Review

CellPress

Genomic selection: genome-wide prediction in plant improvement

Zeratsion Abera Desta and Rodomiro Ortiz

Department of Plant Breeding, Swedish University of Agricultural Sciences, Sundsvagen 14, Box 101, Alnarp, SE 23053, Sweden

Association analysis is used to measure relations between markers and quantitative trait loci (QTL). Their estimation ignores genes with small effects that trigger underpinning quantitative traits. By contrast, genome-wide selection estimates marker effects across the whole genome on the target population based on a prediction model developed in the training population (TP). Whole-genome prediction models estimate all marker effects in all loci and capture small QTL effects. Here, we review several genomic selection (GS) models with respect to both the prediction accuracy and genetic gain from selection. Phenotypic selection or marker-assisted breeding protocols can be replaced by selection, based on whole-genome predictions in which phenotyping updates the model to build up the prediction accuracy.

Genomic selection will revolutionize the applications of plant and tree breeding

Marker-assisted selection (MAS; see Glossary) has been used in plant improvement programs since the 1990s, after promising research results for tagging genes or mapping QTL. MAS and association genetics have been used in the detection of underlying major genes in gene pools and in their introgression to improve traits of major crop breeding programs. Nevertheless, they have shown some shortcomings due to long selection cycles and the search for significant marker–QTL associations being unable to capture 'minor' gene effects [1–3].

The introduction of GS [4] has paved the way to overcome these limitations using whole-genome prediction models. The use of high-density markers is one of the fundamental features of GS. Therefore, every trait locus has the probability of being in linkage disequilibrium (LD) with a minimum of one marker locus in the entire target population. Genome-wide selection removes the need to search for significant QTL–marker loci associations individually. Rather, GS accounts for bunches of predictors simultaneously and is characterized by constraining random estimates towards zero. Moreover, GS can accelerate breeding cycles in such a way that the rate of annual genetic gain per unit of time and cost can be enhanced [5].

Corresponding author: Ortiz, R. (rodomiro.ortiz@slu.se).

Keywords: accuracy; breeding cycle; genetic gain; genomic selection; prediction models.

1360-1385/

 http://dx.doi.org/10.1016/j.tplants.2014.05.006

Glossary

Best linear unbiased prediction (BLUP): a statistical approach used to estimate the breeding values of different traits.

Breeding population (BP): the descendants of a TP or introduced variety but related to the TP, in which they are only genotyped but not phenotyped.

Breeding value: the average effects of alleles in the entire loci that are anticipated to be transferred from the parent to the progeny. The breeding value measures how many of the superior alleles or genes are transferred to the progeny.

Cross-validation: a method used to train and develop the prediction model(s) using different sampling techniques in the TP data sets ahead of estimating the GEBVs in the BP. The greater the similarity of the correlation of the two subsets (training set and validation set) to the correlation of the true breeding values in the TP to the expected GEBVs in the BP, the higher the precision and reliability of the prediction model(s).

Double haploids (DH): synthesis of genotypes after the haploid cells have undergone artificial chromosome doubling.

Genetic distance: measurement of relatedness or dissimilarity between samples or populations. The larger the value of genetic distances between samples, the more divergent the samples.

Genetic value: a cumulative effect of genes in the entire loci that affects the performance of the trait. It includes the additive effect and the dominance effect of an allele. In the absence of dominance, genetic value is equal to breeding value.

Genotype × environment (G × E) interaction: estimates (ranks) the differential reaction of the genotype in terms of stability and performance across seasonal and environmental conditions.

Genomic estimation of breeding value (GEBV): the estimation of genotyped populations using statistical model(s) to further predict the breeding values of future phenotypes in the target species.

Genomic selection (GS): estimates marker effects across the whole genome of the target population based on two distinct but related groups, the so-called training and breeding populations. The selection decision will be made on the breeding population depending on the outcomes of breeding values.

Heritability: the degree of genetic variance that affects a phenotypic trait.

High-throughput phenotyping: recording of agro-morphological and physiological traits using image and computer algorithms.

Imputation: computation of missed genotypic data using various statistical methods.

Inbreeding depression: loss of hybrid vigor resulting from the expression of deleterious recessive alleles. This phenomenon affects severely outbreeding species.

Linkage disequilibrium (LD): the nonrandom association of alleles at different loci in a population.

Marker-assisted selection (MAS): a type of indirect selection based on a significant association between a marker and variation for target trait.

Population structure: the formation and distribution of gene pools in a defined population. Analysis of genetic variation among and within a population is the key to determining the extent and degree of variation in the population structure.

Quantitative trait loci (QTL): DNA segments carrying genes controlling quantitative traits.

Rare alleles: alleles with a frequency below or equal to 1% of the population. These can be deleterious or favorable alleles.

Sequencing: the determination of sequential arrangement of nucleotides along the DNA or RNA of any species.

Single nucleotide polymorphism (SNP): DNA sequence variation arising from pairwise differences in nucleotide(s) of the genome between individuals of same species.

Training population (TP): a group of individuals from a population (such as half-sibs or lines) that are both phenotyped and genotyped.

GS has long been practiced in the field of animal breeding, but is in its infancy in crop [1,6,7] and forest tree [8,9] breeding. Genome-wide selection or GS estimates marker effects across the whole genome of the breeding population (BP) based on the prediction model developed in the TP (Figure 1). TP is a group of related individuals (such as half-sibs or lines) that are both phenotyped and genotyped. BP usually includes the descendants of a TP or a new variety that is related to the TP, and is only genotyped not phenotyped. Hence, GS relies on the degree of genetic similarity between TP and BP in the LD between marker and trait loci.

GS identifies the highest genomic estimated breeding values (GEBVs) instead of novel gene(s) in the target species. Given that many of the selections are replaced by selection on predictions, phenotyping can be considered as a key informant in GS to build up the accuracy of statistical models. MAS [10], marker-assisted recurrent selection (MARS) [11], and gene pyramiding [12] are still important methods of selection to identify and further incorporate novel gene(s) in recurrent parents. These methods can be complemented with GS in integrated plant breeding programs (Figure 1). Therefore, with the advent of cutting-edge next-generation sequencing (NGS) and high-throughput phenotyping tools, GS may revolutionize practical applications of crop and forest tree improvement programs.

In this review, we discuss estimating GEBV, the accuracy and gain of selection using genome-wide prediction models, compare GS versus other selection methods of plant breeding, and provide an outlook of GS in plant breeding schemes.

Prediction models

Plant breeding is a science of prediction. Various types of prediction model respond differently because they vary in their assumption(s) when treating the variance of complex traits. The standard linear model equation can be formulated as (Equation 1):

$$y = \mu + \sum_k \chi_k \beta_k + e, \quad [1]$$

where y is a vector of trait phenotype, μ is an overall phenotype mean, k represents the locus, χ_k is the allelic state at the locus k, β_k is marker effect at the locus k, and $e \sim N(0, \sigma_e^2)$ where e is the vector of random residual effects and σ_e^2 is the residual variance. In χ_k, the allelic state of individuals can be coded as a matrix of 1, 0, or -1 to a diploid genotype value of AA, AB, or BB, respectively.

The number of predictors (p) is usually far greater than the number of individuals (n). In such cases, estimates of ordinary least-squares (OLS) have a poor predictive ability because marker effects are treated as fixed effects, which leads to multicolinearity and overfitting among predictors, thereby making the model infeasible. The advent of GS [4] provides an opportunity to confront these challenges using alternative models, such as whole-genome regressions (Table 1, Figure 2). Parametric and nonparametric models can cluster whole-genome regression methods.

Accuracy assessments of genomic selection in crop and tree breeding

The performance of GS depends on the prediction accuracy to select individuals whose phenotype is unknown. In GS, the GEBV can be computed from Equation 1 as (Equation 2):

$$GEBV = x_{new}\hat{\beta}_k, \quad [2]$$

where x_{new} is a matrix comprising the allelic states of individuals in a BP, and $\hat{\beta}_k$ is the estimate of the regression coefficient of β_k.

Cross-validation is used to train and develop the prediction model in the TP (Figure 3A). Then, the best-fitted model can be used to further evaluate the GEBV in a BP (Figure 3B). Therefore, the prediction of GEBVs should mimic the alternatives of cross-validation strategies [13].

Prediction accuracy (r_A) is the Pearson's correlation (r) between the selection criterion (GEBV) and the true breeding value (TBV) (Figure 3B). The expected prediction accuracy (r_A) can be computed as in [14] (Equation 3):

$$r_A = \sqrt{\frac{h^2}{h^2 + \frac{M_e}{N_p}}}, \quad [3]$$

where h^2 is the narrow sense heritability, N_p is the number of individuals in a TP, and M_e is the number of independent chromosome segments, which depends on both the effective population size (N_e) and the genome length in Morgan (L) that was derived in [15] as $M_e \approx 2N_eL$. Ideally, M_e is related to the effective number of QTL. The combined use of both N_p and h^2, rather than their individual assessment, is key to regulating the expected prediction accuracy [14,16]. This is more pronounced when dealing with low trait heritability, where increasing the number of individuals in the TP may maintain the reduction in the expected prediction accuracy. In this situation, a higher N_p than M_e leads to a reduction in the value of $\frac{M_e}{N_p}$, thereby increasing prediction accuracy.

Factors affecting the prediction accuracy of GS models

The response of GS is the output of various factors affecting the accuracy of GEBVs. These factors are interrelated in a complex and comprehensive manner. They include model performances, sample size and relatedness, marker density, gene effects, heritability and genetic architecture, and the extent and distribution of LD between markers and QTL.

Model performances

Accuracy varies among GS models according to their assumptions and treatments of marker effects (Table 1). For example, it has been established that both Bayesian least absolute shrinkage and selector operator [Bayesian LASSO (BL)] and ridge regression (RR) models outperform support vector regression for predicting GEBVs for host plant resistance to wheat rusts [17], because these traits are controlled by additive gene effects. Another study compared 11 GS models on wheat (*Triticum aestivum*), maize (*Zea mays*), and barley (*Hordeum vulgare*) and all models, except the support vector machine, recorded similar average prediction accuracies using cross-validation [18]. In this study, cluster analysis of the GS models using Euclidean distance led to separate groupings of nonparametric versus parametric regressions.