FACTORS AFFECTING RESPONSE TO RECURRENT GENOMIC SELECTION IN SOYBEANS

Vishnu Ramasubramanian[*] [†] **and William D Beavis**[*†]

[*] Department of Agronomy

[†]Bioinformatics and Computational Biology Graduate Program

Iowa State University, Ames, Iowa, USA - 50010

1) Vishnu Ramasubramanian

1525, Agronomy Hall, ISU

Ames, IA 50011

Phone: 515 598 6376

Email: vishnur@iastate.edu

2) William Beavis[*]

1208, Agronomy Hall, ISU

Ames, IA 50011

Phone: 800 450 4854 (4412)

Fax: 515 294 730

E-mail: wdbeavis@iastate.edu

Factors affecting Response to Recurrent Genomic Selection in Soybeans

Keywords: Genomic Selection, Genetic Response, Recurrent Selection, Soybean, Nested Association Mapping

Corresponding Author:

Vishnu Ramasubramanian

1525, Agronomy Hall, ISU

Ames, IA 50011

Phone: 515 598 6376

Email: vishnur@iastate.edu

Abstract

Herein we report the impacts of applying five selection methods across 40 cycles of recurrent selection and identify interactions with other factors on genetic response using simulated families of recombinant inbred lines derived from 21 homozygous soybean lines used for the Soybean Nested Association Mapping study. The other factors we investigated included the number of quantitative trait loci, broad sense heritability on an entry mean basis, selection intensity, and training sets. Both the rates of genetic improvement in the early cycles and limits to genetic improvement in the later cycles are affected by interactions among the factors. All genomic selection methods provided greater rates of genetic improvement (per cycle) than phenotypic selection, but phenotypic selection provided the greatest long term responses. Model updating significantly improved prediction accuracy and genetic response for three parametric genomic prediction models. Ridge Regression, if updated with training sets consisting of data from prior cycles, achieved greater rates of response relative to BayesB and Bayes LASSO GP models. A Support Vector Machine method, with a radial basis kernel, resulted in lowest prediction accuracies and the least long term genetic response. Application of genomic selection in a closed breeding population of a self-pollinated crop such as soybean will need to consider the impact of these factors on trade-offs between short term gains and conserving useful genetic diversity in the context of goals for the breeding program.

Background

Plant breeding programs consist of 1) recurrent genetic improvement projects, 2) variety development projects 3) trait introgression projects and 4) product placement projects (Fehr, 1991). Genetic improvement is assessed using realized genetic gain, which is an estimate of change of the average genotypic value for traits of interest across cycles of selection and inter-mating. Perhaps the most relevant trait for assessment of genetic gain is yield per unit land. Yield gains in many crop species in both developing and some developed countries have stagnated (Bhatia et al. 2008; Van Ittersum et al. 2013; Liu et al. 2016)

Since 2007, plant breeding programs have investigated genomic prediction (GP) models for application to genetic improvement (Bernardo & Yu 2007; Heffner et al. 2009). Three aspects of GP have been investigated: i) estimation of accuracies from prediction models (Habier et al. 2007; Goddard 2009; Zhong et al. 2009; Jannink 2010; Heffner et al. 2011; Bastiaansen et al. 2012; Bijma 2012;Wimmer et al. 2013; Lorenz 2013; Hickey et al. 2014, 2017), ii) selection of experimental lines to include in a crossing nursery (Cochran 1951; Bertan et al. 2007; Bos and Caligari 2008; Bernardo 2014) and iii) decisions about which lines to cross to create a new cycle of evaluation and selection (Akdemir & Sánchez 2016; Xu et al. 2017; Goiffon et al. 2017; Gorjanc et al. 2018).

Genetic improvement and cultivar development in crops that are primarily propagated through self-pollination (e.g., barley, canola, maize, oat, rice, soybean, sorghum, wheat, etc.,) involves the derivation of homozygous lines, i.e., replicable genotypes, for phenotypic evaluation across many environments and selection. Creation of replicable lines prior to evaluation and crossing assures that every line selected for initiating a cycle of genetic improvement will produce identical gametes in which all alleles are completely linked. Benefits from creating replicable

lines include: 1) ability to estimate repeatability of line performance across geography and time, 2) ability to estimate genotype by environment interactions (GxE) that can be used to distribute lines adapted to specific types of environments and 3) additive genetic variance among homozygous lines will be about twice as large as it would be if replicable lines were not created. As a consequence, genomic selection was not immediately adopted for line development after it was introduced in 2001 (Meuwissen et al. 2001). Animal breeders made most of the initial developments and improvements in application of genomic selection (Meuwissen 1997; Li et al 2008; Dekkers 2010; Dekkers 2012; de los Campos et al. 2013; Henryon et al. 2014). None-the-less, several crop breeders have successfully applied GS and demonstrated its short term advantage for genetic gain relative to phenotypic selection in barley, oats, wheat, soybean and maize (Bernardo 2008, 2014; Jannink et al. 2010; Asoro et al. 2011; Heslot et al. 2012; Nakaya and Isobe 2012; Hagan et al. 2012 ; Emily and Bernardo 2013; Crossa et al. 2014; Heslot et al. 2015; Liu et al. 2015; Beyene et al. 2015; Bassi et al. 2016; Marulanda et al. 2016; Jonas and de Koning 2013, 2016; Hickey et al. 2017; Goiffon et al. 2017).

Robertson (Robertson 1960) demonstrated that the rate of genetic gain for selection of traits with additive genetic architectures is greatest in the first few cycles, while later cycles asymptotically approach a limit. Hill and Robertson (1966) demonstrated that linkage can affect both the initial rate of genetic gain and selection limits. Felsenstein (1974) termed this the Hill-Robertson (HR) effect, where the magnitude of HR effect is determined by selection intensity and initial additive gene frequencies. The HR effect plays a role only in moderate/weak selection, where effective population size (Ne) plays a role in determining the probability of fixation through drift (Comeron et al. 2008). In the absence of epistasis and strong selection pressure, linkage dominates and the effect of linkage disequilibrium (LD) on rate of gain under selection can be

ignored (Felsenstein 1965, 1974; Kimura 1965; Nagylaki 1974). However, the HR effect can increase negative linkage disequilibrium as the linkage distance between two loci decreases (Comeron et al. 2008). In addition, greater selection intensities reduce the genetic potential of founder populations. The drop in genetic potential during the selection process can be prevented by reintroduction of discarded favorable alleles (Robertson 1960; Hill and Robertson 1966; Maynard and Haigh 1974; Comeron et al. 2008).

Previous studies point to differences among models for prediction accuracy and genetic response, but indicate that specific outcomes depend on conditions of genome organization, population structure, genetic architecture and selection intensity (Lorenz 2013; Rutkoski et al. 2015; Goiffon et al. 2017; Matei et al. 2018; Norman et al. 2018). In addition, training population size affects prediction accuracy and genetic gain. Larger training population sizes result in more accurate predictions and greater short term genetic improvement (Akdemir et al. 2015; Jarquin et al. 2016; Xavier et al. 2017). However, estimates of accuracy approach a limit depending on genome organization and population structure. Prediction accuracy is improved if the training set is selected to represent a larger proportion of a heterogeneous population for a targeted population of environments (Xavier et al. 2016; Jarquin et al. 2018).

Marker density also impacts accuracy and short term genetic gains, with a dense marker set performing better than a sparse set. But the improvement in accuracy and gain reaches a limit within a range of marker density depending on LD associated with the population structure. Empirical and simulation studies on maize, soybean and other crops have identified such a threshold value of marker density above which increasing marker density doesn't result in significant improvements in accuracy and genetic gain (de Roos et al. 2009; Schulz-Streeck et al. 2012; Hickey et al. 2014; Xavier et al, 2016; Norman et al. 2018)

Accuracy and gain are greater for traits with high heritability as a larger fraction of genotypic variance is represented in the phenotypic variance. Genetic architecture also impacts accuracy and gain (Wimmer et al. 2013). One of the important factors that contribute to higher rates of short term genetic improvement is shorter time per cycle of GS relative to PS. Phenotyping requires an evaluation phase conducted across multiple years, which is skipped with GS, which can reduce the interval between evaluation and crossing to create a new cohort of progeny by two or three years for most crops, depending on the number of generations that can be grown in continuous nurseries (Heffner et al. 2009).

The type of GP model used affects both prediction accuracy and the magnitudes of residual deviations from predicted values (Long et al. 2011; Heslot et al. 2012; Howard et al. 2014). Comparison of various models for GP have established the important role of genetic architecture (de los Campos et al. 2010, 2013; Howard et al. 2014). Howard et al (2014) compared the performance of a set of 14 GP models on F2 and Backcross simulated populations with additive and epistatic genetic architectures. Parametric methods such as Ridge Regression-BLUP and Bayesian regression methods in the mixed effects modeling framework perform well for traits with additive genetic architecture, whereas non-parametric machine-learning methods such as Neural Networks and Support Vector Machines provide more accurate predictions for traits with epistatic genetic architectures (Howard et al. 2014, 2016, and 2017). Prediction accuracies are essentially the same for all GP models applied to data with additive genetic architectures (Long et al. 2010, 2011; Guo et al. 2012; Howard et al. 2014). This was also the case in a hybrid maize line development program for a set of six traits in recombinant inbred lines derived from bi-parental families (Bernardo and Yu 2007). However, a Reproducing Kernel Hilbert Space (RKHS) model (De Los Campos et al. 2009, 2010), a semi-parametric method, showed a higher

accuracy in F2:3 populations that are comprised of a larger proportion of heterozygotes (Liu et al. 2018). A bagging method that combined RKHS and Bayes-B (BB) demonstrated the best prediction accuracy in SoyNAM population for yield, height and maturity (Xavier et al. 2016). In another study of experimental data, Bayesian methods had similar or better prediction accuracies than SVM and Multi-Layer Perceptron (MLP) (Montesinos-López et al. 2019).

Given the difficulty of conducting long-term genomic selection in experimental systems, simulations have been used to examine trends across multiple cycles of recurrent genomic selection (Habier et al. 2007; Goddard 2009; Jannink 2010; Bastiaansen et al. 2012; Bijma 2012; de los Campos et al. 2013; Liu et al. 2015; Michel et al. 2016). These studies have shown that selection response, represented as standardized genotypic values, is expected to be faster with GS in early cycles and then decrease in later cycles relative to standardized genetic response with phenotypic selection (PS) (Jannink 2010).

Parametric and non-parametric GP models have different impacts on simulated genetic gains (Bernardo and Yu 2007; Habier et al. 2007). While, Ridge Regression (RR) (Endelman 2011) and Bayesian methods (Pérez and de los Campos 2014) were associated with similar short-term genetic gains, but different long-term gains. In the initial cycles of recurrent selection genetic gains and estimates of accuracy were similar between Bayes LASSO (BL) and RR methods, whereas long-term limits to genetic response were better with BL and genetic variance was maintained through later cycles (Liu et al. 2015). In the initial cycles of recurrent selection genetic gains and estimates of accuracy were similar between BL and RR, whereas long-term limits to genetic response were better with BL. It was also noted that genetic variance was maintained through later cycles when selection was based on predicted values from BL (Liu et al. 2015).

Decreased prediction accuracy of GP models in recurrent GS is often due to decay of LD between marker loci and QTL, loss of relationships between lines in early and later cycles of selection or a combination of both (Habier et al. 2007; Zhong et al. 2009; Hickey et al. 2014; Liu et al. 2015; Müller and Melchinger 2017, 2018). Shrinkage based methods used for estimation of marker effects have an impact on relatedness of lines that are selected. The impact of relationships on prediction accuracies is greater for RR than BL because RR is more dependent on relationships within a population, whereas accuracy of bayesian models such as BL is more dependent on LD between marker loci (ML) and QTL, whereas accuracy of BayesB is dependent on both components (Habier et al. 2007; Zhong et al. 2009; Liu et al. 2015). Closely related lines tend to be selected as parents with RR, whereas BL maintains a lower rate of inbreeding. Maintaining a lower rate of inbreeding results in greater long-term genetic gains. (Meuwissen 1997; Li et al. 2008; Akdemir and Sánchez 2016).

Prediction accuracies can be maintained across cycles of selection by updating GP models with new genotypic and phenotypic information from each cycle of selection (Jannink 2010; Liu et al. 2015; Müller et al. 2017, 2018). One approach is to use training data from only the current cycle of selection. With this approach, predictions do not take into account relationships between the current population and the founder population or populations from previous cycles of selection. At the other extreme, data from all prior cycles of selection can be included with data from the current cycle in the training set. However, there are practical computation limits to the number of prior cycles of selection that can be included for training GP models. In particular we found that Bayesian and SVM GP model training time (run time in hours/cycle) increases exponentially with the size of training populations and requires intensive computing resources that are difficult to obtain (Figure S1).

158 Recurrent cycles of selection could generate populations with little genetic covariance with the
159 founder population, so inclusion of data from early cycles in the training set of later cycles may
160 have limited value. The actual structure of genetic covariance that emerges over cycles of
161 recurrent selection will affect number of cycles of data that need to be included in the training set
162 to obtain accurate predictions. However, as noted, the actual impact of including data from prior
163 cycles on accuracies of GP models depends on whether the models rely on relationships among
164 lines or on LD between ML and QTL. Moreover, the practice of model updating involves
165 phenotyping, which can adversely affect the relative advantage of GS over PS in terms of gain
166 per unit of time as phenotyping takes additional growing season(s) for each cycle (Heffner et al.
167 2009; Rutkoski et al. 2015; Matei et al. 2018). In practice, animal breeders use training data from
168 up to three prior cycles of selection including training data from current cycle (personal
169 communication, Jack Dekkers). Practical guidelines for training sets have not been established
170 for recurrent selection of crop species, in particular soybeans.

171 Recognizing the relatively slow genetic improvement of yield and other polygenic traits of
172 soybeans in the corn-soybean agricultural systems of the primary soybean production region in
173 the United States, the North Central Soybean Research Program (NCSRP) supported public
174 soybean breeders to utilize information from the SoyNAM genome wide association study to
175 evaluate implementation of GS in soybean breeding populations
176 (https://www.ncsrp.com/NCSRP_research.html#yield). As a precursor to experimental
177 investigations of recurrent GS in Soybean we utilized simulations (Cooper et al. 2002) based on
178 the genomic organization and population structure of the SoyNAM founders to evaluate genetic
179 responses to five selection methods, three selection intensities, three genetic architectures
180 responsible for 0.3 or 0.7 of the total phenotypic variance (H - broad sense heritability on an

entry mean basis) and four types of training sets across 40 cycles of recurrent selection. While the outcomes are specific to soybean genomes adapted to the primary soybean production region, there are implications for genetic improvement of all line development programs of diploids that utilize derivation of homozygous lines for evaluation and selection.

Methods

*Simulations and Treatment Design***.** The impact of number of QTL, selection intensity, heritability, training set and selection methods on response to selection across 40 cycles of recurrent selection were evaluated using 306 combinations of factors. Explicitly the treatments consisted of three numbers of simulated QTL and three selection intensities, two values for non-genetic variance, five selection methods and four types of training sets used to update four genomic prediction models. In summary the treatment design consists of 18 combinations of factors for phenotypic selection (PS) plus 288 combinations of factors for genomic selection (GS) methods for a total of 306 combinations of factors. Each set of factor combinations was replicated with ten simulated recurrent selections across 40 cycles resulting in 3060 simulations with 122400 outcomes. Note that different training sets are irrelevant for PS and thus the treatment design is not a complete factorial.

Simulated soybean RILs were generated by crossing *in silico* 20 homozygous SoyNAM founder lines with IA3023 to generate 20 distinct F_1 progeny. The F_1 progeny from each of the 20 crosses were self-pollinated *in silico* for five generations to generate 100 RILs per family. The resulting 2000 RILS from 20 families had segregating genotypic information at 4289 genetic loci. Based on NAM genotypic data, we simulated alleles from common founder line with a frequency of 0.9 and alleles from other founder lines with a frequency of 0.1.

Subsets of 40, 400, and 4289 SNP marker loci were designated as QTL. The QTL were distributed evenly throughout the genome, and each contributed equal additive effects of 5/-5, 0.5/-0.5, or 0.05/-0.05 units respectively to the total genotypic value the simulated RILs. Thus, all three genetic architectures had the same potential to create genotypic values ranging from +200 to -200 genotypic units in the initial founder sets of RILs. Phenotypic values were simulated by adding non-genetic variance sampled from an N (0, σ) distribution to the simulated genotypic values, where σ was determined by the heritability on an entry mean basis among the initial sets of founder sets of RILs. Broad sense heritability on an entry mean basis (H) values of 0.7 and 0.3 were simulated for each of the three sets of QTL. After the phenotypic values were simulated in the initial founding sets of RILs, the non-genetic variance was held constant across subsequent cycles of selection.

For each cycle of recurrent selection, 1%, 2.5% or 10% of the most positive phenotypic or predicted genotypic values among 2000 simulated RILs were selected as parents to inter-mate for the next cycle (Figure 1). This corresponds to selection intensities of 2.67, 2.34 and 1.75 in terms of standardized selection differential, $\imath$.

Based on previous results from Howard et al (2014), four GS methods were evaluated. Ridge Regression (RR) was selected to represent a frequentist parametric model. Bayes-B (BB) and Bayesian LASSO (BL) were selected to represent parametric bayesian models and Support Vector Machine with Radial Basis Kernel (SVM-RBF) represented a non-parametric method of machine learning. Ridge regression was implemented with a method that employs expectation maximization to obtain Restricted Maximum Likelihood estimates of marker effects (Xavier 2019). This computational method is faster than the popular implementation of ridge regression in rrBLUP package (Endelman 2011) and produces values that are highly correlated with the

predictions based on the rrBLUP package (Figure S2). The BGLR package (Perez and de los Campos 2014) provided implementations of BB and BL models. The 'Rgtsvm' package in R was used as an implementation of the SVM with RBF kernel method (Wang et al. 2017). 'Rgtsvm' implements SVM training on GPUs with computing time several hundred times less than that required for the implementation in 'caret' package on high performance computing clusters, with similar prediction accuracies and estimates of mean squared errors (Figure S3). The parameters used to train GP models are provided in Table 3.

A preliminary analysis of training sets on genotypic values and prediction accuracies was conducted using RR models trained with data from the current cycle as well as 3, 5, 6, 8, 10, 12, and 14 prior cycles. The results were compared with responses from the RR model updated with cumulative training set comprised of data from all prior cycles and with no updating using prior cycles. Training sets for each cycle were obtained by randomly sampling 1600 RILs from the set of 2000 simulated RILs in each cycle. The most accurate predictions and maximum genetic response was obtained with training data that is cumulatively added every cycle (Figures S4 and S5). The results indicate that including 3-5 prior cycles of training data did not significantly improve prediction accuracies and responses relative to models that were not updated. Also, the standardized genotypic values and prediction accuracies obtained using 10 to 14 prior cycles of data in the training set were not significantly different than results based on training sets consisting of all prior cycles. Based on the results of this preliminary study, we investigated responses to recurrent selection using training sets consisting of up to 14 prior cycles of selection as well as data from the current cycle. After the 14^{th} cycle, training data consisted of only the 14 prior cycles of recurrent selection and before the 14^{th} cycle, training data from all prior cycles

were included. For purposes of this manuscript we use the phrase 'model updating' to refer to retraining GP models with up to 14 previous cycles of training data (Figure 2).

Evaluation Metrics. The standardized genotypic value, R_s (1), was estimated every cycle as the change in genotypic value from the average genotypic value of 2000 RILs derived from the initial founders and standardized to the maximum genotypic potential (200 units) among the founders (Meuwissen et al. 2001; Liu et al. 2015).

$$R_s = \frac{R_c}{(R_m - R_0)} \quad (1)$$

R_s - Standardized genotypic value
R_0 - Average genotypic value of RILs produced by founders
R_c - Average genotypic value in cycle c – R_0
R_m - Maximal possible genotypic value (=200)

The maximum genotypic value (M_{gv}) among the RIL's selected in cycle c is a metric used to evaluate the best RIL produced each cycle, while the standardized genotypic variance (SVg) defined as the estimated genotypic variance divided by the estimated genotypic variance of the initial population, was used to evaluate the loss of genotypic variability. Note that values for the SV_g range from zero to one. Estimated Linkage disequilibrium (LD) among pairs of marker loci on all 20 chromosomes was evaluated as the deviation of observed gametic frequency of alleles at a pair of loci from the product of the individual allele frequencies, assuming independence (Weir 1996). GP models were assessed using the estimated prediction accuracies (r_{ps}), defined as the estimated linear correlation (Pearson) between predicted and simulated genotypic values and with estimated Mean Squared Error (MSE), defined as the sum of the squared deviations of the predicted genotypic values from the simulated values.

Modeled response to recurrent selection. The averaged Rs for each cycle, c, of recurrent selection were modeled with a linear first order recurrence equation:

$$f_0(c)y_{(c+1)} + f_1(c)y_{(c)} = g(c)$$

Where c is a sequence of integers from 0 to 39 representing each cycle of recurrent selection and f_0, f_1 and g are constant functions of c. By rearranging the equation we note that the response in cycle c+1 can be represented as

$$y_{(c+1)} = -\frac{f_1(c)}{f_0(c)}y_{(c)} + \frac{g(c)}{f_0(c)}$$

Since the ratios $f_1(c)/f_0(c)$ and $g(c)/f_0(c)$ are constants, we can represent the response in cycle c+1 as

$$y_{(c+1)} = \alpha y_{(c)} + \beta \qquad (2)$$

If y_0 specifies the average phenotypic value of the first generation of RILs derived from the founders, then (2) has a unique solution (Goldberg 1958):

$$\begin{aligned} y_c &= \alpha^c y_0 + \beta\frac{1-\alpha^c}{1-\alpha} \quad \text{if } \alpha \neq 1 \\ y_c &= \alpha^c y_0 + \beta c \qquad \text{if } \alpha = 1 \end{aligned} \qquad (3)$$

An alternative representation of (3) for the situation of $\alpha \neq 1$ is

$$\begin{aligned} y_c &= \alpha^c(y_0 - y') + y' \\ \textit{with } y' &= \frac{\beta}{1-\alpha}, \end{aligned}$$

, where α is less than 1 for genotypic response to recurrent selection and y' represents the asymptotic limit to selection (Goldberg 1958). To illustrate, values of the sequence of c=0 to 39,

with $y_0 = 0$, $\alpha = 0.9$ and $\beta = 15$, are plotted in Figure 3. The curve can be interpreted as response to selection as a function of the frequencies of alleles with additive selective advantage, selection intensity, time and effective population size (Robertson 1960).

The parameters, y_0, α and β, were estimated with a non-linear least squares method implemented in the 'nls' function of the R base package and the 'nlsList' and 'nlme' functions in the nlshelper and 'nlme' packages (Pinheiro and Bates 2000; Baty et al. 2015; Pinheiro et al. 2019).

Analyses of variance (ANOVA) *across 40 cycles of recurrent selection.* The purpose of the ANOVA is to evaluate significant differences in the modeled response pattern of PS and four GP models, based on three genetic architectures, with two levels of non-genetic contributions to the phenotypes, three selection intensities and four training sets for the GP models. The influence of multiple factor treatment combinations on estimated non-linear regression models have not been implemented in standard statistical software packages that report the analysis of variance in terms of sums of squares and traditional 'F-tests'. For discussions on the challenges of using standard F-test for non-linear mixed effects models see (Pinheiro et al. 2000; Baty et al. 2015; Pinheiro et al. 2019). Consequently, we analyzed the variance among modeled responses using AIC, BIC and Likelihood metrics that were grouped based on combinations of treatment variables consisting of selection methods, training sets, selection intensities, number of simulated QTL and H (Table 1).

We conducted analyses of variance using non-linear least squares on modeled (3) responses grouped by treatment factors (Table 3). In order to provide a balanced data table for analyses by nlme responses from PS were assumed constant for training set levels resulting in responses for 360 combinations of treatment factors. The impact of these factors and their interactions on the modeled response was analyzed using the groupedData function in R to generate data partitions

conditioned on groups of factors. Estimates of modeled parameters were retained as fixed effects and deviations from estimated means conditioned on grouping variables were modeled as random effects using 'nlme' R package. Multiple analyses of 'nlme' objects representing the models were used to identify combinations of factors with significant effects on the non-linear response model. More information on the analyses can be found in the R package 'SoyNAMPredictionMethods' (http://gfspopgen.agron.iastate.edu/SoyNAM_PredictionMethods.html).

Analyses and Data Availability

Simulated data and codes are available as part of R package 'SoyNAMPredictionMethods' (File S1). All supplemental material including the R package has been uploaded to Figshare and can be found at https://figshare.com/s/8dba182a46fe1a28c1af. Documentation to use package is available at http://gfspopgen.agron.iastate.edu/SoyNAM_PredictionMethods.html. SoyNAM genotypic and phenotypic data is available in SoyBase database (Grant et al. 2010).

Results

Modeled genotypic values across 40 cycles of recurrent selection. Average genotypic values grouped on each of five factors provide an overview of differences in rate of response and limits to response among levels within a factor (Figure S6 - S10). The response of averaged Rs (Figure 3) were modeled with recurrence equation (3) and the results are consistent with predicted values and theory by Robertson (1960). There is strong evidence from the analyses of variance (Table 3) that the response of modeled genotypic values across cycles of selection depend on interactions among all simulated factors. The most parsimonious model requires unique

330 estimates of α, and β in equation (3) for each of the 306 combinations of factors (File S2).
331 Estimates of α, and β for all factor combinations are provided in File S2.

332 *Prediction accuracies in the founding sets of RILs:* Estimates of prediction accuracies, r_{ps}, of GP
333 models trained with the initial set of 2000 F_5-derived RILs from 20 crosses ranged from 0.75-
334 0.82 for H of 0.7 and ranged from 0.38 - 0.49 for H of 0.3 (Figure 4). The initial r_{ps} for both H
335 values was best with BB and poorest with the SVM-RBF. The nQTL had little effect on r_{ps}
336 within either value of 0.7 or 0.3 for H. RR and BL had smaller magnitude MSE values than BB
337 and SVM RBF for all numbers of simulated QTL and both values for H (Figure 4).

338 *Comparisons of Selection Methods*. Both the rates and limits of response from selection in terms
339 of R_s are influenced by the five selection methods and their interactions with training sets,
340 selection intensities, nQTL, and H (Figure 5 and S11). Small nQTL, high values for H and high
341 selection intensities resulted in the greatest initial rates (per cycle) of response while large nQTL,
342 low values of H and relaxed selection intensity showed less rapid initial responses but realized
343 greater R_s values for all PS and GS methods with and without updated training sets. Relative to
344 PS, most GS methods provided greater initial rates of response per cycle, but limits to selection
345 responses depended on the other factors (Figure 5). In this section, we describe results from top
346 10% selected fraction for all the GS methods as the differences among GS methods are more
347 pronounced with relaxed selection intensity. Results from top 1% and 2.5% selected fraction are
348 described in later sections.

349 When GP models are not updated, BB demonstrated the largest R_s in the early cycles for 40
350 QTL responsible for 70% of phenotypic variability in the initial population, whereas for 400 and
351 4289 QTL PS demonstrated greater responses than all GS methods after the 10^{th} cycle (Figure 5
352 and 6). For H value of 0.3 in the initial populations, BB also had the largest Rs values for 40, 400

353 and 4289 QTL in the early cycles. Whereas after 10th cycle, PS demonstrated the largest Rs
354 values for 40, 400 and 4289 simulated QTL (Figure S11 and S12; File S3). In contrast, when the
355 GP models are updated with training sets consisting of data from up to 14 prior cycles of
356 recurrent selection, GS using RR models demonstrated largest the Rs values for 40, 400 and
357 4289 QTL for both heritabilities (Figure 7 and S13; File S3).

358 If the BB models are not updated with data from prior cycles, then the Rs from 40 simulated
359 QTL and 0.7 H in the initial population were 10 to 16% greater than they were with PS in the
360 first five cycles. For the same genetic architecture, recurrent selection with BL and RR models
361 resulted in Rs values that were 4 to 13% greater than Rs values from PS in the first five cycles
362 (Figure 5, 6 and File S3). After five cycles, PS resulted in greater responses than all of the GS
363 methods when training sets were not updated. If training sets were not updated after the initial
364 evaluations of RILs and the genetic architectures consisted of 400 and 4289 QTL responsible for
365 0.7 H in the initial sets of RILs, then RR and Bayesian GP models provided greater genetic
366 responses than PS only in the first 2-3 cycles and thereafter PS demonstrated 5-50 % greater
367 standardized genetic responses (File S3).

368 When the RR GP model is updated with data from up to 14 previous cycles of recurrent
369 selection, the Rs values for selection on 40 simulated QTL responsible for 70% of phenotypic
370 variability in the initial sets of RILs was 1.5 to 10% greater than PS for the first 10 cycles of
371 recurrent selection, but after the 10th cycle it was similar to PS (Figure 5 and 7; File S3). For the
372 same genetic architecture the Rs values from BB and BL with model updating were 14.4% and
373 12% greater than PS respectively for the first five cycles of recurrent selection. After the fifth
374 cycle PS resulted in greater Rs values. If the RR model is updated with data from up to 14
375 previous cycles of recurrent selection, the responses to selection of RILs with 400 simulated

QTL responsible for 70% of phenotypic variability in the initial sets of RILs, then the Rs values were 3 to 15% greater than PS across all 40 cycles. The limits of response to selection using BB and BL models were 1 to 14% greater than PS for up to 20 cycles. If the RR model is updated with data from 14 previous cycles of recurrent selection, the Rs values with 4289 QTL were 10-15% greater than PS for 40 cycles (Figure 5 and 7; File S3). Likewise recurrent selection using BB and BL models resulted in greater responses than PS for 40 cycles (Figure 5 and 7; File S3). Similar, albeit distinctive, comparisons among outcomes from GP models with model updating for genetic architectures responsible for 0.3 of the phenotypic variance in the initial sets of RILs are described in File S4.

As noted, if GP models are not updated, PS provides greater responses than GS in the early cycles of recurrent selection, whereas if the parametric GP models are updated then for some combinations of treatment factors, Rs values are greater than PS across many cycles of recurrent selection.

In addition to comparing outcomes of PS with GS, we next consider comparisons of the evaluation metrics with and without updated training sets among the GS methods. When GP models are updated, RR models resulted in 10% greater response than RR without updating for 40 simulated QTL responsible for 70% of phenotypic variability among RILs in the initial cycle. Model updating resulted in 30% and 60% greater responses for 400 and 4289 QTL respectively (Figure S14). Recurrent selection with updated BB models resulted in 4% greater Rs values than without updating for 40 QTL and resulted in 22% and 57% greater responses for 400 and 4289 simulated QTL respectively (Figure S14). Recurrent selection with updated BL models resulted in 3% greater responses than without updating for 40 simulated QTL. If the genetic architecture consisted of 400 QTL and 4289 QTL, updated BL models resulted in 21% and 51% greater

399 responses respectively (Figure S14). SVMRBF when updated with training sets demonstrated no
400 significant improvement in Rs values relative to SVMRBF without updating for all genetic
401 architectures (Figure S14 and S15). If the genetic architecture explains only 30% of the
402 phenotypic variability in the initial sets of RILs, the relative improvements in Rs values using
403 updated training sets are greater than simulated QTL that explain 0.7 of the phenotypic variance
404 (Figure S15). Percent gain in response in GS with model updating relative to response from GS
405 without updating for forty cycles of selection are provided in File S5.

406 When GP models are not updated, Mgv were consistently greater with PS than the four GP
407 models. Among GP models without updating, BB provided the best Mgv, while SVM-RBF had
408 the smallest Mgv (Figure 8 and S16). When GP models are updated, the pattern depends mostly
409 on the number of QTL. For initial population H values of both 0.7 and 0.3 and 40 simulated
410 QTL, Mgv are similar for RR, Bayesian GP models and PS, whereas for 400 QTL, RR produces
411 greater Mgv than PS and Bayesian GS methods. For 4289 QTL, RR and Bayesian models
412 produce greater Mgv with PS. GS with SVMRBF produced the least desirable Mgv for 40, 400
413 and 4289 QTL.

414 When GP models are updated, the standardized genotypic variance (Sgv) declines at a rate
415 similar to the rate of decrease when the models are not updated (Figure 9 and S17). Model
416 updating significantly improved estimated prediction accuracies, $r_{ps,}$ for all GP models except
417 SVMRBF (Figure 10 and S18). Among RR and Bayesian GP models, model updating has a
418 slightly larger impact on estimated accuracies with RR than with Bayesian GP models (Figure 10
419 and S18). For all selection methods pairwise LD among markers on the same chromosome
420 decreases across cycles of recurrent selection (Figure S19-S23). LD decreased slowest with PS
421 (Figure S19). Decay of LD in early and late cycles of selection are similar among parametric GP

models and SVMRBF with relaxed selection intensities. By the 20th cycle of recurrent selection, LD approached zero for all selection methods and there was no evidence that selection methods affected linkage disequilibrium (LD) differentially in the earlier cycles. The rates of LD decay are lower when GP models are updated compared to GP models without updating (Figure S19-S23).

When GP models are not updated, rates and limits of response standardized to change in genotypic variance (RsVar) with RR and Bayesian GS methods are similar to PS for 40, 400 and 4289 simulated QTL responsible for both 70% and 30% of phenotypic variability in the initial population. There are no significant differences among GS methods as well as among GS and PS for most treatment combinations. Both rates and limits of RsVar are also comparable across the three nQTL and selection intensity levels. However, with top 10% selected fraction, PS demonstrated greater limits of RsVar for 400 and 4289 QTL (Figure S24 and S25).

However when GP models are updated, the patterns of RsVar are significantly different among GS methods and PS and are also dependent on nQTL, selection intensity and heritability (Figure S26 and S27). With 0.7 heritability, there are no significant difference in RsVar among GS methods for 40 simulated QTL for top 1%, 2.5% and 10% selected fraction. Whereas for 400 and 4289 QTL with top 2.5% and 10% selected fraction, RR GS method demonstrated greater limits of RsVar followed by PS and Bayesian GS methods. Gain in RsVar with RR GS is even larger for 0.3 H treatment with relaxed selection intensities (Figure S26 and S27). SVMRBF demonstrated the least limits of RsVar for treatment combinations with and without model updating (Figure S24 -S27).

Selection Intensity. Truncation selection using selection intensities of 2.67 and 2.34 resulted in similar rapid responses across the early cycles of recurrent selection (Figure S28 – S31). These

selection intensities associated with selection of 1% and 2.5% top RIL's limited responses to the early cycles (first 10-20 cycles) for all combinations of selection methods and number of simulated QTL. In contrast a selection intensity of 1.75, associated with retaining top 10% of the RILs each cycle, provided continued opportunities for response to selection for additional cycles depending on the number of simulated QTL and selection method (Figure 5-7 and S11- S15).

Selection intensity also impacted the effectiveness of updating GP models. With relaxed selection intensity, the proportional gains are greater with model updating than without updating for the three parametric GP methods, whereas stringent selection intensities with model updating for these methods resulted in proportionally lower gains (Figure S14, S15 and S32 – S35). Percent gain in responses from GP model updating for all three selection intensities are described in File S6.

Mgv's also increased with the relaxed selection intensities and differences of Mgv's among GS methods due to selection intensities were affected by model updating (Figure 8, S16 and S36-S39). As expected, the standardized genotypic variance, Sgv, decreased rapidly with increasing selection intensities (Figure 9, S17, and S40–S43). However, model updating didn't have any significant effect on rate of decrease in Sgv. Coincident with the relationship between loss of Sgv and selection intensities, LD also decays with increasing selection intensities (Figure S19). In contrast to Sgv, the rates of loss of LD among the GS methods are slowed by including training sets that are updated (Figure S20-S23).

By adjusting selection intensity, genetic variance in the population can be maintained for longer number of cycles without contribution from other sources of variation such as mutation or migration. Rate of decrease in genetic variance increases with increasing selection intensity Relaxed selection intensity of top 10% showed the least rate of decay of genetic variance,

468 whereas selection with top 1.0% demonstrated the largest rate of decay of genetic variance
469 (Figure 9, S17, S40 – S43).

470 The rates at which estimated prediction accuracies decline under relaxed selection intensities
471 result are less than the loss with stringent selection intensities. As with other metrics, the impact
472 of selection intensity on estimated predictions accuracies depend on the GS method, model
473 updating, number of QTL and heritability. Differences among GP models with updating are most
474 pronounced for high selection intensities, 4289 QTL and high heritability (Figure 10, S18, S44 -
475 S47).

476 *Number of simulated QTL*. The number of simulated QTL had the largest consistent impact on
477 differences among the response curves for Rs values and Mgv's. This is most obvious by noting
478 that the Rs values (Figure 5, S11, S28- S31) are as high as 80% of the maximum value of 200
479 and reach the limit in less than ten cycles of recurrent selection if there are 40 simulated QTL. In
480 contrast, Rs values are no greater than 40% of the maximum value of 200 and reach the limit in
481 10-15 cycles of recurrent selection for most selection methods if there are 400 simulated QTL
482 while Rs values are no greater than 15% of the maximum value and only begin to approach the
483 limit after 20 cycles if there are 4289 simulated QTL (Figures 5, S11 and S28 – S31).

484 The loss of Sgv's across cycles for the simulated number of QTL is consistent with the rate at
485 which limits to response from selection are approached (Figure 9, S17, S41 -43). As genotypic
486 variance is eliminated, response to selection approaches a limit. Likewise the estimated
487 prediction accuracies approach zero as the genotypic variance approaches zero (Figure 10, S18
488 ,S44 –S47) although the covariance between the two metrics depend on the other simulated
489 factors.

Heritability Most of the differences in response metrics between the two simulated H values have been reported above. In summary, Rs values are greater for simulated QTL responsible for 70% of the initial phenotypic variance than Rs values for all genetic architectures responsible for 30% of the initial phenotypic variance (Figure 5 and S11). These trends are correlated with the other response metrics, in particular prediction accuracies of the GP models.

The loss of estimated prediction accuracies are greater with H values of 0.3 than 0.7 with relaxed selection intensities. Other combinations of selection intensity and heritability require model updating to provide reasonable GP model prediction accuracies and achieve greater responses across more cycles of selection. As we would expect, for all combinations of selection intensity and number of QTL, limits of response and loss of genotypic variance are greater with H values of 0.7 than 0.3 (Figure 10, S18, and S44 – S47).

Discussion

We did not use a coalescent process to establish a set of founders. Rather we used a set of publicly available founders to create a breeding population similar to that found in soybean variety development projects in the primary soybean production region of North America. In both academic and commercial soybean development projects it is not unusual to cross multiple lines to a single exceptional variety and generate 50 to 150 RILs from each cross. Of the 4289 SNP markers with genotypic scores for the SoyNAM population 3818 were polymorphic among the 20 families that were used as founders for the simulations. On average, 773 were polymorphic within a family with a variance among families of 34 polymorphic loci. In the initial founding set of RILs, the average heterozygosity per SNP locus across 20 families was 0.09. The average estimated G_{st} value across the genome for the initial founding set of RILs from 20 families was 0.32. 'Gst' is a measure of sub-population differentiation estimated as ratio of difference between sub population expected heterozygosity and total expected heterozygosity to total expected heterozygosity (Jombart 2008; Ryman and Leimar 2009; Jombart and Ahmed 2011). Relative to previous reported founders in self-pollinated crops derived using a coalescent process, our simulations began with a structure more likely to be found in actual soybean breeding populations and with much less, albeit more realistic, genetic diversity.

Recurrent selection was conducted in the context of a cultivar development process in which RILs are created and phenotypically evaluated each cycle. Field evaluations of replicated RILs provide information on the repeatable performance of RIL's across environments before they are used for creating a new cycle of genetic improvement. Comparisons among selection methods assumed equal time required to develop and evaluate RIL's for each cycle of recurrent selection. In practice, one of the advantages of using recurrent GS relative to recurrent PS is that

526 response/per year will be greater with GS. Even if both GS and PS require the same amount of
527 time to develop RILs, selection with GS methods can be conducted without multiple years of
528 phenotypic evaluations (Heffner et al. 2009).

529 An alternative recurrent selection strategy is to decouple genetic improvement from variety
530 development (Gaynor et al. 2017). By separating the two types of breeding projects, GS can be
531 applied every generation using training sets composed of genotypic and phenotypic data obtained
532 on RILs derived in a previous cycle of recurrent selection. The consequence of using GP models
533 to select and cross individual F1 plants instead of RILs should be to create more opportunities for
534 recombination. In such a system training sets are updated with data obtained at regular intervals
535 from annual field trials, although the training sets may be several selection cycles removed from
536 the cycle used to create the RILs used in current field evaluations. Implementing such a two part
537 strategy in soybean would require significant changes. In particular, intercrossing soybeans is
538 labor intensive and expensive. Therefore, a two part system that requires intercrossing individual
539 plants, rather than replicable lines, three times per year will require significant investments.
540 Whether such investments can be justified will need to be investigated. The results reported
541 herein provide a basis for comparing alternative breeding strategies with the established
542 strategies for genetic improvement and cultivar development.

543 We simulated only simple additive genetic architectures. Alleles at adjacent QTL were assigned
544 alternating positive genotypic values to alleles from each founder. Also recognize that some
545 marker alleles are identical to the QTL alleles. Thus loss of prediction accuracies across cycles is
546 not due to loss of LD, rather it is entirely due to loss of genotypic variance. Thus, the results
547 indicate best case scenarios for implementation of GS methods. For experimental applications of
548 GS, maintaining relationships among RILs in selected populations and LD between marker-QTL

549 will help maintain prediction accuracy even when the genotypic variance is reduced in later
550 cycles of selection (Meuwissen 1997; Zhong et al. 2009; Wimmer et al. 2013; Müller et al. 2017,
551 2018).

552 Based on estimates of the number of effectively segregating genomic segments among RILs 400
553 and 4289 simulated QTL were associated with only about 40 and 400 effective haplotype blocks
554 respectively (data not shown). The reduced number of effective linkage blocks means that the net
555 magnitude of the allelic effects for each linkage block is less than for the case with 40 simulated
556 QTL. For example by summing +/- alleles across a linkage block consisting of five QTL the
557 simulated genetic effects will be 2a (+ - + - +) or -2a (- + - + -). Because the maximum net effect
558 is 200 units regardless of the number of QTL, a large number of linkage blocks with small net
559 effects requires more recombination for RILs in the population to realize the genetic potential of
560 the founders. Hence we hypothesize if GS is applied to individuals in every generation such as
561 proposed by (Gaynor et al. 2017) the increased recombination will effectively release useful
562 genetic variability that is locked up in linkage blocks when RIL's are used as the selection units.

563 Our simulations also used constant values of non-genetic variance to produce phenotypic
564 variance every cycle. While it is possible to simulate constant values of H, rather than σ_r, across
565 all cycles of selection, the translation of such simulated values to experimental field plots is
566 equivalent to planting (exponentially) increasing numbers of field plots as the genotypic variance
567 decreases. Because closed soybean genetic improvement programs cannot afford to increase the
568 number of plots to offset the loss of genotypic variance, we decided that it would be more
569 realistic to assume that the non-genetic variance rather than H will be constant over cycles of
570 recurrent selection.

Modeled genotypic responses across 40 cycles of recurrent selection. The use of a first order recurrence equation (3) to model genotypic responses (Rs) to recurrent selection provided a method for comprehensive analyses of variance including interactions among the combinations of factors. Curiously, we have not been able to find previous applications of recurrence equations to model responses from genomic selection methods in the literature. We hope that our explanation of how to implement such models in available R packages will encourage others to investigate the dynamics of recurrent selection across multiple cycles in both closed and open genetic populations.

Prediction accuracies of GP models trained with founder populations are consistent with prediction accuracies observed in previous studies (Habier et al. 2007; Goddard 2009; Howard et al. 2014). Prediction accuracies of RR and Bayesian methods with the population structure and genomic architecture created with SoyNAM founders are similar to previously reported accuracies for additive genetic architectures responsible for 0.7 and 0.3 of the phenotypic variance while the non-parametric SVM-RBF model produced the least accurate predictions from the founding sets of RILs (Habier et al. 2007; Howard et al. 2014).

The numbers of QTL also impacted prediction accuracy, with greater prediction accuracies for smaller numbers of QTL and a given training set size. With large numbers of QTL and smaller simulated additive effects, larger training sets are required to maintain similar levels of accuracy as compared to small number of QTL with large additive effects (Goddard 2009; Dekkers 2010).

Comparisons of Selection Methods. The comparisons of GS methods using all metrics suggest that RR when updated with training data from prior cycles will provide the best long-term response to selection in a closed breeding population derived from founders of SoyNAM. Without updating the models using training data from prior cycles, responses to recurrent

selection using parametric GS methods are greater than PS in early cycles, but are not better in later cycles, which is of concern for closed population improvement. RR models that use updated training sets achieve greater responses than Bayesian GP models and PS in both the early and late cycles of recurrent selection, whereas Bayesian GP models, when updated, result in greater responses than PS only in the early cycles. With relaxed selection intensities and large number of QTL, it is possible to achieve similar or greater limits of response than PS with RR and Bayesian GS methods with updating without compromising on rate of response with GS in early cycles given certain conditions.

It has been suggested that long-term response in GS will never be greater than PS because PS maintains genotypic variance for a longer number of cycles of recurrent selection (Goddard 2009; Zhong et al. 2009; Jannink 2010). First, prediction accuracies of GS methods will decrease with every cycle of selection, whereas decay of prediction accuracies is irrelevant for PS. Second, GS results in rapid loss of genetic variance in the initial cycles, which results in approaching the asymptotic response limit in early cycles. However, response per cycle in the breeder's equation is also dependent on other factors such as selection intensity and loss of standardized genotypic variance in addition to prediction accuracy. Depending on other factors, GS still has the potential to realize greater gains for some treatments.

Model updating with training sets from prior cycles improves the relationship between training sets and validations sets compared to responses to GS without updated training sets resulting in greater prediction accuracies for RR GS method in late cycles of recurrent selection. Model updating also resulted in greater responses standardized to the rate of decrease of genetic variance in selected populations with updated RR GP models. Response standardized to change

in genotypic variance is very similar to efficiency of converting lost genetic diversity into genetic gain discussed by Gorjanc et al (Gorjanc G et al. 2018).

Instead of estimating efficiency as a slope of genetic gain regressed linearly on change in genotypic variance, we evaluated the non-linear pattern of changes. Given similar rates of decrease of genetic variance among RR and Bayes GS methods, the differences in responses among methods is possibly due to greater efficiency of translating loss of genetic diversity to gain with RR GS in later cycles of selection, although for most treatment factors the limits of response with RR and Bayesian GS methods are about the same.

It is not clear from our examination of patterns of relationship among selected lines and LD decay why prediction accuracies are greater in late cycles for RR compared to Bayesian methods when GP models are updated. Even though model updating resulted in significantly different rates of change of genotypic variance in selected populations, average heterozygosity, rate of inbreeding and rate of LD decay relative to GS methods without updating, there are no significant differences among GS methods.

Selection increased linkage disequilibrium (LD) while decreasing genetic variance in early cycles of selection, whereas in late cycles, LD decayed due to recombination. In PS, LD is influenced mostly by selection intensity, whereas the effects of LD and linkage are complicated in GS. Prediction accuracies of GP models are dependent on LD among marker-QTL and affect selection of lines and genotypic variance of selected populations, which in turn affect prediction accuracies in subsequent cycles of selection. In addition to the factors investigated in this study additional forces can affect LD including selection, drift, epistasis, and GxE effects (Roze and Barton 2006; Hickey et al. 2014). From these data it is not clear which of these forces has the greatest impact on LD, but use of simulations could be used to address this question.

Partitioning the contribution of LD and linkage blocks to prediction accuracy could help estimate the contribution of each of these factors to prediction accuracy of GP models. However, it requires a design similar to that employed by (Müller et al. 2017, 2018) for synthetic populations. In their study, populations with unrelated training and prediction sets with LD and SNP based relationship estimates showed low prediction accuracy and low genetic response in recurrent GS. This is similar to GS without updating in this study. Whereas populations with relationship between training and prediction sets with LD and SNP based relationship estimates is similar to GS with model updating, which showed greater prediction accuracy and greater genetic response in both of their published studies.

The genetic covariance among cycles of RILs depends on the number of QTL and influences the number of prior cycles that needs to be included in the training data set. If the genetic covariance among cycles of RILs changes rapidly, then data from fewer prior cycles are needed for accurate genomic predictions. Small numbers of QTL with larger effects requires fewer prior cycles of data to achieve good prediction accuracies and responses, whereas large numbers of QTL each with smaller additive effects produced slower changes to the genetic co-variance structure among cycles and required training data from more prior cycles to maintain good prediction accuracies and responses.

The rate at which genetic variance is reduced from cycle to cycle depends on number of QTL, heritability, selection intensity and GS methods (Figures 9 and S16) and we could model the loss of genotypic variance using a first order recurrence equation (3). However, it is likely that the estimated genotypic variance from each cycle is underestimated because the covariance among QTL, due to LD among QTL, was not taken into account (Bulmer 1971, 1976; Lehermeier et al.

2017). None-the-less, the changes of genotypic variance from cycle to cycle mirrored the response of Rs and Mgv, decreasing to zero when the Rs and Mgv reached their limits.

We also assumed that the training set size increases with every cycle when we cumulatively add data from prior cycles to the training data set. Increasing training population size confounds estimation of effects of maintaining relationship between training and prediction sets on accuracy. But, there are some approaches to remedy this situation which still needs to be evaluated. One of the potential approaches is to randomly sample subsets of data from each of the prior cycles to maintain a constant cumulative training population size. It is also possible to assign weights to the samples from prior cycles to place more weight on data from more recent cycles. This essentially involves determining the optimal trade-off for sample sizes and weights that will assure maximum prediction accuracy with minimal computational requirements. Some aspects of this optimization problem have been addressed. Akdemir et al (2015) devised a genetic algorithm for selecting optimal training populations to minimize prediction error variance and Xavier et al (2017) developed sampling methods for training bayesian GP models. Another approach is to retrain GP models at regular frequencies instead of updating every cycle, while maintaining a constant training population size. Application of sampling and optimization methods for selection of training populations for GP model updating requires further study in the context of recurrent genomic selection.

Selection Intensity. Replicated responses to high values of selection intensity quickly reach a limit in five to ten cycles of recurrent selection. Replicated responses to lower values of selection intensity consistently result in greater gains over more cycles, indicating that genetic drift is the most likely mechanism for loss of genotypic variance. These constraints on plant breeding programs are well characterized (Brisbane and Gibson 1995; Hayes et al. 2009; Jannink 2010;

Hung et al. 2012; Liu et al. 2015; Akdemir and Sánchez 2016; Yabe et al. 2016). The optimal trade-off between achieving genetic gains and maintaining genetic diversity depend on the objectives. For example, if the objective is to enter and capture market share in a short time frame, then maintenance of genetic diversity is not important.

Most simulation studies apply a constant selection intensity across cycles of selection. The effects of applying a dynamic selection strategy is an interesting problem. We hypothesize that a strategy consisting of applying different selection intensities, optimized for each cycle, will achieve improved long-term genetic response by differentially emphasizing genetic variance and genetic response across multiple cycles of selection.

Conclusion

Using simulations we examined the impact of five factors on genetic response through 40 cycles of recurrent selection in simulated Soybean populations for 306 unique combinations of treatment factors in a factorial design. Two of these factors, number of QTL and heritability are characteristics that the plant breeder cannot alter in a closed breeding population, whereas the plant breeder can alter selection intensity, GP models and GP model updating.

Interactions among the five factors significantly affect the rate of response and limits to selection response. Responses to selection approached a limit after 10 to 20 cycles of recurrent selection for most combinations of the five factors. If GP models are not updated after an initial training set, then bayesian parametric methods performed better than ridge-regression and machine learning methods in terms of prediction accuracy and limits of response to selection. If GP models are updated by re-training using current and prior cycles of genotypic and phenotypic data, then ridge regression models demonstrated greater prediction accuracies and limits of

706 responses. Relaxed selection intensity resulted in greater limits of responses but at slightly lower
707 rates of genetic gain in early cycles compared to stringent selection intensities.

708 By utilizing a first order recurrence equation to model response to selection and evaluating
709 factors that could affect the response to selection our results provide an objective framework for
710 further investigations of selection methods for genetic improvement and line development
711 projects in self-pollinated crops. For example, it has been suggested that the occasional
712 emergence of a line with exceptional characteristics that dominate all other genotypes for several
713 cycles of selection is likely due to a unique combination of genetic alleles, i.e., epistasis. If
714 epistasis is simulated, will the first order recurrence equation be sufficiently robust to model
715 response to selection? Second, we employed truncation selection with a crossing strategy that
716 uses high contributions from top ranked RILs among the selected population of RILs in each of
717 the simulated cycles of selection. We did not consider relationships among selected RILs nor the
718 trade-offs between genetic gain, genetic variance (inbreeding) when selecting RILs to cross.
719 There exist multi-objective optimization breeding methods such as genomic mating and optimal
720 cross selection (Rutten et al. 2002; Woolliams et al. 2015; Akdemir and Sánchez 2016; Gorjanc
721 et al. 2018), that could provide both greater rates and limits of responses across cycles.
722 Importantly, we simulated a closed breeding population in which culled lines were not resampled
723 for discarded favorable alleles, nor did we simulate a more open breeding population with
724 exchange of lines among breeding programs. There are evolutionary algorithms such as island
725 models with parameters that fit these more realistic breeding programs (Yabe et al. 2016).
726 Evaluation of these other types of selection strategies and design of hybrid strategies that
727 incorporate elements of two or more distinct algorithmic approaches is a promising direction of
728 exploration.

729

730

731 **Acknowledgements**

732 Funding for this research was provided by the Department of Agronomy, Iowa State University,
733 the North Central Soybean Research Program and a grant (1830478) from the Engineering
734 Directorate of NSF. Supplemental funding for large scale computing was enabled by the
735 Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by
736 National Science Foundation. XSEDE resources consisted of research allocations (DMS190015
737 & DMS190018) on PSC-Bridges Large Memory nodes for the parametric GP model update
738 simulations. The Iowa State University -Pronto GPU cluster enabled computation of SVM model
739 update simulations. We also want to acknowledge Matheus de Krause for discussions on fitting
740 non-linear models using ‘nlme’ package, Lizhi Wang for efficient programs to simulate meiosis
741 and Alencar Xavier for sharing an efficient expectation maximization method for fitting ridge
742 regression GP models.

743

744

745

746

747

Reference

Akdemir D, Sánchez JI: **Efficient Breeding by Genomic Mating.** *Frontiers in Genetics* 2016, **7**.

Akdemir D, Sanchez JI, Jannink JL: **Optimization of genomic selection training populations with a genetic algorithm.** *Genet Sel Evol* 2015, **47:**38.

Bassi FM, Bentley AR, Charmet G, Ortiz R, Crossa J: **Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.).** *Plant Science* 2016, **242:**23-36.

Bastiaansen JWM, Coster A, Calus MPL, van Arendonk JAM, Bovenhuis H: **Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures.** *Genetics Selection Evolution* 2012, **44:**3.

Baty F, Ritz C, Charles S, Brutsche M, Flandrois J-P, Delignette-Muller M-L: **A Toolbox for Nonlinear Regression in R: The Package nlstools.** *Journal of Statistical Software; Vol 1, Issue 5 (2015)* 2015.

Bernardo R: **Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years.** *Crop Science* 2008, **48:**1649.

Bernardo R: **Genomewide Selection of Parental Inbreds: Classes of Loci and Virtual Biparental Populations.** *Crop Science* 2014, **54:**2586.

Bernardo R, Yu J: **Prospects for Genomewide Selection for Quantitative Traits in Maize.** *Crop Science* 2007, **47:**1082-1090.

Bertan I, Carvalho F, Oliveira Ad: **Parental selection strategies in plant breeding programs.** *J Crop Sci Biotechnol* 2007, **10:**211-222.

Beyene Y, Semagn K, Mugo S, Tarekegne A, Babu R, Meisel B, Sehabiague P, Makumbi D, Magorokosho C, Oikeh S, et al: **Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress.(RESEARCH)(Author abstract).** 2015, **55:**154.

Bhatia VS, Singh P, Wani SP, Chauhan GS, Rao AVRK, Mishra AK, Srinivas K: **Analysis of potential yields and yield gaps of rainfed soybean in India using CROPGRO-Soybean model.(Report).** *Agricultural and Forest Meteorology* 2008, **148:**1252.

Bijma P: **Long-term genomic improvement – new challenges for population genetics.** *Journal of Animal Breeding and Genetics* 2012, **129:**1-2.

Bos I: *Selection methods in plant breeding by Izak Bos and Peter Caligari.* 2nd ed.. edn. Dordrecht: Dordrecht : Springer; 2008.

Brisbane J, Gibson J: **Balancing selection response and rate of inbreeding by including genetic relationships in selection decisions.** *International Journal of Plant Breeding Research* 1995, **91:**421-431.

Bulmer MG: **The Effect of Selection on Genetic Variability.** *The American Naturalist* 1971, **105:**201-211.

Bulmer MG: **The effect of selection on genetic variability: a simulation study.** *Genet Res* 1976, **28:**101-117.

Cochran WG: **Improvement by Means of Selection.** In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*; *1951; Berkeley, Calif.* University of California Press; 1951: 449-470.

Comeron JM, Williford A, Kliman RM: **The Hill-Robertson effect: evolutionary consequences of weak selection and linkage in finite populations.** *Heredity (Edinb)* 2008, **100:**19-31.

Cooper M, Podlich D, Micallef K, Smith O, Jensen N, Chapman S, Kruger N: **Complexity, quantitative traits and plant breeding: a role for simulation modelling in the genetic improvement of crops.** *Quantitative genetics, genomics and plant breeding'(Ed MS Kang) pp* 2002**:**143-166.

Crossa J, Pérez P, Hickey J, Burgueño J, Ornella L, Cerón-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y, et al: **Genomic prediction in CIMMYT maize and wheat breeding programs.** *Heredity* 2014, **112:**48-60.

Crow JF, Kimura M: **An introduction to population genetics theory.** *An introduction to population genetics theory* 1970.

De Los Campos G, Gianola D, Rosa GJM, Weigel KA, Crossa J: **Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods.** *Genetics Research* 2010, **92:**295-308.

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL: **Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding.** *Genetics* 2013, **193:**327-345.

de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM: **Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree.** *Genetics* 2009, **182:**375-385.

de Los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D: **Prediction of complex human traits using the genomic best linear unbiased predictor.** *PLoS Genet* 2013, **9:**e1003608.

de Roos APW, Hayes BJ, Goddard ME: **Reliability of genomic predictions across multiple populations.** *Genetics* 2009, **183:**1545-1553.

Dekkers JCM: **Use of high-density marker genotyping for genetic improvement of livestock by genomic selection.** *CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources* 2010, **5:**1-13.

Dekkers JCM: **Application of Genomics Tools to Animal Breeding.** *Current Genomics* 2012, **13:**207-212.

Diers BW, Specht J, Rainey KM, Cregan P, Song Q, Ramasubramanian V, Graef G, Nelson R, Schapaugh W, Wang D, et al: **Genetic Architecture of Soybean Yield and Agronomic Traits.** *G3: Genes|Genomes|Genetics* 2018, **8:**3367.

Emily C, Rex B: **Accuracy of Genomewide Selection for Different Traits with Constant Population Size, Heritability, and Number of Markers.** *The Plant Genome* 2013, **6**.

Endelman JB: **Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP.** *The Plant Genome Journal* 2011, **4:**250.

Fehr W: **Principles of cultivar development: theory and technique**. Macmillian Publishing Company; 1991.

Felsenstein J: **The Effect of Linkage on Directional Selection.** *Genetics* 1965, **52:**349.

Felsenstein J: **The Evolutionary Advantage of Recombination.** *Genetics* 1974, **78:**737.

Franco GA, Mark AN, William DB, Scott MP, Jean-Luc J: **Accuracy and Training Population Design for Genomic Selection on Quantitative Traits in Elite North American Oats.** *The Plant Genome* 2011, **4:**132-144.

Gao N, Li J, He J, Xiao G, Luo Y, Zhang H, Chen Z, Zhang Z: **Improving accuracy of genomic prediction by genetic architecture based priors in a Bayesian model.** *BMC Genet* 2015, **16:**120.

Goddard M: **Genomic selection: prediction of accuracy and maximisation of long term response.** *Genetica* 2009, **136:**245-257.

Goiffon M, Kusmec A, Wang L, Hu G, Schnable PS: **Improving Response in Genomic Selection with a Population-Based Selection Strategy: Optimal Population Value Selection.** *Genetics* 2017, **206:**1675.

Goldberg S: *Introduction to difference equations, with illustrative examples from economics, psychology, and sociology.* New York: New York, Wiley; 1958.

Gorjanc G, Gaynor RC, Hickey JM: **Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection.** *Theoretical and Applied Genetics* 2018, **131:**1953-1966.

Grant D, Nelson RT, Cannon SB, Shoemaker RC: **SoyBase, the USDA-ARS soybean genetics and genomics database.** *Nucleic acids research* 2010, **38:**D843.

Guo B, Sleper DA, Beavis WD: **Nested Association Mapping for Identification of Functional Markers.** *Genetics* 2010, **186:**373-383.

Guo Z, Tucker DM, Lu J, Kishore V, Gay G: **Evaluation of genome-wide selection efficiency in maize nested association mapping populations.** *Theoretical and Applied Genetics* 2012, **124:**261-275.

Habier D, Fernando RL, Dekkers JCM: **The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values.** *Genetics* 2007, **177:**2389.

Hagan S, Knowles J, Kell DB: **Exploiting Genomic Knowledge in Optimising Molecular Breeding Programmes: Algorithms from Evolutionary Computing (Evolutionary Computing for Molecular Breeding).** 2012, **7:**e48862.

Hayes BJ, Visscher PM, Goddard ME: **Increased accuracy of artificial selection by using the realized relationship matrix.** *Genetics Research* 2009, **91:**47-60.

Heffner EL, Jannink J-L, Iwata H, Souza E, Sorrells ME: **Genomic selection accuracy for grain quality traits in biparental wheat populations.(RESEARCH)(Author abstract)(Report).** *Crop Science* 2011, **51:**2597.

Heffner EL, Sorrells ME, Jannink J-L: **Genomic Selection for Crop Improvement.** *Crop Science* 2009, **49:**1.

Henryon M, Berg P, Sørensen AC: **Animal-breeding schemes using genomic information need breeding plans designed to maximise long-term genetic gains.** *Livestock Science* 2014, **166:**38-47.

Heslot N, Akdemir D, Sorrells ME, Jannink J-L: **Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions.** *Theoretical and Applied Genetics* 2014, **127:**463-480.

Heslot N, Jannink J-L, Sorrells ME: **Perspectives for Genomic Selection Applications and Research in Plants.** *Crop Science* 2015, **55:**1-12.

Heslot N, Yang H-P, Sorrells ME, Jannink J-L: **Genomic Selection in Plant Breeding: A Comparison of Models.** *Crop Science* 2012, **52:**146-160.

Hickey JM, Chiurugwi T, Mackay I, Powell W: **Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery.** *Nature genetics* 2017, **49:**1297.

Hickey JM, Dreisigacker S, Crossa J, Hearne S, Babu R, Prasanna BM, Grondona M, Zambelli A, Windhausen VS, Mathews K, Gorjanc G: **Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation.(RESEARCH)(Author abstract).** 2014, **54:**1476.

Hickey JM, Dreisigacker S, Crossa J, Hearne S, Babu R, Prasanna BM, Grondona M,

Zambelli A, Windhausen VS, Mathews K, Gorjanc G: **Evaluation of Genomic Selection Training Population Designs and Genotyping Strategies in Plant Breeding Programs Using Simulation.** *Crop Science* 2014, **54:**1476-1488.

Hill WG, Robertson A: **The effect of linkage on limits to artificial selection.** *Genetics Research* 2008, **89:**311-336.

Howard R, Beavis WD, Carriquiry A: **Evaluation of a Computational Diagnostic for Epistasis in Plant Breeding Populations.** *bioRxiv* 2016.

Howard R, Carriquiry AL, Beavis WD: **Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures.** *G3 (Bethesda)* 2014, **4:**1027-1046.

Howard R, Carriquiry AL, Beavis WD: **Application of Response Surface Methods To Determine Conditions for Optimal Genomic Prediction.** *G3: Genes|Genomes|Genetics* 2017, **7:**3103.

Hung HY, Browne C, Guill K, Coles N, Eller M, Garcia A, Lepak N, Melia-Hancock S, Oropeza-Rosas M, Salvo S, et al: **The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population.** *Heredity (Edinb)* 2012, **108:**490-499.

Jannink J-L: **Dynamics of long-term genomic selection.** *Genetics Selection Evolution* 2010, **42:**35.

Jannink JL: **Selection dynamics and limits under additive X additive epistatic gene action.** *Crop science* 2003, **v. 43:**pp. 489-497-2003 v.2043 no.2002.

Jannink JL, Lorenz AJ, Iwata H: **Genomic selection in plant breeding: from theory to practice.** *Brief Funct Genomics* 2010, **9:**166-177.

Jarquin D, Howard R, Xavier A, Das Choudhury S: **Increasing Predictive Ability by Modeling Interactions between Environments, Genotype and Canopy Coverage Image Data for Soybeans.** *Agronomy* 2018, **8:**51.

Jarquin D, Specht J, Lorenz A: **Prospects of Genomic Prediction in the USDA Soybean Germplasm Collection: Historical Data Creates Robust Models for Enhancing Selection of Accessions.** *G3: Genes|Genomes|Genetics* 2016.

Jombart T: **adegenet: a R package for the multivariate analysis of genetic markers.** *Bioinformatics* 2008, **24:**1403-1405.

Jombart T, Ahmed I: **adegenet 1.3-1: new tools for the analysis of genome-wide SNP data.** *Bioinformatics* 2011, **27:**3070-3071.

Jonas E, de Koning D-J: **Does genomic selection have a future in plant breeding?** *Trends in Biotechnology* 2013, **31:**497-504.

Jonas E, de Koning DJ: **Goals and hurdles for a successful implementation of genomic selection in breeding programme for selected annual and perennial crops.** *Biotechnology & genetic engineering reviews* 2016, **32:**18.

Kimura M: **Attainment of Quasi Linkage Equilibrium When Gene Frequencies are Changing by Natural Selection.** *Genetics* 1965, **52:**875.

Lehermeier C, de los Campos G, Wimmer V, Schön CC: **Genomic variance estimates: With or without disequilibrium covariances?** *Journal of Animal Breeding and Genetics* 2017, **134:**232-241.

Li Y, Kadarmideen HN, Dekkers JCM: **Selection on multiple QTL with control of gene diversity and inbreeding for long-term benefit.** *Journal of Animal Breeding and Genetics* 2008, **125:**320-329.

Liu H, Meuwissen TH, Sorensen AC, Berg P: **Upweighting rare favourable alleles increases long-term genetic gain in genomic selection programs.** *Genet Sel Evol* 2015, **47:**19.

Liu H, Sørensen AC, Meuwissen THE, Berg P: **Allele frequency changes due to hitch-hiking in genomic selection programs.** *Genetics Selection Evolution* 2014, **46:**8.

Liu X, Wang H, Wang H, Guo Z, Xu X, Liu J, Wang S, Li W-X, Zou C, Prasanna BM, et al: **Factors affecting genomic selection revealed by empirical evidence in maize.** *The Crop Journal* 2018.

Liu Z, Yang X, Lin X, Hubbard K, Lv S, Wang J: **Maize yield gaps caused by non-controllable, agronomic, and socioeconomic factors in a changing climate of Northeast China.** *Science of The Total Environment* 2016, **541:**756-764.

Long N, Gianola D, Rosa GJM, Weigel KA: **Application of support vector regression to genome-assisted prediction of quantitative traits.** *Theoretical and Applied Genetics* 2011, **123:**1065.

Long N, Gianola D, Rosa GJM, Weigel KA, Kranis A, GonzÁLez-Recio O: **Radial basis function regression methods for predicting quantitative traits using SNP markers.** *Genetics Research* 2010, **92:**209-225.

Lopez-Cruz M, Crossa J, Bonnett D, Dreisigacker S, Poland J, Jannink J-L, Singh RP, Autrique E, de los Campos G: **Increased Prediction Accuracy in Wheat Breeding Trials Using a Marker × Environment Interaction Genomic Selection Model.** *G3: Genes|Genomes|Genetics* 2015, **5:**569.

Lorenz AJ: **Resource Allocation for Maximizing Prediction Accuracy and Genetic Gain of Genomic Selection in Plant Breeding: A Simulation Experiment.** *G3: Genes|Genomes|Genetics* 2013, **3:**481.

Lorenz AJ, Smith KP, Jannink J-L: **Potential and Optimization of Genomic Selection for**

Fusarium Head Blight Resistance in Six-Row Barley. *Crop Science* 2012, **52:**1609-1621.

Marulanda J, Mi X, Melchinger A, Xu J-L, Würschum T, Longin C: **Optimum breeding strategies using genomic selection for hybrid breeding in wheat, maize, rye, barley, rice and triticale.** *Theor Appl Genet* 2016, **129:**1901-1913.

Matei G, Woyann LG, Milioli AS, de Bem Oliveira I, Zdziarski AD, Zanella R, Coelho ASG, Finatto T, Benin G: **Genomic selection in soybean: accuracy and time gain in relation to phenotypic selection.** *Molecular Breeding* 2018, **38:**117.

Maynard J, Haigh J: **The hitch-hiking effect of a favourable gene.** *Genetics Research* 2008, **89:**391-403.

Meuwissen T, Hayes B, Goddard M: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157:**1819-1829.

Meuwissen T, Hayes B, Goddard M: **Accelerating improvement of livestock with genomic selection.** *Annu Rev Anim Biosci* 2013, **1:**221-237.

Meuwissen TH: **Maximizing the response of selection with a predefined rate of inbreeding.** *Journal of animal science* 1997, **75:**934-940.

Meuwissen TH, Goddard ME: **Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data.** *Genetics Selection Evolution* 2004, **36:**261.

Michel S, Ametz C, Gungor H, Epure D, Grausgruber H, Loschenberger F, Buerstmayr H: **Genomic selection across multiple breeding cycles in applied bread wheat breeding.(Original Article)(Report).** *Theoretical and Applied Genetics* 2016, **129:**1179.

Mikel MA, Diers BW, Nelson RL, Smith HH: **Genetic Diversity and Agronomic Improvement of North American Soybean Germplasm.** *Crop Science* 2010, **50:**1219-1229.

Montesinos-López OA, Martín-Vallejo J, Crossa J, Gianola D, Hernández-Suárez CM, Montesinos-López A, Juliana P, Singh R: **A Benchmarking Between Deep Learning, Support Vector Machine and Bayesian Threshold Best Linear Unbiased Prediction for Predicting Ordinal Traits in Plant Breeding.** *G3: Genes|Genomes|Genetics* 2019, **9:**601.

Müller D, Schopp P, Melchinger AE: **Persistency of Prediction Accuracy and Genetic Gain in Synthetic Populations Under Recurrent Genomic Selection.** *G3: Genes|Genomes|Genetics* 2017, **7:**801.

Müller D, Schopp P, Melchinger AE: **Selection on Expected Maximum Haploid Breeding Values Can Increase Genetic Gain in Recurrent Genomic Selection.** *G3: Genes|Genomes|Genetics* 2018, **8:**1173.

Nagylaki T: **The evolution of one- and two-locus systems.** *Genetics* 1976, **83:**583.

Nakaya A, Isobe SN: **Will genomic selection be a practical method for plant breeding?** *Annals of Botany* 2012, **110:**1303-1316.

Norman A, Taylor J, Edwards J, Kuchel H: **Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy.** *G3: Genes|Genomes|Genetics* 2018.

Ortiz-Perez E, Cianzio SR, Wiley H, Horner HT, Davis WH, Palmer RG: **Insect-mediated cross-pollination in soybean [Glycine max (L.) Merrill] : I. Agronomic performance.** Iowa State University Digital Repository; 2007.

Pérez P, de los Campos G: **Genome-Wide Regression and Prediction with the BGLR Statistical Package.** *Genetics* 2014, **198:**483-495.

Pinheiro JC: **Mixed-effects models in S and S-PLUS** / *José C. Pinheiro, Douglas M. Bates.* New York: New York : Springer; 2000.

Ray J, Kilen T, Abel C, Paris R: **Soybean natural cross-pollination rates under field conditions.** *Environmental biosafety research* 2003, **2:**133-138.

Riedelsheimer C, Melchinger AE: **Optimizing the allocation of resources for genomic selection in one breeding cycle.** *Theoretical and Applied Genetics* 2013, **126:**2835+.

Rincker K, Nelson R, Specht J, Sleper D, Cary T, Cianzio SR, Casteel S, Conley S, Chen P, Davis V, et al: **Genetic Improvement of U.S. Soybean in Maturity Groups II, III, and IV.** *Crop Science* 2014, **0:**0.

Robertson A: **A Theory of Limits in Artificial Selection.** *Proceedings of the Royal Society of London Series B Biological Sciences* 1960, **153:**234.

Roze D, Barton NH: **The Hill–Robertson Effect and the Evolution of Recombination.** *Genetics* 2006, **173:**1793-1811.

Rutkoski J, Singh RP, Huerta-Espino J, Bhavani S, Poland J, Jannink JL, Sorrells ME: **Genetic Gain from Phenotypic and Genomic Selection for Quantitative Resistance to Stem Rust of Wheat.** *The Plant Genome* 2015, **8**.

Rutten MJM, Bijma P, Woolliams JA, van Arendonk JAM: **SelAction: Software to Predict Selection Response and Rate of Inbreeding in Livestock Breeding Programs.** *Journal of Heredity* 2002, **93:**456-458.

Ryman N, Leimar O: **GST is still a useful measure of genetic differentiation — a comment on Jost's D.** *Molecular Ecology* 2009, **18:**2084-2087.

Sukumaran S, Crossa J, Jarquin D, Lopes M, Reynolds MP: **Genomic Prediction with Pedigree and Genotype × Environment Interaction in Spring Wheat Grown in South and West Asia, North Africa, and Mexico.** *G3: Genes|Genomes|Genetics* 2017, **7:**481.

Takuno S, Terauchi R, Innan H: **The power of QTL mapping with RILs.** *PLoS One* 2012, **7:**e46545.

van Ittersum MK, Cassman KG, Grassini P, Wolf J, Tittonell P, Hochman Z: **Yield gap analysis with local to global relevance—A review.** *Field Crops Research* 2013, **143:**4-17.

Wang Z, Chu T, A Choate L, Danko C: **Rgtsvm: Support Vector Machines on a GPU in R**. 2017.

Wimmer V, Lehermeier C, Albrecht T, Auinger H-J, Wang Y, Schön C-C: **Genome-Wide Prediction of Traits with Different Genetic Architecture Through Efficient Variable Selection.** *Genetics* 2013, **195:**573.

Woolliams JA, Berg P, Dagnachew BS, Meuwissen TH: **Genetic contributions and their optimization.** *J Anim Breed Genet* 2015, **132:**89-99.

Xavier A: **Efficient Estimation of Marker Effects in Plant Breeding.** *G3: Genes|Genomes|Genetics* 2019, **9:**3855.

Xavier A, Jarquin D, Howard R, Ramasubramanian V, Specht JE, Graef GL, Beavis WD, Diers BW, Song Q, Cregan P, et al: **Genome-Wide Analysis of Grain Yield Stability and Environmental Interactions in a Multiparental Soybean Population.** *G3: Genes|Genomes|Genetics* 2017.

Xavier A, Xu S, Muir W, Rainey KM: **Genomic prediction using subsampling.** *BMC Bioinformatics* 2017, **18:**191.

Xu Y, Li P, Zou C, Lu Y, Xie C, Zhang X, Prasanna BM, Olsen MS: **Enhancing genetic gain in the era of molecular breeding.** *Journal of Experimental Botany* 2017, **68:**2641-2666.

Yabe S, Yamasaki M, Ebana K, Hayashi T, Iwata H: **Island-Model Genomic Selection for Long-Term Genetic Improvement of Autogamous Crops.** *PLoS One* 2016, **11:**e0153945.

Yu J, Buckler ES: **Genetic association mapping and genome organization of maize.** *Curr Opin Biotechnol* 2006, **17:**155-160.

Yu J, Holland JB, McMullen MD, Buckler ES: **Genetic Design and Statistical Power of Nested Association Mapping in Maize.** *Genetics* 2008, **178:**539.

Zhong S, Dekkers JCM, Fernando RL, Jannink J-L: **Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study.** *Genetics* 2009, **182:**355-364.

Figures

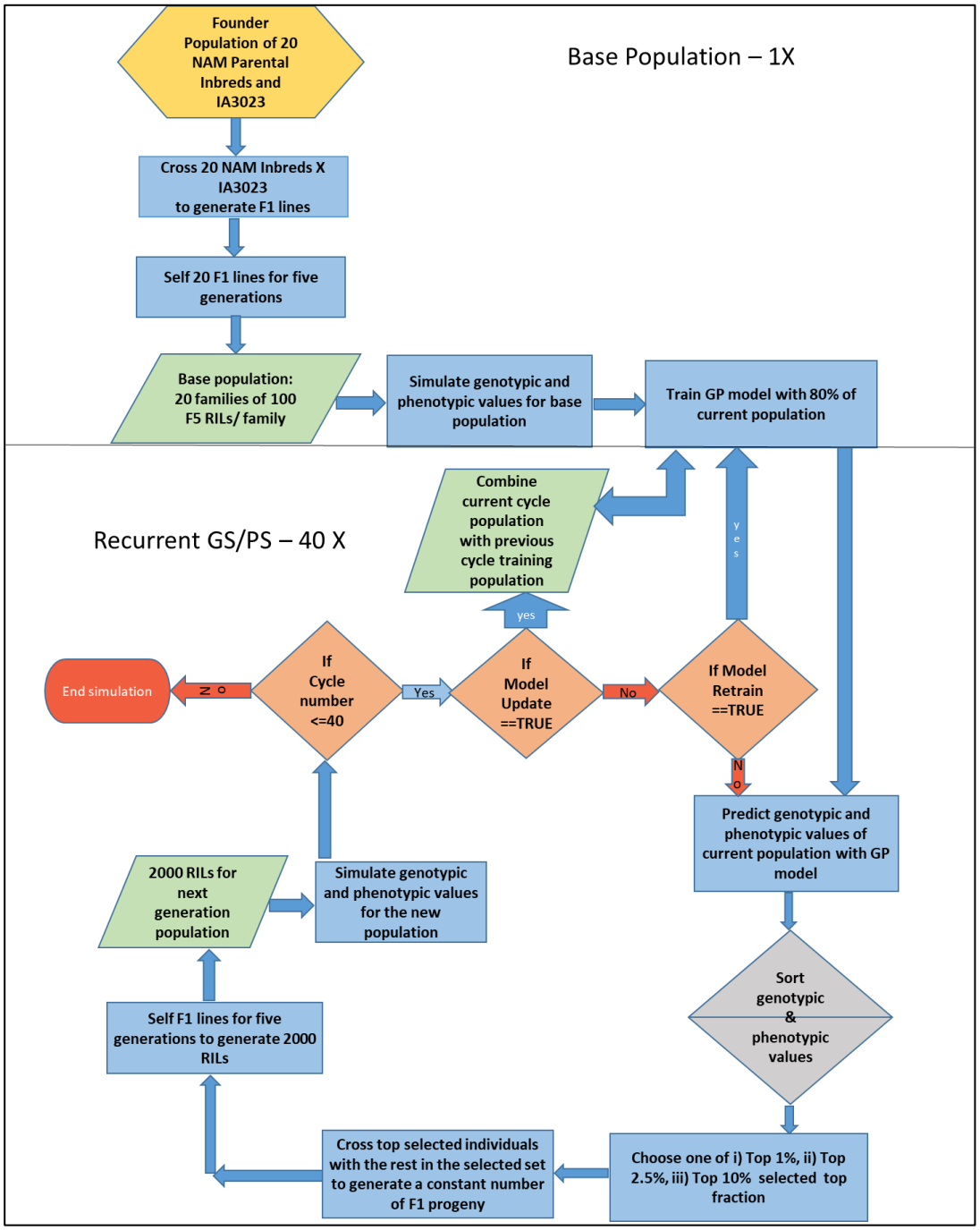

Figure 1 Flow Chart for Simulations of Recurrent Genomic Selection. The upper half panel represents the steps involved in generating the base population of 2000 F_5 RILs derived from 20 NAM founder lines crossed, *in silico*, to IA3023. It includes the model training step for genomic prediction models. The lower half panel represents recurrent steps of prediction, sorting, truncation selection, crossing, and generation of F_5 RILs for each cycle as well as the decision steps to check if the training set should be updated and if the recurrent process is to be continued for another cycle.

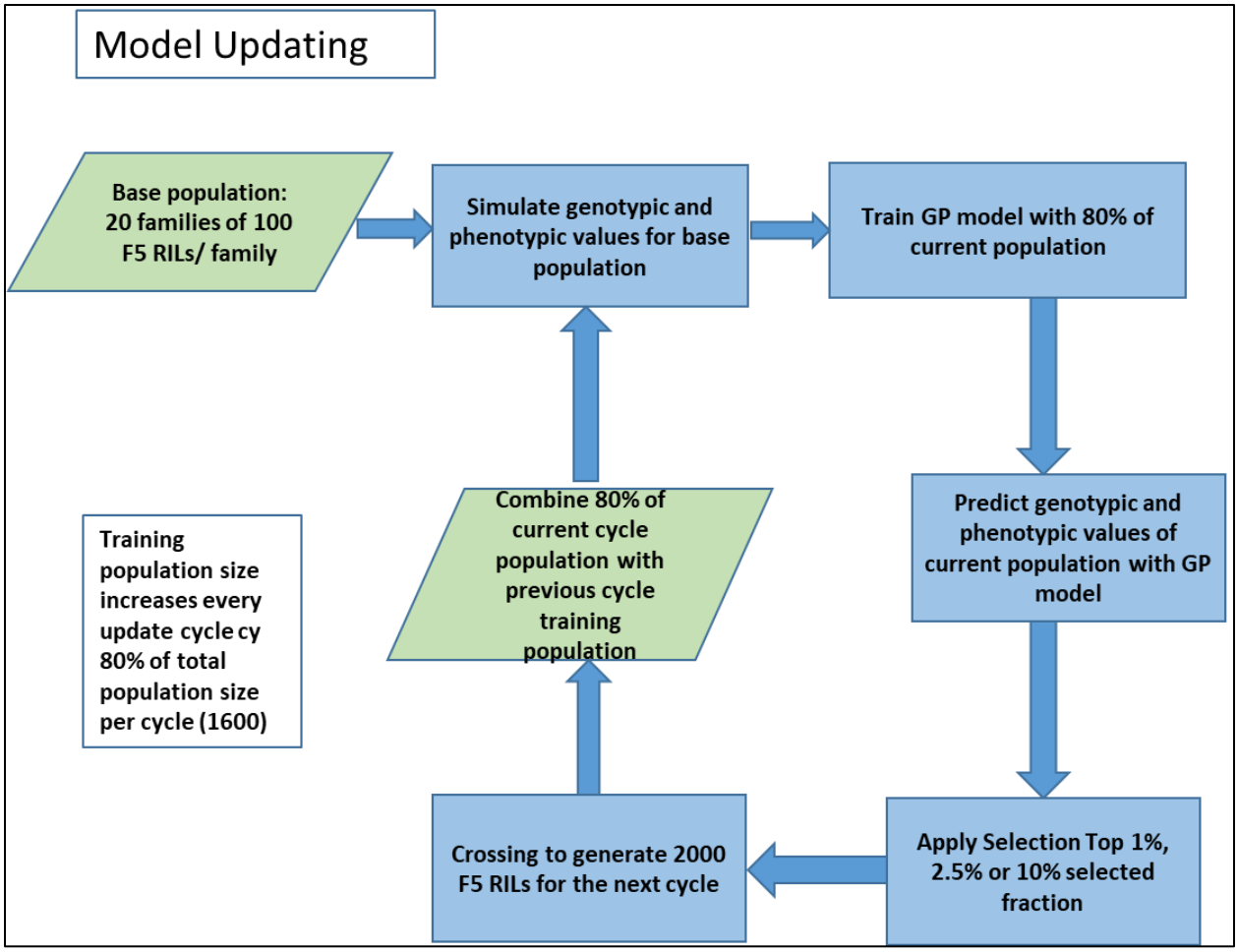

Figure 2 Flow Chart for Model Updating in Simulations of Recurrent Genomic Selection. Model updating involves combining training data from 'n-1' previous cycles (t-n-1...t-1) with training data from the t^{th} cycle to retrain genomic prediction models. 't' refers to the selection cycle and ranges from 1 to 40 and 'n' refers to number of prior cycles that are included in the training set. The treatment design has four levels for 'n' ranging from 0 - 14. 'n=0' refers to no updating, whereas 'n= 10', 'n=12' and 'n=14' refer to inclusion of training data from 9, 11 and 13 prior cycles along with data from t^{th} cycle of selection.

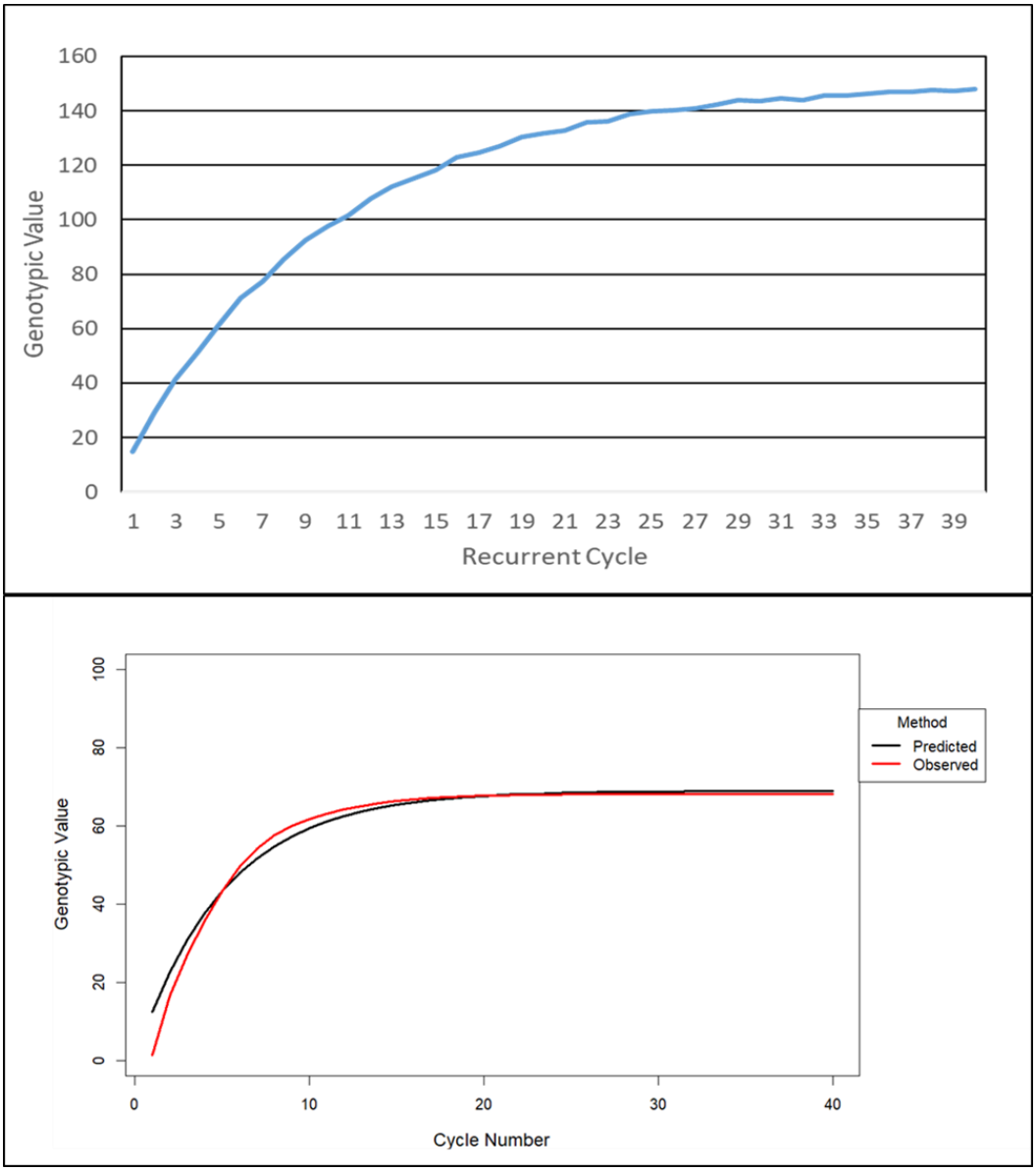

Figure 3 Theoretical and Simulated Genotypic Values: A) (Top Panel) Theoretical Genotypic Values from 40 cycles of recurrent selection modeled with the recurrence equation,

$$y_c = \alpha^c y_0 + \beta\frac{1-\alpha^c}{1-\alpha}$$

where y_c represents the genotypic value in cycle c, with c= 1, 2… 40 and values of α and β are 0.9 and 15 respectively. B) (Bottom Panel) Averaged genotypic values for 40 cycles of simulated recurrent selection. Genotypic values are averaged across selection methods, training sets, selection intensities, number of simulated QTL and simulated heritabilities. Predicted curve is modelled with $\alpha = 0.82$, $\beta = 12.37$, that were obtained from 'nls' fit with completed dataset without any grouping.

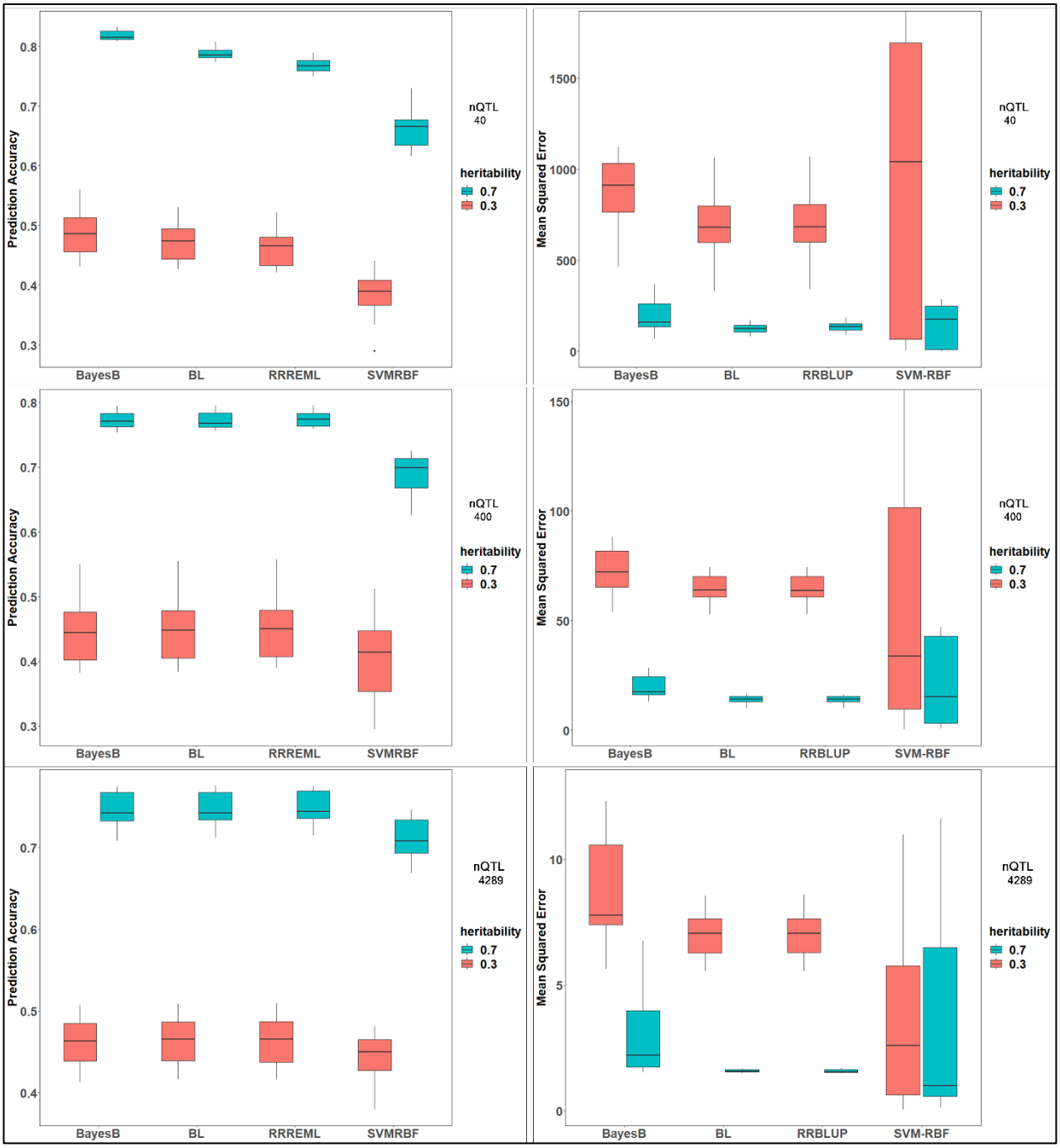

Figure 4 Estimated prediction accuracies and MSE in Founding Set of RILs: Estimated prediction accuracies (left panel) and mean squared errors (right panel) for four genomic prediction (GP) models: BayesB, BL (Bayes LASSO), RRREML (Ridge Regression with REML) and SVMRBF (Support Vector Machines with Radial Basis Function Kernel) trained with F_5 RILs derived from crosses of 20 homozygous founder lines with IA3023. Phenotypes used to train the GP models consisted of genetic architectures comprised of 40, 400 and 4289 simulated QTL (top, middle and bottom) that were responsible for 70% (blue) and 30% (red) of phenotypic variability in the initial populations.

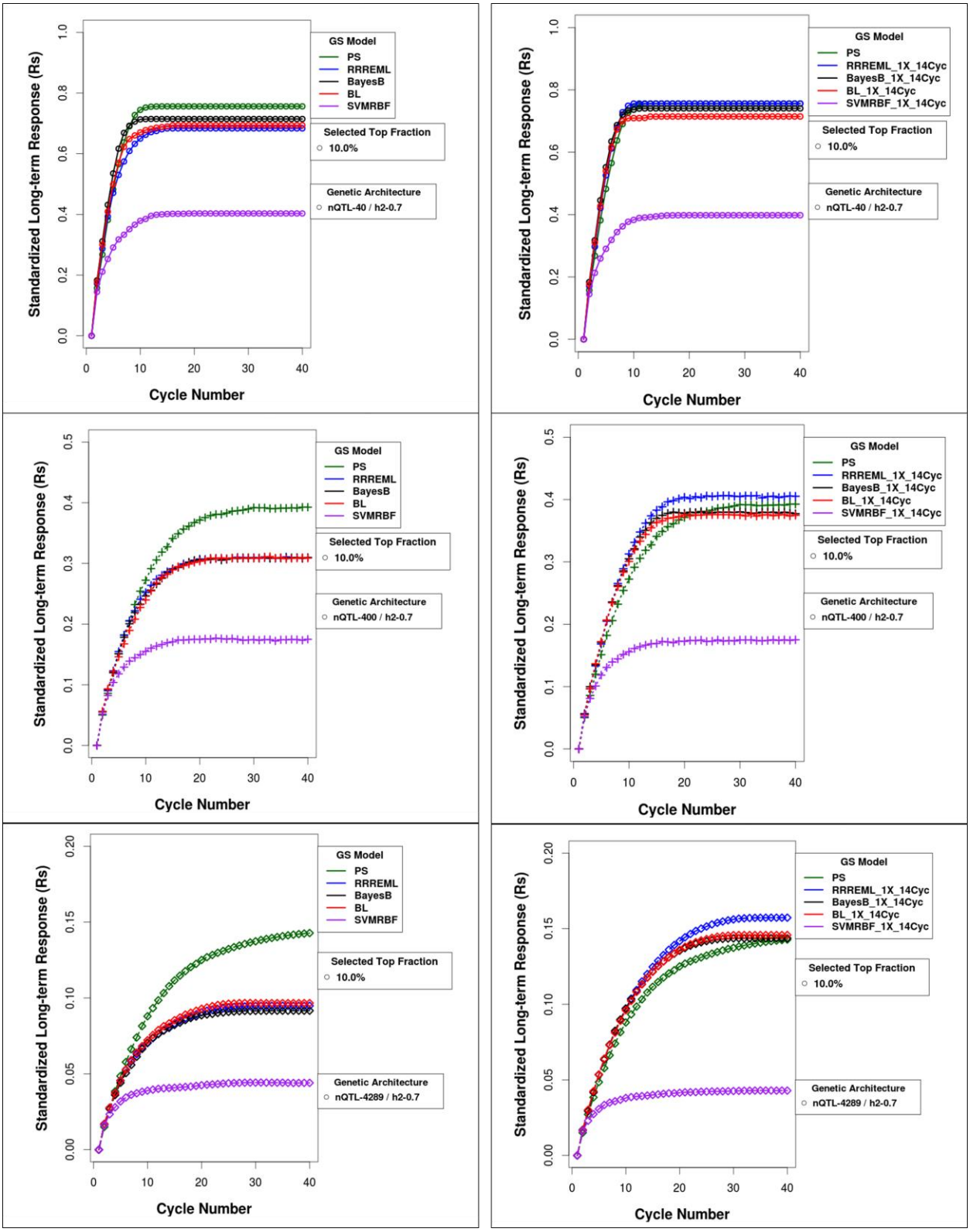

Figure 5 Standardized Responses for Comparison of GS methods with and without Updating for 0.7 H and Top 10% Selected Fraction Forty cycles of standardized responses to selection of 10% of 2000 soybean RILs per cycle. Standardized responses are plotted by selection methods without (left panels) and with (right panels) model updating using prior cycles as training sets for four genotypic prediction models. Phenotypic selection (PS) is not updated and hence is the same in the left and right panels. The top panels consist of responses for genetic architectures consisting of 40 simulated QTL. Middle panels consist of responses for genetic architectures consisting of 400 simulated QTL and the bottom panels consist of responses for genetic architectures consisting of 4289 simulated QTL. All 40, 400, and 4289 simulated QTL are responsible for 70% of phenotypic variability in the initial population. PS – Phenotypic Selection, RR-REML- Ridge Regression with Restricted Maximum Likelihood, BL – Bayes LASSO, and SVMRBF- Support Vector Machine with Radial Basis Kernel.

Method	QTL	heritability	TH5	TH10	TH15	TH20	TH25	TH30	TH35	TH40
RR-REML	40	0.7	-2.36	-12.71	-9.63	-9.61	-9.61	-9.61	-9.61	-9.61
BayesB	40	0.7	10.96	-4.18	-5.49	-5.49	-5.49	-5.49	-5.49	-5.49
BL	40	0.7	3.46	-9.96	-8.77	-8.24	-8.24	-8.24	-8.24	-8.24
SVMRBF	40	0.7	-39.67	-49.09	-46.90	-46.67	-46.67	-46.67	-46.67	-46.67
RR-REML	400	0.7	2.42	-7.37	-14.57	-17.25	-19.61	-20.96	-20.93	-21.11
BayesB	400	0.7	0.10	-9.53	-14.89	-17.58	-19.94	-21.26	-21.05	-21.11
BL	400	0.7	-3.35	-11.95	-15.15	-18.26	-19.62	-21.21	-20.80	-21.37
SVMRBF	400	0.7	-21.91	-43.06	-49.93	-52.95	-54.26	-55.48	-55.46	-55.42
RR-REML	4289	0.7	-6.26	-20.12	-25.33	-27.60	-29.16	-30.98	-32.60	-33.57
BayesB	4289	0.7	-9.07	-19.79	-26.28	-29.14	-31.30	-33.37	-34.90	-35.84
BL	4289	0.7	-5.85	-17.94	-23.50	-25.70	-27.24	-29.66	-31.36	-32.35
SVMRBF	4289	0.7	-34.49	-55.74	-63.38	-66.01	-66.85	-67.79	-68.60	-69.15

Figure 6 Heat Map for Percent Gain in Rs Relative to PS for 0.7 H for GP Models without Updating: Heat map indicating standardized response relative to PS as percentage gain after 40 cycles of recurrent selection using genomic prediction models without updated training sets for 40, 400, 4289 simulated QTL responsible for 70% of phenotypic variability in the initial population. Blue to red shaded cells represent increasing gain in response relative to PS. RR-REML- Ridge Regression with Restricted Maximum Likelihood, BL – Bayes LASSO, and SVMRBF- Support Vector Machine with Radial Basis Kernel.

Method	QTL	heritability	No. Prior Cycles	TH5	TH10	TH15	TH20	TH25	TH30	TH35	TH40
RR-REML	40	0.7	14	9.04	1.55	0.00	0.00	0.00	0.00	0.00	0.00
BayesB	40	0.7	14	14.41	-1.05	-2.06	-2.06	-2.06	-2.06	-2.06	-2.06
BL	40	0.7	14	11.89	-4.70	-5.49	-5.49	-5.49	-5.49	-5.49	-5.49
SVMRBF	40	0.7	14	-39.85	-48.58	-47.63	-47.36	-47.36	-47.36	-47.36	-47.36
RR-REML	400	0.7	14	11.34	14.84	12.30	8.71	5.73	3.29	3.50	3.22
BayesB	400	0.7	14	13.05	12.00	8.38	1.75	-1.11	-2.96	-3.02	-3.91
BL	400	0.7	14	13.85	10.44	6.45	1.02	-1.89	-4.24	-3.92	-4.55
SVMRBF	400	0.7	14	-21.66	-42.81	-50.39	-53.52	-54.86	-55.38	-55.48	-55.40
RR-REML	4289	0.7	14	10.01	10.51	11.79	13.39	14.52	13.67	11.70	10.13
BayesB	4289	0.7	14	10.10	10.41	9.09	8.31	7.20	4.56	2.06	0.60
BL	4289	0.7	14	10.32	9.22	8.87	8.78	8.22	6.11	3.59	2.10
SVMRBF	4289	0.7	14	-36.73	-56.79	-64.13	-66.87	-68.08	-68.97	-69.49	-69.96

Figure 7 Heat Map for Percent Gain in Rs Relative to PS for 0.7 H for GP Models with Updating: Heat map indicating standardized response relative to PS as percentage gain after 40 cycles of recurrent selection using genomic prediction models with updated training sets from up to 14 prior cycles of selection for 40, 400, and 4289 simulated QTL responsible for 70% of phenotypic variability in the initial population. Blue to red shaded cells represent increasing gain in response relative to PS. RR-REML- Ridge Regression with Restricted Maximum Likelihood, BL – Bayes LASSO, and SVMRBF- Support Vector Machine with Radial Basis Kernel.

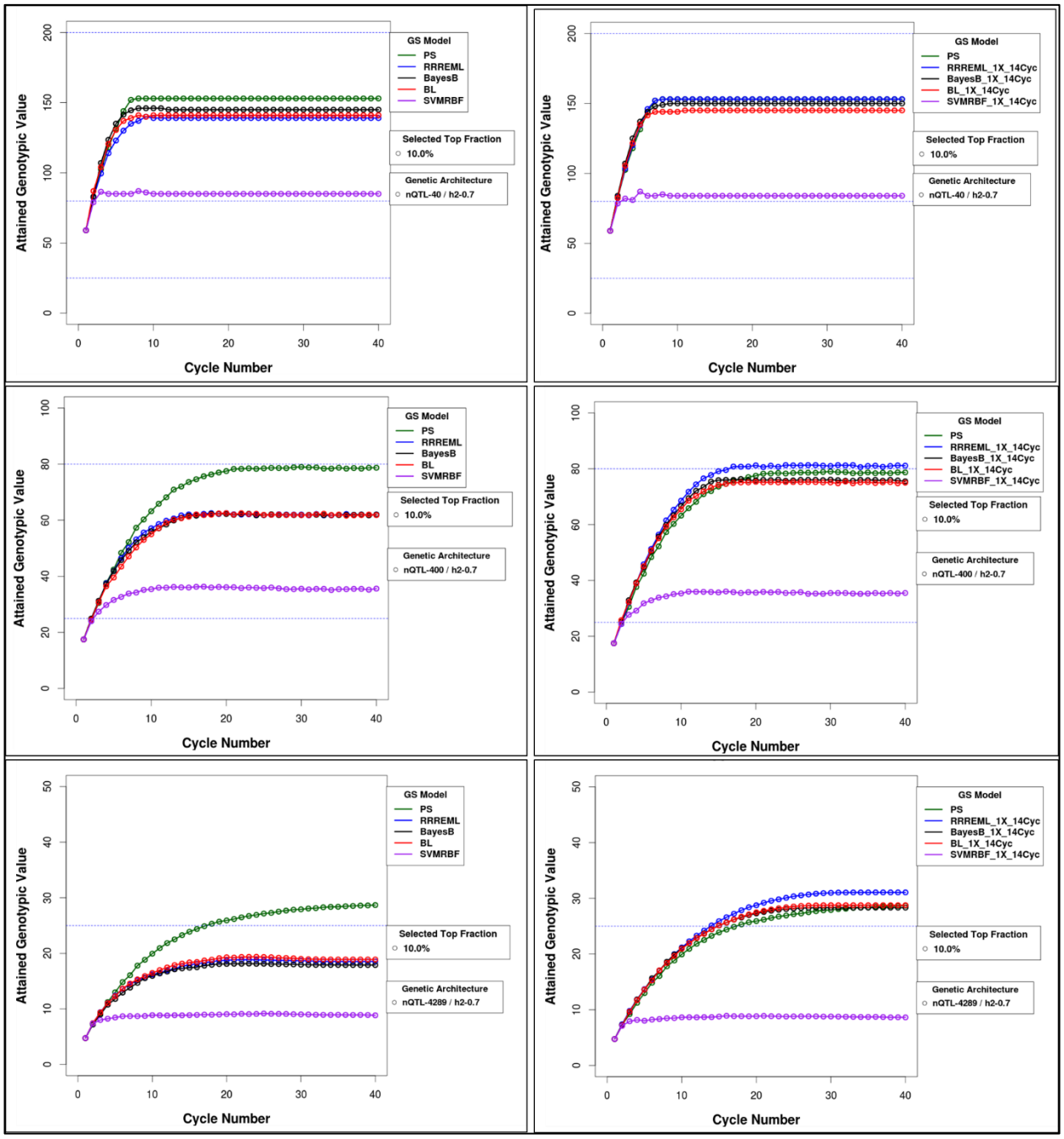

Figure 8 Attained Genotypic Value for Comparison of GS methods with and without Updating for 0.7 H and Top 10% Selected Fraction: Maximum attained genotypic values (Mgvs) in recurrent genomic selection and phenotypic selection (PS) without updating the training sets in the left panels and with training set updates from up to 14 prior cycles in the right panels. PS has no training sets and hence does not change between the left and right panels. a) 40 QTL (top), b) 400 QTL (middle) and c) 4289 QTL (bottom) responsible for 70% of phenotypic variability in the initial population and selection of 10% of the RILs in each cycle. PS – Phenotypic Selection, RR-REML- Ridge Regression with Restricted Maximum Likelihood, BL – Bayes LASSO, and SVMRBF- Support Vector Machine with Radial Basis Kernel.

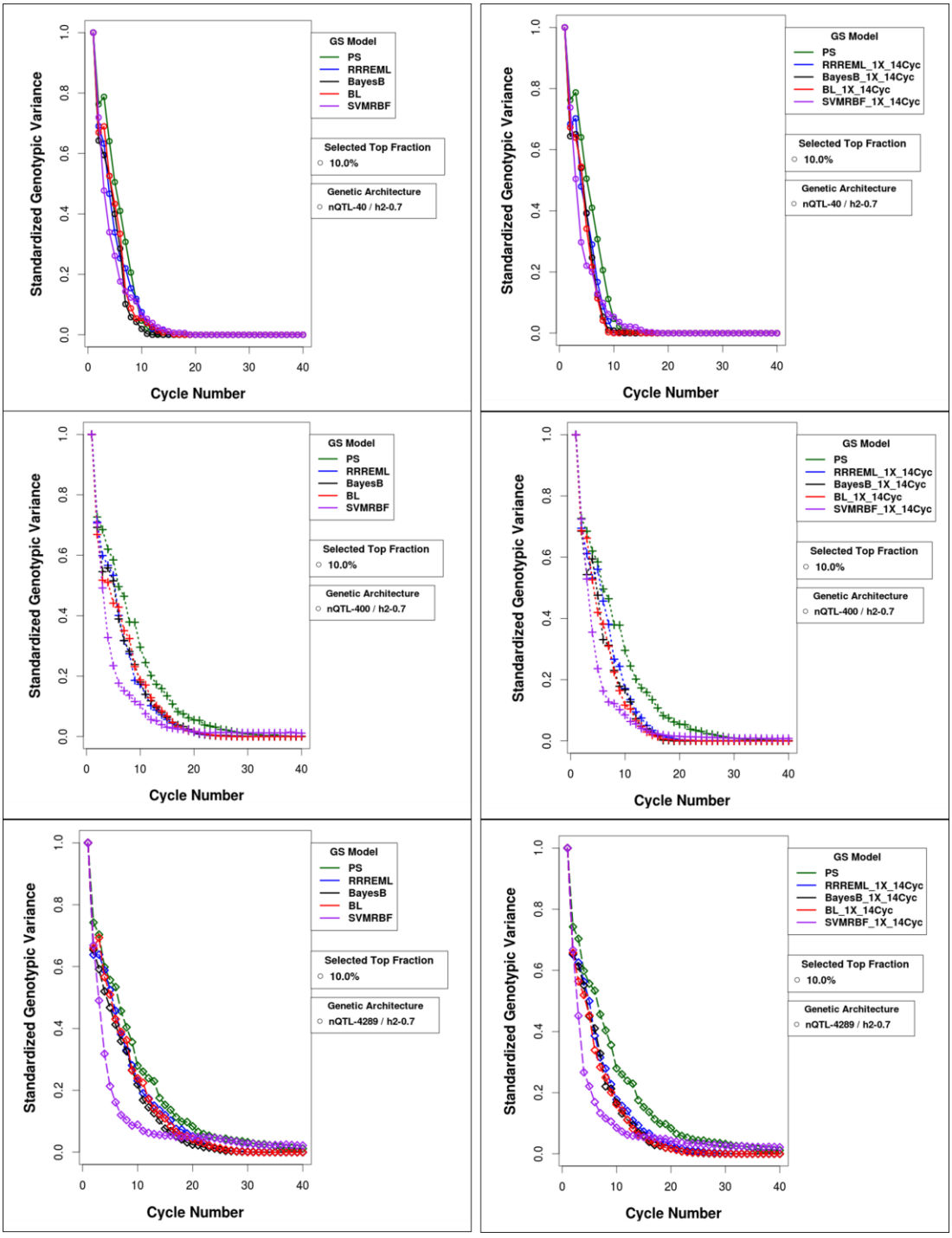

Figure 9 Standardized Genotypic Variance (Sgv) for Comparison of GS methods with and without Updating for 0.7 H and Top 10% Selected Fraction Standardized genotypic variance without training set updating (left panels) and with training set updating with prior cycle training data (right panels) for the four GP models. PS has no updating and hence is the same in both left and right panels. A) Training data from up to 14 prior cycles for 40 simulated QTL (top), 400 simulated QTL (middle) and 4289 simulated QTL (bottom) responsible for 70% of phenotypic variability in the initial population and top 10% of RILs with the greatest predicted values. PS – Phenotypic Selection, RR-REML- Ridge Regression with Restricted Maximum Likelihood, BL – Bayes LASSO, and SVMRBF- Support Vector Machine with Radial Basis Kernel.

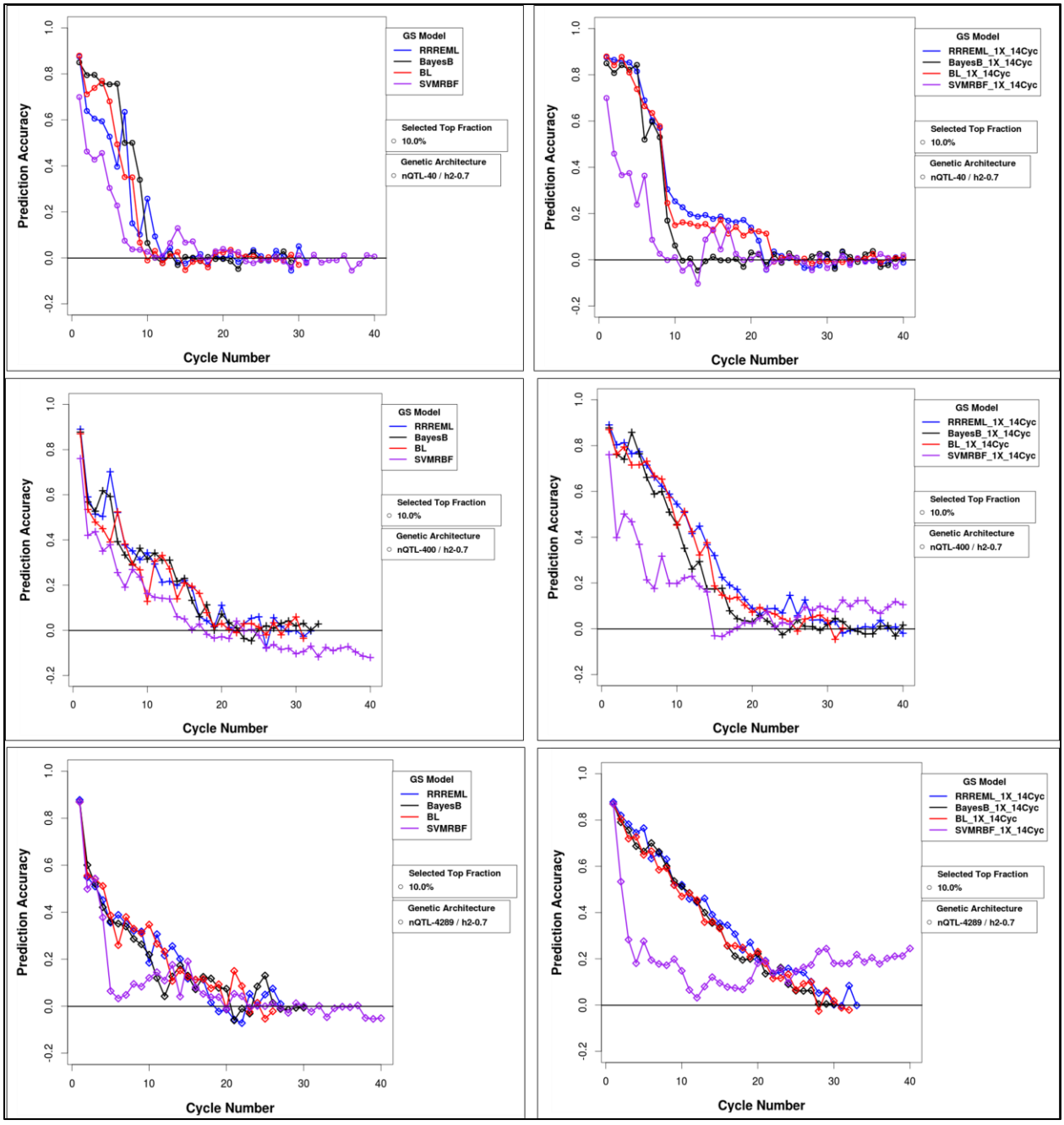

Figure 10 Estimated Prediction Accuracies for Comparison of GS methods with and without Updating for 0.7 H and Top 10% Selected Fraction: Estimated prediction accuracies with updates to the training sets used in genomic prediction (GP) models. Training data from up to 14 prior selection cycles were used to update all four GP models for 40 QTL (top), 400 QTL (middle) and 4289 QTL (bottom) responsible for 70% of phenotypic variability in the initial population and top 10% of RILs with the greatest predicted values. RR-REML- Ridge Regression with Restricted Maximum Likelihood, BL – Bayes LASSO, and SVMRBF- Support Vector Machine with Radial Basis Kernel.

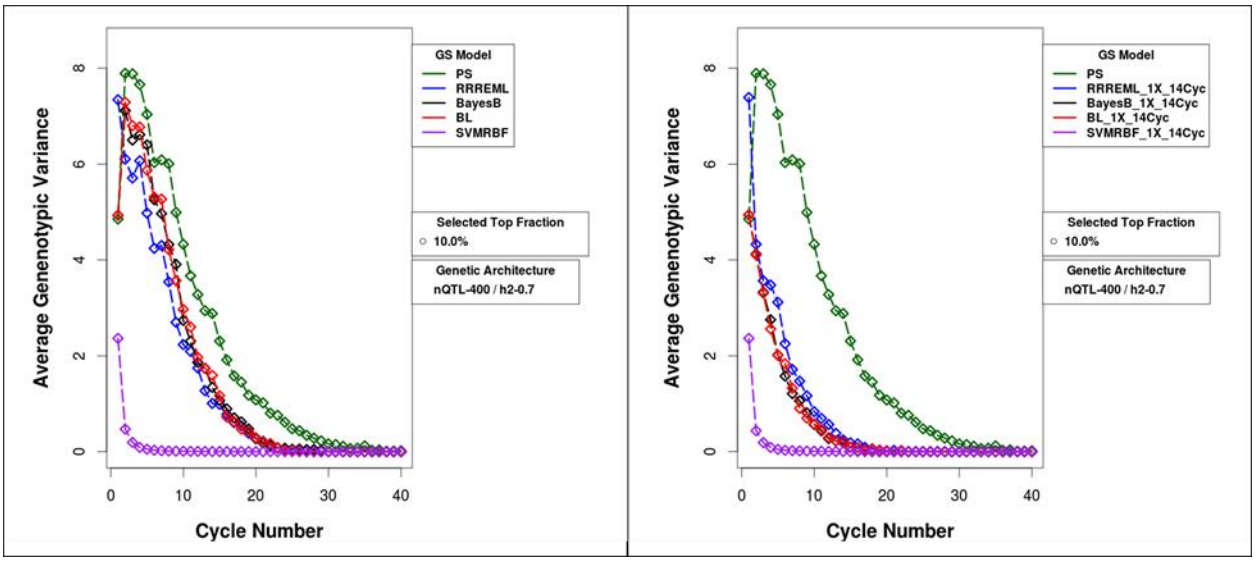

Figure 11 Average Genotypic Variance in Selected Populations for Comparison of GS methods with and without Updating for 400 QTL, 0.7 H and Top 10% Selected Fraction: Average genotypic variance for 40 cycles of recurrent selection of 10% of RILs using five selection methods without updated training sets (left panel) and with updated training sets (right panel). Training sets consisted of genotypic and phenotypic data from up to14 prior cycles of recurrent selection. Simulated phenotypic values of the RILs consisted of 400 simulated QTL responsible for 70% of phenotypic variability in the initial population. PS – Phenotypic Selection, RR-REML- Ridge Regression with Restricted Maximum Likelihood, BL – Bayes LASSO, and SVMRBF- Support Vector Machine with Radial Basis Kernel.

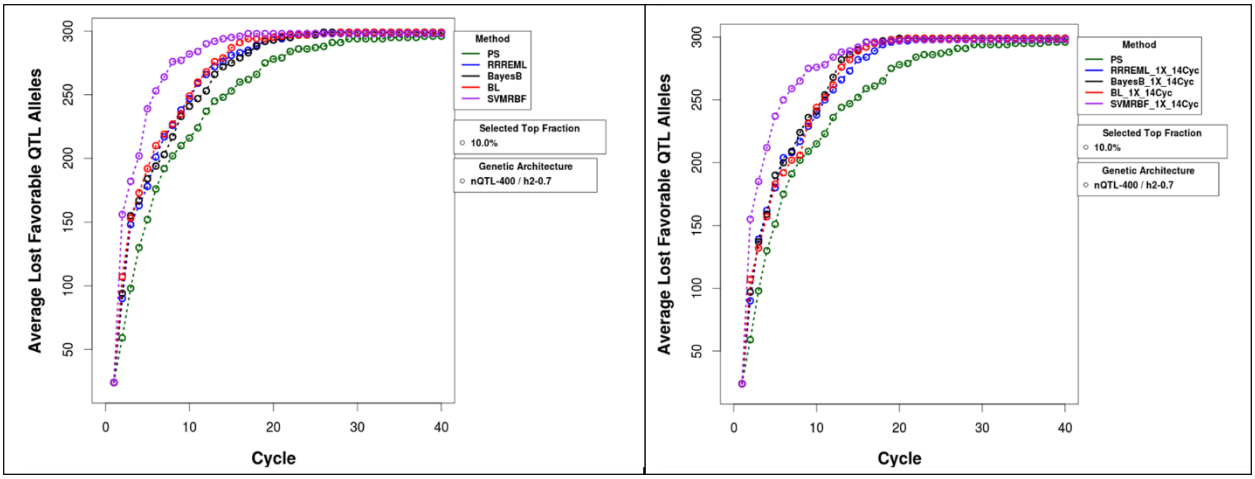

Figure 12: Number of Lost Favorable Alleles for Comparison of GS methods with and without Updating for 400 QTL, 0.7 H and Top 10% Selected Fraction: Number of favorable alleles that are cumulatively discarded across 40 cycles of recurrent selection of 10% of the RILs created every generation using five selection methods without updated training sets (left panel) and with updated training sets (right panel). Training sets consisted of genotypic and phenotypic data from up to14 prior cycles of recurrent selection. Simulated phenotypic values of the RILs consisted of 400 simulated QTL responsible for 70% of phenotypic variability in the initial population. PS – Phenotypic Selection, RR-REML- Ridge Regression with Restricted Maximum Likelihood, BL – Bayes LASSO, and SVMRBF- Support Vector Machine with Radial Basis Kernel.

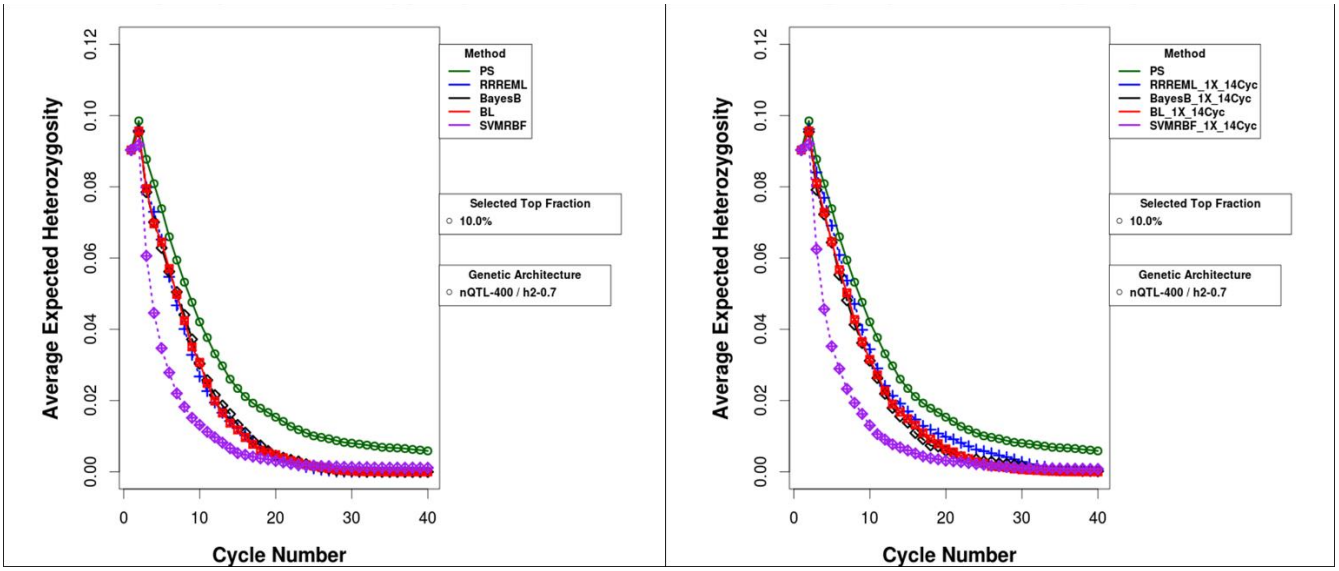

Figure 13 Average Expected Heterozygosity for Comparison of GS methods with and without Updating for 400 QTL, 0.7 H and Top 10% Selected Fraction: Average Expected Heterozygosity in recurrent GS without GP model updating (left panel) and GP models updated every cycle with training data from upto14 prior cycles (right panel). All treatment combinations have 400 simulated QTL responsible for 70% of phenotypic variability in the initial population and 10% top selected fraction. PS – Phenotypic Selection, RR-REML- Ridge Regression with Restricted Maximum Likelihood, BL – Bayes LASSO, and SVMRBF- Support Vector Machine with Radial Basis Kernel.

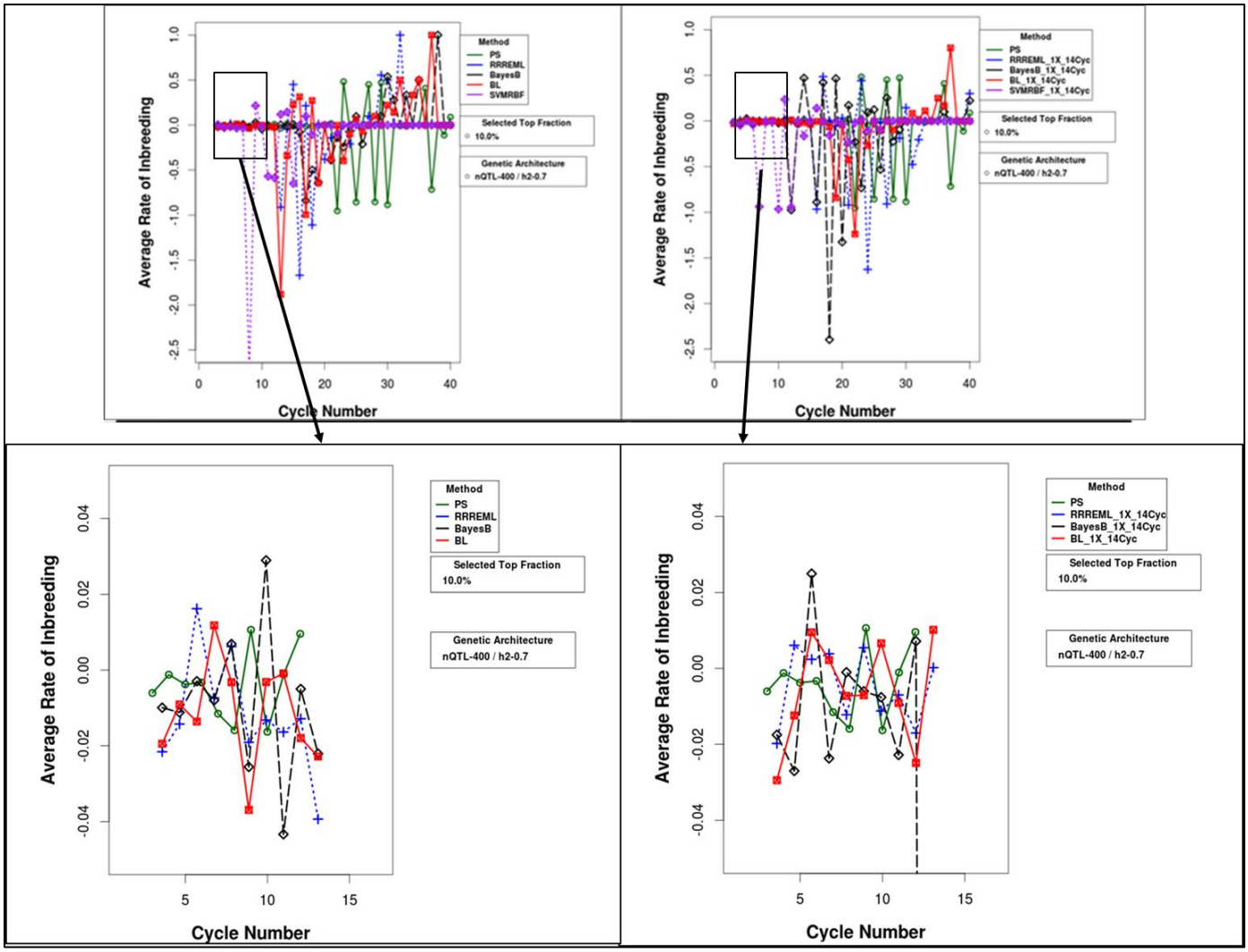

Figure 14 Average Rate of Inbreeding for Comparison of GS methods with and without Updating for 400 QTL, 0.7 H and Top 10% Selected Fraction: Average rate of inbreeding in recurrent GS without GP model updating (left panel) and GP models updated every cycle with training data from upto14 prior cycles. Inbreeding co-efficient is estimated as the harmonic mean of all individuals in the population for 40 cycles (top panel). Inset plot (bottom panel) shows magnified region from 3-12 cycles in average rate of inbreeding for PS, RR, BayesB, and BL GS methods. All treatment combinations have 400 simulated QTL responsible for 70% of phenotypic variability in the initial population and 10% top selected fraction. PS – Phenotypic Selection, RR-REML- Ridge Regression with Restricted Maximum Likelihood, BL – Bayes LASSO, and SVMRBF- Support Vector Machine with Radial Basis Kernel.

Tables

TABLE 1: FACTORIAL DESIGN

Factors	Number of Levels	Values for Levels
Number of QTL	3	40, 400, 4289
Heritability	2	0.7, 0.3
Selection Intensity	3	2.67, 2.34, 1.75
Selection Model	5	i) PS- Phenotypic value ii) GS - GP(RR) iii) GS - GP(Bayes B) iv) GS - GP(Bayes LASSO) v) GS - GP(SVM, Radial basis Function Kernel)
Model Update: number of prior cycles used in training sets	4	i) 0 previous cycles ii) 10 previous cycles iii) 12 previous cycles iv) 14 previous cycles
Total Number of unique combinations	360	
Total Number of Simulations	3600 with 10 reps /condition	

TABLE 2 PACKAGES IN R FOR PARAMETRIC AND NON-PARAMETRIC MODELS WITH TUNING PARAMETERS

GS Model	Package (R)	Model Tuning Parameters (BGLR package)
Ridge Regression	REML-EM (custom R script Xavier A, 2018)	EM algorithm for estimation of parameters with REML method without using matrix inversion
Bayesian LASSO	BGLR (Perez P et al, 2014)	Priors for varE (df=3,S=0.25); varU (df=3,S=0.63); lambda(shape=0.53,rate=5e-5) type='random',value=30), nIter=20000, burnIn=2000, thin=1
Bayes B	BGLR (Perez P et al, 2014)	nIter=41000, burnIn =1000, df0=4, R2=0.7
SVM	Rgtsvm (Wang Zhong et al, 2017)	SVM with Radial basis function kernel on GPU

TABLE 3 ANOVA FOR NON-LINEAR MIXED MODELS

Model Description	df	AIC	BIC	logLik
M1: 2 curves, one for each simulated h2	6	1490883.86	1490943.12	-745435.93
M2: 3 curves, one for each simulated SI	6	1487947.33	1488006.60	-743967.67
M3: 3 curves, one for each nQTL level	6	1271056.31	1271115.57	-635522.15
M4: 4 curves, one for each number of prior cycle TS levels	6	1491707.23	1491766.50	-745847.61
M5: 5 curves, one for each of SM levels	6	1482602.97	1482662.24	-741295.49
M6: 6 curves, one for each of h2 and nQTL levels	9	1262672.35	1262761.25	-631327.17
M7: 6 curves, for each of h2 and SI levels	9	1486685.15	1486774.05	-743333.58
M8: 8 curves, one for each combination of h2 and TS levels	9	1490439.19	1490528.09	-745210.59
M9: 9 curves, one for each level of SI and nQTL	9	1243669.76	1243758.65	-621825.88
M10: 10 curves, one for each level of SM and h2	9	1481466.69	1481555.59	-740724.34
M11: 12 curves, one for each level of TS and nQTL	9	1268945.50	1269034.40	-634463.75
M12: 12 curves, one for each level of TS and SI	9	1487510.42	1487599.31	-743746.21
M13: 15 curves, one for each combination of SM and nQTL	9	1194913.36	1195002.26	-597447.68
M14: 15 curves, one for each combination of SM and SI	9	1477442.03	1477530.93	-738712.01
M15: 20 curves , one for each combination of SM and TS	9	1481835.46	1481924.36	-740908.73
M16: 18 curves, one for each combination of SI, nQTL and h2	12	1233180.40	1233298.93	-616578.20
M17: 24 curves, one for each combination of SI, TS and h2	12	1486217.33	1486335.86	-743096.66
M18: 24 curves, one for each combination of TS, nQTL and h2	12	1260244.61	1260363.14	-630110.31
M19: 30 curves, one for each combination of SM, h2 and SI	12	1179040.14	1179158.67	-589508.07
M20: 30 curves, one for each combination of SM, nQTL and h2	12	1476040.79	1476159.32	-738008.40
M21: 36 curves, one for each combination of SI, nQTL, and TS	15	1229771.09	1229919.25	-614870.54
M22: 40 curves, one for each combination of SM, h2 and TS	12	1480446.72	1480565.25	-740211.36
M23: 45 curves, one for each combination of SI, nQTL and SC	15	1107823.41	1107971.57	-553896.70
M24: 60 curves, one for each combination of SM, nQTL and TS	12	1188346.20	1188464.74	-594161.10
M25: 60 curves, one for each combination of SI, TS and SM	12	1476793.27	1476911.80	-738384.64
M26: 72 curves, one for each combination of TS, h2, nQTL, SI	15	1230064.21	1230212.38	-615017.11
M27: 90 curves, one for each combination of SM, h2, nQTL, SI	15	1107717.81	1107865.98	-553843.91

M28: 120 curves, one for each combination of SM, h2, TS, SI	15	1475380.15	1475528.31	-737675.07
M29: 120 curves, one for each combination of SM, nQTL, TS, h2	15	1171315.89	1171464.05	-585642.94
M30: 180 curves, one for each combination of SM, nQTL, TS, SI	15	1123398.62	1123546.78	-561684.31
M31: 360 curves, one for each combination of SM, nQTL, TS, SI, and h2	6	1094075.09	1094134.35	-547031.54

Table 3: Non-linear Mixed Models fit with increasing number of factors from M1 to M31. Factors include selection method with five levels comprising of PS and four GS models ('SM'), two levels for heritability ('h2'), three levels for number of QTL ('nQTL'), three levels for selection intensity ('SI'), and four levels for prior cycles in training set ('TS')

End Of File