2018

Three new approaches to genomic selection

Lizhi Wang
Iowa State University, lzwang@iastate.edu

Guodong Zhu
Iowa State University

Will Johnson
Iowa State University

Mriga Kher
Iowa State University, makher@iastate.edu

Three new approaches to genomic selection

Abstract
Conventional genomic selection approaches use breeding values to evaluate individual plants or animals and to make selection decisions. Multiple variants of breeding values and selection approaches have been proposed, but they suffer two major limitations. First, selection decisions are not responsive to changes in time and resource availability. Second, selection decisions are not coordinated with related decisions such as mating and resource allocation. We present three new genomic selection approaches that attempt to address these two limitations, which were designed by engineering students in a class project at Iowa State University. Compared with previous approaches using the same data set from the literature, two of these engineering approaches were found to be comparable to the state of the art, and the third one significantly dominated all the previous approaches.

Keywords
engineering, genomic selection, optimization

Disciplines
Genomics | Operational Research | Systems Engineering

Three New Approaches to Genomic Selection

Lizhi Wang*, Guodong Zhu, Will Johnson, Mriga Kher
Iowa State University

August 10, 2018

Abstract

Conventional genomic selection approaches use breeding values to evaluate individual plants or animals and to make selection decisions. Multiple variants of breeding values and selection approaches have been proposed, but they suffer two major limitations. First, selection decisions are not responsive to changes in time and resource availability. Second, selection decisions are not coordinated with related decisions such as mating and resource allocation. We present three new genomic selection approaches that attempt to address these two limitations, which were designed by engineering students in a class project at Iowa State University. Compared with previous approaches using the same data set from the literature, two of these engineering approaches were found to be comparable to the state-of-the-art, and the third one significantly dominated all the previous approaches.

1 Introduction

Genomic selection takes advantage of genetic data from whole-genome single-nucleotide polymorphism (SNP) markers to accelerate genetic gains in plant or animal breeding programs. Genetic prediction techniques [Goddard, 2009, Meuwissen et al., 2001] can be used to accurately estimate the additive effects of all quantitative trait loci (QTL), which will be subsequently used to make selection decisions. The conventional genomic selection (GS) approach [Meuwissen et al., 2001] makes selection decisions based on the genetic estimated breeding value (GEBV), which is the cumulative effect of all marker loci. Weighted genetic estimated breeding value (WGEBV) is a variation of GEBV proposed in [Goddard, 2009, Jannink, 2010], which uses marker frequency to amplify the marker effect of rare and desirable alleles. These approaches have been shown to be effective in achieving genetic improvements in the short term, but they suffer the limitation of losing genetic diversity and growth potential in the longer term. The optimal haploid value (OHV) was another variation of GEBV that addressed this limitation [Daetwyler et al., 2015]. It evaluates the breeding value of an individual by the sum of effects of the better haplotype block from the two chromosomes (assuming diploid species), rather than the cumulative effects of all alleles on both chromosomes.

*Corresponding author: Lizhi Wang, Department of Industrial and Manufacturing Systems Engineering, Iowa State University. Email: lzwang@iastate.edu

Recently, the genome building (GB) [Kemper et al., 2012] and optimal population value (OPV) [Goiffon et al., 2017] approaches suggested a different selection strategy from truncation selection. Rather than selecting individuals that have high breeding values separately, these two new approaches define new metrics to evaluate how complementary the selected individuals are as a group, and use optimization techniques to select a sub-population of individuals to maximize the metrics. Simulation results from [Goiffon et al., 2017] suggest that both GB and OPV outperformed GS, WGS, and OHV approaches.

2 Materials

In this section, we describe a genomic selection project in a diagram in Figure 1, define the mathematical notations that will be used in our analysis, and briefly review the five previous approaches for genomic selection.

- **The** (Start) **point**

  A genomic selection project typically starts with an initial population of plant or animal individuals. The genotype of the initial population can be represented by a three-dimensional binary matrix $G \in \mathbb{B}^{L\times M\times N_0}$, where L is the number of SNP markers, M is the ploidy of the species, and N_0 is the number of individuals in the initial population. For convenience of presentation, we consider diploid species ($M = 2$) in this paper, but the analysis can be extended to more general polyploid species. We use N_t to denote the number of individuals in the population of generation t. The value $G_{i,j,n}$ indicates whether the allele at locus i from chromosome j of individual n is the desirable ($G_{i,j,n} = 1$) or undesirable ($G_{i,j,n} = 0$) variation. The effects of undesirable alleles are normalized to be zero, whereas the effects of desirable alleles are assumed to be known and denoted as $\beta \in \mathbb{R}_+^{L\times 1}$. The deadline T and total budget B for the project should also be determined at this point.

- **The** [Selection] **step**

  A number of individuals are selected from the current population, which will be crossed to produce the next generation. Four types of selection decisions need to be made: how many crosses should be made, which individuals should be selected to make the crosses, how should the selected individuals be paired up, and how many progeny should each cross produce. Genomic selection approaches mainly influence the decisions made in this step of the process.

- **The** [Reproduction] **step**

  The selected individuals are mated to produce the next generation according to the decisions from the Selection step. The genotype of a random progeny from crossing two individuals (or selfing one) is described as follows. Let $P \in \mathbb{B}^{L\times 2}$ denote the genotype matrix of a random progeny from crossing individuals n_1 and n_2. Then P is determined as

$$P_{i,j} = G_{i,J_i^j+1,n_j}, \forall i \in \{1, ..., L\}, j \in \{1, 2\}.$$

Here, $J^1, J^2 \in \mathbb{B}^{L\times 2}$ are two identical and independent random vectors following the inheritance distribution with recombination frequency vector r. A random binary vector $J \in \mathbb{B}^L$ follows an *inheritance distribution* [Han et al., 2017] with parameter vector $r \in [0, 0.5]^{L-1}$ if

$$J_1 = \begin{cases} 0, & \text{w.p. } 0.5 \\ 1, & \text{w.p. } 0.5 \end{cases},$$

$$J_i = \begin{cases} J_{i-1} & \text{w.p. } 1 - r_{i-1} \\ 1 - J_{i-1} & \text{w.p. } r_{i-1} \end{cases}, \forall i \in \{2, ..., L\}.$$

Here, "w.p." stands for "with probability".

- **The** $t \geq T?$ **condition**

The genomic selection project will finish after T generations of breeding, a pre-determined project deadline, and then the effectiveness of the project will be evaluated.

- **The** Finish **point**

The final population is compared with the initial population to assess its genetic gains. Outstanding individuals from the final population will proceed to further stages of new cultivar development.

Genomic selection approaches focus on the Selection component in the diagram. A salient feature about genomic selection is that it is easy to come up with approaches that achieve genetic gains, but it is hard to improve these approaches by overcoming their limitations without introducing new ones. In the following, we interpret the five previous approaches using the aforementioned definitions, and discuss how they have evolved to address the limitations of their predecessors and what limitations still remain to be overcome.

- The GEBV of individual n is defined as

$$\texttt{GEBV}(n) = \sum_{i=1}^{L}\sum_{j=1}^{2} \beta_i G_{i,j,n}.$$

The GS approach calculates the GEBVs for all individuals in the current population and selects those with the highest GEBVs as breeding parents [Meuwissen et al., 2001]. If GEBV is considered as the predicted phenotype of an individual when genetic effects are additive with no interactions and non-genetic effects are negligible, then the GS approach is similar with phenotypic selection, which selects individuals with the best phenotypes as breeding parents. The underlying assumption of these two approaches is that taller parents produce taller children.

A limitation of this approach is its focus on short-term genetic gains at the cost of losing genetic diversity and longer-term growth potential. This is because long-term growth potential

relies on the combination of desirable alleles scattered within different individuals, but GS eliminates all individuals with lower GEBVs even if they contain rare desirable alleles that do not exist elsewhere.

- The WGEBV of an individual $n \in \{1, ..., N\}$ is defined as

$$\texttt{WGEBV}(n) = \sum_{i=1}^{L} \sum_{j=1}^{2} \frac{\beta_i}{\sqrt{\max\{w_i, \frac{1}{N}\}}} G_{i,j,n},$$

where w_i is the percentage of desirable alleles at locus i. The WGS approach selects individuals with the highest WGEBVs [Goddard, 2009, Jannink, 2010]. WGEBV differs from GEBV by amplifying β_i of rare desirable alleles, which is intended to increase the chances for individuals containing rare desirable alleles to be selected. The $\max\{\}$ function prevents the denominator from being zero when $w_i = 0$. This is equivalent to dropping the term for allele i when it is monomorphic, because it would have the same effect on all individuals.

Although WGS and its similar variations address a limitation of GS to certain extent, its effectiveness in maintaining the long-term growth potential does not have a theoretical guarantee and may be practically very limited. This is because allele frequencies may take a few generations to become noticeably bimodal towards either 0 or 1, but when the amplifying allele frequencies start to kick in hard enough to preserve rare alleles after several generations of selection, it might already be too late.

- The OHV of an individual $n \in \{1, ..., N\}$ is defined as

$$\texttt{OHV}(n) = \sum_{b \in \mathcal{B}} 2 \max_{j \in \{1,2\}} \left\{ \sum_{i \in H(b)} \beta_i G_{i,j,n} \right\},$$

where $\mathcal{B}$ is the set of haplotype blocks and $H(b)$ is the set of SNPs that belong to block b. The OHV approach selects individuals with the highest OHVs [Daetwyler et al., 2015]. The OHV of an individual measures the GEBV of its best possible progeny from self-pollination, assuming that recombination events may occur between haplotype blocks but not within them. As such, OHV represents a shift of selection criterion from the individuals with the best genetic achievement themselves to the individuals with the best possible gametes to pass on to the next generation. The effectiveness of the OHV approach is sensitive to the lengths of the haplotype blocks, as suggested in [Goiffon et al., 2017], thus parameter tuning may be a necessary step to achieve optimal performance.

The improvement from GS to WGS and OHV leaves another limitation unaddressed, which is treating the contributions of selected individuals as additive and separable. However, an individual with a lower breeding value (be it GEBV, WGEBV, or OHV) may be a better choice than another one with a higher breeding value, if, for example, the latter has a same (or similar) genotype with one of the already selected individuals whereas the former is genetically unique.

- The GB value of a sub-population of individuals, $S \subseteq \{1, ..., N\}$, is defined as

$$\texttt{GB}(S) = \sum_{b \in \mathcal{B}} \max_{(n_1 \neq n_2) \in S} \max_{j \in \{1,2\}} \left\{ \sum_{i \in H(b)} \beta_i (G_{i,j,n_1} + G_{i,j,n_2}) \right\}.$$

The GB value measures the GEBV of an ideal progeny that takes two best haplotype segments from two individuals in the sub-population for each block. The GB approach selects a sub-population of individuals with the highest GB value [Kemper et al., 2012].

Using similarly defined haplotype blocks as in OHV to separate likely and unlikely recombination events, the GB approach examines the complementarity of the selected sub-population by allowing the two haplotype blocks of the ideal progeny to come from any individual within the sub-population. As such, the approach tends to select individuals that contain rare desirable alleles in different haplotype blocks, effectively preserving genetic diversity for the long-term growth potential.

A challenge of the GB approach is its computational requirement. For a given sub-population, the definition of its GB value already contains two layers of maximization, and selecting the optimal sub-population requires a third layer. In fact, Xu et al. (2011) [Xu et al., 2011] proved that the gene stacking problem, a similar and related optimal selection problem, is NP-hard, meaning that an efficient algorithm that solves the problem in polynomial time may not exist at all.

- The OPV of a sub-population of individuals, $S \subseteq \{1, ..., N\}$, is defined as

$$\texttt{OPV}(S) = 2 \sum_{b \in \mathcal{B}} \max_{n \in S} \max_{j \in \{1,2\}} \left\{ \sum_{i \in H(b)} \beta_i G_{i,j,n} \right\}.$$

The OPV measures the GEBV of the best possible progeny that can be produced from crossing individuals in the sub-population over an unlimited number of generations, assuming recombination events are possible between but not within haplotype blocks. The OPV approach selects a sub-population of individuals with the highest OPV [Goiffon et al., 2017]. By definition, this approach has a built-in focus on the long-term potential, and the tradeoff between short-term and long-term gains is adjustable by the lengths of the haplotype blocks.

The OPV approach suffers a similar computational challenge as GB. This challenge was partially ameliorated in [Goiffon et al., 2017], which presented a more efficient algorithm than the one in [Kemper et al., 2012] for selecting the optimal sub-population with respect to OPV or GB value.

A comprehensive computational experiment was conducted in Goiffon et al. (2017) [Goiffon et al., 2017] to compare the performances of these five approaches; multiple sets of parameters were used for OHV, GB, and OPV approaches and the best parameters were used in the comparison. The results are given in Figure 2, which is Figure 2 in [Goiffon et al., 2017].

We point out two limitations of the five previous genomic selection approaches. The first limitation is the lack of responsiveness to time and resource constraints. The five previous approaches were designed to make the same selection decisions regardless if the project deadline is only one generation away or ten generations away, and regardless if the project can only afford to maintain a small population size or has a large budget to make many crosses and produce a large number of progeny in each generation. However, as demonstrated in Figure 3, the performance of a genomic selection approach is very sensitive to the time and resource constraints. In the left subfigure, whether we should use Approach 1 or Approach 2 depends on if the project deadline is t_1 or t_2; in

the right subfigure, the best progeny from which distribution is more likely to have a higher genetic value depends on the number of progeny we can afford to produce.

Although saving time and resources was identified as one of the fundamental advantages of marker assisted selection over conventional phenotypic selection [Collard and Mackill, 2008], the emphasis on time and resource constraints in selection was quite limited. Dekkers and van Arendonk [Dekkers and Van Arendonk, 1998] used optimal control theory to "optimize response to selection over multiple generations" with the assumption of one major gene and a population of infinite size. Their results were later extended and generalized in [Dekkers and Chakraborty, 2001]. Frisch et al. [Frisch et al., 1999] addressed the question of "How many individuals must be generated and genotyped with molecular markers to reduce the undesirable donor genome below a certain threshold?" Riedelsheimer and Melchinger [Riedelsheimer and Melchinger, 2013] proposed an approach for "the allocation of resources in genomic selection (GS) for one breeding cycle" by "optimally split[ting] the total budget between expenditure for the training set on the one hand and the prediction set on the other hand."

The second limitation is the lack of coordination between selection and other related decisions, such as mating and resource allocation. The previous approaches all focused on which individuals to select, but did not inform breeders how many crosses to make, how many progeny to produce from each cross, or how to pair up the selected individuals. However, these mating and resource allocation decisions should be made by the genomic selection approach in coordination with the selection decisions rather than left for the breeder to figure out.

3 Methods

As a first attempt to address the two limitations of previous approaches, we present three new genomic selection approaches that were created by engineering students at Iowa State University. In the fall of 2016 and spring of 2017, Wang assigned genomic selection as a competition in three courses that he taught in the Department of Industrial and Manufacturing Systems Engineering at Iowa State University: IE 312, IE 534, and IE 634. IE 312 is *Optimization* for undergraduate students, IE 534 is *Linear Programming* for junior graduate or senior undergraduate students, and IE 634 is *Computational Optimization* for senior graduate students. The enrollments for IE 312 and IE 534 in fall 2016 and IE 634 in spring 2017 when the competition took place were 116, 46, and 12, respectively. The purpose of the competition was two-fold. First, it was an experiment to reveal how engineers could overcome disciplinary boundaries and help advance research frontiers in plant breeding. Second, it was an opportunity for engineering students to develop skills for solving complex and non-conventional problems that require knowledge beyond their educational background.

The problem definition for the competition was given as follows. A breeding project has a total budget of $B = \$30,750$ and a deadline of $T = 10$ generations. The costs of making each cross and producing each progeny are \$5 and \$10, respectively. The input for each selection approach includes t, the current generation number, G_t, the genotype of generation t, β, the effect of all alleles, and the remaining budget. The output is a matrix $S \in \mathbb{Z}_+^{K_t \times 3}$, where K_t is the number of crosses to be made in generation t, the first two columns are the indices of the breeding parents, and the third column indicates the numbers of progeny to be produced from the crosses. For example,

$S = \begin{bmatrix} 12 & 45 & 10 \\ 3 & 27 & 17 \\ 45 & 67 & 15 \\ 38 & 38 & 20 \end{bmatrix}$ means that the genomic selection approach decided to make the following four crosses: individuals #12 and #45 are crossed to produce 10 progeny, #3 and #27 are crossed to produce 17 progeny, #45 and #67 are crossed to produce 15 progeny, and #38 is self-pollinated to produce 20 progeny. Such a selection decision would cost $\$5 \times 4 + \$10 \times (10 + 17 + 15 + 20) = \$640$.

The same maize data set used in [Goiffon et al., 2017] was used for this competition, which was a combination of SNPs from [Leiboff et al., 2015] and additional ones genotyped using tGBS [Schnable et al., 2013] and phased using Beagle [Browning and Browning, 2008]. The data set consisted of $L = 1,406,757$ SNPs distributed across 10 chromosomes. The data set also included r, a vector of recombination frequencies, and β, a vector of genetic effects for desirable alleles. As reported in [Goiffon et al., 2017], the recombination rates "were estimated using the genetic map developed from the maize nested association mapping (NAM) [Yu et al., 2008] population." The average value of recombination frequencies was 1.38×10^{-5} with a standard deviation of 0.0014. The genetic effects were estimated using the 369 shoot apical meristem phenotypes [Leiboff et al., 2015] and the BayesB model [Meuwissen et al., 2001] implemented in GenSel [Fernando and Garrick, 2009]. We assumed that marker effects were additive with no interactions and that inaccuracies in marker effect estimation affected all selection approaches equally.

The total budget of $B = \$30,750$ was enough to keep the same population size of 300 for 10 generations by making 15 crosses and producing 20 progeny per cross, costing $\$5 \times 15 + \$10 \times 15 \times 20 = \$3,075$ per generation. However, a genomic selection approach may choose to allocate the total budget to the 10 generations in any other manner. As such, except for the initial population with $N_0 = 300$, the population sizes may vary from generation to generation and depend on the decisions of the selection approach.

3.1 IE 312 winning approach

Ms. Mriga Kher, a senior in the Department of Industrial and Manufacturing Systems Engineering was the winner of the competition in the IE 312 class. Her approach is summarized as follows.

- Resource allocation: total resource is evenly allocated to 10 generations, so \$3,075 is spent in each generation.
- Number of crosses: the number of crosses to make is twice the number of remaining generations: $K = 2 \times (11 - t)$. The motivation is that in earlier generations more individuals should be selected to maintain genetic diversity, whereas in later generations fewer crosses should be made and a larger number of progeny produced per cross to increase the probability of creating outstanding outlier progeny by the deadline.
- Number of progeny: produce the same number of progeny from each cross and make this number as large as resource allows. Table 1 summarizes the number of crosses K_t, number of progeny per cross M_t, and population size N_t for each generation t.
- Selection strategy: selection is based on the GEBVs and time. In earlier generations, individuals with the highest GEBVs are selected, whereas in later generations, lower GEBVs are also included. Suppose the individuals are indexed in the descending order of their GEBVs,

then the indices of the $2K$ selected individuals in generation t are $\{1, 2, ..., K-1, K, 8+t, 9+t, ..., 6+K+t, 7+K+t\}$. This strategy was motivated by the observation that genetic diversity deteriorates quickly over time, and including individuals with lower GEBVs was intended to help maintain the diversity.

- Mating strategy: The K individuals with higher GEBVs are paired up and mated with the other K with lower GEBVs: $\{(1, 8+t), (2, 9+t), ..., (K-1, 6+K+t), (K, 7+K+t)\}$. The motivation is to pair up individuals that are not very similar with each other, and the differences in the GEBV rankings were used as an indication of the similarity of individuals.

3.2 IE 534 winning approach

Mr. Will Johnson, a graduate student in the Department of Aerospace Engineering, was the winner of the competition in the IE 534 class. His approach is summarized as follows.

- Resource allocation: same as IE 312.
- Number of crosses: the number of crosses to make is six times the number of remaining generations: $K = 6 \times (11 - t)$.
- Number of progeny: produce the same number of progeny from each cross and make this number as large as resource allows. Table 2 summarizes the number of crosses K, number of progeny per cross M, and population size N for each generation t.
- Selection and mating strategies: these decisions were made jointly using the following three steps.

  Step 1: Let $\mathcal{N}$ denote the set of indices of $0.05N_t$ progeny with the highest GEBVs: $\mathcal{N} \subseteq \{1, ..., N_t\}$, $|\mathcal{N}| = 0.05N_t$, and $\texttt{GEBV}(\hat{n}) \geq \texttt{GEBV}(n), \forall \hat{n} \in \mathcal{N}, n \notin \mathcal{N}$.

  Step 2: For all $n_1, n_2 \in \mathcal{N}$, evaluate crossing individuals n_1 and n_2 by the following measure:

$$v^{534}(n_1, n_2) = \sum_{i=1}^{L} \max\{G_{i,1,n_1}, G_{i,2,n_1}, G_{i,1,n_2}, G_{i,2,n_2}\}.$$

  Step 3: Select K crosses with the highest $v^{534}(n_1, n_2)$. As such, an individual may be used in multiple crosses and/or self pollination.

  The motivation of these steps was to select K most complementary pairs of individuals. The complementarity of two individuals is measured by function $v^{534}(n_1, n_2)$, which is the number of loci where at least one of the couple possesses a desirable allele. The complementarity of a pair is an indication of the long-term potential of their offspring, which could inherent desirable alleles from both parents. The removal of the 95% individuals with lowest GEBVs propels the improvement of the population's GEBV and reduces computational time.

3.3 IE 634 winning approach

Mr. Guodong Zhu, a graduate student in the Department of Aerospace Engineering, was the winner of the competition in the IE 634 class. His approach is summarized as follows.

- Resource allocation: same as IE 312.
- Number of crosses: same as IE 312.
- Number of progeny: same as IE 312.
- Selection and mating strategies: these decisions were made jointly. The K pairs of breeding parents are selected using the following three iterative steps.

  Step 0: Initialize $k = 1$. Let $\mathcal{N}$ denote the set of indices of 100 progeny with the highest GEBVs: $\mathcal{N} \subseteq \{1, ..., N_t\}$, $|\mathcal{N}| = 100$, and $\texttt{GEBV}(\hat{n}) \geq \texttt{GEBV}(n), \forall \hat{n} \in \mathcal{N}, n \notin \mathcal{N}$.

  Step 1: Select a pair of individuals (n_1^k, n_2^k) so that n_1^k has the highest GEBV in $\mathcal{N}$:

  $$\texttt{GEBV}(n_1^k) \geq \texttt{GEBV}(n), \forall n \in \mathcal{N},$$

  and $n_2^k \in \mathcal{N}$ is the most complementary with n_1^k:

  $$v^{634}(n_1^k, n_2^k, t) \geq v^{634}(n_1^k, n, t), \forall n \in \mathcal{N}.$$

  Here the complementarity function for individuals n_1 and n_2 in generation t is defined as

  $$v^{634}(n_1, n_2, t) = \sum_{i=1}^{L} \beta_i f^t(G_{i,1,n_1} + G_{i,2,n_1} + G_{i,1,n_2} + G_{i,2,n_2}),$$

  where the $f^t(v)$ function is defined in Table 3.

  Step 2: If $k \geq K$, then stop. Otherwise update $k \leftarrow k + 1$. If $t \leq 8$, also update $\mathcal{N} \leftarrow \mathcal{N} \backslash \{n_1^k, n_2^k\}$. Go to Step 1.

  The motivation of these steps was to select K pairs that represent a good tradeoff between achieved GEBVs and potential for further genetic gains. For each pair of breeding parents, one parent should have the highest GEBV among all selectable ones (focusing on achievement), and the other parent should be the most complementary to its spouse (focusing on potential, which is positively correlated to the complementarity of the couple). The removal of the lowest GEBVs propels the improvement of the population's GEBV and reduces computational time. Monogamy is imposed in the first 8 generations to maintain genetic diversity and relaxed later on to increase the chance of producing outstanding offspring by the terminal generation. With a Masters degree in aircraft control, Mr. Zhu borrowed ideas from gain scheduling in nonlinear control [Khalil, 1996] when designing the complementarity function $v^{634}(n_1, n_2, t)$.

4 Results

4.1 Simulation tool

An Octave [Eaton et al., 2015] based simulation tool was developed by Wang and his research team to implement the selection process described in Figure 1. Figure 4 shows the result of a random simulation using the conventional GS approach. The dark blue bars, light blue area, and red curve are, respectively, the histogram, range, and mean of the population's GEBVs. The boundaries of the

white and gray areas are the upper and lower selection limits [Cole and VanRaden, 2011] defined as $\sum_{i=1}^{L} \beta_i \max_{n\in\{1,\ldots,N\}} \max_{j\in\{1,2\}} G_{i,j,n}$ and $\sum_{i=1}^{L} \beta_i \min_{n\in\{1,\ldots,N\}} \min_{j\in\{1,2\}} G_{i,j,n}$, respectively. In this figure, the maximum GEBV in generation 10 was 7.88.

Since the conventional GS approach only specifies the selection strategy, we make the following assumptions on the resource allocation and mating strategies. In each generation, \$3,075 is spent to select 30 individuals and make 15 crosses, producing 20 progeny per cross. The 30 selected individuals are paired up in descending order of their GEBVs, i.e., the individual with the highest GEBV is crossed with the second, the third with the fourth, and so on.

4.2 Comparison with GS

The three engineering approaches were compared with the conventional GS approach. Figure 5 shows a random simulation result for each of these four approaches. The maximum GEBVs achieved by GS, IE 312, IE 534, and IE 634 approaches in generation 10 were, respectively, 7.88, 8.35, 8.53, and 8.36.

Since simulation results are affected by uncertain recombination events in the Reproduction step of the breeding process, we further examine the performances of the four approaches under uncertainty by running 500 independent simulation repetitions. Figure 6 plots the cumulative distribution functions (CDFs) of the population maximum GEBV in generation 10, which compares the performances of the four approaches across percentiles. Ideally, the best genomic selection approach would have high GEBV values in all percentiles, positioning vertically on the far right side of the figure.

Results from Figures 5 and 6 suggest that all three engineering approaches outperformed the conventional GS approach. Between the 30th and the 100th percentiles, the CDFs of all four approaches were roughly vertical; the average GEBVs of IE 312, IE 534, and IE 634 approaches within such range of percentiles were 8.34, 8.45, 8.38, respectively, outperforming that of the GS approach, which was 7.91. IE 534 achieved 8.54 at the 100th percentile, which compared favorably with 8.40 for IE 312, 8.43 for IE 634, and 7.94 for GS. IE 312 fell behind GS between the 5th and 15th percentiles, so did IE 534 between the 5th and 30th. IE 634 maintained its GEBV at 8.28 even at the 1st percentile, significantly higher than IE 534 at 7.32, IE 312 at 6.99, and GS at 6.74.

4.3 Comparison with GS, WGS, OHV, GB, and OPV

We also conducted another experiment to compare the three engineering approaches with all five previous approaches. We used a slightly different simulation setting in order to eliminate the potential advantages of the three engineering approaches, which were fine tuned for the competition data set. In each simulation, 200 individuals were randomly selected from the 369 lines in the original data set to form an initial population, and each random initial population was used once for all eight approaches. A deadline of $T = 10$ generations and a total budget of $B = \$20,500$ was used, which was the cost to keep the same population size of 200 for 10 generations by making 10 crosses and producing 20 progeny per cross, costing $\$5 \times 10 + \$10 \times 10 \times 20 = \$2,050$ per generation. This is the same simulation setting as was used in [Goiffon et al., 2017], which was to the advantage of OHV, GB, and OPV approaches, since their parameters were fine tuned for such data set and simulation setting.

For the previous approaches from the literature, similar assumptions made for GS in the previous simulation were also made here, e.g., \$2,050 is spent in each generation, selecting 20 individuals to

make 10 crosses, each producing 20 progeny. The mating strategies for WGS and OHV were based on descending orders of the individuals' WGEBVs and OHVs. The heuristic algorithm proposed in [Goiffon et al., 2017] was used for the selection and mating decisions for the two population-based selection approaches GB and OPV.

We conducted 2,000 independent random simulation repetitions, and results are summarized in Figure 7. The IE 312 approach had very similar but slightly weaker performance than GB, which was one of the most efficient genomic selection approaches in the literature. The IE 534 approach behaved comparably with GB above the 80th percentile and below the 40th, but it underperformed most of the other approaches otherwise. The IE 634 approach significantly outperformed all other approaches by dominating at almost every percentile. In particular, it maintained an 8.03 GEBV at the 25th percentile, whereas the second highest GEBV at this percentile was 7.17 from OPV, an approach recently proposed by Wang and his collaborators. Results of the IE 634 approach had a much larger variability in Figure 7 than in Figure 6, which was due to the different simulation settings. In Section 4.2, all the 369 lines in the data set were used as the initial population for all repetitions, and uncertainty mainly came from random recombination events; the performance of the IE 634 approach was very robust and consistent. In this section, however, 200 out of 369 lines were randomly selected to form the initial population in each repetition, thus uncertainty originated from the randomness in both initial populations and recombination events. Due to the consistency of IE 634 results in Section 4.2, uncertain initial populations were likely to be responsible for the majority of the variability of IE 634 results in Figure 7. These observations appeared to suggest that the IE 534 approach was well-calibrated and optimized for the particular data set with 369 lines, which slightly outperformed IE 634 approximately 65% of the time; in the other 35% of the time, however, it significantly underperformed. The IE 634 approach appeared to be designed for more general data sets. When tested with different initial populations, the robustness of the IE 634 approach paid off and led to dominating performances against not only IE 534 but also other approaches that were compared.

5 Conclusions

Our work made three new contributions to the research field of genomic selection. First, we pointed out two limitations that previous genomic selection approaches have and presented three new approaches as a first attempt to address these limitations. Second, the effectiveness of the three new approaches, especially IE 634, suggested new directions for future research in the design of more effective genomic selection approaches. Third, this research demonstrated that engineers can overcome disciplinary barriers and contribute at the forefront of research innovation in plant breeding by developing effective decision-making methods and tools.

6 Acknowledgement

The authors thank the Editor and three anonymous reviewers for their insightful feedback. Wang was partially supported by the U.S. Department of Agriculture NIFA Award 2017-67007-26175 and the Plant Sciences Institute at Iowa State University.

7 Contribution of authors

Wang formulated the problem, collected data, conducted computational experiments and analysis, and wrote the manuscript. All the other three co-authors contributed equally by designing their own algorithm and providing feedback to the draft of the manuscript.

Conflict of interest: The authors declared that they have no conflict of interest.

8 Figures and Table Legends

Figure 1 Flowchart of the genomic selection process.

Figure 2 Cumulative distribution functions of population maximums after 10 generations of selection over 2,000 replications for each selection approach. Adopted from Figure 2 in [Goiffon et al., 2017].

Figure 3 The performance of a genomic selection approach depends on the availability of time and resources of the breeding project. In the left subfigure, Approach 2 outperforms Approach 1 if compared at time t_1 and otherwise at t_2. In the right subfigure, a random progeny from Distribution 2 is expected to have a higher genetic value than that from Distribution 1, but if a large number of random progeny are produced from each distribution, then the best one from Distribution 1 is expected to be superior to the best one from Distribution 2.

Figure 4 A sample simulation result using the GS approach. The dark blue bars, light blue area, and red curve are, respectively, the histogram, range, and mean of the population's GEBVs. The boundaries of the white and gray areas are the upper and lower selection limits. The maximum GEBV in generation 10 was 7.88.

Figure 5 Sample simulation results using the GS, IE 312, IE 534, and IE 634 approaches, whose maximum GEBVs in generation 10 were 7.88, 8.35, 8.53, and 8.36, respectively.

Figure 6 Cumulative distribution functions of population maximum from with 500 repetitions using GS, IE 312, IE 534, and IE 634 approaches. Results were obtained using the first 300 individuals from the data set as the initial population.

Figure 7 Cumulative distribution functions of population maximum from 2,000 repetitions using GS, WGS, OHV, GB, OPV, IE 312, IE 534, and IE 634 approaches. Results were obtained using randomly selected 200 individuals from the data set as the initial population.

Table 1 Number of crosses K_t, number of progeny per cross M_t, and population size N_t for each generation t.

Table 2 Number of crosses K_t, number of progeny per cross M_t, and population size N_t for each generation t.

Table 3 Definition of function $f^t(v)$.

References

[Browning and Browning, 2008] Browning, B. and Browning, S. (2008). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics*, 84(2):210–223.

[Cole and VanRaden, 2011] Cole, J. and VanRaden, P. (2011). Use of haplotypes to estimate mendelian sampling effects and selection limits. *Journal of Animal Breeding and Genetics*, 128(6):446–455.

[Collard and Mackill, 2008] Collard, B. C. and Mackill, D. J. (2008). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1491):557–572.

[Daetwyler et al., 2015] Daetwyler, H., Hayden, M., Spangenberg, G., and Hayes, B. (2015). Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. *Genetics*, 200:1341–1348.

[Dekkers and Chakraborty, 2001] Dekkers, J. and Chakraborty, R. (2001). Potential gain from optimizing multigeneration selection on an identified quantitative trait locus. *Journal of animal science*, 79(12):2975–2990.

[Dekkers and Van Arendonk, 1998] Dekkers, J. and Van Arendonk, J. (1998). Optimizing selection for quantitative traits with information on an identified locus in outbred populations. *Genetics Research*, 71(3):257–275.

[Eaton et al., 2015] Eaton, J. W., Bateman, D., Hauberg, S., and Wehbring, R. (2015). *GNU Octave version 4.0.0 manual: a high-level interactive language for numerical computations*.

[Fernando and Garrick, 2009] Fernando, R. and Garrick, D. (2009). Gensel – user manual for a portfolio of genomic selection related analyses. Technical report.

[Frisch et al., 1999] Frisch, M., Bohn, M., and Melchinger, A. A. (1999). Minimum sample size and optimal positioning of flanking markers in marker-assisted backcrossing for transfer of a target gene. *Crop Science*, 39(4):967–975.

[Goddard, 2009] Goddard, M. (2009). Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica*, 136(2):245–257.

[Goiffon et al., 2017] Goiffon, M., Kusmec, A., Wang, L., Hu, G., and Schnable, P. (2017). Optimal population value selection: A population-based selection approach for improving response in genomic selection. Technical report. to appear in *Genetics*.

[Han et al., 2017] Han, Y., Cameron, J., Wang, L., and Beavis, W. (2017). The predicted cross value for genetic introgression of multiple alleles. Technical report. to appear in *Genetics*.

[Jannink, 2010] Jannink, J.-L. (2010). Dynamics of long-term genomic selection. *Genetics Selection Evolution*, 42(1):1–11.

[Kemper et al., 2012] Kemper, K. E., Bowman, P. J., Pryce, J. E., Hayes, B. J., and Goddard, M. (2012). Long-term selection strategies for complex traits using high-density genetic markers. *Journal of Dairy Science*, 95(8):4646–4656.

[Khalil, 1996] Khalil, H. K. (1996). *Noninear Systems*. Prentice-Hall, New Jersey.

[Leiboff et al., 2015] Leiboff, S., Li, X., Hu, H.-C., Todt, N., Yang, J., Li, X., Yu, X., Muehlbauer, G. J., M. C.P. Timmermans, J. Yu, P. S. S., and Scanlon, M. (2015). Genetic control of morphometric diversity in the maize shoot apical meristem. *Nature Communications*, 6:1–10.

[Meuwissen et al., 2001] Meuwissen, T., Hayes, B., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829.

[Riedelsheimer and Melchinger, 2013] Riedelsheimer, C. and Melchinger, A. E. (2013). Optimizing the allocation of resources for genomic selection in one breeding cycle. *Theoretical and applied genetics*, 126(11):2835–2848.

[Schnable et al., 2013] Schnable, P., Liu, S., and Wu, W. (2013). Device for the treatment of hiccups.

[Xu et al., 2011] Xu, P., Wang, L., and Beavis, W. (2011). An optimization approach to gene stacking. *European Journal of Operational Research*, 214(1):168–178.

[Yu et al., 2008] Yu, J., Holland, J., McMullen, M., and Buckler, E. (2008). Genetic design and statistical power of nested association mapping in maize. *Genetics*, 178(1):539–551.

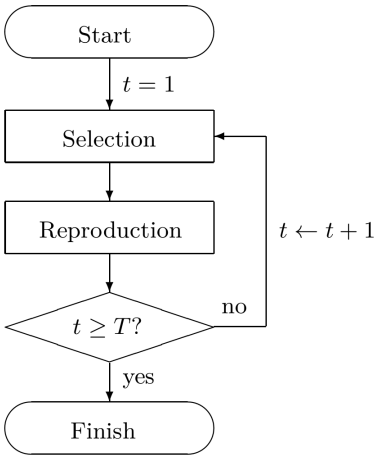

Figure 1: Flowchart of the genomic selection process

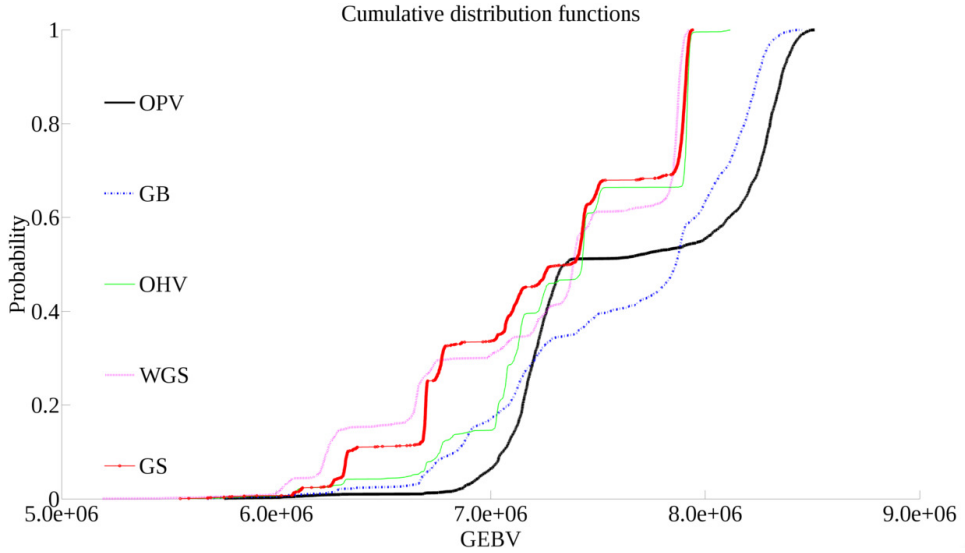

Figure 2: Cumulative distribution functions of population maximums after 10 generations of selection over 2,000 replications for each selection approach. Adopted from Figure 2 in [Goiffon et al., 2017].

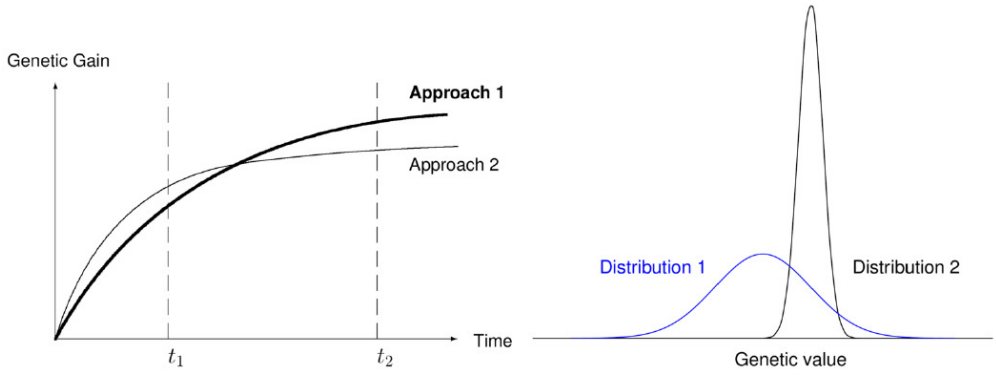

Figure 3: The performance of a genomic selection approach depends on the availability of time and resources of the breeding project. In the left subfigure, Approach 2 outperforms Approach 1 if compared at time t_1 and otherwise at t_2. In the right subfigure, a random progeny from Distribution 2 is expected to have a higher genetic value than that from Distribution 1, but if a large number of random progeny are produced from each distribution, then the best one from Distribution 1 is expected to be superior to the best one from Distribution 2.

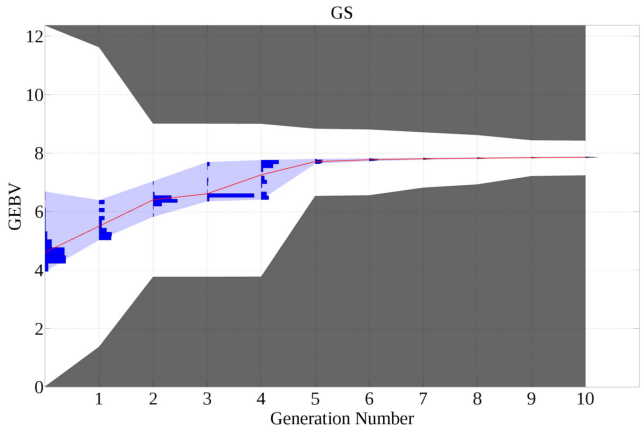

Figure 4: A sample simulation result using the GS approach. The dark blue bars, light blue area, and red curve are, respectively, the histogram, range, and mean of the population's GEBVs. The boundaries of the white and gray areas are the upper and lower selection limits. The maximum GEBV in generation 10 was 7.88.

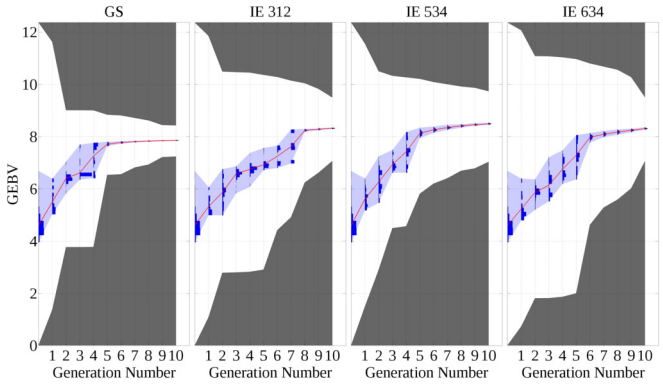

Figure 5: Sample simulation results using the GS, IE 312, IE 534, and IE 634 approaches, whose maximum GEBVs in generation 10 were 7.88, 8.35, 8.53, and 8.36, respectively.

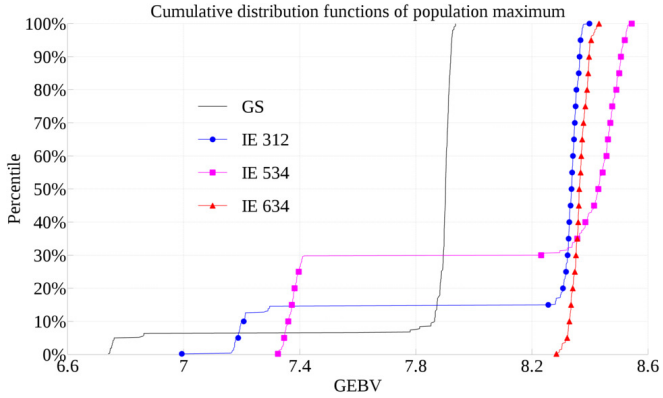

Figure 6: Cumulative distribution functions of population maximum from with 500 repetitions using GS, IE 312, IE 534, and IE 634 approaches. Results were obtained using the first 300 individuals from the data set as the initial population.

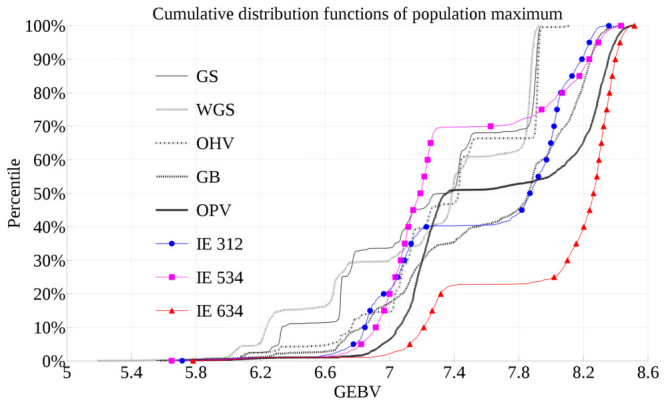

Figure 7: Cumulative distribution functions of population maximum from 2,000 repetitions using GS, WGS, OHV, GB, OPV, IE 312, IE 534, and IE 634 approaches. Results were obtained using randomly selected 200 individuals from the data set as the initial population.

Table 1: Number of crosses K_t, number of progeny per cross M_t, and population size N_t for each generation t.

t	1	2	3	4	5	6	7	8	9	10
K_t	20	18	16	14	12	10	8	6	4	2
M_t	14	16	18	21	25	31	39	52	79	160
N_t	280	288	288	294	300	310	312	312	316	320

Table 2: Number of crosses K_t, number of progeny per cross M_t, and population size N_t for each generation t.

t	1	2	3	4	5	6	7	8	9	10
K_t	60	54	48	42	36	30	24	18	12	6
M_t	4	5	6	6	8	10	13	17	27	55
N_t	240	270	288	252	288	300	312	306	324	330

Table 3: Definition of function $f^t(v)$.

$f^t(v)$	$v = 0$	$v = 1$	$v = 2$	$v = 3$	$v = 4$
$1 \leq t \leq 8$	0.0	2.4	2.8	3.6	4.0
$t = 9$	0.0	2.0	2.4	3.4	4.0
$t = 10$	0.0	1.0	2.0	3.0	4.0