Available online at www.sciencedirect.com
ScienceDirect

Agricultural Sciences in China
2007, 6(8): 101-105

August 2007

Simulation Modeling in Plant Breeding: Principles and Applications

WANG Jian-kang[1] and Wolfgang H Pfeiffer[2]

[1] *Institute of Crop Sciences, the National Key Facility for Crop Gene Resources and Genetic Improvement, and CIMMYT China Office, Chinese Academy of Agricultural Sciences, Beijing 100081. P.R.China*
[2] *HarvetPlus, c/o the International Center for Tropical Agriculture (CIAT), A. A. 6713, Cali, Colombia*

Abstract

Conventional plant breeding largely depends on phenotypic selection and breeder's experience, therefore the breeding efficiency is low and the predictions are inaccurate. Along with the fast development in molecular biology and biotechnology, a large amount of biological data is available for genetic studies of important breeding traits in plants, which in turn allows the conduction of genotypic selection in the breeding process. However, gene information has not been effectively used in crop improvement because of the lack of appropriate tools. The simulation approach can utilize the vast and diverse genetic information, predict the cross performance, and compare different selection methods. Thus, the best performing crosses and effective breeding strategies can be identified. QuLine is a computer tool capable of defining a range, from simple to complex genetic models, and simulating breeding processes for developing final advanced lines. On the basis of the results from simulation experiments, breeders can optimize their breeding methodology and greatly improve the breeding efficiency. In this article, the underlying principles of simulation modeling in crop enhancement is initially introduced, following which several applications of QuLine are summarized, by comparing the different selection strategies, the precision parental selection, using known gene information, and the design approach in breeding. Breeding simulation allows the definition of complicated genetic models consisting of multiple alleles, pleiotropy, epistasis, and genes, by environment interaction, and provides a useful tool for breeders, to efficiently use the wide spectrum of genetic data and information available.

Key words: breeding simulation, genetic model, breeding strategy, design breeding

INTRODUCTION

Phenotype of a biological individual is attributed to genotypic and environmental effects. The major breeding objective is to develop new genotypes that are genetically superior to those currently available, for a specific target population of environments (Fehr 1987; Falconer and Mackay 1996; Lynch and Walsh 1998). To achieve this objective, breeders face many complex choices in the design of efficient crossing and selection strategies aimed at combining the desired alleles into a single target genotype. For example, in the bread wheat breeding program of the International Maize and Wheat Improvement Center (CIMMYT), two major breeding strategies are commonly used and thousands of crosses are made every season. Though breeders spend great efforts in choosing parents to make the targeted crosses, approximately 50-80% of the crosses are discarded in generations F_1 to F_8, following the selection for agronomic traits (e.g., plant height, lodging tolerance, tillering, appropriate heading

This paper is translated from its Chinese version in *Scientia Agricultura Sinica.*
Correspondence WANG Jian-kang, E-mail: wangjk@caas.net.cn, jkwang@cgiar.org, w.pfeiffer@cgiar.org

date, and balanced yield components), disease resistance (e.g., stem rust, leaf rust, and stripe rust), and end-use quality (e.g., dough strength and extensibility, protein quantity and quality). Then, after two cycles of yield trials (i.e., preliminary yield trial in F_8 and replicated yield trial in F_9), only 10% of the initial crosses remain, among which 1-3% of the crosses originally made are released as cultivars from CIMMYT's international nurseries (Wang *et al*. 2003, 2005). Significant resources can therefore be saved if the potential performance of a cross, using a defined selection strategy, can be accurately predicted.

On the other hand, a great amount of studies on QTL mapping have been conducted for various traits in plants and animals in recent years (Zeng 1994; Tanksley and Nelson 1996; Frary *et al*. 2000; Barton and Keightley 2002; Li *et al*. 2003). As the number of published genes and QTLs for various traits continues to increase, the challenge for plant breeders is to determine how to best utilize this multitude of information for the improvement of crop performance. Quantitative genetics provides much of the framework for the design and analysis of selection methods used within breeding programs (Falconer and Mackay 1996; Lynch and Walsh 1998; Goldman 2000). However, there are usually associated assumptions, some of which can be easily tested or satisfied by experimentation; others can seldom, if ever, be met. Computer simulation provides us with a tool to investigate the implications of relaxing some of the assumptions and the effect this has on the conduct of a breeding program (Kempthone 1988). Breeding simulation allows the definition of complicated genetic models consisting of multiple alleles, pleiotropy, epistasis, and genes by environment interaction, and provides a useful tool to breeders, who can efficiently use the wide spectrum of genetic data and information available. This approach will be very helpful when the breeders want to compare breeding efficiencies from different selection strategies, to predict the cross performance with known gene information, and to investigate the efficient use of identified QTLs in conventional breeding, and so on.

In this article, the principles of simulation modeling in plant breeding are introduced initially, and then several applications using the simulation tool of QuLine are summarized.

PRINCIPLES OF SIMULATION MODELING IN PLANT BREEDING

The genetics and breeding simulation module of QuLine

QU-GENE is a simulation platform for quantitative analysis of genetic models, which consists of a two-stage architecture (Podlich and Cooper 1998). The first stage is the engine, and its role is to: (1) define the genotype by environment (GE) system (i.e., all the genetic and environmental information of the simulation experiment), and (2) generate the starting population of individuals (base germplasm) (Fig.1). The second stage encompasses the application modules, whose role is to investigate, analyze, or manipulate the starting population of individuals within the GE system defined by the engine. The application module usually represents the operation of a breeding program. A QU-GENE strategic application module, QuLine, has therefore been developed to simulate the breeding procedure deriving inbred lines (Fig.1).

Built on QU-GENE, QuLine (previously called QuCim) is a genetics and breeding simulation tool, which can integrate various genes with multiple alleles operating within epistatic networks and differentially interacting with the environment, and predict the outcome from a specific cross following the application of a real selection scheme (Wang *et al*. 2003; Wang *et al*. 2004). It therefore has the potential to provide a bridge between the vast amount of biological data and the breeder's queries on optimizing selection gain and efficiency. QuLine has been used to compare two selection strategies (Wang *et al*. 2003), to study the effects on selection of dominance and epistasis (Wang *et al*. 2004), to predict cross performance using known gene information (Wang *et al*. 2005), and to optimize marker-assisted selection to efficient pyramid multiple genes (Kuchel *et al*. 2005; Wang *et al*. 2007).

Genetic models used in simulation

The simulation principles are illustrated by using CIMMYT's wheat breeding program as an example. Two breeding strategies are commonly used in CIMMYT's wheat breeding programs. The MODPED

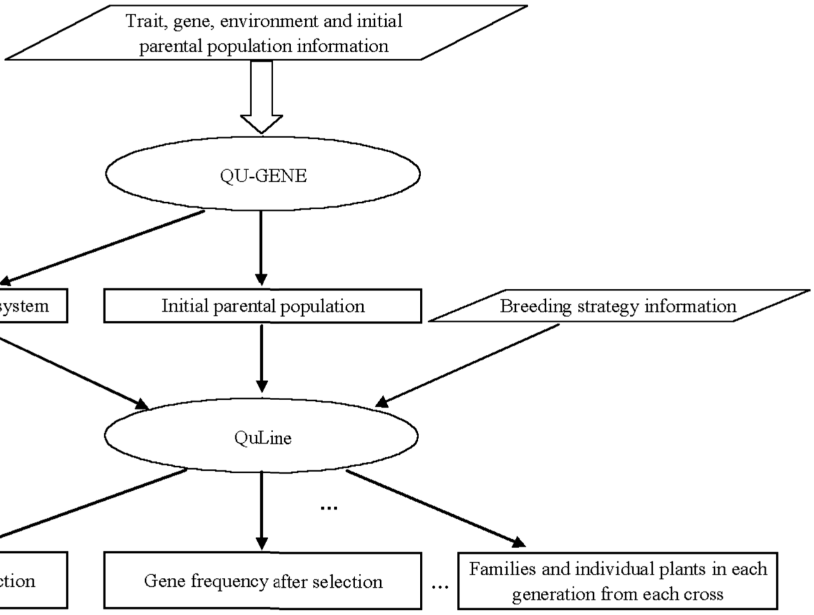
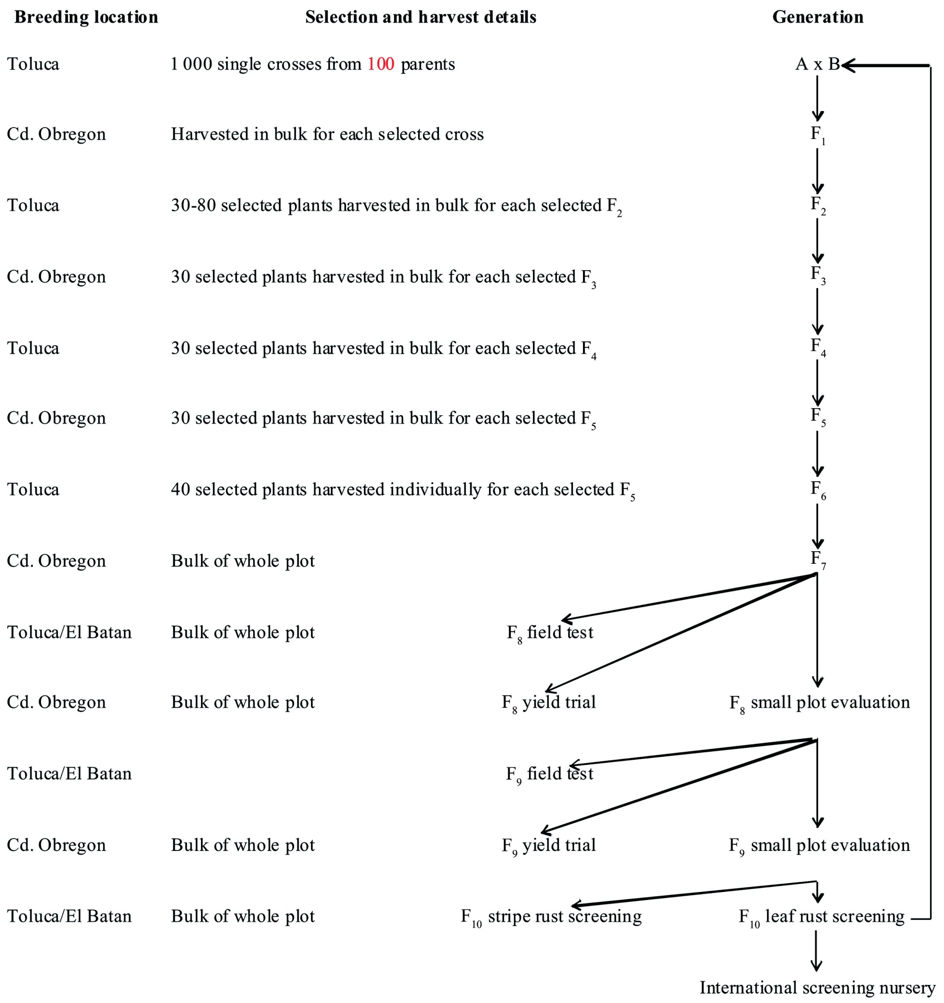

Trait, gene, environment and initial parental population information
QU-GENE
Trait, gene and environment system
Initial parental population
Breeding strategy information
QuLine
...
Genotypic values after selection
Gene frequency after selection
...
Families and individual plants in each generation from each cross

Fig. 1 Flowchart of the breeding simulation tool QuLine. The two ellipses represent the two computer programs, i.e., QU-GENE and QuLine; the parallelograms represent inputs for QU-GENE and QuLine; and the rectangles represent outputs from QU-GENE and QuLine.

(modified pedigree) method begins with pedigree selection of individual plants in the F_2, followed by three bulk selections from F_3 to F_5, and pedigree selection in the F_6; hence the name modified pedigree/bulk. In the SELBLK (selected bulk) method, spikes of selected F_2 plants within one cross are harvested in bulk and threshed together, resulting in one F_3 seed lot per cross. This selected bulk selection is also used from F_3 to F_5, whereas, pedigree selection is used only in the F_6. A major advantage of SELBLK compared to MODPED is that fewer seed lots need to be harvested, threshed, and visually selected for seed appearance, leading to significant saving of time, labor, and costs associated with nursery preparation, planting, and plot labeling ensue (van Ginkel *et al*. 2002). The flowchart of SELBLK is shown in Fig.2.

Seven agronomic traits and three rust resistances are the major traits used in selection in CIMMYT's wheat breeding programs. The gene number and genetic values are derived from discussions with breeders and from analyses of past unpublished experiments. In total it is postulated that 59 independently segregating genes control these traits (Table 1). The genetic effects of traits other than yield are considered fixed. Pleiotropic effects are included to account for trait correlations, and they are also considered fixed. Two kinds of pleiotropic effects are included, although more complicated pleiotropic interaction can also be defined within the QU-GENE engine. The first kind is positive pleiotropy, such as, the pleiotropic effects on lodging from genes for grains per spike. The second kind is the negative pleiotropy, such as, the pleiotropic effects on kernel weight from genes for grains per spike. As shown in Table 1, at Cd. Obregon the three lodging genes, the stem rust genes, and the leaf rust genes have some degree of negative effect on the yield, and the five kernel weight genes have a positive pleiotropic effect. Stem rust, leaf rust, heading, tillering, and grains-per-spike genes have a negative pleiotropic effect on kernel weight (Table 1). Stripe rust rarely occurs at Cd. Obregon, hence, there is no selection for stripe rust when the nursery is grown there and the genetic effects of stripe rust genes are considered to be zero in this environment (Table 1).

Apart from the pleiotropic effects of genes affecting other traits, it is postulated that there are 20 genes yield *per se*(italic is necessary?), even though their very existence has been debated. Four gene effect models were considered for yield, those are, pure additive [AD0, *Aa* = (*AA* + *aa*)/2, where *A* and *a* represent the two alleles at each locus affecting the yield], partial dominance [AD1, *Aa* ¹ (*AA*+*aa*)/2, but is between *AA*



Breeding location	Selection and harvest details	Generation
Toluca	1 000 single crosses from 100 parents	A x B
Cd. Obregon	Harvested in bulk for each selected cross	F_1
Toluca	30-80 selected plants harvested in bulk for each selected F_2	F_2
Cd. Obregon	30 selected plants harvested in bulk for each selected F_3	F_3
Toluca	30 selected plants harvested in bulk for each selected F_4	F_4
Cd. Obregon	30 selected plants harvested in bulk for each selected F_5	F_5
Toluca	40 selected plants harvested individually for each selected F_5	F_6
Cd. Obregon	Bulk of whole plot	F_7
Toluca/El Batan	Bulk of whole plot	F_8 field test
Cd. Obregon	Bulk of whole plot	F_8 yield trial F_8 small plot evaluation
Toluca/El Batan		F_9 field test
Cd. Obregon	Bulk of whole plot	F_9 yield trial F_9 small plot evaluation
Toluca/El Batan	Bulk of whole plot	F_{10} stripe rust screening F_{10} leaf rust screening
		International screening nursery

Fig. 2 Germplasm flow in CIMMYT's Wheat Breeding Program. The breeding strategy described was called selected bulk selection method.

and *aa*], a combination of partial, complete, and overdominance (AD2, the genetic values of *AA*, *Aa* and *aa* are independent), and digenic interaction (ADE) (Wang *et al*. 2004).

Definition of breeding strategies in QuLine

By defining breeding strategy, QuLine translates the complicated breeding process in a way that the computer can understand and simulate. QuLine allows for several breeding strategies, which were contained in one input file, to be defined simultaneously. The program then makes the same virtual crosses for all the defined strategies at the first breeding cycle. Hence, all strategies start from the same point (the same initial population, the same crosses and the same genotype and environment system), allowing appropriate comparison. A breeding strategy in QuLine is defined as all the crossing, seed propagation, and selection activities in an entire breeding cycle. A breeding cycle begins with crossing and ends at the generation when the selected advanced lines are returned to the crossing block, as new parents. SELBLK (Fig.2) is defined in Tables 2 and 3.

Number of generations in MODPED and number of selection rounds in each generation

In the breeding program in Fig.2, the best advanced lines developed from the F_{10} generation will be returned to the crossing block to be used for new crosses; that is to say a new breeding cycle starts after the F_{10} leaf rust screening at El Batan. Therefore, the number of generations in one breeding cycle is 10 for SELBLK

(Fig.2 and Table 2). The crossing block (viewed as F_0) and the 10 generations need to be defined in SELBLK. The parameters to define a generation consist of the number of selection rounds in the generation, an indicator for seed source (explained later), and the planting and selection details for each selection round (Table 2). Most generations in this breeding program have just one selection round, for example, F_1 to F_6, whereas, some generations have more than one selection round as they are grown simultaneously at different sites or under different conditions, for example, F_7, F_8, and F_9 (see

Table 1 Number of segregating genes and their genetic effects in the Cd. Obregon environment type[1)]

Gene classification	Number of genes	Traits affected	Individual gene effects		
			AA	Aa	aa
Yield	20	Yield (t/ha)	Four genetic models for yield: AD0 (pure additive), AD1(partial dominance), AD2 (overdominance), ADE (digenic epistasis)		
Lodging	3	Lodging (%)	0.00	5.00	10.00
		Yield (t ha^{-1})	0.00	-0.40	-0.80
Stem rust	5	Stem rust (%)	0.00	0.50	1.00
		Yield (t ha^{-1})	0.00	-0.25	-0.50
		Kernel weight (g)	0.00	-0.75	-1.50
Leaf rust	5	Leaf rust (%)	0.00	5.00	10.00
		Yield (t ha^{-1})	0.00	-0.25	-0.50
		Kernel weight (g)	0.00	-0.75	-1.50
Stripe rust	5	Stripe rust	0.00	0.00	0.00
Height	3	Height (cm)	40.00	30.00	20.00
		Lodging (%)	5.00	2.50	0.00
Maturity	5	Maturity (day)	20.00	16.00	12.00
		Kernel weight (g)	-1.00	-0.50	0.00
Tillering	3	Tillering (no.)	5.00	3.00	1.00
		Lodging	2.00	1.00	0.00
		Maturity (day)	1.00	0.50	0.00
		Grains per ear	-1.00	-0.50	0.00
		Kernel weight (g)	-1.50	-0.75	0.00
Grains per ear	5	Gains per ear	14.00	10.00	6.00
		Lodging (%)	2.00	1.00	0.00
		Kernel weight (g)	-1.00	-0.50	0.00
Kernel weight	5	Kernel weight (g)	12.00	8.50	5.00
		Yield (t ha^{-1})	1.00	0.50	0.00
		Lodging (%)	2.00	1.00	0.00

[1)] There is no stripe rust in the Cd. Obregon environment type, so the effects of the 5 genes for stripe rust were set at 0. However, these genes have effects in the other two environment types.

Table 2 Definition of the selected bulk method for developing inbred lines in QuLine

Number of selection rounds	Seed source	Generation title[1)]	Seed propagation type	Generation advance method	Number of replications	Individual plants in a plot	Number of test locations	Environment type
1		F_0	*self*	*bulk*	1	20	1	Toluca
1		F_1	*singlecross*	*bulk*	1	20	1	Cd. Obregon
1		F_2	*self*	*bulk*	1	1000	1	Toluca
1		F_3	*self*	*bulk*	1	500	1	Cd. Obregon
1		F_4	*self*	*bulk*	1	625	1	Toluca
1		F_5	*self*	*bulk*	1	625	1	Cd. Obregon
1		F_6	*self*	*pedigree*	1	750	1	Toluca
4	0	F_7	*self*	*bulk*	1	70	1	Cd. Obregon
		F_8(T)	*self*	*bulk*	1	70	1	Toluca
		F_8(B)	*self*	*bulk*	1	70	1	El Batan
		F_8(YT)	*self*	*bulk*	1	100	1	Cd. Obregon
4	0	F_8(SP)	*self*	*bulk*	1	30	1	Cd. Obregon
		F_9(T)	*self*	*bulk*	1	70	1	Toluca
		F_9(B)	*self*	*bulk*	1	70	1	El Batan
		F_9(YT)	*self*	*bulk*	2	100	1	Cd. Obregon
1		F_9(SP)	*self*	*bulk*	1	30	1	Cd. Obregon
2	0	F_{10}(LR)	*self*	*bulk*	1	30	1	El Batan
		F_{10}(YR)	*self*	*bulk*	1	30	1	Toluca

[1)] T, the breeding location of Toluca; B, the breeding location of El Batan; YT: yield trial; SP: small plot evaluation; LR, leaf rust; YR, stripe rust.

Table 3 Traits and their selected proportions in each generation in the selected bulk method

Generation	Selection	Yield	Lodging	Stem rust	Leaf rust	Stripe rust	Height	Maturity	Tillering	Grains per ear	Kernel weight	Total selected proportion
	mode	Top	Bottom	Bottom	Bottom	Bottom	Middle	Middle	Top	Top	Top	
F_1	Among-family		0.98	0.99	0.85		0.99	0.98	0.90	0.97		0.70
F_2	Among-family		0.99	0.99		0.90	0.99	0.99	0.99	0.99		0.85
	Within-family		0.95	0.99		0.40	0.85	0.90	0.60	0.50		0.08
F_3	Among-family		0.99		0.90				0.95			0.85
	Within-family		0.90		0.70		0.90	0.90	0.80	0.25	0.60	0.06
F_4	Among-family		0.99			0.96			0.95			0.90
	Within-family		0.90			0.65	0.95	0.90	0.80	0.20	0.60	0.05
F_5	Among-family		0.99		0.6				0.95			0.90
	Within-family		0.90		0.70		0.90	0.90	0.80	0.20	0.60	0.05
F_6	Among-family		0.99			0.96			0.95			0.90
	Within-family		0.90			0.70	0.90	0.98	0.95	0.10		0.05
F_7	Among-family		0.85		0.70		0.98	0.96	0.85	0.70	0.75	0.25
F_8(T)	Among-family		0.55			0.70	0.99	0.99	0.98	0.90		0.55
F_8(B)	Among-family				0.90							0.90
F_8(YT)	Among-family	0.40										0.40
F_8(SP)	Among-family											1.00
F_9(T)	Among-family		0.97			0.95			0.99	0.99		0.90
F_9(B)	Among-family				0.95							0.95
F_9(YT)	Among-family	0.40										0.40
F_9(SP)	Among-family											1.00
F_{10}(YR)	Among-family					0.98						0.98
F_{10}(LR)	Among-family				0.98							0.98

the first column in Table 2).

Seed propagation type for each selection round

The seed propagation type describes how the selected plants in a retained family, from the previous selection round or generation, are propagated, to generate the seed for the current selection round or generation. There are nine options for seed propagation, presented here in the order of increasing genetic diversity (F_1 excluded): (i) *clone* (asexual reproduction), (ii) *DH* (doubled haploid), (iii) *self* (self-pollination), (iv) *singlecross* (single crosses between two parents), (v) *backcross* (back crossed to one of the two parents), (vi) *topcross* (crossed to a third parent, also known as three-way cross), (vii) *doublecross* (crossed between two F_1s), (viii) *random* (random mating among the selected plants in a family), and (ix) *noself* (random mating but self-pollination is eliminated). The seed for F_1 is derived from crossing among the parents in the initial population (or crossing block). QuLine randomly determines the female and the male parents for each cross from a defined initial population, or alternately, one may select some preferred parents from the crossing block. The selection criteria used to identify such preferred parents (grouped here as the male and female master lists) can be defined in terms of among-family and within-family selection descriptors (see below for details) within the crossing block (referred to as F_0 generation). By using the parameter of seed propagation type, most, if not all methods of seed propagation in self-pollinated crops can be simulated in QuLine.

Two seed propagation types were used in SELBLK, which were, *singlecross* (only used for F_1 generation) and *self* (Table 2).

Generation advance method for each selection round

The generation advance method describes how the selected plants within a family are harvested. There are two options for this parameter: *pedigree* (the selected plants within a family are harvested individually, therefore each selected plant will result in a distinct family in the next generation), and *bulk* (the selected plants in a family are harvested in bulk, resulting in just one family in the next generation). This parameter and the seed propagation type allow QuLine to simulate not only the traditional breeding methods, such as, pedigree breeding and bulk population breeding, but also many combinations of different breeding methods (e.g., pedigree selection until the F_4 and then doubled haploid production on selected F_4 plants). The *bulk* generation advance method will not change the number of families in the following generation if no among-family selection is applied in the current generation, whereas, the *pedigree* method increases the number of families rapidly if among-family selection intensity is weak, and several plants are selected within each retained family. For a generation with more than one selection round, the generation advance method for the first selection round can be either *pedigree* or *bulk*. The subsequent selection rounds are used to determine which families derived from the first selection round will advance to the next generation. In the majority of cases, *bulk* generation advance is the preferred option for the subsequent selection rounds.

It can be seen from Table 2 that *pedigree* is only used in F_6 and *bulk* is used in the other generations in SELBLK.

Field experimental design for each selection round

The parameters used to define the virtual field experimental design in each selection round include the number of replications for each family, the number of individual plants in each replication, the number of test locations, and the environment type for each test location (Table 2). Each environment type defined in the genotype and environment system has its own gene action and gene interaction, which provides the framework for defining the genotype by environment interaction. Therefore, by defining the target population of environments as a mixture of environment types, genotype by environment interactions are defined as a component of the genetic architecture of a trait.

It can be seen from Table 2, for example, that F_7 is grown in the Cd.Obregon environment, F_8(T) in Toluca, F_8(B) in El Batan, and F_8(YT) in Cd. Obregon.

Among-family selection and within-family selection for each selection round

Ten traits have been included as relevant (Table 1) for the selection process in the breeding program described in Fig.2. Among-family selection and within-family selection are distinct processes in a breeding strategy. However, the definition of these two types of selections is essentially the same: the number of traits to be selected is followed by the definition of each trait (Table 3; Wang *et al*. 2004).

Apart from the trait code there are two parameters that define a trait used in the selection: selected proportion and selection mode. Among-family selection, the selected proportion is the percentage of families to be retained, and within-family selection it is the percentage of individual plants to be selected in each retained family. There are four options for the trait selection mode: (i) *top* (the individuals or families with highest phenotypic values for the trait of interest will be selected, for example, yield, tillering, grains per spike, and kernel weight), (ii) *bottom* (the individuals or families with the lowest phenotypic values will be selected, for example, lodging, stem rust, leaf rust, and stripe rust), (iii) *middle* (individuals or families with medium trait phenotypic values will be selected, for example, height and heading), and (iv) *random* (individuals or families will be randomly selected). Independent culling is used if multiple traits are considered for among-family or within-family selection. If there is no among-family or within-family selection for a specific selection round, the number of selected traits is noted as 0. The traits for both among-family and within-family selections can be the same or different, as is the case for selected proportions (Table 3). The traits for selection may also differ from generation to generation, as may the selected proportions for traits.

Taking F_6 as an example, three traits are used for among-family selection, and they are, the 2 (lodging), 5 (leaf rust), and 8 (tillering) traits. Six traits are used for within-family selection, and they are the 2 (lodging), 5 (leaf rust), 6 (height), 7 (heading), 8 (tillering), and 9 (grains per spike) traits. The selected proportions of these traits can be seen from Table 3.

It should be noted that some new functionalities have just been added to QuLine to select families or individuals with trait values above or below some preassigned values, or to select a predefined number of families or individuals.

Phenotypic value of a genotype and family mean of a family

For the purpose of simulation, the genotypic value of a genotype can be calculated from the definition of gene actions. However, breeders select on the basis of phenotypic value. Therefore, the phenotypic value of a genotype in a specific environment needs to be defined from its genotypic value and some associated environmental errors. For example, if there are *n* plots (or replications) for a family and the plot size is *m*, there will be $n \times m$ individual plants (or genotypes) for this family. The genotypic value g_{ij} $i = 1, \ldots, n$; $j = 1, \ldots m$ can be determined from the defined genetic models, and the phenotypic value p_{ij} can then be calculated from the formula $p_{ij} = g_{ij} + e_{bi} + e_{wij}$, where e_{bi} is the between-plot error for plot *i*, e_{wij} is the within-plot error for the genotype *j* in the plot *i*, and both e_{wij} and e_{bi} are assumed to be normally distributed. The variance (σ_e^2) of e_{wij} is calculated from the definition of heritability in the broad sense $h_b^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$, where the genetic variance (σ_g^2) is calculated from the genotypic values of individuals in the reference population. Once the error variance is determined, it will be used for all generations without change. The genetic variance changes from generation to generation, therefore, heritability may be different in different generations.

APPLICATIONS OF THE BREEDING SIMULATION MODULE QuLine

Comparison of modified pedigree (MODPED) and selected bulk (SELBLK)

Some small-scale field experiments were conducted comparing the efficiencies of MODPED and SELBLK (Singh *et al*. 1998), however, the efficiency of SELBLK compared with that of MODPED remains untested on a larger scale. The genetic models developed accounted for epistasis, pleiotropy, and genotype by environment (GE) interaction (Table 1). For both breeding strategies, the simulation experiment comprised of the same 1 000 crosses developed from 200 parents. A total of 258 advanced lines remained following 10 generations of selection. The two strategies were each applied 500 times on 12 GE systems.

The average adjusted genetic gain on yield across all genetic models was 5.83 for MODPED and 6.02 for SELBLK, a difference of 3.3% (Fig.3-A). This difference is not large and therefore unlikely to be detected using field experiments (Singh *et al*. 1998). However, it can be detected through simulation, which indicates that the high level of replication (50 models by 10 runs in this experiment) is feasible with simulation and can better account for the stochastic properties from a run of a breeding strategy, and from the sources of experimental errors. The average adjusted gains for the two yield gene numbers 20 and 40 were 6.83 and 5. 02, respectively, suggesting that genetic gain decreases with increasing yield gene number.

The number of crosses remaining after one breeding cycle was significantly different among models and strategies, but not among runs. The number of crosses remaining from SELBLK was always higher than that from MODPED, which means that delaying pedigree selection favors diversity.

On an average, 30 more crosses were maintained in SELBLK (Fig.3-B). However, there was a crossover between the two breeding strategies (Fig.3-B). Prior to F_5 the number of crosses in MODPED was higher than that in SELBLK. The number of crosses became smaller in MODPED after $F_{5,}$ when pedigree selection was applied in F_6. Among-family selection from F_1 to F_5 in SELBLK was equal to among-cross selection, and resulted in a greater reduction in the cross numbers for SELBLK compared to MODPED, in the early generations. In general, only a small proportion of crosses remained at the end of a breeding cycle (11.8% for MODPED and 14.8% for SELBLK); therefore, intense among-cross selection in early generations was unlikely to reduce the genetic gain. On the contrary, breeders would tend to concentrate on fewer but “higher probability” crosses. The fact that just a few crosses of the many generated remained after the final yield trial stage, was common in most breeding programs. As more crosses remained in SELBLK, the population following selection from SELBLK might have a larger genetic diversity than that from MODPED. In this context also, SELBLK is superior to MODPED.

As the number of families and selection methods after F_8 were basically the same for both MODPED and SELBLK, only the resources allocated from F_1 to F_8 were compared. The total number of individual plants from F_1 to F_8 was calculated to be 5,155,090 for MODPED and 3,358,255 for SELBLK (Fig.3-C).

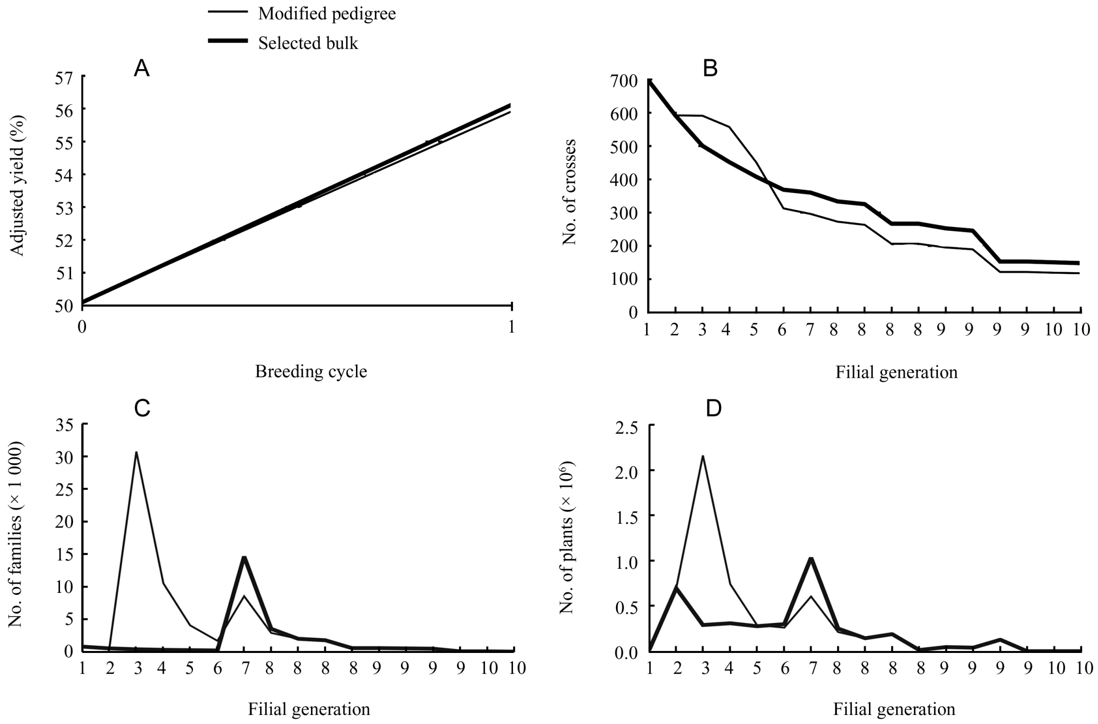


Fig. 3 Comparision of modified pedigree and selected bulk from the simulation experiment. A, adjusted genetic gain after one breeding cycle across all experimental sets; B, number of crosses after each generation's selection across all experimental sets; C, number of families in each generation in one breeding cycle; D, number of individual plants in each generation in one breeding cycle.

Assuming that planting intensity is similar, SELBLK will use approximately two thirds of the land allocated to MODPED. Furthermore, SELBLK produced smaller number of families compared to MODPED. From F_1 to F_8, there were 63,188 families for MODPED, but only 24,260 for SELBLK, approximately 40% of the number for MODPED (Fig.3-D). Therefore when SELBLK is used, fewer seed lots need to be handled at both harvest and sowing, resulting in a significant saving in time, labor, and cost.

Parental selection using known gene information

Selecting parents to make crosses is the first and essential step in plant breeding (Fehr 1987). Because of incomplete gene information (that is, only some resistance genes and their effects on phenotype are known, whereas, some are not. Most genes for agronomic traits are unknown), many seemingly good crosses are discarded during the segregating phase of a breeding program. Generally speaking, the cross with the highest progeny mean and largest genetic variance has the most potential to produce the best lines (Bernardo 2002). Under an additive genetic model, the midparent value is a good predictor of the progeny mean, but the variance cannot be deduced from the performance of the parents alone. The best way to estimate the progeny variance is to generate and test the progeny. Breeders normally use one of two types of parental selection: one based on parental information, such as, parental performance or the genetic diversity among parents; the other based on parental and progeny information. In the first case, previous studies found that both high × high and high × low crosses have the potential to produce the best lines, and the correlation between the genetic distance of parents and their progeny performance is not high. In the second case, the progeny needs to be grown and tested, which precludes parental selection. Because of complicated intra-genic, inter-genic, and gene-by-environment interactions, no method has given a precise prediction of cross performance (Wang *et al.* 2005).

Cross performance can be accurately predicted when information about the genes controlling the traits of interest is known. If progeny arrays after selection in a breeding program could be predicted, then the efficiency of plant breeding would be greatly increased. For the majority of economically important traits in wheat breeding, the genes controlling their expression remain unknown. However, for wheat quality this information is known, though incompletely, for certain aspects of wheat quality (Eagles *et al*. 2002, 2004). How cross performance, following selection, can be predicted in wheat quality breeding by using QuLine, under the condition that all the gene information of key selection traits is known is demonstrated here.

The eight Silverstar wheat sister lines are morphologically very similar, but have different values for two important quality traits, Rmax and extensibility. Supposing it is intended to use Silverstar in crosses with other adapted wheat cultivars, such as, Westonia, Krichauff, Machete, and Diamondbird, without losing grain quality, which sister line should one use? Relevant single crosses were made by QuLine between the four selected parents and the eight Silverstar sister lines. For each cross, 1 000 F_8 lines were developed from 1 000 F_2 individual plants by single seed descent. Forty F_8 lines were finally selected, based on line performance for Rmax and/or extensibility, resulting in a selected proportion of 0.04. Four selection schemes were considered: (1) the 40 lines were selected based only on line performance for Rmax (R0.04); (2) 200 lines were first selected based on line performance for Rmax and subsequently 40 lines were selected based on extensibility (R0.2E0.2); (3) 200 lines were first selected based on line performance for extensibility and then the 40 lines were selected based on Rmax (E0.2R0.2); (4) 40 lines were selected based only on line performance for extensibility (E0.04).

When using crosses with Westonia, Silverstar 3 and 7 show the largest improvement in Rmax, when Rmax is used in selection (i.e., R0.04, R0.2E0.2, and E0.2R0.2) (Table 4). They can also improve extensibility in combination with Westonia, particularly when selecting for extensibility (i.e., R0.2E0.2 and E0.2R0.2). When high Rmax and extensibility together are the required quality traits, but Rmax is more important, they are both parents of choice; however, Silverstar 3 is the better of the two (Table 4).

For crosses with Krichauff, if selection is solely for Rmax, or if it is selected first when both traits are targeted for selection (i.e., R0.04 and R0.2E0.2), Silverstar 1, 3, 5, and 7 can result in similar improvements in Rmax and extensibility. In crosses with Krichauff, if selection is solely for extensibility, or if extensibility is selected first, when both traits are targeted for selection (i.e., E0.2R0.2 and E0.04), then Silverstar 3 and 7 are the best parents for improving both traits (Table 4).

For crosses with Machete, Silverstar 3, 4, 7, and 8 are the best parents to improve Rmax if it is the only trait selected, or if it is selected first when both traits are targeted for selection (i.e., R0.04 and R0.2E0.2). However, to improve extensibility simultaneously, Rmax should be selected first and then extensibility (i.e., R0.2E0.2). If extensibility is selected before Rmax, then Silverstar 4 and 8 should be chosen to improve both traits in crosses with Machete (Table 4).

For crosses with Diamondbird, the use of Silverstar 1, 2, 3, and 4 can cause a slight increase in Rmax and extensibility, if Rmax is the trait targeted for selection (i.e., R0.04 and R0.2E0.2). If extensibility is targeted for selection (i.e., E0.2R0.2 and E0.04), then only Silverstar 3 and 4 can improve both traits slightly. Clearly, parental selection depends on the breeding objective and definition of the selection scheme. In

Table 4 The best Silverstar sister lines for the four selected parents, under different breeding objectives

Parent to be improved	Breeding objective	Selection scheme [1]			
		R0.04	R0.2E0.2	E0.2R0.2	E0.04
Westonia	High Rmax (BU)	3, 7	3, 7	3, 7	1, 3
	High extensibility (cm)	1	1, 5	1, 3, 5	1, 3, 5, 7
Krichauff	High Rmax (BU)	1, 3, 5, 7	1, 3, 5, 7	3, 7	3, 7
	High extensibility (cm)	1, 3, 5, 7	1, 3, 5, 7	1, 5	1, 5
Machete	High Rmax (BU)	3, 4, 7, 8	3, 4, 7, 8	4, 8	None
	High extensibility (cm)	1, 2, 5, 6	1, 2, 5, 6	1, 2, 3	1, 2, 3, 4
Diamondbird	High Rmax (BU)	1, 2, 3, 4	1, 3, 4	3, 4	3, 4
	High extensibility (cm)	None	None	1, 2, 5, 6	1, 2, 5, 6

[1] R, Rmax; E, extensibility; trait followed by selected proportion.

most instances, the lines that can improve Rmax are not the best lines for improving extensibility (Table 4).

Design breeding using identified QTL-marker associations

The concept of design breeding was proposed in recent years as the fast development in molecular marker technology (Bernardo 2002; Peleman and Voort 2003; Wan 2006). Three steps are involved in design breeding. The first step is to identify the genes for breeding traits, the second step is to evaluate the allelic variation in parental lines, and the third step is to design and conduct breeding. Genotypic selection is used in design breeding based on identified gene-marker associations. Here QuLine is used to demonstrate the design breeding in improving rice grain quality.

Rice quality is a complex character consisting of many components, such as, milling, appearance, nutritional, cooking, and eating qualities. For the improvement of appearance, milling, and eating qualities, the endosperm of high-quality rice varieties should be free of chalkiness (low or zero area of chalky endosperm or ACE), as chalky grains have a lower density of starch granules compared to the vitreous ones, and are therefore more prone to breakage during milling. Meanwhile, it has been well known that amylose content (AC) is the most important factor affecting rice eating quality. Therefore, low ACE and high AC are generally favored in rice quality breeding. Some QTL for ACE and AC have been identified using 65 chromosome segment substitution (CSS) lines (Table 5). These CSS lines were generated from a cross between the japonica rice variety Asominori (the background parent, denoted as P_1) and the indica rice variety IR24 (the donor parent, denoted as P_2) (Wan *et al*. 2005, 2006).

Table 5 shows the significant markers (representing chromosome segments) for ACE and AC through a likelihood ratio test based on stepwise regression (Wang *et al*. 2006). It is impossible to derive an inbred with the minimum of ACE and the maximum of AC, as QTL on segments M35, M57, and M59 have unfavorable pleiotropic effects on ACE and AC. But the ideal inbred with relatively low ACE and high AC can be identified through simulation. This designed inbred contains four segments from IR24, which are, M19, M35, M57, and M60, and another genome is from the background parent Asominori (Table 6). The value of ACE in this inbred is 9.2%, where the theoretical minimum ACE is 0. The value of AC is 17.73%, whereas, the theoretical maximum of AC is 22.3%. Among the 65 CSS lines, the three lines, CSSL15, CSSL29, and CSSL49, have the required target segments, therefore, can be used as the parental lines in breeding (Table 6).

Three possible topcrosses can be made among the

Table 5 QTL mapping results of ACE and AC in the population consisting of 65 CSS lines

	QTL for ACE								
Marker	M19*	M35**	M38*	M39*	M43*	M57**	M59**		
LOD score	0.94	2.16	1.19	1.54	1.23	16.86	10.02		
Additive effect (%)	-1.80	-1.63	1.20	-1.31	-0.88	5.93	4.96		
Percentage of variance explained (%)	1.10	2.66	1.43	1.70	1.47	35.00	16.56		
	QTL for AC								
Marker	M6*	M14**	M21*	M35*	M38*	M57**	M59**	M60**	M63**
LOD score	1.07	2.60	1.40	0.92	1.37	7.24	4.66	4.34	1.48
Additive effect (%)	0.47	-0.61	-0.35	-0.36	-0.43	1.12	1.03	0.71	0.45
Percentage of variance explained (%)	1.89	4.83	2.48	1.62	2.41	15.97	9.28	8.59	2.59

* Significance level 0.05; ** significance level 0.01.

Table 6 Marker types and predicted genetic values on AC and ACE of a designed genotype and three CSS lines

Chromosome	3	3	5	8	9	Predicted value	
Marker	M19	M21	M35	M57	M60	ACE (%)	AC (%)
Designed genotype	2	1	2	2	2	9.27	17.73
CSSL15	2	2	1	1	1	0.55	14.09
CSSL29	1	1	2	1	1	0.88	14.07
CSSL49	1	1	1	2	2	16.13	18.44

1 and 2 represent the chromosome segment from background parent Asominori and donor parent IR24, respectively.

three parental lines, Topcross 1: (CSSL15 × CSSL29) × CSSL49, Topcross 2: (CSSL15 × CSSL49) × CSSL29, and Topcross 3: (CSSL29 × CSSL49) × CSSL15. Different marker assisted selection (MAS) schemes can be used to select the target inbred line. Here two schemes are considered. Scheme 1:200 topcross F_1 (TCF_1) were first generated. Then 20 doubled haploid (DH) were derived from each TCF_1 individual. The target inbred lines were selected from the 4000 DH lines. Scheme 2: 200 topcross F_1 (TCF_1) were first generated. An enhancement selection (Wang *et al*. 2007), was conducted among the 200 TCF_1 individuals. Then 20 doubled haploid (DH) were derived from each selected TCF_1 individual. The target inbred lines are selected from those derived DH lines. QuLine was used to implement the above selection procedure.

From 100 simulation runs, it was found that by using Scheme 1, 27 target inbred lines were selected from Topcross 1, 13 from Topcross 2, and 8 from Topcross 3 (Table 7). Therefore Topcross 1 had the largest probability to select the target inbred line, and should be used in breeding low ACE and AC inbred lines. The two MAS schemes resulted in significant difference in cost when genotyping for MAS. Scheme 1 required 4000 DNA samples for each topcross. On the contrary, Scheme 2 required 462 DNA samples for Topcross 1, 324 for Topcross 2, and 691 for Topcross 3. Topcross 1 combined with Scheme 2 resulted in the least DNA samples per selected line (Table 7), and therefore was the best crossing and selection scheme.

Table 7 Comparison of the three top crosses and the two marker selection schemes

Marker selection scheme	Individuals in TCF_1 before selection	Individuals in TCF_1 after selection	Lines before selection	Lines after selection (S.E.)	DNA samples to be tested	DNA samples per selected line
Top cross 1: (CSSL15 × CSSL29) × CSSL49						
Scheme 1	200	200	4000	27.1 (6.6)	4000	148
Scheme 2	200	13	262	16.7 (6.2)	462	28
Top cross 2: (CSSL15 × CSSL49) × CSSL29						
Scheme 1	200	200	4000	12.9 (4.9)	4000	310
Scheme 2	200	6	124	7.9 (4.5)	324	41
Top cross 3: (CSSL29 × CSSL49) × CSSL15						
Scheme 1	200	200	4000	7.5 (3.1)	4000	536
Scheme 2	200	25	491	7.7 (3.1)	691	89

DISCUSSION

Breeding strategies used by CIMMYT breeders have evolved with time. Pedigree selection was used primarily from 1944 to 1985. From 1985 until the second half of the 1990s, the main selection method was a modified pedigree/bulk method (MODPED) (van Ginkel *et al*. 2002), which successfully produced many of the widely adapted wheats now being grown in the developing world. This method was replaced in the late 1990s by the selected bulk method (SELBLK) (van Ginkel *et al*. 2002) in an attempt to improve resource-use efficiency. Before simulation, the breeders already knew that SELBLK could save costs compared to MODPED. The simulation not only confirmed this knowledge, but also gave a clear answer to the breeder that the adoption of SELBLK would not cause a yield gain penalty. Simulation also indicated a fact that CIMMYT's breeders did not realize. The fact was that SELBLK could retain more crosses in the final selected population. When this result came out, CIMMYT's historical breeding books were checked and it was found that this was true. Therefore simulation can not only confirm breeders' intuitive experiences, but can also find out some facts which breeders do not realize.

In field-based breeding, the breeder selects the phenotype. However, in simulation the genotype must be defined. The genotypic value of the genotype can be calculated from the definition of gene actions. The phenotypic value and family mean can be found from the genotypic value and its associated error (environmental deviation). QuLine then conducts within-family selection from phenotypic values and among-family selection from family means. A sensible definition of genetic models is thus essential for any such simulation, as it determines the phenotypic value of a genotype and then the phenotypic mean of a population to which the selection is applied. However, given the current state of the knowledge of gene-to-phenotype relationships for complex traits, it is difficult to comprehensively define a real genetic model.

In the future it will be possible to build more realistic genetic models if advances in genomics improve the understanding of the genotype to phenotype relationship and genotype by environment interactions (Bernardo 2002; Cooper *et al*. 2005). Conclusions on the relative merits of breeding strategies based on simple gene-to-phenotype models may have to be re-evaluated in the context of an exponentially growing knowledge base. This information will aid in determining gene number and gene effects on phenotype. In addition, conventional plant breeding provides a wealth of information about trait heritability and trait correlation. This information, once determined, will help define errors, linkage, and pleiotropic effects. In addition, crop physiological models may also help fine-tune the genetic models for breeding modeling (Reymond *et al*. 2003; Yin *et al*. 2004; Hammer *et al*. 2005).

As there is accumulation in the knowledge of the genetics for most breeding traits, simulation modeling will become more and more important, as computer simulation can help to investigate many what-if crossing and selection scenarios, and allows many scenarios to be tested *in silico* in a short period of time, which in turn helps breeders make important decisions before conducting highly resource demanding field experiments.

Acknowledgements

The development of QuLine (freely available from the senior author) was supported by GRDC (Grains Research and Development Corporation) of Australia (2000-2004), Generation and HarvestPlus Challenge Programs of the Consultative Group on International Agricultural Research (2005-2007). This work was supported in part by the National 863 Program of China (and 2006AA10Z1B1).

References

Barton N H, Keightley P D. 2002. Understanding quantitative genetic variation. *Nature Review Genetics*, **3**, 11-21.

Bernardo R. 2002. *Breeding for Quantitative Traits in Plants*. Stemma Press, Woodbury, Minnesota.

Cooper M, Podlich D W, Smith O S. 2005. Gene-to-phenotype and complex trait genetics. *Australian Journal of Agricultural Research*, **56**, 895-918.

Eagles H A, Eastwood R F, Hollamby G J, Martin E M, Cornish G B. 2004. Revision of the estimates of glutenin gene effects at the Glu-B1 locus form southern Australian wheat breeding programs. *Australian Journal of Agricultural Research*, **55**, 1093-1096.

Eagles H A, Hollamby G J, Gororo N N, Eastwood R F. 2002. Estimation and utilization of glutein gene effects from the analysis of unbalanced data from wheat breeding programs. *Australian Journal of Agricultural Research*, **53**, 367-377.

Falconer D S, Mackay T F C. 1996. *Introduction to Quantitative Genetics*. 4th ed. Longman, Essenx, England.

Fehr W R. 1987. *Principles of Cultivar Development*. Vol. 1. *Theory and Technique*. Macmillian Publishing Company, New York.

Frary An, Nesbitt T C, Frary Am, Grandillo S, Knaap E V D, Cong B, Liu J P, Meller J, Elber R. Alpert K B, Tanksley S D. 2000. *fw2.2*: A quantitative trait locus key to the evolution of tomato fruit size. *Science*, **289**, 85-88.

Goldman I L. 2000. Prediction in plant breeding. *Plant Breeding Reviews*, **19**, 15-40.

Hammer G L, Chapman S C, van Oosterom E, Podlich D. 2005. Trait physiology and crop modeling as a framework to link phenotypic complexity to underlying genetic systems. *Australian Journal of Agricultural Research*, **56**, 947-960.

Kempthorne O. 1988. An overview of the field of quantitative genetics. In: Weir B S, Eisen E J, Goodman M M, Namkoong G, eds. Proceedings of the 2nd International Conference on Quantitative Genetics. Sinauer Associates, Inc. Sunderland, MA. pp.47-56.

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.

Li Z K, Yu S B, Lafitte H R, Huang L, Courtois B, Hittalmani S, Vijayakumar C H M, Liu G F, Wang G C, Shashidhar H E, Zhuang J Y, Zheng K L, Singh V P, Sidhu J S, Srivantaneeyakul S, Khush G S. 2003. QTL × environment interactions in rice. I. Heading date and plant height. *Theoretical and Applied Genetics*, **108**, 141-153.

Lynch M, Walsh B. 1998. *Genetics and Analysis of Quantitative Genetics*. Sinauer Associates, Inc. Sunderland, MA.

Peleman J D, Voort J R. 2003. Breeding by design. *Trends in Plant Science*, **8**, 330-334.

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

Reymond M, Muller B, Leonardi A, Charcosset A, Tardiew F. 2003. Combining quantitative trait loci and an ecophysiological model to analyze the genetic variability of responses of maize leaf growth to temperature and water deficit. *Plant Physiology*, 131, 664-675.

Singh R P, Rajaram S, Miranda A, Huerta-Espino J, Autrique E. 1998. Comparison of two crossing and four selection schemes for yield, yield traits, and slow rusting resistance to leaf rust in wheat. *Euphytica*, **100**, 35-43.

Tanksley S D, Nelson J C. 1996. Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadpated germplasm into elite breeding lines. *Theoretical and Applied Genetics*, **92**, 191-203.

van Ginkel M, Trethowan R, Ammar K, Wang J, Lillemo M. 2002. Guide to bread wheat breeding at CIMMYT (rev). Wheat Special Report, CIMMYT, D.F. Mexico. No. 5.

Wan J M. 2006. Perspectives of molecular design breeding in crops. *Acta Agronomica Sinica*, **32**, 455-462. (in Chinese)

Wan X Y, Wan J M, Jiang L, Wang J K, Zhai H Q, Weng J F, Wang H L, Lei C H, Wang J L, Zhang X, Cheng Z J, Guo X P. 2006. QTL analysis for rice grain length and fine mapping of an identified QTL with stable and major effects. *Theoretical and Applied Genetics*, 112, 1258-1270.

Wan X Y, Wan J M, Weng J F, Jiang L, Bi J C, Wang C M, Zhai H Q. 2005. Stability of QTLs for rice grain dimension and endosperm chalkiness characteristics across eight environments. *Theoretical and Applied Genetics*, **110**, 1334-1346.

Wang J, Chapman S C, Bonnett D G, Rebetzke G J, Crouch J. 2007. Application of population genetic theory and simulation models to efficiently pyramid multiple genes via marker-assisted selection. *Crop Science*, (in press).

Wang J, Eagles H A, 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.

Wang J, Ginkel M, Trethowan R, Ye G, DeLacy I H, Podlich D, Cooper M. 2004. Simulating the effects of dominance and epistasis on selecting response in the CIMMYT wheat breeding program using QuLine. *Crop Science*, **44**, 2006-2018.

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

Wang J K, Wan X Y, Crossa J, Crouch J, Weng J F, Zhai H Q, Wan J M. 2006. QTL mapping of grain length in rice (*Oryza sativa* L.) using chromosome segment substitution lines. *Genetical Research*, **88**, 93-104.

Yin X, Struik P C, Kropff M J. 2004. Role of crop physiology in predicting gene-to-phenotype relationships. *Trends in Plant Science*, **9**, 426-432.

Zeng Z B. 1994. Precision mapping of quantitative trait loci. *Genetics*, **136**, 1457-1468.

(Edited by ZHANG Yi-min)