Published March 13, 2015

ORIGINAL RESEARCH

Efficient Use of Historical Data for Genomic Selection: A Case Study of Stem Rust Resistance in Wheat

J. Rutkoski,* R. P. Singh, J. Huerta-Espino, S. Bhavani, J. Poland, J. L. Jannink, and M. E. Sorrells

Abstract

Genomic selection (GS) is a methodology that can improve crop breeding efficiency. To implement GS, a training population (TP) with phenotypic and genotypic data is required to train a statistical model used to predict genotyped selection candidates (SCs). A key factor impacting prediction accuracy is the relationship between the TP and the SCs. This study used empirical data for quantitative adult plant resistance to stem rust of wheat (*Triticum aestivum* L.) to investigate the utility of a historical TP (TP_H) compared with a population-specific TP (TP_{PS}), the potential for TP_H optimization, and the utility of TP_H data when close relative data is available for training. We found that, depending on the population size, a TP_{PS} was 1.5 to 4.4 times more accurate than a TP_H, and TP_H optimization based on the mean of the generalized coefficient of determination or prediction error variance enabled the selection of subsets that led to significantly higher accuracy than randomly selected subsets. Retaining historical data when data on close relatives were available lead to a 11.9% increase in accuracy, at best, and a 12% decrease in accuracy, at worst, depending on the heritability. We conclude that historical data could be used successfully to initiate a GS program, especially if the dataset is very large and of high heritability. Training population optimization would be useful for the identification of TP_H subsets to phenotype additional traits. However, after model updating, discarding historical data may be warranted. More studies are needed to determine if these observations represent general trends.

Published in The Plant Genome 8
doi: 10.3835/plantgenome2014.09.0046

5585 Guilford Rd., Madison, WI 53711 USA
An open-access publication

GENOMIC SELECTION (GS) (Haley and Visscher, 1998; Meuwissen et al., 2001) is a breeding methodology that can increase rates of genetic gain by reducing the breeding cycle duration or by increasing the selection accuracy. With GS, a training population (TP) consisting of individuals having both phenotypic and genotypic observations is used to train a model that predicts breeding values of selection candidates (SCs) based on genotype. The accuracy of this prediction depends on the TP size (N_p), heritability (h^2), effective number of loci, and

J. Rutkoski, International Programs in the College of Agriculture and Life Sciences, and Plant Breeding and Genetics Section in the School of Integrative Plant Science, 240 Emerson Hall, Cornell Univ., Ithaca, NY 14853, USA, and International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600 El Batan, Mexico; R.P. Singh and J. Huerta-Espino, International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600 El Batan, Mexico; J. Huerta-Espino, Campo Experimental Valle de México INIFAP, Apdo. Postal 10, 56230 Chapingo, Edo de México, Mexico; S. Bhavani, CIMMYT, ICRAF House, United Nations Ave., Gigiri, Village Market-00621, Nairobi, Kenya; J. Poland, Wheat Genetics Resource Center, Dep. of Plant Pathology and Dep. of Agronomy, Kansas State Univ. (KSU), 4011 Throckmorton Hall, Manhattan, KS 66506, USA; J.L. Jannink, USDA–ARS and Plant Breeding and Genetics Section in the School of Integrative Plant Science, 240 Emerson Hall, Cornell Univ., Ithaca, NY 14853, USA; M.E. Sorrells, Plant Breeding and Genetics Section in the School of Integrative Plant Science, 240 Emerson Hall Cornell Univ., Ithaca, NY 14853, USA. Received 8 Sept. 2014. Accepted 5 Jan. 2015. *Corresponding author (jer263@cornell.edu).

Abbreviations: BLUP, best linear unbiased prediction; CD_{mean}, coefficient of determination; FA, factor analytic; Fst, F-statistic; G-BLUP, genomic best linear unbiased prediction; GBS, genotyping-by-sequencing; GS, genomic selection; G×E, genotype-by-environment interaction; LD, linkage disequilibrium; MAF, minor allele frequency; N_p, training population size; PEV_{mean}, prediction error variance; QTL, quantitative trait loci; SC, selection candidate; TP, training population; TP_H, historical training population; TP_{PS}, population-specific training population.

the level of linkage disequilibrium (LD) between genetic markers and quantitative trait loci (QTL) (Goddard, 2009; Daetwyler et al., 2010). If the TP and SCs are from different populations, the genetic relationship between these two populations is another major factor affecting GS accuracy (Habier et al., 2007; de Roos et al., 2009; Hayes et al., 2009; Long et al., 2011; Pszczola et al., 2012). As the relationship between the TP and SC decreases, the forces of selection, recombination, and drift change the pattern of LD between markers and QTL. Furthermore, markers that capture family effects rather than QTL effects contribute much less to the GS accuracy as relationship between the TP and SCs declines (Habier et al., 2007). Quantitative trait loci interaction effects may also contribute to the decrease in accuracy as the relationship between the TP and SCs decreases.

In plant breeding, there is considerable interest in the use of historical data for GS model training to predict breeding values of new SCs (Crossa et al., 2010; Asoro et al., 2011; Dawson et al., 2013). By historical data, we mean preexisting data from a breeding program that was not specifically generated with GS modeling training in mind. Compared to a population-specific TP (TP_{PS}) that consists of a subset of the SC population, a historical TP (TP_H) enables predictions to be generated sooner in the breeding cycle because phenotyping the TP occurs before the selection candidates are developed. In addition, a TP_H could offer higher phenotypic accuracy through advanced replicated trials and sample more environments compared to a newly generated TP_{PS}. On the other hand, compared to a TP_{PS}, a TP_H consists of more distant relatives, which can lead to reduced accuracy. Studies that have assessed GS accuracy from historical data in crop species have measured accuracy using either random cross-validation or forward validation, where an older subset of the data is used to predict a newer subset (Asoro et al., 2011; Dawson et al., 2013). Accuracies from random cross-validation are likely to be overestimated because the TP and SCs are from the same population. On the other hand, accuracies from forward validation may be driven largely by the level of genotype-by-environment interaction (G×E) between the historical (training) and current (validation) environments rather than the relationship between the TP and SCs. As a result, there are no empirical studies that can clearly demonstrate the utility of historical data for the prediction of new SCs assuming that the historical set of environments represent those environments of interest to the breeding program.

The purpose of this case study was to assess the utility of historical data for the prediction of new, early generation SCs. We used empirical data from a recurrent genomic selection program for stem rust (*Puccinia graminis* f. sp. *tritici*) adult plant resistance in wheat (*Triticum aestivum* L.) to (i) determine the relative accuracies achieved using historical and population-specific training sets for the prediction of new SCs, (ii) determine the potential to use TP optimization methods to identify the best subsets of historical individuals to use for training, and (iii) determine if historical data should remain part of the TP if data on close relatives becomes available for model training.

Materials and Methods

Genetic Material

A set of 365 advanced CIMMYT wheat lines was used as the historical population. These lines were developed using a selected bulk-breeding scheme that includes early generation, single-plant selection against late maturity, tallness, and susceptibility to diseases including stem rust. The same seed stock used for phenotyping was also used for genotyping. A second population of 503 new SCs was generated by two rounds of random mating between 14 founder lines from the historical population, followed by one round of selfing for seed increase. No selection was intentionally applied during development of the SCs. Each SC was phenotypically evaluated based on its S_1 or S_2 progeny. Each SC was genotyped using bulk DNA from six S_1 progenies.

Phenotypic Data

Individuals were evaluated for quantitative adult plant resistance to stem rust at the Kenya Agricultural Research Institute, Njoro, Kenya and the Ethiopian Institute of Agricultural Research, Debre Zeit, Ethiopia under the conditions and methods described in Yu et al. (2011). The historical population was evaluated across 10 seasons in Kenya and three seasons in Ethiopia from 2007 and 2013, with each individual appearing in approximately four of the 13 environments. The S_1 or S_2 families derived from each SC were evaluated in Kenya during the 2012 main and off-season and during the 2013 main season. Each family was planted in a twin row field plot of 70- and 30-cm spacing surrounded by a 1-m border of spreader plants. Hills of spreader plants were planted in rows perpendicular to the entry rows. Just before booting (growth stage Z35–Z37; Zadoks et al., 1974), individual spreader plants of the border rows were inoculated with fresh urediniospores of *Puccinia graminis* f. sp. *tritici* race TTKST (Sr24 virulent race) suspended in distilled water using a hypodermic syringe on at least two occasions. Spreaders were also sprayed with suspension of urediniospores in light mineral oil Soltrol 170 to ensure successful infection. Disease severity was determined according to modified Cobb scale (Peterson et al., 1948), and a Box–Cox transformation (Box and Cox, 1964) was applied before analysis. For both the historical and selection candidate populations, heritability on a line-mean basis was calculated according to Hallauer et al. (2010). Variance components were estimated using the R package lme4 (Bates and Maechler, 2010).

Genotypic Data

Genotyping-by-sequencing (GBS, Elshire et al., 2011) was implemented according to the protocol described in Poland et al. (2012a). Out of the total of 27,434 polymorphic markers generated, 17,168 unique markers with less than 80% missing data in the historical population and

polymorphic in the selection candidates were selected. Before marker filtering, missing data was imputed using random forest imputation described in Poland et al. (2012b) as recommended by Rutkoski et al. (2013).

Relationship Matrix

The relationship matrix (**A**) was calculated according to Leutenegger et al. (2003), Amin et al. (2007), and Astle and Balding (2009). Relationship estimates for a pair of individuals, *i* and *j*, were calculated as follows:

$$f_{ij} = \frac{1}{n}\sum_{k=1}^{n}\frac{\left(a_{ik} - p_k\right)\left(a_{jk} - p_k\right)}{p_k\left(1 - p_k\right)}$$

where a_{ik} is the genotype of individual *i* at marker *k* coded as 0, 0.5, and 1; p_k is the frequency of the major allele, and *n* is the number of markers used for kinship estimation. Before relationship matrix calculation, markers with a minor allele frequency (MAF) less than 0.05 were excluded.

Population Characterization

Population differentiation between the 365 lines and the 503 SCs was measured using the F-statistic (Fst) (Weir and Cockerham, 1984). For each marker only nonimputed data points were used. Statistical significance of the median Fst across all markers was assessed using 1000 permutations. For each iteration, the population assignment of the individuals was randomly shuffled before calculating median Fst. The distribution of the 1000 median Fst values was used as the null distribution for p-value calculation.

To visualize the population structure of the combined historical and selection candidate population, principal component analysis of the relationship matrix was implemented in R (R Development Core Team, 2010).

Linkage disequilibrium decay in historical and the SC population was investigated by fitting a cubic smoothing spline to the r^2 vs. genetic distance in centimorgans for pairs of markers on the same chromosome. Estimates of marker position from the Synthetic W9784 × Opata85 genetic map (Poland et al., 2012a) were available for 2425 markers. Markers with unknown map position and markers with MAF <0.05 were excluded, leaving 2050 markers. For each pairwise r^2 calculation, only nonimputed data points were used and marker pairs were excluded if there were <30 pairwise complete observations.

Genomic Selection Model

A single-stage, genomic best-linear unbiased prediction (G-BLUP) model (Bernardo, 1994; Piepho, 2009), was used for all genomic predictions:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

$$\mathbf{u} \sim \mathbf{N}\left(0, \mathbf{A}\sigma_u^2\right)$$

$$\boldsymbol{\varepsilon} \sim \mathbf{N}\left(0, \mathbf{R}\sigma_\varepsilon^2\right)$$

where **y** is a vector of phenotypes, $\boldsymbol{\beta}$ is a vector of environment effects treated as fixed, **u** is a vector of genotype effects treated as random, **X** and **Z** are the design matrices relating $\boldsymbol{\beta}$ and **u** to the observations in **y**, $\boldsymbol{\varepsilon}$ is the residual error, σ_u^2 is the genetic variance, σ_ε^2 is the error variance and **R** was the residual covariance matrix, which was equal to the identity matrix unless specified otherwise. The G-BLUP solutions for the breeding values were obtained using the mixed model equations (Henderson, 1984):

$$\begin{pmatrix} \mathbf{X}\mathbf{R}^{-1}\mathbf{X}' & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} + \lambda\mathbf{A}^{-1} \end{pmatrix}\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}$$

where $\hat{\boldsymbol{\beta}}$ is the vector of fixed effect solutions, $\hat{\mathbf{u}}$ is the vector of estimated breeding values, and $\lambda = \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_u^2}$. The variance components, $\hat{\sigma}_\varepsilon^2$ and $\hat{\sigma}_u^2$, were estimated with the training set using restricted estimation maximum likelihood implemented in the R package rrBLUP (Endelman, 2011).

Training Population Accuracy Comparison

Out of the 503 SCs, 138 selected to be representative of the entire population based on pedigree were set aside as the validation population. The remaining 365 SCs were designated as the TP_{PS}. The 365 historical lines formed the TP_H. The TP_{PS} and TP_H were compared in terms of accuracy for N_p values: 73, 146, 219, 292, and 365. For each accuracy calculation, 1000 random samples of size N_p, drawn without replacement, were used for model training, validation, and accuracy calculation. For each level of N_p, a 95% confidence interval for accuracy was constructed by sorting the 1000 accuracies from smallest to largest and using the 24th and 974th accuracy values as the lower and upper confidence limits. Lastly, an average λ across the 1000 samples for each N_p was computed (λ_{Np}) for use in later analyses.

The validation set was evaluated across two environments: Kenya main-season 2012 and Kenya main-season 2013. For model training, data from Kenya main-season 2012 and Kenya main-season 2013 were excluded so that the training and validation environments would not overlap. Accuracies are reported as the Pearson's correlation between the G-BLUPs (with validation individuals' own phenotypic records excluded) and the genetic values of the validation set calculating using individuals' own phenotypic records only. To estimate the genetic values of the validation set, the R package rrBLUP (Endelman, 2011) was used to fit the mixed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

$$\mathbf{u} \sim \mathbf{N}\left(0, \mathbf{I}\hat{\sigma}_u^2\right)$$

$$\boldsymbol{\varepsilon} \sim \mathbf{N}\left(0, \mathbf{I}\hat{\sigma}_\varepsilon^2\right)$$

where **I** is an identity matrix.

Correlation Between Model Training and Validation Environments

A factor analytic (FA) model, implemented in ASreml-R (Gilmour et al., 2009), was fit to parsimoniously model the covariance among environments. The FA model estimates the unobserved common factors, *k*, that give rise to the correlations between the environments, *e*. The environmental covariance matrix is modeled as follows:

$$\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Gamma}' + \mathbf{\Psi}$$

where $\mathbf{\Gamma}$ is an $e \times k$ matrix of factor loadings, and $\mathbf{\Psi}$ is an $e \times e$ diagonal matrix of environment specific variances. The FA variance models including genomic relationship information were fit for $k = 1$, 2, and 3 according to Beeck et al., (2010). Data from 16 environments between 2005 and 2012 were used to fit the FA models. The FA $k = 2$ model was selected based on the Akaike information criterion. To estimate BLUPs of the SCs in each environment, including environments where they were not phenotypically observed, estimates of variance parameters were used in the mixed model equations to estimate empirical BLUPs of each individual *i* in each environment *j*, $\hat{\mathbf{u}}_{ij}$, according to Thompson et al. (2003). The genetic value of each validation individual, *i*, across a set of *N* environments was predicted as follows:

$$\bar{\mathbf{u}} = \frac{1}{N}\sum_{j}^{N}\hat{\mathbf{u}}_{ij}$$

This was calculated for the validation individuals across the set of historical training environments, $\bar{\mathbf{u}}_{\mathrm{H}}$, and across the set of population-specific training environments $\bar{\mathbf{u}}_{\mathrm{PS}}$. Correlations between each of the training environments and the set of validation environments were calculated as $cor\left(\mathbf{u},\bar{\mathbf{u}}_{\mathrm{H}}\right)$, and $cor\left(\mathbf{u},\bar{\mathbf{u}}_{\mathrm{PS}}\right)$, were $\mathbf{u}$ was a vector of genetic values for the selection candidates calculated as described previously.

Training Population Optimization

Two approaches were tested for TP optimization: (i) minimize the genetic differentiation between the training and validation populations and (ii) maximize the precision of the prediction of the difference between each validation set individual and the mean of the validation population. For the first approach, the median Fst across all markers was the TP optimization criterion. For the second approach the mean prediction error variance ($\mathrm{PEV}_{\mathrm{mean}}$) and the mean coefficient of determination ($\mathrm{CD}_{\mathrm{mean}}$) were tested as TP optimization criteria as suggested by Rincent et al. (2012). The $\mathrm{PEV}_{\mathrm{mean}}$ and $\mathrm{CD}_{\mathrm{mean}}$ were recommended by Kennedy and Trus (1993) and Laloë (1993), respectively, as measures of the predictability of contrasts for breeding value estimation by best linear unbiased prediction (BLUP). Precise estimation of the contrasts (differences) between the overall selection candidate population mean and the individual breeding values is key for the identification of the best individuals for selection.

For each population consisting of a potential training set of size N_{p} and the validation set, according to Rincent et al. (2012), $\mathrm{PEV}_{\mathrm{mean}}$ or $\mathrm{CD}_{\mathrm{mean}}$ of the validation individuals were computed as follows:

$$\mathrm{PEV}_{\mathrm{mean}} = \frac{\sum_{i=1}^{N_v}\frac{\mathbf{c}_i'\left(\mathbf{Z}'\mathbf{M}\mathbf{Z}+\lambda^{-1}\right)^{-1}\mathbf{c}_i'}{\mathbf{c}_i'\mathbf{c}_i}\times\hat{\sigma}_{\varepsilon}^2}{N_v}$$

$$\mathrm{CD}_{\mathrm{mean}} = \frac{\sum_{i=1}^{N_v}\frac{\mathbf{c}_i'\left(\mathbf{A}-\lambda\left(\mathbf{Z}'\mathbf{M}\mathbf{Z}+\lambda\mathbf{A}^{-1}\right)^{-1}\right)\mathbf{c}_i'}{\mathbf{c}_i'\mathbf{A}\mathbf{c}_i}}{N_v}$$

where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. λ was set equal to λ_{Np} according to the size of the TP tested, N_v is the number of validation individuals, and $\mathbf{c}_i$ is a contrast vector of length $N_{\mathrm{p}} + N_v$. Our contrast of interest was between the individuals in the validation set and the overall mean of the validation set. For each validation individual *i*, the element in $\mathbf{c}_i$ corresponding to the individual *i* was $1 - 1/N_v$, the elements corresponding to the other validation individuals were $-1/N_v$, and the remaining values were zero. Contrasts were specified in this way because in this case individuals in the TP are not candidates for selection. The objective was to select the TP so that either $\mathrm{PEV}_{\mathrm{mean}}$ is minimized or $\mathrm{CD}_{\mathrm{mean}}$ is maximized.

An exchange algorithm was used for the selection of optimal TPs. Step one, a random sample of size N_{p} is selected and the optimization criterion of interest is calculated. Step two, a randomly selected individual is removed and then replaced by a new randomly selected individual. Step three, this change is accepted if the TP is improved based on the optimization criteria or rejected if not. Steps two and three are repeated for a maximum of 2000 iterations or until changes to the TP are rejected for 200 consecutive iterations. The exchange algorithm was repeated 100 times, and the overall optimal TP was selected. Genomic selection accuracies using the optimal TPs were computed.

As an external validation, the optimized TPs were used to predict an additional population that was derived by intermating 10 individuals selected from the SCs as part of a recurrent selection experiment. Accuracies with optimized TPs from TP_{H} were compared to accuracies with randomly selected TPs from TP_{H}. Phenotypic and genotypic data for this external validation population was generated as described for the SC population, except only one season of phenotypic data was available, and mean imputation was used before relationship matrix calculation.

Combined Training Population Analysis

A population-specific training population combined with random samples of size N_{p} from TP_{H} was compared to a $\mathrm{TP}_{\mathrm{PS}}$ alone in terms of GS accuracy. Different values for heritability on a line-mean basis were simulated for $\mathrm{TP}_{\mathrm{PS}}$ and TP_{H} individuals. To vary the heritability, a random

error vector with mean zero and standard deviation, σ'_ε was added to the observations in TP_{PS} and TP_H according to the simulated heritability, H^2_{sim} for both populations:

$$\sigma'_\varepsilon = \frac{\sigma^2_g}{H^2_{sim}} - \left(\sigma^2_g + \frac{\sigma^2_{ge}}{e} + \frac{\sigma^2_e}{er}\right)$$

σ^2_g, σ^2_{ge}, σ^2_e are the genetic, G×E, and error variances, e is the number of environments and r is the number of replicates within an environment. Heritabilities of 0.2 and 0.6 were simulated for TP_{PS} and for each of these heritability levels, N_p individuals from TP_H were added with H^2_{sim} of 0.2, 0.3, 0.4, or 0.6. The GS accuracies were calculated using each combined TP. To determine if accuracy could be improved by weighting the observations from TP_{PS} and TP_H according to the H^2_{sim} of their population of origin, the combined TP analysis was repeated except in the mixed model used for genomic prediction described previously, the diagonal of the residual covariance matrix, **R**, was $1 - H^2_{sim}$.

Results

Phenotypic Data Characterization

Line-mean heritability was 0.82 and 0.61 for the historical and SC populations, respectively. The correlation between the validation set evaluation environments with the historical and TP_{PS} evaluation environments was 0.81 and 0.83, respectively.

Population Characterization

The historical and SC populations were significantly differentiated based on the median Fst across all markers, p-value = 0. Populations also formed distinct but partially overlapping groups together based on the first two principle components calculated from the genetic relationship matrix (Fig. 1). The rate of LD decay with physical distance was similar for the historical and SC population; however, there was more long-range LD in the SC population (Fig. 2).

TP Comparison and Optimization

The TP_{PS} always lead to higher accuracies than TP_H, and for N_p = 73 and 146, accuracies were significantly different (Fig. 3). As N_p increased, the difference between accuracies from TP_{PS} and TP_H decreased. For example, when N_p = 73, TP_{PS} was 4.4 times more accurate than TP_H, and when N_p = 292, TP_{PS} was only 1.5 times more accurate than TP_H. For TP_H, optimally selected TPs lead to significantly higher accuracies than randomly selected TPs for N_p = 73, 146, 219, and 292 (Fig. 4). The optimization criteria PEV_{mean} and CD_{mean} performed similarly and both outperformed Fst. For N_p = 73, 146, 219, and 292 optimally selected TPs based on PEV_{mean} and CD_{mean} lead to accuracies higher than that of the full TP with N_p = 365.

When validated using a second population derived from the SC population, the TPs that were optimally selected from TP_H based on CD_{mean} and PEV_{mean} lead to consistently higher accuracies compared to randomly selected TPs (Fig. 5). For this validation experiment, the improvement in accuracy provided by CD_{mean} optimization was most consistent, followed by PEV_{mean}. The TPs selected based on Fst performed worse than random TPs for N_p = 146 and 219. Although optimized TPs selected using CD_{mean} or PEV_{mean} consistently outperformed random TPs, no significant differences were detected due to the large variation of the random TP accuracies due to sampling. In contrast with the results from validation using the SC population, we observed increasing accuracy as N_p increased for TPs optimized using CD_{mean} and PEV_{mean}. However, when N_p = 292 and TPs were selected based on Fst, PEV_{mean}, or CD_{mean}; and when N_p = 73 and TPs were selected based on Fst, accuracy was higher than that of the complete TP.

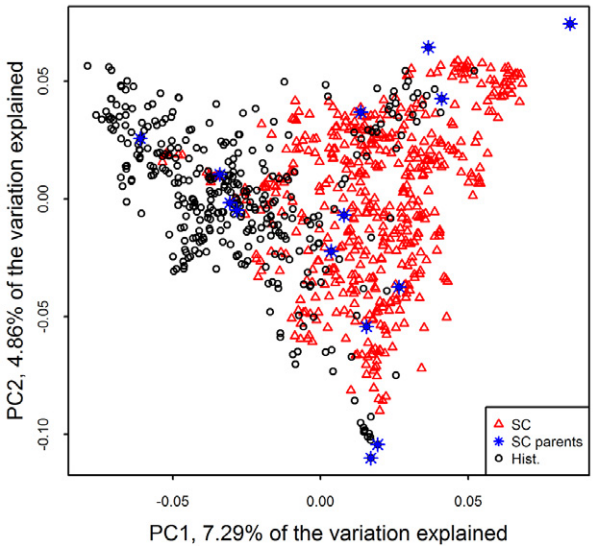

Figure 1. Principal components analysis including the historical lines, selection candidates (SCs), and SC parents.

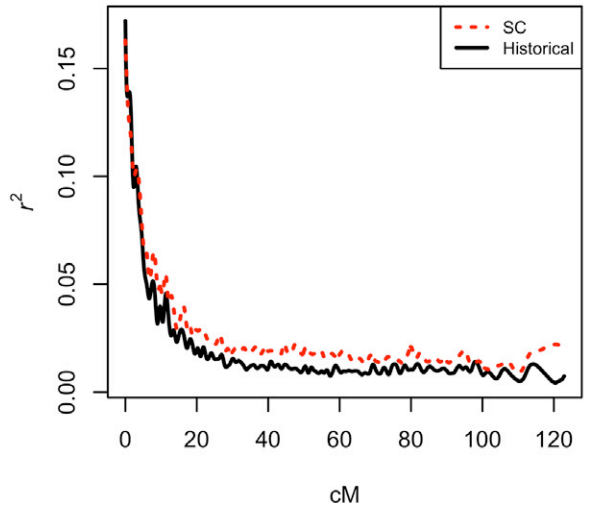

Figure 2. Relationship between linkage disequilibrium measured as r^2 and genetic distance in centimorgans (cM) for historical and selection candidate (SC) populations.

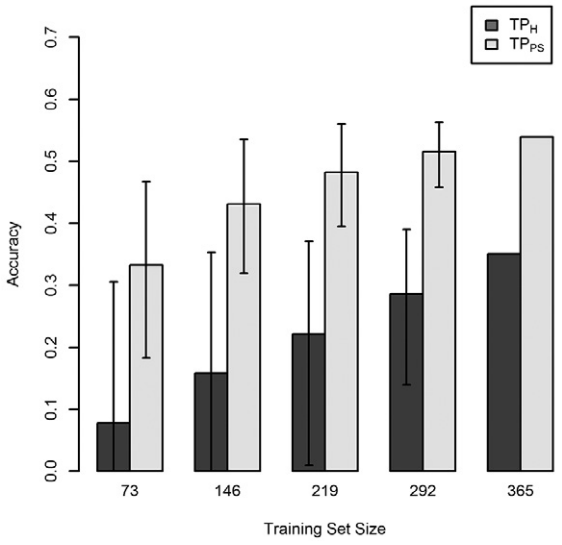

Figure 3. Prediction accuracies for the selection candidate (SC) population based on population-specific training population (TP_{PS}) and historic training population (TP_H) with varying population sizes.

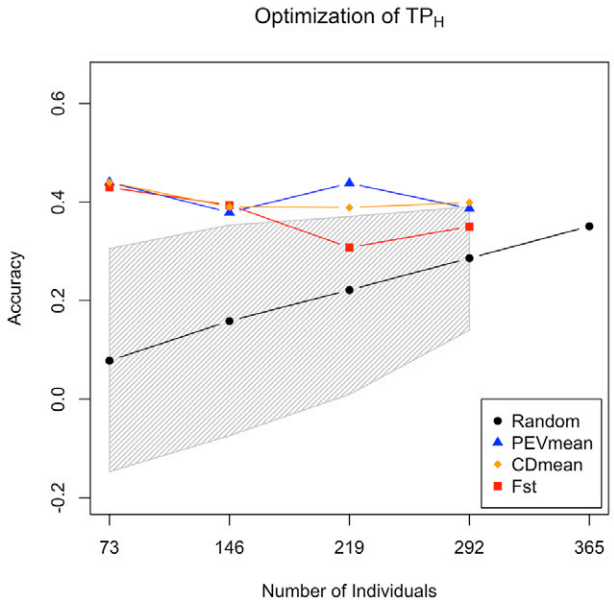

Figure 4. Prediction accuracies for the selection candidate population based on optimized training populations from historic training population (TP_H) in comparison with accuracies from randomly sampled TPs from TP_H. The 95% confidence interval for accuracy from randomly sampled TPs is shaded in gray.

Combined TP analysis

When H^2_{sim} of TP_{PS} was low (H^2 = 0.2) adding samples from TP_H of H^2_{sim} = 0.3, 0.4, and 0.6, led to a small but constant improvement in accuracy as N_p increased (Fig. 6). When H^2_{sim} of TP_H was also low (H^2 = 0.2), adding individuals from TP_H to TP_{PS} led to an initial decrease in accuracy, followed by a slight increase with increasing N_p. For the maximum number of TP_H samples added,

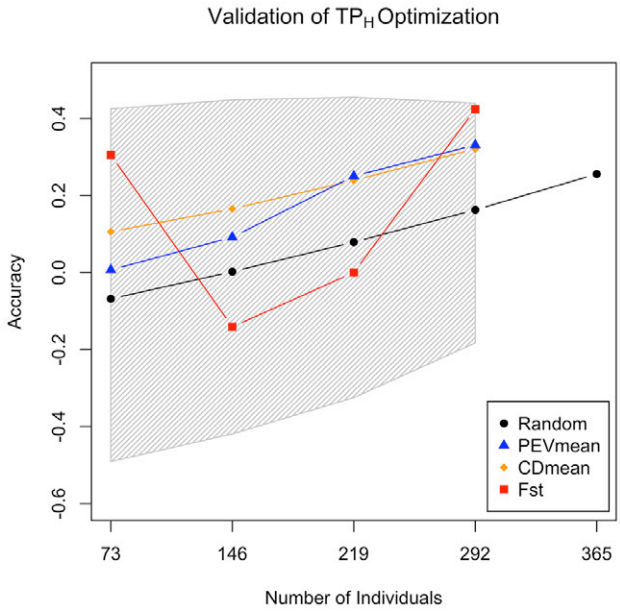

Figure 5. Prediction accuracies for an additional validation population based on optimized training populations from historic training population (TP_H) in comparison with accuracies from randomly sampled TPs from TP_H. The 95% confidence interval for accuracy from randomly sampled TPs is shaded in gray.

365, accuracy improved by 1.2, 7.6, 10.9, and 11.9% for H^2_{sim} = 0.2, 0.3, 0.4, and 0.6, respectively. Adding a weight of 1 – H^2_{sim} to the diagonal of the residual covariance only affected accuracy by up to 1.02% (Fig. 6B).

When H^2_{sim} of TP_{PS} was high, 0.6 (Fig. 7), adding samples from TP_H of equal heritability led to small and constant increases in accuracy with increasing N_p. When H^2_{sim} of added samples from TP_H was moderate, 0.3 and 0.4, there was an initial decrease in accuracy followed by a slow increase with increasing N_p. However, even for the largest N_p, 365, adding individuals from TP_H led to a –4.5 and 0% change in accuracy for H^2_{sim} = 0.3 and 0.4, respectively. When H^2_{sim} of added samples from TP_H was low, 0.2, accuracy declined by 12% and did not show an eventual increase with increasing N_p. Adding a weight of 1 – H^2_{sim} to the diagonal of the residual led to improved accuracy when H^2_{sim} of TP_H was 0.2, 0.3, and 0.4 (Fig. 7B). Improvements in accuracy due to weighting ranged from 2.4 to 9.7%. The lower the H^2_{sim} of TP_H, the greater the benefit of using TP specific weights. However when H^2_{sim} of TP_H was very low at 0.2, adding individuals from TP_H never led to a net increase in accuracy, even when weighing was used.

In summary, using TP_H individuals when TP_{PS} individuals were available for training was always beneficial when H^2_{sim} of TP_H was greater than H^2_{sim} of TP_{PS}. In some cases, when H^2_{sim} of TP_{PS} was high and H^2_{sim} of TP_H was at least moderate, using TP_H and TP_{PS} individuals for training was beneficial when observations were properly weighted according to the heritability of their TP of origin.

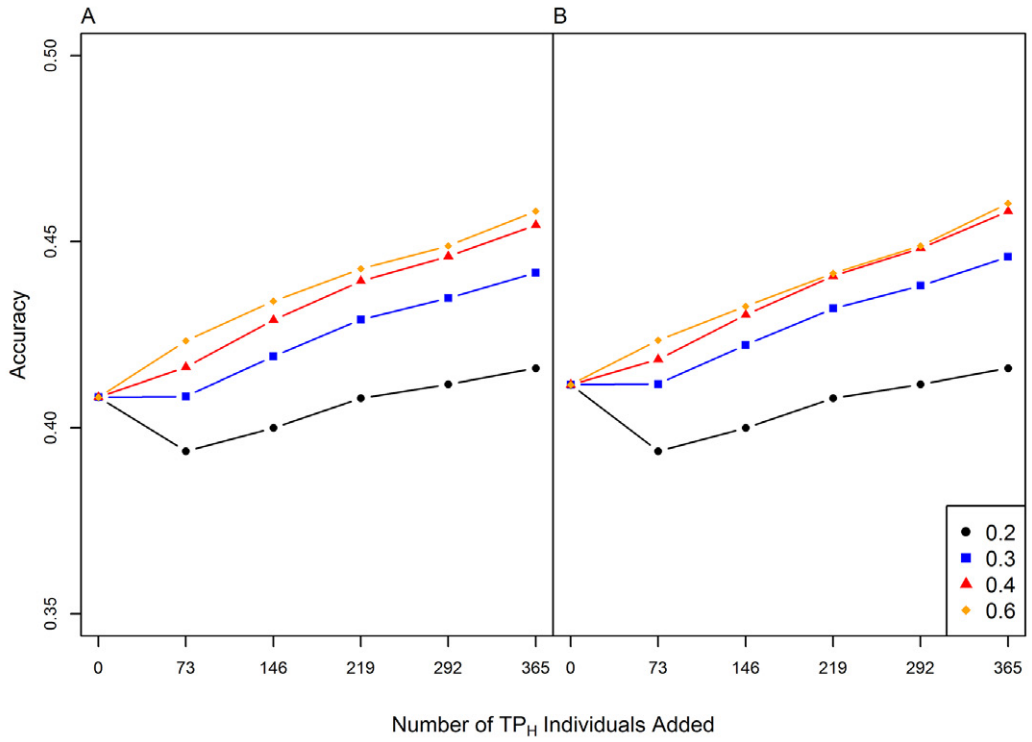

Figure 6. The effect of adding historic training population (TP_H) individuals to population-specific training population (TP_{PS}) when simulated heritability of TP_{PS} is 0.2 and simulated heritability of TP_H is 0.2, 0.3, 0.4, and 0.6. (A) Populations are weighted equally. (B) populations weighted according to simulated heritability.

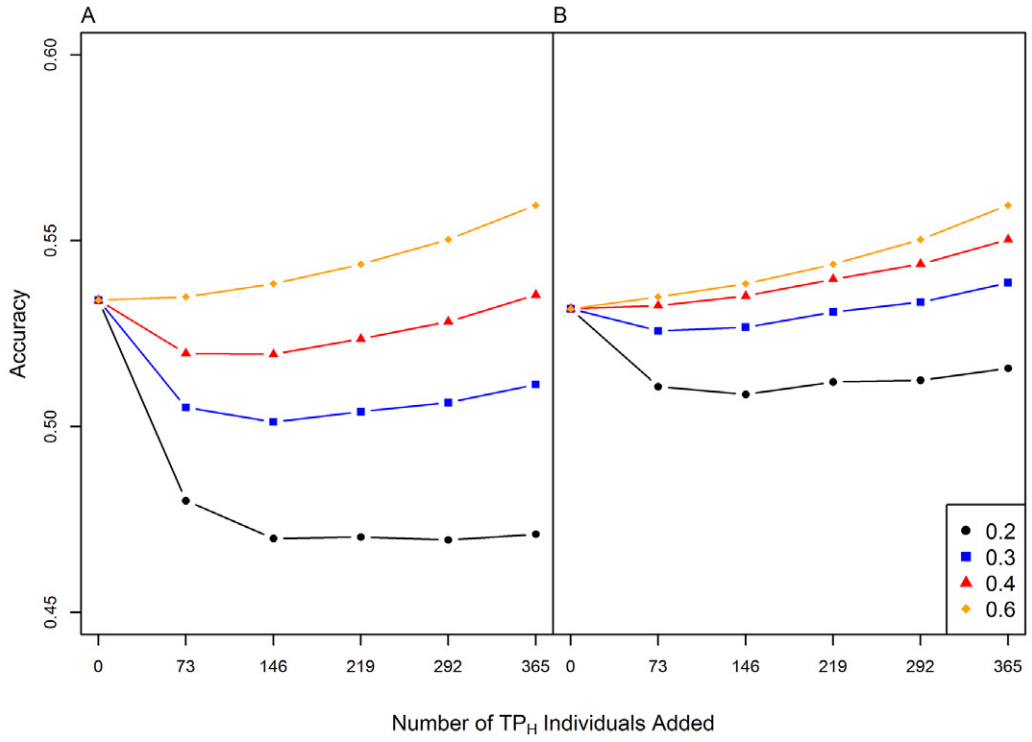

Figure 7. The effect of adding historic training population (TP_H) individuals to population-specific training population (TP_{PS}) when simulated heritability of TP_{PS} is 0.6 and simulated heritability of TP_H is 0.2, 0.3, 0.4, and 0.6. (A) Populations are weighted equally. (B) populations weighted according to simulated heritability.

Discussion

Populations

The significant population differentiation between the historical and selection candidate populations was a consequence of selection and genetic drift that occurred because the SC population was generated from only 14 founder lines from the historical population that were selected because they had at least moderate stem rust resistance and good agronomic performance. Selection and drift lead to a reduced rate of LD decay with physical distance in the SC population. The level of differentiation between historical and selection candidate populations due to selection and drift would be expected in plant breeding programs because each cycle of selection is founded by a small number of parents selected for intermating. However, breeding programs that use a lower selection intensity may experience less differentiation between historical and SC populations. Thus, breeding programs with lower selection intensities may be able to use historical data more successfully compared those that use higher selection intensities.

Accuracy Comparison

In general, the relative performance of a TP_H and TP_{PS} depends on the relative population sizes, trait heritability, levels of G×E, and genetic differentiation between the historical TP and the SCs. In this study, the lower accuracy from TP_H was primarily driven by the genetic differentiation between TP_H and the validation set. Accuracy from a TP_H could be as high or higher than that of a TP_{PS} in some scenarios. For example, based on linear regression of accuracy on TP size for both TP_{PS} and TP_H, if we were to add 225 more historical individuals to TP_H, TP_{PS} and TP_H accuracies may have been equivalent. Furthermore, if this study focused on a trait such as yield with low heritability on a single-plot basis and high G×E, population-specific training data from very few environments will be of low line-mean heritability and may not adequately sample the target environments of the breeding program. Lastly, a TP_H would likely be more effective if it has not undergone selection and subsequently a reduction in genetic variance (Bulmer, 1971) for the trait of interest. However, it may not be possible to find historical data from a breeding program where traits of interested have not undergone selection.

Training Population Optimization

The TP optimization methods based on PEV_{mean} and CD_{mean} enabled the selection of TPs from TP_H that were more accurate than those selected based on random sampling. Other studies evaluating TP optimization (Isidro et al., 2015; Rincent et al., 2012 found similar results; however, according to Isidro et al. (2015), TP optimization could be less accurate than random sampling if it resulted in a reduction in the phenotypic variance. Training population optimization would be useful if the historical dataset used for model training does not contain phenotypic data for all traits of interest. A subset of individuals from a historical dataset, selected to be predictive of the selection candidates based on CD_{mean} or PEV_{mean}, could be phenotyped for new traits of interest. This could reduce costs with potentially little to no sacrifice in accuracy compared to phenotyping all historical individuals.

Optimizing with respect to population progenitors appeared to be an effective way to select the appropriate subsample individuals for phenotyping and model updating in a GS program; however, the N_p that leads to the highest accuracy cannot be identified in advance. If phenotyping resources are limited, one could select N_p based on resource constraints and select individuals for phenotyping and model updating by optimizing with respect to the progenitors of the future SCs.

Interestingly, some optimal TPs lead to greater accuracy compared to the complete TP, suggesting that TP optimization could increase the overall GS prediction accuracy if it were possible to know in advance what N_p value could maximize accuracy. However, we caution that the ability to select an optimal TP that leads to greater accuracy compared to the complete TP is expected to be highly dependent on a given population and trait due to differences in population and family structure, nonadditive genetic variance, and LD between markers and causal loci. Assuming there is perfect linkage between markers and QTL, increasing N_p values will lead to an asymptotic increase in accuracy (Daetwyler et al., 2010; de Los Campos et al., 2013), and the complete TP will lead to higher accuracy than an optimal subset of the TP. Assuming imperfect linkage between markers and QTL and population or family structure, optimizing the TP for the SCs could increase accuracy because the estimated relationships between pairs of closely related individuals will be more accurate than the estimated relationships between less-related individuals (de Los Campos et al., 2013), and eliminating less related individuals could reduce noise in the relationship matrix. Aside from population genetic factors, the importance of nonadditive genetic variance for the trait of interest may also partially determine if TP optimization improves accuracy. Nonadditive genetic variance contributes to the covariance among close relatives only. When the TP is selected to be closely related to the SCs, more nonadditive genetic variance may be captured in G-BLUP, thus for traits where nonadditive genetic variance is important, TP optimization may lead to higher accuracy compared to the complete TP. This is similar to the effect of using a Gaussian kernel, where the genetic covariance can decrease more rapidly with genetic distance (Endelman, 2011). However, with optimization, the relationship between some pairs of individuals is effectively set to zero. Because of the various factors that affect the potential gain from training with an optimized TP rather than the complete TP, selecting optimal subsets from a TP is not yet a reliable way to improve accuracy.

Combining Training Population Data Sources

Our results showed that retaining historical data when data on close relatives was available reduced accuracy especially when the heritability of the historical data was low, the heritability of the close relative training data was high, and the observations were not weighted properly according to heritability. This has implications for prediction model updating. In a selection program, it may be better to discard older training data that is less relevant to the selection candidates as newer training data becomes available. However, when to discard training data will need to be determined empirically because it will depend on the selection intensity of the breeding program, the availability of data on close relatives, and quality of the historical data. For example, Asoro et al. (2011) evaluated the utility of adding historical oat (*Avena sativa* L.) lines to a training population going back in time and found that historical lines did not decrease accuracy, though the increase in accuracy they provided was small.

Conclusions

This case study found that historical data could be useful for initializing a GS-based breeding program where the selection candidates are founded by historical individuals. Although the highest accuracy could be achieved by phenotyping and model training with a subset of the selection candidate population itself, such an approach would require at least 2 yr of additional time to collect multilocation and multiyear data for all traits of interest, and may be less robust to G×E effects because data will be collected across relatively few environments. While historical data may be useful initially, this study suggests that once GS model updating can occur, it may be best to discard historical data and simply use the most recent data for model training.

Optimization of the historical TP was promising for selection of data subsets that were more predictive than randomly selected subsets. This would be useful when using a historical TP that lacks data for some key traits. To save resources, a subset of the TP_H, rather than the entire TP, could be phenotyped while the selection candidates are being developed.

We note that our conclusions are relevant to the germplasm and trait used in this study, and individual breeding programs will need to initiate GS programs to empirically determine the utility of historical data and at what point data should be discarded from the model training dataset. The utility of TP optimization should also be empirically studied in the context of a GS breeding program. More publicly available data generated by GS selection experiments and breeding programs will enable many such studies that will lead to the discovery of common trends across datasets.

Supplemental Information Available

Supplemental material is available at http://www.crops.org/publications/tpg.

Supplemental File S1. Genotyping-by-sequencing data for the historical and SC populations.

Supplemental File S2: Phenotypic data for the historical and SC populations.

Acknowledgments

This research was funded by The Bill and Melinda Gates Foundation (grants: Durable Rust Resistance in Wheat and Genomic Selection: The Next Frontier for Rapid Gains in Maize And Wheat Improvement), the USDA–ARS (Appropriation No. 5430-21000-006-00D), USDA National Institute of Food and Agriculture (NIFA)–Agriculture and Food Research Initiative (AFRI) grant support award number 2011-68002-30029, and United States Agency for International Development (USAID) support to the Feed the Future Innovation Lab for Applied Wheat Genomics (Cooperative Agreement No. AID-OAA-A-13-00051). Partial support for J. Rutkoski was provided by a USDA National Needs Fellowship Grant #2008-38420-04755 and an American Society of Plant Biology (ASPB)–Pioneer Hi-Bred Graduate Student Fellowship. Partial support was also provided by USDA–NIFA Hatch Project 149-430. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. The USDA is an equal opportunity provider and employer.

References

Amin, N., C.M. van Duijn, and Y.S. Aulchenko. 2007. A genomic background based method for association analysis in related individuals. PLoS ONE 2(12):E1274. doi:10.1371/journal.pone.0001274

Asoro, F.G., M.A. Newell, W.D. Beavis, M.P. Scott, N.A. Tinker, and J.L. Jannink. 2011. Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. Plant Gen. 4:132–144. doi:10.3835/plantgenome2011.02.0007

Astle, W., and D.J. Balding. 2009. Population structure and cryptic relatedness in genetic association studies. Stat. Sci. 24:451–471. doi:10.1214/09-STS307

Bates, D., and M. Maechler. 2010. lme4: Linear mixed-effects models using S4 classes. http://cran.r-project.org/package=lme4 (accessed 1 Dec. 2013).

Beeck, C.P., W.A. Cowling, A.B. Smith, and B.R. Cullis. 2010. Analysis of yield and oil from a series of canola breeding trials. Part I. Fitting factor analytic mixed models with pedigree information. Genome 53:992–1001. doi:10.1139/G10-051

Bernardo, R. 1994. Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci. 34:20–25. doi:10.2135/cropsci1994.0011183X003400010003x

Box, G.E., and D.R. Cox. 1964. An analysis of transformations. J. R. Stat. Soc., B 26:211–252.

Bulmer, M.G. 1971. The Effect of selection on genetic variability. Am. Nat. 105:201–211. doi:10.1086/282718

Crossa, J., G.D.L. Campos, P. Pérez, D. Gianola, J. Burgueño, J.L. Araus, D. Makumbi, R.P. Singh, S. Dreisigacker, J. Yan, V. Arief, M. Banziger, and H.J. Braun. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724. doi:10.1534/genetics.110.118521

Daetwyler, H.D., R. Pong-Wong, B. Villanueva, and J.A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021–1031. doi:10.1534/genetics.110.116855

Dawson, J.C., J.B. Endelman, N. Heslot, J. Crossa, J. Poland, S. Dreisigacker, Y. Manès, M.E. Sorrells, and J.L. Jannink. 2013. The use of unbalanced historical data for genomic selection in an international wheat breeding program. F. Crop. Res. 154:12–22. doi:10.1016/j.fcr.2013.07.020

de Los Campos, G., A.I. Vazquez, R. Fernando, Y.C. Klimentidis, and D. Sorensen. 2013. Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet. 9(7):E1003608.

de Roos, A.P.W., B.J. Hayes, and M.E. Goddard. 2009. Reliability of genomic predictions across multiple populations. Genetics 183:1545–1553. doi:10.1534/genetics.109.104935

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:e19379. doi:10.1371/journal.pone.0019379

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. 4:250–255.
Gilmour, A.R., B.J. Gogel, B.R. Cullis, and R. Thompson. 2009. ASReml user guide release 3.0. VSN International, Hemel Hempstead, UK.
Goddard, M. 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. Genetica (The Hague) 136:245–257.
Habier, D., R.L. Fernando, and J.C.M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397.
Haley, C.S., and P.M. Visscher. 1998. Strategies to utilize marker-quantitative trait loci associations. J. Dairy Sci. 81:85–97. doi:10.3168/jds.S0022-0302(98)70157-2
Hallauer, A.R., M.J. Carena, and J.B. Miranda Filho. 2010. Quantitative genetics in maize breeding. Iowa State Univ. Press, Ames.
Hayes, B.J., P.M. Visscher, and M.E. Goddard. 2009. Increased accuracy of selection by using the realized relationship matrix. Genet. Res. 91:47–60. doi:10.1017/S0016672308009981
Henderson, C.R. 1984. Applications of linear models in animal breeding. University of Guelph Press, Guelph, ON, Canada.
Isidro, J., J.L. Jannink, D. Akdemir, J. Poland, N. Heslot, and M.E. Sorrells. 2015. Training set optimization under population structure in genomic selection. Theor. Appl. Genet. 128:145–158. doi:10.1007/s00122-014-2418-4
Kennedy, B.W., and D. Trus. 1993. Considerations on genetic connectedness between management units under an animal model. J. Anim. Sci. 71:2341–2352.
Laloë, D. 1993. Precision and information in linear models of genetic evaluation. Genet. Sel. Evol. 25:1–20. doi:10.1186/1297-9686-25-6-557
Leutenegger, A.L., B. Prum, E. Génin, C. Verny, A. Lemainque, F. Clerget-Darpoux, and E.A. Thompson. 2003. Estimation of the inbreeding coefficient through use of genomic data. Am. J. Hum. Genet. 73:516–523. doi:10.1086/378207
Long, N., D. Gianola, G.J.M. Rosa, and K.A. Weigel. 2011. Long-term impacts of genome-enabled selection. J. Appl. Genet.: 467–480.
Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.
Peterson, R.F., A.B. Campbell, and A.E. Hannah. 1948. A diagrammatic scale for estimating rust intensity on leaves and stems of cereals. Can. J. Res. 26c(5):496–500. doi:10.1139/cjr48c-033
Piepho, H.P. 2009. Ridge regression and extensions for genomewide selection in maize. Crop Sci. 49:1165–1176. doi:10.2135/cropsci2008.10.0595
Poland, J.A., P.J. Brown, M.E. Sorrells, and J.L. Jannink. 2012a. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7(2):E32253. doi:10.1371/journal.pone.0032253
Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu, Y. Manes, S. Dreisigacker, J. Crossa, H. Sánchez-Villeda, M. Sorrells, and J.L. Jannink. 2012b. Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Gen. 5:103–113. doi:10.3835/plantgenome2012.06.0006
Pszczola, M., T. Strabel, H.A. Mulder, and M.P.L. Calus. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. 95:389–400. doi:10.3168/jds.2011-4338
R Development Core Team. 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Rincent, R., D. Laloë, S. Nicolas, T. Altmann, D. Brunel, P. Revilla, V.M. Rodríguez, J. Moreno-Gonzalez, A. Melchinger, E. Bauer, C.C. Schoen, N. Meyer, C. Giauffret, C. Bauland, P. Jamin, J. Laborde, H. Monod, P. Flament, A. Charcosset, and L. Moreau. 2012. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: Comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). Genetics 192:715–728. doi:10.1534/genetics.112.141473
Rutkoski, J.E., J. Poland, J.L. Jannink, and M.E. Sorrells. 2013. Imputation of unordered markers and the impact on genomic selection accuracy. G3 (Bethesda). 3:427–439.
Thompson, R., B. Cullis, A. Smith, and A. Gilmour. 2003. A sparse implementation of the average information algorithm for factor analytic and reduced rank variance models. Aust. NZ. J. Stat. 45:445–459.
Weir, B., and C. Cockerham. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370. doi:10.2307/2408641
Yu, L.X., A. Lorenz, J. Rutkoski, R.P. Singh, S. Bhavani, J. Huerta-Espino, and M.E. Sorrells. 2011. Association mapping and gene-gene interaction for stem rust resistance in CIMMYT spring wheat germplasm. Theor. Appl. Genet. 123:1257–1268. doi:10.1007/s00122-011-1664-y
Zadoks, J., T. Chang, and C.F. Konzak. 1974. A decimal code for the growth stages of cereals. Weed Res. 14:415–421. doi:10.1111/j.1365-3180.1974.tb01084.x