Genet. Res., Camb. (2005), **85**, *pp.* 151–157. *With* 1 *figure.* © 2005 Cambridge University Press
doi:10.1017/S0016672305007421 Printed in the United Kingdom

Using markers to reduce the variation in the genomic composition in marker-assisted backcrossing

BERTRAND SERVIN*
UMR de Génétique Végétale INRA/CNRS/UPS/INPA-G Ferme du moulon 91190 Gif-sur-Yvette, France

(*Received 19 November 2003 and in revised form 9 February and 3 March 2005*)

Summary

Marker-assisted introgression or backcrossing is a widely used method to improve commercial breeding lines or study the effects of genes in a homogeneous genetic background. In this context, the recovery of the recipient parent genome is a major objective of backcrossing. Selection on markers has been shown to be very useful to accelerate the rate of recovery of the recipient parent genome in backcrossing. In this study we show how much information markers give on the true genetic composition of individuals by deriving the variance and estimating the distribution of the genetic composition of individuals sharing a known genotype at markers. These calculations enable predictions of the number of individuals carrying an ideal genotype at markers that must be produced to fulfil background selection objectives.

1. Introduction

Backcrossing is a widely used method for the improvement of varieties. The use of molecular markers to increase selection efficiency in so-called marker-assisted backcrossing (MAB) has been studied for a long time. Markers can be used to: (i) assess the presence of donor-type alleles at target genes (Melchinger, 1990; Hospital & Charcosset, 1997), (ii) reduce the length of the intact donor segment retained around the target gene (Frisch & Melchinger, 2001; Hospital, 2001) and (iii) accelerate the return to a fully recipient genotype on non-carrier chromosomes (e.g. Hospital *et al.*, 1992; Visscher *et al.*, 1996). Basically, the principle of MAB is to define, prior to selection, an ideal genotype at markers (*ideotype*) and then to perform generations of backcrossing while selecting individuals on their genotype at markers to obtain the ideotype as fast and cheaply as possible. In this context, markers are useful because they allow estimation of the proportion of recipient genome (PRG, also denoted as Π herein) of individuals. It is then interesting to determine the optimal number and positions of markers to use so that individuals obtained by selection on markers have recovered sufficient of the recipient parent genome.

In MAB, the recovery of the recipient parent genome is a major objective of selection. The extent to which this recovery is required may depend on the difference between the agronomic performances of the starting lines (donor and recipient parents). For example, if the donor line is of poor agronomic performance and only used to introduce a particular trait, such as resistance to diseases, the selection objective on the PRG is high. However, if the donor parent also has a good agronomic performance, the selection objective on the PRG might be lower.

In order to ensure a sufficient recovery of the recipient parent genome, markers to be used for selection are chosen and then individuals which carry homozygous recipient genotype at these markers are selected. This selection is particularly efficient on chromosomes that do not carry target genes (*non-carrier chromosomes*). The optimal number and positions of markers on non-carrier chromosomes has been studied by Visscher (1996) and then by Servin & Hospital (2002). Visscher (1996) has developed a method to compute the part of the variation in the genomic composition of individuals explained by markers. His method allows evaluation of how much information markers give regarding the genomic composition of individuals. However, this method assumes no selection on markers as the effect of selection is indeed hard to cope with when using that

Present address: Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195, USA. Tel: +1 (206) 5438265. e-mail: servin@stat.washington.edu

approach. Servin & Hospital (2002) used a different optimization criterion to compute the optimal positions of markers on non-carrier chromosomes. They assumed selection on markers is very efficient so that individuals carrying a fully recipient genotype at markers can be obtained. Hence, from the known genotype at markers, the principle of their method is to compute the average PRG ($\widehat{\Pi}$) of individuals and to find positioning of markers corresponding to the maximal value of $\widehat{\Pi}$. This allows the effect of selection to be taken partially into account as $\widehat{\Pi}$ is computed conditionally on the success of selection. However, their method does not take into account the variance of Π, which would give information on the probability that the selected individuals actually have recovered enough recipient genome. Here, we derive a method to compute the variance of the PRG of individuals sharing a known genotype at markers. We will denote the standard deviation of Π as SD_Π. We then derive a method to estimate the distribution function of Π, f_Π, and the cumulative distribution of Π, Φ_Π, in individuals carrying the ideotype at markers. This permits estimation of the number of individuals that must be produced to meet background selection objectives.

2. Methods

Here, we first show how to calculate the mean and variance of the PRG of an ideotype obtained after t generations of backcrossing. We then describe a method to obtain the empirical distribution of this PRG by Monte Carlo simulations.

(i) *Mean and variance of* Π

Let us first note that, as pairs of homologous chromosomes segregate independently, we only need to derive the mean and variance of the PRG for each pair in order to be able to compute the PRG on all non-carrier chromosomes.

We consider a pair of homologous non-carrier chromosomes of an ideotype in a population derived by t generations of backcrossing. The pair is composed of one chromosome inherited from the backcross population at generation $t-1$ (herein *segregating chromosome*), and a chromosome inherited from the recurrent parent (herein *non-segregating chromosome*).

By definition, the PRG of the non-segregating chromosome is 1. We denote by $\overline{Z}^{(t)}$ the PRG of the segregating chromosome. Then, the overall PRG for the chromosome pair, $\Pi^{(t)}$, is

$$\Pi^{(t)} = \frac{\overline{Z}^{(t)}+1}{2} \tag{1}$$

and we can calculate the mean and variance of $\Pi^{(t)}$ as

$$E(\Pi)^{(t)} = \frac{E(\overline{Z}^{(t)})+1}{2} = \widehat{\Pi}, \tag{2}$$

$$Var\ (\Pi^{(t)}) = \frac{1}{4} Var\ (\overline{Z}^{(t)}). \tag{3}$$

We compute the mean variance of the PRG on the segregating chromosome ($\overline{Z}^{(t)}$) as follows. Let $Z_i^{(t)}$ denote whether the ith locus at generation t originates from the recurrent parent ($Z_i^{(t)}=1$) or the donor parent ($Z_i^{(t)}=0$). Assuming a uniform distribution of loci on a chromosome of length L, with map positions z_i, we have

$$\overline{Z}^{(t)} = \frac{1}{L}\int_0^L Z_i^{(t)} dz_i. \tag{4}$$

From this, we can compute the mean of $\overline{Z}^{(t)}$ as

$$E(Z^{(t)}) = \frac{1}{L}\int_0^L E(Z_i^{(t)}) dz_i, \tag{5}$$

$$E(Z_i^{(t)}) = P(Z_i^{(t)}=1|M), \tag{6}$$

where $P(Z_i^{(t)}=1)$ is the probability that locus i originates from the recurrent parent, conditional on the genotype at markers M, here the ideotype.

When considering the ideotype at markers, $E(\overline{Z}^{(t)})$ and therefore $E(\Pi^{(t)})$ are maximized when markers are located at the optimal positions described by Servin & Hospital (2002). Note that the values for the PRG given in table 1 of Servin & Hospital (2002) are $E(\overline{Z}^{(t)})$ and not $E(\Pi^{(t)})$ as is, mistakenly, indicated.

The variance of $\overline{Z}^{(t)}$ has not been calculated yet. This variance is:

$$Var(\overline{Z}^{(t)}) = \frac{1}{L^2}\int_0^L\int_0^L Cov(Z_i^{(t)}, Z_j^{(t)})\ dz_i\, dz_j,$$

$$Var(\overline{Z}^{(t)}) = \frac{1}{L^2}\int_0^L\int_0^L [E(Z_i^{(t)}, Z_j^{(t)}) - E(Z_i^{(t)})E(Z_j^{(t)})]\ dz_i\, dz_j,$$

$$Var(\overline{Z}^{(t)}) = \frac{1}{L^2}\int_0^L\int_0^L P(Z_i^{(t)}=1, Z_j^{(t)}=1|M) - P(Z_i^{(t)}=1|M)P(Z_j^{(t)}=1|M)\ dz_i\, dz_j. \tag{7}$$

Again, we condition the probabilities of Z_i and the joint probability of Z_i and Z_j on the genotype at markers, i.e. the ideotype. These probabilities can be derived using the method of Visscher & Thompson (1995), for simple cases, or computed numerically

https://doi.org/10.1017/S0016672305007421 Published online by Cambridge University Press

using the MDM program (Servin *et al.*, 2002) in the general case. In the next section we show how the former approach can be used to calculate equation (7) in BC1, with one marker on a non-carrier chromosome.

(a) *Derivation of* $Var(\overline{Z}^{(1)})$ *with a single marker on the chromosome.*

We now show how to derive $Var(\overline{Z}^{(1)})$, when considering a single marker on the chromosome, at map position d. Denoting $Cov(Z_i^{(1)}, Z_j^{(1)})$ by c_{ij} for simplicity, equation (7) can then be written as

$$\begin{aligned} Var(Z^{(1)}) = \frac{1}{L^2}\Bigg[& \int_0^d \left(\int_{z_i}^d c_{ij}\, dz_j \right) dz_i + \int_0^d \left(\int_{z_j}^d c_{ij}\, dz_i \right) dz_j \\ & + \int_d^L \left(\int_{z_i}^L c_{ij}\, dz_j \right) dz_i + \int_d^L \left(\int_{z_j}^L c_{ij}\, dz_i \right) dz_j \\ & + \int_0^d \left(\int_d^L c_{ij}\, dz_j \right) dz_i + \int_0^d \left(\int_d^L c_{ij}\, dz_i \right) dz_j \Bigg]. \end{aligned} \tag{8}$$

The six terms in the sum relate to the six different possible orderings of loci, i, j and M on the chromosome: respectively, for each term from left to right, these orderings are $z_i \leqslant z_j \leqslant m$, $z_j \leqslant z_i \leqslant m$, $m \leqslant z_i \leqslant z_j$, $m \leqslant z_j \leqslant z_i$, $z_i \leqslant m \leqslant z_j$ and $z_j \leqslant m \leqslant z_i$. Gathering terms of the same value, we can simplify this expression to

$$Var(Z^{(1)}) = \frac{2}{L^2}(\gamma(d) + \gamma(L-d) + \gamma') \tag{9}$$

with

$$\gamma(x) = \int_0^x \left(\int_{z_i}^x c_{ij}\, dz_j \right) dz_i \tag{10}$$

and

$$\gamma' = \int_0^d \left(\int_d^L c_{ij}\, dz_j \right) dz_i \tag{11}$$

Computation of $\gamma(x)$. We denote the map position of the marker by x, which can be either d or $L-d$, depending on which side of the marker loci i and j are. In that case, the ordering of the loci on the chromosome is such that $z_i \leqslant z_j \leqslant x$ and we have

$$\begin{aligned} c_{ij} &= P(Z_i^{(1)} = 1, Z_j^{(1)} = 1 | M) - P(Z_i^{(1)} = 1 | M) P(Z_j^{(1)} = 1 | M) \\ &= (1 - r_{ij})(1 - r_{jM}) - (1 - r_{iM})(1 - r_{jM}) \\ &= (1 - r_{jM})(r_{iM} - r_{ij}) \end{aligned} \tag{12}$$

where r_{ab} denotes the recombination fraction between loci a and b. Using the Haldane mapping function to relate recombination fraction to map locations we have

$$r_{ij} = 1/2(1 - \exp(-2(z_j - z_i))),$$

$$r_{iM} = 1/2(1 - \exp(-2(x - z_i))),$$

$$r_{jM} = 1/2(1 - \exp(-2(x - z_j))),$$

so that

$$\begin{aligned} c_{ij} = {} & 1/4(1 + \exp(-2(x - z_j)))(\exp(-2(z_j - z_i)) \\ & - \exp(-2(x - z_i))). \end{aligned} \tag{13}$$

Integrating, we obtain

$$\begin{aligned} \gamma(x) &= \int_0^x \left(\int_{z_i}^x c_{ij}\, dz_j \right) dz_i \\ &= (1/8)x - (3/32) + (1/8)\exp(-2x) \\ &\quad - (1/32)\exp(-4x). \end{aligned} \tag{14}$$

Computation of γ'. In that case, the ordering of the loci on the chromosome is $z_i \leqslant m \leqslant z_j$ and we have

$$\begin{aligned} c_{ij} &= P(Z_i^{(1)} = 1, Z_j^{(1)} = 1 | M) - P(Z_i^{(1)} = 1 | M) P(Z_j^{(1)} = 1 | M) \\ &= (1 - r_{iM})(1 - r_{jM}) - (1 - r_{iM})(1 - r_{jM}) \\ &= 0. \end{aligned} \tag{15}$$

This shows that in BC1, the genotypes of two loci flanking a marker are independent, conditional on the genotype at the marker.

Finally, we get

$$\begin{aligned} Var(Z^{(1)}) = {} & \frac{1}{16L^2}(4L - 6 + 4e^{-2m} + 4e^{-2(L-m)} \\ & - e^{-4m} - e^{-4(L-m)}). \end{aligned} \tag{16}$$

This is minimized when $m = L/2$, which is also the value of m maximizing $E(\overline{Z}^{(1)})$.

(b) *Computation of* $Var(\overline{Z}^{(t)})$

When we consider more than one marker per chromosome and more advanced backcross generations, the calculations of $Var(\overline{Z}^{(t)})$ become tedious and we need to use numerical computations. We used the MDM program (Servin *et al.*, 2002) to approximate SD_{Π}.

(ii) *Distribution of* Π

Together with the mean and variance of Π, we are interested in the distribution of Π values in individuals presenting the ideotype at markers. The distribution function of Π, f_{Π}, is not known.

https://doi.org/10.1017/S0016672305007421 Published online by Cambridge University Press

Table 1. *Expected proportion of recipient genome* ($\widehat{\Pi}$) *on a chromosome of 100 cM and its corresponding standard deviation* (SD_{Π}) *on a genotype recipient at all m markers for different backcross generations* (*BC*). *The theoretical proportion of recipient genome on the chromosome when no selection is performed* (π), *its corresponding standard deviation* (SD_{π}*; from Hill, 1993*) *and optimal positioning of markers* (d^**; from Servin & Hospital, 2002*) *are also tabulated*

m	BC	π (%)	SD_{π} (%)	d^* (cM)	$\widehat{\Pi}$ (%)	SD_{Π} (%)
2	1	75	18·8	18·6	96·7	5·7
	2	87·5	15·4	21·4	97·5	4·8
	3	93·75	11·0	22·9	98·4	3·8
	4	96·9	7·6	24·0	99·0	2·9
3	1	75	18·8	8·4	98·6	3·4
	2	87·5	15·4	11·0	98·8	3·0
	3	93·75	11·0	12·6	99·2	2·5
	4	96·9	7·6	14·0	99·4	2·0
4	1	75	18·8	4·5	99·2	2·2
	2	87·5	15·4	6·5	99·3	2·1
	3	93·75	11·0	7·8	99·5	1·8
	4	96·9	7·6	9·0	99·6	1·4

Table 2. *Expected proportion of recipient genome of an ideotype at m markers, when the ideotype is obtained at generation* t_I *and the backcross programme pursued until generation t. The probabilities of obtaining an ideotype at one non-carrier chromosome* ($P_I(1)$) *and 10 non-carrier chromosomes* ($P_I(10)$) *are indicated. The chromosome length is 100 cM, and the markers are positioned as indicated in Table 1*

		t					
m	t_I	BC1	BC2	BC3	BC4	$P_I(1)$	$P_I(10)$
2	BC1	96·7	98·4	99·2	99·6	0·32	$1{\cdot}1\times10^{-5}$
	BC2	–	97·5	98·8	99·4	0·61	$7{\cdot}0\times10^{-3}$
	BC3	–	–	98·4	99·2	0·79	$9{\cdot}2\times10^{-2}$
	BC4	–	–	–	99·0	0·89	$3{\cdot}1\times10^{-1}$
3	BC1	98·6	99·3	99·7	99·8	0·26	$1{\cdot}3\times10^{-6}$
	BC2	–	98·8	99·4	99·7	0·54	$2{\cdot}0\times10^{-3}$
	BC3	–	–	99·2	99·6	0·73	$4{\cdot}5\times10^{-2}$
	BC4	–	–	–	99·4	0·85	$2{\cdot}1\times10^{-1}$
4	BC1	99·2	99·6	99·8	99·9	0·23	$4{\cdot}2\times10^{-7}$
	BC2	–	99·3	99·7	99·8	0·50	$9{\cdot}0\times10^{-4}$
	BC3	–	–	99·5	99·8	0·70	$2{\cdot}8\times10^{-2}$
	BC4	–	–	–	99·6	0·83	$1{\cdot}6\times10^{-1}$

However, we can perform a Monte Carlo integration of f_{Π} by drawing values from the distribution of Π with computer simulations as follows. We have simulated chromosomes of 100 centimorgans (cm) carrying markers and many evenly spread loci on the chromosomes, which allows assessment of their true proportion of recipient genome (Π). After a given number of generations, we stopped the simulations and kept only chromosomes carrying fully recipient genotype at markers. This allows estimation of f_{Π} and the cumulative distribution function of Π (Φ_{Π}), that is the probability:

$$\Phi_{\Pi}(\alpha)=P(\Pi>\alpha) \quad \text{for } 0\leqslant\alpha\leqslant1. \tag{17}$$

The computer programs used to perform the numerical computations and Monte Carlo simulations mentioned in this section can be obtained from the author on request.

3. Results and discussion

Using the MDM program, we have computed the variance in the PRG of individuals presenting a fully recipient genotype at markers for all marker positioning, with a 0·1 cM step, on a 100 cM chromosome. We have found that the positions minimizing the variance in the PRG are different from the positions maximizing the mean (data not shown). As our main objective is to maximize the recipient genome recovery, for the following derivations we have chosen to consider the marker positionings maximizing the mean recipient genome recovery, i.e. those described by Servin & Hospital (2002).

Table 1 shows $\widehat{\Pi}$ and SD_{Π} of individuals presenting a fully recipient genotype at markers on a non-carrier chromosome of 100 cM. Figures are shown for two, three and four markers per non-carrier chromosome (m) and for one, two, three and four backcross generations (BC). The tabulated values are successively:

- The mean proportion of recipient genome in non-selected populations (π) and the corresponding standard deviation (SD_{π}), computed using the formula from Hill (1993).
- The optimal positioning of markers (d^*) computed as in Servin & Hospital (2002), the position of the ith marker on the chromosome being given by $d^*+(i-1)(L-2d^*)/(m-1)$, where m is the total number of markers on the chromosome.
- The mean proportion (Π) of recipient genome on the chromosome given it carries a fully recipient genotype at markers, and the corresponding standard deviation (SD_{Π}).

The results presented in Table 1 show that, for a given number of markers, the expected PRG is higher for more advanced backcross generations. This suggests that the backcross programme might be pursued even after the ideotype is obtained to increase the PRG. The question is then to know to what extent the PRG would be increased by performing more backcross generations. Table 2 shows the expected PRG of

https://doi.org/10.1017/S0016672305007421 Published online by Cambridge University Press

an ideotype when the backcross programme is pursued until generation BC4: the results are shown for two, three and four markers per non-carrier chromosome for programmes where the ideotype is obtained at generation t_I, ranging from BC1 to BC4. These results show that the expected PRG is higher when t_I is small: the sooner the ideotype is obtained the higher the expected PRG is in BC4. However, this must be contrasted with the probability of obtaining the ideotype for the different programmes considered. The last two columns of Table 2 show the probabilities of obtaining the ideotype at one non-carrier chromosome (P_I (1)) and 10 non-carrier chromosomes (P_I (10)) respectively, at generation BC4. Note that, when an ideotype is obtained on a non-carrier chromosome, it is obtained with probability 1 in subsequent generations, so the probabilities P_I (1) and P_I (10) are just the probabilities of obtaining the ideotype at generation t_I. When considering only one non-carrier chromosome, the probabilities P_I (1) are of the same order of magnitude for all the programmes considered. However, for an ideotype at 10 non-carrier chromosomes, waiting one generation more to get it increases P_I (10) by at least a factor 10. The probabilities P_I (1) and P_I (10) given in Table 2 do not include the probability of keeping the target genes throughout the backcross programme. Taking into account selection for the target genes and selection to reduce the linkage drag around the target genes would further reduce P_I (1) and P_I (10). On the other hand, selection on non-carrier chromosomes during the backcross programme would increase these probabilities. However, it must be noted that selection is not possible before BC1 and, therefore, in BC1 P_I (1) and P_I (10) cannot be greater than the values indicated in Table 2. The results of Table 2 suggest that selection on markers on non-carrier chromosomes should be performed starting from the first generation of the backcross programme, as it will increase the expected PRG reached at the end of the programme.

Table 1 shows that the standard deviation in the true genomic composition of genotypes fully recipient at markers decreases with increasing the number of generations. This is expected as, when performing more backcross generations, individuals tend all to be equivalent with a proportion of recipient genome of about 100%. This standard deviation also decreases when increasing the number of markers, showing that using more markers leads to a better estimation of the true PRG of individuals, which is also expected. If we use few markers per non-carrier chromosome, selection on markers will be very efficient. However, the values of SD_{Π} in Table 1 show that the actual PRG (Π) of an individual obtained at the end of the selection process might be much lower than the average PRG, $\widehat{\Pi}$. In order to cope with this problem, it is possible to use more markers per non-carrier chromosome, but this will concomitantly increase the cost of the backcross programme. In order to choose precisely the number of markers that must be genotyped on non-carrier chromosomes, we have used an approach based on the estimation of the distribution of Π in individuals carrying the ideotype at markers.

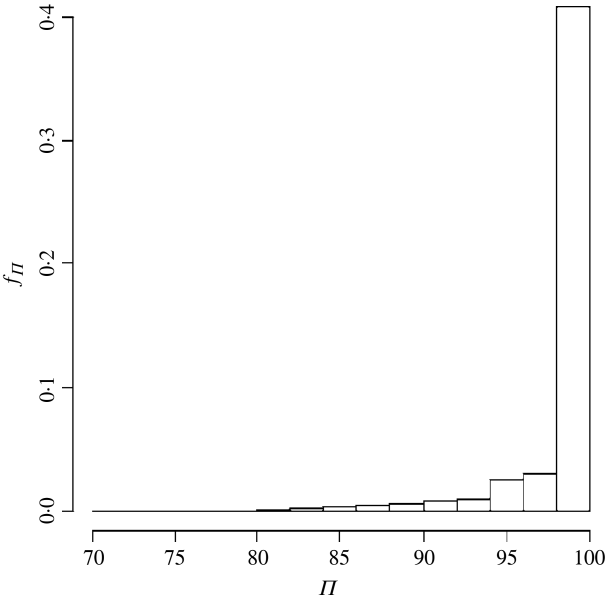

Fig. 1. The distribution (f_{Π}) of the proportion of recipient genome (Π) in a population of individuals presenting homozygous recipient genotypes at three markers at generation BC2 on a non-carrier chromosome of 100 cM. Positions of markers are as described in Servin & Hospital (2002). Results are based on 50 000 simulated chromosomes.

At the end of a MAB programme a population of individuals sharing the same ideal genotype at markers is obtained. To complete the programme, it is possible to select the best individual from this population. As individuals share the same genotype at markers, the selection can be done by either (i) genotyping more markers on non-carrier chromosomes to have a more precise estimate of Π for each individual or (ii) evaluating the agronomic performance of each individual. Note that, in the first case, as the recipient genome recovery of all individuals in the population is already quite high, the number of additional markers to genotype before getting a sufficient discrimination power inside the population can be very large.

The cost of the selection process at this last step depends on the number of individuals that must be evaluated. In order to limit that cost, we can estimate the minimal number of individuals (N_{I^*}) required that carry a fully recipient genotype at markers, so that at least one of them has a sufficient PRG in its genetic background.

N_{I^*} is computed from the probability Φ_{Π} that an individual carrying the ideotype at markers has Π above given value. Fig. 1 shows the distribution of

https://doi.org/10.1017/S0016672305007421 Published online by Cambridge University Press

Table 3. *Number of individuals to produce in order to have a probability of 0·99 of obtaining an individual presenting a true proportion of recipient genome* (Π) *above a given threshold* (*95%, 97% and 99%*), *as a function of the number of non-carrier chromosomes* (N_C), *number of markers per non-carrier chromosome* (*m*) *and backcross generation at which the genotype at markers is obtained* (*BC*)

		$N_C=1$			$N_C=5$			$N_C=10$		
m	BC	$\Pi \geqslant 95\%$	$\Pi \geqslant 97\%$	$\Pi \geqslant 99\%$	$\Pi \geqslant 95\%$	$\Pi \geqslant 97\%$	$\Pi \geqslant 99\%$	$\Pi \geqslant 95\%$	$\Pi \geqslant 97\%$	$\Pi \geqslant 99\%$
2	2	3	4	5	12	19	30	41	92	216
	3	3	3	4	7	10	14	16	28	53
	4	2	3	3	5	6	8	9	13	21
3	2	2	3	4	5	8	14	10	21	53
	3	2	3	3	4	6	9	7	13	24
	4	2	2	3	4	5	6	6	9	13
4	2	2	2	3	4	5	8	6	10	21
	3	2	2	3	3	4	7	5	7	15
	4	2	2	2	3	4	4	4	6	7

Π drawn with simulations performed as presented in Section 2. We can see that the distribution of Π is L-shaped, particularly because we only retain the individuals that carry a fully recipient genotype at markers: typically, as seen in Fig. 1, only a few individuals have a low Π while most of them are nearly fully recipient even outside the markers. From f_Π we can estimate the cumulative distribution function of Π (Φ_Π). We have estimated Φ_Π with two to four markers per non-carrier chromosome, obtained at backcross generations BC2 to BC4 and for genomes composed of one, five or 10 non-carrier chromosomes. In each case, the estimate of Φ_Π is based on the PRG of 50 000 simulated chromosomes with the ideotype at markers. Table 3 gives the corresponding N_{I^*}, computed with a probability of success of 0·99, for selection objectives of 95%, 97% or 99% of recipient genome. The mean and variance of Π in the simulations were very close to the values given in Table 1 (data not shown).

Table 3 shows that N_{I^*} is large at generation BC2, except when using four markers per chromosome. For example, the N_{I^*} needed to achieve a selection objective of 97% at generation BC2, when considering a genome composed of 10 non-carrier chromosomes, each with two markers, is 92. It is interesting to note that the corresponding average rate of the recipient parent genome recovery is 97·5% (see Table 1). This shows that $\widehat{\Pi}$ is not always a sufficient criterion to choose the number of markers to use for background selection, because SD_Π is large. N_{I^*} is much smaller at generation BC2 when using four markers per non-carrier chromosome. However, obtaining N_{I^*} individuals with fully recipient genotypes at four markers per non-carrier chromosome as early as generation BC2 is difficult and might necessitate huge population sizes.

For more advanced backcross generations BC3 and BC4, the results presented in Table 3 show that if the number of non-carrier chromosomes is one or five, N_{I^*} is close to 10, whatever the selection objective is. Hence, when the number of non-carrier chromosomes is low, the whole genetic background can be controlled successfully with only two markers per 100 centiMorgans. However, when considering 10 non-carrier chromosomes, using more than two markers is mandatory for obtaining a recipient genome recovery above 97% while limiting the cost of the last selection step. Using only two markers per non-carrier chromosome is thus sensible only if the selection objective on Π is low *and* the number of non-carrier chromosomes is small.

The choice between using three or four markers per non-carrier chromosome for a large genome (10 non-carrier chromosomes) is not obvious. N_{I^*} is larger when using only three markers per non-carrier chromosome, but these individuals are easier to obtain by selection on markers than when using four markers. Hence, using three markers will limit the genotyping cost at early generations of backcross but will imply a high cost to select the best individual at the end of the backcross programme. On the other hand, using four markers per non-carrier chromosome will increase genotyping cost in all generations but necessitate less expense to identify an individual that meets a given selection objective at the end of the MAB programme.

The ability to obtain the N_{I^*} ideotypes inferred using our method will depend on the means available for a particular backcross programme, such as the population sizes that can be handled and the genotyping technology available. Furthermore, factors such as the number of target genes introgressed and the reduction of the linkage drag around the target genes will affect the probability of obtaining those

https://doi.org/10.1017/S0016672305007421 Published online by Cambridge University Press

ideotypes. This must be optimized by choosing appropriate selection strategies which require more developments that are beyond the scope of this study.

The author wishes to thank Steve Openshaw, Frédéric Hospital, Peter Visscher and an anonymous referee for their useful comments and help with the manuscript.

References

Frisch, M. & Melchinger, A. E. (2001). The length of the intact donor chromosome segment around a target gene in marker-assisted backcrossing. *Genetics* **157**, 1343–1356.

Hill, W. G. (1993). Variation in genetic composition in backcrossing programs. *Journal of Heredity* **84**, 212–213.

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

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

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

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

Servin, B. & Hospital, F. (2002). Optimal positioning of markers to control genetic background in marker-assisted backcrossing. *Journal of Heredity* **93**, 214–216.

Servin, B., Dillmann, C., Decoux, G. & Hospital, F. (2002). MDM: a program to compute fully informative genotype frequencies in complex breeding schemes. *Journal of Heredity* **93**, 227–228.

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

Visscher, P. M. & Thompson, R. (1995). Haplotype frequencies of linked loci in backcross population derived from inbred lines. *Heredity* **75**, 644–649.

Visscher, P. M., Haley, C. S. & Thompson, R. (1996). Marker assisted introgression in backcross breeding programs. *Genetics* **144**, 1923–1932.

https://doi.org/10.1017/S0016672305007421 Published online by Cambridge University Press