
Saba Moeinizade
*Iowa State University*, sabamz@iastate.edu

Guiping Hu
*Iowa State University*, gphu@iastate.edu

Lizhi Wang
*Iowa State University*, lzwang@iastate.edu

Patrick Schnable
*Iowa State University*, schnable@iastate.edu

Follow this and additional works at: [https://lib.dr.iastate.edu/imse_pubs](https://lib.dr.iastate.edu/imse_pubs)

Part of the Genomics Commons, and the Operational Research Commons

The complete bibliographic information for this item can be found at [https://lib.dr.iastate.edu/imse_pubs/220](https://lib.dr.iastate.edu/imse_pubs/220). For information on how to cite this item, please visit [http://lib.dr.iastate.edu/howtocite.html](http://lib.dr.iastate.edu/howtocite.html).

This Article is brought to you for free and open access by the Industrial and Manufacturing Systems Engineering at Iowa State University Digital Repository. It has been accepted for inclusion in Industrial and Manufacturing Systems Engineering Publications by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

Abstract
New genotyping technologies have made large amounts of genotypic data available for plant breeders to use in their efforts to accelerate the rate of genetic gain. Genomic selection (GS) techniques allow breeders to use genotypic data to identify and select, for example, plants predicted to exhibit drought tolerance, thereby saving expensive and limited field-testing resources relative to phenotyping all plants within a population. A major limitation of existing GS approaches is the trade-off between short-term genetic gain and long-term potential. Some approaches focus on achieving short-term genetic gain at the cost of reduced genetic diversity necessary for long-term gains. In contrast, others compromise short-term progress to preserve long-term potential without consideration of the time and resources required to achieve it. Our contribution is to define a new “look-ahead” metric for assessing selection decisions, which evaluates the probability of achieving high genetic gains by a specific time with limited resources. Moreover, we propose a heuristic algorithm to identify optimal selection decisions that maximize the look-ahead metric. Simulation results demonstrate that look-ahead selection outperforms other published selection methods.

Keywords
Genetic gain, Genomic Selection, Look-ahead Selection, Simulation, Optimization, Genomic Prediction, GenPred, Shared Data Resources

Disciplines
Genomics | Operational Research

Comments

Creative Commons License
This work is licensed under a Creative Commons Attribution 4.0 License.

Saba Moeinizade,*1 Guiping Hu,* Lizhi Wang,* and Patrick S. Schnable†
*Department of Industrial and Manufacturing Systems Engineering, Iowa State University and †Department of Agronomy, Iowa State University
ORCID IDs: 0000-0001-7402-3385 (S.M.); 0000-0001-8392-8442 (G.H.); 0000-0002-5527-4047 (L.W.); 0000-0001-9169-5204 (P.S.S.)

ABSTRACT New genotyping technologies have made large amounts of genotypic data available for plant breeders to use in their efforts to accelerate the rate of genetic gain. Genomic selection (GS) techniques allow breeders to use genotypic data to identify and select, for example, plants predicted to exhibit drought tolerance, thereby saving expensive and limited field-testing resources relative to phenotyping all plants within a population. A major limitation of existing GS approaches is the trade-off between short-term genetic gain and long-term potential. Some approaches focus on achieving short-term genetic gain at the cost of reduced genetic diversity necessary for long-term gains. In contrast, others compromise short-term progress to preserve long-term potential without consideration of the time and resources required to achieve it. Our contribution is to define a new “look-ahead” metric for assessing selection decisions, which evaluates the probability of achieving high genetic gains by a specific time with limited resources. Moreover, we propose a heuristic algorithm to identify optimal selection decisions that maximize the look-ahead metric. Simulation results demonstrate that look-ahead selection outperforms other published selection methods.

KEYWORDS Genetic gain Genomic Selection Look-ahead Selection Simulation Optimization Genomic Prediction GenPred Shared Data Resources

Feeding the world’s growing population remains a significant challenge. Advances in plant breeding have been instrumental in improving agricultural output. Classical plant breeding programs rely on the phenotyping of progenies in field trials to identify superior individuals. The number of individuals that can be phenotyped is resource limited (Rincent et al. 2017), which limits genetic gain. Genomic selection (GS) refers to using a set of markers distributed across the genome to estimate the breeding value of selection candidates for quantitative traits (Goddard 2009). GS makes it possible to predict the performance of unphotted individuals from readily available genotyping data (Rincent et al. 2017; Meuwissen et al. 2001). Genomic Estimated Breeding Value (GEBV) of individual plants (or animals) has been widely adopted as the selection criteria; it selects individuals based on the sum of their estimated marker effects (Meuwissen et al. 2001). This approach has been widely adopted in GS practice due to its effectiveness in achieving short-term genetic improvements. More recently, two methods have been proposed to improve conventional GS (CGS): the optimal haploid value (OHV) (Daetwyler et al. 2015) and the optimal population value (OPV) (Goiffon et al. 2017). Simulation experiments and some empirical studies have shown that CGS selection results in rapid genetic gains (Hayes et al. 2009; Lorenzana and Bernardo 2009; VanRaden et al. 2009; Jannink 2010). However, CGS focuses on one or two cycles of selection and does not guarantee long-term gain (Sonesson et al. 2012; Lin et al. 2017; Gorjanc et al. 2018; Akdemir et al. 2019). The OHV method, calculates the GEBV of the best possible doubled haploid (DH) derived from an individual (Daetwyler et al. 2015). This method focuses selection on haplotypes and optimizes the breeding program toward the end goal of generating an elite fixed line (Daetwyler et al. 2015). Simulation studies have shown that OHV...
selection results in more genetic gain and diversity as compared to CGS (Daetwyler et al. 2015). CGS and OHV are truncation selection approaches in that they rank individuals and select the top fraction of the population. In contrast, OPV is a group-based selection strategy. Specifically, OPV selects the best group of individuals based on their interactive effects and calculates the GEBV of the best possible progeny from this group produced after an unlimited number of generations, which may require a large amount of time and resources to achieve (Goiffon et al. 2017). In this paper, we extend OPV by again selecting groups of individuals as a unit, but propose an innovative method for selecting groups, "look-ahead selection" (LAS). This new selection method can improve genetic gain by maximizing the expected GEBV of the best offspring in the terminal generation. It makes the optimal trade-off between short-term gain and long-term potential to achieve the highest genetic gain within a specified time.

MATERIALS AND METHODS

Generic formulation for GS methods

In this section, we present a generic formulation for existing GS methods namely, CGS, OHV, OPV, and the new selection method, LAS. Equations (1), (2), and (3) show this genetic optimization formulation.

\[
\max_x f(x) \quad (1)
\]

such that \( \sum_{n=1}^{N} x_n = S \quad (2) \)

\( x_n \in \{0, 1\}, n \in \{1, \ldots, N\} \quad (3) \)

Here,

- \( N \) is the number of individuals in the population.
- \( x_n \) is a binary decision variable that shows whether individual \( n \) is selected \((x_n = 1)\) or not \((x_n = 0)\).
- \( S \) is the number of individuals that are to be selected out of the current population.

It should be observed that the only difference among the three previous methods is in their objective functions as they aim to maximize different objectives. The objective function of the optimization problem, \( f(x) \) is formulated as \( f(x)_{\text{CGS}} \), \( f(x)_{\text{OHV}} \), and \( f(x)_{\text{OPV}} \) in equations (4), (5), and (6) respectively.

Conventional genomic selection: Meuwissen et al. (2001) proposed to evaluate an individual as a breeding parent by its genomic estimated breeding value (GEBV), which is the sum of all marker effects across the entire genome, as defined in equation (4). The CGS method selects individuals with the highest GEBVs.

\[
f(x)_{\text{CGS}} = \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{m=1}^{2} G_{l,m,n} \beta_l x_n. \quad (4)
\]

Here, the notations are defined as follow:

- \( L \): The number of marker loci.
- \( G_{l,m,n} \in \{0, 1\}, \forall l \in \{1, 2, \ldots, L\}, \forall m \in \{1, 2\} \) and \( \forall n \in \{1, 2, \ldots, N\} \): The genotypic information of locus \( l \) from chromosome \( m \) of individual \( n \), with 1 and 0 representing the major and minor allele, respectively.
- \( \beta_l \): The normalized effect of the major allele at locus \( l \), with that for the minor allele being 0.
- \( M \): The ploidy of the plants. We use diploid species \((M = 2)\) as an example in this paper.

To maximize long-term response, the weighted genomic selection (Goddard 2009; Jannink 2010) was proposed as a variation of the CGS method by emphasizing the preservation of rare favorable alleles. It replaced the allele effect \( \beta_l \) in equation (4) with \( \frac{\sqrt{\max(w_l, 1/N)}}{\sqrt{\max(w_l, 1/N)}} \), where \( w_l \) is the frequency of favorable alleles at locus \( l \) among all individuals in the population. As such, this variation gives a higher weight to low-frequency favorable alleles. Notice that the denominator \( \sqrt{\max(w_l, 1/N)} \) is equal to \( \sqrt{w_l} \) except for \( w_l = 0 \) when \( G_{l,m,n} = 0 \) for all \( m \) and \( n \).

Optimal haploid value: More than a decade after the CGS method, OHV was proposed to combine the creation of doubled haploids with GS methods and evaluates the potential of producing elite doubled haploids (Daetwyler et al. 2015). Equation (5) shows the objective function for OHV selection. This method selects individuals with the highest OHVs.

\[
f(x)_{\text{OHV}} = \sum_{n=1}^{N} \max_{w \in \{1, 2\}} \sum_{l \in H(b)} G_{l,m,n} \beta_l x_n. \quad (5)
\]

Here, segments of adjacent markers are clustered into haplotypes, which are defined as follows:

- \( B \): The number of haplotype blocks per chromosome.
- \( H(b), \forall b \in \{1, \ldots, B\} \): The set of marker loci that belong to haplotype block \( b \).

The OHV of an individual is the GEBV of its best possible DH progeny. Recombination events are assumed to be possible between haplotypes but not within them. This assumption reduces the computational effort of the algorithm.

What also makes CGS and OHV computationally efficient is the fact that they are both truncation selection methods, which assumes that the contribution of breeding parents are separable and additive. Mathematically, the summation operator \( \sum_{n=1}^{N} \) in equation (4) and (5) suggests that the maximization of the objective functions \( f(x)_{\text{CGS}} \) or \( f(x)_{\text{OHV}} \) can be easily achieved by evaluating each individual \( n \) separately and setting \( x_n = 1 \) for the ones with the highest GEBVs or OHVs. Compared with CGS, OHV represents an important shift of the selection objective from maximizing genetic achievement of the parents to that of their progeny.

Optimal population value: OPV selection is an extension to OHV which evaluates the breeding merit of a set of individuals instead of evaluating the breeding value of a single individual (Goiffon et al. 2017). The OPV of breeding population \( S \) is the GEBV of the best possible progeny produced after an unlimited number of generations. The objective function for the OPV method is defined as follows:

\[
f(x)_{\text{OPV}} = \max_{n \in \{1, \ldots, N\}} \max_{m \in \{1, 2\}} \sum_{l \in H(b)} G_{l,m,n} \beta_l x_n. \quad (6)
\]

OPV represents another important shift of the selection objective from individual-based truncation selection to group-based selection. The contribution of a breeding parent is evaluated based on not only the favorable alleles that it carries but also the favorable alleles that it carries but are missing in other selected breeding parents. A limitation of OPV is that the objective function \( f(x)_{\text{OPV}} \) is a lot harder to optimize, since it is no longer separable with respect to \( x \). As a result, heuristic algorithms were used to identify good but not necessarily optimal selections.

Potential improvements

The success of CGS has been demonstrated in numerous simulation and field experiments, especially in achieving short-term genetic gains in
both plant and animal breeding (Meuwissen 1997; Rosvall 1999; Hayes et al. 2002; Ulrich 2007; Hayes et al. 2009; Lorenzana and Bernardo 2009; VanRaan et al. 2009; Jannink et al. 2010; Mujibi et al. 2011; Nakaya and Isobe 2012; Hallatschek and Geyrhofer 2015). OHV and OPV were proposed as extensions of CGS to improve long-term genetic gains, which have been shown to be effective in simulation studies. Herein, we identify three areas in genomic selection that can be made more efficient and present a new genomic selection method that attempts to address each of these three areas.

First, time management. For a given population of individuals, the optimal selection decision should depend on whether the deadline of the breeding project is in the near future or far down the road. However, none of the aforementioned three methods take deadlines into consideration.

Second, mating strategy. All three methods focus on selecting breeding parents without explicitly indicating how they should be mated in pairs, but several studies have observed that different mating decisions may affect genetic gain (Toro and Varona 2010; Kinghorn 2011; Sun et al. 2013; Akdemir and Sánchez 2016; Liu et al. 2017; Wang et al. 2018).

Third, resource allocation. Intuitively, making more crosses and producing more progenies leads to a higher chance of creating outstanding individuals from the progeny population, but this also requires more resources. Allocating a fixed total budget over a period of time to achieve the best final outcome is therefore a strategic decision that needs to be optimized (Lorenz 2013).

Look-ahead selection

The cornerstone of the LAS method is a new definition of the objective function, $f^{\text{LAS}}(x, y, r, T - t)$, that reflects what truly matters in genomic selection. The input of this function includes selected breeding parents ($x$), mating decisions ($y$), recombination frequencies ($r$), and remaining number of generations ($T - t$), the difference between the current generation number $T$ and the deadline $t$). The former two input terms are decision variables that need to be optimized by the model, whereas the latter two are parameters that the model needs to take into account when searching for the optimal solution. We define $f^{\text{LAS}}$ as the expected GEBV of the best offspring in the terminal generation. In comparison, $f^{\text{CGS}}$ can be interpreted as the genetic achievement of the breeding parents measured in terms of GEBV; and $f^{\text{OHV}}$ and $f^{\text{OPV}}$ represent the best possible progeny that can be produced by, respectively, self-pollination and cross-pollination, both assuming unlimited time and resources. The models for these three methods only differ in the objective functions but share the same constraints (2) and (3), whereas the LAS model requires additional constraints. The LAS method can be formulated as follows.

$$
\max_{x, y} f^{\text{LAS}}(x, y, r, T - t)
$$

such that Constraints (2) and (3)

$$
x_n = \sum_{j=1}^{N} y_{nj} \quad \forall n \in \{1, \ldots, N\}
$$

$$
y_{ij} \in \{0, 1\} \quad \forall i, j \in \{1, \ldots, N\}
$$

The new variables and parameters are defined as follows.

- $y_{ij}$: A binary variable that shows whether individual $i$ is mated with individual $j$ ($y_{ij} = 1$) or not ($y_{ij} = 0$).
- $r \in [0, 0.5]^{l-1}$: The recombination frequency vector.

The remainder of this section will explain how to numerically evaluate the objective function $f^{\text{LAS}}(x, y, r, T - t)$ for any given solution $(x, y)$, how to search for the optimal (or close to optimal) solution $(x^*, y^*)$ that achieves the maximal value in the objective function, and how to allocate resources to improve the rate of genetic gains.

Evaluation of the objective function $f^{\text{LAS}}$: The exact evaluation of the objective function $f^{\text{LAS}}$ is challenging both computationally and analytically due to uncertain recombination events over $T - t$ generations as well as the selection, mating, and resource allocation decisions that will be made therein. To overcome this challenge, we designed a novel simulation method that provides a computationally tractable yet reasonable approximation of the true objective function. Figure 1 illustrates the look-ahead simulation that is based on two simplifying assumptions.

Assumption 1: The selected pairs of breeding parents will each produce one progeny in generation $t + 1$.

Assumption 2: All progenies from generation $t + 1$ to $T - 1$ were crossed with each other (including selfing) in the same generation, each producing one progeny.

As such, the objective function $f^{\text{LAS}}$ can be approximated by taking a random sample of the population in generation $T$ of the look-ahead simulation and calculating the highest GEBV of all individuals.

The following theorem defines the distribution of the progenies in the final generation $T$, which allows efficient evaluation of the approximated objective function.

Theorem 1. Let $G \in \{0, 1\}^{l \times 2^{n_S}}$ denote the genotype of a population in generation $t$ with an even number, $S$, of individuals. Suppose all individuals with odd indices, $\{1, 3, \ldots, S - 1\}$, are respectively mated with the next individuals, $\{2, 4, \ldots, S\}$. These individuals are mated according to Assumptions 1 and 2. Let $g \in \{0, 1\}^{l}$ denote a random gamete produced by breeding parents in meiosis of the $(T - 1)$st generation. The distribution of $g$ can be described by the following equations (11) and (12).

$$
P(g_1 = G_{1, m}) = \frac{1}{2^S} \quad \forall i \in \{1, 2, \ldots, S\}, \forall m \in \{1, 2\}. \quad (11)
$$

$$
P(g_{l+1} = G_{l+1, m', r} | g = G_{l, m'} r) =
$$

$$
\begin{cases}
(1 - r_{l+1})^2(1 - R_{l+1}), & \text{if } \beta^0 = \beta^1 \text{ and } m^0 = m^1 \\
\frac{1}{2} r_{l+1}(1 - R_{l+1}), & \text{if } \beta^0 = \beta^1 \text{ and } m^0 \neq m^1 \\
\frac{R_{l+1}}{2(S - 2)}, & \text{if } \beta^0 \neq \beta^1 \text{ and } m^0 = m^1
\end{cases}
$$

$$
\forall l \in \{1, \ldots, L - 1\}, \forall \beta^i, \beta^0, m^0, m^1 \in \{1, 2\}. \quad (12)
$$

Here, $r \in [0, 0.5]^{l-1}$ is the given vector of recombination frequencies and $R_l$ is the recombination frequency between allele 1 and allele $l + 1$ between generations $t + 2$ and $T$ for all $l \in \{1, \ldots, L - 1\}$, which can be derived as:
\[ R_t = \frac{(S - 2) \left[ 1 - (1 - r_t)^{T-t} \right]}{S}, \]  \hspace{1cm} (13)
phenotypes (Leiboff et al. 2015) using the BayesB model (Meuwissen et al. 2001). Similar to Goiffon et al. (2017), we assumed that marker effects were known and that errors in marker effects have an equal effect on all selection methods. The genetic map developed from maize nested association mapping (NAM) population is used for estimating recombination rates (Yu et al. 2008). To facilitate comparisons, genetic data were scaled such that the maximum potential of the initial breeding population is 100.

The same simulation process (shown in Figure 2) as (Goiffon et al. 2017) was used to compare the four methods in our study. Each of the components in Figure 2 is explained as follows:

- **The initial population start point:** In plant breeding, the genomic selection process starts with an initial population of individuals. The genotypes and marker effects are given at this point. In each simulation run, 200 individuals were selected randomly from the 369 maize inbred lines to make the initial population. Furthermore, the same set of 200 individuals were used as the initial population for all methods to make comparisons consistent.

- **The selection step:** All four methods were used to make selection decisions in this step, including mating strategies, number of crosses to make \( NC \) and number of progenies per cross \( NP \).

- **The reproduction step:** The selected individuals were crossed to make the breeding population for the next generation. A random progeny inherits the genetic information from its breeding parents according to the look-ahead algorithm to make \( NC = 10 \) crosses. The number of progenies for each cross was determined by the heuristic strategy described in Section 2.3.3 with the constraint that the total number of progenies remains 200.

- **The \( t \geq T \) condition:** The breeding cycle repeats itself until generation \( T \), a predetermined deadline.

The selection diagram adopted from Goiffon et al. (2017).

**Figure 3** Cumulative genetic gains over 10 generations for four GS methods.
The final population end point: After the terminal generation, the population will be assessed to determine its genetic improvement over the initial population.

Simulation results
One thousand independent simulation repetitions were performed for each of the four selection approaches. Simulations were conducted on a computer with 256GB RAM and a processor with the following specifications: Intel(R) Xeon(R) CPU E5-4650 0 @2.70GHz 2.70GHz (2 processors). The computation time required for one simulation (including 4 methods) was 6248 sec. Hence, it takes almost 1735 Hours (72 days) to conduct 1000 simulations. Ten different simulations have ran in parallel to reduce the CPU calendar time to 7 days. The LAS method is modestly more computationally intensive. LAS requires approximately two times more computational time than the other three methods. Major results are summarized as follows.

Genetic gains over ten generations: Figure 3 shows the average cumulative genetic gains over ten generations. We define the cumulative genetic gain as the difference between the mean GEBV of the current population and that of the initial population. Because this figure shows genetic gains for each of the four methods averaged across 1,000 simulation repetitions, the comparison reflects their different performances in general. CGS achieved a high rate of genetic gain in the first three generations before gradually reaching a plateau. OHV maintained a relatively high rate of genetic gain throughout ten generations due to its emphasis on the progenies rather than the parents. OPV managed to achieve an even higher genetic gain throughout ten generations due to its emphasis on the progeny performance. The other three methods always make the same selection decisions. LAS adjusts selection decisions based on the user-defined deadline whereas other three methods always make the same selection decisions.

Genetic gains with varying deadlines: LAS is the only method that adjusts selection decisions based on the user-defined deadline. Figure 5 shows the performance of LAS with varying deadlines from \( T = 1 \) to \( T = 10 \). In all ten cases, LAS used a similar strategy to patiently accumulate desirable alleles in early generations and make big leaps in the final two generations. As a result, LAS outperformed all other methods for all tested deadlines. The other three methods make the same selection decisions and thus result in the same performance under different deadlines.

Variable performance across different simulation repetitions: The average values and standard deviations (among the 1,000 simulation repetitions) for population minimum, mean, and maximum in the 10th generation are summarized in Table 1. Figure 6 compares the cumulative distribution functions (CDFs) of the population maximum in generation 10. Here, the horizontal axis shows the GEBV of an individual (representing genetic gains) whereas the vertical axis is the percentile of the simulation repetitions. By definition, the 1st percentile is one of the worst performances within the 1,000 simulation repetitions, the 99th percentile is one of the best, and the 50th percentile is the median value. As such, the further toward the right and bottom directions of the figure a CDF curve, the better performance a method has. The figure shows the improvements of different methods from CGS to LAS. In particular, LAS-X is a reduced version of LAS using the same resource allocation strategy with all previous methods (producing the same number of progenies from each cross), rather than using the heuristic strategy for resource allocation described in Section 2.3.3. These results demonstrated the effectiveness of LAS in making selection, mating, and resource allocation decisions.
Behavior of LAS in the final two generations: LAS has an interesting behavior in the final two generations when it makes big leaps in genetic gain (Figures 3 and 5). This happens because LAS accumulates desirable alleles in the early generations to utilize in the final generations.

Figure 7 presents histograms of population GEBVs over time for one sample simulation using the LAS method. The yellow triangles show the GEBV of selected breeding parents from the population in each generation. This demonstrates how the breeding value rankings of the individuals selected by LAS change by generation. Note that in the last two generations LAS selects individuals with high GEBVs. This explains the behavior of LAS in the final two generations.

CONCLUSIONS

Genomic selection has been instrumental in improving the efficiency of plant breeding. In this study, we introduced a new selection method, LAS, which has the potential to further improve the efficiency of breeding given limited resources and specific user-defined project duration.

Unlike previous methods which try to maximize the genetic achievement of breeding parents or the best possible progeny without considering time and resource constraints, LAS is maximizing what exactly matters in a GS problem by aiming at the right objective. The objective of LAS is to maximize the expected GEBV of the best offspring in the terminal generation given a limited amount of resources. As such, this method is much more computationally challenging than previous ones, due to multiple complex factors such as recombination frequencies, mating strategy, time management, and resource allocation that are explicitly accounted for. To deal with these challenges, we designed a simulation optimization algorithm that estimates and maximizes the LAS objective function by exploring the selection and mating solution space efficiently.

LAS makes three major contributions to the literature on genomic selection. First, LAS is deadline sensitive. Selection decisions adjust to the project duration to make a trade-off between achieving short-term genetic gains and maintaining genetic diversity long-term. Second, LAS optimizes both selection and mating strategies. It recognizes the importance of mating strategies and assigns selected individuals into pairs of breeding parents to achieve further genetic gains. Third, LAS involves resource allocation decisions. Rather than producing the same number of progenies from each cross, it allows breeding parents with higher genetic diversity to produce more progenies to increase the chance of producing high performers.

LAS was compared with previous genomic selection methods in a comprehensive simulation study using empirical data from a population of inbred maize lines. Computational results demonstrated the improvements of LAS over other methods in three perspectives: (1) LAS achieved the highest genetic gain by the deadline of the breeding project, which varied from one generation to ten generations. (2) LAS preserved the highest level of genetic diversity at the end of the breeding project. (3) LAS outperformed all other methods in almost all percentiles in the 1,000 simulation repetitions.

Future research is needed to address the limitations of the LAS method. The first assumption described in Section 2.3.1 is allowing only one progeny to be produced from the selected pairs of breeding parents in generation $t + 1$ and the second assumption is allowing the crosses to be made within the same generation each producing one progeny from generation $t + 1$ to $T - 1$. These two assumptions were made to simplify the computational requirement of estimating the objective function, which inevitably reduced its accuracy. Moreover, future studies can explore more comprehensive comparisons by performing simulations by: 1. using other methods for estimating marker effects such as GBLUP and ridge regression; 2. considering populations with...
different LD structures; and 3. applying different resource allocation strategies.

ACKNOWLEDGMENTS
This work is supported by Agriculture and Food Research Initiative Grant no. 2017-67007-26175/Accession No. 1011702 from the USDA National Institute of Food and Agriculture. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the U.S. Department of Agriculture. This work is also supported by the Plant Sciences Institute’s Faculty Scholars program at Iowa State University.

LITERATURE CITED


Leiboff, S., X. Li, H.-C. Hu, N. Todt, J. Yang et al., 2015 Genetic control of morphometric diversity in the maize shoot apical meristem. Nat. Commun. 6: 8974. https://doi.org/10.1038/ncomms9974


Rosvall, O., 1999 Enhancing gain from long-term forest tree breeding while conserving genetic diversity. Doctor’s Dissertation, Swedish University of Agricultural Sciences Umeå, Sweden, Sverstria.


Communicating editor: J.-L. Jannink
APPENDIX

PROOF FOR THEOREM 1

This appendix proves theorem 1 through an example to provide a more insightful description for four different possibilities of recombination. Let’s assume we start with three pairs of breeding parents (S=6). We represent the genotypic information of these individuals with the following matrices:

```
pair one:   [ G_{1,1} G_{1,2} ]   [ G_{1,1} G_{1,2} ]
                   [ G_{2,1} G_{2,2} ]   [ G_{2,1} G_{2,2} ]
                   [ G_{L,1} G_{L,2} ]   [ G_{L,1} G_{L,2} ]

pair two:   [ G_{1,3} G_{1,2} ]   [ G_{1,4} G_{1,2} ]
                   [ G_{2,3} G_{2,2} ]   [ G_{2,4} G_{2,2} ]
                   [ G_{L,3} G_{L,2} ]   [ G_{L,4} G_{L,2} ]

pair three: [ G_{1,5} G_{1,2} ]   [ G_{1,6} G_{1,2} ]
                   [ G_{2,5} G_{2,2} ]   [ G_{2,6} G_{2,2} ]
                   [ G_{L,5} G_{L,2} ]   [ G_{L,6} G_{L,2} ]
```

The individuals in each pair are crossed to produce one progeny. The resulting progenies are then randomly mated for \( T - t - 1 \) generations. \( g \in \{ 0, 1 \}^L \) is the random gamete produced by breeding parents in meiosis of the \( (T - 1) \)st generation. From equation (12) we see that four possibilities exist for recombination. Here, we illustrate these four cases with color coding.

Proof: We divide the process into two phases: Phase 1: generation 0 until 2 and Phase 2: generation 2 until \( T \). Let \( h \in \{ 0, 1 \}^L \) denote the genotype of a specific gamete produced in meiosis by a progeny of a specific pair of breeding parents from the breeding population. This specific gamete contains the allele \( G_{i,n^e} \) that is passed on to the gamete \( g \) at locus \( l \), i.e., \( h_l = g_l = G_{i,n^e} \). We know that such a gamete uniquely exists because of the way the two phases are defined. The four cases are as follow:

Case 1: No recombination happens \( (g_2 \) comes from the same chromosome as \( g_1 \)).

```
[ g_1 ]   [ G_{1,1} G_{2,1} ]   [ g_1 ]
[ g_2 ]   [ G_{1,2} G_{2,2} ]   [ g_2 ]
[ g_L ]   [ G_{L,1} G_{L,2} ]   [ g_L ]
```

According to equation (12), when \( i^0 = i^1 \) and \( m^0 = m^1 \), we have:

\[
P(\overline{g_{l+1}} = G_{l+1,n^0}; g_l = G_{l,m^0}) = (1 - r_{l+1})^2(1 - R_{l+1})
\]

\[
\forall l \in \{1, \ldots, L-1\}, \forall i^0, i^1 \in \{2, \ldots, S\}, \forall m^0, m^1 \in \{1, 2\}.
\]

Using this definition equation (17) can be calculated as follow:

\[
P(\overline{g_{l+1}} = G_{l+1,n^0}; g_l = G_{l,m^0})
\]

\[
= P(\overline{g_{l+1}} = G_{l+1,n^0}; g_l = G_{l,m^0})
\]

\[
= P(h_{l+1} = G_{l+1,n^0}; g_l = G_{l,m^0})
\]

\[
= P(h_{l+1} = G_{l+1,n^0}; g_l = G_{l,m^0}; g_l = h_l)
\]

\[
= P(h_{l+1} = G_{l+1,n^0}; h_l = G_{l,n^0}; g_l = h_l) \cdot P(\overline{g_{l+1}} = h_{l+1}; g_l = G_{l,m^0}; g_l = h_l)
\]

\[
= P(h_{l+1} = G_{l+1,n^0}; h_l = G_{l,n^0}; P(\overline{g_{l+1}} = h_{l+1}; g_l = G_{l,m^0}; g_l = h_l)
\]

\[
= (1 - r_{l+1})^2(1 - R_{l+1})
\]

Equation (19) comes from the fact that \( i^0 = i^1 \) and \( m^0 = m^1 \). Equation (20) is derived from equation (19) because of the way \( h \) is defined. To find equation (21) from (20) independency is used. Finally, equation (22) is derived from (21) due to the fact that \( h_{l+1} = G_{l+1,n^0} \) is independent from \( g_l = h_l \) and also \( g_{l+1} = h_{l+1} \) is independent from \( h_l = G_{l,n^0} \).
Similarly, equation (31) can be calculated as follow:

\[ R_l = 1 - P(g_{l+1} = h_{l+1} | g_l = h_l) \]  

(24)

**Proof.**

\[ R^2_l = 0 \]

\[ R^3_l = 1 - (1 - R^2_l) (1 - r_l) + \frac{r_l}{S/2} \] \( \forall i \in \{3, 4, \ldots, \tau\} \)

Where \( r_l \) is the \( l^{th} \) recombination frequency for \( l \in \{1, 2, \ldots, L - 1\} \) and \( S \) is number of breeding parents. From the above equations we obtain:

\[ R_l = \frac{(S - 2) (1 - (1 - r_l)^{T - 1})}{S} \]  

(25)

This provides the proof for equation (13).

**Case 2:** Recombination happens within an individual (\( g_2 \) is coming from the other chromosome of the same individual where \( g_1 \) is coming from).

\[
\begin{bmatrix}
g_1 \\
g_2 \\
\vdots \\
g_l \\
\end{bmatrix} = \begin{bmatrix}
G_{1,1.1} \\
G_{2,2.1} \\
\vdots \\
G_{L,m,s} \\
\end{bmatrix}
\]

According to equation (12), when \( \beta = i^l \) and \( m^0 = m^1 \), we have:

\[ P(g_{l+1} = G_{l+1,m^0,m^1} | g_l = G_{l,m^0,m^1}) = r_{l+1}(1 - r_{l+1})(1 - R_{l+1}) \]  

(26)

\( \forall l \in \{1, \ldots, L - 1\}, \forall \beta, i^l \in \{1, 2, \ldots, S\}, \forall m^0, m^1 \in \{1, 2\} \).

Similarly, equation (26) can be calculated as follow:

\[ P(g_{l+1} = G_{l+1,m^0,m^1} | g_l = G_{l,m^0,m^1}) \]

(27)

\[ = P(h_{l+1} = G_{l+1,m^0,m^1}, g_{l+1} = h_{l+1} | h_l = G_{l,m^0,m^1}, g_l = h_l) \]  

(28)

\[ = P(h_{l+1} = G_{l+1,m^0,m^1}, h_l = G_{l,m^0,m^1})P(g_{l+1} = h_{l+1} | g_l = h_l) \]  

(29)

\[ = r_{l+1}(1 - r_{l+1})(1 - R_{l+1}) \]  

(30)

**Case 3:** Recombination happens within the paired individual.

\[
\begin{bmatrix}
g_1 \\
g_2 \\
\vdots \\
g_l \\
\end{bmatrix} = \begin{bmatrix}
G_{1,1.1} \\
G_{2,2.1} \\
\vdots \\
G_{L,m,s} \\
\end{bmatrix}
\]

or

\[
\begin{bmatrix}
g_1 \\
g_2 \\
\vdots \\
g_l \\
\end{bmatrix} = \begin{bmatrix}
G_{1,1.1} \\
G_{2,2.1} \\
\vdots \\
G_{L,m,s} \\
\end{bmatrix}
\]

According to equation (12), when \( \beta = \frac{\beta}{2} \), we have:

\[ P(g_{l+1} = G_{l+1,m^0,m^1} | g_l = G_{l,m^0,m^1}) = \frac{1}{2} r_{l+1}(1 - R_{l+1}) \]  

(31)

\( \forall l \in \{1, \ldots, L - 1\}, \forall \beta, i^l \in \{1, 2, \ldots, S\}, \forall m^0, m^1 \in \{1, 2\} \).

Similarly, equation (31) can be calculated as follow:

\[ P(g_{l+1} = G_{l+1,m^0,m^1} | g_l = G_{l,m^0,m^1}) \]

(32)

\[ = P(h_{l+1} = G_{l+1,m^0,m^1}, g_{l+1} = h_{l+1} | h_l = G_{l,m^0,m^1}, g_l = h_l) \]  

(33)

\[ = P(h_{l+1} = G_{l+1,m^0,m^1}, h_l = G_{l,m^0,m^1})P(g_{l+1} = h_{l+1} | g_l = h_l) \]  

(34)
\[ = \frac{1}{2} r_{t+1}(1 - R_{t+1}) \]  

Case 4: This case considers all possible remaining recombination.

\[
\begin{bmatrix}
g_1 \\ g_2 \\ \vdots \\ g_L 
\end{bmatrix} = \begin{bmatrix}
G_{1,1,1} \\ G_{2,1,3} \\ \vdots \\ G_{L,m,s} 
\end{bmatrix}, \text{ or } \begin{bmatrix}
g_1 \\ g_2 \\ \vdots \\ g_L 
\end{bmatrix} = \begin{bmatrix}
G_{1,1,1} \\ G_{2,2,3} \\ \vdots \\ G_{L,m,s} 
\end{bmatrix}, \text{ or }
\]

According to equation (12), when \( \frac{i}{2} \neq \frac{j}{2} \), we have:

\[ P\left( g_{i+1} = G_{i+1,m^i,j^i} \mid g_l = G_{l,m^l,j^l} \right) = R_{l+1} \frac{2}{2(S - 2)} \]  

\[ \forall l \in \{1, \ldots, L - 1\}, \text{ and } \forall \frac{i}{2}, \frac{j}{2} \in \{1, 2, \ldots, S\}, \forall m^0, m^1 \in \{1, 2\}. \] 

Similarly, equation (36) can be calculated as follow:

\[ P\left( g_{i+1} = G_{i+1,m^i,j^i} \mid g_l = G_{l,m^l,j^l} \right) \]

\[ = P\left( h_{i+1} = G_{i+1,m^i,j^i}, g_{i+1} = h_{i+1} \mid h_l = G_{l,m^l,j^l}, g_l = h_l \right) \]

\[ = P\left( h_{i+1} = G_{i+1,m^i,j^i} \mid h_l = G_{l,m^l,j^l} \right) P\left( g_{i+1} = h_{i+1} \mid g_l = h_l \right) \]

\[ = \frac{1}{4} \times \frac{R_{l+1}}{\frac{2}{2} - 1} \]

\[ = \frac{R_{l+1}}{2(S - 2)} \]