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].

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.

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

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):

$$y = \mu + \sum\nolimits_k \chi_k \beta_k + e, \qquad [1]$$

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):

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

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):

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

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.

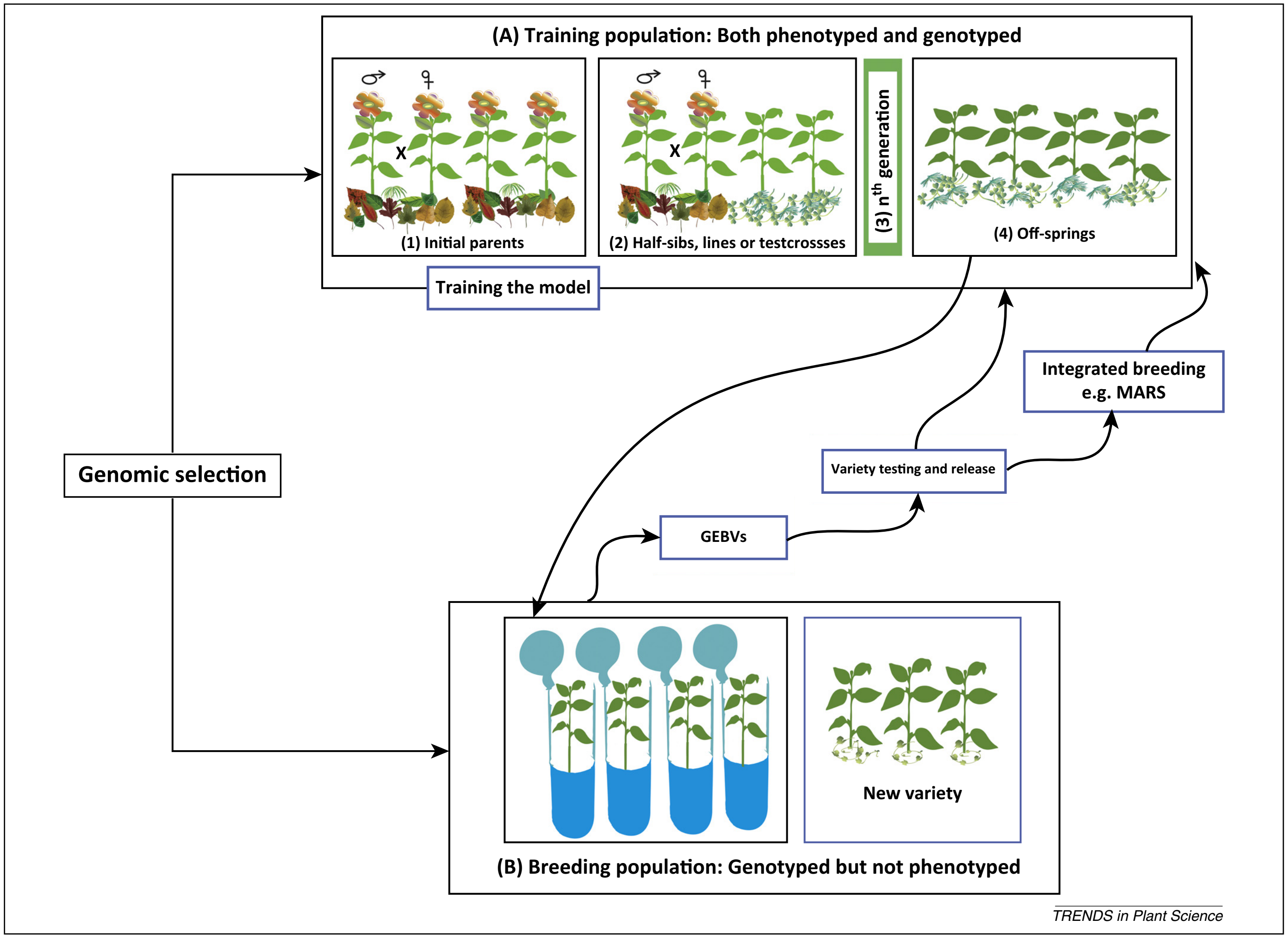
(A) Training population: Both phenotyped and genotyped
♂
♀
X
(1) Initial parents
♂
♀
X
(2) Half-sibs, lines or testcrossses
(3) n th generation
(4) Off-springs
Training the model
Genomic selection
Integrated breeding e.g. MARS
Variety testing and release
GEBVs
New variety
(B) Breeding population: Genotyped but not phenotyped
TRENDS in Plant Science

Figure 1. Genomic selection (GS) steps and applications in plant breeding. **(A)** The training population (TP) is the population set being phenotyped and genotyped. The initial parents used to produce the next generation (1) by crossing selected parents, half-siblings, lines, or test crosses are included in (2), and continue until the *n*th generation (3) that delivers offspring to be used as a validation set to train the model against the training sets in the TP (4). **(B)** Breeding populations (BP) are only genotyped but not phenotyped. These can also comprise new varieties introduced as BP but related to the TP. The breeding lines with highest genomic estimated breeding values (GEBVs) are selected and this will routinely continue as a turn cycle of GS to the TP. The selected candidates with high GEBVs can be integrated with other breeding schemes, such as marker-assisted recurrent selection (MARS) to introgress the required agro-morphological trait(s) to well-adapted crop species.

Table 1. Main features of genome-wide prediction models

Model acronym[a]	Features	Refs
RR-BLUP	Assumes that all markers have equal variances with small but non-zero effect	[4,19]
	Applies homogeneous shrinkage of predictors towards zero, but allows for markers to have uneven effects	
	Computed from a realized-relation matrix based on markers	
	Some QTL are in LD to marker loci, whereas others are not	
LASSO	Combines both shrinkage and variable selection methods	[70,71]
	RR-BLUP does not use variable selection, but outsmarts LASSO when there is multicolinearity between the predictors	
EN	Double regularization using ℓ_1 and ℓ_2 penalty norms combines the merited features of these norms to confront the challenge of high-dimensional data	[72]
BRR	Induces homogeneous shrinkage of all marker effects towards zero and yields a Gaussian distribution of marker effects	[73]
	Similar to RR-BLUP, there is a problem of QTL linkages to the marker lloci	
BL	Applies to both shrinkage and variable selection	[71,74]
	Has an exponential prior on marker variances resulting in a double exponential (DE) distribution	
	The DE distribution has a higher mass density at zero and heavier prior tails compared with a Gaussian distribution	
Bayes A	Utilizes an inverse chi-square (χ^2) on marker variances yielding a scaled *t*-distribution for marker effects	[4,74]
	Similar to BL and in contrast to BRR, it shrinks tiny marker effects towards zero and larger values survive	
	Has a higher peak of mass density zero compared with the DE distribution	
Bayes B	Similar to Bayes A, uses an inverse χ^2 resulting in a scaled *t*-distribution	[4,20]
	Unlike Bayes A, utilizes both shrinkage and variable selection methods	
	When π = 0, then it is similar to Bayes A	
Bayes C	Applies both shrinkage and variable selection methods	[74–76]
	Characterized by a Gaussian distribution	
	Bayes B and Bayes C consist of point of mass at zero in their slab priors	
Bayes Cπ	A modified variant of Bayes B	[75]
	Used to alleviate the shortcomings of Bayes A and Bayes B	
	Unlike Bayes B, π is not fixed, but estimated from the data	
RKHS	Based on genetic distance and a kernel function with a smoothing parameter to regulate the distribution of QTL effects	[77,78]
	Effective for detecting nonadditive gene effects	
RF	Uses the regression model rooted in bootstrapping sample observations	[55,66,79]
	Takes the average of all tree nodes to find the best prediction model	
	Captures the interactions between markers	

[a]EN, elastic net; RF, random forest; RHKS, reproducing kernels Hilbert spaces regression.

Sample size and relatedness

Generally as sample size increases prediction accuracy increases even though other influencing factors are crucial to consider. An increase in the TP of a biparental wheat population (TP comprising of 96, 48, and 96) [19] and a multifamily wheat breeding program (TP comprising of 96, 192, and 256) [20] resulted in an increase in the prediction accuracy.

Designing the composition of the TP in relation to the BP is important in maintaining a high degree of accuracy in GS. A few studies have shown that merging different groups of related populations enhances selection accuracy [21,22]. Combining multiple groups as part of the TP attained maximum and statistically significant prediction accuracy compared with a single group in both dent and flint traits of maize [23]. Studies in oat (*Avena sativa*) [24], maize [25], and sugar beet (*Beta vulgaris*) [26] showed similar results, whereas a study with barley found that the use of combined groups did not respond as expected [27]. In a biparental-crossing maize breeding population, the incorporation of half-sib representatives from both parents, rather than increasing the number of individuals arbitrarily, in the TP led to an increase in the prediction accuracy [28,29].

The formation of the population structure can influence the performance of genomic-wide predictions in stratified populations [24,30–32]. Research in a maize breeding program showed a very low prediction performance for dissimilar subpopulations [32]. Similarly, the accuracy of RR best linear unbiased predictor (RR-BLUP) declined as the genetic distance increased between TP and BP [26]. The presence of genetically dissimilar subpopulations in the TP resulted in insignificant prediction accuracy regardless of high marker application.

GS has a dual effect by estimating trait-marker effects based on a relation matrix and generating predictions for the target population. Hence, in the presence of population structure, GS has the ability to identify the extent of relations of individuals within and between subpopulations. However, to secure accuracy across subpopulations, it is better to design the TP by pooling multiple subpopulations of stable LD between markers and QTL [21,24].

Testing multiple traits across multienvironments is one impediment to the improved predictive ability of the GS models. Genotype × environment interaction (G×E) models based on phenotypes or markers may improve the prediction accuracy [17,33–36]. In this context, 2437 winter wheat lines were genotyped using 1287 single

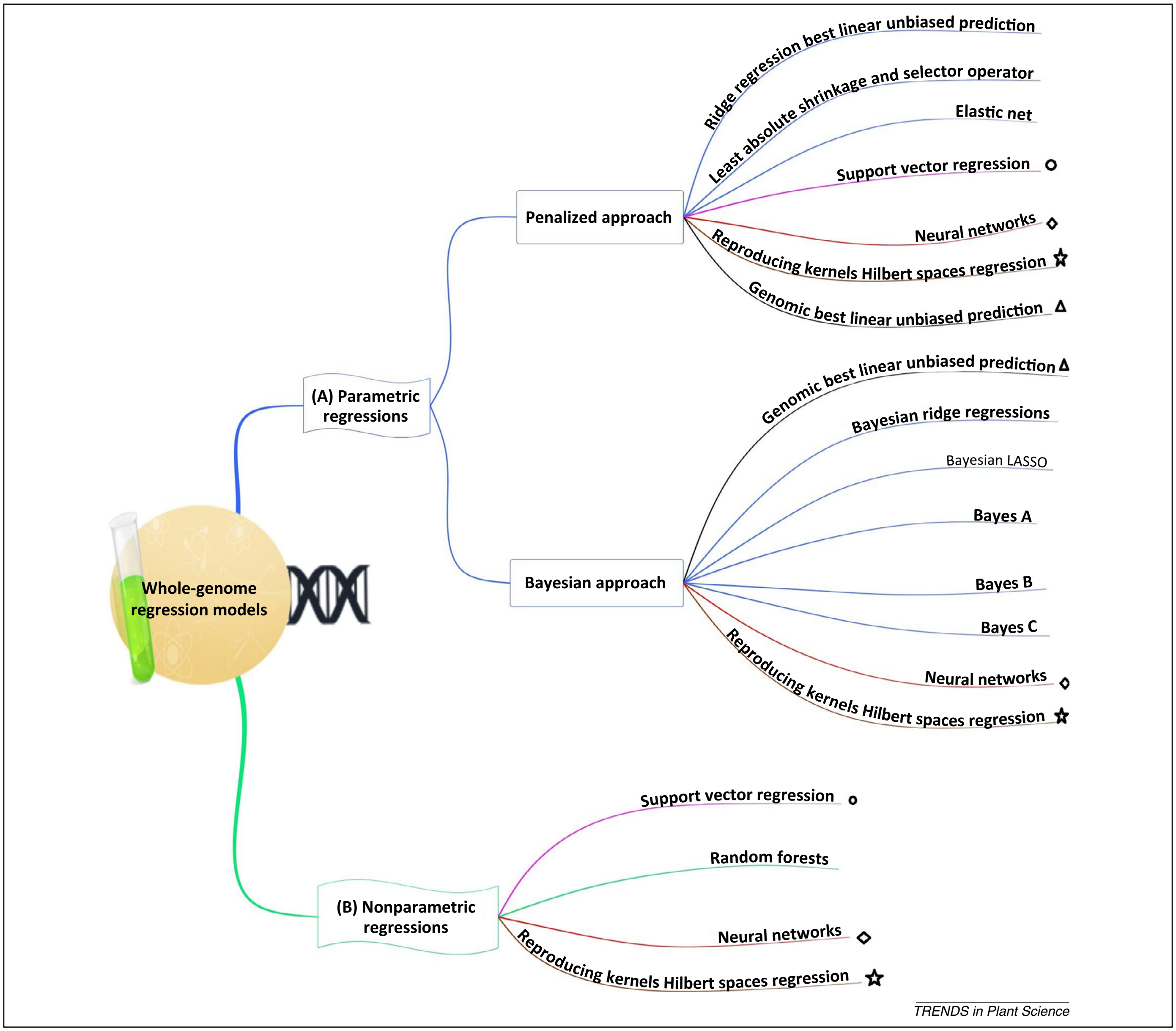
Whole-genome regression models
(A) Parametric regressions
Penalized approach
Ridge regression best linear unbiased prediction
Least absolute shrinkage and selector operator
Elastic net
Support vector regression ○
Neural networks ◇
Reproducing kernels Hilbert spaces regression ☆
Genomic best linear unbiased prediction △
Bayesian approach
Genomic best linear unbiased prediction △
Bayesian ridge regressions
Bayesian LASSO
Bayes A
Bayes B
Bayes C
Neural networks ◇
Reproducing kernels Hilbert spaces regression ☆
(B) Nonparametric regressions
Support vector regression ○
Random forests
Neural networks ◇
Reproducing kernels Hilbert spaces regression ☆
TRENDS in Plant Science

Figure 2. Classification of whole-genome regression models. Broadly, these models are categorized as parametric regressions **(A)** (in blue) and nonparametric regressions **(B)** (in green). Models that are indicated by multiline colors are additionally tagged with symbols for further identification because they are classified in different whole-genome regressions. Blue is used to show the following models: ridge regression best linear unbiased prediction, Bayes A, Bayes B, Bayes C [75], Bayesian ridge regression [4], least absolute shrinkage and selector operator [80], and elastic net [81]. Support vector machines [82] are coded in a pink with an oval; Neural networks [83] are coded with red and a diamond; reproducing kernels Hilbert spaces regression [84] is coded in brown with a star; genomic best linear unbiased prediction [85] is coded with black and a triangle; and random forest [86] is coded in green.

nucleotide polymorphisms (SNP) in 44 environments to predict an unknown environment using cross-validation [37]. The authors used 22 environments as a training set and 22 environments as a validation set and the best model resulted in average increase of 11.1% in accuracy. GEBVs cannot explain the impact of correlated environmental interactions when avoiding G×E interactions, especially in multienvironment trials. Therefore, the G×E effects could bias prediction accuracy in GS [11,38,39].

Marker density

Increasing marker density ensures the conservation of marker–QTL associations and achieves a high prediction accuracy. Marker density is mainly determined by the LD span and sample size. Maize has a shorter LD span compared with barley or wheat and, therefore, a higher marker density is preferred for maize than for both these small grain cereals. Research in a biparental bread wheat population genotyped with 485 markers showed that accuracy plateaued with a minimum number of markers (128–256), beyond which accuracy started to decline [19]. Accuracy plateaued when 800 markers were used for genotyping elite maize populations [22]. The effect of marker density to secure optimal prediction accuracy follows similar trends across species. For example, prediction accuracy for height in humans (with a short LD span) increased rapidly with marker density (approximately 150 000 markers) but plateaued at between 200 000 and 400 000 markers.

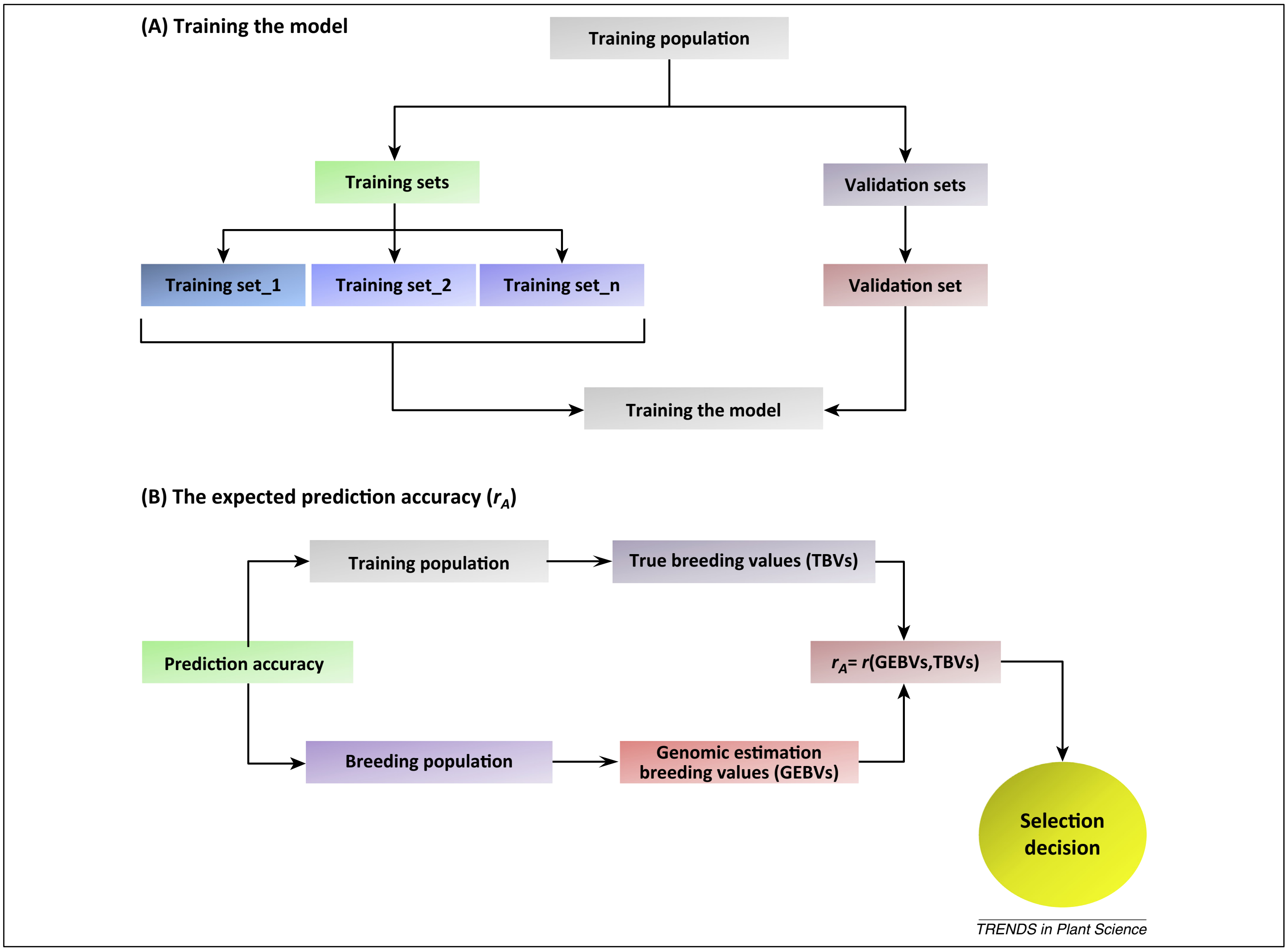

Figure 3. Training the model and prediction accuracy. **(A)** Training the model: different groups of the training population (TP) are represented as training sets to correlate against the validation set. The training sets along with validation set are used for cross-validation with *K*-folds to train the prediction models. **(B)** Prediction accuracy: the selected prediction model(s) are used to estimate the expected prediction accuracy or genomic estimated breeding values (GEBVs) in the selected candidates of the target species. The highest GEBVs are selected as a result of the selection decision.

The minimum number of markers across a family can be determined by $N_e \times L$, where N_e and L are the effective population size and genome size in Morgan, respectively [40]. Nevertheless, in biparental GS, the marker number needed is reduced vis-à-vis multifamily GS [20]. Doubling the TP size, rather than increasing marker density, may be preferable when dealing with related individuals in the TP and BP.

Gene effects

GS models should be assessed based on trait complexity and sample size. In this regard, a recent study estimated prediction accuracy of various models for days to maturity and grain yield in 306 elite wheat lines [41]. The study included penalized regressions [Bayesian ridge regression (BRR), BL, Bayes A, and Bayes B] versus nonlinear regressions [reproducing kernels Hilbert spaces regression (RKHS), Bayesian regularized neural networks (BRNN), and radial basis function neural networks (RBFNN)]. The authors noticed that nonlinear models had a maximum prediction accuracy higher than that of penalized models. Similar results were obtained when using RKHS and RBNN in maize against BL [42].

Nonlinear models could capture the nonadditive genetic effects (i.e., dominance and epistasis), making them suitable for improving the accuracy of GS models [33,43,44]. However, if the trait is additive, the use of nonparametric models may not yield the expected accuracy; for example, addition of dominance to additive effects decreased the accuracy of GEBV in hybrid wheat [45].

Heritability and genetic architecture

Ideally, high-heritability traits are positively correlated with higher GEBV prediction. Prediction accuracy for wheat resistance to yellow and stem rust was influenced by their low and high heritability, respectively [17]. Similar results were noted for grain yield (low heritability) versus grain moisture (high heritability) in maize, the respective accuracy of which was estimated to be 0.58 and 0.90, respectively [45]. Nevertheless, there are some irregularities and no compelling reason(s) to interpret results based on heritability only. For example, the prediction accuracy for flour protein content (heritability = 0.56) and sucrose solvent retention (heritability = 0.45) was 0.64 and 0.74, respectively in doubled-haploid biparental wheat lines [19].

The accuracy of GEBVs seems to be inversely related to the number of QTL [46]. Small numbers of major QTL effects are better suited to Bayesian regressions than to linear regressions. As the number of QTL per sample size decreased, Bayes B outperformed linear regression, whereas the opposite was the case with an increasing number of QTL per sample size [4,14,46]. The choice of whether to use variable selection approaches (e.g., LASSO, elastic net, and Bayes B) versus RR-BLUP often depends on the genetic architecture and heritability of the trait, as well as sample size. Under scenarios of low LD, high heritability, large sample size, and lower causal mutation relative to sample size variable selection methods had better accuracy than RR-BLUP [47]. However, this was reversed when these criteria were not met. By contrast, the accuracy of genomic BLUP (G-BLUP) was insensitive to QTL number and trait heritability [14]. Similarly, increasing QTL number had a minor impact on prediction accuracy in forest tree breeding [9,48] and in elite maize [49]. Bayes B had higher accuracy than G-BLUP with lower numbers of QTL, but declined and performed with lower accuracy compared with G-BLUP when the QTL number increased [14].

The extent of LD between marker and QTL

Absence of LD between markers and trait loci resulted in reduction of accuracy in future generations [33]. Using several loci or a whole genome may affect the trait but might have little impact unless they bear segregating QTL. QTL segregation is affected by the number of possible QTL and effective population size; for example, a small effective population size leads to few polymorphic loci and, thus, results in few polymorphic QTL.

Genomic selection versus other plant breeding methods

The use of genome-wide selection has increased significantly in animal breeding and is an emerging approach for plant improvement. Plant breeding for many crop species, unlike animal breeding, generates a large population size over a very short period of time; in addition, various mating designs can be implemented and it is easy to produce pure line hybrids or clones. However, G × E interactions remain a major issue to exploiting GS in plant breeding [50].

MAS only uses markers significantly linked to target traits, which are often qualitative. Association genetics utilizes LD mapping instead of testing and selecting to identify and validate alleles [51]. The success of a new variety depends on 'minor' genes of small effects that underlie most important traits, but these may be missed by MAS or when using association genetics. The use of GS [4] considers complex quantitative traits using whole-genome markers to predict GEBVs in the unknown phenotypes of target species.

Gain from selection (ΔG) is directly related to the selection intensity (i), prediction accuracy (r_A), and square root of additive genetic variance $\left(\delta_A^2\right)$, and is indirectly proportional to the breeding cycle time (Δt), represented as a general formula (Equation 4):

$$\Delta G = \frac{i r_A \delta_A}{\Delta t} \qquad [4]$$

Accuracy in GS is the key parameter because it is related to gains in selection and compares various GS models [52]. One example stems from a biparental wheat experiment comparing the phenotypic prediction accuracy with multiple linear regression (MLR) versus GS accuracy (using RR and Bayes Cπ models) [19]. The authors estimated the marker-based prediction accuracy and found that, in MLR and GS models, this ranged from 0.33 to 0.50 and from 0.44 to 0.68, respectively. Accuracy of the GS models outperformed MLR on average by 47%. Similarly, wheat breeding research on *Fusarium* head blight resistance revealed that the average GS accuracy rate was 57% higher than that of the MLR [53].

Accelerating the breeding cycles by reducing the time and cost associated with phenotyping would increase genetic gains in GS [5,54]. For example, three cycles of GS require three sets of genomic DNA harvesting before flowering. The faster the breeding cycle time, the higher the rate of annual genetic gain, especially for plants with long breeding cycles [8]. For instance, if GS reduces the breeding cycle time from 9 to 3 years, it means that GS is three times faster than phenotype selection (PS).

Relative annual genetic gains among GS, MAS, and PS plus MAS were assessed in maize and wheat multifamily populations [5]. This research showed that a GS accuracy of 0.53 led to threefold and twofold annual genetic gains relative to MAS and PS in maize and wheat, respectively.

The use of many cycles per year can accelerate recurrent selection [55] and intensify selection gains per unit time and cost over MAS and PS, particularly for complex traits associated with low heritability [56,57]. This superiority of GS in terms of genetic gain over PS has also been noticed in trees [9,38,39,48,58]. However, the gain of GS per cycle may not be better than PS. A study using empirical data from *Arabidopsis*, maize, and barley showed that gain from GS per cycle was half that of the gain from PS [59]. For example, when PS and GS accuracy rates are 0.9 and 0.3, respectively, then GS must be three times faster to match PS.

Concluding remarks

Breeding value has a key role in genome-wide prediction because it is helpful to select parents that produce superior progenies. In plant breeding, the commercial value of individuals usually depends on the performance of the genotype and not on their utilization as a parent. However, with the advent of GS, plant breeders have begun selecting parents based on the superiority of their progeny in addition to their genetic values. This may facilitate parental selection after some generation cycles instead of going back to the original parents. Nevertheless, the GS techniques established in animal breeding are difficult to apply directly to plant breeding programs [60]. Moreover, the comprehensive nature of population structure, especially in inbreeding species, is a major barrier to implementing GS in plant breeding.

Neither insertion-deletions (InDels) nor gene function are recognized when using GS. Other genomic tools, such as forward genetic approaches (e.g., map-based cloning), detect the unknown gene(s) from a known phenotype [61]. By contrast, the reverse genetic approaches [e.g., targeting induced local lesions in genomes (TILLING)] connect the polymorphisms in the unknown phenotype based on a predetermined gene of interest [62]. However, GS predicts GEBVs in the target species depending the phenotype and genotype effects on the TP using genome-wide markers (Figure 1). Therefore, GS cannot stand lonely or replace conventional breeding approaches, but instead can be used in integrated breeding programs.

It is impossible to determine the number and distribution of QTL and the heritability of the trait ahead of an experiment. However, the number of individuals and genotyping platforms in the TP can be adjusted in plant breeding. Therefore, thorough consideration should be given to the design and composition of TP based on BP and missing data input, especially with highly structured populations [63], to improve the reliability and credibility of GS models.

The level of accuracy achieved in GS lacks an easy biological interpretation, as noted in [14]. Therefore, it will be relevant to consider further studies on mechanisms to minimize computational complexity and handle the major QTL effects in GS. R statistical software (http://www.cran.r-project.org) can be used for predicting GEBVs, although it requires prior skills and is time consuming [64], particularly when sample size and marker density are large. The decay in accuracy, especially as the target species becomes further separated from the TP, is a crucial issue in GS [24]. Therefore, further research is required to determine the number of generation cycles between TP and BP

The cost of genotyping has declined dramatically, especially in the era of NGS [65], whereas the cost of phenotyping is increasing due to labor and land-use expenses. The application of GS does not eliminate phenotyping but replaces many of the selections associated with phenotyping based on whole-genome prediction. Phenotyping serves as a messenger to improve the accuracy of the prediction model. Hence, reliable and rapid high-throughput phenotyping that analyzes the whole plant improves the efficiency of GS [66–68]. Upgrading and updating image technology and its computational algorithms will be also required [68].

As selection cycles increase, the expected losses in diversity from GS are unavoidable [63]. Hence, measurements should be taken to circumvent the problem in tradeoffs between gain in selection and keeping diversity [69]. Rare alleles in a population can be lost when performing GS [63]. Importantly, GS can be adjusted to mitigate the risks of genetic losses in breeding programs. Thus, the introduction of new parental lines containing favorable agro-morphological traits in the cycles of selection [54,55] may compensate this drawback of GS. Furthermore, research on the implementation of GS in inbreeding species is needed. Trends of instability in their population structure and lack of polymorphisms impede the creation of new parents in GS generation cycles.

Few studies have been conducted to determine the implementation of GS for selection decision purposes in future phenotypes. Most GS research has given priority to prediction of the validation set using cross-validation. Most prediction accuracy in the validation set does not reflect the actual patterns of LD and population structure in the target species. Hence, more emphasis should be given to whole-genome prediction to capitalize GS in plant breeding. Moreover, in-depth research regarding the incorporation of G × E interactions into models may determine whether GS could be a better strategy to predict performance in new environments.

Acknowledgments

We thank the editor and anonymous referees for helpful comments on the manuscript. This work benefits from funding by Mistra–Stiftelsen för miljöstrategisk forskning and SLU.

References

1 Heffner, E.L. *et al.* (2009) Genomic selection for crop improvement. *Crop Sci.* 49, 1–12
2 Goddard, M.E. and Hayes, B.J. (2007) Genomic selection. *J. Anim. Breed. Genet.* 124, 323–330
3 Xu, Y. *et al.* (2012) Whole-genome strategies for marker-assisted plant breeding. *Mol. Breed.* 29, 833–854
4 Meuwissen, T.H.E. *et al.* (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829
5 Heffner, E.L. *et al.* (2010) Plant breeding with genomic selection: gain per unit time and cost. *Crop Sci.* 50, 1681–1690
6 Bernardo, R. and Yu, J. (2007) Prospects for genomewide selection for quantitative traits in maize. *Crop Sci.* 47, 1082–1090
7 Lorenz, A.J. *et al.* (2011) Genomic selection in plant breeding. *Adv. Agron.* 110, 77–123
8 Wong, C. and Bernardo, R. (2008) Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. *Theor. Appl. Genet.* 116, 815–824
9 Grattapaglia, D. and Resende, M.D.V. (2011) Genomic selection in forest tree breeding. *Tree Genet. Genomes* 7, 241–255
10 Collard, B.C. and Mackill, D.J. (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 363, 557–572
11 Crossa, J. *et al.* (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186, 713–724
12 Servin, B. *et al.* (2004) Toward a theory of marker-assisted gene pyramiding. *Genetics* 168, 513–523
13 Pérez-Cabal, M.A. *et al.* (2012) Accuracy of genome-enabled prediction in a dairy cattle population using different cross-validation layouts. *Front. Genet.* 3, 27
14 Daetwyler, H.D. *et al.* (2010) The impact of genetic architecture on genome-wide evaluation methods. *Genetics* 185, 1021–1031
15 Goddard, M. (2009) Genomic selection: prediction of accuracy and maximisation of long-term response. *Genetica* 136, 245–257
16 Combs, E. and Bernardo, R. (2013) Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. *Plant Genome* 6, 6
17 Ornella, L. *et al.* (2012) Genomic prediction of genetic values for resistance to wheat rusts. *Plant Genome J.* 5, 136–148
18 Heslot, N. *et al.* (2012) Genomic selection in plant breeding: a comparison of models. *Crop Sci.* 52, 146–160
19 Heffner, E.L. *et al.* (2011) Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Sci.* 51, 2597–2606
20 Heffner, E.L. *et al.* (2011) Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *Plant Genet.* 4, 65–75

21 Schulz-Streeck, T. *et al.* (2012) Genomic selection using multiple populations. *Crop Sci.* 52, 2453

22 Zhao, Y. *et al.* (2012) Accuracy of genomic selection in European maize elite breeding populations. *Theor. Appl. Genet.* 124, 769–776

23 Technow, F. *et al.* (2013) Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. *G3* 3, 197–203

24 Asoro, F.G. *et al.* (2011) Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. *Plant Genome J.* 4, 132–144

25 Ogutu, J.O. *et al.* (2012) Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proc.* 6 (Suppl. 2), S10

26 Würschum, T. *et al.* (2013) Genomic selection in sugar beet breeding populations. *BMC Genet.* 14, 85

27 Lorenz, A.J. *et al.* (2012) Potential and optimization of genomic selection for fusarium head blight resistance in six-row barley. *Crop Sci.* 52, 1609–1621

28 Riedelsheimer, C. *et al.* (2013) Genomic predictability of interconnected biparental maize populations. *Genetics* 194, 493–593

29 Jacobson, A. *et al.* (2014) General combining ability model for genomewide selection in a biparental cross. *Crop Sci.* 54, 895–905

30 Riedelsheimer, C. *et al.* (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat. Genet.* 44, 217–220

31 Wientjes, Y.C.J. *et al.* (2013) The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics* 193, 621–631

32 Windhausen, V.S. *et al.* (2012) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. *G3* 2, 1427–1436

33 Kumar, S. *et al.* (2011) Towards genomic selection in apple (*Malus* × *domestica* Borkh.) breeding programmes: prospects, challenges and strategies. *Tree Genet Genomes* 8, 1–14

34 Guo, Z. *et al.* (2013) Accuracy of across-environment genome-wide prediction in maize nested association mapping populations. *G3* 3, 263–272

35 Burgueño, J. *et al.* (2012) Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. *Crop Sci.* 52, 707–719

36 Ly, D. *et al.* (2013) Relatedness and genotype × environment interaction affect prediction accuracies in genomic selection: a study in cassava. *Crop Sci.* 53, 1312–1325

37 Heslot, N. *et al.* (2014) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theor. Appl. Genet.* 127, 463–480

38 Resende, M.F., Jr *et al.* (2012) Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. *New Phytol.* 193, 617–624

39 Resende, M.D. *et al.* (2012) Genomic selection for growth and wood quality in *Eucalyptus*: capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytol.* 194, 116–128

40 Meuwissen, T. (2009) Accuracy of breeding values of 'unrelated'individuals predicted by dense SNP genotyping. *Genet. Sel. Evol.* 41, 35

41 Perez-Rodriguez, P. *et al.* (2012) Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3 (Bethesda)* 2, 1595–1605

42 Gonzalez-Camacho, J.M. *et al.* (2012) Genome-enabled prediction of genetic values using radial basis function neural networks. *Theor. Appl. Genet.* 125, 759–771

43 Akdemir, D. (2013) Locally epistatic genomic relationship matrices forgenomic association, prediction and selection. arXiv 1302.3463v 1306. http://arxiv.org/pdf/1302.3463v6.pdf.

44 Sun, X. *et al.* (2012) Nonparametric method for genomics-based prediction of performance of quantitative traits involving epistasis in plant breeding. *PLoS ONE* 7, e50604

45 Zhao, Y. *et al.* (2013) Genomic prediction of hybrid wheat performance. *Crop Sci.* 53, 802–810

46 Zhong, S. *et al.* (2009) Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. *Genetics* 182, 355–364

47 Wimmer, V. *et al.* (2013) Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics* 195, 573–587

48 Iwata, H. *et al.* (2011) Prospects for genomic selection in conifer breeding: a simulation study of *Cryptomeria japonica*. *Tree Genet. Genomes* 7, 747–758

49 Riedelsheimer, C. *et al.* (2012) Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. *BMC Genomics* 13, 452

50 Nakaya, A. and Isobe, S.N. (2012) Will genomic selection be a practical method for plant breeding? *Ann. Bot. (Lond.)* 110, 1303–1316

51 Breseghello, F. and Sorrells, M.E. (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. *Genetics* 172, 1165–1177

52 Daetwyler, H.D. *et al.* (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* 193, 347–365

53 Rutkoski, J. *et al.* (2012) Evaluation of genomic prediction methods for fusarium head blight resistance in wheat. *Plant Genome J.* 5, 51–61

54 Oliveira, E.J. *et al.* (2012) Genome-wide selection in cassava. *Euphytica* 187, 263–276

55 Rutkoski, J.E. *et al.* (2011) Genomic selection for durable stem rust resistance in wheat. *Euphytica* 179, 161–173

56 Yabe, S. *et al.* (2013) Potential of genomic selection for mass selection breeding in annual allogamous crops. *Crop Sci.* 53, 95–105

57 Ziyomo, C. and Bernardo, R. (2013) Drought tolerance in maize: indirect selection through secondary traits versus genomewide selection. *Crop Sci.* 53, 1269–1275

58 Denis, M. and Bouvet, J-M. (2012) Efficiency of genomic selection with models including dominance effect in the context of *Eucalyptus* breeding. *Tree Genet. Genomes* 9, 37–51

59 Lorenzana, R.E. and Bernardo, R. (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theor. Appl. Genet.* 120, 151–161

60 Jonas, E. and de Koning, D.J. (2013) Does genomic selection have a future in plant breeding? *Trends Biotechnol.* 9, 497–504

61 Alonso, J.M. and Ecker, J.R. (2006) Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in *Arabidopsis*. *Nat. Rev. Genet.* 7, 524–536

62 Nagy, A. *et al.* (2003) Tailoring the genome: the power of genetic approaches. *Nat. Genet.* 33, 276–284

63 Jannink, J-L. (2010) Dynamics of long-term genomic selection. *Genet. Sel. Evol.* 42, 35

64 Xu, S. (2013) Genetic mapping and genomic selection using recombination breakpoint data. *Genetics* http://dx.doi.org/10.1534/genetics.1113.155309

65 Davey, J.W. *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12, 499–510

66 Jannink, J.L. *et al.* (2010) Genomic selection in plant breeding: from theory to practice. *Brief. Funct. Genomics* 9, 166–177

67 Cabrera-Bosquet, L. *et al.* (2012) High-throughput phenotyping and genomic selection: the frontiers of crop breeding converge. *J. Integr. Plant Biol.* 54, 312–320

68 Dhondt, S. *et al.* (2013) Cell to whole-plant phenotyping: the best is yet to come. *Trends Plant Sci.* 18, 428–439

69 Li, Y. *et al.* (2008) Selection on multiple QTL with control of gene diversity and inbreeding for long-term benefit. *J. Anim. Breed. Genet.* 125, 320–329

70 Friedman, J. *et al.* (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33, 1–22

71 Li, Z. and Sillanpää, M.J. (2012) Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. *Theor. Appl. Genet.* 125, 419–435

72 Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* 67, 301–320

73 de los Campos, G. *et al.* (2013) Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet.* 9, e1003608

74 de los Campos, G. *et al.* (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182, 375–385

75 Habier, D. *et al.* (2011) Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* 12, 186

76 de Los Campos, G. *et al.* (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193, 327–345

77 Gianola, D. and van Kaam, J.B. (2008) Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178, 2289–2303

78 de los Campos, G. *et al.* (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet. Res.* 92, 295–308

79 Holliday, J.A. *et al.* (2012) Predicting adaptive phenotypes from multilocus genotypes in Sitka spruce (*Picea sitchensis*) using random forest. *G3* 2, 1085–1093

80 Usai, M.G. *et al.* (2009) LASSO with cross-validation for genomic selection. *Genet. Res.* 91, 427–436

81 Croiseau, P. *et al.* (2011) Fine tuning genomic evaluations in dairy cattle through SNP pre-selection with the Elastic-Net algorithm. *Genet. Res.* 93, 409–417

82 Moser, G. *et al.* (2009) A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genet. Sel. Evol.* 41, 56

83 Okut, H. *et al.* (2011) Prediction of body mass index in mice using dense molecular markers and a regularized neural network. *Genet. Res.* 93, 189–201

84 Gianola, D. *et al.* (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* 173, 1761–1776

85 VanRaden, P. (2008) Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423

86 González-Recio, O. and Forni, S. (2011) Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. *Genet. Sel. Evol.* 43, 1–12