CHAPTER 4

ACCOUNTING FOR VARIABILITY IN THE DETECTION AND USE OF MARKERS FOR SIMPLE AND COMPLEX TRAITS

S.C. CHAPMAN[#], J. WANG[##], G.J. REBETZKE[###] AND D.G. BONNETT[###]

[#] *CSIRO Plant Industry, Queensland Bioscience Precinct, St. Lucia, QLD 4067, Australia.*
[##] *Crop Research Informatics Laboratory, International Maize and Wheat Improvement Center (CIMMYT), Institute of Crop Science, and The National Key Facility for Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences, Beijing 100081, China.*
[###] *CSIRO Plant Industry, P.O. Box 1600, Canberra, ACT 2601, Australia.*
E-mail: scott.chapman@csiro.au

Abstract. There are many sources of variability in gene–phenotype associations. During the measurement of genotype and phenotype and during selection, researchers must deal with experimental error in trials; gene-gene interaction (epistasis) for sub-traits and observed traits; trait-trait interaction (pleiotropy) and gene- or genotype-by-environment interaction. These effects can be structured in a framework that allows simulation of the entire gene-environment 'landscape'. Studies of these landscapes have been published by others. Here we aim to explain with simple examples some of the types of insights that can be made. A current challenge for breeders working with simple marker–phenotype associations is to design selection strategies that can rapidly create new combinations of multiple marker-based traits. For a real-world example in wheat, we have used simulation to show how gene enrichment during early generations (selection of homozygotes and heterozygotes with desirable alleles) can greatly reduce resource requirements when combining 9 genes into one genotype through marker-assisted selection. Another wheat example compares phenotypic and QTL-based selection for coleoptile length where the QTL also had a pleiotropic association with plant height. These simulations show the relative negative effects of either low heritability, or less than complete detection of QTL associated with traits. Finally, we revisit a marker-assisted selection (MAS) example whereby a QTL study is undertaken on a population for a complex trait, and then those QTL are used in selection. This process is subject to all sources of error described above. If the trait is complex, then interactions among sub-traits; between sub-traits and the environment; or between the chromosomal locations of controlling genes, create an extremely 'rugged' selection landscape that slows breeding progress. In this situation, a detailed understanding of some of these interactions is required if MAS is to be able to exceed the progress of conventional breeding.

J.H.J. Spiertz, P.C. Struik and H.H. Van Laar (eds.), Scale and Complexity in Plant Systems Research: Gene-Plant-Crop Relations, 37–44.

INTRODUCTION AND BACKGROUND

Many breeding programmes are now utilizing marker–trait associations as part of their selection process. Some of the typical applications include introgression of traits from donor ('unadapted' lines) into parental germplasm; broadening the genetic base of a crop (Xu et al. 2004); selection of parental combinations, based on marker profiles (Wang et al. 2005); selection of cross progeny during early and late generations of selfing and evaluation (Eagles et al. 2001); and recurrent selection based on marker–trait associations (Podlich et al. 2004). The value of markers is heightened when the target trait is difficult or expensive to screen, like resistance to cereal cyst nematodes (Ogbonnaya et al. 2001). With the continued expansion of information on quantitative trait loci (QTL) for more complex traits, there is an increasing desire to implement these efficiently in plant-breeding programmes, with new strategies being proposed for this (Podlich et al. 2004).

In considering how to utilize markers, several issues arise around the association of markers with genes that affect trait expression, and around the precision of the estimates of the relationship between alleles and trait expression. In perhaps the simplest relationship, the presence of a single allele at a single locus explains 100% of the observed phenotype in a particular environment, such as in the case of a gene that confers resistance to a single rust pathotype. This association is effectively the same as a qualitative gene effect like seed colour, apart from the need to have the 'rust environment' to see the effect. If the gene sequence is known, and/or the phenotype has been carefully mapped in crosses or screened across a large number of resistance and susceptible lines, then a 'perfect' allele marker may be available (Ogbonnaya et al. 2001). So, the gene-trait relationship is 100% explained; there is no genetic-background effect (the marker works in different pedigrees); and the relationship can be predicted without error by the presence/absence of a marker for the desired allele. The challenge for breeders then is to combine sets of essential alleles into single backgrounds.

At the other extreme is the case where many genes interact with each other (epistasis) in different environments (gene-by-environment interaction) and affect sub-traits that interact to determine the desirable trait (pleiotropy). For a particularly complex trait like yield, there are networks of interactions, including recursive effects, among these components of control of the desirable trait. A QTL study will never explain 100% of the genetic variation typically observed. During selection these effects will be apparent as low heritability for the trait and/or poor linkage between QTL and their markers. But it may not be clear how the main sources of error (epistasis, gene-by-environment or pleiotropy) result in the residual variance that is not explained by QTL.

Recently, Cooper et al. (2005) proposed a gene-to-phenotype modelling framework to utilize molecular breeding for complex traits. This illustrates, for a large number of genetic models, how the 'context-dependent' relationships between genes (epistasis, gene-by-environment interaction and pleiotropy) impact on genetic progress in both molecular and phenotypic breeding strategies. They propose, as an alternative to 'traditional' quantitative genetic models (say, comprised of genotype and genotype-by-environment interaction effects), to work with models where

phenotype is described as a function of 'explained' and 'unexplained' sources of variation, and these sources are associated with vectors of 'known' and 'unknown' gene (or QTL) and gene-by-environment effects. For example, a simulation may use the predicted effects and the QTL/marker locations from a QTL study to simulate genetic progress in a breeding programme, assuming that 100% of the genetic variance is explained by the QTL. In an actual study, for all but the simplest gene–trait relationships, the 'unexplained' variance for a complex trait is typically 20 to 80%. The simulation can then be re-run multiple times, while adding in each an ensemble of different gene effects (representing epistasis and G×E etc.) to determine the potential effect of this 'unexplained variance' on expected genetic progress. Not so surprisingly, real-world QTL studies where the unexplained variance was high suffered more in terms of potential impact on selection, but now there is a method to quantify this effect in terms of expected context dependencies. These methods can help breeders to decide on the likely usefulness of the QTL in their selection scheme, given a better understanding of how robust the QTL are for expected (or unexpected) levels of complexity in the 'unexplained' variance.

While there has been some application of simulation approaches to examine the value of QTL for complex traits in Australian sorghum (Hammer et al. 2005), marker technology is still being developed for application in that breeding programme, but is focusing on utilization for QTL associated with complex traits such as midge resistance and stay-green (see Hammer et al., Chapter 5). Markers for single-gene/single-trait applications have been used in wheat-breeding programmes in Australia for over 10 years, e.g., Ogbonnaya et al. (2001) and Eagles et al. (2001). In general, their use has been in introgressing into breeding lines (in BCF_1) and in screening progeny in early (F_2) and later generations of evaluation. A pertinent task for these breeding programmes is devising strategies to combine these many 'simple' genes together into breeding lines.

Cooper et al. (2005) explored a large number of breeding scenarios, focusing on QTL for complex traits, and were able to summarize from these that gene-by-environment effects were still a substantial impediment to marker selection for complex traits. We aimed to present three practical scenarios of applying marker-assisted selection:

1. from a 3-way cross, recovering a target genotype comprising 9 desirable genes that have near-perfect markers;
2. from a 2-way cross, selecting for a quantitative trait (coleoptile length), given different levels of knowledge of the genetic variation explained by the QTL;
3. for a sorghum-breeding programme, selection for yield based on QTL detected in a single environment, compared to progress based on knowledge of underlying 'physiological pleiotropy' controlling yield.

Throughout the chapter, we aim to demonstrate how these approaches can account for sources of variability and assist breeders to deal with them.

MATERIALS AND METHODS

While some of the calculations presented here can be applied quite simply to sets of unlinked genes, the QU-GENE simulation platform was used for more complex scenarios (Podlich and Cooper 1998). The programme generates populations of genotypes and provides a library of subroutines to develop simulation modules of breeding programmes. For the wheat examples, we simulated selection using QuLine, a breeding module used to simulate wheat-breeding programmes (Wang et al. 2003), and to predict cross performance for quality traits (Wang et al. 2005). For sorghum, the original simulations were done using a proprietary breeding module (Hammer et al. 2005). More details of the examples are given in the Results section.

RESULTS

Example 1: Single-gene control of traits – Using F_2 enrichment to combine 'simple' genes in a complex cross

This example is the subject of a paper (Wang et al. in press) that explores additional details beyond those given here. Where 5 genes are unlinked and a simple cross is considered, the frequency (f) of the desired homozygote in the F_2 can be estimated as $0.25^5 = 0.00098$. To select one target genotype at an acceptance probability (α) of 0.01, this would require an F_2 population size of about 4,700 individuals, estimated from log (α) / log (1 – f). Delaying selection until lines are homozygous requires only 145 individuals as the frequency becomes 0.5^5. For 12 independent loci, > 77 million lines are needed to identify a single homozygote in the F_2, or > 18,000 in fixed lines. In this case, F_2 'gene enrichment' (selection of homozygotes and heterzygotes (Bonnett et al. 2005)) is a useful strategy as only 144 F_2s would need to be screened to retain the desired gene combination ($f = 0.75^{12} = 0.03168$), followed by screening of 596 fixed lines to recover then a homozygote individual ($f = (2/3)^{12} = 0.00771$).

Using simulation (QU-GENE/QuLine, Wang et al. 2003), we examined progeny from a 3-way cross ((Silverstar + *tin* × HM14BS) × Sunstate) segregating at 9 loci (7 independent). The aim was to recover a target genotype (at overall acceptance α = 0.01) that had the required alleles (bottom line of Table 1).

In the TCF_1, selection of *Rht-B1a* and *Glu-B1i* homozygotes could be fixed, and enrichment (selection for heterozygotes) done for *Rht8*, *Cre1*, and *tin*. If no selection was applied in the F_2, then a total of > 3500 lines (> 25,000 marker screens) were needed to recover the target genotype (Table 2). This was reduced to < 600 lines (< 3500 marker screens) if F_2 enrichment was used for the 7 loci that had not been fixed in the TCF_1. The effect of linkage between the *Glu-A3* and *tin* loci, and the non-perfect marker for *tin* (Table 1) resulted in a final frequency of the *tin* gene of 0.79, while other genes were all fixed at frequencies of 1.0 or > 0.98. Therefore, the presence of *tin* would still need to be confirmed by phenotyping after production of the fixed lines. So, in this example of multiple gene selection, the desired gene combinations can be achieved with a relatively small number of screens, even given

slightly imperfect markers for three trait loci, and linkage-in-repulsion for two of the loci.

Example 2: Polygenic control of quantitative traits – Selection for increased coleoptile length and reduced height

The GA-insensitive, height-reducing gene, *Rht-D1b*, also reduces coleoptile length (cl) by about 20% in wheat seedlings, while the *Rht8* gene has virtually no effect. Long coleoptiles are desirable so that seeds can be planted deeper to access soil water better at sowing. The screening of the cl phenotype is taken after a set period growing in dark conditions in a controlled-temperature environment.

Based on QTL-mapping studies (Rebetzke et al. 2001; in press), 8 QTL were considered to affect height (ht) and coleoptile length in addition to the major height genes. Supposing the reduced height alleles at *Rht-D1* and *Rht8* reduce the plant height by 10 and 8 cm (explaining 48% and 31% of genetic variance, respectively; Ellis et al. 2002), then these additional QTL affecting plant height by 2 to 3 cm each explain between 2 and 5% of the genetic variance (data not shown). The QTL for coleoptile length explain similar proportions of genetic variance (equating to –3 to +4 mm), while the 18-mm reduction due to *Rht-D1b* explains about 80% of the

***Table 1**. Selected genes, their chromosomal location and the genotypes for the three parents*

Gene symbol	Rht-B1	Rht-D1	Rht8	Sr2	Cre1	VPM	Glu-B1	Glu-A3	tin
Chromosome	*4BS*	*4DS*	*2DL*	*3BS*	*2BL*	*7DL*	*1BL*	*1AS*	*1AS*
Marker type	*Cod.*	*Cod.*	*Cod.*	*Cod.*	*Dom.*	*Dom.*	*Cod.*	*Cod.*	*Cod.*
Marker–gene distance (cM)	*0*	*0*	*0.6*	*1.1*	*0*	*0*	*0*	*0*	*0.8*
Silverstar+ *tin*	***Rht-B1b***	*Rht-D1a*	*rht8*	*sr2*	***Cre1***	*vpm*	***Glu-B1i***	*Glu-A3c*	***tin***
HM14BS	*Rht-B1a*	*Rht-D1a*	***Rht8***	*sr2*	*cre1*	*vpm*	*Glu-B1a*	*Glu-A3e*	*Tin*
Sunstate[a]	*Rht-B1a*	***Rht-D1b***	*rht8*	***Sr2***	*cre1*	***VPM***	***Glu-B1i***	***Glu-A3b***	*Tin*
Target genotype	*Rht-B1a*	*Rht-D1a*	***Rht8***	***Sr2***	***Cre1***	***VPM***	***Glu-B1i***	***Glu-A3b***	***tin***

[a] The bold-printed alleles at *Rht-B1*, *Rht-D1* and *Rht8* reduce plant height; those at *Sr2*, *Cre1*, and *VPM* confer resistance to rusts or cereal-cyst nematode; those at *Glu-B1* and *Glu-A3* improve dough quality; and, the bold-printed allele at *tin* reduces the tiller number. The genes are all unlinked, except for *Glu-A3* and *tin*, which are linked in repulsion at 3.8 cM apart on chromosome 1AS.

***Table 2**. Selected proportion and number of individuals (or families) selected in each marker selection scheme*

Breeding population	*No enrichment selection in TCF_2*		*Enrichment selection for all target genes in TCF_2*	
	Selected proportion	*Minimum population size*	*Selected proportion*	*Minimum population size*
TCF_1	0.0313	145	0.0316	144
TCF_2			0.1190	37
DHs	0.0013	3440	0.0112	408

genetic variance. We undertook a series of simulations of a cross between HM14BS (ht 82 cm; cl 125 mm) and Sunstate (ht 78 cm; cl 75 mm) to attempt to recover a target genotype with increased coleoptile length and reduced height, but with a greater proportion of the desirable Sunstate genetic background.

In all cases, the process was to make the cross, produce ten F_1 plants, and then produce 1000 doubled haploid (fixed) lines prior to selection by either phenotype or by combinations of the markers for the two major and eight minor QTL. In an initial simulation (1), we made a single cross between HM14BS and Sunstate, assuming broad-sense heritabilities of 0.7 and 0.8 for height and coleoptile length, respectively, and undertook selection for coleoptile length in the 1000 DH lines, with no selection for height (Table 3). As might be expected, this led to a taller phenotype with a long coleoptile, i.e. a greater proportion of lines carrying both the *Rht-D1b* and *Rht8* alleles, and minor QTL for both coleoptile length and height. The next two simulations (2 and 3) show the effect of experimental precision in the measurement of coleoptile length. Compared to the initial simulation ($H_b = 0.8$), the final length of the selected lines decreased or increased by 5 mm or more as the phenotyping was made less precise ($H_b = 0.5$) or more precise ($H_b = 1.0$).

The remaining simulations (4 to 6) involve selection using the QTL information. For the major QTL, selection was against *Rht-D1b* and for *Rht8*, to increase the coleoptile length while trying to minimize the effect on height. When selection was applied only to these major QTL (simulation 4, Table 3), followed by selection on coleoptile phenotype, the plant height was close to that of HM4BS.

Table 3. *Breeding schemes and final height (ht) and coleoptile length (cl) of top 2% of lines*

Scheme	*Heritability*		*Selection for QTL*		*Selection for cl*	*Mean value*	
	Ht	*Cl*	*Major*	*Minor*		*ht (cm)*	*cl (mm)*
1	0.7	0.8	No	No	Yes	91.1	132.8
2	0.5	0.5	No	No	Yes	90.7	127.8
3	1.0	1.0	No	No	Yes	90.9	138.3
4	0.7	0.8	Yes	No	Yes	82.6	126.6
5	0.7	0.8	Yes	8	No	82.9	123.5
6	0.7	0.8	Yes	4	No	82.0	133.7

Example 3: Polygenic control of complex traits – Selection for 'yield' QTL in sorghum

At the other extreme of gene–trait relationships, is the example of selection of markers linked to QTL controlling complex traits. In this situation, many sources of error exist, which include: experimental error in measuring the phenotype during the QTL study (trait heritability); error in selection of the marker or markers for the QTL (poor linkage); lack of observation (or knowledge) about how 'sub-traits' combine physiologically to affect the trait of interest; and, most critically, lack of knowledge of the gene action of the 'unexplained' variance in the QTL study.

Using simulation, Chapman et al. (2003) and Hammer et al. (2005) illustrated that when simple additive gene action was defined for four sub-traits ('trait

parameters') in a sorghum-cropping system, complex gene-by-gene-by-environment interactions could still be generated for expression of yield. Models of gene action (i.e. for phenotypes associated with QTL) were used to define trait parameters that are input values to a crop simulation model. For example, one trait parameter was the relationship between crop development rate (toward flowering) and temperature. For this trait, a simple additive three-gene model, based on existing knowledge of QTL, calculated the parameter for each of the simulated genotypes created in a population. The calculated parameters (for this and the other three traits) were input to a crop simulation model, and the model was then run for each genotype using soil data for six locations and weather data over 100 years. This generated a complex 'gene–environment landscape' from which environments could be sampled (e.g., several locations in a single year) and genotypes could be selected on the basis of the expression of the trait value as it affected yield. The best lines were crossed to create new generations in a manner similar to a conventional breeding programme. Chapman et al. (2003) quantified how bias in the sampling of environments by the breeding programme (because of variability in rainfall between successive seasons) reduced the efficiency of selection, through the generation of substantial genotype-by-environment interaction. Using the same dataset, Hammer et al. (2005) showed how even 'simple' combinations of traits across genotypes and environments could easily confound detection of QTL associated with yield.

DISCUSSION

In example 1, there was no attempt to select for 'background' alleles during the process of combining the essential genes. In practice, the breeding programme screens a large number of lines (ca. 10 to 20% more than indicated) using F_2 enrichment so that more than one target genotype is recovered. These target lines are then tested for field performance and may then be used as cultivars and/or parents in crossing and selection. In Australian wheat breeding greater disease resistance and grain quality are deemed 'essential' and have often taken priority over selection for yield *per se*, with integration of new sources of yield adaptation taking quite some time. This contrasts with the situation illustrated in sorghum, and in US corn breeding, where a major objective is to maintain and build upon elite combinations of genes for complex traits like yield (Duvick et al. 2004; Podlich et al. 2004).

Using a slightly different simulation approach that studied only marker-assisted recurrent selection, Bernardo and Charcosset (2006) found that if large numbers (say 40 to 100) of QTL affected a trait, it was more advantageous to use only large-effect QTL and to ignore the small-effect QTL in selection, given the small population size typically used in marker-assisted recurrent selection. However, empirical evidence suggests that these large-effect QTL are fixed in early cycles while evidence from other studies (e.g., Openshaw and Frascaroli 1997) show that many of the genetic effects for traits such as yield are indeed small.

Thus, for most important breeding traits it is challenging to implement the large amount of QTL studies through marker-assisted selection to exceed the breeding efficiency of the conventional phenotypic selection.

ACKNOWLEDGEMENTS

D.W. Podlich, M. Cooper, G.L. Hammer are all acknowledged for contributions to discussions about this work, particularly for the sorghum simulation example. This research has been funded or in collaboration with the CGIAR Generation Challenge Program, the Australian Grains Research and Development Corporation, the Australian Research Council, Pioneer Hi-Bred, The University of Queensland and the Queensland Department of Primary Industries.

REFERENCES

Bernardo, R. and Charcosset, A., 2006. Usefulness of gene information in marker-assisted recurrent selection: a simulation appraisal. *Crop Science,* 46 (2), 614-621.

Bonnett, D.G., Rebetzke, G.J. and Spielmeyer, W., 2005. Strategies for efficient implementation of molecular markers in wheat breeding. *Molecular Breeding,* 15 (1), 75-85.

Chapman, S.C., Cooper, M., Podlich, D., et al., 2003. Evaluating plant breeding strategies by simulating gene action and dryland environment effects. *Agronomy Journal,* 95 (1), 99-113.

Cooper, M., Podlich, D.W. and Smith, O.S., 2005. Gene-to-phenotype models and complex trait genetics. *Australian Journal of Agricultural Research,* 56 (9), 895-918.

Duvick, D.N., Smith, J.S.C. and Cooper, M., 2004. Long-term selection in a commercial hybrid maize breeding program. *Plant Breeding Reviews,* 24 (2), 109-152.

Eagles, H.A., Bariana, H.S., Ogbonnaya, F.C., et al., 2001. Implementation of markers in Australian wheat breeding. *Australian Journal of Agricultural Research,* 52 (11/12), 1349-1356.

Ellis, M.H., Spielmeyer, W., Gale, K.R., et al., 2002. "Perfect" markers for the *Rht-B1b* and *Rht-D1b* dwarfing genes in wheat (*Triticum aestivum* L.). *Theoretical and Applied Genetics,* 105 (6/7), 1038-1042.

Hammer, G.L., Chapman, S., Van Oosterom, E., et al., 2005. Trait physiology and crop modelling as a framework to link phenotypic complexity to underlying genetic systems. *Australian Journal of Agricultural Research,* 56 (9), 947-960.

Ogbonnaya, F.C., Subrahmanyam, N.C., Moullet, O., et al., 2001. Diagnostic DNA markers for cereal cyst nematode resistance in bread wheat. *Australian Journal of Agricultural Research,* 52 (11/12), 1367-1374.

Openshaw, S. and Frascaroli, E., 1997. QTL detection and marker-assisted selection for complex traits in maize. *Annual Corn Sorghum Research Conference Proceedings,* 52, 44-53.

Podlich, D.W. and Cooper, M., 1998. QU-GENE: a simulation platform for quantitative analysis of genetic models. *Bioinformatics,* 14 (7), 632-653.

Podlich, D.W., Winkler, C.R. and Cooper, M., 2004. Mapping as you go: an effective approach for marker-assisted selection of complex traits. *Crop Science,* 44 (5), 1560-1571.

Rebetzke, G.J., Appels, R., Morrison, A.D., et al., 2001. Quantitative trait loci on chromosome 4B for coleoptile length and early vigour in wheat (*Triticum aestivum* L.). *Australian Journal of Agricultural Research,* 52 (11/12), 1221-1234.

Rebetzke, G.J., Ellis, M.H., Bonnett, D.G., et al., in press. Molecular mapping of genes for coleoptile growth in bread wheat (*Triticum aestivum* L.). *Theoretical and Applied Genetics*.

Wang, J., Van Ginkel, M., Podlich, D., et al., 2003. Comparison of two breeding strategies by computer simulation. *Crop Science,* 43 (5), 1764-1773.

Wang, J., Eagles, H.A., Trethowan, R., et al., 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 (5), 465-473.

Wang, J., Chapman, S.C., Bonnett, D.B., et al., in press. Application of population genetic theory and simulation models to efficiently pyramid multiple genes via marker-assisted selection. *Crop Science*.

Xu, Y., Beachell, H. and McCouch, S.R., 2004. A marker-based approach to broadening the genetic base of rice in the USA. *Crop Science,* 44 (6), 1947-1959.