International Journal of Plant Breeding ©2008 Global Science Books

Marker-assisted Gene Pyramiding for Inbred Line Development: Basic Principles and Practical Guidelines

Guoyou Ye[1*] • Kevin F. Smith[2]

[1] Bundoora Centre, Biosciences Division, The Department of Primary Industries Victoria, and Molecular Plant Breeding Cooperative Research Centre, 1 Park Drive, Bundoora Vic 3086, Australia

[2] Hamilton Centre, Biosciences Division, The Department of Primary Industries, Mount Napier Road, Hamilton, Vic 3300, Australia

Corresponding author: * **guoyou.ye@dpi.vic.gov.au**

ABSTRACT

Gene pyramiding, which aims to assemble multiple desirable genes into a single genotype, is a commonly used method in breeding for self-pollinated crops. Traditionally, the main use of gene pyramiding is to improve an existing elite cultivar through introgression of a few genes of large effects from other sources, since the presence of the target genes has to be monitored by phenotyping, which is only effective for major genes. Depending on the trait and inheritance of the targeted genes, gene pyramiding may require much labour, time and material resources. The development of modern plant molecular techniques and quantitative genetics in the last two decades has dramatically widened the applicability of gene pyramiding. It provides enhanced knowledge of the genetics of the breeding traits and of the relative genomic location of functionally related as well as neutral markers associated with the genes responsible for the traits. It facilitates the identification of genes with large effect for traits which are traditionally regarded as quantitative and not targeted by gene pyramiding program. Marker-based selection reduces/eliminates extensive phenotyping, provides more effective options to control linkage drag, makes the pyramiding of genes with very similar phenotypic effects possible, and reduces the breeding duration. Marker-based gene pyramiding is now the method of choice for inbred line development targeted at improving traits controlled by major genes. In this review, we focus on aspects of designing an efficient marker-based gene pyramiding strategy for inbred line development. The basic principles of gene pyramiding, the process and useful guidelines for designing an efficient strategy, and the integration of gene discovery and pyramiding are discussed in this paper, while the successful use of gene pyramiding in practical breeding is summarised in a companion paper.

Keywords: functional markers, introgression lines, marker-assisted selection, molecular markers, quantitative trait loci

CONTENTS

INTRODUCTION

Gene pyramiding is defined as a method aimed at assembling multiple desirable genes from multiple parents into a single genotype. The end product of a gene pyramiding program is a genotype with all of the target genes. Generally speaking, the objectives of gene pyramiding include: 1) enhancing trait performance by combining two or more complementary genes, 2) remedying deficits by introgressing genes from other sources, 3) increasing the durability

Received: 8 January, 2008. Accepted: 17 February, 2008.

Invited Review

of disease and/or disease resistance, and 4) broadening the genetic basis of released cultivars.

Traditionally, gene pyramiding is mainly used to improve qualitative traits such as disease and insect resistance. This is associated with the fact that the presence of target trait genes must be confirmed by phenotyping mostly at the individual level and that individual phenotypic performance is a good indicator of the genotype only if genes have a major effect on phenotypic performance and the error of phenotyping is minimal. In addition to the reliability of phenotyping at individual level other factors influencing the success of gene pyramiding are the inheritance model of the genes for the target traits, linkage and/or pleiotropism between the target trait and other traits. For instance, allelic genes cannot be combined in the same genotype. The effect conferred by a recessive gene cannot be evaluated on heterozygous individuals and progeny testing is required. If the target gene is tightly linked to genes with large negative effects on other traits, these undesirable genes may be transferred together with the target gene into the recipient line and result in reduced performance of other traits (linkage drag). Therefore, any improvement in the knowledge of the trait genetics (inheritance, genetic relationship, etc.) and techniques for inferring genotype-phenotype relationship will be useful.

The development of modern plant molecular and quantitative genetics in the last two decades has the potential to revolutionise what has mostly been experienced-based empirical plant breeding. This development has enhanced our knowledge of the genetics of the breeding traits and the relative genomic location of functionally related and neutral markers associated with the genes responsible for the traits. It has also widened several aspects of the practical application of gene pyramiding. Firstly, for traits that are simply inherited, but that are difficult or expensive to measure phenotypically, and/or that do not have a consistent phenotypic expression under certain specific selection conditions, marker-based selection is more effective and /or economic than phenotypic selection (Paterson *et al*. 1991; Stuber *et al*. 1999; Eagles *et al*. 2001; Dekkers and Hospital 2002; Langridge and Barr 2003; Dubcovsky 2004). Secondly, traits which are traditionally regarded as quantitative and not targeted by gene pyramiding program can be improved using gene pyramiding if major genes affecting the trait are identified (Ashikari and Matsuoka 2006). Thirdly, genes with very similar phenotypic effects, which are impossible or difficult to combine in single genotype using phenotypic selection, can be pyramided through marker-assisted selection (MAS). Fourthly, markers provide a more effective option to control linkage drag and speed up the recovery of recurrent genome and make the use of genes contained in unadapted resources easier (Frisch and Melchinger 2005). Marker-assisted gene pyramiding is currently the method of choice for inbred line development targeted at improving traits controlled by major genes. In this review, we focus on aspects of designing an efficient marker-based gene pyramiding strategy for inbred line development. The basic principles of gene pyramiding, the process and useful guidelines for designing an efficient strategy, and the integration of gene discovery and pyramiding are discussed in this paper, while the successful use of gene pyramiding in practical breeding is summarised in the next paper (Ye and Smith 2008). The marker-based recurrent backcrossing, which can be regarded as a simple form of gene pyramiding, is also discussed in Ye and Smith (2008).

GENERAL PRINCIPLES OF MARKER-ASSISTED GENE PYRAMIDING

Basic assumptions and characteristics

Before the general principle of gene pyramiding is discussed, it is necessary to outline the general situation faced by a gene pyramiding program. The basic assumptions are: 1) that the locations of a series of genes of interest (target genes) and thus the linkage relationship between them is known; 2) that a target genotype for these genes is defined prior to selection as the genotype with favourable alleles at all loci of interest; 3) that the genotype of an individual can be identified by these genes or markers linked to them; 4) that a collection of lines containing all the target genes are available.

In principle, pyramiding multiple genes is achieved by crossing parental lines with complementary desirable genes and selecting the desired recombinants from among the progeny population (Allard 1999). Since breeding is a very time-consuming process, breeders aim to combine as many desirable alleles as possible in a single breeding cycle (from crossing to the generation of near-homozygous breeding lines). When the number of genes to be assembled is known, the goal of gene pyramiding is to obtain near-homozygous breeding lines that are fully homozygous for the desirable alleles of the target genes using the minimum number of generations of selection and the lowest genotyping and phenotyping costs. This suggests that the total cost and duration are the two principal criteria when pyramiding strategies are designed and compared. Both the phenotyping and genotyping costs can be roughly quantified using the total number of plants to be screened for phenotypic performance and genotypic status, respectively. Therefore, the cost of a strategy can be roughly estimated with the sum of minimal population sizes at each of the generations. The factors that affect the cost and duration of a gene pyramiding scheme are discussed below.

Frequencies of genotypes in a segregating population

The expected multi-locus genotype frequency in a segregating population is one of the two factors affecting the minimum population size for the successful recovery of the desirable genotype. Since the expected genotype frequency is the product of the corresponding gametes produced by the two parents, the expected frequencies of the gamete types produced by each of the two parents are needed. Assuming the absence of crossover interference, the expected multi-locus gamete frequency produced by a genotype can be computed as follows:

$$\frac{1}{2}\prod_{i=1}^{n-1}\alpha_{i,i+1}$$

where $\alpha_{i,i+1} = r_{i,i+1}$ if the gamete is the recombinant type, whereas $\alpha_{i,i+1} = 1- r_{i,i+1}$ if the gamete is the parental type. $r_{i,i+1}$ is the recombination frequency between the i-th gene and the i+1-th gene. For instance, the expected frequency of the gamete '110' produced by the genotype '111/000' is $0.5(1-r_{1,2})r_{2,3}$, since for genes 1 and 2, the gamete ('11') is the parental type but for genes 2 and 3 the gamete ('10') is the recombinant type.

When there are many genes, it is better to organize them into linkage groups. Genes in a group are linked, while genes in different groups are independent. Then, the computation of gamete frequencies can be done separately for each group. The frequencies of gametes for all genes can be calculated as the product of the frequencies of the corresponding gametes for each of the linkage groups. For instance, if six genes are grouped into three linkage groups, L1 (1 and 2), L2 (3 and 4), and L3 (5 and 6), the gametes and their frequencies produced by the genotype '111100/000000' can be computed as given in **Table 1**.

When the desirable genotype can be produced by the union of different types of gametes, its frequency is computed by summing all the possible paths leading to the genotype. This does not occur if each of the founding parents has only one target gene and is used once in the entire scheme as assumed by Servin *et al.* (2004). However, in a more practical breeding situation, it is likely to occur. Theoretical prediction equations for many situations have

Table 1 Expected frequencies of gametes produced by the genotype '111100/000000'.

L1		L2		L3		Six Loci	
Gamete	**Frequency**	**Gamete**	**Frequency**	**Gamete**	**Frequency**	**Gamete**	**Frequency**
00	$(1-r_1)/2$	00	$(1-r_2)/2$	00	1	00 00 00	$[(1-r_1)/2]\ [(1-r_2)/2]$
		01	$r_2/2$	00	1	00 01 00	$[(1-r_1)/2]\ (r_2/2)$
		10	$r_2/2$	00	1	00 10 00	$[(1-r_1)/2]\ (r_2/2)$
		11	$(1-r_2)/2$	00	1	00 11 00	$[(1-r_1)/2]\ [(1-r_2)/2\ ]$
01	$r_1/2$	00	$(1-r_2)/2$	00	1	01 00 00	$[r_1/2]\ [(1-r_2/)/2]$
		01	$r_2/2$	00	1	01 01 00	$(r_1/2)(r_2/2)$
		10	$r_2/2$	00	1	01 10 00	$(r_1/2)(r_2/2)$
		11	$(1-r_2)/2$	00	1	01 11 00	$(r_1/2)\ [(1-r_2)/2]$
10	$r_1/2$	00	$(1-r_2)/2$	00	1	10 00 00	$(r_1/2)\ [(1-r_2/)/2]$
		01	$r_2/2$	00	1	10 01 00	$(r_1/2)(r_2/2)$
		10	$r_2/2$	00	1	10 10 00	$(r_1/2)(r_2/2)$
		11	$(1-r_2)/2$	00	1	10 11 00	$(r_1/2)\ [(1-r_2)/2]$
11	$(1-r_1)/2$	00	$(1-r_2)/2$	00	1	11 00 00	$[(1-r_1)/2]\ [(1-r_2)/2]$
		01	$r_2/2$	00	1	11 01 00	$[(1-r_1)/2]\ (r_2/2)$
		10	$r_2/2$	00	1	11 10 00	$[(1-r_1)/2]\ (r_2/2)$
		11	$(1-r_2)/2$	00	1	11 11 00	$[(1-r_1)/2]\ [(1-r_2)/2\ ]$

been developed (Bailey 1961). Ye *et al*. (2004) put forward a theoretical prediction method to compute frequencies of genotypes in a population developed from a three-way cross, using a single seed descent (SSD) or doubled haploid approach, and developed a computer tool, CrossPredictor, to allow easy computation from parental genotypic information.

Minimal population size for the recovery of desirable genotype

The minimum population size required to ensure with a probability (i.e. q = 95%, or 99%), that a genotype is present in a segregating population can be computed using the binomial distribution. Assuming that the frequency of the desired genotype is p, the number of individuals of the desirable genotype (m) in a population of size N follows a binomial distribution.

$$P(Y=m)=\binom{N}{m}p^m(1-p)^{N-m}$$

The probability q that at least one individual among N individuals has the desired genotype is

$$q=P(Y>0)=1-P(Y=0)=1-(1-p)^N$$

From this equation, the minimum population size to ensure a predetermined probability (i.e. q = 95%, or 99%) that at least one desired genotype is present in the population can be derived from the following equation.

$$N_q=\frac{\ln(1-q)}{\ln(1-p)} \qquad (1)$$

Although a single individual is theoretically sufficient to achieve the selection objective, it is, nevertheless, risky from both a statistical and practical points of view. Unexpected or unpredictable field problems, such as low germination or poor pollen quality, can have dramatic consequences on gene and genotypic frequencies. In addition, if seed collectable from a single plant is used, progeny size may be so small that chance deviations become relatively large. Based on these considerations, the successful identification of more than one desired individual should be considered for any recommendations to represent breeders' realities.

The probability q that among N individuals at least m individuals have the desired genotype is

$$q_m=P(Y\geq m)=1-\sum_{i=0}^{m-1}P(Y=i)$$

And the minimum population size required can be computed using equation 1 by replacing q with q_m.

When the number of genes is large and/or the linkage relationships are complex, many computations are required if a purely mathematical prediction method is used. The computational requirements will be further increased if markers are not completely linked to the target genes (i.e. are not diagnostic), and if several flanking markers are used. In addition, many generations are involved in a gene pyramiding scheme, and the frequencies of the desirable genotypes and the population sizes in all the generations must be computed. Though theoretically simple the computation is very tedious. Special computer software has been developed to compute the frequencies of all possible genotypes in the segregating population (Servin *et al*. 2002). Alternatively, computer simulation software can be used. Computer simulation gives a very good approximation of the expected value if enough runs are repeated. Such simulations can be used to replace the above theoretical predictions when designing crossing and selection strategies. They also provide a stochastic and statistical framework that will allow statements of relative accuracy of the simulation (Podlich and Cooper 1998; Wang *et al*. 2003; Kuchel *et al*. 2005; Wang *et al*. 2005; Ye *at al*. 2007).

PROCESS OF DESIGNING A GENE PYRAMIDING STRATEGY

Bringing all the desirable alleles into a single genotype is the overall objective of a gene pyramiding program. When the number of parental lines containing the desirable genes (founding parents) is more than three, more than one crossing scheme can result in the generation of the target genotype. Therefore, the gene pyramiding scheme can be divided into two parts. The first part is aimed at cumulating one copy of all target genes in a single genotype (called root genotype). The second part is aimed at fixing the target genes into a homozygous state, that is, to derive the target genotype from the root genotype. Sevrin *et al*. (2004) called these two parts pedigree and fixation, respectively. **Fig. 1** is an example of a gene-pyramiding scheme cumulating six target genes from six parental lines.

Designing the Fixation Scheme

Assuming that a genotype with a copy of the desirable allele at each of the targeted loci (root genotype) is available, the design of an optimal strategy is aimed to find the minimum number of generations for genotyping and/or phenotyping required to fix all the loci for the desirable alleles within the limit of the largest possible population size applicable. The most commonly used methods for the production of homozygous individuals are the development of re-

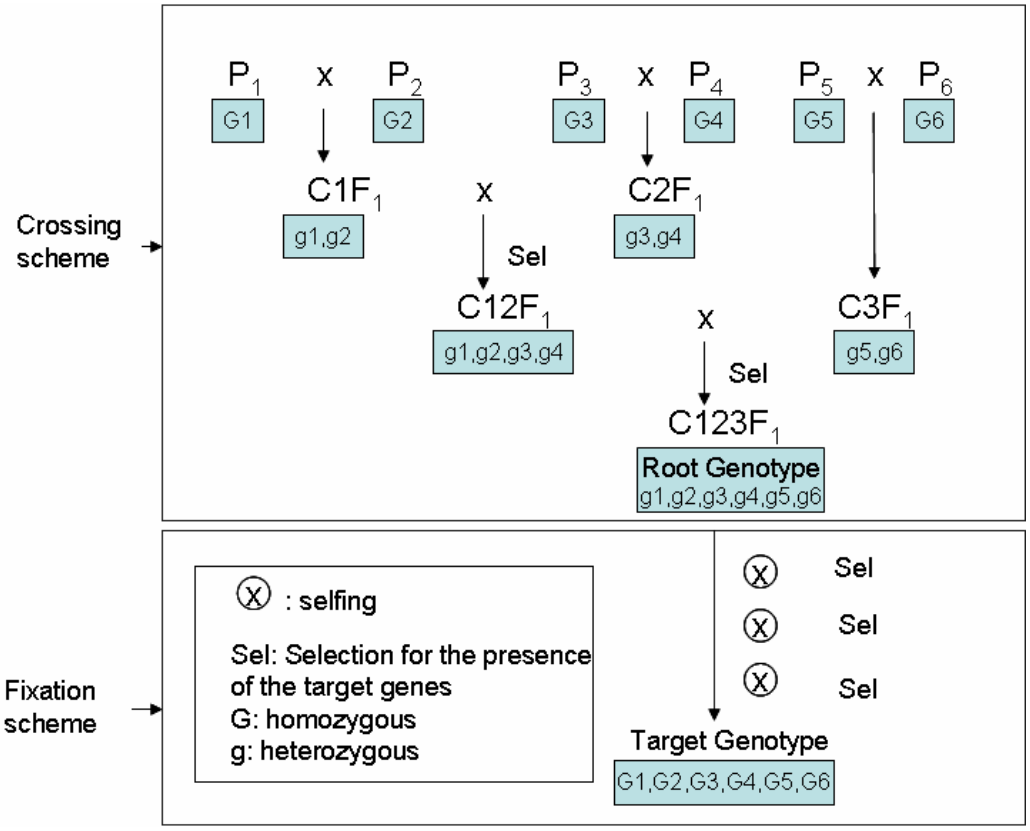

Fig. 1 Example of a gene-pyramiding scheme cumulating six target genes.

combinant inbred lines (RIL), and doubled-haploid (DH) population. Therefore, it is advisable to investigate the feasibility of achieving the objective using RIL or DH. With the recombination frequencies between the target genes, the proportion of the desired genotype in the RIL or DH population can be worked out and the minimum population size required can be determined using binomial distribution. Either RIL or DH can be adopted if this size of population is practically achievable and the cost of genotyping it is affordable, although it may not be the optimal scheme.

If neither RIL nor DH options are feasible, repeated selection in more than one subsequent segregating generation is required. Selection in sequential generations of individuals that have an increasing number of the desired alleles fixed at the desired loci while heterozygous at the remaining desired loci increases the frequency of the targeted recombinant through accumulated recombination. The objective of this step is to identify a selection scheme that leads to the production of the target genotype using the minimum number of generations and the practically allowable population sizes in each of the generations. Ye *et al*. (2007) showed how to define such a selection strategy in steps. They (Ye *et al*. 2007) only considered the use of self-pollination in all subsequent generations, since it is the least expensive mating options in self-pollinated species and produces relatively more progeny. But, it may be less efficient since self-pollination breaks the already established desirable linkages between some of the favourable alleles. When crossing to another genotype can be easily and cheaply conducted, two other options may be taken as suggested by Servin *et al*. (2004). One option is to cross to a founding parent. The advantage of crossing to a founding parent is that the probability of obtaining a genotype that is homozygous for the target genes brought by the founding parent but heterozygous for the other targets is high. Hence, that target gene need not be fixed subsequently, increasing the probability of getting the target genotype. The choice of the parent to use may be subject to particular considerations depending on the value of the founding parents, the position of the loci, etc. The other option is to cross to a blank line containing none of the favourable alleles. The use of blank line increases the chance of obtaining a genotype carrying all favourable alleles in coupling and thereby increasing the frequency of target genotype in subsequent generations. If the number of generations required and/or the total population size is too large and thus genotyping is not acceptable, then the objective is deemed to be too ambitious and unachievable and the number of genes to be pyramided has to be reduced.

Designing the Cross Scheme

A crossing scheme which leads to the production of the root genotype needs to be designed if the objective is achievable based on the above step. With the assumption that every founding parent is involved in only one cross in the gene-pyramiding scheme, Servin *et al*. (2004) described an algorithm for the building of every possible succession of pair crosses leading to the target genotype. They developed a computer program to generate all the possible schemes and associated minimal population size and the largest of the population sizes to be handled at any segregating generations or steps during the pyramiding process. The number of possible satisfactory schemes increases very fast with the number of genes. Even with the computer program, it is impossible to evaluate all the satisfactory schemes when the number of loci is more than a dozen. Ishii and Yonezawa (2007a) suggested some guidelines by investigating the efficiencies of a series of crossing schemes. The following section gives some guidelines for the choice of a crossing scheme.

GUIDELINES FOR DESIGNING AN EFFICIENT GENE PYRAMIDING STARTEGY

Guidelines for designing a gene pyramiding crossing scheme

Founding parents with fewer target markers enter the schedule at earlier stages

This guideline is based on the following facts: 1) Once a target gene has been incorporated into an intermediate genotype, genotyping must be done in all later stages to ensure its presence. Therefore, founding parents with more

target genes should be used in later stage. 2) Target genes containing in a founding parent are in desired linkage phase, which may be broken down due to recombination. The more the meiosis involved the lower the probability of maintaining the desired linkage.

A cross that invokes a strong repulsion linkage should be performed as early as possible

When the target genes are linked, genes linked in repulsion at some stages of the pyramiding is unavoidable and selection for recombinants is required. As the frequency of recombinant type is always lower than that of the parental types, larger population sizes are required to recover the desired recombinant. In the genome of a plant selected at each stage, genes that were newly incorporated via the latest crossing are linked in the repulsion phase with genes that had been incorporated at other stages before the latest. However, all genes incorporated in any stages before the latest are linked in coupling phase because it is converted to a coupling linkage after one round of MAS. On one hand, more plants and markers need to be tested in later stages since the number of target genes increases with the advancement of stages, repulsion linkage of the same strength is more disadvantageous when it occurs in at later stages. On the other hand, a repulsion linkage, once converted to the coupling phase after one round of marker selection, contributes to reduction of the number of tested plants in all subsequent stages. Similarly, the order of crossing should aim at a minimum occurrence of duplicate repulsion linkages.

More crosses should be conducted at each generation if genotyping cost is low and the practically applicable population size is large

When the maximum number of crosses is performed at each generation, the number of generations required to generate the root genotype is reduced and thus the total duration of the pyramiding program is reduced. Servin *et al*. (2004) showed that the number of generations required (h) is between ln *n* and n-1 (n is the number of founding parents) if every founding parent is involved in only one cross. However, the number of individuals (population size) must be large enough to ensure the recovery of the desirable genotype, which necessitates more genotyping.

One cross per generation is required if the practically applicable population size is small or genotyping cost is high

In this type of crossing design, from the second generation the desirable genotype is formed by a recombinant gamete produced by the selected genotype in the last generation and a gamete of the newly introduced parent. The probability of the desirable genotype is much higher than in schemes where the other parent is also a selected individual from the last crossing generation and thus the desirable gamete of this parent is recombinant type as well. The drawback to this crossing design is that the number of generations is large and the production of the new line is delayed.

Using backcrossing before assembling more genes

When the required population size at any stage is too large to be practicable, the use of backcrossing before assembling more genes is advisable. For example, if the selected individual '111000/000100' is crossed to a selected individual '000010/000001'from another cross to assemble the target genes 5 and 6, the probability of the desirable genotype '111100/000011' in the progeny population is $[0.5(1-r_1)(1-r_2)r_3][(0.5r_5)]$. If '111000/000100' is backcrossed to the genotype '000100/000100' and the genotype '111000/000100' is selected and crossed to '000010/000001', the probability of the desirable genotype '111100/000011' in the progeny population is $[0.5(1-r_1)(1-r_2))[(0.5r_5)]$, which is $1/r_3$ times higher than the former case.

Methods for enhancing the efficiency of the fixation step

Crossing between selected individuals

As aforementioned, the success of a gene pyramiding strategy depends on obtaining the target genotype within the time frame and cost defined by breeders irrespective of inter-mating between selected individuals at each generation. If a satisfactory genotype with at least one copy of the desirable alleles in all target loci is missing at any generation, crossing two plants with the best complementary genotypes can be used to secure the program. Crossing between complementary genotypes may also speed up the breeding process even if a satisfactory genotype is present.

Ishii and Yonezawa (2007b) demonstrated that a strategy which combines the use of haplo-diploidization and crossing between selected plants is highly efficient. Haplo-diploidization is used to generate homozygous genotypes in a single generation. The frequency of the target genotype in the DH population equals to the frequency of the gamete with a copy of desirable alleles in all target loci, which is much higher than that in a self-pollinated population. If a DH plant of the target genotype is lacking, two DH plants with the best complementary genotypes are crossed to produce a hybrid, which in turn is haplo-diploidized for the next round of selection. In this strategy, even a plant having as many as 20 target markers can be obtained at an almost perfect certainty in about three rounds of selection and a maximum of 200 tested plants per round. We would not regard any strategy which relies on the production of homozygous lines with complementary compositions of the target genes to achieve the final objective as effective, since theoretically unlimited number of genes can be pyramided if such intermediate genotypes are produced and the breeding program is continued.

Crossing the root genotype to a genotype with several desirable genes

As mentioned before, backcrossing to a genotype with desirable genes provides the opportunity for the selection of individuals homozygous for the genes hosted by the tester and the conversion of an undesirable phase to a desirable one. Crossing the root genotype to an elite line without the desired genes can also be used to convert the undesirable phase to the desirable one. This will effectively reduce the required population size in subsequent generations. This idea can be extended to generations before the formation of root genotype. Rather than crossing two intermediate genotypes selected from two crosses, the two selected individuals can be crossed to one of their parents or an elite genotype without the desired genes. Individuals with all desirable genes in coupling phase are selected from these backcrossing progeny population and are then crossed to combine the desirable genes. The drawback to this two-step procedure is the extra crossing generation. Servin *et al*. (2004) demonstrated that the two-step approach could be more effective under certain circumstances.

Advancing all satisfactory genotypes

In some generations more than one (partially heterozygous) genotype has the potential to produce the genotype desired for the next generation. These satisfactory genotypes may have different frequencies and different progeny sizes are needed to best realise their potential. The most efficient strategy will be the one that promotes the most satisfactory genotype (fewer loci are segregating, and the segregating loci are in the desirable linkage phase) at each generation. However, it may be beneficial in practice to obtain more than one individual of the desirable genotype, since then selection for other traits can also be conducted among these otherwise genetically distinct lines. In this scenario, all the satisfactory genotypes should be advanced to the next generation. Despite the advantages of weighing and considering optimum population sizes, the benefit of using different population sizes might be limited due to practical considera-

tions. An estimate of the actual number of seeds produced per plant is usually known and thus defines an upper population size limit.

MAIN FACTORS AFFECTING GENE PYRAMIDING

Characteristics of the target traits/genes

When the genes to be pyramided are functionally well characterised and markers used for selection equal to the gene itself (perfect markers), gene pyramiding will be more successful. For qualitative traits controlled by one or a few genes, the identification of the genes and tightly linked markers is easier provided phenotyping is carefully conducted. One or two markers per gene can be used for tracing the presence/absence of the target genes. Bulk segregant analysis (BSA) is the preferred method for the identification of markers tightly linked to a major gene (Michelmore *et al*. 1991). For BSA plants from a segregating population are grouped according to phenotypic expression of the trait into two bulks. The bulks are screened with a large number of markers to identify those that distinguish the bulks and, by inference, must be genetically linked to the trait locus. When the target genes are QTL with moderate or small effects, pyramiding may be less successful due to the following reasons. Firstly, the identified QTL may be more likely to be a false positive. Secondly, inaccurate QTL localizations result in the need to select for more marker loci covering large genomic segments to be certain that target QTL alleles are retained in selected progeny (Hospital and Charcosset 1997). Thirdly, QTL effects may be specific to a particular genetic background. Moreover, markers identified for a QTL can be ineffective in monitoring the QTL since the marker-QTL association might be different from population to population. Fourthly, more QTL need to be pyramided to achieve a significant improvement.

Reproductive characteristics

The propagation capability of a crop is determined by the number of seeds produced by a single plant. This capacity determines the population size applicable if seed has to be collected from only a single plant. In a gene pyramiding program, in most generations this is the case, since the chance of selecting two or more individuals of exactly the same genotype in previous generation is very low. For example, although a fairly large F_2 population can be obtained by collecting seed from many F_1 plants of the cross between two homozygous parents, from the F_3 generation seed can only be collected from a single plant. The fact that F_1 plants of the cross between two homozygous parents are genetically the same can also be used to increase the size of a progeny population of the F_1 plants of two crosses (double cross) or of the F_1 plants of one cross and an inbred line (Three-way cross or testcross).

The efficiency of hybridization may be an important constraint for some crop species. When wild relatives are used as the donor of desirable genes, many more reproduction related constraints may exist including cross incompatibility between the wild species and cultivated crop. F_1-hybrid sterility, infertility of the segregating generations, reduced recombination between the chromosomes of the two species. Appropriate techniques that may include chemical treatment and immature embryo culture for overcoming these problems must be established.

A breeder's capability to identify the 'desired' genotypes

It is obvious that the desirable genes must be present in all generations leading to the target genotype. To ensure the presence of the target genes individuals of desired genotype (which may change with generation advance) must be identified among all individuals in each generation. Breeder's capability to identify the desired genotypes has been greatly enhanced by the use of tightly linked or diagnostic markers. It might be appropriate to consider the importance of marker and trait gene linkage here.

The efficiency of marker-based gene pyramiding will decrease substantially if the markers are not perfectly or tightly linked with useful trait genes. Association of a marker with a trait allele and consequently the reliability of the marker-based selection decreases with increasing cycles of meiosis. With a recombination value of r between a marker and trait allele, the probability of this linkage being maintained across m cycles of meiosis is equal to $(1 - r)^m$. To keep this probability higher than a certain critical value, say P, m must not exceed lnP/ ln(1 - r), suggesting that a phenotypic test should be performed every m generations of selection to confirm the persistence of the initial linkage. Put in another way, the probability of losing the target allele by recombination is $1 - (1- r)^t$. For example, if the marker locus exhibits 10% recombination with the target gene, there is a 10% chance of losing the target allele each generation, and a 27% chance of losing the target allele after three generations of meiosis. However, if the recombination frequency is 1%, there is only a 3% chance of losing the target allele after three generations of meiosis.

When tightly linked markers are not available, selection on a pair of markers flanking the target locus can be very effective. If two marker loci *M1* and *M2* flank the target locus, one would select progeny that have both *M1* and *M2* alleles. The probability of losing the target allele with flanking marker selection is equal to the probability of selecting a double recombinant progeny from among the doubly heterozygous backcross progeny. If the flanking loci have recombination frequencies r_1 and r_2, respectively, with the target locus, the probability of losing the target allele due to double crossovers within the selected region is:

$$\frac{r_1 r_2}{1 - r_1 - r_2 + 2 r_1 r_2} .$$

This probability can be much lower than the probability of losing the target allele based on selection for a single marker. For example, if the flanking markers each have 10% recombination frequency with the target locus, there is only a 1.2% chance of losing the target allele after a single generation. In any case, with tighter linkage, the chance of losing the target allele is reduced. However, this requires more plants to be tested and higher cost per plant. It is imperative to use markers that are tightly linked with trait genes.

Multiple marker loci closely linked to the target gene, permits discrimination on the basis of the haplotype of several markers rather than just the genotype at one marker. For example, Cregan *et al*. (1999) developed two SSR markers tightly linked to the *rhg1* gene. Neither marker alone could distinguish all resistant from all susceptible genotypes, because of identity in state alleles shared by some resistant and susceptible lines, but the two markers together could discriminate almost all resistant and susceptible lines. One resistant cultivar carried the susceptible allele at both loci, presumably due to recombination between marker and resistance loci during line development. Thus, recombination can change the linkage phase between markers, but if MAS is used first to select putatively resistant lines, followed by phenotypic evaluation of resistance, the linkage phase will remain intact in all selected progeny. Therefore, MAS can be self-reinforcing, ensuring that the same set of markers will be effective in future crosses.

Operating capital

All breeding programmes are operated within the limits of available operating capital. Therefore, reducing the overall cost is always an important consideration when choosing a strategy. In addition to the use of the most economic mating and testing approaches, other factors affecting the cost also need to be considered. In the context of gene pyramiding,

cost affects both what can be achieved and how to achieve it. Increasing the number of generations (duration) will reduce the pressure on population size required in each generation and may result in the reduction of the total cost. However, increasing the duration delays the release of the new cultivar and consequently reduced market share. The well-known trade-off between duration and cost in breeding has no exception in gene pyramiding. To find an optimum balance between duration and cost is desired but very difficult to achieve. In practice, the best strategy may be defined as the one that enabled the breeder to achieve the objectives with the shortest duration and within a fixed expected investment.

INTEGRATING GENE DISCOVERY, VALIDATION AND PYRAMIDING

The principles of gene pyramiding discussed above assume that parental lines containing the target genes and markers linked to the target genes are available. This assumption also implies that the effects of the target genes and the linkage phase between the target genes and their linked markers are also known. It is relatively easier to identify genes and markers linked to them for qualitative traits, since the inheritance is usually very simple and the effect of gene is highly predictable. When the expression of a trait is controlled by multiple genes with relatively small effects, it is a quantitatively trait. Most of the important agronomic traits such as yield, stress resistance and quality are quantitative traits. Genes for quantitative traits are more difficult to identify. Quantitative traits loci (QTL) mapping using purposely generated mapping populations such as F_2 plants, backcross plants, Recombinant Inbred Lines (RIL), Backcross Inbred Lines (BIL) or Doubled Haploid Lines (DHL), as well as a linkage map constructed using molecular markers are currently the standard approach for identifying QTL controlling quantitative trait. A large population size is required to provide sufficient detection power. For example, nearly 300 F_2 progeny are required to detect a QTL responsible for at least 10% of the total variance. The QTL are localized with relatively poor resolution, typically approximately 20 cM. Since many QTL segregates within the mapping population contributing 'phenotypic noise', it is difficult to ascertain whether a given plant has inherited a specific QTL allele. Moreover, the lines (plants) of the mapping population may be difficult to use directly as parents in a pyramiding program. Methods based on the combined use of advanced backcrossing and marker-assisted selection, advanced backcross QTL analysis (AB-QTL) and the use of introgression lines (ILs), have been proposed and used in practice to better integrate QTL identification and the production of materials more suitable for breeding use. The rapid development of plant genomics makes it possible to develop markers that identify the alleles directly (that is, functional markers (FMs)) and thus eliminate the unknown/or changing linkage phase problems faced when using linked markers.

Advanced backcross QTL analysis (AB-QTL)

The advanced backcross QTL analysis (AB-QTL) was proposed by Tanksley and Nelson (1996) to simultaneously identify and introgress favourable alleles from unadapted donors into elite background. The general AB-QTL analysis is comprised of the following experimental phases:

1) Generating an elite by donor hybrid,

2) Backcrossing to the elite parent to produce BC_1 population which is subjected to marker/or phenotypic selection against undesirable donor alleles,

3) Genotyping BC_2 or BC_3 population with polymorphic molecular markers,

4) Evaluating the segregating BC_2F_2 or BC_2F_3 population for traits of interest and QTL analysis,

5) Selecting target genomic regions containing useful donor alleles for the production of NILs in the elite genetic background and

6) Evaluation of the agronomic traits of the NILs and elite parent controls in replicated environments.

The AB-QTL approach has been evaluated in many crop plant species to determine whether genomic regions derived from wild or unadapted germplasm have the potential to improve yield. However, the donor genome may mask the magnitude of some favourable effects that were identified for certain introgressed alleles. Thus, the trait-promoting QTL may not make a substantial contribution to the phenotype and the best lines may be inferior to commercial cultivars. A major limitation to AB-QTL is the difficulty in maintaining an adequate population size in selected backcross populations so that useful alleles are not lost and the QTL can be accurately mapped (Varshney *et al.* 2005).

Introgression lines (ILs)

The use of introgression lines (ILs) aimed at capitalizing on the genetic diversity in exotic germplasm and its use for breeding as well as gene discovery was proposed by Eshed and Zamir (1994a, 1994b). ILs are produced by systematic backcrossing and introgression of marker-defined exotic segments in the background of elite varieties. An example of ILs development scheme is given in **Fig. 2**. These ILs can be considered to be similar to a genomic library with a huge genome insert. Phenotypic characterization of each line can reveal the chromosome fragment from the donor with the gene(s) associated with a trait of interest. Multiple traits can be studied in one population using the same genotypic data. Since identifying QTL genes using ILs does not require linkage map construction or sophisticated statistical analysis for QTL, this is a more user-friendly method for practical breeding programs and also for biological science. ILs enable the phenotypic analysis of specific QTL and offer a common genetic background in which direct comparison of two lines can be used to evaluate the phenotype conditioned by a single introgressed exotic segment (Tanksley *et al*. 1996). The resolution and statistical power of QTL mapping is increased, because excluding extra genetic factors reduces phenotypic variation. ILs facilitate fine mapping of QTL, because the location of a QTL can be narrowed to a smaller genomic interval by evaluating a series of ILs that differ for overlapping regions of the genome (Paterson *et al*. 1990). ILs are also a valuable resources for the unravelling of gene function by expression profiling or map-based cloning (Eshed and Zamir 1995). ILs can be easily evaluated for all the important traits to identify any undesirable traits linked to the target gene(s) due to the relatively large chromosome segment introgressed and these traits can be identified before pyramiding. If necessary, undesirable genes should and can be eliminated by chromosome recombination in the progeny between the IL and the recurrent parent and screened by MAS. Since ILs contain only a low percentage of exotic germplasm the elimination of unfavourable exotic alleles can be easily and rapidly accomplished. This will speed up the transfer of the desirable alleles into the elite varieties (Ashikari and Matsuoka 2006).

Functional markers

The maintenance of the linkage phase between the target gene and its linked markers across multiple populations presents a serious problem for selecting for the target gene using markers. Markers linked to the QTL identified by linkage mapping using one or a few populations may or may not be useful in gene pyramiding because different subsets of QTL will be polymorphic in each population, and the linkage phases between the marker and QTL alleles can differ even between closely related genotypes. The linkage phase also tends to be more consistent if the source of QTL is from a gene pool which is very distinct from the one used by the breeders. Thus, markers linked to novel alleles from exotic germplasm or wild relatives are more likely to be

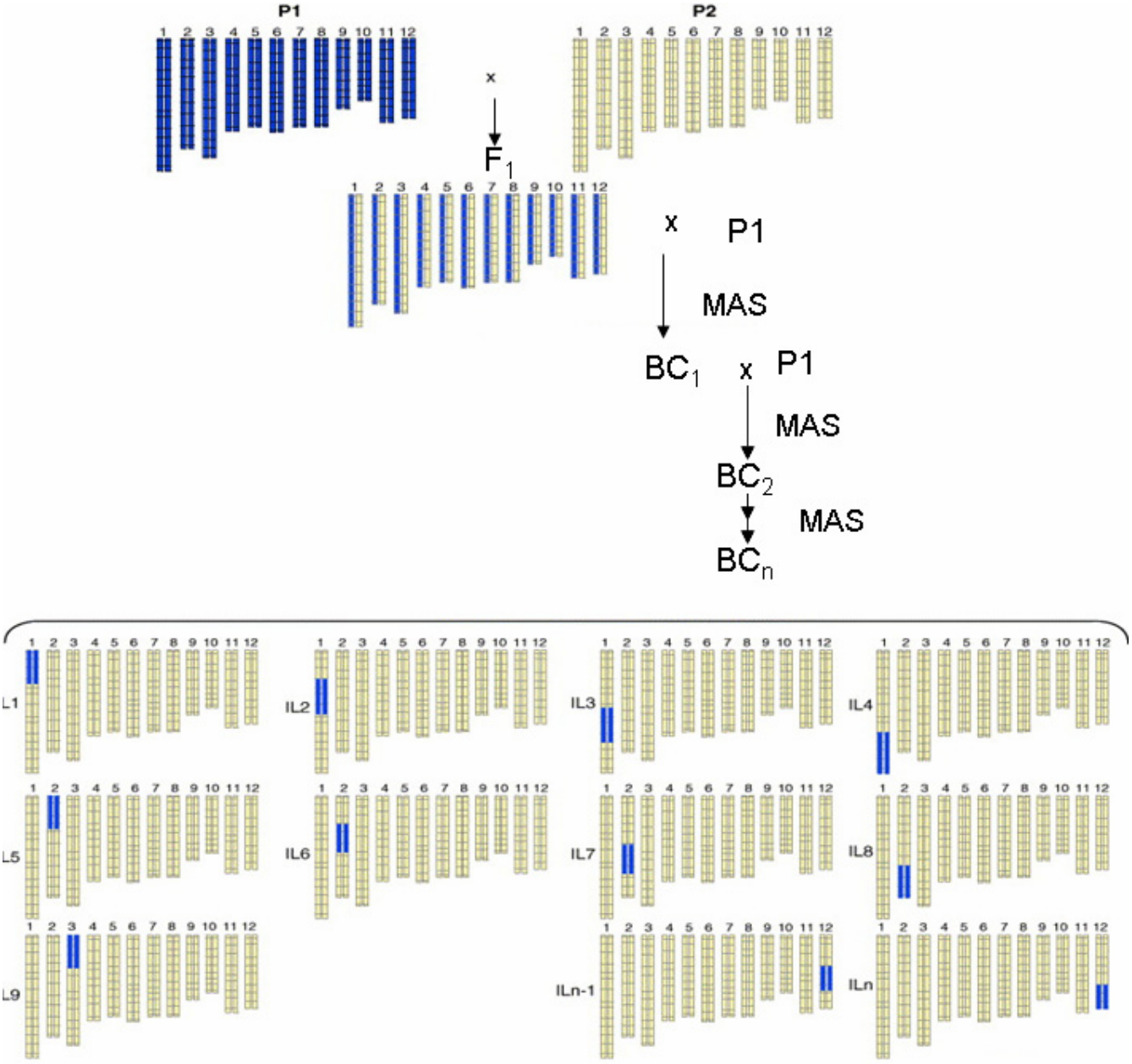

Fig. 2 Production of Introgression lines (ILs). The donor plant is repeatedly backcrossed with a recurrent parent several times. The whole genome genotype of each backcrossed line is analyzed using molecular markers to identify the remaining chromosome segments from the donor plant. The backcrossed lines are arranged so that theyare successively overlapping and covering the whole donor genome from the top of chromosome 1 to the bottom of chromosome 12 (IL1 to ILn) (Modified from Ashikari and Matsuoka 2006).

successfully implemented (Tanksley and McCouch 1997). We expect that the use of ILs for the identification and pyramiding of favourable genes contained in wild relatives into an elite parent will prove an efficient method for improving quantitative traits. The tighter the linkage the more consistent the linkage phase across populations. When the linkage between marker and target gene is complete, the markers identify the alleles directly and are called perfect markers. A class of perfect markers known as functional markers (FMs) can be developed in two ways. The first is the use of allele sequences of functionally characterized genes. This involves the identification of polymorphic, functional motifs that affect plant phenotype within these genes, and the validation of associations between DNA polymorphisms and trait variation. During the past few years, functionally characterized genes, EST and genome sequencing projects have been rapidly developed for many important crop species, which provides the genomics base for the development of FMs from the transcribed regions of the genome (Gupta and Rustgi 2004). Putative functions can be deduced for the markers derived from ESTs or genes using homology searches with protein databases. Direct proof of sequence motif function can be obtained by comparing isogenic genotypes differing in single sequence motifs.

The other approach is the use of association mapping (AM) approach. AM was originally established in human genetics as a gene identification (mapping) approach based on linkage disequilibrium (LD) (the nonrandom occurrence of allele haplotypes in the genome) (Risch 2000). If the LD declines rapidly in the causative genes as demonstrated in maize, association studies have the potential to identify sequence motifs, such as a few nucleotides or insertions/deletions that affect trait expression (Thornsberry *et al* 2001; Osterberg *et al.* 2002; Borevitz and Nordborg 2003) and as a result the identification of FMs (Anderson and Lübberstedt 2003). For species with extensive LD, of the order of several hundreds of kilobases or more, AM cannot be used directly for the development of FMs. However, it might be feasible to identify genetic regions that are associated with a particular trait of interest by scanning the entire genome with closely linked markers. This will facilitate the

development of FMs in future. Anderson and Lübberstedt (2003) suggested that association studies could be applied to select candidate sequence motifs for further testing in isogenic comparisons yielding true FMs.

The drawbacks of association mapping stem from the fact that it is not a controlled experiment. Power is unpredictable, partly because the decay of linkage disequilibrium is noisy, and partly because the genetic architecture of the trait is unknown (the latter is always a problem in mapping complex traits, but it is likely to be worse in association mapping because genetic heterogeneity is not limited by a small number of founders) (Weiss and Terwilliger 2000; Nordborg and Tavaré 2002; Zondervan and Cardon 2004). One of the primary limitations of LD-based association mapping in plant species has been the frequent occurrence of related subgroups in the sample, which results in a high probability of type I error. Pritchard *et al*. (2000a, 2000b) proposed a Bayesian approach for inferring population structure based on unlinked markers. The estimated probabilities for group membership can then be used to assign genotypes to groups within which marker–trait associations are tested. This method was extended by Thornsberry *et al*. (2001) for the analysis of quantitative traits by using the matrix of population assignments and the quantitative traits as predictors in a logistic regression model, in which the dependent variable is a binary genetic polymorphism. Approaches based on Linear Mixed Model were developed recently (Yu *et al*. 2006; Malosetti *et al*. 2007), which take genome-wide differences in relatedness into account via estimated pairwise kinship coefficients.

PROSPECTS

The design of optimal breeding schemes aimed at accumulating many genes is a complex problem. The genotyping/phenotyping cost, practically applicable segregating population size, and the duration of breeding need to be considered when crossing and selection strategy is designed. It should be realised that the most important factor is the number of genes to be pyramided because the population size necessary to fix the target genes increases exponentially with the number of target loci. In practice, the number of target genes should be kept low so that the overall cost and the length of breeding duration (from parental lines to the successful selection of the target genotype) are acceptable. When several favourable genes are originally hosted by only two or three parents, the crossing scheme can be easily determined. The target genotype is obtained by selecting the most promising genotypes (genotype with the highest number of genes homozygous for the desirable alleles and heterozygous for all other target loci) in each of the selfing generations. Therefore, pyramiding will be more straightforward if parental lines with complementary sets of homozygous loci are available. When genes are dispersed between many parents, crossing scheme needs to be chosen so that the root genotype can be obtained quickly and at low cost.

Traits controlled by major genes should be the primary target of gene pyramiding. QTL mapping studies clearly demonstrated that only QTL with large effects can be estimated and positioned accurately using an affordable size of mapping population. Moreover, it is also true that major genes are usually more stable in different genetic backgrounds and environments. Major genes for agronomically important traits are more likely to be identified in wild relatives rather than the elite gene pool used by breeders. This is because the repeated exploration of the elite gene pool by active breeding should have already fixed most of the major genes. Therefore, it seems that gene pyramiding strategy may suit the exploration of wild germplasm.

When major genes for the target trait are not available, trait improvement will be more difficult. It may be necessary to construct NILs for each of the target QTL before a pyramiding program is started. By transferring QTL into a common background, the effect of each of the QTL can be estimated more precisely and the possible false positive QTL identified. It also offers the possibility to identify markers closer to the QTL. Although it is possible that QTL without additive effect may have sustainable favourable epistatic effect, they will be difficult to identify. It is advisable that only QTL with confirmed relatively large effects are targeted for pyramiding. Step-wise pyramiding QTL one by one may be followed if the interaction between QTL is present or suspected to be present. It should be pointed out that gene pyramiding may not be the most suitable strategy when many QTL with small effects control the trait and other methods such as marker-assisted recurrent selection should be considered.

When gene pyramiding is meant to be used as a strategy to utilise identified genes, it makes sense to think of gene identification, validation and pyramiding as the components of an integrated process. Ideally, the gene identification and validation process should also lead to the creation of good parental lines for later pyramiding. This can be achieved by using introgression lines or the advanced backcrossing QTL mapping methods.

ACKNOWLEDGEMENTS

We thank three anonymous reviewers for their valuable comments and suggestions.

REFERENCES

Allard RW (1999) *Principle of Plant Breeding* (2nd Edn), John Wiley and Sons, New York, 272 pp

Andersen JR, Lübberstedt T (2003) Functional markers in plant. *Trends in Plant Science* **8**, 554-560

Ashikari M, Matsuoka M (2006) Identification, isolation and pyramiding of quantitative trait loci for rice breeding. *Trends in Plant Science* **11**, 344-350

Bailey NTJ (1961) *Introduction to the Mathematical Theory of Genetic Linkage*, The Clarendon Press, Oxford, 298 pp

Borevitz JO, Nordborg M (2003) The impact of genomics on the study of natural variation in *Arabidopsis*. *Plant Physiology* **132**, 718-725

Cregan PB, Mudge J, Fickus EW, Danesh D, Denny R, Young ND (1999) Two simple sequence repeat markers to select for soybean cyst nematode resistance conditioned by the *rhg1* locus. *Theoretical and Applied Genetics* **99**, 811-818

Dekkers JCM, Hospital F (2002) The use of molecular genetics in the improvement of agricultural populations. *Nature Review Genetics* **3**, 22-32

Dubcovsky J (2004) Marker assisted selection in public breeding programs: The wheat experience. *Crop Science* **44**, 1895-1898

Eagles HA, Bariana HS, Ogbonnaya FC, Rebetzke GJ, Hollamby GJ, Henry RJ, Henschke PR (2001) Implementation of markers in Australian wheat breeding. *Australian Journal of Agricultural Research* **52**, 1349-1356

Eshed Y, Zamir D (1994a) A genomic library of *Lycopersicon pennellii* in *L. esculentum*: a tool for mapping of genes. *Euphytica* **79**, 175-179

Eshed Y, Zamir D (1994b) Introgression from *Lycopersicon pennellii* can improve the soluble-solid yield of tomato hybrids. *Theoretical and Applied Genetics* **88**, 891-897

Eshed Y, Zamir D (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

Frisch M, Melchinger AE (2005) Selection theory for marker-assisted backcrossing. *Genetics* **170**, 909-917

Gupta PK, Rustgi S (2004) Molecular markers derived from expressed/transcribed portion of the genome in higher plants. *Functional and Integrative Genomics* **4**, 139-162

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

Ishii T, Yonezawa K (2007a) Optimization of the marker-based procedures for pyramiding genes from multiple donor lines: I. Schedule of crossing between the donor lines. *Crop Science* **47**, 537-546

Ishii T, Yonezawa K (2007b) Optimization of the marker-based procedures for pyramiding genes from multiple donor lines: II. Strategies for selecting the objective homozygous plant. *Crop Science* **47**, 1878-1886

Kuchel H, Ye G, Fox R, Jefferies S (2005) Genetic and genomic analysis of a targeted marker-assisted wheat breeding strategy. *Molecular Breeding* **16**, 67-78

Langridge P, Barr A (2003) Preface to 'better barley faster: the role of marker assisted selection'. *Australian Journal of Agricultural Research* **54**, 1-4

Malosetti M, van der Linden CG, Vosman B, van Eeuwijk FA (2007) A mixed-model approach to association mapping using pedigree information with an illustration of resistance to *Phytophthora infestans* in potato. *Genetics* **175**, 879-889

Michelmore RW, Paran I, Kesseli RV (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: A rapid method to detect markers in specific genomic regions by using segregating populations. *Proceedings of the National Academy of Sciences USA* **88**, 9828-9832

Nordborg M, Tavaré S (2002) Linkage disequilibrium: What history has to tell us. *Trends in Genetics* **18**, 83-90

Osterberg MK, Sharvorska O, Lasloux M, Lagercrantz U (2002) Naturally occurring indel variation in the *Brassica nigra COL1* gene is associated with variation in flowering time. *Genetics* **161**, 299-306

Paterson AH, Deverna JW, Lanini B, Tanksley SD (1990) Fine mapping of quantitative trait loci using selected overlapping recombinant chromosomes, in an interspecies cross of tomato. *Genetics* **124**, 735-742

Paterson AH, Tanksley SD, Sorrells ME (1991) DNA markers in plant improvement. *Advances in Agronomy* **46**, 39-90

Podlich DW, Cooper M (1998) QU-GENE: A platform for quantitative analysis of genetic models. *Bioinformatics* **14**, 632-653

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000a) Association mapping in structured populations, *American Journal of Human Genetics* **67**, 170-181

Pritchard JK, Stephens M, Donnelly P (2000b) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959

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

Risch NJ (2000) Searching for genetic determinants in the new millennium. *Nature* **405**, 847-856

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

Servin B, Martin OC, Mezard M, Hospital F (2004) Toward a theory of marker-assisted gene pyramiding. *Genetics* **168**, 513-523

Stuber CW, Polacco M, Senior ML (1999) synergy of empirical breeding, marker-assisted selection, and genomics to increase crop yield potential. *Crop Science* **39**, 1571-1583

Tanksley SD, McCouch SR (1997) Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* **277**, 1063-1068

Tanksley SD, Nelson JC (1996) Advanced backcross QTL analysis: A method for simultaneous discovery and transfer of valuable QTL from unadapted germplasm into elite breeding. *Theoretical and Applied Genetics* **92**, 191-203

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES (2001) JM *Dwarf8* polymorphisms associate with variation in flowering time. *Nature Genetics* **28**, 286-289

Varshney RK, Granera A, Sorrells ME (2005) Genomics-assisted breeding for crop improvement. *Trends in Plant Science* **10**, 621-630

Wang J, van Ginkel M, Podlich D, Ye G, Trethowan R, Pfeiffer W, DeLacy IH, Cooper M, Rajaram S (2003) Comparison of two breeding strategies by computer simulation. *Crop Science* **43**, 1764-1773

Wang, J, Eagles HA, Trethowan R, van Ginkel M (2005) Using computer simulation of the selection process and known gene information to assist in parental selection in wheat quality breeding. *Australian Journal of Agricultural Research* **56**, 465-473

Weiss KM, Terwilliger JD (2000) How many diseases does it take to map a gene with SNPs? *Nature Genetics* **26**, 151-157

Ye G, Eagles HA, Dieters MJ (2004) Parental selection using known genes for inbred line development. In 'Cereals 2004, Proceedings of 54th Australian Cereal Chemistry Conferences and 11th Wheat Breeders Assembly', pp 245-248

Ye G, Moody D, Livinus L, van Ginkel M (2007) Designing an optimal marker-based pedigree selection strategy for parent building in barley in the presence of repulsion linkage, using computer simulation. *Australian Journal of Agricultural Research* **58**, 243-251

Ye G, Smith KF (2008) Marker-assisted gene pyramiding for inbred line development: practical applications. *International Journal of Plant Breeding* **2**, 11-22

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**, 203-208

Zondervan KT, Cardon LR (2004) The complex interplay among factors that influence allelic association. *Nature Review Genetics* **5**, 89-100