

Application of Population Genetic Theory and Simulation Models to Efficiently Pyramid Multiple Genes via Marker-Assisted Selection

Jiankang Wang,* Scott C. Chapman, David G. Bonnett, Greg J. Rebetzke, and Jonathan Crouch

ABSTRACT

Breeders face many complex choices in the design of efficient crossing and selection strategies aimed at combining desired alleles into a single target genotype. Both population genetic theory and a breeding simulation tool were used to study the effects of different strategies on population size and number of marker assays required to recover a target genotype in wheat (*Triticum aestivum* L.). Enriching the frequency of desirable alleles in the F_2 of single-cross and in the F_1 of backcross and topcross populations greatly reduced the minimum required population size, but the gain from another enrichment selection is minor. General equations were developed to determine appropriate crossing strategies, and sequential culling was proposed to minimize total marker screening costs. For a topcross of three adapted lines from an existing breeding program, simulation of changes in allele frequencies at nine target genes (seven unlinked) showed that population size was minimized with a three-stage selection strategy in the F_1 generation of the topcross (TCF_1), the F_2 generation of the topcross (TCF_2), and doubled haploid lines (DHs). Enrichment of allelic frequencies in TCF_2 reduced the total number of lines screened from >3500 to <600. Eight of the genes were present at frequencies >0.97 after selection, while the *tin* reduced-tillering allele was only at 0.77 in the final selected population due to its strong repulsion-phase linkage to the grain quality gene *Glu-A3* in this cross and the incomplete linkage of the *tin* marker. Therefore, the presence of the *tin* gene needs to be further confirmed by other methods.

J. Wang and J. Crouch, Crop Research Informatics Lab., and Genetic Resources Enhancement Unit, International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico; S.C. Chapman, CSIRO Plant Industry, 306 Carmody Rd, St. Lucia, QLD 4067, Australia; D.G. Bonnett and G.J. Rebetzke, CSIRO Plant Industry, P.O. Box 1600, Canberra, ACT 2601, Australia; and J. Wang, Institute of Crop Science, and The National Key Facility for Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences, Beijing 100081, China. Received 28 May 2006. *Correspondence author (wangjk@caas.net.cn or j.k.wang@cgiar.org).

Abbreviations: DHs, doubled haploid lines; MAS, marker-assisted selection; QTL, quantitative trait locus; RIL, recombination inbred lines; TCF_1 , F_1 generation of the topcross; TCF_2 , F_2 generation of the topcross.

MANY BREEDING programs in a range of crops are using molecular markers to screen for one to several alleles of interest. The availability of an increasing number of useful molecular markers is allowing accurate selection at a greater number of loci than has been previously possible (Paterson et al., 1991; Dekkers and Hospital, 2002; Dubcovsky, 2004). However, larger population sizes are required to ensure with reasonable certainty that an individual with the target genotype is present. Different crossing and selection strategies may require vastly different population sizes to recover a target genotype with the same certainty even when the same parents are used (Bonnett et al., 2005). Determination of the most efficient strategy has the potential to dramatically decrease the amount of resources (plants, plots, marker assays, and labor) required to combine a set of target alleles into a new genotype.

Marker-assisted selection (MAS) may utilize markers that are closely or completely linked with target genes of interest or markers that are associated with quantitative trait loci (QTLs) and explain only part of the variance for a trait that may be under complex

Published in *Crop Sci* 47:xx-xx (2007).

doi: 10.2135/cropsci06.05.0341

© Crop Science Society of America

677 S. Segoe Rd., Madison, WI 53711 USA

genetic control. If markers are not completely linked with the target genes, two flanking markers (on either side of the gene or QTL) may still be useful. Although molecular markers may allow more accurate selection in early generations than conventional phenotypic selection, the large number of individuals needed to recover a target homozygote at multiple loci at this stage can make this approach impracticable and/or too expensive. Conversely, screening in later generations often provides little or no advantage over conventional selection techniques (Bonnett et al., 2005). Considerable efficiency gains can be achieved if plant breeders are able to choose the most appropriate crossing (e.g., single cross, backcross, or topcross) and best MAS methods (Lande and Thompson, 1990; Delphin Koudande et al., 2000; Bonnett et al., 2005; Kuchel et al., 2005). Calculation of the distribution of desirable alleles among an initial set of genotypes can considerably assist the breeder.

Under simplified conditions (i.e., gene-based markers where the association between gene and marker is complete), some general recommendations were given by Bonnett et al. (2005). Using population genetic theory and the QU-GENE (developed at the University of Queensland, Australia) application module QuLine (previously called QuCim) (Podlich and Cooper, 1998; Wang et al., 2003; Wang et al., 2005), we have extended this theory to identify principles for design of efficient selection strategies where there is recombination between marker and gene, and where there is repulsion-phase linkage between desirable alleles. Note that we focus on crosses between “generally adapted” parents and therefore do not consider the process of “background” selection (Frisch and Melchinger, 2005), whereby markers are used to both select for target genes and to maximize recovery of the recurrent parent genome.

In this study, population genetic theory was used to establish general rules for the numbers of markers required, the best crossing strategies, and the level of inbreeding to maximize the efficiency of marker implementation where there was no recombination between marker and gene of interest. When the scenario was extended to linked markers, we adopted simulation analysis to develop rules for selection. A topcross among three Australian wheat lines was used to demonstrate the outcomes from the population genetic theory and simulation models, while considering both completely and incompletely linked markers, as well as linkage between target alleles.

MATERIALS AND METHODS

In using markers, several scenarios are commonly faced by breeders: (i) pyramiding alleles at multiple loci including consideration of most appropriate cross type; (ii) minimizing marker screening costs by sequential culling; (iii) use of incompletely linked markers to combine target alleles; and (iv) combining alleles linked in repulsion in crosses segregating for other unlinked target alleles. Population genetic theory was

used to investigate Scenarios i and ii, while the QU-GENE breeding simulation platform was used for Scenarios iii and iv, where population genetic theory becomes intractable.

Calculating Minimum Population Size

Where α is the probability of not having at least one target genotype present in the population sampled, and f is frequency of the genotype to be selected, the minimum population size (N) to ensure at least one target genotype is present in the population with the given level of certainty can be calculated as

$$N = \frac{\log \alpha}{\log(1-f)} \quad [1]$$

In all examples, a probability of $\alpha = 0.01$ was used. For strategies with multiple selection stages, population sizes were calculated to achieve a cumulative probability of α at 0.01 across all selection stages.

Comparing Biparental, Back-, and Topcrosses

If n loci differ between two parents with n_1 favorable alleles in the first parent P_1 , and n_2 in the second parent P_2 , then relative proportions of the target genotype in DHs or recombination inbred lines (RILs) derived from F_1 , P1BC1 (backcrossed to P_1), and P2BC1 (backcrossed to P_2) are

$$\begin{aligned} f_{F_1} &= \left(\frac{1}{2}\right)^n, & f_{P1BC1} &= \left(\frac{3}{4}\right)^{n_1} \left(\frac{1}{4}\right)^{n_2}, \\ f_{P2BC1} &= \left(\frac{1}{4}\right)^{n_1} \left(\frac{3}{4}\right)^{n_2} \end{aligned} \quad [2]$$

The three proportions were used as a guide as to whether a backcross reduced population size and to indicate which parent should be used as the recurrent parent.

If target alleles are dispersed among three parents, i.e., P_1 , P_2 , and P_3 , a topcross (or three-way cross), e.g., $(P_1 \times P_2) \times P_3$, is required to combine all alleles. If each parent carries different alleles, the alleles contributed by parents P_1 and P_2 in the first cross will be present at frequencies of 0.25 following a topcross with P_3 , and the alleles contributed by P_3 will each have a frequency of 0.5. If n_1 , n_2 , and n_3 are the numbers of target alleles in the three parents, respectively, under the condition of no selection, the expected proportion of individuals with the target genotype in DHs/RILs is

$$f_{TC} = \left(\frac{1}{4}\right)^{n_1+n_2} \left(\frac{1}{2}\right)^{n_3} = 2^{n_3-2n} \quad [3]$$

where $n = n_1 + n_2 + n_3$. Equation [3] was used to determine the order in which to cross parents to minimize the population sizes required in a topcross.

Minimizing the Total Number of Marker Assays with Sequential Culling

In a population of N individuals to be screened sequentially with markers at n independent loci, and where only those with the target genotype are retained for screening with the next marker, the total number of assays (M) required to identify the target genotype at all loci can be calculated according to the formula

$$M = N + Nf_1 + Nf_1f_2 + \dots + Nf_1f_2 \dots f_{n-1} \quad [4]$$

where f_1, f_2, \dots , and f_n are the proportions of individuals retained after screening with each marker. For any set of markers, M will be minimized if the marker with the lowest retained fraction f

Table 1. Nine genes, their locations on chromosomes, and the genotypes for the three selected parents.

Gene (locus) [†]	<i>Rht-B1</i>	<i>Rht-D1</i>	<i>Rht8</i>	<i>Sr2</i>	<i>Cre1</i>	<i>VPM</i>	<i>Glu-B1</i>	<i>Glu-A3</i>	<i>tin</i>
Chromosome	4BS	4DS	2DL	3BS	2BL	7DL	1BL	1AS	1AS
Marker type	Codominant	Codominant	Codominant	Codominant	Dominant	Dominant	Codominant	Codominant	Codominant
Distance between marker and gene (cM)	0.0	0.0	0.6	1.1	0.0	0.0	0.0	0.0	0.8
HM14BS	<i>Rht-B1a</i>	<i>Rht-D1a</i>	<i>Rht8</i>	<i>sr2</i>	<i>cre1</i>	<i>vpm</i>	<i>Glu-B1a</i>	<i>Glu-A3e</i>	<i>Tin</i>
Sunstate	<i>Rht-B1a</i>	<i>Rht-D1b</i>	<i>rht8</i>	<i>Sr2</i>	<i>cre1</i>	<i>VPM</i>	<i>Glu-B1i</i>	<i>Glu-A3b</i>	<i>Tin</i>
Silverstar+ <i>tin</i>	<i>Rht-B1b</i>	<i>Rht-D1a</i>	<i>rht8</i>	<i>sr2</i>	<i>Cre1</i>	<i>vpm</i>	<i>Glu-B1i</i>	<i>Glu-A3c</i>	<i>tin</i>
Target genotype [‡]	<i>Rht-B1a</i>	<i>Rht-D1a</i>	<i>Rht8</i>	<i>Sr2</i>	<i>Cre1</i>	<i>VPM</i>	<i>Glu-B1i</i>	<i>Glu-A3b</i>	<i>tin</i>

[†] Alleles *Rht-B1b*, *Rht-D1b*, and *Rht8* reduce plant height. Allele *Sr2* confers resistance to stem rust, and alleles *Cre1* and *VPM* confer resistance to cereal cyst nematode. Alleles *Glu-B1i* and *Glu-A3b* improve dough quality, and allele *tin* reduces the tiller number. The genes are all unlinked, except for *Glu-A3* and *tin*, which are 3.8 cM apart on chromosome 1A.

[‡] The target genotype is determined when all the nine genes are considered together. Alleles in the target genotype contribute to semidwarfing with long coleoptile length, multiple disease resistances, good grain quality, and less tillering. The three semidwarfing alleles can all produce the required plant height. However, *Rht-B1b* and *Rht-D1b* also reduce the coleoptile length, which is unfavorable for breeding drought-resistant wheat cultivars. *Rht8* reduces the plant height without affecting the coleoptile length and therefore is the favorable dwarfing allele. Other alleles in the target genotype are easily understood as they increase the resistance to some diseases, increase the grain quality, or reduce the number of tillers.

(or the highest cull rate) is used first, followed by the next lowest, and so on. The total cost (*C*) of marker assays can be determined from Eq. [4] by inclusion of the cost of each assay,

$$C = Nc_1 + Nf_1c_2 + Nf_1f_2c_3 + \dots + Nf_1f_2 \dots f_{n-1}c_n \quad [5]$$

where $c_1, c_2, \dots,$ and c_n are the cost of each of the marker assays. From Eq. [5], it can be shown that *C* is minimized when $\frac{c_1}{1-f_1} < \frac{c_2}{1-f_2} < \dots < \frac{c_n}{1-f_n}$.

Equations [1] to [5] can be used to address the first two scenarios when no gene linkages exist. Simulation is needed for the other scenarios. The analytic expression for the cost of sequential culling ignores the costs of plant/line handling (tagging, leaf sampling, etc.) and DNA extraction, which are fixed with total sample size and cannot be reduced by sequential culling. If these fixed costs are major parts of the expense for genotyping, the order of markers used in the sequential culling may become less important.

The Genetics and Breeding Simulation Tools

QU-GENE is a simulation platform for quantitative analysis of genetic models. The program generates populations of genotypes and provides a library of subroutines to develop simulation modules for real-world breeding programs (Podlich and Cooper, 1998). QuLine is a QU-GENE application module that was specifically developed to simulate breeding programs developing inbred lines (Wang et al., 2003) and has also been used to predict cross performance for quality traits using known gene information (Wang et al., 2005). The software is available to researchers via arrangements with the International Maize and Wheat Improvement Center (CIMMYT) (contact the corresponding author) or The University of Queensland, Australia (contact Dr. Mark Dieters: m.dieters@uq.edu.au).

Use of Simulation Modeling to Examine the Strategies to Minimize Population Sizes while Combining Target Alleles

Equations [1] to [5] do not consider genetic linkage between the marker and target gene, or different target genes. While the equations can be readily extended to accommodate recombina-

tion, they become difficult to evaluate algebraically as gene number increases. To illustrate the effect of linkage, we simulated a topcross among three wheat lines: Sunstate (a commercial Australian line), HM14BS (a source of the “long coleoptile” trait that utilizes the *Rht8* allele for reduced height), and Silverstar+*tin* (a modified Australian variety that is a source of the *tin* “reduced-tillering” trait). Genotypic and marker data at the nine polymorphic loci are shown in Table 1. Alleles at seven of the nine loci are independently inherited, while *Glu-A3* and *tin* are linked in repulsion on the short arm of chromosome 1A at a distance of 3.8 cM ($r = 0.0366$) (Spielmeyer and Richards, 2004). Haldane’s mapping function was used to transform the mapping distance into recombination frequency.

The target alleles (Table 1, last row) at the *Rht-B1*, *Rht-D1*, and *Rht8* loci all affect plant height (Rebetzke and Richards, 2000). Other genes include *Sr2* for adult plant stem rust resistance, *Cre1* for cereal cyst nematode resistance, *VPM*, an *Aegilops ventricosa* chromosome translocation carrying genes for leaf (*Lr37*), stem (*Sr38*), and stripe (*Yr17*) rust resistance (Bariana and McIntosh, 1993), the *Glu-B1* and *Glu-A3* grain storage protein loci, and the *tin* gene, affecting tiller number. Completely linked molecular markers are available for all loci except *Rht8*, *Sr2*, and *tin*, where markers are 1.1 cM or less from the gene (Korzun et al., 1998; Spielmeyer et al., 2003). Except for *Cre1* and *VPM*, the molecular markers are codominant (Korzun et al., 1998; Ogonnaya et al., 2001; Ellis et al., 2002; Ma et al., 2003; Spielmeyer et al., 2003; Zhang et al., 2004).

RESULTS AND DISCUSSION

Efficient Pyramiding of Alleles at Multiple Loci: Biparental Cross

When many (unlinked) markers are targeted in selection, the frequency of a target homozygous genotype will be low, and a large population size will be required. For example, in the F_2 of a biparental cross between two inbred parents segregating at five unlinked (independent) loci, the frequency of the target genotype is $0.25^5 = 0.00098$, and the minimum population size (Eq. [1]) to recover at least one target genotype is 4714 ($\alpha = 0.01$). If selection is

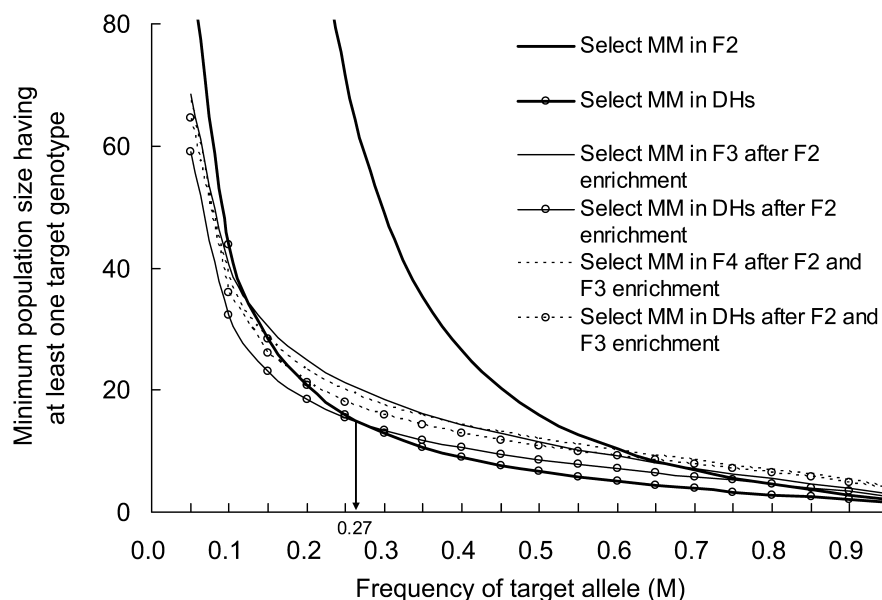


Figure 1. The minimum population size having at least one selected individual. A probability of not having at least one target genotype present in the population sampled $\alpha = 0.01$ is assumed. If more than one selection stage is involved, the summation of the minimum population sizes of all stages is used. The probability used for each stage is $1 - (1 - \alpha)^{\frac{1}{2}}$ when there are two selection stages, and $1 - (1 - \alpha)^{\frac{1}{3}}$ when there are three selection stages.

made among homozygous lines [i.e., DHs or RILs] from the same cross, the frequency of the target genotype is $0.5^5 = 0.03125$ with a minimum population size of only 146 ($\alpha = 0.01$), i.e., the target genotype is more readily recovered with smaller population size if selection is delayed until greater homozygosity has been achieved.

For more segregating loci, population sizes quickly increase even in DH or RIL populations. For example, in a biparental population with eight unlinked segregating loci, the frequency of the target genotype in a homozygous population is $0.5^8 = 0.0039$, and the minimum population size 1177. In these instances, Bonnett et al. (2005) proposed a two-stage selection strategy. The first stage is “ F_2 enrichment,” where F_2 individuals carrying the entire set of target alleles in either homozygous or heterozygous form are selected. F_2 enrichment takes advantage of the high expected frequency of carriers (either homozygous or heterozygous) at each locus of 0.75. The value of the technique can be seen in a population segregating at 12 loci, where the frequency of genotypes selected in an F_2 enrichment step is $0.75^{12} = 0.03168$, resulting in the minimum population size of 144 F_2 generations (cf. frequency of 0.25^{12} and a population size >77 million to identify a single homozygous individual in the F_2). After F_2 enrichment, the frequency of each of the 12 target alleles in the selected population is increased from 0.5 to 0.67. The second step is to generate a population of more or less homozygous lines from the selected F_2 . The frequency of the target genotype in DHs or RILs generated from the enriched F_2 will have been increased from $0.5^{12} = 0.00002$ to $0.67^{12} = 0.00771$, resulting in a decrease in

minimum population size from 18 861 to 596. Thus, with enrichment, both the F_2 and DH/RIL populations are of a more practical size for breeding.

The point at which population sizes become unmanageable will vary from one breeding program to another, and for high-value trait combinations, breeders may be prepared to apply molecular screens to larger numbers (say tens of thousands of lines). However, to simplify further discussion, in our studies we set a relatively modest maximum population size of 1000 at $\alpha = 0.01$ at any given selection stage. With this limitation, direct selection of the target genotype in F_2 will allow no more than three alleles to be combined. If the target genotype is selected in DHs or RILs, only seven alleles can be combined. Use of F_2 enrichment allows target alleles at 12 or 13 loci to be combined in derived homozygous lines. Linkage of genes in coupling will have a positive effect on the frequency of the target genotype, while linkage in repulsion will have a negative effect related to the level of recombination. Wherever linkage occurs, simulation approaches (see later) can assist in determining optimum selection strategies.

While our initial findings support those of Bonnett et al. (2005) on the benefit of F_2 enrichment using conventional formulae, we were able to extend this work to investigate possible benefits of enrichment in later segregating generations or combining F_2 enrichment with enrichment in F_3 and/or F_4 populations. Given that minimum population size is determined by the frequency of the target genotype and that the same genotype frequency can result from different numbers and frequencies of target alleles, it is possible to study the relative efficiencies of different selection methodologies with a single-locus model. If a target allele, M, has the frequency p in an F_2 population, then the frequencies of the three marker types MM, Mm, and mm, are p^2 , $2p(1-p)$, and $(1-p)^2$, respectively, under Hardy–Weinberg expectations (Falconer and Mackay, 1996). Varying the frequency of M will result in different genotype frequencies for which alternative selection schemes were compared: (i) target genotype, i.e., MM, selected in F_2 ; (ii) target genotype selected in DHs or RILs (say $>F_3$); (iii) target genotype selected in F_3 after F_2 enrichment; (iv) the target genotype selected in DHs or RILs after F_2 enrichment; (v) target genotype selected in F_4 after F_2 and F_3 enrichment; and (vi) target genotype selected in DHs or RILs after F_2 and F_3 enrichment.

The formula for calculating gene and genotype frequencies after selection for each of these methodologies can be readily derived via Eq. [1], based on which we calculated the minimum population size with $\alpha = 0.01$ for each scheme (Fig. 1). For Schemes 3 to 6, with more than one selection stage, the minimum population sizes were summed across stages. The probability used for each stage was $1 - (1 - \alpha)^{\frac{1}{2}}$ when there were two selection stages (Schemes 3 and 4), and $1 - (1 - \alpha)^{\frac{1}{3}}$ when there were three selection stages (Schemes 5 and 6), to have the same cumulative probability of α for each scheme.

Direct selection of the target genotype in the F_2 generation requires a substantially greater minimum population size, unless the frequency of the target genotype in the F_2 exceeds about 0.60. When the frequency of the target genotype exceeds 0.27, unenriched DHs/RILs (Scheme 2) require the smallest population size. Otherwise, selecting the target genotype in DHs/RILs after F_2 enrichment (scheme 4) results in the lowest numbers (Fig. 1). In a biparental cross, the point at which frequency of the target genotype falls below 0.27 in an unenriched DH/RIL population and F_2 enrichment (Scheme 4) offers potentially useful reductions in numbers occurs with only three segregating loci. Therefore, in most cases, F_2 enrichment followed by selection of homozygotes in DHs/RILs results in the greatest reduction in minimum population sizes.

Enrichment at two selection stages (in F_2 and F_3) always required greater assay numbers than simple F_2 enrichment (Fig. 1). As indicated by Bonnett et al. (2005), F_2 enrichment increased the frequency of selected alleles, allowing large reductions in minimum population size for recovery of target genotypes (commonly around 90%) and/or selection at a greater number of loci. So the gain from another enrichment selection in F_3 after the enrichment in F_2 is at best minor and often results in a small net increase in minimum population size.

Comparison of Biparental, Backcross, and Topcross Populations

Backcrossing is an effective method to reduce population size compared with a biparental cross where one parent contributes more target alleles than the other (Bonnett et al., 2005). However, when each parent has a similar number of target alleles, the magnitude of the reduction may not be sufficient to compensate for the added cost, complexity, and time involved in generating a backcross population. If $f_{P1BC1} / f_{F1} = 3^{n_1} / 2^n > 1$ (Eq. [2]), a backcross will reduce population sizes using P_1 as the recurrent parent; if $f_{P2BC1} / f_{F1} = 3^{n_2} / 2^n > 1$, P_2 should be the recurrent parent; otherwise, no backcross is needed. For example, if $n = 5$, $n_1 = 3$, and $n_2 = 2$, then $f_{P1BC1} / f_{F1} = 0.84$, and $f_{P2BC1} / f_{F1} = 0.28$, and backcrossing is not helpful. If $n_1 = 4$ and $n_2 = 1$, $f_{P1BC1} / f_{F1} = 2.53$, and therefore, a backcross should be used with P_1 as the recurrent parent.

If the target alleles are dispersed among three parents, i.e., P_1 , P_2 , and P_3 , a topcross (or three-way cross) is often used, e.g., $(P_1 \times P_2) \times P_3$. Equation [3] shows that f_{TC} is maximized when n_3 is the largest number, i.e., when a topcross is required, the parent with the largest number of favorable alleles should be used as the third parent.

Effects of Incompletely Linked Markers on Allele Frequencies following Selection

It takes substantial effort to develop markers that are completely linked to target alleles. The usefulness of incompletely linked markers depends on the level of recombination between the marker and the target allele and the minimum frequency of target genotypes considered acceptable following selection. If the minimum acceptable frequency of target genotypes is taken to be 0.95, a single marker will be suitable if its distance to the gene is less than 5 cM and homozygotes are to be selected in the F_2 generation (Table 2). Single markers with a genetic distance of 10 cM will result in a frequency of the target allele of 0.91 (Table 2). However, selection in the F_2 for flanking markers at 10 cM results in an allele frequency of 0.99, equivalent to that of a single marker 1 cM from the target gene. Such flanking markers will be better than a single marker at 5 cM in all cases, including where homozygotes are selected in F_{10} (0.959) or where allele enrichment is applied in F_2 , followed by selection of homozygotes in F_{10} (0.963).

Prediction of Selection Outcomes for more Complex Genetic Models

A topcross between lines HM14BS, Sunstate, and Silverstar+*tin* (Table 1) was simulated to determine the minimum population sizes required to recover a target genotype, given selection among DHs with and without prior enrichment in the F_2 generation. The target genotype given in Table 1 was determined by semidwarfing with long coleoptile length, multiple resistances, good grain quality, and reduced tillering. Any of the three semidwarfing alleles, i.e., *Rht-B1b*, *Rht-D1b*, and *Rht8*, will be able to produce the required plant height, and multiple dwarfing alleles make the plant too short to be useful.

Table 2. Gene frequency after marker-assisted selection using incompletely linked markers.

Selection method	Marker type	Distance between marker and gene		
		1 cM	5 cM	10 cM
Homozygous selection in F_2	Single marker	0.991	0.954	0.910
	Flanking markers	1.000	0.998	0.990
Homozygous selection in F_{10}	Single marker	0.980	0.912	0.846
	Flanking markers	0.999	0.988	0.959
Enrichment selection in F_2 , and homozygous selection in F_{10}	Single marker	0.982	0.914	0.847
	Flanking markers	0.999	0.987	0.963

However, *Rht-B1b* and *Rht-D1b* also reduce the coleoptile length as well as plant height, contributing to reduced drought resistance. *Rht8* reduces the plant height without affecting the coleoptile length (Rebetzke and Richards, 2000; Botwright et al. 2001). Therefore *Rht8* is the favorable dwarfing allele and should be present in our target genotype. Other alleles in the target genotype are easily understood as they increase the resistance to some diseases, increase the grain quality, or reduce the number of tillers. Target alleles are distributed unequally between the three parents, with HM14BS carrying three target alleles, Sunstate carrying five target alleles, and Silverstar+*tin* carrying four target alleles. The frequency of the target genotype will be maximized if Sunstate is used as the third parent in topcrossing (Eq. [3]), so the other two topcrosses were not considered.

Selection in the F₁ Generation of the Topcross

In the F₁ generation of the topcross (TCF₁), *Rht-B1*, *Rht8*, *Cre1*, *Glu-B1*, and *tin* are segregating. The target genotypes of *Rht-B1aRht-B1a* and *Glu-B1iGlu-B1i* have a frequency of 0.5 in TCF₁, and all other target alleles exist in heterozygous form at frequencies of 0.5. Therefore selection of *Rht-B1a* and *Glu-B1i* homozygotes and allele enrichment for *Rht8*, *Cre1*, and *tin* can be applied in TCF₁, and the theoretical selected proportion in TCF₁ is $0.5^5 = 0.0313$. Considering this high proportion and for simplicity, no other selection option was applied in TCF₁.

Selection in the F₂ and F₂-Derived DH Generation of the Topcross

The target genotype lacks *Rht-B1b* and *Rht-D1b* and is homozygous for *Rht8*, *Sr2*, *Cre1*, *VPM*, *Glu-B1i*, *Glu-A3b*, and *tin* (Table 1, last row). We considered three options for selection in TCF₂: (i) no selection in TCF₂, (ii) F₂ enrichment for all genes except *Rht-B1* and *Glu-B1* (as *Rht-B1a* and *Glu-B1i* have been fixed after selection of the homozygotes in TCF₁ at the two loci), and (iii) selection of *Rht8* homozygotes and F₂ enrichment of all remaining alleles. Selection of homozygotes at two loci in TCF₂ was also

simulated, but a much larger minimum population size in TCF₂ was required (results not shown).

For the three options considered, selection of target homozygotes was conducted in DHs, i.e., the first option (no selection in TCF₂) consists of two selection stages, one in TCF₁, the other in DHs. The simulation shows the proportion selected in TCF₁ is close to the theoretical upper limit of 0.0313 (Table 3). The selected proportion in DHs is about 0.0009, requiring quite a large DH population to select the target genotype. The second and the third options both consist of three selection stages, one in TCF₁, one in TCF₂, and one in DHs. For the second option, the selected proportion is 0.1190 in TCF₂ and 0.0071 in DHs. The third option has a more evenly distributed selected proportion over stages and requires the smallest number of lines overall (Table 3). In practice, if multistage selection is applied, the general rule to minimize population size would be to minimize differences in selection intensity at the different stages, which will minimize cost if markers are equal in cost. Multiplexing appropriate sets of markers provides further cost savings.

Final Target Allele Frequencies following MAS

Due to the complete linkage of genes *Rht-B1*, *Rht-D1*, *Cre1*, *VPM*, *Glu-B1*, and *Glu-A3* with their markers (Table 1), the frequencies of alleles *Rht-B1a*, *Rht-D1a*, *Cre1*, *VPM*, *Glu-B1i*, and *Blu-A3b* are 1.0 after MAS in the final selected population. *Rht8* has a distance of 0.6 cM to its marker, and *Sr2* 1.1 cM to its marker. Through simulation, we found the allele frequency is near 0.99 for *Rht8* and 0.98 for *Sr2* after MAS selection, which should be acceptable in practical breeding.

Given that *tin* and its microsatellite marker are 0.8 cM apart, the estimated allele frequency of *tin* is at 0.77 in the final selected population. The reason for the lower than expected frequency is due to its linkage in repulsion with the important glutenin allele, *Glu-A3b*, in parents Sunstate and Silverstar+*tin* (Table 1). The haplotype frequency from the biparental cross between Sunstate and Silverstar+*tin* illustrates the effect of repulsive linkage on

Table 3. Selected proportion and number of individuals (or doubled haploid lines [DHs]) selected in each marker selection scheme.

Breeding population	No enrichment selection in TCF ₂ [†]		Enrichment selection for all target genes in TCF ₂		Homozygous selected for <i>Rht8</i> , and enrichment selection for others in TCF ₂	
	Selected proportion	Minimum population size	Selected proportion	Minimum population size	Selected proportion	Minimum population size
TCF ₁ [‡]	0.0313	145	0.0316	144	0.0313	145
TCF ₂			0.1190	37	0.0397	114
DHs derived from TCF ₂	0.0013	3440	0.0112	408	0.0160	286
Total population size required		3585		589		545

[†] TCF₂, F₂ generation of the topcross.

[‡] In the F₁ generation of the topcross (TCF₁), homozygous selection is conducted for *Rht-B1a* and *Glu-B1i*, and enrichment selection for *Rht8*, *Cre1*, and *Tin*. The other loci are not segregating in TCF₁. The homozygous frequency for *Rht-B1a* and *Glu-B1i*, and the heterozygous frequencies for *Rht8*, *Cre1*, and *tin* are all equal to 0.5. So the theoretical selected proportion in TCF₁ is $0.5^5 = 0.0313$.

allele frequency. When three linked loci, *Glu-A3*, *tin*, and the marker for *tin* (denoted as *Mtin*), are considered, there are eight haplotypes (Table 4). When no crossover interference is assumed, the frequency of each haplotype can be calculated from the recombination frequency between *Glu-A3* and *tin*, and between *tin* and its marker (Table 4, last column). After MAS for *Glu-A3b* and *tin*, only Haplotypes 2 and 3 are retained, with a frequency for *tin* of $0.01488/(0.01488 + 0.00388) = 0.79318$, which in turn confirms our simulation results. The frequency of *tin* may not be sufficient, and therefore the presence of the *tin* allele following MAS must be confirmed by other methods.

Optimum Strategy to Combine Nine Genes from a Topcross

In summary, the optimum strategy to combine the nine target alleles in the topcross Silverstar+*tin*/HM14BS//Sunstate can be divided into four steps:

- Step 1.** Selection of Sunstate as the final parent (having largest number of favorable alleles) in the topcross
- Step 2.** Selection for *Rht-B1a* and *Glu-B1i* homozygotes, and enrichment of *Rht8*, *Cre1*, and *tin* in TCF₁
- Step 3.** Selection of homozygotes for one target allele, e.g., *Rht8*, and enrichment of remaining target alleles in TCF₂
- Step 4.** Selection of the target genotype (Table 1, last row) in DHs/RILs

The selected proportion in Table 3 can be used to determine the minimum population size for each selection stage. At this point, the presence of the *tin* gene needs to be reconfirmed by phenotyping. Currently, laboratory progeny marker screening and field selection experiments are underway with these populations so that we can validate the simulation results.

To identify the best strategy with the smallest minimum population size to recover one target genotype does not solve all the problems facing breeders when using MAS. Sometimes, breeders may want to know how many target genotypes can be selected at the end of the selection process. This is important if breeders want to select on other segregating traits for which no markers are available. For example, there are 500 individuals in the TCF₁, 50 seeds are taken from each selected individual after Step 1. After the selection of Step 2, 50 DHs are developed from each selected individual in TCF₂, based on which the selection of Step 3 is applied. From 1000 simulation runs, we found on average 15.73 individuals were selected in TCF₁, 31.43 were selected in TCF₂, and 16.50 DHs with the target genotype (Table 1) were selected at the end.

In practice, breeders can seldom repeat a breeding process. But simulation has the advantage of being able to investigate the outcome of a crossing/selection process for a large number of replications, from which the variation can be estimated. From the 1000 simulation runs, we found

the standard errors of selected individuals in TCF₁, TCF₂, and DHs are 4.00, 10.01, and 11.25, respectively. The frequency distribution of the number of selected individuals in TCF₁ and DHs are shown in Fig. 2. The number of selected individuals has a range from 5 to 31 in TCF₁, and a range from 0 to 76 in DHs. Simulation cannot determine the exact number of selected individuals for a single selection experiment but can determine the probability of selecting a certain number of target genotypes. For the selection process previously described, the probability is 0.995 to select one or more target genotypes, 0.645 to select 10 or more, and 0.287 to select 20 or more (Fig. 2). Thus a larger population may be required if the breeders want to select no less than 20 DHs, based on which the selection for other important traits can be applied.

Usefulness of Simulation Approaches in Breeding

As the number of published genes and QTLs for various traits increases, the challenge for plant breeders is to determine how to best utilize this knowledge to increase the efficiency of crop improvement and enhance genetic gain. Two types of selection involving markers are widely utilized (Bernardo, 2002). One is based on an index comprising both phenotypic value (usually for quantitative traits) and marker type (Lande and Thompson, 1990; Servin et al., 2004; Bernardo and Charcosset, 2006). The other is based on whether the marker is present or not (Young, 1999; Eagles et al., 2001; Kuchel et al., 2005) and is used to select for important genes in crosses between largely adapted parents or to backcross specific genes into adapted backgrounds. This article largely concerns the latter use of marker selection: the efficient combination of multiple, favorable alleles into lines that will typically be used as parents, or to release “converted” sister lines from crosses that already possess largely elite agronomic backgrounds.

Table 4. Haplotype, genotype, and frequency from the biparental cross between Sunstate and Silverstar+*tin*.

	<i>Glu-A3</i>	<i>tin</i>	<i>Mtin</i>	Frequency [†]
Haplotype 1	<i>Glu-A3b</i>	<i>Tin</i>	<i>MTin</i>	$(1 - r_1)(1 - r_2) / 2 = 0.48112$
Haplotype 2 [‡]	<i>Glu-A3b</i>	<i>Tin</i>	<i>Mtin</i>	$(1 - r_1)r_2 / 2 = 0.00388$
Haplotype 3 [‡]	<i>Glu-A3b</i>	<i>tin</i>	<i>Mtin</i>	$r_1(1 - r_2) / 2 = 0.01488$
Haplotype 4	<i>Glu-A3b</i>	<i>tin</i>	<i>MTin</i>	$r_1r_2 / 2 = 0.00012$
Haplotype 5	<i>Glu-A3c</i>	<i>Tin</i>	<i>Mtin</i>	$r_1r_2 / 2 = 0.00012$
Haplotype 6	<i>Glu-A3c</i>	<i>Tin</i>	<i>MTin</i>	$r_1(1 - r_2) / 2 = 0.01488$
Haplotype 7	<i>Glu-A3c</i>	<i>tin</i>	<i>MTin</i>	$(1 - r_1)r_2 / 2 = 0.00388$
Haplotype 8	<i>Glu-A3c</i>	<i>tin</i>	<i>Mtin</i>	$(1 - r_1)(1 - r_2) / 2 = 0.48112$
Frequency of <i>tin</i> after marker-assisted selection for <i>Glu-A3b</i> and <i>Mtin</i>				0.7932

[†] $r_1 = 0.03$ is the recombination frequency between *Glu-A3* and *tin* (3cM), $r_2 = 0.008$ is the recombination frequency between *tin* and its marker locus (0.8cM). The two alleles at *Glu-A3* are *Glu-A3b* and *Glu-A3c*, the two alleles at *tin* are *tin* and *Tin*, and the alleles for marker at *tin* are *Mtin* and *MTin*.

[‡] Haplotypes 2 and 3 only will be retained through marker-assisted selection, and *tin* frequency can be calculated from these (see text).

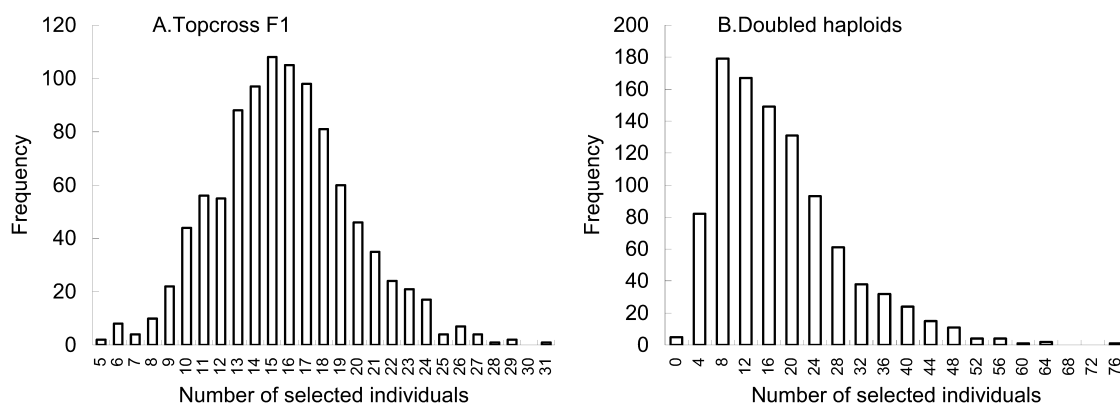


Figure 2. Frequency distribution of the number of selected individuals in the F_1 generation of the topcross (TCF_1) (A) and doubled haploid lines (DH) (B). The parents for the topcross were HM14BS, Sunstate, and Silverstar+*tin*. There were 500 individuals grown in the TCF_1 , 50 seeds were taken from each selected TCF_1 individual to produce the F_2 generation of the topcross (TCF_2), and 50 DHs were produced from each selected TCF_2 individual. In the TCF_1 generation, homozygous selection was applied for *Rht-B1a* and *Glu-B1i*, and enrichment selection for *Rht8*, *Cre1*, and *tin*. In the TCF_2 generation, homozygous selection was applied for *Rht8*, and enrichment selection for remaining target alleles. In DHs, only the target genotype was selected.

Computer simulation can help to investigate many possible crossing and selection scenarios. This allows many scenarios to be tested *in silico* in a relatively very short amount of time and helps breeders make some decisions before conducting highly resource-demanding field experiments.

In this article, we give practical guidelines and a specific example of combining alleles related to several traits into the same target genotype. In practice, our breeding program described uses population sizes slightly greater than those given, as our program attempts to recover more than a single genotype during recombination. These guidelines are most relevant when the genes of interest are already present in genotypes that have relatively “adapted” backgrounds for other complex agronomic traits, as we have not considered here the effects of random background selection in the donor parents (Frisch and Melchinger, 2005). An extension of this work to optimize selection where a quantitative trait of interest is associated with multiple QTLs and has complex gene action (including genotype by environment interaction) is currently underway.

Acknowledgments

This research was funded by the Generation Challenge Programme of CGIAR (www.generationcp.org), while the development of the simulation tools was funded by the GRDC (Grain Research and Development Corporation, Australia). The authors would like to acknowledge valuable discussions at various times with Richard Trethowan, Maarten van Ginkel, Howard Eagles, Mark Cooper, Dean Podlich and Mark Dieters in formulating ideas and methods for this article.

References

Bariana, H.S., and R.A. McIntosh. 1993. Cytogenetics studies in wheat XV: Location of rust resistance genes in VPM1 and their genetic linkage with other disease resistance genes in chromosome 2A. *Genome* 36:476–482.

- Bernardo, R. 2002. *Breeding for quantitative traits in plants*. Stemma Press, Woodbury, MN.
- Bernardo, R., and A. Charcosset. 2006. Usefulness of gene information in marker-assisted recurrent selection: A simulation appraisal. *Crop Sci.* 46:614–621.
- Bonnert, D.G., G.J. Rebetzke, and W. Spielmeyer. 2005. Strategies for efficient implementation of molecular markers in wheat breeding. *Mol. Breed.* 15:75–85.
- Botwright, T.L., G.J. Rebetzke, A.G. Condon, and R.A. Richards. 2001. The effect of *rht* genotype and temperature on coleoptile growth and dry matter partitioning in young wheat seedlings. *Aust. J. Plant Physiol.* 15:417–423.
- Dekkers, J.C.M., and F. Hospital. 2002. The use of molecular genetics in the improvement of agricultural populations. *Nat. Rev. Genet.* 3:22–32.
- Delphin Koudande, O., F. Iraqi, P.C. Thomson, A.J. Teale, and J.A.M. van Arendonk. 2000. Strategies to optimize marker-assisted introgression of multiple unlinked QTL. *Mamm. Genome* 11:145–150.
- Dubcovsky, J. 2004. Marker-assisted selection in public breeding programs: The wheat experience. *Crop Sci.* 44:1895–1898.
- Eagles, H.A., H.S. Bariana, F.C. Ogonnaya, G.J. Rebetzke, G.J. Hollamby, R.J. Henry, P.H. Henschke, and M. Carter. 2001. Implementation of markers in Australian wheat breeding. *Aust. J. Agric. Res.* 52:1349–1356.
- Ellis, M.H., W. Spielmeyer, K. Gale, G.J. Rebetzke, and R.A. Richards. 2002. Perfect markers for the *Rht-B1b* and *Rht-D1b* dwarfing mutations in wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* 105:1038–1042.
- Falconer, D.S., and T.F.C. Mackay. 1996. *Introduction to quantitative genetics*. 4th ed. Longman Group, Essex, UK.
- Frisch, M., and A.E. Melchinger. 2005. Selection theory for marker-assisted backcrossing. *Genetics* 170:909–917.
- Korzun, V., M.S. Roder, M.W. Ganal, A.J. Worland, and C.N. Law. 1998. Genetic analysis of the dwarfing gene (*Rht8*) in wheat: I. Molecular mapping of *Rht8* on the short arm of chromosome 2D of bread wheat (*Triticum aestivum*). *Theor. Appl. Genet.* 96:1104–1109.
- Kuchel, H., G. Ye, R. Fox, and S. Jefferies. 2005. Genetic and genomic analysis of a targeted marker-assisted wheat breeding

- strategy. *Mol. Breed.* 16:67–78.
- Lande, R., and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124:743–756.
- Ma, W., W. Zhang, and K.R. Gale. 2003. Multiplex-PCR typing of high molecular weight glutenin alleles in wheat. *Euphytica* 134:51–60.
- Ogbonnaya, F.C., N.C. Subrahmanyam, O. Moullet, J. de Majnik, H.A. Eagles, J.S. Brown, R.F. Eastwood, J. Kollmorgan, R. Appels, and E.S. Lagudah. 2001. Diagnostic DNA markers for cereal cyst nematode resistance in bread wheat. *Aust. J. Agric. Res.* 52:1367–1374.
- Paterson, A.H., S.D. Tanksley, and M.E. Sorrells. 1991. DNA markers in plant improvement. *Adv. Agron.* 46:39–90.
- Podlich, D.W., and M. Cooper. 1998. QU-GENE: A platform for quantitative analysis of genetic models. *Bioinformatics* 14:632–653.
- Rebetzke, G.J., and R.A. Richards. 2000. Gibberellic acid-sensitive dwarfing genes reduce plant height to increase kernel number and grain yield of wheat. *Aust. J. Agric. Res.* 51:235–245.
- Servin, B., O.C. Martin, M. Mezard, and F. Hospital. 2004. Toward a theory of marker-assisted pyramiding. *Genetics* 168:513–523.
- Spielmeyer, W., and R.A. Richards. 2004. Comparative mapping of wheat chromosome 1AS carrying tiller inhibition gene with corresponding rice chromosome 5. *Theor. Appl. Genet.* 109:1303–1310.
- Spielmeyer, W., P.W. Sharp, and E.S. Lagudah. 2003. Identification and validation of markers linked to broad-spectrum stem rust resistance gene *Sr2* in wheat. *Crop Sci.* 43:333–336.
- Wang, J., H.A. Eagles, R. Trethowan, and M. van Ginkel. 2005. Using computer simulation of the selection process and known gene information to assist in parental selection in wheat quality breeding. *Aust. J. Agric. Res.* 56:465–473.
- Wang, J., M. van Ginkel, D. Podlich, G. Ye, R. Trethowan, W. Pfeiffer, I.H. DeLacy, M. Cooper, and S. Rajaram. 2003. Comparison of two breeding strategies by computer simulation. *Crop Sci.* 43:1764–1773.
- Young, N.D. 1999. A cautiously optimistic vision for marker-assisted selection. *Mol. Breed.* 5:505–510.
- Zhang, W., M.C. Gianibelli, L.R. Rampling, and K.R. Gale. 2004. Characterisation and marker development for low molecular weight glutenin genes from *Glu-A3* alleles of bread wheat (*Triticum aestivum*. L). *Theor. Appl. Genet.* 108:1409–1419.

