Chapter 5
Historical and Prospective Applications of 'Quantitative Genomics' in Utilising Germplasm Resources

Adrian Hathorn and Scott C. Chapman

Contents

Abstract The last 30 years have seen major changes in the field of plant breeding. In this relatively short time frame we have witnessed the transition from a solely pedigree-based approach to genetic improvement, to one based almost entirely on genome-wide sequence information. We have also witnessed the evolution of dominant genetic theory, including the adoption of new statistical techniques necessary to accommodate the plethora of genomic information now available. In this chapter we review the past, present and future of plant breeding in terms of the three distinct "eras": "the pedigree era", "the molecular era" and the "genomics era".

A. Hathorn (✉) · S. C. Chapman
CSIRO Plant Industry, Queensland Bioscience Precinct, 306 Carmody Rd.,
St. Lucia, QLD 4067, Australia
e-mail: adrian.hathorn@csiro.au

R. Tuberosa et al. (eds.), *Genomics of Plant Genetic Resources,*
DOI 10.1007/978-94-007-7572-5_5,

Keywords MAS · MARS · GWS · QTL mapping · Plant breeding · Genomics

5.1 Introduction

High-throughput genotyping technologies, in particular the single nucleotide polymorphism (SNP) chip, have prompted a revolution in the field of genetics and breeding. With the potential of genotyping literally hundreds of thousands of molecular markers at an affordable price, the once distant prospect of establishing an individual's genetic value without need of its pedigree has now become a reality. In the following chapter we consider the three distinct eras of quantitative genetics that led to this genotyping revolution, with final emphasis on genome-wide selection (GWS). We begin with "the pedigree era", to describe analysis prior to DNA markers and revisit the fundamentals of quantitative genetic theory. In "the molecular era" we review the birth of molecular markers and the advent of marker assisted selection (MAS). Finally, in "the genomic era", we describe the development of GWS and its current and future implications for plant breeding and utilization of genetic resources.

5.2 The Pedigree Era

5.2.1 The Infinitesimal Model

Prior to the era of molecular markers and genomics, genetic variation in quantitative traits was explained by modeling an individual's phenotype as the sum of an infinite number of infinitesimally small genetic effects plus an interaction between genotype and environmental values:

$$y_{ij} = \mu + g_i + e_{ij} \tag{5.1}$$

where; y_{ij} is the phenotype of individual i observed in environment j, μ refers to the fixed environmental effects of individual i, g_i is the total genetic value of individual i, and e_{ij} is the sum of random environmental effects affecting individual i in environment j. This is more commonly known as the infinitesimal model. The total genetic value of an individual g_i can be further partitioned into additive (g_A), dominance (g_D) and epistatic (g_E) components, with g_D and g_E representing the non-additive component of the genetic variation. Elaborations on this model include interaction effects of genotype with environment and also aim to consider interactions of different trait phenotypes using selection indices, as discussed below.

5.2.2 The Concept of Breeding Values

Substantial genetic improvement in animal breeding has been achieved by selecting on estimated breeding values (EBVs). However it has not been a popular index in plant breeding due mainly to the fact that the EBV is based on a simplified definition of heritability tailored towards selection on individual animals and does not take into account the diversity of observational units and mating systems used in plant breeding (Holland et al. 2010). However, EBVs are central to applications of genomic selection and are introduced here for that purpose.

The EBV represents the sum of the additive effects of an individual's genes (Falconer and Mackay 1996; Lynch and Walsh 1998) and is typically used to determine an animal's genetic potential when used as a parent. In its simplest form an EBV is estimated using the individual's phenotype and the population narrow sense heritability, calculated as $h^2 = {}^{\sigma^2_A}\!/_{\sigma^2_P}$. The difference between an individual's phenotypic value and the mean value of its population is adjusted according to h^2 in the following way:

$$EBV_i = m_0 + h^2(y_i - m_0) \tag{5.2}$$

where y_i is the phenotypic value of individual i and m_0 is the mean phenotypic value of the population. In this case, adjusting the phenotype according to the population narrow-sense heritability is a way of recognising that only a fraction of an individual's phenotype is heritable. Furthermore, as each parent contributes a sample half of its genes to its progeny, it can also only transmit one-half of its genetic value. Thus the expected breeding value of the offspring of parents 1 and 2 is equal to:

$$E(Progeny_{P1\times P2}) = m_0 + \tfrac{1}{2}EBV_{P1} + \tfrac{1}{2}EBV_{P2} \tag{5.3}$$

5.2.3 Selection Indices

Selection indices are a way of combining information across pedigrees and across traits. Equation 5.3 shows how the expectation of a breeding value can be defined using parental EBV's. Depending on the accuracy of those breeding values, it may be useful to include information from more distant relatives such as grandparents. A selection index allows us to use this information in the one prediction thereby increasing the accuracy of genetic evaluation (Falconer and Mackay 1996).

The simplest example of using a selection index is the calculation of EBV's based on own performance (Eq. 5.2). This takes the form of $\boldsymbol{I = EBV_i = b_{AP}P_1}$ where $\boldsymbol{b_{AP}}$ is the simple regression of breeding value (A) on phenotype (P), and in the absence of any interaction between genotype and environment (GxE) is equal to ${}^{cov}(\boldsymbol{A},\boldsymbol{P})/_{\boldsymbol{\sigma^2_P}} = {}^{\boldsymbol{\sigma^2_A}}\!/_{\boldsymbol{\sigma^2_P}} = \boldsymbol{h^2}$ (Falconer and Mackay 1996). A selection index including information from a number of relatives therefore corresponds to a multiple

regression of breeding value on all the sources of information and the linear index of any one individual becomes:

$$I = b_{AP:1}P_1 + b_{AP:2}P_2 + b_{AP:3}P_3 + \ldots \tag{5.4}$$

5.2.4 Best Linear Unbiased Prediction (BLUP)

A limitation of the selection index approach is that the method does not adjust the data for fixed environment effects; this must be done separately before the analysis. Henderson (1976) devised an efficient method to simultaneously estimate genetic and environmental effects in a single analysis. Henderson's method, called best linear unbiased prediction (BLUP), uses a mixed model approach and rapidly became the most widely accepted method of genetic evaluation due to its many desirable statistical properties.

The mixed model approach to estimating breeding values consists of modeling each trait value as the sum of all fixed environmental effects and a residual component comprised of the sum of all random genetic effects (BV's) plus a random error. In matrix notation,

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Za} + \boldsymbol{e} \tag{5.5}$$

where $\boldsymbol{y}$ is a vector of trait values, $\boldsymbol{\beta}$ is a vector of fixed environmental effects with incidence matrix $\boldsymbol{X}$, $\boldsymbol{a}$ is a vector of random genetic effects with incidence matrix $\boldsymbol{Z}$ and $\boldsymbol{e}$ is a vector of errors. In plant breeding, alternative formulations of Eq. 5.5 are commonly used to incorporate random genotype by environment interaction effects as variance-covariance structures (Piepho 1997).

The basic mixed model used in both animal and plant breeding incorporates information from all relatives with or without phenotypic records to estimate BV's. Henderson (1976) showed that the expectation of fixed environmental effects $\hat{\boldsymbol{\beta}}$ and the expectation of random genetic effects $\hat{\boldsymbol{a}}$, are solutions to the mixed model equations:

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z} + \boldsymbol{A}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{a}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'y} \end{bmatrix} \tag{5.6}$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{a}}$ are referred to as the best linear unbiased estimate (BLUE) and best linear unbiased predictor (BLUP), respectively.

5.3 The Molecular Era

The discovery of DNA markers marked the dawn of a new era in agricultural breeding, with some expectation that the ability to select directly on genotypes (MAS—marker assisted selection) would lead to the redundancy of pedigree based selection methods.

While DNA-based markers have now been deployed extensively for the tracking and introgression of simple traits (Eathington et al. 2007), their use in the selection for complex traits such as crop yield has so far been largely constrained to pyramiding of quantitative trait loci (QTL) in single cross populations.

The first DNA-based genetic markers were restriction fragment length polymorphisms (RFLPs) (Botstein et al. 1980), first used for the improvement of qualitative traits in crops by Beckmann and Soller (1986). Other early generation markers were enhanced by the introduction of DNA amplification-based procedures, such as; random amplified polymorphic DNAs (RAPDs) (Williams et al. 1990), AFLPs (Vos et al. 1995), as well as another broad class of DNA markers categorized as simple sequence repeats (SSRs) (Akkaya et al. 1992). SSR's are typically the most widely used markers in major cereals as they are highly reliable, co-dominant in inheritance and highly polymorphic (Collard and Mackill 2008).

5.3.1 QTL Mapping

One of the first applications of molecular marker technology was the discovery and mapping of quantitative trait loci (QTL). There are two distinct approaches to finding and mapping QTL. The first involves testing for marker-trait associations in a segregating population using marker genotypes located across the entire genome. The second, commonly referred to as the candidate gene approach, intuitively proposes previously sequenced genes of known function as potentially containing molecular polymorphisms related to the trait of interest.

The discovery and mapping of QTL in plants via marker trait associations, typically starts with the development of a mapping population, say 100 to 500 segregating individuals derived from an F2 or backcross population. Individuals (or, for hybrid crops, test-cross progeny) are then phenotyped for each trait of interest and genotyped with evenly spaced markers across the genome (linkage mapping). A variation on this is association mapping, where the individuals to be mapped represent a diverse set of relevant germplasm, e.g. historical (founder) and current breeding lines and potential donor lines for useful traits (Lynch and Walsh 1998).

The standard process of establishing significant marker trait associations is to use ordinary least squares where markers are treated as fixed effects and selected for inclusion into a prediction model using a stepwise regression approach based on arbitrary significance thresholds (Lande and Thompson 1990). The effects of markers below this threshold are set to zero, whilst those above the threshold are included in the model. The stepwise approach is useful to the extent that it minimizes the complexity of the model and ensures that there remain sufficient degrees of freedom for the estimation of marker effects. Once significant marker-trait associations have been made and major QTL identified, in theory at least, these major QTL are then ready to be introgressed into elite germplasm, hopefully leading to the development of new and improved cultivars.

One variation of this approach involves the genotyping of only that part of the population exhibiting extreme phenotypes for the target trait, creating two distinct

pools of DNA (Michelmore et al. 1991). Association is then inferred by finding allelic frequency differences between the groups of plants with contrasting phenotypes (Lebowitz et al. 1987). This approach is also referred to as "bulked segregant analysis" and has been successful in genetic mapping in plants using RFLP and SSR markers (Xu and Crouch 2008). Despite numerous reports for single major genes and major QTL using this method, bulked DNA analysis reports have been challenged by problems relating to insufficient marker density, low power of QTL detection and high false positive rates for marker-trait associations (Xu and Crouch 2008). See Van Eeuwijk et al. (2010) for a summary of analysis methods to derive performance estimates from different populations and statistical models.

5.3.2 The Candidate Gene Approach

The candidate gene approach has three chronological steps (Pflieger et al. 2001). First, candidate genes are proposed based on molecular and physiological studies of a trait. Then, a molecular polymorphism is identified so that statistical correlations between candidate gene polymorphisms and phenotypic variation can be calculated in a set of genetically unrelated individuals. The final step is the validation step and involves conducting complementary experiments to confirm the actual involvement of the candidate gene in the trait variation.

Although the candidate gene approach has been successfully used to characterize disease resistance genes and has led to the isolation of many new putative function resistance genes (R-genes), it is generally regarded as an expensive alternative to QTL mapping, especially for complex growth related traits. A recognized problem with the approach is that there are often a large number of candidate genes affecting a trait, so many genes must be sequenced in many individuals. The cost of carrying out so many association studies in a large sample of individuals is both expensive and time consuming. Furthermore, there is always a chance that the true causative mutation(s) may lie in a gene that would not intuitively have been selected as a candidate gene (Pflieger et al. 2001).

5.3.3 Gene Introgression and QTL Pyramiding

One of the more successful applications of molecular marker technology is in the introgression of major genes via marker assisted backcrossing (MABC). Although backcrossing has been successfully used in plant breeding to integrate disease resistance into numerous crop species such as maize (Hooker 1977) and wheat (Sharma and Gill 1983), prior to the invention molecular markers it was often a slow and complicated process. Without markers, phenotypic selection had to be done at each stage of the process. Fortunately, the implementation of MABC circumvented much of this process by using marker information to track target alleles from the donor parent

(Lamkey and Lee 2006). Large crop breeding programs, such as those for soybean and maize, have been redesigned to accommodate a seven-fold increase in data and analysis demands to implement accelerated breeding based on markers (Eathington et al. 2007). The same authors describe the complexities of using MABC to transfer transgene segments into multiple adapted genetic backgrounds.

A major problem in traditional backcrossing techniques is linkage drag. The identification of plants possessing the target trait and a high level of resemblance to the recurrent parent was complicated due to the fact that unfavorable alleles closely flanking the target allele would often "hitch a ride" into the recurrent parent. The implementation of MABC significantly helped researchers manage linkage drag by using marker information to identify plants with a high proportion of desirable genome from the recurrent parent. There are many examples of the successful use of MABC in rice, in particular bacterial blight resistance (Chen et al. 2000) and submergence tolerance (Toojinda et al. 2005), but also in other cereals such as barley, maize and wheat (see Table 1 in Collard and Mackill (2008)). This was further supplemented by simulation studies which found methods to optimize the recovery of recurrent parent alleles in just a few generations of backcrossing (Hillel et al. 1990; Hospital and Charcosset 1997; Visscher et al. 1996). For most crops, over 90 % of the recurrent parental genotype can now be recovered within two generations (Xu and Crouch 2008).

MABC is usually conducted in conjunction with phenotypic selection for other adaptive traits such as yield. However it remains a difficult challenge to introgress and pyramid multiple genes into a single cultivar even if they can be identified. Most plant breeders are forced to juggle selection on multiple traits and this often involves the more difficult process of selecting for many QTL simultaneously. The complexity of this task increases exponentially with increasing numbers of QTL. For example, if the frequency of 10 favorable alleles between two inbred parental lines is 0.5, then assuming they are unlinked, the frequency of the ideotype in the progeny will be equal to 1 in every 1024 recombinants. Pyramiding genes from more than two parents is an even tougher challenge. If those same 10 QTL were evenly distributed amongst three parents, the frequency of the ideotype would be around 1 in 60,000 recombinants! Some notable examples of successful QTL pyramiding in cereals include bacterial blight resistance in rice (Huang et al. 1997) and yellow mosaic virus in barley (Okada et al. 2004).

A common method to increase the frequency of target genotypes is through 'enrichment' of target alleles in segregating generations (e.g. F_2), followed by inbreeding (Bonnett et al. 2005; Wang et al. 2007). If the frequency of the ideotype is rare in early generations, a compromise is to select on heterozygotes at some, or all of the loci in the F_2 generation. By selecting for both target homozygotes and heterozygotes, this filter removes non-target homozygotes from the population. Through further inbreeding, or the development of doubled-haploid (DH) populations, the frequency of target homozygotes in the population can be increased along with the frequency of the target ideotype. The benefits of F_2 enrichment have also been demonstrated through simulation (Wang et al. 2009).

5.4 The Genomic Era

The molecular era was characterized by extensive searches for individual QTL using early generation markers such as RFLPs, RAPDs, AFLPs and SSRs. We also witnessed the advent of MAS. Consequently, substantial effort was directed to issues of marker density, population size, selection fractions and the combining of QTL across different genetic backgrounds.

In a sense, the transition to the genomic era came about largely through technological necessity. Most of the early generation markers were developed using the Sanger sequencing method (Sanger et al. 1978), which was both expensive and labor intensive. With the discovery of single nucleotide polymorphism (SNP[1]) markers, there was increasing interest in developing a high-throughput low-cost assay that could make use of the relative abundance of these markers in both animal and plant genomes. A technological breakthrough came in the form of high density oligonucleotide arrays and quickly led to the development of massively parallel sequencing platforms, otherwise known as next-generation sequencing (NGS) platforms. The versatility of these arrays also allowed for the development of novel marker systems like single feature polymorphisms (SFPs), diversity array technology (DArT) and restriction site-associated DNA (RAD) markers (Gupta et al. 2008).

Although the emergence of NGS technologies has significantly reduced the cost of marker scoring (Shendure and Ji 2008), the development of new markers still requires significant investment (Deschamps et al. 2012). This is especially the case for crop species such as maize (SanMiguel et al. 1996) and wheat (Li et al. 2004), where the efficiency of the SNP discovery process is often hampered by large numbers of repetitive sequences. However, a new concept called genotype-by-sequencing (GBS) is beginning to emerge whereby massively parallel sequencing platforms are used to simultaneously develop and score SNP markers within a segregating population (Elshire et al. 2011). Since GBS can also be performed through a reduced representation approach (Van Orsouw et al. 2007), polymorphism discovery in larger and more complex genomes (e.g. allotetraploid durum wheat) is now becoming a simpler and more cost effective process (Trebbi et al. 2011).

5.4.1 Genome-Wide Selection

In 2001, (Meuwissen et al. 2001) proposed a method called "genome wide selection" or GWS, a simplification of the two step model selection approach detailed in (Lande and Thompson 1990). Rather than selecting a subset of markers for inclusion into the prediction model based on arbitrary significance thresholds, GWS proposed to exploit linkage disequilibrium (LD) within the genome by using all marker information in a single step to estimate individual genomic breeding values (GEBVs).

[1] SNP markers are point mutations commonly occurring throughout plant and animal genomes, whereby alleles differ by only one base position.

The implementation of GWS requires that a population of individuals (in structured or unstructured populations) be initially phenotyped for the trait of interest and genotyped for a pre-defined set of markers. This is referred to as the "training population". The purpose of the training population is to accurately calibrate the prediction model by correlating marker effects with phenotypic values. The markers can then be used to estimate the genetic value of successive generations of individuals without need of phenotyping.

The method is theoretically superior to MAS for several reasons. The traditional MAS approach of fitting only the largest QTL is subject to a degree of upward bias known as the Beavis effect, an unavoidable consequence of selecting *a posteriori* among many estimates (Beavis 1998; Xu 2003). Lande and Thompson (1990) proposed a method to avoid this bias by using one half of the data to select the loci with the largest effects, and the other half to re-estimate the effects, although this was deemed to be a suboptimal use of the information (Meuwissen et al. 2001).

Furthermore, by fitting all markers into the prediction model these analyses should capture all (or most) of the additive genetic variance. This is in contrast to traditional MAS, where the estimation of a subset of significant QTL results in only a portion of the genetic variance being captured (Goddard and Hayes 2007). One consequence of this is that there is a greater chance that the largest and most accurately estimated QTL will be fixed in the first cycle of selection, leaving insufficient residual variation to maintain genetic gain in the cycles thereafter (Moreau et al. 2004).

5.4.2 Stepwise Regression, BLUP and the Bayesian Alphabet

Although stepwise regression is the technique of choice for selecting and fitting QTL markers in MAS, the choice of statistical technique for fitting all markers simultaneously in GWS is the topic of continuous debate. In GWS analyses, the number of marker effects to estimate will almost always be greater than the number of records (Goddard and Hayes 2007), and estimating a large number of marker effects in a data set of limited size leads to the problem of there not being enough degrees of freedom to fit all of the effects simultaneously via ordinary least squares (OLS). This is sometimes referred to as the 'large p, small n' problem. If the significance thresholds in stepwise regression are sufficiently relaxed, thereby allowing for a greater spectrum of QTL effects (e.g. additive x additive interactions) to be included in the model, it has been shown that prediction accuracies as high as 0.61 can be achieved using this method (Habier et al. 2007). This is still not ideal however, since in order to exploit the full potential of GWS we must make use of all available marker information.

The only way to use all marker information is to treat the markers as random effects within a BLUP or Bayesian framework. Meuwissen et al. (2001) used simulation to compare the accuracy of GEBVs using ridge regression BLUP (RR-BLUP) and two Bayesian methods called BayesA and BayesB. Whilst RR-BLUP assumes that marker effects are normally distributed with constant variance (Whittaker et al.

2000), both BayesA and BayesB assume slightly different forms of an inverted Chi-square prior distribution of marker effects. Although RR-BLUP was highly accurate ($\gamma TBV;EBV = 0.73$) BayesA and BayesB were clearly superior ($\gamma TBV;EBV = 0.80$ and 0.85 respectively), especially when QTL vary in magnitude. Other variations on the "Bayesian alphabet" have also been shown to be highly effective (Habier et al. 2011). For an in-depth discussion of Bayesian methodology in a breeding context see (Gianola et al. 2009).

More recently, Hayes et al. (2009a) proposed using dense marker information to predict realized relationship coefficients between pairs of individuals using BLUP. The method, which we will refer to as genomic BLUP (G-BLUP), is therefore analogous in principle to traditional pedigree BLUP in that it attempts to calculate in each case the proportion of the genome that is identical by descent (IBD). It is superior to pedigree BLUP because it calculates this value directly, rather than relying on an expectation derived through lineage and selection, and is therefore able to account for Mendelian sampling during gamete formation.

5.4.3 How Many Markers Do We Need?

Marker density affects both the prediction of individual marker effects in all forms of GWS and the estimation of realized relationship coefficients in G-BLUP. This is because the accuracy of prediction for both methods relies in large part, on the ability of markers to serve as proxies for QTL from generation to generation. So what is the ideal marker density?

Firstly, the ability of markers to act as proxies for QTL, is a function of average LD within the genome. The greater the expanse of LD in the genome, the fewer markers that will be required to tag QTL located in any particular region. Furthermore, for each generation that GWS is practiced, the proportion of genetic variance explained by each marker decreases and the accuracy of GWS will tend to decline for each successive generation that it is practiced (Muir 2007).

The rates of LD decay are known to vary considerable between species, depending on a range of population characteristics, including those affected by selection history (Gaut and Long 2003). From a genetic perspective, species LD will depend on the population recombination rate, $4N_e r$ where N_e is equal to the effective population size, and r is the recombination rate per base pair. If this is known, the target marker density for GWS can be approximated by using the average r^2 between adjacent markers as a measure of their marker density relative to the decay of LD (Calus and Veerkamp 2007; Heffner et al. 2009).

The importance of marker density can also be demonstrated in terms of the relationship between the effective population size N_e and the number of independent chromosome segments q, defined as $q = 2N_e \times L$ where L is equal to the total genomic map length in Morgans (Hayes et al. 2009b). Obviously if N_e is large, the number of 'independent' chromosome segments is also large, and the extent of LD in the population will be limited, requiring a very large number of markers to capture all QTL effects.

5.4.4 The Use of Low Density SNP Chips

Despite rapid advancements in genotyping technology, the cost of genotyping (including sample collection and DNA extraction) remains a potential limitation in the implementation of GWS in smaller breeding programs (Ibañez-Escriche and Gonzalez-Recio 2011). This is especially the case when the number of selection candidates per generation is high, or the economic benefit per selection candidate is low compared to the cost of genotyping (Habier et al. 2009).

One solution is to use a reduced set of markers in the selection candidates to reduce overall cost while minimising loss of accuracy. One adaptation of this strategy involves the use of variable selection methods to identify a small set of markers that are predictive of trait phenotype or breeding value. Although variable selection methods have been shown to have good predictive ability (Cleveland et al. 2010; Iwata and Jannink 2010; Vazquez et al. 2010; Weigel et al. 2009), this approach is less attractive for multiple trait selection and across populations since it requires specific SNPs for each trait and population (Ibaez-Escriche and Gonzalez-Recio 2011).

A more flexible (but less accurate) approach was proposed by Habier et al. (2009) who suggested using evenly spaced low-density markers to obtain GEBVs in the selection candidates. This way, co-segregation of high and low density SNPs within families can be used to impute missing SNP genotypes in the selection candidates. The method is similar to the use of TAG SNPs to identify haplotype blocks segregating across human populations (Servin and Stephens 2007), the primary difference being that Habier et al. (2009) used pedigrees to estimate haplotype blocks within families rather than across populations.

5.4.5 Training Population Size and Design

Another avenue to reduce genotyping costs is to limit the size of the training population through selective genotyping. Selective genotyping has been previously proposed to improve the efficiency of QTL detection in a linkage mapping context (Sun et al. 2010) and more recently in a GWS context (Zhao et al. 2012). Genotyping only those candidates with high or low phenotypic values of the target trait (bidirectional selection) has been shown to lead to only a marginal decrease in the prediction accuracy of genomic breeding values (Zhao et al. 2012) and may therefore be a useful way to conduct GWS under a restricted budget.

Further concessions can also be made in crops when conducting GWS within large bi-parental populations, since bi-parental populations have extensive LD and allow for complete genome coverage with only a few hundred markers (Heffner et al. 2011). In bi-parental GS, the training population is made up of a subset of the progeny and the resulting prediction models are then used for predicting genetic value of the remaining progeny or for subsequent cycles of marker assisted recurrent selection (Bernardo and Yu 2007).

5.4.6 Marker Assisted Recurrent Selection (MARS)

Compared to crop breeders, animal breeders have been more accepting of GWS, perhaps due to a long history of selecting on breeding values as a surrogate for aggregate performance (Nakaya and Isobe 2012). In actual fact, the concept of selecting on an index is not at all foreign to many modern plant breeders. Many breeders practice a 'simplified' version of GWS referred to as "marker assisted recurrent selection" or MARS (Eathington et al. 2007; Hospital and Charcosset 1997; Koebner 2003). Like GWS, MARS focuses on individual performance rather than marker genotypes and selects on "marker scores" rather than GEBVs.

MARS was made possible by Lande and Thompson (1990) who first derived optimal selection indices for the improvement of quantitative traits, using both molecular and phenotypic information. Simply put, MARS refers to the improvement of typically an F_2 population by a single cycle of MAS (based on phenotypic and marker scores) followed by three cycles of selection based on marker scores only (Bernardo and Yu 2007). Thus there are two distinct steps involved in the application of MARS: model selection and model estimation. The model selection step involves identifying F_2 or F_2 derived progeny with a high proportion of favorable alleles at target marker loci. Markers are typically chosen on the basis of a statistical test for significance. In the model estimation step, each marker is given a weight based on its estimated effect and individual candidates are then ranked based on selection index (molecular score). The selection index represents a prediction of genetic value of each line, much the same as a GEBV.

The primary benefit of MARS, as opposed to traditional MAS, is that once the linear model for estimation of breeding values has been derived, it can be used to predict the breeding value of other marker genotyped individuals within the population or in future generations. MARS has been widely used in plant breeding programs in bi-parental crosses (Eathington et al. 2007) but is now being replaced by approaches that utilize data compiled across the entire breeding program. As LD in successive generations is slowly eroded due to recombination, the prediction value of each marker decreases. With tight linkage between marker and trait loci, the breakdown of LD can be minimized and the model can be used for several cycles of selection. Theoretically, if all markers were to be attributed the same effect, MARS would actually be equivalent to F_2 enrichment (Bernardo 2008). However since there is variation in marker effects, an individual with two QTL with large effects can be prioritized for selection in MARS ahead of an individual with four QTL with small effects.

5.4.7 Maintaining Genetic Diversity

The loss of genetic diversity in elite breeding programs due to factors such as selection, small population sizes and genetic drift remains an issue for concern for plant breeders. The importance of genetic diversity can be thought of in terms of its role

in generating additive genetic variance (σA) and thus genetic gain. It may useful to think of the additive genetic variance as stored genetic potential, and selection as being the process through which this genetic potential is converted to genetic gain. This can be seen in the following equation:

$$R = ih^2\sigma_p$$

where R = response to selection, i = intensity of selection, h^2 = heritability, equal to $^{\sigma A}/_{\sigma_p}$ and σ_p equals the phenotypic standard deviation (Falconer and Mackay 1996).

Selection can alter the genetic variance in a population by either changing allele frequencies and/or generating linkage disequilibrium (Lynch and Walsh 1998). However depending on the type of selection being applied to the population, the genetic variance may either increase due to the generation of coupling disequilibrium (e.g. disruptive selection), or decrease due to the generation of repulsion disequilibrium (e.g. directional and stabilizing selection). This reduction in additive genetic variance, otherwise known as the Bulmer effect (Bulmer 1976), has been shown to adversely affect the response to both GWS and pedigree based BLUP selection (Van Grevenhof et al. 2012). The challenge for breeders is therefore to practice selection whilst preserving genetic variance or risk severely limiting their opportunities to force further adaptation in the long term.

One approach to this issue involves the introgression of new alleles for quantitative traits from unadapted germplasm. Plant breeders have been understandably hesitant of this approach primarily due to the risk of inadvertently breaking up favorable linkage blocks in elite germplasm. Many of these linkage blocks have been formed through the gradual accumulation of favorable genes linked in coupling over many generations. Reassembling favorable linkage blocks can be difficult, especially for small effect genes whose positive effects can be masked by the negative effects of other linked genes introduced from the unadapted parent (Jordan et al. 2011).

With the availability of high-density marker platforms to better understand the genetic architecture of both donor and recipient populations (Klein et al. 2008), it is becoming increasingly possible to utilize crossing strategies such as nested association mapping (Buckler et al. 2009; Yu et al. 2008) and modified backcrossing strategies (e.g. Jordan et al. 2011) to more quickly and efficiently deliver novel genetic segments into adapted cultivars. Indeed, high-density marker platforms are now being used extensively for assessments of genetic diversity in cereals (Chen et al. 2012; Pan et al. 2012), as well as in the discovery of allelic variants of known genes (Deschamps and Campbell 2010). Furthermore, with the ability to select directly on high-density markers, GWS allows for the option of preferentially weighting low frequency favorable markers in the early cycles of selection so as to avoid losing low frequency favorable QTL. Simulation has shown this to be a potentially useful way of maximizing response over the long term, whilst sacrificing little or no response in the short term (Jannink 2010).

5.5 GWS or MAS/MARS?

Since its inception in 2001, GWS has had a limited uptake in public plant breeding programs, although an internet search will show that large seed companies have been hiring many staff in these areas in the last 2-3 years. The slow uptake in public breeding is despite a growing body of evidence in both plants and animals suggesting that GWS is the most efficient way of making use of the availability of dense marker information. For example, Bernardo and Yu (2007) simulated the response due to GWS compared with MARS over three cycles of selection in maize and found the response due to GWS to be 18 to 43 % larger than the response due to MARS. Furthermore, through reduction in time and costs needed to prove the value of a bull, Schaeffer (2006) showed that GWS could provide a twofold increase in response to selection and save 92 % of the costs of the current progeny test based breeding programs. Finally, (Wong and Bernardo 2007) showed through simulation that the 19 year selection cycle in oil palm could be reduced to 6 years, and that GWS would outperform MARS and phenotypic selection on a gain per unit cost and time basis, even with very small population sizes ($N = 50$). It may be that some of the reticence to implement GWS is due to lack of demonstration of the realised results. Most of the published studies are related to prediction of target populations from training populations, rather than the realisation of recombination over several cycles of selection, and it will be some time before supporting evidence is accumulated in this area.

Further motivation for the adoption of GWS should lie in the fact that MAS has not performed up the expectations set on it over two decades ago. Since the first application of molecular markers to crop improvement in 1986, an interesting dichotomy has developed between the number of publications purporting to have found significant marker-trait associations and the number of publications describing the successful development of finished breeding products. By the year 2000, the number of publications containing the term “quantitative trait loci” outnumbered those that contained the term “marker assisted selection” on Google Scholar by a factor of 3 (Xu and Crouch 2008), and the gap appears to have widened since. Thus although the discovery of marker trait associations in crops has been successful (Price 2006), the application of this knowledge into developing new plant varieties has not, at least in the public sector. Potential reasons for this disparity are presented by Collard and Mackill (2008).

MAS and its various adaptations should be viewed collectively as interim methodologies, developed to make the best use of limited available information at a time when genotyping technologies were very much in their infancy. Much has changed since this time. Over the last 20 years, the number of base pairs sequenced per dollar has increased exponentially, and by extension, so has the amount of genomic information available for analysis. With the cost of DNA sequencing now dropping by half every 5 months (Stein 2010), DNA sequencing throughput is outpacing advancements in both computer speed and storage capacity. A major limitation for smaller breeding programs is that they need efficient and well-designed information systems

together with skilled staff to exploit these opportunities. Given these capabilities are available, it is therefore opportune for plant breeding programs to transit from traditional and MAS breeding to GWS and strive to exploit this influx of genomic information.

Acknowledgment This publication has been partially supported by external funding from the Generation Challenge Program and the Bill and Melinda Gates Foundation as part of the Integrated Breeding Platform project which supports training and implementation of marker technologies and open-source tools in public breeding programs (https://www.integratedbreeding.net/).

References

Akkaya MS, Bhagwat AA, Cregan PB (1992) Length polymorphisms of simple sequence repeat DNA in soybean. Genetics 132:1131–1139

Beavis WD (1998) The power and deceit of QTL experiments: lessons from comparative QTL studies. 49th annual corn and sorghum industry research conference. ASTA, Washington, pp 145–162

Beckmann JS, Soller M (1986) Restriction fragment length polymorphisms in plant genetic improvement. Oxf Surv Plant Mol Cell Biol 3:196–250

Bernardo R (2008) Molecular markers and selection for complex traits in plants: learning from the Last 20 Years. Crop Sci 48:1649–1664

Bernardo R, Yu J (2007) Prospects for genome-wide selection for quantitative traits in maize. Crop Sci 47:1082–1090

Bonnett DG, Rebetzke GJ, Spielmeyer W (2005) Strategies for efficient implementation of molecular markers in wheat breeding. Mol Breed 15:75–85

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

Buckler ES, Holland JB, Bradbury PJ et al (2009) The genetic architecture of maize flowering time. Science 325:714–718

Bulmer MG (1976) The effect of selection on genetic variability: a simulation study. Genet Res 28:101–117

Calus M, Veerkamp RF (2007) Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. J Anim Breed Genet 124:362–368

Chen S, Lin XH, Xu CG, Zhang Q (2000) Improvement of bacterial blight resistance of 'Minghui 63', an elite restorer line of hybrid rice, by molecular marker-assisted selection. Crop Sci 40:239–244

Chen X, Min D, Yasir TA, Hu Y-G (2012) Genetic diversity, population structure and linkage disequilibrium in elite chinese winter wheat investigated with SSR markers. PLoS ONE 7:e44510

Cleveland M, Forni S, Deeb N, Maltecca C (2010) Genomic breeding value prediction using three bayesian methods and application to reduced density marker panels. BMC Proc 4:S6

Collard BCY, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos T T Soc A 363:557–572

Deschamps S, Campbell MA (2010) Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. Mol Breed 25:553–570

Deschamps S, Llaca V, May GD (2012) Genotyping-by-sequencing in plants. Biology 1:460–483

Eathington SR, Crosbie TM, Edwards MD et al (2007) Molecular markers in a commercial breeding program. Crop Sci 47:S154–S163

Elshire RJ, Glaubitz JC, Sun Qi et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for High diversity species. PLoS ONE 6:e19379
Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics. Benjamin Cummings
Gaut BS, Long AD (2003) The lowdown on linkage disequilibrium. Plant Cell 15:1502–1506
Gianola D, De Los Campos G, Hill WG et al (2009) Additive genetic variability and the Bayesian alphabet. Genetics 183:347–363
Goddard ME, Hayes BJ (2007) Genomic selection. J Anim Breed Genet 124:323–330
Gupta PK, Rustgi S, Mir RR (2008) Array-based high-throughput DNA markers for crop improvement. Heredity 101:5–18
Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397
Habier D, Fernando RL, Dekkers JCM (2009) Genomic selection using low-density marker panels. Genetics 182:343–353
Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12:186
Hayes BJ, Visscher PM, Goddard ME (2009a) Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res 91:47
Hayes BJ, Daetwyler HD, Bowman P et al (2009b) Accuracy of genomic selection: comparing theory and results. In: Proceedings of the 18th conference: association for the advancement of animal breeding and genetics, Barossa Valley, South Australia, pp 34–37
Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. Crop Sci 49:1–12
Heffner EL, Jannink J-L, Iwata H et al (2011) Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Sci 51:2597–2606
Henderson CR (1976) A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69–83
Hillel J, Schaap T, Haberfeld A et al (1990) DNA fingerprints applied to gene introgression in breeding programs. Genetics 124:783–789
Holland JB, Nyquist WE, Cervantes-Martínez CT (2010) Estimating and interpreting heritability for plant breeding: an update. In: Janick J (ed) Plant breeding reviews, Wiley, Oxford, pp 9–112
Hooker AL (1977) A plant pathologist's view of germplasm evaluation and utilization. Crop Sci 17:689–694
Hospital F, Charcosset A (1997) Marker-assisted introgression of quantitative trait loci. Genetics 147:1469–1485
Huang N, Angeles ER, Domingo J et al (1997) Pyramiding of bacterial blight resistance genes in rice: marker-assisted selection using RFLP and PCR. Theor Appl Genet 95:313–320
Ibañez-Escriche N, Gonzalez-Recio O (2011) Review promises, pitfalls and challenges of genomic selection in breeding programs. Span J Agric Res 9:404–413
Iwata H, Jannink J-L (2010) Marker genotype imputation in a low-marker-density panel with a high-marker-density reference panel: accuracy evaluation in barley breeding lines. Crop Sci 50:1269–1278
Jannink J-L (2010) Dynamics of long-term genomic selection. Genet Sel Evol 42:35
Jordan DR, Mace ES, Cruickshank AW et al (2011) Exploring and exploiting genetic variation from unadapted sorghum germplasm in a breeding program. Crop Sci 51:1444–1457
Klein RR, Mullet JE, Jordan DR et al (2008) The effect of tropical sorghum conversion and inbred development on genome diversity as revealed by high-resolution genotyping. Crop Sci 48:S12–26
Koebner R (2003) MAS in cereals: green for maize, amber for rice, still red for wheat and barley. In: Marker assisted selection: A fast track to increase genetic gain in plant and animal breeding? Turin, Italy. 17–18 Oct 2003
Lamkey KR, Lee M (2006) Plant breeding: the Arnel R Hallauer international symposium. Wiley-Blackwell
Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756

Lebowitz RJ, Soller M, Beckmann JS (1987) Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. Theor Appl Genet 73:556–562

Li W, Zhang P, Fellers JP, Friebe B, Gill BS (2004) Sequence composition, organization, and evolution of the core Triticeae genome. Plant J 40:500–511

Lynch M, Walsh B (1998) Genetics and analysis of quantitative Traits. Sinauer Associates

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

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. PNAS 88:9828–9832

Moreau L, Charcosset A, Gallais A (2004) Experimental evaluation of several cycles of marker-assisted selection in maize. Euphytica 137:111–118

Muir WM (2007) Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J Anim Breed Genet 124:342–355

Nakaya A, Isobe SN (2012) Will genomic selection be a practical method for plant breeding? Ann Bot 110:1303–1316

Okada Y, Kanatani R, Arai S, Ito K (2004) Interaction between barley yellow mosaic disease-resistance genes *rym1* and *rym5*, in the response to BaYMV strains. Breed Sci 54:319–325

Pan Q, Ali F, Yang X et al (2012) Exploring the genetic characteristics of two recombinant inbred line populations via high-density SNP markers in maize. PLoS ONE 7:e52777

Pflieger S, Lefebvre V, Causse M (2001) The candidate gene approach in plant genetics: a review. Mol Breed 7:275–291

Piepho HP (1997) Analyzing genotype-environment data by mixed models with multiplicative terms. Biometrics 761–766

Price AH (2006) Believe it or not, QTLs are accurate!. Trends Plant Sci 11:213–216

Sanger F, Coulson AR, Friedmann T et al (1978) The nucleotide sequence of bacteriophage φX174. J Mol Biol 125:225–246

SanMiguel P, Tikhonov A, Jin YK et al (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science 274:765–768

Schaeffer LR (2006) Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet 123:218–223

Servin B, Stephens M (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 3:e114

Sharma HC, Gill BS (1983) Current status of wide hybridization in wheat. Euphytica 32:17–31

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145

Stein L (2010) The case for cloud computing in genome informatics. Genome Biol 11:207

Sun Y, Wang J, Crouch JH, Xu Y (2010) Efficiency of selective genotyping for genetic analysis of complex traits and potential applications in crop improvement. Mol Breed 26:493–511

Toojinda T, Tragoonrung S, Vanavichit A et al (2005) Molecular breeding for rainfed lowland rice in the mekong region plant production. Science 8:330–333

Trebbi D, Maccaferri M, Heer P de et al (2011) High-throughput SNP discovery and genotyping in durum wheat (*Triticum durum* Desf). Theor Appl Genet 123:555–569

Van Eeuwijk FA, Bink MC, Chenu K, Chapman SC (2010) Detection and use of QTL for complex traits in multiple environments. Curr Opin Plant Biol 13:193–205

Van Grevenhof EM, Van Arendonk JA, Bijma P (2012) Response to genomic selection: the bulmer effect and the potential of genomic selection when the number of phenotypic records is limiting. Genet Sel Evol 44:26

Van Orsouw NJ, Hogers RCJ, Janssen A et al (2007) Complexity reduction of polymorphic sequences (CRoPS^{TM}): a novel approach for large-scale polymorphism discovery in complex genomes. PLoS ONE 2:e1172

Vazquez AI, Rosa GJM, Weigel KA et al (2010) Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. J Dairy Sci 93:5942–5949

Visscher PM, Haley CS, Thompson R (1996) Marker-assisted introgression in backcross breeding programs. Genetics 144:1923–1932

Vos P, Hogers R, Bleeker M et al (1995) AFLP: a new technique for DNA fingerprinting. Nucl Acids Res 23:4407–4414

Wang J, Chapman SC, Bonnett DG et al (2007) Application of population genetic theory and simulation models to efficiently pyramid multiple genes via marker-assisted selection. Crop Sci 47:582–588

Wang J, Chapman S, Bonnett D, Rebetzke G (2009) Simultaneous selection of major and minor genes: use of QTL to increase selection efficiency of coleoptile length of wheat (*Triticum aestivum L*). Theor Appl Genet 119:65–74

Weigel KA, De Los Campos G, González-Recio O et al (2009) Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J Dairy Sci 92:5248–5257

Whittaker JC, Thompson R, Denham MC (2000) Marker-assisted selection using ridge regression. Genet Res 75:249–252

Williams JGK, Kubelik AR, Livak KJ et al (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucl Acids Res 18:6531–6535

Wong C, Bernardo R (2007) Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. Theor Appl Genet 116:815–824

Xu S (2003) Theoretical basis of the beavis effect. Genetics 165:2259–2268

Xu Y, Crouch JH (2008) Marker-assisted selection in plant breeding: from publications to practice. Crop Sci 48:391–407

Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178:539–551

Zhao Y, Gowda M, Longin F et al (2012) Impact of selective genotyping in the training population on accuracy and bias of genomic selection. Theor Appl Genet 125:707–713