GS Methods

April 17, 2020

Genomic Selection (GS):

Genomic selection (GS) promises to accelerate the rate of genetic gain in plant breeding for genetically complex traits, including yield, by selecting individuals of high breeding value earlier in the breeding cycle.

Genomic estimated breeding value (GEBV):

The GS approach is novel in that it simultaneously estimates all genome-wide marker effects to predict an individual's genomic estimated breeding value (GEBV), contrary to estimate marker effects based on their level of significance (e.g MAS).

Genomic information:

As the cost and efficiency of obtaining genomic information on plants (or animals) below the cost and efficiency of evaluating individuals over years and locations, genomic information can more affordably be leveraged to predict phenotypic performance.

Shortening of the breeding cycle:

- GS can facilitate a shortening of the breeding cycle and enable earlier selection and intercrossing of early-generation breeding material.
- GS-based cycle time may be reduced to one or two years from the traditional five to seven years as selection of lines prior to field testing allows shortening of the generation interval typical with phenotypic selection

Potential advantages of GS:

- One of the greatest potential advantages of GS is its ability to identify individuals with higher breeding values without the requirement of collecting phenotypes pertaining to those individuals.
- It has been shown that selection of individuals based on GEBV can substantially increase the rate of genetic gain in plant breeding compared to traditional marker-assisted selection (MAS) or phenotypic selection.
- GS can facilitate a shortening of the breeding cycle and enable earlier selection and intercrossing of early-generation breeding material.
- GS-based cycle time may be reduced to one or two years from the traditional five to seven years as selection of lines prior to field testing allows shortening of the generation interval typical with phenotypic selection.

Training Population:

- GS prediction models are developed using a training population consisting of individuals with both genome-wide marker genotypes and phenotypes of interest.
- These individuals are used to train a model by simultaneously estimating the contribution of marker effects to their phenotypic value.
- GS models utilize information gathered from the training population to estimate a breeding value and predict the performance of breeding lines without phenotypes.
- Dense marker coverage is needed in order to maximize the number of QTL in linkage disequilibrium (LD) with at least one marker, thereby maximizing the number of QTL effects captured by the molecular markers.

Large p, small n problem:

- With a large number of genetic markers there will often be more effects to be estimated than there are phenotypic data points for which to estimate them.
- It is a concern for GS as it can cause over-fitting of the model which exaggerates minor fluctuations in the data and is due to collinearity between markers (Lorenz et al., 2011).
- This creates a model that is highly accurate when evaluating the training population but has poor predictive ability when applied to a different test population.
- Research has shown that GS models have diminishing returns for additional markers once the populations have reached the point of 'marker saturation'.

Composition of_training_population:

- One of the most important factors in developing a GS model is the composition of individuals included in the training population set.
- Prediction accuracies are maximized when the training population and the test population (selection candidates) are closely related.
- More closely related individuals share a common ancestry fewer generations back, and, therefore, fewer opportunities for recombination between markers and QTL, thus preserving QTL-marker linkage phases.
- Several studies have shown that training population size has a greater impact on GS prediction accuracy than marker number, with a smaller training population size having a strong negative effect on GS prediction accuracy.
- Studies have shown that combining genotypes from multiple populations in order to create a larger training population results in higher prediction accuracies than analysing individual populations with fewer genotypes in the training population

Genomic Selection Models:

- The GS concept encompasses a broad range of methods
- Their common feature is to estimate the breeding values of individuals for quantitative traits using whole genome genotypes through the simultaneous estimation of marker effects in a single step
- With the increased popularity of GS in plant breeding, numerous models have been proposed:
  - Ridge regression-best linear unbiased prediction (RR-BLUP)
  - Bayesian regression
  - Kernel regression, and
  - Machine learning

Prediction of GEBV:

$$\mathbf{GEBV} = \sum_{\mathbf{i}}^{\mathbf{n}} \mathbf{X_i} \hat{\mathbf{g}}_{\mathbf{i}}$$

- Where:
  - n is the number of markers across the genome,
  - X_i is a design matrix allocating animals to genotypes at marker i, and
  - g_i is the effect of the genotype at marker i.

Ridge regression-best linear unbiased prediction (RR-BLUP):

- Assumption: marker effects are all very small, and are normally distributed

$$\boldsymbol{y} = \mathbf{1}_n\mu + \boldsymbol{X}\boldsymbol{g} + \boldsymbol{e}$$

- Where:
  - y is a vector of phenotypes (number of records x 1)
  - 1_n is a vector of 1s, allocating the effect of the mean to each record
  - μ is the overall mean
  - X is a design matrix, allocating records to genotypes for m markers (number or records x m)
  - g is a vector of the effects of the m markers
  - e is a vector of random residuals, assumed normally distributed, variance σ^2_e
- Makes the assumption that markers are random effects having nonzero effects with equal marker variance
- This assumption does not mean the effects of all markers are

Bayesian models:

- Bayesian models address the simple but likely unrealistic assumptions that all markers have nonzero effects and that markers have equal variance.
- Bayesian models relax these two assumptions and better model marker effects of differing sizes
- Bayesian models estimate a separate variance for each marker, and the variances are assumed to follow a specified prior distribution

BayesA:

- Allow variance toward zero, but did not permit the value of zero itself

BayesB:

- Allows markers to have a variance of zero.

BayesCπ:

- Assumes common marker variance and allows for some markers to have no effect

Bayesian vs RR-Blup:

- Bayesian methods outperformed RR-BLUP through better estimation of large-effect QTL by allowing for unequal variances.
- Studies show only slight differences between their accuracies (RR-BLUP, BayesA, BayesB, and BayesCπ).
- Studies concluded that GS accuracy was not strongly influenced by model choice.
- One of the draw backs of the Bayesian approach is its computational complexity which results in long run times.
- Computational complexity has led many researchers and breeders to rely on the RR-BLUP model.

The best approach:

- The best approach for using molecular markers in GS largely depends on the genetic architecture of the trait
- For a given number of markers (N_M), RR-BLUP assumes that each marker accounts for $(\frac{1}{N_M})^{th}$ of the total genetic variation (V_G).
- If one of the markers corresponds to a known major gene, the assumption of common variance for the known major gene leads to an underestimation of the estimated effects of the major gene.
- An alternative to the modelling all genes as random effects with equal variance is to model known major genes or QTLs as fixed effects in a model with genome-wide markers as random effects.

RR-BLUP with major genes:

- RR-BLUP models with fixed effects of major genes were shown to provide greater GS prediction accuracy when trait heritability was high and a large percentage of V_G was explained by the major genes.
- It was also shown that as the number of training population individuals decreased, it became more advantageous to consider major genes as fixed effects rather than random effects.
- These results were consistent whether there was a single major gene or whether there were multiple major genes
- Although treating a major gene as a fixed effect can increase GS prediction accuracy, using a major gene as a fixed effect puts a stronger selection pressure on the major gene which can lead to more drastic changes in gene frequency