The Length of the Intact Donor Chromosome Segment Around a Target Gene in Marker-Assisted Backcrossing

Matthias Frisch and Albrecht E. Melchinger

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

Manuscript received May 19, 2000
Accepted for publication November 20, 2000

ABSTRACT

Recurrent backcrossing is an established procedure to transfer target genes from a donor into the genetic background of a recipient genotype. By assessing the parental origin of alleles at markers flanking the target locus one can select individuals with a short intact donor chromosome segment around the target gene and thus reduce the linkage drag. We investigated the probability distribution of the length of the intact donor chromosome segment around the target gene in recurrent backcrossing with selection for heterozygosity at the target locus and homozygosity for the recurrent parent allele at flanking markers for a diploid species. Assuming no interference in crossover formation, we derived the cumulative density function, probability density function, expected value, and variance of the length of the intact chromosome segment for the following cases: (1) backcross generations prior to detection of a recombinant individual between the target gene and the flanking marker; (2) the backcross generation in which for the first time a recombinant individual is detected, which is selected for further backcrossing; and (3) subsequent backcross generations after selection of a recombinant. Examples are given of how these results can be applied to investigate the efficiency of marker-assisted backcrossing for reducing the length of the intact donor chromosome segment around the target gene under various situations relevant in breeding and genetic research.

RECURRENT backcrossing with selection for presence of a target gene is a well-established breeding method for introgressing desirable genes from a donor into the genetic background of a recipient genotype used as recurrent parent. With the development of high-density linkage maps in most crop species, it became possible to monitor the parental origin of alleles at DNA markers throughout the entire genome. Selection of individuals, which not only carry the target gene but also are homozygous for the recurrent parent alleles at a large portion of markers, can accelerate recovery of the recurrent parent genome and reduce the number of backcross generations required for gene introgression. This approach is called background selection and was first proposed by Tanksley *et al.* (1989).

The goal of background selection is to reduce the recurrent parent genome proportion across the whole genome (Frisch *et al.* 1999b). However, special attention must be paid to the donor chromosome segment around the target gene. Without selection, this segment can remain fairly long over a large number of backcross generations and, hence, contribute a major part to the donor genome still present in the final breeding product. For example, Young and Tanksley (1989) found lengths up to 51 cM of the segment attached to a resistance gene after six backcross generations in tomato. Their experiments confirmed theoretical results of Stam and Zeven (1981), who showed that the length of the donor chromosome segment attached to a target gene on a 100-cM chromosome after six backcross generations without background selection is expected to be 32 cM. Moreover, there are numerous examples of undesirable traits tightly linked to a target gene, which were introgressed together with the gene into near-isogenic lines (Zeven *et al.* 1983).

Naveira and Barbadilla (1992) reviewed the early theoretical studies on the length of the intact donor chromosome segment around the target gene in recurrent backcrossing without marker-assisted selection. The problem was first addressed by Bartlett and Haldane (1935), but their approach was limited because they used recombination frequencies instead of map distances. Using Haldane's (1919) classical definition of map distance, Fisher (1949, pp. 49–50) derived the probability that the donor chromosome segment attached on one side of the target gene is after t backcross generations greater than a certain value x as $p = e^{-tx}$ and the probability density function (pdf) of random variable X describing its length as $df/dx = te^{-tx}$. Calculating the expectation $E_t(X)$ by assuming a chromosome of infinite length yields Fisher's (1949, p. 50) formula

$$E_t(X) = \int_0^\infty xte^{-tx}dx = 1/t. \tag{1}$$

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

Hanson (1959) extended Fisher's results by abandoning the latter assumption, which results in an overestimation of the expected length of the attached chromosome segment. He considered a chromosome of length l and took the discrete character of X for $x = l$ into account:

$$E_t(X) = \int_0^l xte^{-tx}dx + le^{-lx} = (1 - e^{-tl})/t. \qquad (2)$$

Stam and Zeven (1981) derived the expected donor genome proportion on the carrier chromosome in backcrossing with selection for presence of a target gene with known as well as unknown map position. In contrast to Hanson's (1959) study, their approach also includes donor chromosome segments on the carrier chromosome, which are not directly linked to the target gene. Hospital *et al.* (1992) extended Stam and Zeven's (1981) approach to background selection with exactly two markers on the carrier chromosome. However, the probability distribution of the chromosome segment attached to the target gene is unknown.

Frisch *et al.* (1999a) derived the minimum population size required to find with probability q single or double recombinants between the target gene and flanking markers in marker-assisted backcrossing, assuming known distances to the flanking markers. However, the effect of the marker distances and generation of selecting a recombinant between the target gene and a flanking marker on the probability distribution of the chromosome segment attached to the target gene is unknown.

The objective of this study was to extend earlier results concerning the length of the intact donor chromosome segment around the target gene to backcross programs with selection for (a) presence of a target gene and (b) homozygosity of the recurrent parent allele at flanking markers. Knowledge of the probability distribution of the length of the chromosome segment around a target gene is useful in (i) choosing the flanking markers depending on the effect of their position on the intact chromosome segment length and (ii) estimating the intact chromosome segment length on the basis of the marker genotype during a backcross program.

We derived the cumulative density functions (cdf's) and the pdf's of the intact donor chromosome segment length for (1) backcross generations prior to detection of a recombinant between the target gene and the marker(s), (2) the backcross generation in which a recombinant is first detected and selected for further backcrossing, and (3) backcross generations after selection of a recombinant. The respective expected values and variances can be calculated from these density functions.

ASSUMPTIONS, DEFINITIONS, AND SOME BASIC RESULTS

We assume that the process of crossover formation during meiosis in a diploid species is completely described by the following two properties: (A) The location of crossovers is uniformly distributed along the chromosome, and (B) the number of crossovers, which occur during one meiosis on a chromosome region of length l, follows a Poisson distribution with parameter l. Assumptions A and B are mathematically equivalent to those made by Haldane (1919), when he introduced the idea of relating recombination frequencies to map distances with a mapping function. This models implies that an odd number of crossovers between two loci results in recombination between them and, hence, meiosis is modeled by involving only two chromatids (see Zhao and Speed 1996 for discussion). A further consequence of assumptions A and B is the absence of interference in crossover formation (Stam 1979).

For our derivations we use an algebra of events, the union and intersection operators $\cup$ and $\cap$, and the subset relation $\subseteq$. Let u and v denote two positions on a chromosome, measured in a scale in morgan units with the coordinate origin at the target locus. For all calculations concerning only one side of the target locus, we assume without loss of generality $u < v$ and u, $v > 0$. The map distance between the target locus and the end of the chromosome is l. The following notation is used for events in generation BC$_i$:

$Z_{u,v,i}$: No crossover occurred in the interval $[u, v)$.
$O_{u,v,i}$: An odd number of crossovers occurred in $[u, v)$.
$E_{u,v,i}$: An even number (including zero) of crossovers occurred in $[u, v)$.

Note that $Z_{u,v,i} \subseteq E_{u,v,i}$. Furthermore, we define

$N_{u,v,t} = \cap_{i=1}^{t} E_{u,v,i}$: No recombination occurred between the loci at positions u and v in any of the generations BC$_1$ to BC$_t$.
$R_{u,v,t} = \cap_{i=1}^{t-1} E_{u,v,i} \cap O_{u,v,t}$ for $t > 1$ and $R_{u,v,1} = O_{u,v,1}$ for $t = 1$: Recombination between the loci at positions u and v occurred for the first time in generation BC$_t$.

Note that the events $Z_{u,v,i}$, $O_{u,v,i}$, and $E_{u,v,i}$ describe recombination in the interval $[u, v)$ in generation BC$_i$, whereas $N_{u,v,t}$ and $R_{u,v,t}$ refer to the accumulation of events in generations BC$_1$ to BC$_t$.

From assumptions A and B follows directly (Haldane 1919), for the probabilities of the events $Z_{u,v,s}$, $O_{u,v,s}$, and $E_{u,v,s}$ in generation BC$_s$,

$$P(Z_{u,v,s}) = e^{-d}, \qquad (3)$$

$$P(O_{u,v,s}) = e^{-d}\sinh d, \qquad (4)$$

$$P(E_{u,v,s}) = e^{-d}\cosh d, \qquad (5)$$

where $d = v - u$. Following Haldane's original derivation, we use hyperbolic functions because they are easier to handle in subsequent derivations than the more common formulas with exponential functions only.

Formation of crossovers in different generations is stochastically independent. Moreover, assumptions A and B imply that for every generation crossover formation in nonoverlapping intervals is independent (Stam 1979). Hence, we obtain for $I, J \in \{Z, O, E\}$, $s \neq s'$, and arbitrary intervals $[u, v)$ and $[u', v')$,

$$P(I_{u,v,s} \cap J_{u',v',s'}) = P(I_{u,v,s})P(J_{u',v',s'}), \tag{6}$$

and for $K, L \in \{Z, O, E, N, R\}$, arbitrary generations s and s', and $[u, v) \cap [u',v') = \emptyset$,

$$P(K_{u,v,s} \cap L_{u',v',s'}) = P(K_{u,v,s})P(L_{u',v',s'}). \tag{7}$$

Equations 6 and 7 can be used to calculate the probabilities of events N and R as

$$P(N_{u,v,t}) = \prod_{i=1}^{t} P(E_{u,v,i}) = e^{-td}\cosh^t d, \tag{8}$$

$$P(R_{u,v,t}) = \prod_{i=1}^{t-1} P(E_{u,v,i})P(O_{u,v,t}) = e^{-td}\cosh^{t-1} d \sinh d. \tag{9}$$

We define a random variable X, which describes the length of the donor chromosome segment attached on one side of the target gene. The event "the donor chromosome segment attached on one side of the target gene is greater than a certain value x" is denoted by $\{X > x\}$. Hence, the cdf of the random variable X is $F(x) = 1 - P(\{X > x\})$.

SEGMENT ATTACHED ON ONE SIDE OF THE TARGET GENE

Under assumptions A and B, the two random variables that describe the length of the intact donor chromosome segments attached on each side of the target gene are stochastically independent. We use this property to first derive the distribution of each random variable and then combine the results to obtain formulas for the total length. The core of the approach is the derivation of the cdf's from conditional probabilities; further properties of the distribution such as pdf, expectation, and variance can be derived with standard methods.

Generations prior to detection of a recombinant: We first investigate the length of the intact chromosome segment in backcross generation BC$_t$ under the condition that no recombination between the target gene and a marker at position y occurred in any backcross generation BC$_i$ ($i \leq t$). We distinguish three cases: (1) The attached chromosome segment is smaller than the flanking marker distance; (2) the attached chromosome segment is greater than the flanking marker distance but smaller than the distance between target gene and the end of the chromosome; and (3) the attached chromosome segment comprises the complete distance between target gene and the end of the chromosome.

Case 1. $x \in [0, y)$: The event "the intact chromosome segment is greater than a certain value $x \in [0, y)$ in generation BC$_t$ and no recombination was observed between the target gene and the marker" occurs if and only if no crossover event happened in the interval $[0, x)$ in all backcross generations BC$_i$ ($i \leq t$) and an even number of crossovers occurred in the interval $[x, y)$ in all backcross generations BC$_i$ ($i \leq t$):

$$\{X_t > x\} \cap N_{0,y,t} = \bigcap_{i=1}^{t} (Z_{0,x,i} \cap E_{x,y,i}). \tag{10}$$

With Equations 3, 5, and 7 we obtain

$$P(\{X_t > x\} \cap N_{0,y,t}) = e^{-ty}\cosh^t(y - x), \tag{11}$$

which is required for calculation of the conditional probability

$$P(\{X_t > x\}|N_{0,y,t}) = \frac{P(\{X_t > x\} \cap N_{0,y,t})}{P(N_{0,y,t})} = \frac{\cosh^t(y - x)}{\cosh^t y}. \tag{12}$$

Hence, the cdf of the attached chromosome segment length is

$$F_t(x|N_{0,y,t}) = 1 - \frac{\cosh^t(y - x)}{\cosh^t y} \quad \text{for } x \in [0, y). \tag{13}$$

Case 2. $x \in [y, l)$: The event "the attached chromosome segment is greater than a certain value $x \in [y, l)$ in generation BC$_t$ and no recombination was observed between the target gene and the marker" occurs if and only if no crossover happened in the interval $[0, x)$ in all backcross generations BC$_i$ ($i \leq t$). From $x \in [y, l)$ follows $\{X_t > x\} \subseteq N_{0,y,t}$ and hence,

$$\{X_t > x\} \cap N_{0,y,t} = \bigcap_{i=1}^{t} Z_{0,x,i}. \tag{14}$$

In analogy to the calculations in Equations 10–13 we obtain the cdf

$$F_t(x|N_{0,y,t}) = 1 - \frac{e^{t(y-x)}}{\cosh^t y} \quad \text{for } x \in [y, l). \tag{15}$$

Case 3. $x = l$: The event "the attached chromosome segment takes its maximum value in generation BC$_t$, $\{X_t = l\}$ and no recombination was observed between the target gene and the marker" occurs if and only if no crossover occurred in the interval $[0, l)$ in all backcross generations BC$_i$ ($i \leq t$). From $\{X_t = l\} \subseteq N_{0,y,t}$ follows

$$\{X_t = l\} \cap N_{0,y,t} = \bigcap_{i=1}^{t} Z_{0,l,i} \tag{16}$$

and therefrom we get

$$P(\{X_t = l\}|N_{0,y,t}) = \frac{e^{t(y-l)}}{\cosh^t y}. \tag{17}$$

The discrete character of X_t for the value $x = l$ must be taken into account when calculating the expectation and variance of X_t.

Pdf for Cases 1 and 2: Differentiation of Equations 13 and 15 with respect to x yields the pdf

$$f_t(x|N_{0,y,t}) = \begin{cases} \dfrac{t\cosh^{t-1}(y-x)\sinh(y-x)}{\cosh^t y} & \text{for } x \in [0, y) \\ \dfrac{te^{t(y-x)}}{\cosh^t y} & \text{for } x \in [y, l). \end{cases} \tag{18}$$

Note that $f_t(x|N_{0,y,t})$ is not continuous for $x = y$.

Generation in which a recombinant is detected and selected: Let us now assume that recombination between the target gene and the marker occurred for the first time in generation BC_s. The event "the chromosome segment attached on this side of the target gene is greater than a certain value $x \in [0, y)$ and recombination is observed between the target gene and the marker" occurs if and only if no crossover happened in the interval $[0, x)$ in all backcross generations BC_i $(i \leq s)$, an even number of crossovers occurred in the interval $[x, y)$ in all backcross generations BC_i $(i < s)$, and an odd number of crossovers occurred in the interval $[x, y)$ in generation BC_s:

$$\{X_s > x\} \cap R_{0,y,s} = \bigcap_{i=1}^{s} Z_{0,x,i} \cap \bigcap_{i=1}^{s-1} E_{x,y,i} \cap O_{x,y,s}. \tag{19}$$

In analogy to the calculations in Equations 10–13 we obtain the cdf

$$F_s(x|R_{0,y,s}) = \begin{cases} 1 - \dfrac{\sinh(y-x)\cosh^{s-1}(y-x)}{\sinh y\cosh^{s-1} y} & \text{for } x \in [0, y) \\ 1 & \text{for } x \in [y, l]. \end{cases} \tag{20}$$

Differentiation with respect to x yields the corresponding pdf of the attached chromosome segment length under the condition that recombination between the target gene and the marker occurred for the first time in generation BC_s:

$$f_s(x|R_{0,y,s}) = \begin{cases} \dfrac{s\cosh^{s}(y-x) - (s-1)\cosh^{s-2}(y-x)}{\sinh y\cosh^{s-1} y} & \text{for } x \in [0, y) \\ 0 & \text{for } x \in [y, l]. \end{cases} \tag{21}$$

Note that $f_s(x|R_{0,y,s})$ is not continuous for $x = y$.

Subsequent generations after selection of a recombinant: We now investigate the distribution of the length of the attached segment on one side of the target, when selection of a recombinant individual, on the basis of a flanking marker at position y, was carried out in generation BC_s and backcrossing is continued for another $t - s$ generations. The event "the attached chromosome segment is greater than a certain value x in generation BC_t and recombination is observed between the target gene and the marker in generation BC_s $(s \leq t)$" occurs if and only if no crossover occurred in the interval $[0, x)$ in all generations BC_i $(i \leq t)$, an even number of crossovers occurred in the interval $[x, y)$ in all backcross generations BC_i $(i < s)$, and an odd number of crossovers occurred in the interval $[x, y)$ in generation BC_s:

$$\{X_t > x\} \cap R_{0,y,s} = \bigcap_{i=1}^{t} Z_{0,x,i} \cap \bigcap_{i=1}^{s-1} E_{x,y,i} \cap O_{x,y,s}. \tag{22}$$

In analogy to the calculations in Equations 10–13 we obtain the cdf

$$F_t(x|R_{0,y,s}) = \begin{cases} 1 - \dfrac{\sinh(y-x)\cosh^{s-1}(y-x)}{\sinh y\cosh^{s-1} y} e^{(s-t)x} & \text{for } x \in [0, y) \\ 1 & \text{for } x \in [y, l]. \end{cases} \tag{23}$$

Differentiation with respect to x yields the corresponding pdf of the attached chromosome segment length in generation BC_t under the condition that recombination between the target gene and the marker occurs for the first time in generation BC_s:

$$f_t(x|R_{0,y,s}) = \begin{cases} \dfrac{e^{(s-t)x}\cosh^{s-2}(y-x)}{\sinh y\cosh^{s-1} y} \\ \quad \times [\tfrac{1}{2}(t-s)\sinh(2y-2x) \\ \qquad + s\sinh^2(y-x) + 1] & \text{for } x \in [0, y) \\ 0 & \text{for } x \in [y, l]. \end{cases} \tag{24}$$

Note that for $s = t$, Equation 24 simplifies to Equation 21 and that $f_t(x|R_{0,y,s})$ is not continuous for $x = y$.

Expected values and variances: From the presented pdf's, expected values and variances of the distribution of X on one side of the target gene can be obtained with standard methods of calculus,

$$E_t(X|R_{0,y,s}) = \int_0^y xf(x|R_{0,y,s})\,dx \tag{25}$$

$$E_t(X|N_{0,y,t}) = \int_0^l xf(x|N_{0,y,t})\,dx + lP(X = l|N_{0,y,t}) \tag{26}$$

$$V_t(X|R_{0,y,s}) = E_t(X^2|R_{0,y,s}) - [E_t(X|R_{0,y,s})]^2 \tag{27}$$

$$V_t(X|N_{0,y,t}) = E_t(X^2|N_{0,y,t}) - [E_t(X|N_{0,y,t})]^2 \tag{28}$$

with

$$E_t(X^2|R_{0,y,s}) = \int_0^l x^2 f(x|R_{0,y,s})\,dx \tag{29}$$

$$E_t(X^2|N_{0,y,t}) = \int_0^l x^2 f(x|N_{0,y,t})\,dx + l^2 P(X = l|N_{0,y,t}). \tag{30}$$

Note that integration must be performed separately for the intervals of definition of $f(x|N_{0,y,t})$.

The expectation and variance of the length of the intact chromosome segment attached on one side of the target gene, when selecting in generations BC_1 and BC_2 for recombinants between the target gene and the marker, are presented in Table 1. In APPENDIX A we demonstrate how these equations were derived using $E_t(x|R_{0,y,1})$ as an example.

Selection of recombinants without marker analyses in previous generations: There are situations in practice when the genotype of flanking markers is not examined right from the beginning of a backcrossing program but

TABLE 1

Expected values (E_t) and variances (V_t) of the chromosome segment attached on one side of the target gene in generations BC_1 ($t = 1$), BC_2 ($t = 2$), and BC_t

$E_1(X \mid N_{0,y,1}) = \dfrac{1 + \sinh y - e^{2y-2l}}{\cosh y}$	$E_2(X \mid N_{0,y,2}) = \dfrac{2 + 2y + \sinh y - 2e^{2y-2l}}{4 \cosh^2 y}$
$E_1(X \mid R_{0,y,1}) = \tanh\left(\dfrac{y}{2}\right)$	$E_2(X \mid R_{0,y,2}) = \dfrac{\tanh y}{2}$
$E_t(X \mid R_{0,y,1}) = \begin{cases} \dfrac{te^{-2y} - 2e^{-yt} - t + 2}{t(t-2)\,(e^{-2y} - 1)} & \text{if } t \neq 2 \\ \dfrac{(1 + 2y)e^{-2y} - 1}{2(e^{-2y} - 1)} & \text{if } t = 2 \end{cases}$	$E_t(X \mid R_{0,y,2}) = \begin{cases} \dfrac{te^{-4y} - 4e^{-yt} - t + 4}{t(t-4)\,(e^{-4y} - 1)} & \text{if } t \neq 4 \\ \dfrac{(1 + 4y)e^{-4y} - 1}{4(e^{-4y} - 1)} & \text{if } t = 4 \end{cases}$
$V_1(X \mid N_{0,y,1}) = 2\dfrac{\cosh y - (l + 1)e^{y-l} + y}{\cosh y} - [E_1(X \mid N_{0,y,1})]^2$	$V_2(X \mid N_{0,y,2}) = \dfrac{\cosh 2y - (4l + 2)e^{2y-2l} + 2y^2 + 4y + 1}{4 \cosh^2 y} - [E_2(X \mid N_{0,y,2})]^2$
$V_1(X \mid R_{0,y,1}) = 2\dfrac{\sinh y - y}{\sinh y} - [E_1(X \mid R_{0,y,1})]^2$	$V_2(X \mid R_{0,y,2}) = \dfrac{\sinh 2y - 2y}{4 \cosh y \sinh y} - [E_2(X \mid R_{0,y,2})]^2$
$V_t(X \mid R_{0,y,1}) = \begin{cases} 2\dfrac{t^2e^{-2y} - [g(y, t)]e^{-yt} - (t-2)^2}{t^2(t-2)^2(e^{-2y} - 1)} - [E_t(X \mid R_{0,y,1})]^2 & \text{if } t \neq 2 \\ g(y, t) = 2yt^2 - 4yt + 4t - 4 & \\ \dfrac{2e^{-4y} - (8y^2 + 4)e^{-2y} + 2}{8(e^{-2y} - 1)^2} & \text{if } t = 2 \end{cases}$	$V_t(X \mid R_{0,y,2}) = \begin{cases} 2\dfrac{t^2e^{-4y} - [h(y, t)]e^{-yt} - (t-4)^2}{t^2(t-4)^2(e^{-4y} - 1)} - [E_t(X \mid R_{0,y,2})]^2 & \text{if } t \neq 4 \\ h(y, t) = 4yt^2 - 16yt + 8t - 16 & \\ \dfrac{2e^{-8y} - (32y^2 + 4)e^{-4y} + 2}{32(e^{-4y} - 1)^2} & \text{if } t = 2 \end{cases}$

One of the following conditions apply: (1) recombination between the target gene and a marker at position y occurred for the first time in generation BC_s ($R_{0,y,s}$), or (2) no recombination occurred until generation BC_s ($s = 1, 2$) ($N_{0,y,s}$). For detailed definitions of $R_{0,y,s}$ and $N_{0,y,s}$ see text.

only in advanced generations. Frequently, the flanking markers used for identification of recombinants are positioned fairly distant from the target gene to have a high probability of success for recovering at least one recombinant with a limited population size (Frisch *et al.* 1999b). However, if several recombinants are found, the experimenter may decide to assay these with additional markers closer to the target gene to identify the one with the shortest intact donor chromosome segment. We derive the distribution of the length of the donor chromosome segment attached on one side of the target gene for these situations in APPENDIX C.

Numerical results: Figure 1 shows the expected length $E_t(X)$ and the standard deviation $\mathrm{SD}_t(X)$ of the intact chromosome segment attached on one side of the target gene in generations BC_1 to BC_{15} ($t = 1 \ldots 15$) for backcross programs with and without background selection at flanking markers in generation BC_1. The target locus is positioned at distance $l = 1.0$ M from the chromosome end and the flanking marker is located at distance $y = 0.1, 0.2, 0.3, 0.4, 0.5$ M from the target locus.

In generation BC_1, $E_1(X \mid R_{0,y,1}) = \tanh(y/2) \approx y/2$ (Table 1). Hence, the expected length of the intact chromosome segment is ~0.25 M when selecting for a flanking marker at 0.5 M distance. Without marker-assisted selection, a value of ~0.25 M is reached only in generation BC_4. The standard deviation of the length of the intact chromosome segment is distinctly smaller with marker-assisted selection in early backcross generations than without. For example, $\mathrm{SD}_1(x \mid R_{0,0.1,1}) = 0.015$ while without selection at the flanking marker $\mathrm{SD}_1(X) = 0.375$. In advanced backcross generations, the differences between the two schemes become smaller. However, an expected length of the intact chromosome segment of ~0.05 M, as reached in generation BC_1 with a flanking marker 0.1 M distant, is not reached even after 15 backcross generations without background selection.

Figure 2 shows the pdf's and cdf's of the length of the intact chromosome segment attached on one side of the target gene in generation BC_5 for backcross programs with background selection at a flanking marker positioned at $y = 0.1$ or $y = 0.5$ M distant from the target gene, when a recombinant is selected in generations BC_1 to BC_5. The pdf for selection in generation BC_5 at a marker with distance $y = 0.1$ M has only a small negative slope, whereas for selection in generation BC_1 the pdf for large attached chromosome segments (x is near y) is only about half the absolute value of the pdf of small

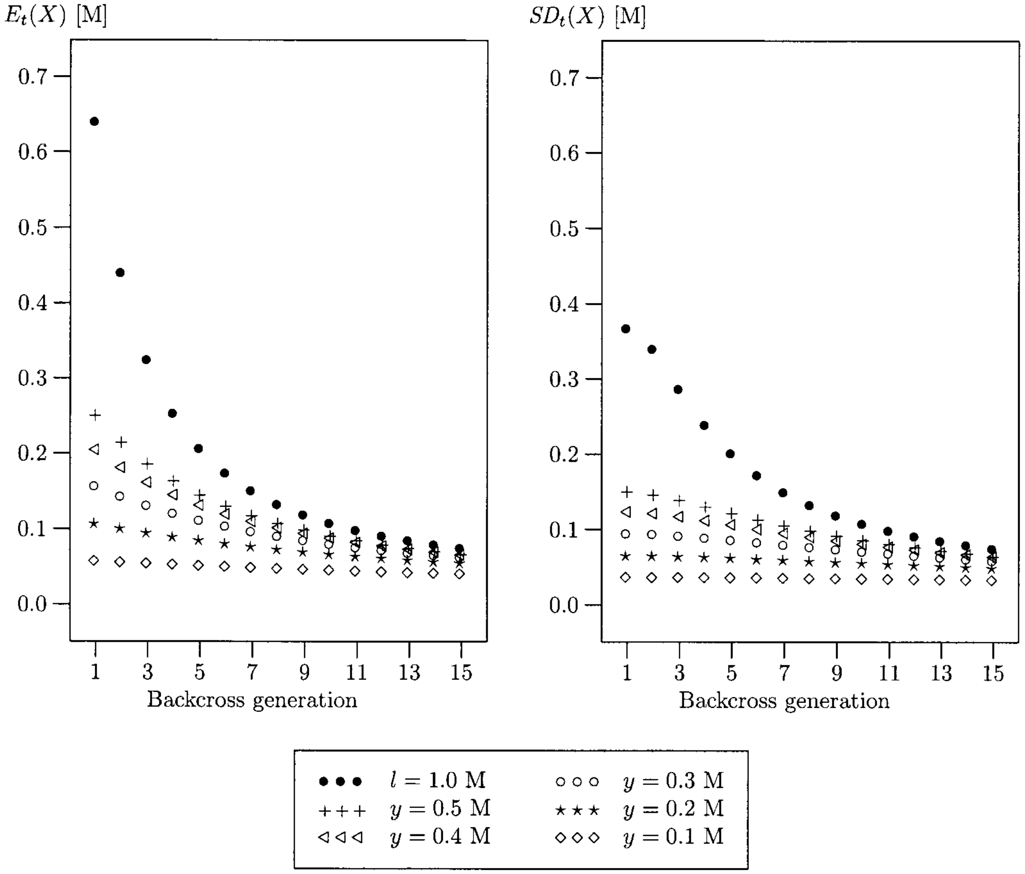

Figure 1.—Expected value [$E_t(X)$, left] and standard deviation [$\mathrm{SD}_t(X)$, right] of the length of the intact donor chromosome segment attached on one side of the target gene in generations BC_t ($t = 1 \ldots 15$) (1) in the absence of marker-assisted selection ($l = 1$) and (2) when selection for recombinants at a flanking marker with distance $y = 0.1$, $y = 0.2$, $y = 0.3$, $y = 0.4$, or $y = 0.5$ M is carried out in generation BC_1.

linked segments (x is near 0). The effect of the generation of selection on the difference in the pdf for small compared with large x values is even greater for a marker at position $y = 0.5$ M. Here, the probability of having large attached chromosome segments (x is near y) is almost zero when selecting for the flanking marker in generation BC_1. The cdf for selection in generation BC_1 is greater than for selection in BC_2 to BC_5, the difference being larger for $y = 0.5$ M than for $y = 0.1$ M. Consequently, the probability of having a smaller intact chromosome segment is greater with selection in an early compared with a late generation.

TOTAL LENGTH OF THE INTACT DONOR CHROMOSOME SEGMENT

In this section we use an abbreviated notation. The two sides of the target locus are named a and b, which are used as subscripts to mark parameters for the respective side. Parameters without subscript a or b refer to sums of both sides: $X = X_a + X_b$, $l = l_a + l_b$, $y = y_a + y_b$. The subscripts for the generation were dropped, and $f_a(x_a)$ and $f_b(x_b)$ can be any of the previously derived pdf's, referring to a certain backcross generation t. The events R_a and R_b denote that recombination between the target gene and the marker occurred on the respective side of the target gene in backcross generation $s \leq t$; N_a and N_b denote that no recombination occurred in any backcross generation $s \leq t$. Note that the generation s, in which recombination occurred, can be different for sides a and b.

Expected values and variances: Calculation of expected values and variances is straightforward due to the stochastic independence of X_a and X_b:

$$E(X) = E(X_a) + E(X_b) \tag{31}$$

$$V(X) = V(X_a) + V(X_b). \tag{32}$$

For derivation of the pdf's we distinguish the following three cases. The cdf's can be obtained by integrating the pdf's.

Case 1. Recombination on both sides of the target gene: Under condition $R_a \cap R_b$ both random variables X_a and X_b are continuous [*i.e.*, f_a and f_b are either $f_t(x|R_{0,y,t})$ or $f_t(x|R_{0,y,s})$]. Without loss of generality, we assume $y_a \geq y_b$. Because of the stochastic independence of X_a and X_b, the joint density of (x_a, x_b) is calculated by multiplying the marginal densities. Consider a certain length x of the intact donor chromosome segment: from $x = x_a + x_b$ follows $x_b = x - x_a$. Hence, the probability density for any x is obtained by integration of the joint density of $(x_a, x - x_a)$ over all possible values for x_a that result in $x = x_a + x_b$. We denote this integral by

$$i(\alpha, \beta) = \int_\alpha^\beta f_a(x_a) f_b(x - x_a)\, dx_a. \tag{33}$$

The limits of integration, *i.e.*, the minimum and maxi-

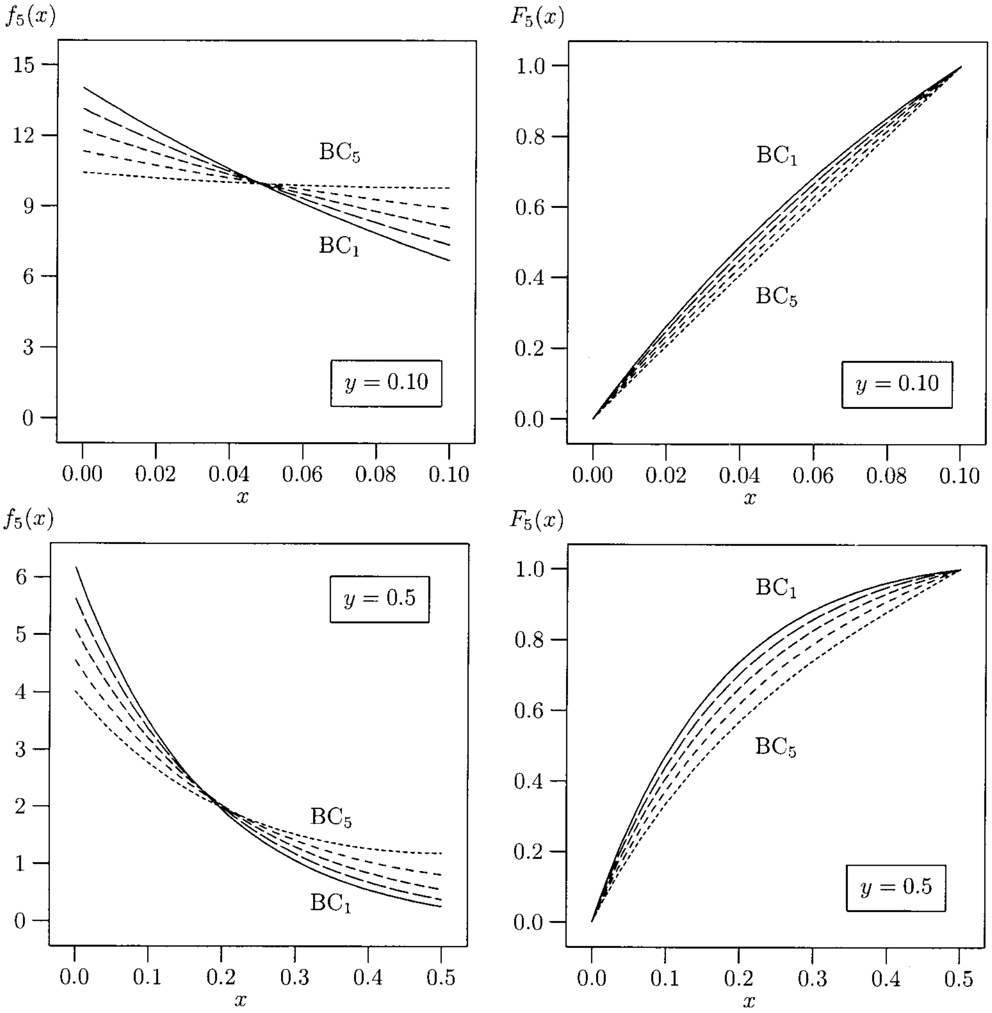

FIGURE 2.—Probability density functions [$f_5(x)$, left] and cumulative density functions [$F_5(x)$, right] of the length of the intact donor chromosome segment attached on one side of the target gene in generation BC$_5$, when recombinants at a flanking marker with distance $y = 0.1$ M (top) and $y = 0.5$ M (bottom) were detected and selected in generation BC$_1$ (solid line) and generations BC$_2$–BC$_5$ (subsequent dashed lines).

mum length of the chromosome segment attached on side a, depend on the total length of the intact donor chromosome segment x. For $x < y_b$ the whole donor chromosome segment may be on either side of the target gene; hence, x_a can range from 0 to x. If $y_b \leq x$, the length on side a has to be at least $x - y_b$. The maximum length of x_a under condition R_a is y_a. Hence, the pdf can be written in terms of i as

$$f(x|R_a \cap R_b) = \begin{cases} i(0, x) & \text{for } x \in [0, y_b) \\ i(x - y_b, x) & \text{for } x \in [y_b, y_a) \\ i(x - y_b, y_a) & \text{for } x \in [y_a, y) \\ 0 & \text{for } x \in [y, l). \end{cases} \qquad (34)$$

This principle is illustrated in APPENDIX B, using as an example the pdf of the length of the intact donor chromosome segment around the target gene in generation BC$_1$, when selection is for recombinants at flanking markers on both sides.

Case 2. Recombination on one side of the target gene: Without loss of generality we consider $N_a \cap R_b$ [*i.e.*, f_a is $f_t(x|N_{0,y,t})$ and f_b is either $f_t(x|R_{0,y,t})$ or $f_t(x|R_{0,y,s})$] and distinguish $l_a > y_b$ and $l_a \leq y_b$. For $l_a > y_b$, the value of x_a can range between 0 and x if $x < y_b$; otherwise, the minimum of x_a is $x - y_b$ and the maximum is l_a. If $x > l_a$, it is also possible that no recombination on side a occurred ($X_a = l_a$) and the probability density for this case adds to the integral because of the discrete character of X_a for $x_a = l_a$. Hence, we obtain for $l_a > y_b$

$$f(x|N_a \cap R_b) = \begin{cases} i(0, x) & \text{for } x \in [0, y_b) \\ i(x - y_b, x) & \text{for } x \in [y_b, l_a) \\ i(x - y_b, l_a) & \\ \quad + P(X_a = l_a)f_b(x - l_a) & \text{for } x \in [l_a, l_a + y_a) \\ 0 & \text{for } x \in [l_a + y_a, l). \end{cases} \qquad (35)$$

In analogy, the pdf for $l_a \leq y_b$ is

$$f(x|N_a \cap R_b) = \begin{cases} i(0, x) & \text{for } x \in [0, l_a) \\ i(0, l_a) & \\ \quad + P(X_a = l_a)f_b(x - l_a) & \text{for } x \in [l_a, y_b) \\ i(x - y_b, l_a) & \\ \quad + P(X_a = l_a)f_b(x - l_a) & \text{for } x \in [y_b, l_a + y_b) \\ 0 & \text{for } x \in [l_a + y_b, l). \end{cases} \qquad (36)$$

Note that for condition N_a the function f_a is defined depending on the value of x (Equation 18). This must be taken into account for the integration.

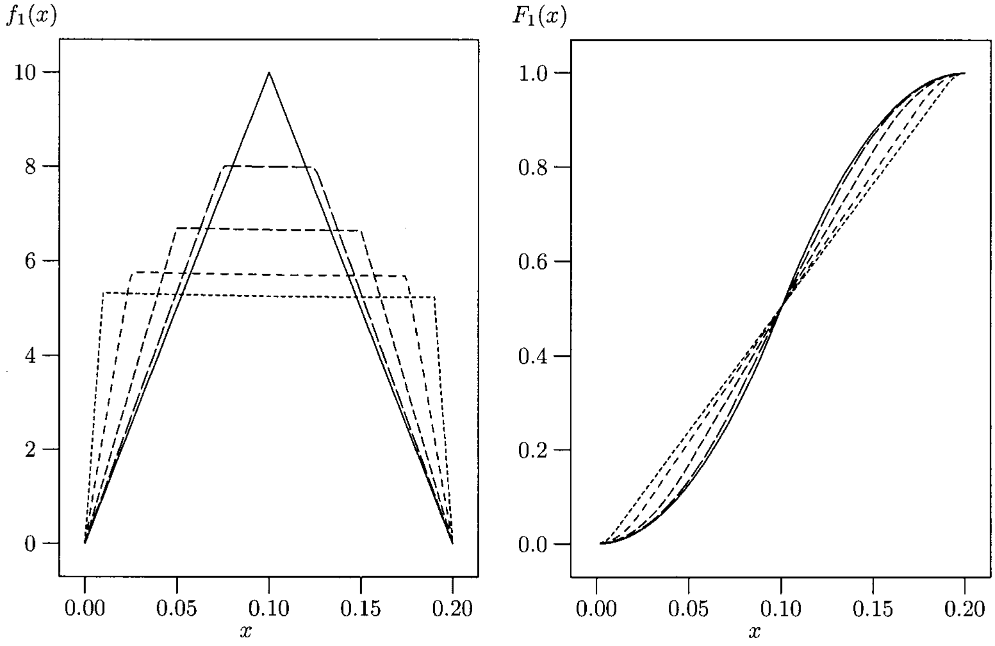

Figure 3.—Probability density functions [$f_1(x)$, left] and cumulative density functions [$F_1(x)$, right] of the intact donor chromosome segment around the target gene in generation BC$_1$ when recombinants were detected simultaneously at two flanking markers with distances (y_a = 0.1, y_b = 0.1) (solid line) or (y_a = 0.125, y_b = 0.075), (y_a = 0.15, y_b = 0.05), (y_a = 0.175, y_b = 0.025), (y_a = 0.19, y_b = 0.01) (subsequent dashed lines).

Case 3. No recombination on either side: We have N_a and N_b [*i.e.*, f_a and f_b are $f_l(x|N_{0,y,l})$]. Without loss of generality we assume $l_a \geq l_b$. In analogy to the above cases, the pdf of the length of the intact chromosome segment can be derived as

$$f(x|N_a \cap N_b) = \begin{cases} i(0, x) & \text{for } x \in [0, l_b) \\ i(x - l_b, x) & \\ \quad + f_a(x - l_b)P(X_b = l_b) & \text{for } x \in [l_b, l_a) \\ i(x - l_b, x) & \\ \quad + f_a(x - l_b)P(X_b = l_b) & \\ \quad + P(X_a = l_a)f_b(x - l_a) & \text{for } x \in [l_a, l). \end{cases} \tag{37}$$

The random variable X is discrete for $x = l$:

$$P(X = l|N_a \cap N_b) = P(X_a = l_a)P(X_b = l_b). \tag{38}$$

Numerical results: Figure 3 shows the pdf's and cdf's of the total length of the intact donor chromosome segment around a target gene for various combinations of marker distances y_a and y_b ,which sum up to 0.2 M. As reflected by the shape of the pdf's, for asymmetric marker bracket there is a high chance of detecting recombinants with a medium length of the intact chromosome segment, while for an asymmetric marker bracket, the probability of finding recombinants with a short or long intact chromosome segment increases.

DISCUSSION

Genetic model: Following earlier studies (Fisher 1949; Hanson 1959; Stam and Zeven 1981; Hospital *et al.* 1992), we used a Poisson process for modeling crossover formation during meiosis, as proposed by Haldane (1919). It is well known that the Poisson process is a simplified model of crossover formation because of the assumption of no interference (Stam 1979). Since Haldane's pioneering article, numerous researchers (*e.g.*, Kosambi 1944; Karlin and Liberman 1978, 1979; Zhao and Speed 1996; Browning 2000) proposed alternative mathematical models, which include interference. Most of the resulting map functions can be modeled by a stationary renewal process, the interevent distribution of which can be approximated by gamma distributions (Zhao and Speed 1996). McPeek and Speed (1995) compared the fit of various crossover formation models and concluded that gamma interevent distribution fit best the Drosophila dataset of Morgan *et al.* (1935).

We used Haldane's (1919) Poisson model due to its mathematical simplicity, its exponential interevent distribution, and the stochastic independence of crossover formations in adjacent chromosome regions, which allowed us to derive closed analytical formulas for the problems addressed in this article. Applying gamma interevent distributions would in most instances yield unwieldy formulas, which could only be numerically approximated. Moreover, as pointed out by Stam and Zeven (1981), dropping the assumption of no interference would reduce the generality of the presented approach because for each target gene it would be necessary to know the type and degree of interference.

Under the assumption of positive chiasma interference (Stam 1979), multiple crossovers in a given chromosome region occur less frequently than under the assumption of no interference. Consequently, if the target gene is located in a region with positive interference, the cdf of the chromosome length attached at one side of the target gene is greater than the presented cdf and the expected values are underestimated. The reverse holds true under the assumption of negative interference. In conclusion, the reader should be aware that the presented model (as most mathematical models of biological systems) is not capable of capturing every detail of the underlying biological process and the presented results should be interpreted with this in mind.

Comparison with earlier studies: Our results for marker-assisted selection of recombinants can readily be used to derive the cdf, when selection is only for presence of the target gene. For any $y \in (0, l)$ and generation BC_t, the disjoint events $R_{0,y,i}$ $(i = 1, \ldots, t)$ and $N_{0,y,t}$ represent a mutually exclusive partition of the entire probability space. Using the theorem of total probability, we obtain for $y < x < l$:

$$P(\{X_t > x\}) = \sum_{i=1}^{t} [P(\{X_t > x\}|R_{0,y,i})P(R_{0,y,i})] + P(\{X_t > x\}|N_{0,y,t})P(N_{0,y,t}). \tag{39}$$

Inserting Equations 8, 9, 15, and 20, we obtain FISHER's (1949, p. 50) probability p presented in the Introduction, which is also the basis of HANSON's (1959) formula. Summarizing, the relation between the three studies can be described as follows: HANSON (1959) corrected FISHER's (1949) results for a finite length of the chromosome. We extended HANSON's results to the case of marker-assisted selection for recovery of the recurrent parent genome at markers flanking the target locus.

HOSPITAL *et al.* (1992) extended earlier results of STAM and ZEVEN (1981) and investigated backcrossing with background selection at exactly two markers on the carrier chromosome, one on each side of the target gene. They derived the expected donor genome proportion on the carrier chromosome and derived an equation to calculate the marker positions, which maximize the expected donor genome proportion. However, the application of their approach in practical breeding programs is limited, because usually several markers on the carrier chromosome are available and used for background selection (see *e.g.*, RAGOT *et al.* 1995). In this case, the markers flanking the target gene are used to control the intact chromosome segment around the target gene and more distant markers are used to control the parental origin of the remainder of the chromosome. In contrast to the study of HOSPITAL *et al.* (1992), our approach considers only the donor chromosome segment directly linked to the target gene, but not those on the remainder of the carrier chromosome, and, hence, can be applied in backcross programs with more than two markers on the carrier chromosome. Furthermore, it yields a complete description of the underlying probability distribution.

Applications of the theory: Marker-assisted backcrossing is applied to the following tasks in breeding and genetic research: (1) transfer of a target gene, which may be a transgene or another major gene (*e.g.*, a disease resistance gene); (2) transfer of a chromosome region, which contains a favorable allele at a putative quantitative trait locus (QTL); and (3) development of near-isogenic lines (NILs). Our theoretical results can be applied to the experimental design of such backcross programs and for monitoring the length of the attached chromosome segment in various generations.

Transfer of a gene: To optimize marker-assisted selection for transfer of a gene, FRISCH *et al.* (1999a) proposed selection of backcross individuals on the basis of the ordering of genotypes,

$$y_a^- z^+ y_b^- \succ y_a^- z^+ y_b^+ \succ y_a^+ z^+ y_b^- \succ y_a^+ z^+ y_b^+, \tag{40}$$

where z^+, y_a^+, y_b^+ denote heterozygosity at the target locus and two flanking markers at distance y_a and y_b, respectively, and y_b^-, y_b^- denote homozygosity for the recurrent parent allele at the respective loci. Without loss of generality we assume $y_a \leq y_b$. If several individuals of the most preferable genotype (according to the above ordering) are found, selection of the best among them is based on a selection index calculated from the genotype at additional markers on the carrier chromosome of the target gene and on the noncarrier chromosomes as proposed by FRISCH *et al.* (1999b).

Before starting a t-generation backcross program, our results can be used to determine *a priori* the effect of the population size $n_1, \ldots, n_t$ in generations BC_1 to BC_t and the flanking marker distances y_a and y_b on the probability distribution of the intact chromosome segment in the selected individual in generation BC_t. The pdf of the attached chromosome segment length on one side of the target gene in generation BC_t is a mixture of the conditional pdf's for selection of a recombinant in one of generations BC_1 to BC_t and the conditional pdf for the case that no recombinant is selected up to generation BC_t. The respective weights are calculated from the multinomial distribution, following the principle described in detail in Equations 37–39 of FRISCH *et al.* (1999a). Hence, the pdf of the attached chromosome segment on side $c \in \{a, b\}$ in generation BC_t is

$$f_t(x_c) = \begin{cases} \sum_{i=1}^{t} \left[\prod_{0<j<i} (1 - \gamma_i) \right] \gamma_i \, f_t(x_c|R_{0,y_c,i}) & \\ \quad + \left[\prod_{1 \leq j \leq t} (1 - \gamma_i) \right] f_t(x_c|N_{0,y_c,i}) & \text{for } x_c \in [0, y_c) \\ f_t(x_c|N_{0,y_c,i}) & \text{for } x_c \in [y_c, l), \end{cases} \tag{41}$$

where

$$\gamma_i = \begin{cases} 1 - P(E_{0,y_a,i})^{n_i} & \text{for } c = a \\ P(E_{0,y_a,i})^{n_i}[1 - P(E_{0,y_b,i})^{n_i}] & \\ \quad + 1 - [P(E_{0,y_a,i})P(E_{0,y_b,i})]^{n_i} & \text{for } c = b. \end{cases} \tag{42}$$

The probability that the attached chromosome segment comprises the complete distance between the target gene and the end of the chromosome is

$$P(X_c = l) = P\left(\bigcap_{i=1}^{t} Z_{0,l_c,i}\right) = e^{-tl_c}. \qquad (43)$$

The cdf, expectation, and variance can be obtained from Equations 41 and 43 with standard methods, and the distribution of the total length of the intact chromosome segment is obtained according to the principle described in theory. These formulas can be used before starting the backcross program to calculate the following:

1. The expected length of the intact chromosome segment for given flanking marker distances y_a, y_b, and population sizes n_t;
2. the population sizes $n_1, \ldots, n_t$ required for given flanking marker distances y_a and y_b to obtain a desired value for the expected intact donor chromosome segment length, or to obtain with a given probability α an intact chromosome segment length shorter than a value u by using $F(u) \leq \alpha$; and
3. the flanking marker distances y_a and y_b required for given population sizes n_t to obtain a desired value for the expected intact donor chromosome segment length, or to obtain with a given probability α an intact chromosome segment length shorter than a value u by using $F(u) \leq \alpha$.

During the breeding program, our results can be used to infer the length of the intact chromosome segment from the known genotype of an individual (*a posteriori* situation). We illustrate this by a three-generation backcross program. Consider a single recombinant in generation BC$_1$. On the side of no recombination, the probability distribution of the length of the chromosome segment attached to the target gene is described by the equations derived in *Generations prior to detection of a recombinant*, whereas on the side of recombination, the results derived in *Generation in which a recombinant is detected and selected* apply. These results also apply in generation BC$_2$ to the second side of the target gene, when recombination occurs. The results derived in *Subsequent generations after selection of a recombinant* apply in generation BC$_2$ to the side, where recombination occurred already in generation BC$_1$, as well as to both sides of the target gene in more advanced backcross generations. Consequently, the given formulas allow a complete description of the length of the chromosome segment attached to the target gene in such a backcross program.

Introgression of favorable alleles at quantitative trait loci: Marker-assisted selection in QTL introgression usually comprises selection for presence of the donor allele at two markers z_a and z_b delimiting the interval in which the putative QTL was detected and for the recurrent parent allele at markers y_a and y_b flanking the QTL interval $[z_a, z_b]$ (Hospital and Charcosset 1997). Obviously, the formulas presented in Segment attached on one side of the target gene apply to the chromosome segment attached to markers z_a and z_b on the remote side of the QTL. Hence, our results can be applied in QTL introgression programs to reduce the donor chromosome segment attached to a QTL interval in exactly the same way as described for the transfer of a target gene.

Development of near-isogenic lines: A set of NILs, of which each line differs from any other line in one chromosome region, can be employed for confirmation, reanalysis, and fine mapping of QTL (Eshed and Zamir 1995). To generate such a set of NILs, recurrent backcrossing is carried out with a set of individuals that carry the donor alleles at different markers covering the whole genome. As for transfer of a target gene or QTL introgression, the derived formulas apply to the length of the donor chromosome segment attached to a marker at which selection is carried out for the donor allele, when selection at a flanking marker is for the recipient allele. In addition to the applications described above, our results can be used to calculate the probability distribution of the length of overlapping chromosome segments for two NILs.

Selection of several individuals and application in animal breeding: In developing the presented theory, we assumed that in each generation one individual was backcrossed to the recurrent parent. However, especially in an animal breeding context lower selection intensities may be desirable, for example, by backcrossing all recombinant individuals recovered in a backcross population. Since our results on pdf and cdf are valid irrespective of the number of recombinant individuals selected per generation, they also apply to such breeding plans.

Furthermore, our approach can be extended to derive the distribution of the intact donor chromosome segment in the "best" of several recombinant individuals for two important special cases using results from order statistics. The latter requires the stochastic independence of the length of the intact donor chromosome segment for the individuals under consideration. This holds true (a) for BC$_1$ populations or (b) in advanced backcross generations s, if each BC$_s$ individual traces back to a different ancestor in generation BC$_1$. Consider one side of the target gene and suppose that m recombinant individuals are found. Then, the pdf of the first order statistic is obtained as

$$g_1(x) = m[1 - F_s(x|R_{0,y,s})]^{m-1} f_s(x|R_{0,y,s}) \qquad (44)$$

(Shao 1999, p. 72), which yields

$$g_1(x) = \begin{cases} \dfrac{m \sinh^{m-1}(y - x)\cosh^{ms-m-1}[s \cosh^2(y - x) - (s - 1)]}{\sinh^m y \cosh^{m(s-1)} y} \\ \quad \text{for } x \in [0, y) \\ 0 \quad \text{for } x \in [y, l]. \end{cases} \qquad (45)$$

This result can be used to calculate the moments of the first order statistic, referring to the length of the attached donor chromosome segment in the recombinant individual with the shortest segment in a sample of *m*.

Placement of flanking markers: If several markers on both sides of the target gene are available, it is of interest to compare the effect of symmetric versus asymmetric placement of flanking markers on the intact donor chromosome segment length. As reflected by the shape of the pdf's shown in Figure 3, for a symmetric marker bracket there is a high chance of detecting recombinants with a medium length of the intact chromosome segment, while for an asymmetric marker bracket, the probability of finding recombinants with a short or long intact chromosome segment increases. Larger population sizes are required in a backcross program with an asymmetric rather than a symmetric marker bracket (Frisch *et al.* 1999a). Consequently, symmetric marker brackets are preferable, especially when the population size is a limiting factor. In addition to requiring fewer individuals, the probability that a recovered recombinant has a relatively large intact donor chromosome segment is lower than for an asymmetric marker bracket.

Generation of selection: A marker-assisted backcross program usually comprises three or more generations. Hence, it is of interest to compare the effect of the generation in which a recombinant is selected on the intact donor chromosome segment length in the final breeding product. The probability of having a smaller intact chromosome segment is greater with selection in an early generation than with selection in an advanced generation (Figure 2), because crossover events in subsequent generations after selection may result in a further reduction of the intact chromosome segment. The shape of the pdf of *X* in the final backcross generation (Figure 2) reveals that with a closely linked flanking marker ($y = 0.1$ M) and selection in an advanced generation, individuals with a short intact chromosome segment occur almost as frequently as individuals with a long intact segment (compared with *y*). However, with increasing marker distance ($y = 0.5$ M) and selection in early generations, the chance of recovering individuals with relatively short segments is considerably increased.

These results show that in practical breeding programs selection of recombinants between marker and target in early generations is not only advantageous with respect to the resources required (Frisch *et al.* 1999b) but also with respect to obtaining a short intact donor chromosome segment around the target gene.

Donor and recipient are elite: In backcross programs for transfer of a desirable gene from one elite line to another, it is not necessary to have a maximum reduction of the attached chromosome segment because tight linkage of undesirable traits is unlikely and there may be even positive effects caused by the attached chromosome segment (Lee 1995). However, introgression of the complete carrier chromosome is also not desirable because this most likely affects the phenotypic characteristics of the recipient line.

In such a breeding program selection for recombinants between the target gene and a flanking marker is effective even when the marker is fairly distant from the target gene. For example, a saving of three backcross generations concerning the expected length of the linked chromosome segment is realized with a marker distance of $y = 0.5$ M (Figure 1). The considerably reduced standard deviation of the linked chromosome segment length with background selection compared to selection only for the target gene (Figure 1) reflects the fact that without marker-assisted selection large intact segments occur quite frequently in early generations. This is due to the absence of crossover events between the target gene and the end of the chromosome and results in the undesired introgression of large intact donor chromosome segments.

Because recombinants between the target gene and fairly distant flanking markers occur with a high probability even in small backcross populations (Frisch *et al.* 1999a), marker-assisted background selection can be used to avoid large intact chromosome segments in transfer of genes between elite lines, even with limited resources for the population size and marker analyses.

Donor is unadapted and recipient is elite: In a backcross program for transfer of a target gene from unadapted material into breeding material used for variety development, a short attached chromosome segment is important. In a classical backcross program more than the generally recommended six backcross generations are required in this case (Fehr 1987, p. 375). In marker-assisted backcross programs, an effective reduction can be achieved by selection for the recurrent parent alleles at tightly linked flanking markers. For example, in the numerical example shown in Figure 1, background selection at a marker with distance $y = 0.1$ M yields a shorter expected attached chromosome segment than 15 generations of backcrossing without background selection.

In a backcross program with tightly linked flanking markers, the sequential analysis of markers surrounding the target gene can assure an economic use of resources: First, a fairly distant flanking marker is analyzed. Assuming a given population size, its distance from the target locus can be determined such that with a high probability at least one single or double recombinant is found (Frisch *et al.* 1999a, Equations 11–13). If several recombinants are found, subsequent analysis of more tightly linked markers can be used to find the individual with the shortest intact chromosome segment. The results given in appendix c apply to this scenario and can be used to monitor the probability distribution of the attached chromosome segment.

Further research needs: Especially in early backcross generations, donor chromosome segments not directly attached to the target gene contribute a substantial amount to the total fraction of the undesirable donor genome in a backcross individual. We are currently investigating whether our approach can be extended to obtain a complete description of the distribution of the total donor genome proportion for a given marker genotype at several markers distributed throughout the genome.

We greatly appreciate the suggestions and comments of two anonymous reviewers, which helped to improve this article.

LITERATURE CITED

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

Browning, S., 2000 The relationship between count-location and stationary renewal models for the chiasma process. Genetics **155:** 1955–1960.

Eshed, Y., and D. Zamir, 1995 An introgression line population of *Lycopersicon pennellii* in the cultivated tomato enables the identification and fine mapping of yield associated QTL. Genetics **141:** 1147–1162.

Fehr, W. R., 1987 *Principles of Cultivar Development, Vol. 1: Theory and Technique*. Macmillan, New York.

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.

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.

Karlin, S., and U. Liberman, 1978 Classification of multilocus recombination distributions. Proc. Nat. Acad. Sci. USA **75:** 6332–6336.

Karlin, S., and U. Liberman, 1979 A natural class of multilocus recombination processes and related measures of crossover interference. Adv. Appl. Prob. **11:** 479–501.

Kosambi, D. D., 1944 The estimation of the map distance from recombination values. Ann. Eugen. **12:** 172–175.

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

McPeek, M. S., and T. P. Speed, 1995 Modeling interference in genetic recombination. Genetics **139:** 1031–1044.

Morgan, T. H., C. B. Bridges and J. Schulz, 1935 Constitution of the germinal material in relation to heredity. Carnegie Inst. Washington Publ. **34:** 284–291.

Naveira, H., and A. Barbadilla, 1992 The theoretical distribution of lengths of intact chromosome segments around a locus held heterozygous with backcrossing in a diploid species. Genetics **130:** 205–209.

Ragot, M., M. Biasiolli, M. F. Delbut, A. Dell'Orco, L. Malgarini *et al.*, 1995 Marker-assisted backcrossing: a practical example, pp. 45–56 in *Techniques et Utilisations des Marqueurs Moleculaires*. Montepellier, France. March 29–31, 1994. Institut National de la Recherche Agronomique, Paris.

Shao, J., 1999 *Mathematical Statistics*. Springer-Verlag, New York.

Stam, P., 1979 Interference in genetic crossing over and chromosome mapping. Genetics **92:** 573–594.

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

Tanksley, S. D., 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.

Young, N. D., and S. D. Tanksley, 1989 RFLP analysis of the size of chromosomal segments retained around the *Tm-2* locus of tomato during backcross breeding. Theor. Appl. Genet. **77:** 353–359.

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

Zhao, H., and T. P. Speed, 1996 On genetic map functions. Genetics **142:** 1369–1377.

Communicating editor: Z-B. Zeng

APPENDIX A

Calculation of $E_t(X|R_{0,y,1})$: We demonstrate calculation of the expected length of the intact chromosome segment attached on one side of the target gene in generation BC$_t$, when selection for a recombinant individual was performed in generation BC$_1$ ($s = 1$). The results in Table 1 are obtained with analogous calculations.

Inserting $s = 1$ in Equation 24 yields the pdf

$$f_t(x|R_{0,y,1}) = \frac{\cosh(y - x) + (t - 1)\sinh(y - x)}{e^{(t-1)x}\sinh y} = \frac{-te^{-tx} - (2 - t)e^{(2-t)x-2y}}{e^{-2y} - 1} \qquad \text{(A1)}$$

for $x \in [0, y)$. Calculation of the expected value requires integration:

$$E_t(X|R_{0,y,1}) = \int_0^y x f_t(x|R_{0,y,1})\,dx = \frac{te^{-2y} - 2e^{-yt} - t + 2}{t(t - 2)(e^{-2y} - 1)} = \frac{1}{t} - \frac{2e^{-2y}(1 - e^{-(t-2)y})}{t(t - 2)(1 - e^{-2y})}. \qquad \text{(A2)}$$

Equation A2 is not defined for $t = 2$ because the denominator and nominator are 0. With the rule of l'Hospital we calculate the limit of $E_t(X|R_{0,y,1})$ for $t \rightarrow 2$ and obtain

$$E_2(X|R_{0,y,1}) = \frac{(1 + 2y)e^{-2y} - 1}{2(e^{-2y} - 1)} = \frac{1}{2} - \frac{ye^{-2y}}{1 - e^{-2y}}. \qquad \text{(A3)}$$

APPENDIX B

Calculation of $f(x|R_a \cap R_b)$: We illustrate calculation of the pdf of the length of the intact donor chromosome segment around the target gene in generation BC$_1$, when recombination between the target gene and markers occurred on both sides. From Equation 21 we obtain

$$f_a(x_a|R_a) = \frac{\cosh(y_a - x_a)}{\sinh y_a} \quad \text{for } x_a \in [0, y_a) \tag{B1}$$

$$f_b(x_b|R_b) = \frac{\cosh(y_b - x_b)}{\sinh y_b} \quad \text{for } x_b \in [0, y_b). \tag{B2}$$

For calculation of the pdf of $X = X_a + X_b$, we need the integral $i(\alpha, \beta)$ (Equation 33). First we calculate the indefinite integral

$$\phi(x_a) = \int \frac{\cosh(y_a - x_a)\cosh(y_b - x_b)}{\sinh y_a \sinh y_b} dx_a. \tag{B3}$$

Substitution of $x_b = x - x_a$ and integrating yields

$$\phi(x_a) = \frac{2x_a\cosh(x - y_a - y_b) - \sinh(x + y_a - 2x_a - y_b)}{4 \sinh y_a \sinh y_b}. \tag{B4}$$

With ϕ, the integral $i(\alpha, \beta) = \phi(\alpha) - \phi(\beta)$ and hence, we obtain the pdf

$$f(x|R_a \cap R_b) = \begin{cases} \phi(x) - \phi(0) & \text{for } x \in [0, y_b) \\ \phi(x) - \phi(x - y_b) & \text{for } x \in [y_b, y_a) \\ \phi(y_a) - \phi(x - y_b) & \text{for } x \in [y_a, y) \\ 0 & \text{for } x \in [y, l). \end{cases} \tag{B5}$$

APPENDIX C

Selection of recombinants without marker analyses in previous generations: In practical breeding programs, there are situations when a flanking marker is analyzed for the first time in an advanced backcross generation BC$_s$. Here, we derive the distribution of the length of the donor chromosome segment attached on one side of the target gene for two such cases. We define the event

$B_{u,v,s} = \overline{N}_{u,v,s} = \cup_{i=1}^{s} R_{u,v,i}$: Recombination between loci at positions u and v was observed in generation BC$_s$ but it is unknown in which generation BC$_i$ ($i \leq s$) recombination occurred for the first time.

The probability of event $B_{u,v,s}$ is

$$P(B_{u,v,s}) = 1 - P(N_{u,v,s}) = 1 - e^{-sd}\cosh^s d\,. \tag{C1}$$

Marker assay only in an advanced backcross generation, $B_{0,y,s}$: In this case, the distribution of the length of the chromosome segment attached on one side of the target gene is a mixture of the distributions under condition $P(R_{0,y,i})$ with weights

$$w_i = \frac{P(R_{0,y,i})}{P(B_{0,y,s})} = \frac{e^{-iy}\cosh^{(i-1)}y \sinh y}{1 - e^{-sd}\cosh^s d}\,. \tag{C2}$$

Hence, we have for $x \in [0, y)$

$$F_t(x|B_{0,y,s}) = \sum_{i=1}^{s} w_i F_t(x|R_{0,y,i})$$

$$= 1 - \sum_{i=1}^{s} \frac{\sinh(y - x)\cosh^{(i-1)}(y - x)}{\sinh y \cosh^{(i-1)}y} e^{(i-t)x} \tag{C3}$$

$$= 1 - \frac{\sinh(y - x)}{\sinh y} e^{-tx} \frac{e^x - (\cosh(y - x)/\cosh y)^s e^{(s+1)x}}{1 - \cosh(y - x)/\cosh y\, e^x}, \tag{C4}$$

$$f_t(x|B_{0,y,s}) = \sum_{i=1}^{s} w_i f_t(x|R_{0,y,i})$$

$$= - \sum_{t=1}^{s}\left\{\frac{e^{(i-t)x}\cosh^{i-2}(y - x)}{\sinh y \cosh^{i-1}y} \times \left[\frac{1}{2}(i - t)\sinh(2y - 2x) + i \sinh^2(y - x) - 1\right]\right\}, \tag{C5}$$

$$E_t(X|B_{0,y,s}) = \sum_{i=1}^{s} w_i E_t(x|R_{0,y,i}), \tag{C6}$$

$$V_t(X|B_{0,y,s}) = \sum_{i=1}^{s} w_i[E_t(x|R_{0,y,i})]^2 + \sum_{i=1}^{s} w_i[V_t(x|R_{0,y,i})]^2 - [E_t(X|B_{0,y,i})]^2. \tag{C7}$$

Marker assay with more closely linked markers after detection of recombinants, $B_{0,y,s} \cap R_{0,y^*,s}$: In generation BC$_s$, recombination between the target gene and a marker at position y^*, which was analyzed in all previous generations, was observed for the first time. Recombination between the target gene and a second marker at position $y < y^*$, which was analyzed for the first time in generation BC$_s$, was also observed.

It is unknown in which generation BC$_i$ ($i \leq s$) recombination between the target gene and the marker at position y occurred. The distribution of the length of the chromosome segment attached on one side of the target is a mixture of the distributions under conditions $R_{0,y,i} \cap R_{0,y^*,s}$ with weights

$$w_i = \frac{P(R_{0,y,i} \cap R_{0,y^*,s})}{P(B_{0,y,s} \cap R_{0,y^*,s})}, \tag{C8}$$

where

$$B_{0,y,s} \cap R_{0,y^*,s} = \dot{\bigcup_{i=1}^{s}} (R_{0,y,i} \cap R_{0,y^*,s}) \tag{C9}$$

and

$$P(B_{0,y,s} \cap R_{0,y^*,s}) = e^{-sy^*}[\cosh^{s-1}y^* \sinh y^* - \cosh^s y \cosh^{s-1}(y^* - y)\sinh(y^* - y)]. \tag{C10}$$

Let BC$_z$ be the generation in which recombination between the target gene and the marker at position y^* occurred. It can be shown that, for $z = s$ and $x \in [0, y)$,

$$P(R_{0,y,s} \cap R_{0,y^*,s}) = e^{-sy^*}\cosh^{s-1}y \cosh^s(y^* - y)\sinh y, \tag{C11}$$

$$P(\{X_t > x\}|R_{0,y,s} \cap R_{0,y^*,s}) = \frac{\cosh^{s-1}(y - x)\sinh(y - x)}{\cosh^{s-1}y \sinh y} e^{-(t-s)x}, \qquad \text{(C12)}$$

$$\frac{\partial P(\{X_t > x\}|R_{0,y,s} \cap R_{0,y^*,s})}{\partial x} = \frac{e^{(s-t)x}\cosh^{s-2}(y - x)}{\sinh y \cosh^{s-1} y} \times \left[\frac{1}{2}(s - t)\sinh(2y - 2x) + s\sinh^2(y - x) - 1\right], \qquad \text{(C13)}$$

and, for $z \neq s$ and $x \in [0, y)$,

$$P(R_{0,y,z} \cap R_{0,y^*,s}) = e^{-sy^*}\cosh^{z-1}y \cosh^{z-1}(y^* - y)\cosh^{s-z-1}y^* \times \sinh y \sinh(y^* - y)\sinh y^*, \qquad \text{(C14)}$$

$$P(\{X_t > x\}|R_{0,y,z} \cap R_{0,y^*,s}) = e^{-(t-s)x} \frac{\cosh^{z-1}(y - x)\cosh^{s-z-1}(y^* - x)\sinh(y - x)\sinh(y^* - x)}{\cosh^{z-1}y \cosh^{s-z-1}y^* \sinh y \sinh y^*}, \qquad \text{(C15)}$$

$$\begin{aligned}\frac{\partial P(\{X_t > x\}|R_{0,y,z} \cap R_{0,y^*,s})}{\partial x} = {} & \frac{e^{(s-t)x}\cosh^{z-1}(x - y)\cosh^{s-z-1}(x - y^*)}{\cosh^{z-1}y \cosh^{s-z-1}y^* \sinh y \sinh y^*} \\ & \times [(s - t)\sinh(x - y)\sinh(x - y^*) \\ & + (z - 1)\sinh^2(x - y)\sinh(x - y^*) \\ & \times \cosh^{-1}(x - y) \\ & + (s - z - 1)\sinh^2(x - y^*) \\ & \times \sinh(x - y)\cosh^{-1}(x - y^*) \\ & + \cosh(x - y)\sinh(x - y^*) \\ & + \sinh(x - y)\cosh(x - y^*)].\end{aligned} \qquad \text{(C16)}$$

With this result we get

$$F_t(x|R_{0,y,i} \cap R_{0,y^*,s}) = 1 - P(\{X_t > x\}|R_{0,y,z} \cap R_{0,y^*,s}), \qquad \text{(C17)}$$

$$f_t(x|R_{0,y,i} \cap R_{0,y^*,s}) = -\frac{\partial P(\{X_t > x\}|R_{0,y,z} \cap R_{0,y^*,s})}{\partial x}, \qquad \text{(C18)}$$

and

$$\begin{aligned}F_t(x|B_{0,y,s} \cap R_{0,y^*,s}) &= \sum_{i=1}^{s} w_i F_t(x|R_{0,y,i} \cap R_{0,y^*,s}) \\ &= 1 - \sum_{i=1}^{s} P(\{X_t > x\}|R_{0,y,z} \cap R_{0,y^*s}),\end{aligned} \qquad \text{(C19)}$$

$$\begin{aligned}f_t(x|B_{0,y,s} \cap R_{0,y^*,s}) &= \sum_{i=1}^{s} w_i f_t(x|R_{0,y,i} \cap R_{0,y^*,s}) \\ &= -\sum_{i=1}^{s} \frac{\partial P(\{X_t > x\}|R_{0,y,z} \cap R_{0,y^*,s})}{\partial x}.\end{aligned} \qquad \text{(C20)}$$

Expectation and variance are calculated as

$$E_t(X|B_{0,y,s} \cap R_{0,y^*,s}) = \sum_{i=1}^{s} w_i E_t(x|R_{0,y,i} \cap R_{0,y^*,s}), \qquad \text{(C21)}$$

$$\begin{aligned}V_t(X|B_{0,y,s} \cap R_{0,y^*,s}) = {} & \sum_{i=1}^{s} w_i[E_t(x|R_{0,y,i} \cap R_{0,y^*,s})]^2 \\ & + \sum_{i=1}^{s} w_i[V_t(x|R_{0,y,i} \cap R_{0,y^*,s})]^2 \\ & - [E_t(X|B_{0,y,s} \cap R_{0,y^*,s})]^2.\end{aligned} \qquad \text{(C22)}$$