u

$$H_{w,u}(b) = \sum_{0 \leq v \leq u} P(V = v) F_{w:v}(b) \tag{12}$$

with

$$P(V = v) = \binom{u}{v} p^v (1-p)^{u-v}, \tag{13}$$

where the probability p that an individual carries all target genes is

$$p = \prod_{(i,j) \in T} (1 - r^*_{i,j}) \tag{14}$$

and $r^*_{i,j}$ is calculated in analogously to Equations 5 and 6 but replacing I with T.

The probability that the w-th best individual carries b donor alleles is

$$h_{w,u}(b) = \begin{cases} H_{w,u}(b) & \text{for } b = 0 \\ \\ H_{w,u}(b) - H_{w,u}(b-1) & \text{for } b > 0. \end{cases} \tag{15}$$

Distribution of the donor genome proportion: In following, we investigate the homologous chromosomes of backcross individuals that originate from the non-recurrent parent. We divide the chromosomes into non-overlapping intervals

$$(a_{i,j}, b_{i,j}) = \begin{cases} (0, x_{i,j}) & \text{for } j = 1 \\ (x_{i,j-1}, x_{i,j}) & \text{for } 1 < j \leq l_i \\ (x_{i,l_i}, y_i) & \text{for } j = l_i + 1 \end{cases} \tag{16}$$

with length

$$d_{i,j} = b_{i,j} - a_{i,j} \tag{17}$$

for each

$$(i, j) \in J = \{(i, j) \,|\, i = 1 \ldots c, j = 1 \ldots l_i + 1\}. \tag{18}$$

Consider a BC_n individual with genotype g_n of which the genotype of the non-recurrent parent in generations BC_s $(1 \leq s < n)$ was g_s. We first derive the expected donor genome proportion $E(Z_{i,j})$ of a chromosome interval delimited by $(a_{i,j}, b_{i,j})$. Assume at first a finite number e of loci equidistantly distributed on the chromosome interval at positions $x^*_{i,1}, \ldots, x^*_{i,e}$, the corresponding random variables indicating the presence of the donor allele are $G^*_{n,i,1}, \ldots, G^*_{n,i,e}$. The expected donor genome proportion in the interval is then

$$E(Z_{i,j}) = \frac{1}{e} \sum_{k=1}^{e} E(G^*_{n,i,k}) \tag{19}$$

According to Hill (1993), who used results of Franklin (1977), Equation 19 can be extended to an infinite number of loci at positions $x^*_{i,k}$:

$$E(Z_{i,j}) = \frac{1}{d_{i,j}} \int_{a_{i,j}}^{b_{i,j}} E(G^*_{n,i,k})\, dx^*_{i,k} \tag{20}$$

with

$$\begin{aligned} E(G^*_{n,i,k}) &= P(G^*_{n,i,k} = 1) \\ &= \prod_{1 \leq s \leq n} P(G^*_{s,i,k} = 1 | g^*_{s-1,i,k} = 1). \end{aligned} \tag{21}$$

The probability $P(G^*_{s,i,k} = 1|g^*_{s-1,i,k} = 1)$ depends on the genotypes of the loci flanking the interval (i, j) in generations BC_{s-1} and BC_s. For telomere chromosome segments ($j = 1$, $j = l_i + 1$)

$$P(G^*_{s,i,k} = 1|g^*_{s-1,i,k} = 1) = \begin{cases} (1 - r^*) & \text{for } (g_{s-1,i,j}, g_{s,i,j}) = (1, 1) \\ r^* & \text{for } (g_{s-1,i,j}, g_{s,i,j}) = (1, 0) \\ 1/2 & \text{for } (g_{s-1,i,j}, g_{s,i,j}) = (0, 0) \end{cases} \quad (22)$$

where

$$r^* = \begin{cases} r(x_{i,1} - x^*_{i,k}) & \text{for } j = 1 \\ r(x^*_{i,k} - x_{i,l_i}) & \text{for } j = l_i + 1. \end{cases} \quad (23)$$

For non-telomere chromosome segments ($1 < j < l_i + 1$) the probability $P(G^*_{s,i,k} = 1|g^*_{s-1,i,k} = 1)$ can be calculated with the equations in Table 2. Table 2

The expected donor genome proportion on the homologous chromosomes originating from the non-recurrent parent of a BC_n individual with genotype g_n can then be determined as

$$z(g_n) = \sum_{(i,j)\in J} \frac{d_{i,j}}{y} E(Z_{i,j}) \quad (24)$$

Response to selection: We define response to selection R as the difference between the expected donor genome proportion μ in the selected fraction of a BC_n population and the expected donor genome portion μ' in the unselected BC_n population

$$R = \mu_n - \mu'_n. \quad (25)$$

We consider a BC_n family of size u_q generated by backcrossing one BC_{n-1} individual of genotype $g_{n-1,q}$. With respect to this family

$$E_{w,u_q}\left(z(G_n|g_{n-1,q})\right) = \sum_{b=0}^{m} \left[h_{w,u_q}(b) \sum_{g_n \in \mathscr{G}_{n,t,b}} \left\{ P(G_n = g_n \,|\, b, g_{n-1,q})\; z(g_n) \right\} \right] \quad (26)$$

denotes the expected donor genome proportion of the w-th best individual, where

$$P(G_n = g_n \,|\, b, g_{n-1,q}) = \frac{P(G_n = g_n \,|\, g_{n-1,q})}{\sum\limits_{g_n \in \mathscr{G}_{n,t,b}} P(G_n = g_n \,|\, g_{n-1,q})}. \tag{27}$$

We now consider p BC_{n-1} individuals with genotypes $g_{n-1,q}$ $(q = 1 \ldots p)$ that are backcrossed to the recurrent parent. Family size of family q is u_q such that the size of the BC_n population is $u = \sum_q u_q$. From family q, the w_q best individuals are selected such that the selected fraction consists of $w = \sum_q w_q$ individuals. We then have

$$\mu'_n = \frac{1}{4} \sum_{1 \le q \le p} \left[\frac{u_q}{u} z(g_{n-1,q}) \right] \text{ and} \tag{28}$$

$$\mu_n = \frac{1}{2} \sum_{1 \le q \le p} \sum_{1 \le j \le w_q} \left[\frac{1}{w} E_{j:u_q}\big(z(G_n | g_{n-1,q})\big) \right] \tag{29}$$

Note that $z(g_n)$ refers to one set of homologous chromosomes whereas, μ_n and μ'_n refer to both homologous chromosome sets. This results in the factors $1/4$ and $1/2$ in Equations 28 and 29.

Numerical implementation: Calculations for Equations 8 and 26 require enumeration of all realizations of the random vector G_n. For large number of markers, a Monte-Carlo method can be used to limit the necessary calculations. Instead of enumerating all realizations of G_n, a random sample of realizations, determined with a random walk procedure from the probability of occurrence of multi-locus genotypes (Equation 3), can be used as basis for the calculations. The routines developed for implementig our theory are available in software Plabsoft (Maurer *et al.* 2004).

DISCUSSION

Comparison to normal distribution selection theory: Normal distribution selection theory can be applied to marker-assisted backcrossing by considering a BC_n population in which indirect selection for low donor genome proportion Z is carried out by selecting individuals with a low count B of donor alleles at markers. Assuming a heritabiltiy of $h^2 = 1$ for the marker score B, response to selection R can be predicted (Bernardo 2002, pp. 264) as

$$R = i_b \, \frac{\mathrm{cov}(Z, B)}{\sqrt{\mathrm{var}(B)}}. \tag{30}$$

where i_b is the selection intensity.

Under the assumptions of (i) no selection in generations BC_s ($s < n$) and (ii) no preselection for the presence of target genes in generation BC_n, we have (Appendix A, using results of Hill 1993 and Visscher 1996)

$$\begin{aligned} \mathrm{var}(B) = m & \left(\frac{1}{2^{n+2}} - \frac{1}{2^{2n+2}} \right) \\ & + 2 \left(\frac{-k}{2^{2n+2}} + \frac{1}{2^{n+2}} \sum_{(i,j) \in M} \sum_{(i,j') \in M'_{i,j}} (1 - r_{i,j,j'}) \right), \end{aligned} \tag{31}$$

where

$$\begin{aligned} k &= \sum_{1 \le i \le c} (m_i - 1)^2 / 2, \\ M'_{i,j} &= \{ (i, j') \,|\, (i, j') \in M, j' > j \}, \\ r_{i,j,j'} &= r(x_{i,j'} - x_{i,j}), \end{aligned} \tag{32}$$

and (Appendix A)

$$\mathrm{cov}(B, Z) = \sum_{1 \le i \le c} \frac{y_i}{y} \, \mathrm{cov}\left(G_{n,i,j}, Z_i\right) \tag{33}$$

with (Visscher, 1996)

$$\mathrm{cov}(G_{n,i,j}, Z_i) = \frac{1}{4^{n+1} y_i} \sum_{1 \le s \le n} \binom{n}{s} \frac{1}{2s} \left(2 - e^{-2s x_{i,j}} - e^{-2s(y_i - x_{i,j})} \right). \tag{34}$$

From a mathematical point of view, applying normal distribution selection theory to marker-assisted backcrossing has the following shortcomings:

(a) The distribution of marker scores is discrete, but the normal approximation is continuous.

(b) The distribution of the marker scores is limited, but the normal approximation is unlimited.

(c) The relationship between marker score and donor genome proportion of an individual is nonlinear (this can be shown by using Equation 20), but normal distribution selection theory assumes a linear relationship.

From a genetical point of view, the presented derivations (Appendix A) of variance and covariance for the normal approximation (Equation 30) are based on the following assumptions:

(d) The BC_n population is generated by recurrent backcrossing of unselected BC_s ($1 \leq s < n$) populations of large size.

(e) No preselection for the presence of target genes was carried out in the BC_n population under consideration.

We illustrate the effects of these shortcomings and assumptions with a model close to the maize genome with 10 chromosomes of length 2 M, markers evenly distributed across the genome, and two target genes located in the center of Chromosomes 1 and 2.

For unselected BC_1 populations and large numbers of markers (e.g., 200), the normal approximation of the distribution of donor alleles fits very well the exact distribution (Figure 1A). However, if only few markers are employed, the discretization Figure 1

of the probability density function of the normal distribution approximates only roughly the exact distribution (Figure 1B). In particular, for donor genome proportions < 0.2, where selection will most likely take place, a considerable underestimation of the exact distribution is observed. This results in an underestimation of the response to selection when normal distribution selection theory is employed. The underestimation is even more severe if an order statistics approach for normal distribution selection theory is applied (Hill 1976), which takes the finite population size into account.

Due to the donor chromosome segments attached to the target genes, the donor genome proportion in backcross populations preselected for the presence of target genes is greater than in unselected backcross populations. This can result in an overestimation of the response to selection, when employing the normal distribution selection theory and using $1/2^{n+1}$ as the population mean of the donor genome proportion (Figure 1C). Note however, that an adaptation of the normal selection approach should be possible by adjusting the population mean with the expected length of the attached donor segment using results of Hanson (1959).

In marker-assisted backcross programs, usually a high selection intensity is employed and only one or few individuals of a backcross population are used as nonrecurrent parents for the next backcross generation. This results in a smaller variance in the donor genome proportion at markers compared with backcrossing the entire unselected population that is assumed by the normal distribution approach (Figure 1D). The result can be a severe overestimation of the response to selection.

The suggested exact approach overcomes the shortcomings and assumptions listed under (a)–(e). In conclusion, it can be applied to a much larger range of situations than the normal distribution approach.

Comparison to simulation studies: Simulation studies were successfully applied for obtaining guidelines for the design of marker-assisted backcrossing (HOSPITAL *et al.* 1992; VISSCHER 1999; FRISCH *et al.* 1999b; RIBAUT *et al.* 2002). According to VISSCHER (1999), one of the most important advantages of simulation studies is that selection is taken into account, whereas previous theoretical approaches yielded only reliable estimates for backcrossing without selection.

Our approach solves the problem of using selected individuals as non-recurrent parents. With respect to two areas, however, simulation studies cover a broader range of scenarios than the presented selection theory approach: (i) Simulations allow the comparison of alternative selection strategies, while in the present study we developed the selection theory approach for using the marker score B as selection index. (ii) Simulations allow to cover an entire backcross program, while we developed our approach only for one backcross generation. Both issues are promising areas for further research.

Prediction of response to selection: Prediction of response to selection with Equation 25 can be employed to compare alternative scenarios with respect to population size and required number of markers. We illustrate this application by the example of a BC_1 population using model genomes close to maize (10 chromosomes of length 2 M) and sugar beet (9 chromosomes of length 1 M). Markers are evenly distributed across all chromosomes and a target gene is located 66 cM from the telomere on Chromosome 1. The donor of the target gene and recurrent parent are completely homozygous. One individual is selected as non-recurrent parent of generation BC_2.

The expected response to selection for maize ranges from approximately 5% donor genome (20 markers, 20 plants) to 12% (120 markers, 1000 plants), for sugar

beet it ranges from approximately 7% to 15% (Figure 2). To obtain a response to selection of about 10% with 60 markers, a population size of 180 is required in maize, corresponding to approximately $180/2 \times 60 = 5400$ marker data points (MDP). By comparison, in sugar beet a population size of 60 is sufficient, resulting in only 30% of the MDP required for maize. This result indicates that the efficiency of marker-assisted backcrossing in crops with smaller genomes is much higher than in crops with larger genomes. STAM (2003) obtained similar results in a simulation study. Figure 2

Using more than 80 markers in maize (corresponding to a marker density of 25 cM) or more than 60 markers in sugar beet (marker density 15 cM) resulted only in a marginal increase of the response to selection, irrespective of the population size employed (Figure 2). Increasing the population size up to 100 plants results in substantial increase in response to selection in both crops, and using even larger populations still improves the expected response to selection. In conclusion, increasing the response to selection by increasing the number of markers employed is only possible up to an upper limit that depends on the number and length of chromosomes. In contrast, increasing response to selection by increasing the population size is possible up to population sizes that exceed the reproduction coefficient of most crop and animal species.

An optimum criterion for the design of marker-assisted selection in a backcross population can be defined by the expected response to selection reached with a fixed number of MDP. For fixed numbers of MDP in sugar beet, designs with large populations and few markers always reached larger values of response to selection than designs with small populations and many markers (Figure 2). For maize, the same trend was observed for 500 and 1000 MDP, while for larger number of MDP the optimum design ranged between 40 and 50 markers. In conclusion, in BC_1

populations of maize and sugar beet and a fixed number of MDP, marker-assisted selection is, within certain limits, more efficient for larger populations than for higher marker densities.

Selecting backcross individuals: Selection of backcross individuals can be carried out by using the number of donor alleles at markers B as a selection index. However, when employing markers not evenly distributed across the genome, the donor genome proportion at markers reflects only poorly the donor genome proportion across the entire genome.

The presented selection theory provides two alternative criteria that can be used as selection index for evaluation of each backcross individual: (1) The expected donor genome proportion $z(g_n)$ (Equation 24) of the backcross individual and (2) the expected donor genome proportion $E_{1,u}\left(z(G_{n+1}|g_n)\right)$ (Equation 26) of the best of the progenies obtained when using the backcross individual as non-recurrent parent of the next backcross generation. Employing $z(g_n)$ is recommended when selecting plants for selfing from the final generation of a backcross program, because the ultimate goal of a backcross program is to generate an individual (carrying the target genes) with low donor genome proportion. In contrast, employing $E_{1,u}\left(z(G_{n+1}|g_n)\right)$ is recommended for selecting individuals as parents for subsequent backcross generations, because here the donor genome proportion in the progenies is more important than the donor genome in the selected individual itself. Both criteria take into account the position of the markers and are, therefore, more suitable than B, if unequally distributed markers are employed.

Comparison of B, $z(g_n)$, and $E_{1,u}\left(z(G_{n+1}|g_n)\right)$ is demonstrated with experimental data from a gene introgression program in sugar beet. The target gene was located on Chromosome 1 with map distance 6 cM from the telomere, and 25

codominant polymorphic markers were employed for background selection. The map positions of the markers were (chromosome number/distance from the telomere in cM): 1/12, 1/28, 1/32, 1/40, 1/46, 1/75, 2/1, 2/16, 2/96, 3/0, 3/55, 3/78, 4/36, 4/64, 4/67, 5/33, 5/65, 6/42, 6/57, 7/4, 7/67, 8/14, 8/74, 9/0, 9/12. The length of Chromsomes 1 to 9 was 90, 102, 78, 84, 102, 89, 75, 94, and 94 cM. After producing the BC_1 generation, 89 plants carrying the target gene were preselected and analyzed for the 25 markers. The criteria B, $z(g_n)$, and $E_{1,u}\left(z(G_{n+1}|g_n)\right)$ for $u = 20, 40, 80$ were calculated and presented for the 25 plants with the smallest marker scores B (Table 3). Table 3

We refer here only to the most interesting results. (1) Plant #6 had $z(g_n) = 9.0\%$ and plant #10 had $z(g_n) = 17.0\%$, inspite of an identical marker score of $B = 6$. (2) Plant #1 was the best with respect to all three criteria. However, plant #6 was second best with respect to the expected donor genome proportion but had only rank 6 with respect to the marker score B. (3) Plant #9 had a considerably larger expected donor genome proportion ($z(g_n) = 14.8\%$) than plant #17 ($z(g_n) = 12.1\%$), but the expected donor genome proportion in the best progeny of plant #9 was lower than that of plant #17 for all three populations sizes.

These results demonstrate that the criteria B, $z(g_n)$, and $E_{1,u}\left(z(G_{n+1}|g_n)\right)$ can result in different rankings of individuals. In conclusion, if markers are not evenly distributed, calculating the proposed selection criteria in addition to the marker score B provides additional information to assess the value of backcross individuals and can assist geneticists and breeders in their selection decision.

QTL introgression: Marker-assisted selection in introgression of favorable alleles at quantitative trait loci (QTL) usually comprises selection for (1) presence of the donor allele at two markers delimiting the interval in which the putative QTL was detected and (2) the recurrent parent allele at markers outside the QTL

interval. Our results can be applied for the latter purpose in exactly the same way as previously described for the transfer of a single target gene. Hence, our approach is applicable to many scenarios in application of marker-assisted backcrossing for qualitative and quantitative traits.

Acknowledgments: We thank Dr. Dietrich Borchardt for critical reading and helpful comments on the manuscript. We are indepted to KWS Saat AG, 75555 Einbeck, Germany, for providing the experimental data on sugar beet. We greatly appreciate the helpful comments and suggestions of an anonymous reviewer.

LITERATURE CITED

Bartlett, M. S., and J. B. S. Haldane, 1935 The theory of inbreeding with forced heterozygosity. J. Genet. **31:** 327–340.

Bernardo, R., 2002 *Breeding for quantitative traits in plants*. Stemma Press, Woodbury, Minn.

David, H. A., 1981 *Order Statistics*. Wiley Inc, New York.

Franklin, I. R., 1977 The distribution of the proportion of genome which is homozygoues by descent in inbred individuals. Theor. Pop. Biol. **11:** 60–80.

Fisher, R. A., 1949 *The Theory of Inbreeding*. Oliver and Boyd, Edinburgh.

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. (erratum: Crop Sci. **39:** 1913).

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

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.

Hanson, W. D., 1959 Early generation analysis of lengths of heterozygous chromosome segments around a locus held heterozygous with backcrossing or selfing. Genetics **44:**

833–837.

Hill, W.G., 1976 Order statistics of correlated variables and implications in genetic selection programmes. Biometrics **32:** 889–902.

Hill, W. G., 1993 Variation in genetic composition in backcrossing programs. J. Heredity **84:** 212–213.

Hillel, J., T. Schaap, A. Haberfeld, A. J. Jeffreys, Y. Plotzky, *et al.*, 1990 DNA Fingerprints applied to gene introgression in breeding programs. Genetics **124:** 783–789.

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

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

Markel, P., Shu, P., Ebeling, C., Carlson, G. A., Nagle, D. L., *et al.*, 1997 Theoretical and empirical issues for marker-assisted breeding of congenic mouse strains. Nature Genetics **17:** 280–284.

Maurer, H. P., A. E. Melchinger, and M. Frisch, 2004 Plabsoft: Software for simulation and data analysis in plant breeding. *In:* Proceedings of the 17th EUCARPIA General Congress, 8-11 September 2004, Tulln, Austria. p. 359–362.

Ribaut, J.-M., C. Jiang, and D. Hoisington, 2002 Simulation experiments on efficiencies of gene introgression by backcrossing. Crop Sci. **42:** 557–565.

Stam, P., 2003 Marker-assisted introgression: Speed at any cost? *In:* EUCARPIA Leafy Vegetables 2003. *Eds.* Th. J. L. van Hintum, A. Lebeda, D. Pink, J. W. Schut. p. 117–124.

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

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

Visscher, P. M, 1999 Speed congenics: accelerated genome recovery using genetic markers. Genetical Research **74:** 81–85.

APPENDIX A

We use here an abbreviated notation: $G_{i,j}$ $(i = 1 \dots c,\ j = 1 \dots m_i)$ is a random variable taking 1 if the j-th marker on the i-th chromosome is heterozygous and 0 otherwise.

We derive the variance of B in a BC_n population under the assumptions of (1) no selection in generations BC_s $(s < n)$ and (2) no preselection for presence of target genes in generation BC_n (i.e., the entire BC_n population is considered, comprising individuals carrying the target genes as well as individuals not carrying the target gene). We have

$$\begin{aligned}\mathrm{var}(B) &= \mathrm{var}\left(\sum_{1\le i\le c}\sum_{1\le j\le m_i} G_{i,j}\right)\\ &= \sum_{1\le i\le c}\sum_{1\le j\le m_i}\mathrm{var}\left(G_{i,j}\right) + 2\sum_{1\le i\le c}\sum_{1\le j\le m_i}\sum_{j<j'\le m_i}\mathrm{cov}\left(G_{i,j},G_{i,j'}\right)\end{aligned}$$

Under assumptions (1) and (2) we have for any $G_{n,i,j}$ (Hill 1993)

$$\begin{aligned}\mathrm{var}\left(G_{i,j}\right) &= \frac{1}{4}\frac{1}{2^n}\left(1-\frac{1}{2^n}\right)\\ &= \frac{1}{2^{n+2}} - \frac{1}{2^{2n+2}}\end{aligned}$$

and further (Visscher 1996)

$$\begin{aligned}\mathrm{cov}\left(G_{i,j},G_{n,i,j'}\right) &= \frac{1}{4}\frac{1}{2^n}\left((1-r_{i,j,j'})^t - \frac{1}{2^n}\right)\\ &= \frac{(1-r_{i,j,j'})^n}{2^{n+2}} - \frac{1}{2^{2n+2}}\end{aligned}$$

with

$$r_{i,j,j'} = r(x_{i,j'} - x_{i,j}).$$

Therefore,

$$\begin{aligned}\mathrm{var}(B) = m&\left(\frac{1}{2^{n+2}} - \frac{1}{2^{2n+2}}\right)\\ &+ 2\left(\frac{-k}{2^{2n+2}} + \frac{1}{2^{n+2}}\sum_{1\le i\le c}\sum_{1\le j\le m_i}\sum_{j<j'\le m_i}(1-r_{i,j,j'})\right)\end{aligned}$$

where

$$k = \sum_{1 \le i \le c} (m_i - 1)^2/2$$

is the number of covariance terms.

We derive $\mathrm{cov}(B, Z)$ under the assumptions (1) and (2). Because

$$\begin{aligned}
\mathrm{E}(BZ) &= \mathrm{E}\left(\left[\sum_{1\le i\le c}\sum_{1\le j\le m_i} G_{i,j}\right]\left[\sum_{1\le i'\le c}\frac{y_{i'}}{y}Z_{i'}\right]\right) \\
&= \mathrm{E}\left(\sum_{1\le i\le c}\sum_{1\le j\le m_i}\sum_{1\le i'\le c} G_{i,j}\frac{y_{i'}}{y}Z_{i'}\right) \\
&= \sum_{1\le i\le c}\sum_{1\le j\le m_i}\sum_{1\le i'\le c} \mathrm{E}\left(G_{i,j}\frac{y_{i'}}{y}Z_{i'}\right)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}(B)\,\mathrm{E}(Z) &= \mathrm{E}\left(\sum_{1\le i\le c}\sum_{1\le j\le m_i} G_{i,j}\right)\mathrm{E}\left(\sum_{1\le i'\le c}\frac{y_{i'}}{y}Z_{i'}\right) \\
&= \left[\sum_{1\le i\le c}\sum_{1\le j\le m_i}\mathrm{E}\left(G_{i,j}\right)\right]\left[\sum_{1\le i'\le c}\mathrm{E}\left(\frac{y_{i'}}{y}Z_{i'}\right)\right] \\
&= \sum_{1\le i\le c}\sum_{1\le j\le m_i}\sum_{1\le i'\le c}\mathrm{E}\left(G_{i,j}\right)\mathrm{E}\left(\frac{y_{i'}}{y}Z_{i'}\right)
\end{aligned}$$

we have

$$\begin{aligned}
\mathrm{cov}(B, Z) &= \mathrm{E}(BZ) - \mathrm{E}(B)\,\mathrm{E}(Z) \\
&= \sum_{1\le i\le c}\sum_{1\le j\le m_i}\sum_{1\le i'\le c}\frac{y_{i'}}{y}\,\mathrm{cov}\left(G_{i,j}Z_{i'}\right)
\end{aligned}$$

and from

$$\mathrm{cov}\left(G_{i,j}Z_{i'}\right) = 0 \quad \text{for } i \neq i'$$

follows

$$\mathrm{cov}(B, Z) = \sum_{1\le i\le c}\sum_{1\le j\le m_i}\frac{y_i}{y}\,\mathrm{cov}\left(G_{i,j}Z_i\right).$$

TABLE 1
Notation

c	Number of chromosomes
$y = \sum_{1 \le i \le c} y_i$	y: Total genome length; y_i: length of chromsome i
$t = \sum_{1 \le i \le c} t_i$	t: Total number of target loci; t_i: Number of target loci on chromosome i
$m = \sum_{1 \le i \le c} m_i$	m: Total number of marker loci; m_i: Number of marker loci on chromosome i
$l = \sum_{1 \le i \le c} l_i$	l: Total number of loci; $l_i = m_i + t_i$: Number of loci on chromosome i
$x_{i,j}$	Map distance between locus j on chromosome i and the telomere
M	Set consisting of the indices (i,j) of marker loci $M = \{(i,j) \mid x_{i,j}$ is the map position of a marker locus$\}$
T	Set consisting of the indices (i,j) of target loci $T = \{(i,j) \mid x_{i,j}$ is the map position of a target locus$\}$
$L = M \cup T$	Set comprising the indices (i,j) of target and marker loci.
$d_{i,j}$	Length of chromosome interval j on chromosome i in map distance. For detailed definition see Equations 16–18
J	Set containing all indices (i,j) of chromosome intervals. $J = \{(i,j) \mid i = 1 \dots c, j = 1 \dots l_i + 1\}$
$G_{n,i,j}$, $g_{n,i,j}$	Indicator variable taking the value 1 if the locus at position $x_{i,j}$ carries the donor allele in generation BC_n and 0 otherwise. Realizations are denoted by $g_{n,i,j}$
G_n, g_n	Random vector denoting the multilocus genotype of a BC_n individual, $G_n = (G_{n,1,1}, G_{n,1,2}, \dots, G_{n,1,l_1}, \dots, G_{n,c,l_c})$. Realizations are denoted by g_n
$Z = \sum_{(i,j) \in J} \frac{y_i}{y} Z_{i,j}$	Z_n: Random variable denoting the donor genome proportion across the entire genome. $Z_{i,j}$: Random variable denoting the donor genome proportion in the chromosome interval corresponding to $d_{i,j}$ †
$Z_i = \sum_{1 \le j \le l_i + 1} \frac{d_{i,j}}{y_i} Z_{i,j}$	Z_i: Random variable denoting the donor genome proportion on chromosome i †
$B = \sum_{(i,j) \in M} G_{n,i,j}$	Random variable counting the number of donor alleles at marker loci

†: Random variables Z, Z_i, and $Z_{i,j}$ refer to the homologuous chromosomes originating from the non-recurrent parent

TABLE 2

Probability $P(G^*_{s,i,k} = 1 \,|\, g^*_{s-1,i,k} = 1)$ depending on flanking marker genotypes $g_{s-1,i,j-1}$, $g_{s-1,i,j}$, $g_{s,i,j-1}$, and $g_{s,i,j}$ for $1 < j < l + 1$.

	$g_{s,i,j-1}$, $g_{s,i,j}$			
$g_{s-1,i,j-1}$, $g_{s-1,i,j}$	1,1	1,0	0,1	0,0
1,1	$\frac{(1-r_1^*\dagger)(1-r_2^*)}{1-r_{i,j}}$	$\frac{(1-r_1^*)r_2^*}{r_{i,j}}$	$\frac{r_1^*(1-r_2^*)}{r_{i,j}}$	$\frac{r_1^* r_2^*}{1-r_{i,j}}$
1,0	0	$(1-r_1^*)$	0	r_1^*
0,1	0	0	$(1-r_2^*)$	r_2^*
0,0	0	0	0	1/2

‡ $r_1^* = r(x^*_{i,k} - x_{i,j-1})$ and $r_2^* = r(x_{i,j} - x^*_{i,k})$

TABLE 3

Selection criteria for the 25 BC_1 plants with highest marker score B in the sample dataset for sugar beet consisting of 89 plants. For details and explanation of $z(g_n)$ and $E_{1,u}\left(z(G_{n+1}|g_n)\right)$ see text.

			$E_{1,u}\left(z(G_{n+1}\|g_n)\right) \times 100$		
Plant #	B	$z(g_n) \times 100$	$u = 20$	$u = 40$	$u = 80$
1	2	7.4	3.4	3.2	3.0
2	4	18.0	6.7	6.1	5.6
3	4	11.5	3.9	3.5	3.2
4	5	14.8	5.4	4.9	4.5
5	5	14.9	5.5	5.0	4.6
6	6	9.0	3.8	3.4	3.1
7	6	16.2	5.9	5.3	4.9
8	6	15.4	5.0	4.4	3.9
9	6	14.8	4.8	4.2	3.7
10	6	17.0	5.6	5.0	4.4
11	7	17.2	6.4	5.7	5.2
12	7	20.2	6.3	5.5	4.9
13	8	19.6	6.5	5.7	5.1
14	8	23.2	8.0	7.1	6.4
15	8	14.5	5.6	4.9	4.3
16	8	17.1	6.0	5.4	4.9
17	8	12.1	4.9	4.3	3.8
18	9	16.7	6.4	5.5	4.7
19	9	17.0	6.7	5.9	5.3
20	9	18.0	7.0	6.4	5.8
21	9	21.4	8.4	7.7	7.1
22	9	27.3	10.7	9.7	8.6
23	9	14.8	5.6	4.9	4.3
24	9	18.3	5.6	4.9	4.3
25	9	18.3	6.1	5.4	4.9

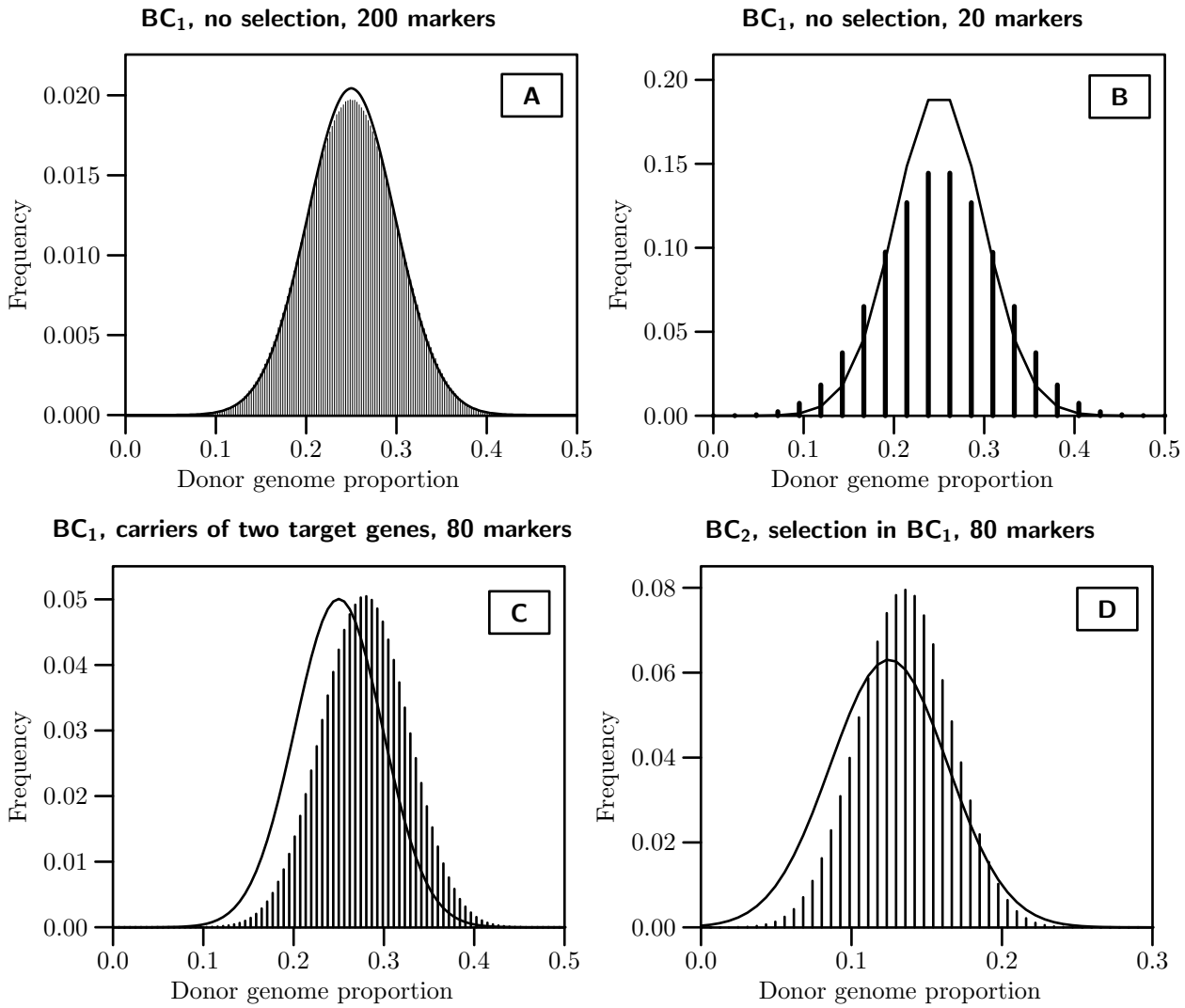

FIGURE 1

Distribution of the donor genome proportion at markers throughout the entire genome (comprising homologous chromosomes originating from the non-recurrent parent and the recurrent parent) calculated with a normal approximation (continous line) and the presented exact approach (histogram) for a model of the maize genome. Diagrams are shown for a BC_1 population without preselection for presence of target genes employing (A) 200 markers and (B) 20 markers, for a BC_1 population after preselection for presence of two target genes located in the center of chromosomes 1 and 2 employing 80 markers (C), and for a BC_2 population after preselection for presence of two target genes employing 80 markers. The BC_2 population was generated by backcrossing one BC_1 individual with donor genome content 0.25 (D).

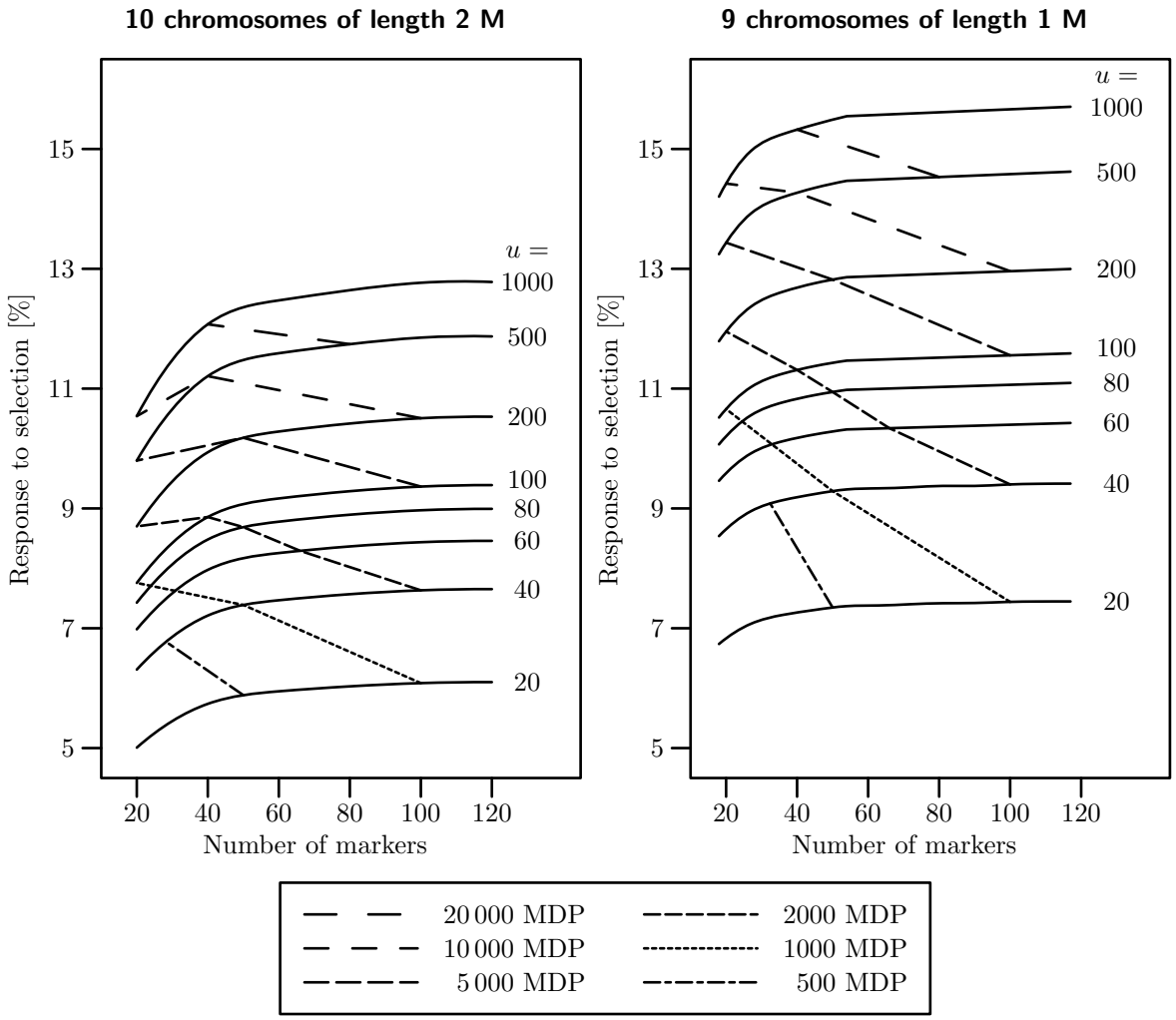

FIGURE 2

Expected response to selection throughout the entire genome (comprising homologous chromosomes originating from the non-recurrent parent and the recurrent parent) and expected number of required marker data points (MDP) when selecting the best out of $u = 20, 40, 60, 80, 100, 200, 500, 1000$ BC_1 individuals. The values depend on the number of markers (20 – 120) and on the number and length of the chromosomes. Left diagram: Model of the maize genome with 10 chromosomes of length 2 M. Right diagram: Model of the sugar beet genome with 9 chromosomes of length 1 M.