IOWA STATE UNIVERSITY
Digital Repository

Graduate Theses and Dissertations

Iowa State University Capstones, Theses and Dissertations

2016

Optimal population value selection: A population-based selection strategy for genomic selection

Matthew Daniel Goiffon
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Agriculture Commons, Genetics Commons, Industrial Engineering Commons, and the Plant Sciences Commons

Recommended Citation
Goiffon, Matthew Daniel, "Optimal population value selection: A population-based selection strategy for genomic selection" (2016). *Graduate Theses and Dissertations*. 15704.
https://lib.dr.iastate.edu/etd/15704

This Thesis is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

www.manaraa.com

Optimal population value selection: A population-based selection strategy for genomic selection

by

Matthew D. Goiffon

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Industrial Engineering

Program of Study Committee:
Lizhi Wang, Major Professor
Guiping Hu
Patrick Schnable

Iowa State University

Ames, Iowa

2016

Copyright © Matthew D. Goiffon, 2016. All rights reserved.

www.manaraa.com

TABLE OF CONTENTS

www.manaraa.com

www.manaraa.com

ACKNOWLEDGMENTS

The following thesis is the culmination of six years at Iowa State University. From my first course as an undergraduate until my final course as a Master's student, the time I have spent here has been more than I could have ever hoped or imagined.

While there are many professors and administrators that have helped along the way, I would specifically like to thank my major professor, Dr. Lizhi Wang, and my committee members, Dr. Guiping Hu and Dr. Patrick Schnable. In the previous months, the time you all took to meet with me was more than many are ever afforded. Beyond that, in these meetings you all demonstrated tremendous patience and provided me with invaluable feedback. For that, I thank you.

For providing data, textbooks, and references to relevant journal articles, as well as for attending meetings, providing feedback, and proofreading, I would like to thank Aaron Kusmec. I do not know how you managed to find time with your own dissertation work in progress, but I truly appreciate it. Additionally, I would like to thank Teri Craven for proofreading my manuscript and providing me with much needed encouragement throughout this process.

Most importantly, I would like to thank my parents. Here, I speak of this thesis as a result of the past six years at Iowa State University. However, in reality it would not have been possible without my upbringing. As the result of years of your dedication and hard work, you have instilled in me a mentality that nothing is beyond my reach. For that, I am truly grateful.

www.manaraa.com

ABSTRACT

In order to feed the world's growing population, an interdisciplinary effort is needed. In this thesis, operations research tools of mathematical modeling, optimization, and simulation are used to improve an existing plant breeding method, genomic selection. To do this, a new method, called optimal population value (OPV) selection, is proposed. In this paper, OPV selection is first defined as an optimization problem that selects a breeding population using a population metric, instead of individual metrics. Then, OPV selection is thoroughly tested in a simulation study against the existing methods of genomic selection, weighted genomic selection, and optimal haploid value selection. From the results of the simulation study, up to an 8.3%, or 0.58 base standard deviations, greater mean response can be expected than when using traditional genomic selection. These results suggest that population-based selection methods are a promising future research direction.

www.manaraa.com

CHAPTER I

GENERAL INTRODUCTION

Ensuring the global food supply for the next century will be an enormous challenge as the world's population continues to grow. By 2050, there will be a predicted nine billion people (Godfray, et al., 2012). In order to feed these people, total food production will need to increase amidst the obstacles of climate change, land scarcity, soil degradation, weeds, disease, and pests, while in an environment with less genetic diversity than in previous decades (The Royal Society of London, 2009).

One field working to feed the future world population is plant breeding. The National Association of Plant Breeders (2016) defines plant breeding as the process of improving plants by combining parent plants and selecting those progeny with the most potential to meet the population's needs. However, the problem of ensuring a sufficient food supply is complex and requires more than just improving cultivars. For example, in developing countries a significant proportion of the current food supply is lost due to poor supply chain infrastructure such as insufficient transportation, cold storage, and finance systems (Godfray, et al., 2012). In order to utilize improvements to food production, these logistical gaps will need to be filled. Additionally, future food production may be plagued by constraints of land, water, and energy (Leaver, 2011). As a result, in order to maximize food production, tradeoffs will need to be fully assessed. Since the scope of this problem goes beyond the role of traditional plant breeding, an interdisciplinary approach is needed.

www.manaraa.com

Operations research is one discipline that might be applicable when facing global food production constraints in the next century. Operations research uses a diverse set of problem solving tools, such as optimization, simulation, mathematical modeling, and statistics, in order to make better decisions and improve efficiency ("What is Operations Research?", 2016). With respect to impending food production problems, operations research can assist in designing transportation systems to minimize food waste or allocate land, water, and energy resources such that yield is maximized. Beyond these problems, operations research techniques can be useful in plant breeding specific problems. For example, operations research approaches have already been applied to gene stacking (Xu, Wang, & Beavis, 2011) and multi-allelic introgression (Han, Wang, Beavis, & Cameron, 2016).

In this thesis, operations research tools are applied to genomic selection (GS) in order to improve the mean response to selection. Chapter II provides a detailed literature review on GS. Within the literature review, three critical parts to GS are discussed: forming the training population, building a prediction model, and using the estimated marker effects for selection. In Chapter III, a journal paper discussing a new approach to GS is given. The paper proposes a new method called optimal population value (OPV) selection and defines OPV selection as an optimization problem. Rather than evaluating and selecting individuals to form a breeding population as in existing methods, sets of individuals are evaluated together and the best set of individuals is selected. To determine whether population-based selection strategies, such as OPV selection, can generate more response on average than individual-based selection strategies, a simulation study was

www.manaraa.com

performed using empirical data. Then, the results are described. Finally, general conclusions of the research and future work are provided in Chapter IV.

www.manaraa.com

CHAPTER II

LITERATURE REVIEW

Marker-assisted selection (MAS) aims to incorporate genotypic information into selection decisions (Lande & Thompson, 1990). In MAS, genetic markers that have strong statistical associations with quantitative trait loci (QTL), or loci controlling the trait of interest, are first identified using arbitrary significance thresholds. Then, marker effects, or the predicted impact of the marker on the trait of interest, are estimated for significant markers (Heffner, Sorrells, & Jannink, 2009). This two-step process results in a response limited by the amount of variance explained by the QTL detected in significance testing (Meuwissen & Goddard, 1996). However, since this two-step process ignores small effect markers deemed insignificant, only a fraction of the total variance will be explained (Goddard & Hayes, 2007).

GS attempts to address this limitation of MAS. In GS, genome-wide genetic markers and phenotypic observations from a training population are used to estimate marker effects (Meuwissen, Hayes, & Goddard, 2001). Instead of first identifying significant markers as in MAS, GS uses all markers to train the prediction model. By doing this, GS avoids one major pitfall of MAS, i.e. identifying QTL. As a result, GS can achieve a high prediction accuracy, or a strong correlation between the sum of marker effects, called genomic estimated breeding values (GEBVs), and the sum of QTL effects, called true breeding values (TBVs), of the validation population (Meuwissen, Hayes, & Goddard, 2001). This has had a profound impact on what is possible within breeding

www.manaraa.com

programs and has revolutionized animal breeding (Hayes, Bowman, Chamberlain, & Goddard, 2009).

As a result, the potential impact of GS in plant breeding is being widely discussed (Desta & Ortiz, 2014; Heffner, Sorrells, & Jannink, 2009; Jannink, Lorenz, & Iwata, 2010; Lorenz, et al., 2011). In particular, a number of simulation and empirical studies have been performed to investigate GS's ability to increase genetic gain in crops. One simulation study compared GS and Marker-Assisted Recurrent Selection (MARS) in a bi-parental maize breeding program. It showed that the response to GS was greater than that to MARS (Bernardo & Yu, 2007), where response is defined as R in the breeder's equation, $R = ir_A\sigma_A$, and i is selection intensity and is a function of the proportion of the population selected, r_A is prediction accuracy, and σ_A is the additive genetic standard deviation (Falconer, 1981). Similarly, in empirical studies on bi-parental (Heffner, Jannink, Iwata, Souza, & Sorrells, 2011) and multi-family (Heffner, Jannink, & Sorrells, 2011) wheat populations, GS resulted in greater prediction accuracy than MAS. From the breeder's equation previously defined, this greater relative accuracy reported should lead to an increase in response. While the findings of these studies show that GS can increase response in plant breeding, several other studies have additionally noted that GS allows for more breeding cycles per unit time. In these studies, this translated into a significant advantage for GS in response per unit time (Beyene, et al., 2015; Heffner, Lorenz, Jannink, & Sorrells, 2010; Wong & Bernardo, 2008).

Although GS has reportedly outperformed more traditional plant breeding methods such as MAS and MARS, successful implementation of GS depends on three conditions: (i) the training population must adequately reflect the test population, (ii)

www.manaraa.com

effects must be estimated accurately, and (iii) the estimated effects must be used in a way that predicts the reproductive merit of the plant or animal with respect to a breeding goal.

In general, for condition (i) to be satisfied the training population must be large, have a sufficient number of markers, and be closely related to the test population. In the earliest GS study, prediction accuracy was greater with a larger training population than with a smaller training population (Meuwissen, Hayes, & Goddard, 2001). Since then, the same observations were made in barley (Zhong, Dekkers, Fernando, & Jannink, 2009) and maize (Rincent, et al., 2012) datasets. Similarly, prediction accuracy in GS tends to improve with increasing marker density (Lorenzana & Bernardo, 2009) because high-density markers tend to be in sufficient linkage disequilibrium (LD) with QTL (Zhong, Dekkers, Fernando, & Jannink, 2009). However, if extensive LD already exists within the population, a relatively small training population and relatively few markers can be used without much reduction in prediction accuracy (Lorenz, Smith, & Jannink, 2012). Habier, Fernando, and Dekkers (2007) explained this phenomenon. There, accuracy was shown to be decomposable into accuracy from genetic relationships and accuracy from LD. Thus, extensive LD within a population is sufficient for relatively high prediction accuracy. Likewise, a high degree of coancestry is sufficient for relatively high prediction accuracy. Further work related to training populations has been completed to optimally select training sets (Isidro, et al., 2014; Rincent, et al., 2012) and to merge historical training sets (Asoro, Newell, Beavis, Scott, & Jannink, 2011; Muir, 2007; Rutkoski, et al., 2015) or training sets from separated populations (De Roos, Hayes, & Goddard, 2009; Lund, et al., 2011).

www.manaraa.com

To address condition (ii), a number of statistical methods have been proposed since the inception of GS. While many other models exist, such as, variations of Bayesian models as discussed in Kärkkäinen and Sillanpää (2012), LASSO methods (Ogutu, Schulz-Streeck, & Piepho, 2012), hybrid LASSO best linear unbiased prediction (BLUP) methods (Li, Wang, & Bao, 2015), and others (Desta & Ortiz, 2014; Gianola & Van Kaam, 2008; Solberg, Sonesson, Woolliams, & Meuwissen, 2009), some of the most common methods are ridge regression BLUP (RR-BLUP; Meuwissen, Hayes, & Goddard, 2001; Whittaker, Thompson, & Denham, 2000), genomic BLUP (VanRaden, 2008), and Bayes B (Meuwissen, Hayes, & Goddard, 2001). In RR-BLUP, each marker is assigned an effect from a normal distribution with the same variance. By using a statistical technique called ridge regression, these effects are shrunk towards zero to avoid collinearity, which can arise from having more explanatory variables than observations. In GBLUP, a genomic relationship matrix is used to predict marker effects. Instead of capturing expected relationships, as in pedigree models, GBLUP captures the realized relationships that occur due to the probabilistic inheritance of alleles (Meuwissen, Hayes, & Goddard, 2013), which is called Mendelian sampling. While the explanation of RR-BLUP and GBLUP models differ here, under a few basic assumptions they have been shown to be equivalent (Habier, Fernando, & Dekkers, 2007). In Bayes B, the constraint of equal marker variances is relaxed. Instead, Bayes B models marker effects as random draws from a normal distribution with a marker-dependent variance drawn from an inverted chi-squared distribution. Additionally, markers are given a probability, π, of receiving no effect (Meuwissen, Hayes, & Goddard, 2001).

www.manaraa.com

While much research has been devoted to conditions (i) and (ii), only three approaches to condition (iii) are known to exist. The typical way of handling condition (iii) is to perform truncation selection on GEBVs (Meuwissen, Hayes, & Goddard, 2001). This procedure is thought to maximize response in the following generation since the selection differential is maximized for a given proportion selected (Falconer, 1981). However, maximizing response of the next generation is not always optimal when trying to maximize long-term response due to premature fixation of alleles (Gibson, 1994). To prevent allele fixation, a method was proposed that weights rare and favorable alleles more than common or unfavorable allele (Goddard, 2009). To do this, markers are given a weight scaled according to allele frequencies (Hayes, Bowman, Chamberlain, & Goddard, 2009). In turn, weights are applied to either the sign of the estimated marker effect (Goddard, 2009) or the estimated marker effects in the GEBV calculations (Jannink, 2010). Truncation selection is then carried out on the weighted GEBVs. This selection method was called weighted genomic selection (WGS) and was tested in a simulation study. The results indicated that weighting markers sacrifices short-term gains somewhat, but quickly makes up for it in subsequent generations (Jannink, 2010). In another approach to condition (iii), truncation selection based on the optimal haploid values (OHVs) was proposed, where OHV is the selection limit of a line (Daetwyler, Hayden, Spangenberg, & Hayes, 2015). In this way, OHV selection combined ideas of an upper selection limit (Cole & VanRaden, 2011) and an ideal genotype (Kemper, Bowman, Pryce, Hayes, & Goddard, 2012). The OHV selection approach to condition (iii) was shown to increase gains and preserve diversity better than selection based on

www.manaraa.com

GEBVs in simulated wheat breeding program that used doubled haploids (Daetwyler, Hayden, Spangenberg, & Hayes, 2015).

Based on the success of WGS and OHV, it is clear that truncation selection of GEBVs is not optimal when considering time horizons longer than one generation. However, neither WGS nor OHV selection methods were proven optimal either. This suggests there are potentially better methods yet to be discovered. With the massive scale of global food production (FAO, 2015), these future discoveries could result in a significant return.

www.manaraa.com

CHAPTER III

SELECTION ON OPTIMAL POPULATION VALUE INCREASES GENETIC GAIN RELATIVE TO EXISTING GENOMIC SELECTION METHODS

Abstract

In the original genomic selection (GS) method, individuals are selected based on the sum of their estimated marker effects, known as genomic estimated breeding values (GEBVs). Due to significant correlation between GEBVs and true breeding values, this approach to GS has resulted in rapid genetic gain. Since then, however, optimal haploid value (OHV) selection and weighted genomic selection (WGS) have been proposed as extensions to the original GS method to facilitate efficient development of doubled haploids and to improve long-term response, respectively. In simulation studies, these methods were shown to separately outperform GS under different assumptions. However, further improvements exist. In this paper, optimal population value (OPV) selection is introduced as selection based on the maximum possible haploid value in a sub-set of the population. Instead of evaluating the breeding merit of individuals, as in GS, OHV selection, and WGS, the proposed method evaluates the breeding merit of a set of individuals together. After testing OPV selection thoroughly across two populations and under 15 parameter combinations, OPV selection was found to achieve up to 8.3%, or 0.58 base standard deviations, more response than GS. Additionally, it statistically outperformed both extensions to GS: WGS and OHV selection. These results suggest a new paradigm for selection methods in which an individual's value is dependent upon its compatibility with others.

www.manaraa.com

Introduction

Genomic selection (GS) was proposed as a method to capture effects of all quantitative trait loci (QTL; Meuwissen, Hayes, & Goddard, 2001). In GS, genome-wide genetic markers and phenotypic observations are used to estimate marker effects that can subsequently be used to accurately predict breeding values of individuals that have only been genotyped (Meuwissen, Hayes, & Goddard, 2001). As a result of this prediction accuracy, GS has been recognized as a potentially viable way to accurately select for cultivar improvement programs in plant breeding (Bernardo & Yu, 2007), as well as for allowing more breeding cycles per unit of time (Heffner, Lorenz, Jannink, & Sorrells, 2010).

Even though GS has been shown to accurately predict breeding values and has allowed for more breeding cycles per unit time, two extensions have been proposed improve it. The first, weighted genomic selection (WGS), was proposed to increase the frequency of rare favorable alleles in the population in order to maximize long-term response (Goddard, 2009). In a simulation study, WGS was shown to increase response after just a few generations (Jannink, 2010). In the second extension, the optimal haploid values (OHVs) of the individual were used for selection. This was proposed and shown to improve response in doubled haploid breeding programs (Daetwyler, Hayden, Spangenberg, & Hayes, 2015).

While GS and both extensions perform well, further improvements are possible. All three methods perform truncation selection on individual metrics. However, after several generations of random crossing and recombination, it is likely that contributions from many founder lines can be found in each line. Therefore, it is important that the

www.manaraa.com

selected lines in each generation are compatible with the other lines selected. This suggests an individual line's value is dependent on the other lines within the selected breeding population.

In this article, an extension of OHV selection (Daetwyler, Hayden, Spangenberg, & Hayes, 2015), optimal population value (OPV) selection is proposed. In OPV selection, instead of selecting the individuals with the greatest optimal haploid values, the sub-set of the population with the combined maximum haploid value is selected. To compare this proposed method with existing methods, OPV selection, GS (Meuwissen, Hayes, & Goddard, 2001), WGS (Jannink, 2010), and OHV selection (Daetwyler, Hayden, Spangenberg, & Hayes, 2015) are first defined mathematically. Then, a simulation study with empirical data from an inbred maize population is used to analyze the methods' relative ability to improve response, maintain a modified upper selection limit of the population (Cole & VanRaden, 2011), and maintain total additive genetic variance. The objectives of this paper are to (i) improve mean response and (ii) investigate the potential of population-based selection methods.

Selection Methods

In this section, four selection methods are described. To start, three existing selection methods are mathematically defined for convenience: GS (Meuwissen, Hayes, & Goddard, 2001), WGS (Jannink, 2010), and OHV selection (Daetwyler, Hayden, Spangenberg, & Hayes, 2015). Then, the proposed selection method, OPV selection, is defined as an extension of OHV selection. For this definition, an optimization formulation is used for clarity.

www.manaraa.com

While there may be some similarities between the formulas of existing methods and the formula of the proposed method, the selection processes differ tremendously. In each of the existing methods, truncation selection is performed on genomic estimated breeding values (GEBVs), weighted GEBVs, or OHVs, respectively. In OPV selection, sub-sets of the population are evaluated as units. After all possible sub-sets of the population have been evaluated, the unit with the best OPV is selected.

Genomic selection

In this selection method, the q lines with the largest GEBVs, where GEBV is defined in (1), are selected (Meuwissen, Hayes, & Goddard, 2001). In (1), the GEBV of line n is given by $\widehat{V_n}$, A_{lmn} is a genotype array with L marker loci, a ploidy of M, and N lines, and x_l is the estimated marker effect vector.

$$\widehat{V_n} = \sum_{lm} A_{lmn} x_l \quad (1)$$

Weighted genomic selection

At least two versions of WGS are known to exist (Goddard, 2009; Jannink, 2010). For this paper, q lines were selected based on the weighted GEBVs, V_n', calculated using (2). For these calculations, the estimated marker effect was weighted using the frequency of the most beneficial allele in the population at that locus, denoted w_l (Jannink, 2010). Additionally, A_{lmn} is a genotype array with L marker loci, a ploidy of M, and N lines, and x_l is the estimated marker effect vector.

$$\widehat{V_n'} = \sum_{lm} A_{lmn} x_l w_l^{-0.5} \quad (2)$$

www.manaraa.com

Since $w_l^{-0.5}$ is undefined when $w_l = 0$, a rule was set to assume $w_l = 1$ for these cases. It is important to note that when $w_l = 0$, every member in the population has the same allele. Thus, assuming $w_l = 1$ has an equal effect on all lines.

Optimal haploid value selection

For this selection method, the q lines with the largest OHV are selected (Daetwyler, Hayden, Spangenberg, & Hayes, 2015). OHV of line n was defined as

$$OHV_n = 2 * \sum_p \max_m(HV_{pmn}) \qquad (3)$$

where $HV_{pmn} = \sum_k h_{kmn} x_k \ \forall p, m, n,\ h_{kmn}$ is the genotype at locus k within the haplotype segment. Here, $k \epsilon \{1,2,..,K\}$ where K is the length of the haplotype segment, p is the number of haplotype segments in the genome, m is the ploidy, n is the line, and x_k is the estimated marker effect at locus k within haplotype segment p.

Optimal population value selection

As an extension to OHV selection (Daetwyler, Hayden, Spangenberg, & Hayes, 2015), OPV selection is introduced. For this method, the breeding population of size q that maximizes OPV, defined in (4), is selected. OPV can be interpreted as a generalization of upper selection limit (Cole & VanRaden, 2011) to varying haplotype lengths, rather than just haplotypes with a length of one marker. Since this generalization can differ from the upper selection limit described by Cole and VanRaden (2011) and incorporates the idea of varying haplotype lengths (Daetwyler, Hayden, Spangenberg, & Hayes, 2015), OPV will be used to describe this value. When upper selection limit is used in the rest of the paper, it will specifically be referring to (4) with a haplotype length of

www.manaraa.com

one marker only. Mathematically, OPV is defined as

$$OPV = 2 * \sum_{p} \max_{n}(\max_{m}(HV_{pmn})\ ) \quad (4)$$

where $HV_{pmn} = \sum_k h_{kmn}x_k\ \ \forall p, m, n,\ \ n$ is the index of the line, and $k,\ \ h_{kmn},\ \ p, m$, and x_k are defined as in OHV selection. An example comparing OHV selection and OPV selection is given in Figure 1.

In this selection method, OPV is maximized under two constraints. The first constraint limits the number of lines that can be selected for crossing. That is, no more than q lines can be used to maximize OPV. The second constraint used with this method restricts consideration of candidate breeding populations of size q to only those that have a GEBV greater than or equal to the mean GEBV of the current population. As an optimization problem, this can be interpreted as maximizing the overall potential effect of selecting a limited number of lines such that the response to selection is expected to be positive. This optimization problem is given in (5). In this optimization problem, $d_n\ \forall\ n$ are the binary indicator variables that determine if line n is selected ($d_n = 1$) or is not selected ($d_n = 0$).

$$\begin{aligned} \max \quad & 2 * \sum_{p} \max_{n}(d_n * \max_{m}(HV_{pmn})\ ) \\ s.t. \quad & \sum_{n} d_n \leq q \\ & \frac{\sum_{lmn} A_{lmn}x_l}{N} \leq \frac{\sum_{lmn} d_n * A_{lmn}x_l}{q} \\ & d_n \epsilon \{0,1\} \end{aligned} \quad (5)$$

From (5), it can be seen that diversity at a particular locus is beneficial in that the maximum allele effect is assumed. That is, loci with both favorable and unfavorable

www.manaraa.com

alleles assume the value of the favorable allele. However, this does not necessarily mean that allele fixation is penalized. In this method, fixed alleles are not penalized when loci are fixed for favorable alleles, but are penalized otherwise.

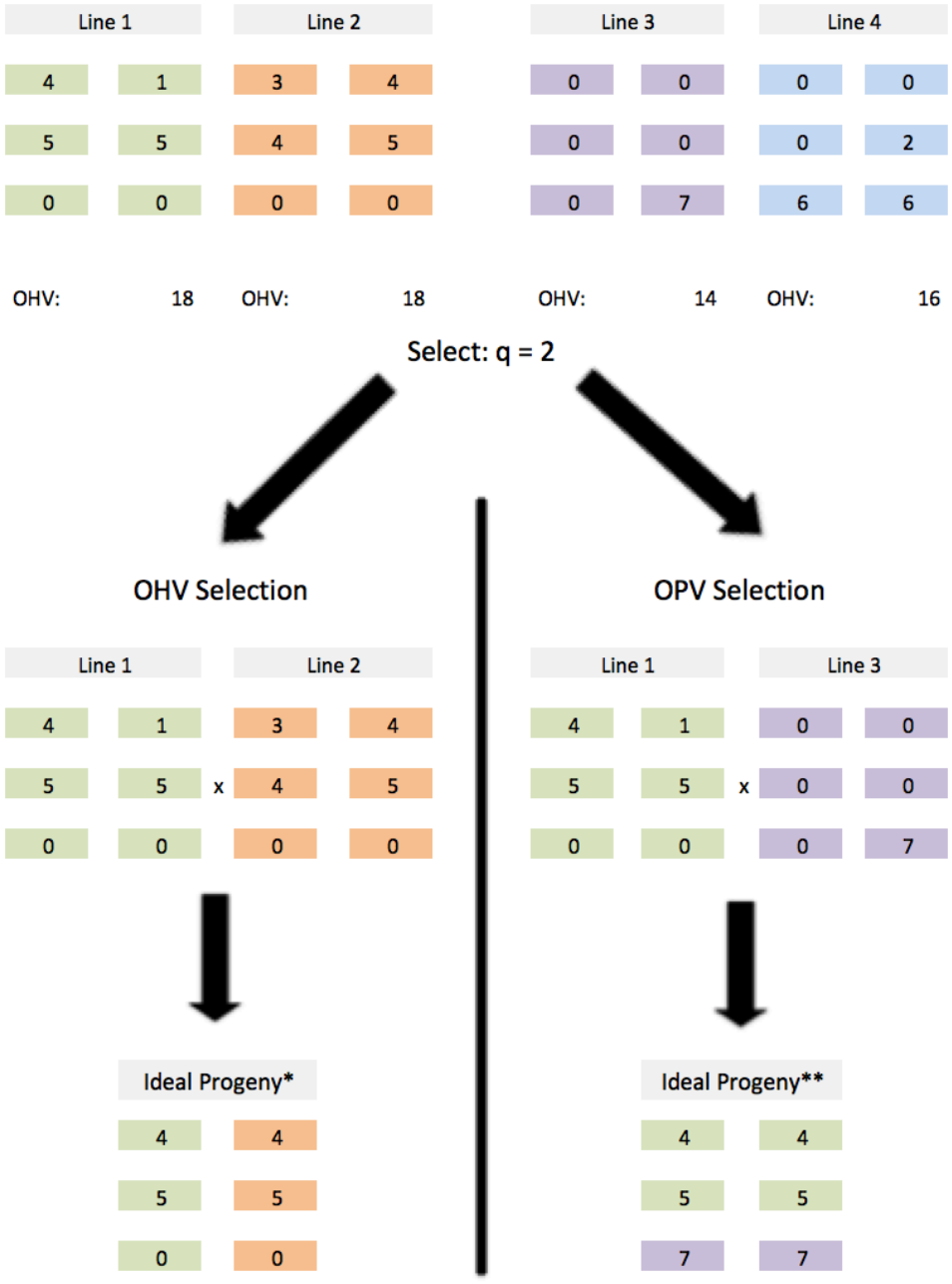

Figure 1: An example comparing OHV selection and a simplified version of OPV selection in which a filter is not added. In each colored box, the haplotype value is given. Ideal progeny* indicates a progeny that is possible after one generation. Ideal progeny** indicates a progeny that is possible after three generations.

www.manaraa.com

Simulation

In silico breeding process model

The plant breeding process considered in this paper is defined as a repetitive process composed of an evaluation, selection, pairing, and reproduction step each cycle and is shown in Figure 2. The process is initialized by inputting an $Initial\ Population$ of size N and setting $t = 0$. Then, the process starts with the evaluation of the current state of the population. Once the current state has been evaluated, q lines are selected with a selection method. After selection, lines are paired randomly. Finally, paired lines are crossed to produce a new population of size N. This population is then considered to be the $Initial\ Population$ of generation $t = t + 1$. This process is repeated until the final generation, $maxt$, has been completed. Doubled haploid production is not considered. In order to compare GS, WGS, OHV selection, and OPV selection, this breeding process was

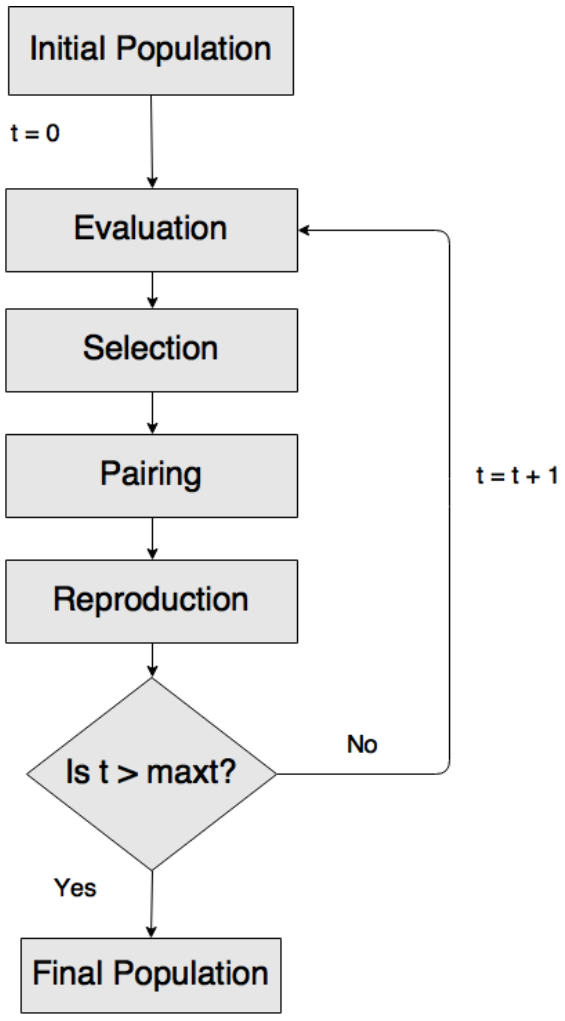

Figure 2: A diagram of the basic breeding process considered in this paper. For this process, no doubled haploid production is considered.

www.manaraa.com

implemented *in silico* in MATLAB. Pseudo-code of the simulation framework is provided in Appendix A.

In each generation of the simulation, the current state of the population was evaluated first. For the evaluation of response, an additive model was assumed to accurately depict the true phenotype of a line (Hill, Goddard, & Visscher, 2008). Furthermore, the fact that the marker effects are estimates of QTL was ignored. Instead, it was assumed that the marker effects that were estimated are the true effect of having that marker. The resulting additive model used is given in (6), where the true breeding value (TBV) of a given line n is V_n and $A_{lmn}$is the genotype array of the population indexed by loci number l, gamete m, and line n. The major allele effect vector is denoted x_l. The constant c can be interpreted as the population mean and ϵ_n is a random error term. For these simulations, an error term of $\epsilon_n = 0$ was used.

$$V_n = c + \sum_{lm} A_{lmn} x_l + \epsilon_n \quad \forall\, n \qquad (6)$$

The rationale for ignoring QTL effects and considering marker effects as true effects is as follows. Since our goal is to improve how estimated marker effects are used, assuming marker effects as true effects allows for the evaluation of the selection methods independent of marker prediction models. If QTL determined TBVs, the evaluations of the selection methods would be subject to Type I and II errors if markers had been estimated inaccurately. Therefore, it is assumed that the lines with the largest sum of estimated marker effects are the best lines and that the selection method that achieves a population with the best lines evaluated in this way is best. This is equivalent to saying a selection method is best if it achieves the best GEBVs after selection. Additionally, for all simulation studies an error term of $\epsilon_n = 0$ was used rather than an error term scaled to

www.manaraa.com

some heritability level. Any other error term would incorporate noise into evaluation of the selection methods. The implementation of (6) is provided in Appendix B.

For these simulations, the response was calculated as the change in mean TBV, where mean TBV in any generation is given as $\sum_n V_n/N$. At times, response per base standard deviation is used. For this, base standard deviation was defined as the standard deviation of TBVs in the initial population. Additionally, in each generation the modified upper selection limit and total additive genetic variance (Falconer, 1981) were evaluated, where the modified upper selection limit was calculated as in (4) with a haplotype segment length set to one marker and where total additive genetic variance was defined as

$$TotVariance = \sum_l 2 * f_l * (1 - f_l) * x_l^2 \quad (7)$$

with f_l defined as the minor allele frequency at loci l and x_l is the effect. The implementations of the upper selection limit and total additive genetic variance evaluation methods are provided in Appendices C and D, respectively.

Following evaluation, selection and paring occur. In the selection step, GS, WGS, OHV selection, or OPV selection were used. For the implementations of these methods it is important to note that although GS, WGS, and OHV selection are computationally simple, solving (5) for OPV selection is difficult. Instead of solving the optimization for OPV selection, a heuristic method was used. The basic heuristic used was to make repeated pairwise swaps of members in the current candidate breeding population with those not in the current candidate breeding population. If a swap improves the OPV of the candidate breeding population, then the swap is kept. Otherwise, it is discarded and the current candidate breeding population is set back to its previous state. This process

www.manaraa.com

continues until no single member of the breeding population can be substituted out. To guarantee the positive genetic gain constraint in (5), a GEBV filter was used prior to OPV selection. This filter restricted the lines considered for selection to some top percentage of lines. This top percentage GEBV filter was treated as a parameter for the simulation studies. The pseudo-code for GS, WGS, OHV selection, and OPV selection implementations is given in Appendices E, F, G and H, respectively. After the breeding population is selected the selected lines are paired randomly. The implementation of random pairing is given in Appendix I.

In order to replicate the true reproduction process, each offspring was determined probabilistically according to a recombination rate vector r. This vector is indexed by locus number l, where element r_l is defined as the probability of a recombination event occurring between locus $l-1$ and locus l. In other words, for some gamete G indexed by l, $r_l = \Pr\left(G_l = A_{l1n} \middle| A_{(l-1)2n}\right) = \Pr\left(G_l = A_{l2n} \middle| A_{(l-1)1n}\right)$. A special case of this is defined for the first locus of each chromosome, for which it is assumed $r_1 = 0.5 = \Pr(G_1 = A_{11n}) = \Pr(G_1 = A_{12n})$. The details of this implementation are given in Appendix J.

Data

For this paper, the 369 inbred maize (Zea mays subspecies mays L.) lines studied in Leiboff et al. (2015) were genotyped using RNA-Seq (Barbazuk, Emrich, Chen, Li, & Schnable, 2007) and tGBS (Schnable, Liu, & Wu, 2013), merged with genotyping by sequencing single nucleotide polymorphisms (SNPs) from Romay et al. (2013), and phased using Beagle (Browning & Browning, 2008). For each of the 369 lines, these

www.manaraa.com

results yielded SNPs at approximately 1.4 million loci spread across ten chromosomes with a total length of approximately 15.5 Morgans. For efficiency reasons, data from only 300 of the lines were selected for further processing. To prepare remaining data for simulation, all of the chromosomes from a line were concatenated along their first dimension to create an approximately 1.4 million x 2 matrix. Each cell in that matrix was then assigned a value of either "1" or "0" for having the major or minor allele, respectively. All lines were then concatenated together along their third dimension to create an approximately 1.4 million x 2 x 300 array, which will be referred to as A. From A, two random samples of 25 starting genotypes were sampled without replacement from the population of 300 genotypes. For each sample, every line was crossed with every other line once. For each of these crosses, only one offspring was generated. This produced two populations of 300 individuals. These populations will be called Initial Population 1 and Initial Population 2 and will be denoted $A^1$and A^2, respectively.

The phenotype data used for this study was retrieved from Leiboff et al. (2015) and consisted of 369 shoot apical meristem volume phenotypes. These phenotypes, along with the corresponding genotypes from above, were used to estimate marker effects using the Bayes B model (Meuwissen, Hayes, & Goddard, 2001) implemented in GenSel (Fernando & Garrick, 2009). This produced a vector of estimated marker effects for major alleles at each of the approximately 1.4 million SNPs, which will be referred to as x.

To estimate the recombination rates used in this study, the maize nested association mapping (NAM; Yu, Holland, McMullen, & Buckler, 2008) population was used as a starting point. Out of the 1,144 genetic markers in the NAM population, 133

www.manaraa.com

were removed because the orderings between physical and genetic positions were inconsistent. The remaining 1,011 markers were used to estimate the genetic positions of the SNPs by linear interpolation between the flanking NAM markers of each SNP. Once the genetic positions of the SNPs were estimated, recombination rates were calculated for each chromosome using Haldane's mapping function. For simulation purposes, a probability of 0.5 was assigned at the beginning of each chromosome's recombination rate vector. Then, all recombination rate vectors were concatenated along the first dimension to form r, an approximately 1.4 million x 1 vector.

Experimental setup

To test the proposed method against existing selection methods, a full factorial experiment was performed with two factors: Initial Population and Selection Method. For the Initial Population factor, the populations $A^1$and A^2 were treated as levels. For the Selection Method factor, 22 different methods were defined and are given in Figure 3. These methods are comprised of on e GS, one WGS, five OHV selection, and 15 OPV

Model Number	Selection Method Name	Selection Method Type	Filter Percentage	Number of Haplotype Segments (per chromosome)
1	GS	Genomic Selection	-	-
2	WGS	Weighted Genomic Selection	-	-
3	OHV CHR	OHV Selection	-	1
4	OHV 2/Chr	OHV Selection	-	2
5	OHV 3/Chr	OHV Selection	-	3
6	OHV 6/Chr	OHV Selection	-	6
7	OHV 12/Chr	OHV Selection	-	12
8	OPV 50 Chr	OPV Selection	50	1
9	OPV 50 2/Chr	OPV Selection	50	2
10	OPV 50 3/Chr	OPV Selection	50	3
11	OPV 50 6/Chr	OPV Selection	50	6
12	OPV 50 12/Chr	OPV Selection	50	12
13	OPV 30 Chr	OPV Selection	30	1
14	OPV 30 2/Chr	OPV Selection	30	2
15	OPV 30 3/Chr	OPV Selection	30	3
16	OPV 30 6/Chr	OPV Selection	30	6
17	OPV 30 12/Chr	OPV Selection	30	12
18	OPV 10 Chr	OPV Selection	10	1
19	OPV 10 2/Chr	OPV Selection	10	2
20	OPV 10 3/Chr	OPV Selection	10	3
21	OPV 10 6/Chr	OPV Selection	10	6
22	OPV 10 12/Chr	OPV Selection	10	12

Figure 3: Descriptions of the selection methods used in the full factorial experiment.

www.manaraa.com

selection methods. For each of these selection methods, 10% of the population is selected. In OHV selection methods and OPV selection methods, the haplotype length parameter can be varied. For this experiment, the haplotype lengths were varied between one chromosome, 1/2 chromosome, 1/3 chromosome, 1/6 chromosome, and 1/12 chromosome as in the research on OHV selection (Daetwyler, Hayden, Spangenberg, & Hayes, 2015). Furthermore, in OPV selection the filter percentage parameter was varied between 10, 30, and 50%. Here, the filter percentage is the top percentage of the population, with respect to GEBVs, to consider for OPV selection. Filters greater than 50% were not considered since no more than 50% of the population is needed to maximize long-term response (Cockerham & Burrows, 1980).

For each experiment, the number of lines selected each generation and population size were held constant at 30 and 300, respectively. Additionally, the constant c was set in the first generation such that the mean TBV of the population was zero. This full factorial of 44 experiments was run for 10 generations and for 60 complete replications.

Data analysis

The mean responses of the 22 different methods were compared using Tukey's Honestly Significant Difference Procedure for multiple comparisons at a significance threshold of 0.05. All significance testing was completed in Matlab (2015a) with the functions "anovan" and "multcompare" from the Statistics and Machine Learning Toolbox. Then, plots were generated to qualitatively compare upper selection limits and total additive genetic variance, as well as to compare trends when using OPV selection at various haplotype lengths and filter percentages.

www.manaraa.com

Results

Selection method comparison

In this simulation study 22 selection methods were compared for their ability to generate genetic gain, maintain the upper selection limit of the population, and maintain total additive genetic variance.

In Figure 4 the graphical results simulation study are given with respect to mean response. From this bar graph, it can be seen that when starting with Initial Population 1 the greatest seven responses are the result of OPV selection methods. The best of these seven methods is OPV 30 2/Chr, followed by OPV 30 3/Chr, OPV 50 3/Chr, OPV 50 2/Chr, OPV 30 6/Chr, OPV 30 Chr, and OPV 50 Chr, respectively. The best method that

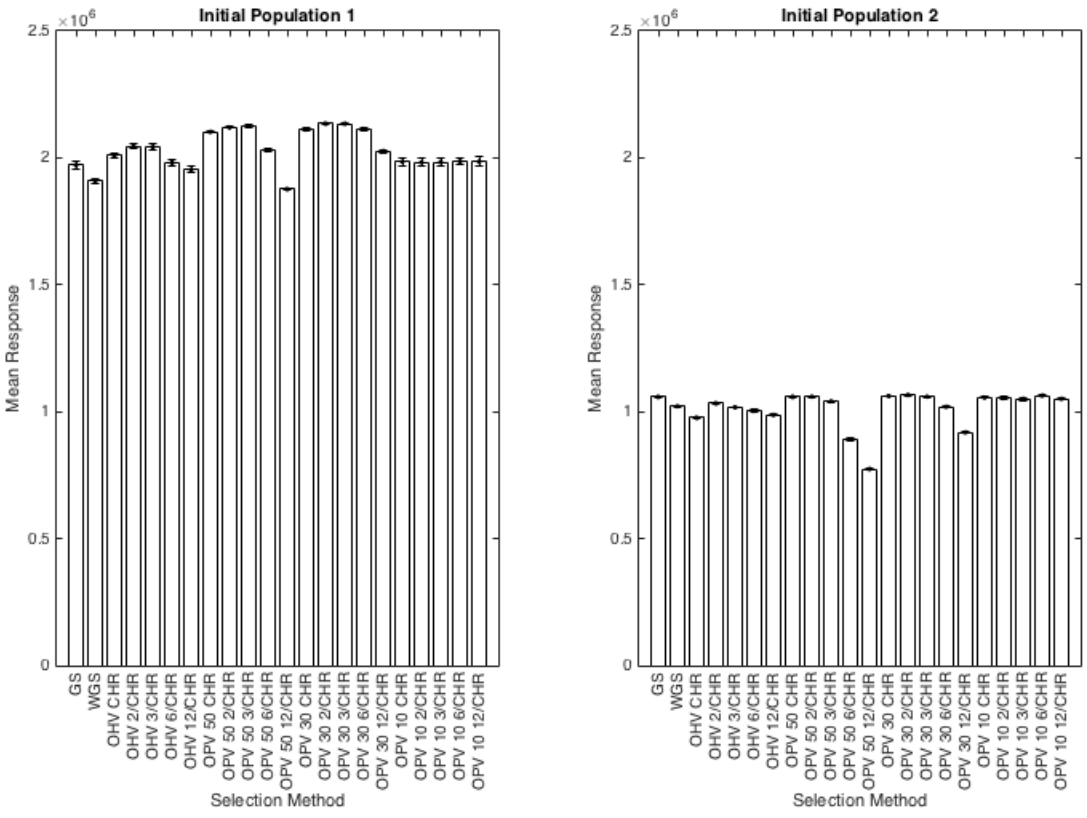

Figure 4: Comparison plots of total mean response and standard error of mean response for all 22 selection methods after 10 generations. On the left, results are given for Initial Population 1. On the right, results are given for Initial Population 2.

www.manaraa.com

was not in the family of OPV selection methods was OHV 2/Chr. When multiple comparison procedures were used to test for significant differences in mean responses, each of the top seven methods are statistically different than the bottom 16 methods for Initial Population 1. Notably, there is a total of 8.3%, or 0.58 base standard deviations, improvement in mean response when comparing OPV 30 2/Chr with GS and a 4.4%, or 0.32 base standard deviations, improvement in mean response when comparing OPV 30 2/Chr with OHV 2/Chr. Similar to these results, the top six selection methods when using Initial Population 2 were OPV selection methods. Listed from best to worst, these were OPV 30 2/Chr, OPV 10 6/Chr, OPV 30 Chr, OPV 50 2/Chr, OPV 30 3/Chr, and OPV 50 Chr. For this starting population, the best method that was not in the family of OPV selection methods was GS. In terms of percent improvement, the best OPV method for Initial Population 2 only out performs GS by 0.7%, or 0.03 base standard deviations, and outperforms the best OHV method by 3.2%, or 0.12 base standard deviations. However, it is important to note that even though the results of Initial Population 2 has six OPV selection methods performing the best, none of these had statistically different results than GS. This point is emphasized by the inclusion of OPV 10 6/Chr in the top six, which is equivalent to GS because of the 10% GEBV filter. When comparing OHV selection and GS, improved mean responses were observed while using OHV selection with one, two, three, and six haplotype segments per chromosome and Initial Population 1. However, this outcome was not observed in Initial Population 2. In Initial Population 2, all OHV selection methods performed statistically worse than GS. WGS performed worse than GS with both Initial Population 1 and Initial Population 2. A table of mean responses

www.manaraa.com

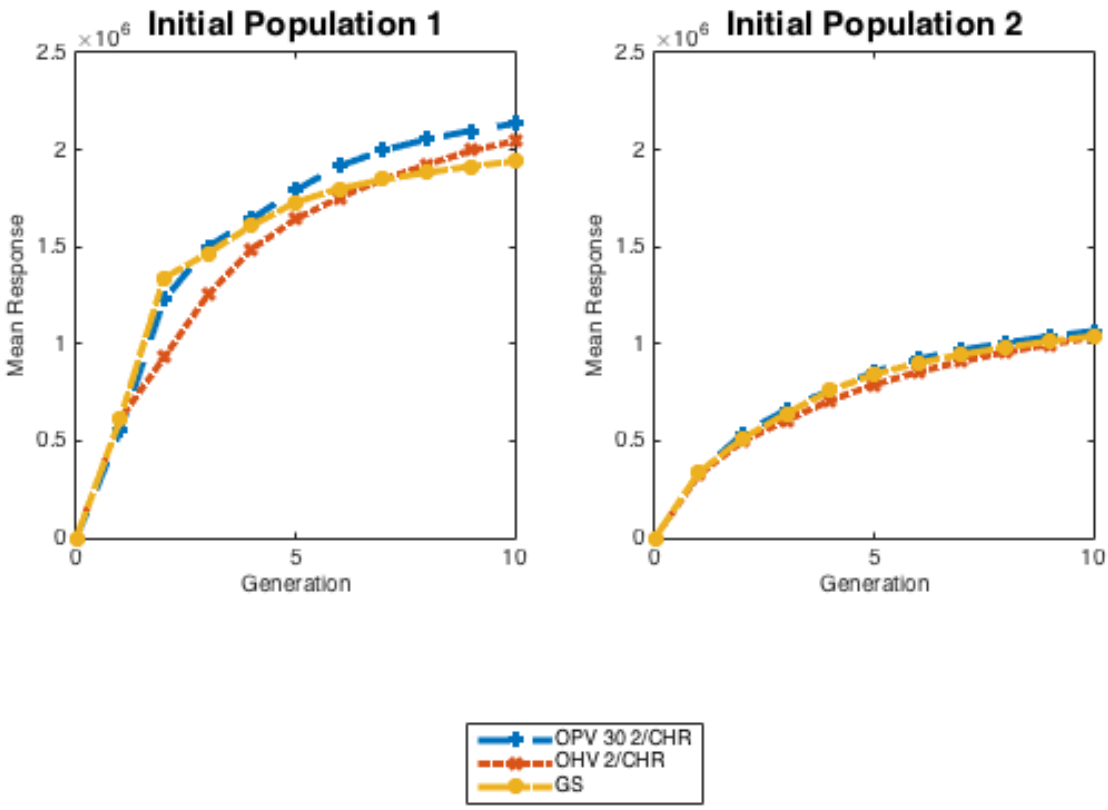

Figure 5: Comparison plots of mean response at different time horizons. On the top, results are given for Initial Population 1. On the bottom, results are given for Initial Population 2.

and standard errors of each selection method with each initial population is provided in Appendix K.

To demonstrate how the overall best performing OPV selection method compares to GS and the overall best performing OHV selection method prior to generation 10, a plot of the overall mean responses is given in Figure 5. In the plot of Initial Population 1, OHV2/Chr and GS perform approximately the same in the first generation, with OPV 30 2/Chr performing worse. By generation 3, OPV 30 2/Chr overtakes both OHV 2/Chr and GS. This advantage in mean response is maintained through generation 10. However, in generation 7 OHV 2/Chr surpasses GS and rapidly increases genetic gain. By generation 10, OHV 2/Chr is approaching OPV 30 2/Chr. When the initial population was set to

www.manaraa.com

Initial Population 2, all three methods perform approximately the same through the first generation. In the second generation, OPV 30 2/Chr generated a small advantage that was maintained through generation 10. In generation 10, OHV 2/Chr nearly reaches the response level of GS. A table of overall mean responses in each generation and with each initial population is provided in Appendix L for OPV 30 2/Chr, OHV 2/Chr, and GS.

With respect to total additive genetic variance, OHV 2/Chr outperformed both OPV 30 2/Chr and GS, as shown in Figure 6. In this figure, OHV 2/Chr total additive genetic variance curve peaks later and higher than the other two selection methods. Additionally, OHV 2/Chr maintains its total additive genetic variance longer. When considering the other two methods, OPV 30 2/Chr tends to decline more slowly after peaking than GS, although their peaks are at similar total additive genetic variance levels.

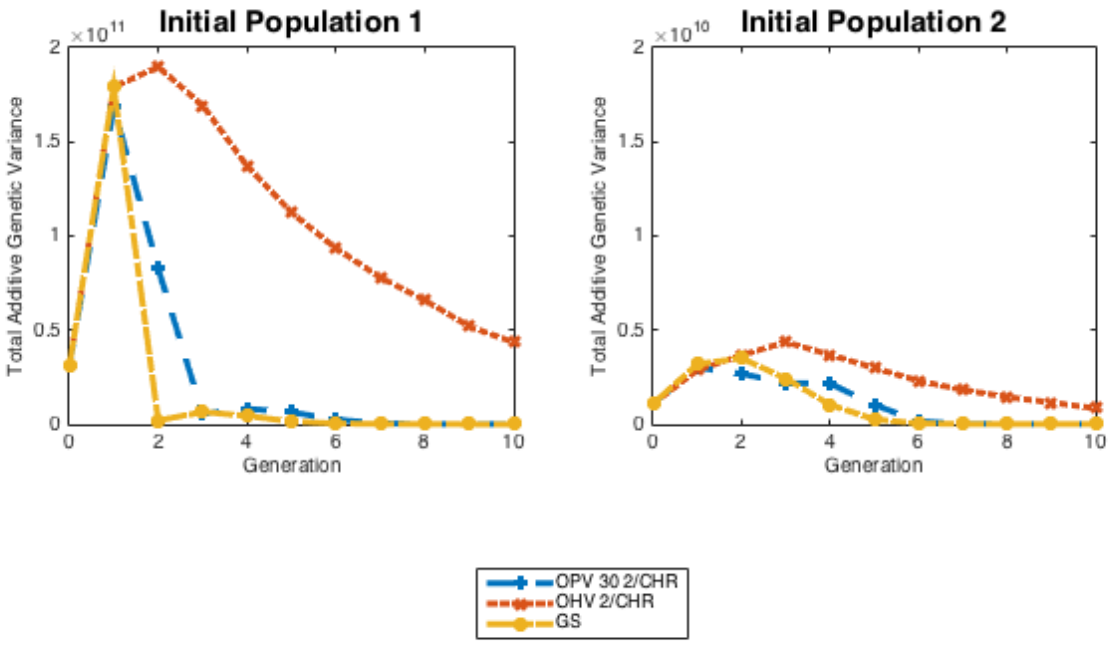

Figure 6: Comparison plots of total additive genetic variance at different time horizons. On the left, results are given for Initial Population 1. On the right, results are given for Initial Population 2. Note the scales of the axes are different for Initial Population 1 and Initial Population 2.

www.manaraa.com

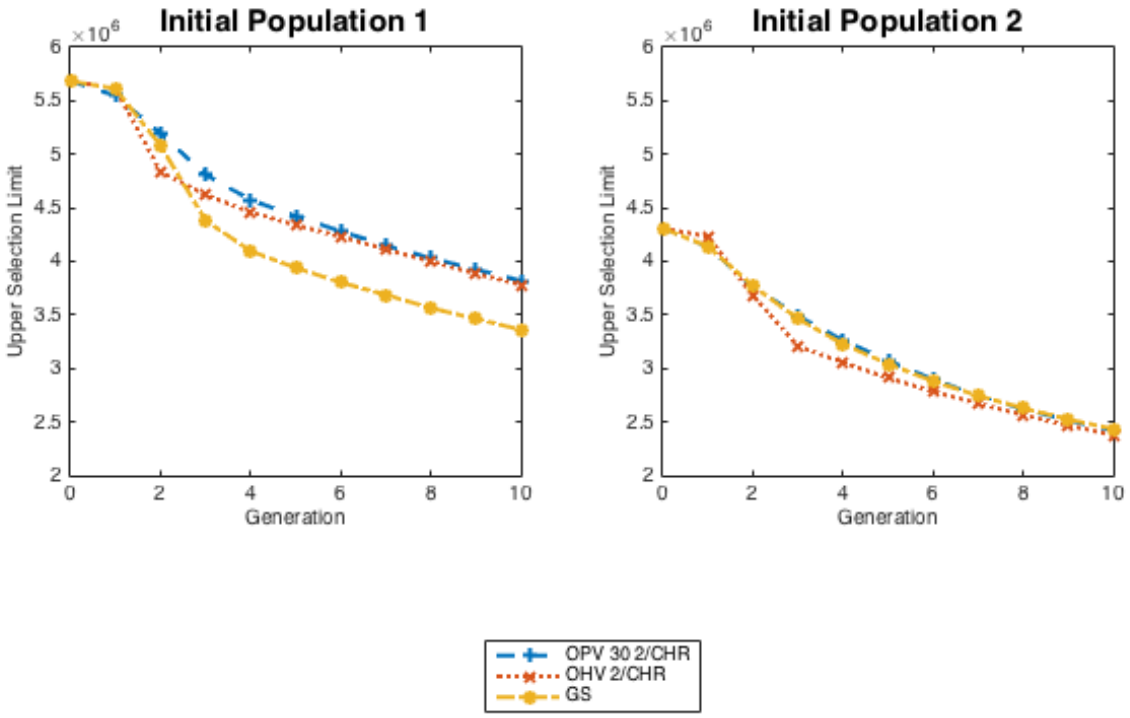

Figure 7: Comparison plots of upper selection limits at different time horizons. On the left, results are given for Initial Population 1. On the right, results are given for Initial Population 2.

In terms of maintaining the population's upper selection limit, these three methods performed similarly. However, on average across both initial populations and all time periods OPV 30 2/Chr appeared to perform slightly better than OHV 2/Chr and GS. A plot of upper selection limit curves with respect to generation is given in Figure 7.

Trends in OPV selection methods

In this section, comparison plots of mean response, upper selection limit, and total additive genetic variance are given to clearly demonstrate the trends that are common to all OPV selection methods. Since these trends are common to all methods, demonstration on only one OPV selection method suffices to show these trends. The methods plotted in this section were specifically chosen for illustration purposes.

www.manaraa.com

In Figure 8, a comparison plot of varying filter percentages is given for Initial Population 1 and 3 haplotype segments per chromosome. In this plot, as filter percentage decreases the rate of short-term response increases, there is a greater reduction in upper selection limit, and total additive genetic variance is maintained for fewer generations. This trend was observed for all quantities of haplotype segments per chromosome, as well as for Initial Population 2.

When a plot of varying haplotype segments was made for OPV selection with a 50% filter and Initial Population 1, a similar trend was observed. In this plot, as the number of haplotype segments per chromosome decreases the rate of short-term response increases, there is a greater reduction in upper selection limit, and total additive genetic variance is maintained for fewer generations. This plot is show in Figure 9 and is similar to plots 50 and 30% filters. With a 10% filter, changes in the number of haplotypes had no effect since the selection percentage is also 10%.

www.manaraa.com

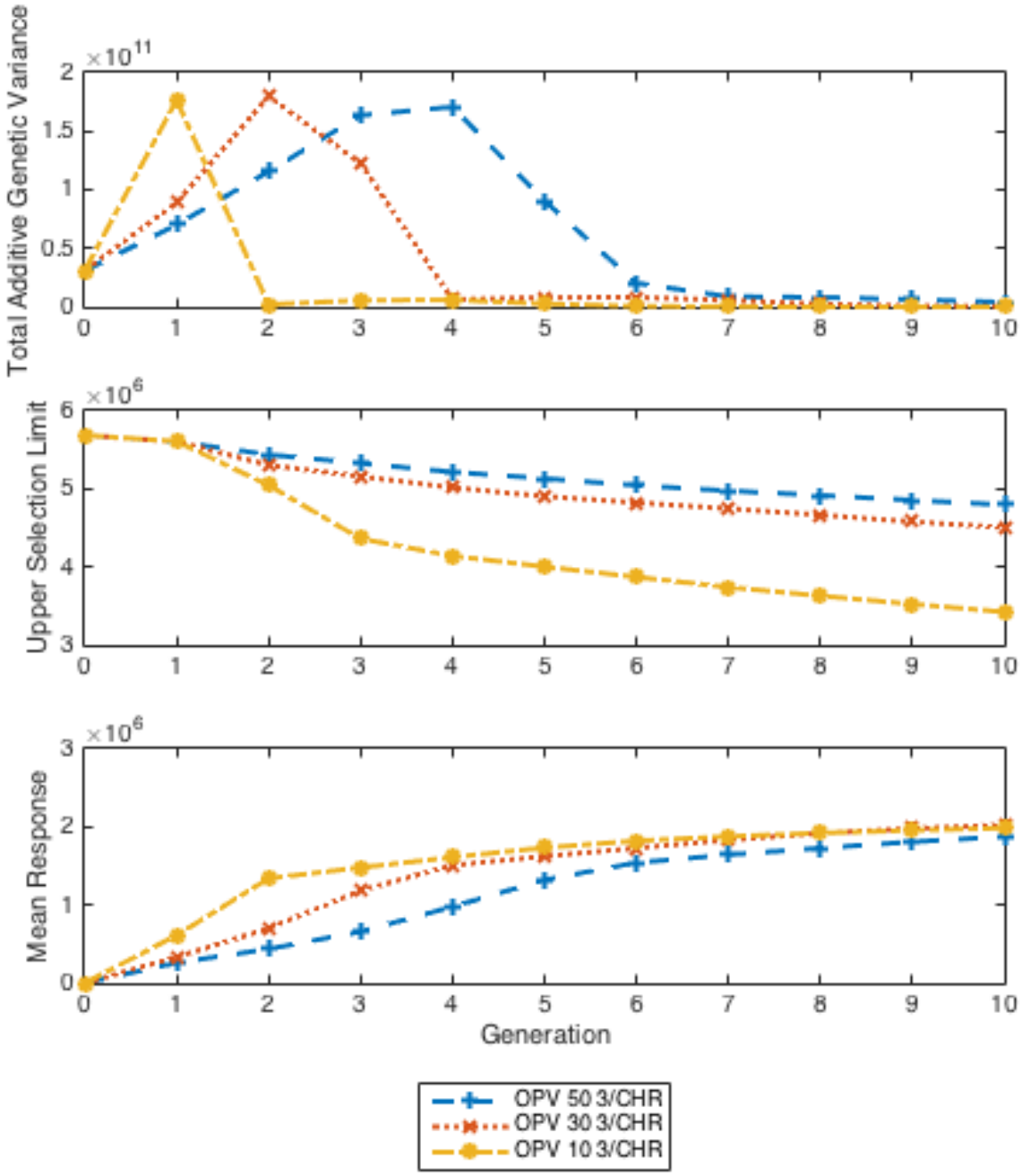

Figure 8: Comparison plots of 50, 30, and 10% filters in OPV selection with three haplotype segments per chromosome and Initial Population 1. On the top, comparisons of total additive genetic variance are made. In the middle, comparisons of upper selection limit are made. On the bottom, comparisons of mean response are made.

www.manaraa.com

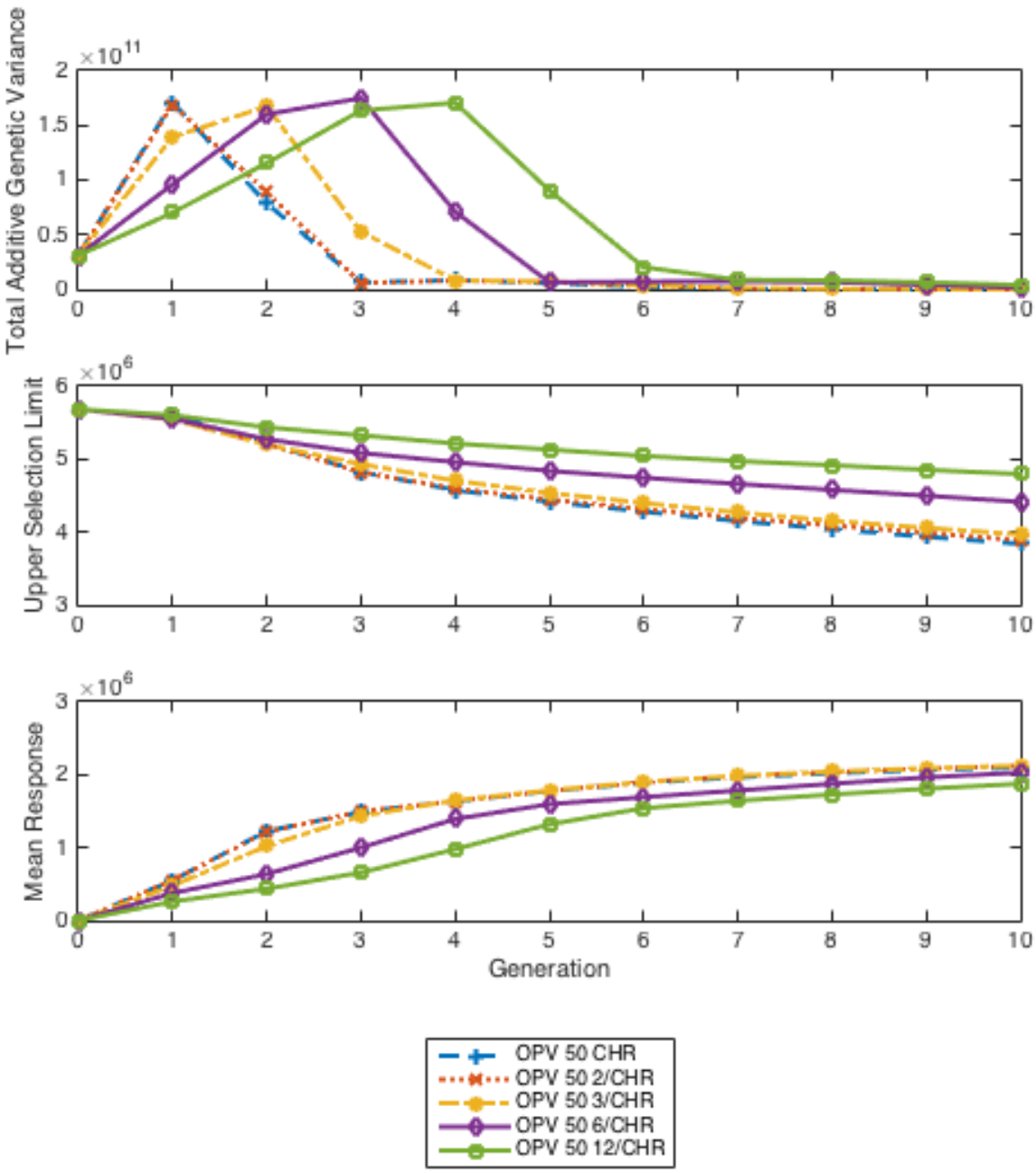

Figure 9: Comparison plots of 1, 2, 3, 6, and 12 haplotype segments per chromosome in OPV selection with a 50% filter and Initial Population 1. On the top, comparisons of total additive genetic variance are made. In the middle, comparisons of upper selection limit are made. On the bottom, comparisons of mean response are made.

www.manaraa.com

Discussion

Selection method comparison

From the results presented, it is clear that OPV 30 2/Chr generates the greatest mean response through 10 generations. However, OPV 30 2/Chr did not preserve the total additive genetic variance better than OHV 2/Chr, nor did OPV 30 2/Chr perform notably better than OHV 2/Chr or GS in terms of maintaining the upper selection limit. For this reason, it is not suspected that either of these metrics individually are the key driving force behind OPV 30 2/Chr's success. Instead, a subtler difference is suspected: OPV selection methods treat the population as a collaborative unit. In each generation, lines with a highly desirable haplotype segment are selected, paired randomly, and crossed. Since the most desirable haplotype segments are maintained each generation and randomly crossed, eventually they will be propagated throughout the population. This will inevitably result in recombinants with multiple selected haplotype segments because selected haplotype segments are from distinct non-overlapping sections of the genome. However, this has some limitations. For instance, in each generation of this simulation only 30 lines were selected. That means, if optimal haplotype segments were all in separate lines, then, at most, the best 30 haplotype segments could be selected. This implies that only selection of at least as many lines as haplotype segments considered can guarantee the optimal haplotype at each segment is selected.

In these simulations, the performance of the OHV selection method was inconsistent. When Initial Population 1 was used, OHV selection with one, two, three, and six haplotype segments outperformed GS with a maximum of +3.7% difference in genetic gain. This result closely matches a previous study in which OHV selection outperformed GS with less than or equal to 3 haplotype segments per chromosome. In

www.manaraa.com

that study, a +3% difference in genetic gain was observed between OHV selection and GS (Daetwyler, Hayden, Spangenberg, & Hayes, 2015). It is important to note again that for these simulations doubled haploid production was not considered, which suggests doubled haploid production may not be necessary to benefit from OPV selection. However, in this study a second starting population was used to test the robustness of all selection methods. With Initial Population 2, it was found that GS outperformed all OHV selection methods. While doubled haploid production is expected to have a minor positive impact on the difference between OHV and GS (Daetwyler, Hayden, Spangenberg, & Hayes, 2015), it is not expected to make up the large difference observed in this experiment. Additionally, OHV with two, three, six, and twelve haplotype segments outperformed OHV with one haplotype segment. This suggests that relatively significant genetic gain is only achieved by accumulating multiple generations of recombination (Daetwyler, Hayden, Spangenberg, & Hayes, 2015). This hypothesis is supported by the relatively low total additive genetic variance of Initial Population 2 compared to Initial Population 1, when the upper selection limits of the two populations are similar. This point indicates that beneficial alleles exist within the population, but are at low frequencies thus potentially require more recombination to see a notable response.

WGS was also considered in this simulation study. The mean response results did not suggest any clear benefit to WGS for this data set, which varies from previously published research (Jannink, 2010). However, in this study three orders of magnitude more markers were used than in the previous study, which could lead to a cumulative effect of many weighted small effect loci drowning out the signal from large effect loci.

www.manaraa.com

Since OPV selection optimizes OPV, which is closely related to the upper selection limit, it is not surprising that OPV 30 2/Chr maintained the upper selection limit better, on average, than OHV 2/Chr and GS. However, this difference was minimal due to the relatively long haplotype segments and low filter percentage. Also, because of the relatively long haplotype segments GS and OHV 2/Chr had, at times, a better upper selection limit. This happens, by chance, when haplotype segments are long, since the optimization of OPV differs from optimization of the upper selection limit.

However, a greater upper selection limit does not perfectly correlate with more variation within the population. This was demonstrated when OHV 2/Chr resulted in significantly more total additive genetic variance than OPV 30 2/Chr and GS. As explained in Daetwyler, Hayden, Spangenberg, and Hayes (2015), OHV selects based on the sum of favorable haplotype segments in a line without incentivizing multiple beneficial haplotypes at a single segment. Therefore, there is not a selection pressure towards homozygosity, which allows for better maintenance of diversity. In a similar way, OPV 30 2/Chr selects a breeding population to form an optimal haploid, but bases selection on the sum of the population-wide best haplotypes segments. That means there is not a strong selection pressure to develop homozygosity, nor is there selection pressure on most of the genotypes. These two observations are expected as the cause of the greater total additive genetic variance in OPV 30 2/Chr compared to GS. Whereas, it is suspected that OPV 30 2/Chr performs worse than OHV 2/Chr because particular haplotypes are strongly selected for individually throughout the population. Therefore, these haplotypes will eventually tend to dominate the population.

www.manaraa.com

Trends in OPV selection methods

When comparing varying filter percentage used with OPV selection, it is clear that as filter percentage decreases the rate of short-term response increases, there is a greater reduction in upper selection limit, and total additive genetic variance is maintained for fewer generations. This is expected and can be understood in terms of the optimization formulation presented in (5). When filter percentage is increased, (5) is relaxed. This leads to an objective function as good or better than the problem that has not been relaxed. Since the objective function maximizes OPV, this will tend to increase the upper selection limit. However, relaxing the filter percentage in (5) implies the mean GEBV of the next generation can be less than the problem that has not been relaxed. Since optimizing OPV does not guarantee the lines with the greatest GEBV will be chosen, the mean GEBV of the next generation will tend to decrease with increasing filter percentage. Likewise, total additive genetic variance is maintained at a higher level for longer when filter percentage is increased. This is because OPV can be thought of loosely as a measure of diversity since OPV is a weighted measure of beneficial alleles that exist in the population. Thus, maximizing OPV maintains significant levels of total additive genetic variance longer.

Another trend that can be seen is for OPV selection to better maintain the upper selection limit with an increasing number of haplotype segments. Additionally, the rate of short-term response is reduced and a significant level of total additive genetic variance is maintained for longer. Once again, this can be understood by considering (5). Upper selection limit is equivalent to OPV with a haplotype segment of one marker. Since (5) optimizes OPV at a given haplotype length, the upper selection limit calculation will tend

www.manaraa.com

to be largest when the haplotype segment length of the OPV calculation is close to one marker. Another way is to view this is to consider OPV as a model of the upper selection limit. When the number of haplotype segments is increased, the fidelity of the model increases. To understand the tendency for short-term response to decrease with more haplotype segments, consider the extremes. Assume each marker was considered a separate haplotype. Then, optimization is based on an optimal line that may require more than a million recombination events and is, therefore, unachievable in all practical situations. Conversely, if only a few haplotype segments are considered in each line, then the optimization is based on a line achievable in the short-term. As a result, better lines may be realized at a nearer time horizon. When considering total additive genetic variance, significant levels were maintained for longer because there is a correlation between a greater upper selection limit and an increasing number of haplotype segments. As stated for varying filter percentages, this is because OPV can be thought of loosely as a measure of diversity.

Impact on total maize production

A possible 8.3% improvement in response over the course of a 10 year period was demonstrated in this paper. While this is a seemingly minor improvement, it has a huge impact when considering the scope of global maize production. For example, maize production in 2003 was estimated to be 645,164,993 metric tons. In 2013, estimated production was 1,017,536,854 metric tons (FAO, 2015). Over the course of this 10 year period, production increased by 372,371,861 metric tons. If improvements were instead 8.3% greater over this time period, in 2013 the world would have produced 30,906,865 more metric tons of maize in 2013. Although this is a simplified example that assumes

www.manaraa.com

perfectly replicable results for all populations of maize, while assuming the same results hold when using real maize with QTL instead of perfectly estimated marker values and when there are varying levels of heritability, it gives some idea of what is possible with these minor improvements in response.

Validity of results

It is important to note that the marker effects, x_l, were not re-estimated at any time throughout these simulations. This means that while a method's ability to accumulate beneficial alleles at QTL is actually of interest, the method's ability to accumulate markers associated with beneficial alleles at QTL in the initial population is measured. This measurement system is expected to be a valid substitute if the model used to estimate the effects in the initial population is accurate, and the prediction model's accuracy largely results from capturing markers in linkage disequilibrium (LD) with QTL. This is because it has been shown that accuracy results from both LD and genetic relationships, and that accuracy from genetic relationships declines rapidly (Habier, Fernando, & Dekkers, 2007). To mitigate losses in accuracy due to declining genetic relationships, Bayes B was used to estimate marker effects. Bayes B tends to capture markers in LD with QTL better than best linear unbiased prediction methods (Habier, Fernando, & Dekkers, 2007; Zhong, Dekkers, Fernando, & Jannink, 2009) and has been shown to remain accurate for up to 10 generations (Meuwissen & Goddard, 2010). Furthermore, a large number of SNPs were used in these simulations to make the accuracy of estimated effects more robust to changes in LD (Zhong, Dekkers, Fernando, & Jannink, 2009), which is likely to happen after the population has been simulated for

www.manaraa.com

several generations. By using Bayes B, simulating only up to 10 generations, and using a large number of SNPs, it is expected that the conclusions made in this paper are valid.

Conclusion

In this paper, a new selection method was presented. Instead of using evaluations of individual lines to select the breeding population, a candidate breeding population was selected as a unit. While this presents some challenges, such as solving a combinatorial optimization problem, it was shown to outperform existing methods in a series of simulation experiments that spanned 10 generations and used data from an inbred maize population. The statistically detectable improvements in mean response, although in the best case only a modest 8.3% better than GS over 10 generations, could result in significant gains in the worldwide production of maize. In addition to improving response, OPV selection has demonstrated the ability to maintain the upper selection limit better than previous methods such as GS, WGS, and OHV selection. This means that while response may begin to plateau, unfixed beneficial alleles still exist in the population and could translate into subsequent gains in the future. Future research related to OPV selection will focus on demonstrating the robustness of the selection method, improving response by accounting for deadlines either by varying selection intensities or haplotype lengths with time, and by logically picking between candidate breeding populations with the same OPV.

www.manaraa.com

CHAPTER IV

GENERAL CONCLUSIONS AND FUTURE WORK

As the world's population continues to grow, so does the challenge of feeding everyone. In order to produce enough food to meet the demand in 2050, an interdisciplinary approach is needed. In this thesis, operations research tools were applied to GS in order to increase response. While operations research tools have been applied to plant breeding systems before, this paper provides further proof of concept for the integration of optimization, simulation, and mathematical modeling into plant breeding systems.

Within this paper, a new approach to GS, called OPV selection, was studied. Rather than evaluating breeding merit on an individual basis, OPV selection evaluates breeding merit on a population basis. This ensured that individuals selected would be, to some degree, complementary. As a result, OPV selection achieved a greater mean response than the other GS techniques. While the improvements to response were only 8.3% in the best case, this could result in significant improvements to the global maize production. More importantly, however, this thesis demonstrated the potential of incorporating population information into selection decisions. By doing this, a new and potentially fruitful direction of GS has been opened to further research.

Although there are many conceivable ways to select on a population basis in the future, OPV selection should be fully vetted first. To do this, three follow-up studies are recommended.

www.manaraa.com

1) *Investigate selection methods beyond 10 generations*. This is important for several reasons. One such reason is to investigate whether OHV ever surpasses OPV or if it just approaches asymptotically, as seen Figure 5. Additionally, it is important to see if OPV methods that maintain variation and upper selection limits more effectively, such as those with a 50% filter or with 6 and 12 haplotype segments per chromosome, result in greater long-term response than OPV 30 2/Chr. Most important, however, several breeding programs need to be compared. For instance, could some combination of OPV selection methods, at varying filter percentages and haplotype numbers, be used in breeding programs to maintain a high upper selection limit while generating short-term rapid gains? The results of this extension could be relevant to commercial breeders that are motivated by short-term gains, while unsure of long-term consequences of selection based on short-term gains.

However, in order to do this the TBV model used in this paper needs to be modified. While this paper's TBV model limits the amount of noise allowed into the comparisons, the reliability of this model may decay below satisfactory levels after 10 generations (Habier, Fernando, & Dekkers, 2007; Meuwissen & Goddard, 2010). In order to extend this work beyond 10 generations, QTL should be simulated and the prediction model should be updated regularly. To do this, the procedure found in Zhong, Dekkers, Fernando, and Jannink (2009) is proposed.

2) *Account for variable time horizons*. Based on the breeder's equation, GS is optimal when selecting for the next generation. However, the results of the journal paper in Chapter III suggest that GS may not be optimal when selecting for time horizons beyond one generation. This indicates a time dependency. Therefore, an optimal selection

www.manaraa.com

method must account for the time horizon. In Liu, Meuwissen, Sørensen, and Berg (2015), a project deadline was incorporated in weighted genomic selection. For this method, more weight is assigned to rare favorable alleles than common favorable alleles in earlier generations. As time approaches the deadline, all weights converge to "1" and the weighted selection method reduces to GS (Liu, Meuwissen, Sørensen, & Berg, 2015). Another possible way to incorporate a time horizon with existing selection strategies could be to vary selection intensity. For this approach, 2^t lines are selected in each generation, where t is the number of generations remaining until the time horizon. The rationale for this method follows from backwards induction. Assume that the goal is to achieve the single best TBV by some time horizon. Then, in the year prior to that generation some pair is best suited to achieve that goal. Likewise, each parent must have a pair of parents that are best suited for achieving them. This rationale continues until the current generation is reached, resulting in selection of 2^t lines for crossing. Another option is to vary haplotype length with t when using OPV selection or OHV selection. This approach is based on the idea that the number of haplotype segments should reflect the number of recombination events that can be expected (Daetwyler, Hayden, Spangenberg, & Hayes, 2015). If the deadline is further away, then cumulatively more recombination events would be expected. Therefore, more haplotype segments should be considered. To take full advantage of this, however, there should also be a proportional change in the size of the breeding population.

3) *Select the population sub-set with the most probabilistic maximum OPV*. The implementation of OPV selection did not include logic for intelligently selecting between multiple sub-sets of the population with the same OPV. In the future, the sub-set with the

www.manaraa.com

most probability of realizing the maximum possible OPV should be selected. One promising way to do this efficiently is with an extension of predicted parental value (Han, Wang, Beavis, & Cameron, 2016) to more than two lines. In this extension, the probability of achieving a perfect line after t generations would be calculated for a candidate breeding population size 2^t. The candidate breeding population with the greatest predicted parental value would then be chosen.

www.manaraa.com

REFERENCES

Asoro, F., Newell, M., Beavis, W., Scott, M., & Jannink, J.-L. (2011). Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. *The Plant Genome Journal, 4* (2), 132-144.

Barbazuk, W., Emrich, S., Chen, H., Li, L., & Schnable, P. (2007). SNP discovery via 454 transcriptome sequencing. *Plant Journal , 51* (5), 910-918.

Bernardo, R., & Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Science, 47* (3), 1082-1090.

Beyene, Y., Semagn, K., Mugo, S., Tarekegne, A., Babu, R., Meisel, B., … Crossa, J. (2015). Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. *Crop Science, 55* (1), 154-163.

Browning, B., & 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.

Cockerham, C., & Burrows, P. (1980). Selection limits and strategies. *Proceedings of the National Academy of Sciences of the United States of America, 77* (1), 546-549.

Cole, J., & VanRaden, P. (2011). Use of haplotypes to estimate Mendelian sampling effects and selection limits. *Journal of Animal Breeding and Genetics, 128* (6), 446-455.

Daetwyler, H., Hayden, M., Spangenberg, G., & Hayes, B. (2015). Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. *Genetics, 200* (August), 1341-1348.

De Roos, A., Hayes, B., & Goddard, M. (2009). Reliability of genomic predictions across multiple populations. *Genetics, 183* (4), 1545-1553.

Desta, Z., & Ortiz, R. (2014). Genomic selection: Genome-wide prediction in plant improvement. *Trends in Plant Science, 19* (9), 592-601.

Falconer, D. (1981). *Introduction to Quantitative Genetics.* New York: Longman Group Limited.

FAO. (2015). *FAOSTAT, Production, Crops*. Retrieved June 11, 2016, from http://faostat3.fao.org/browse/Q/QC/E

Fernando, R., & Garrick, D. (2009). *GenSel – User Manual for a portfolio of Genomic Selection related Analyses* .

www.manaraa.com

Gianola, D., & Van Kaam, J. (2008). Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics, 178* (4), 2289-2303.

Gibson, J. (1994). Short-term gain at the expense of long-term response with selection of identified loci. *5th World Congress on Genetics Applied to Livestock Production. 21*, pp. 201-204. Guelph, Ontario: Department of Animal and Poultry Science, University of Guelph.

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

Goddard, M., & Hayes, B. (2007). Genomic selection. *Journal of Animal Breeding and Genetics, 124*, 323-330.

Godfray, H., Beddington, J., Crute, I., Haddad, L., Lawrence, D., Muir, J., … Toulmin, C. (2012). Food security: The challenge of feeding 9 billion people. *Science , 327*, 812-818.

Habier, D., Fernando, R., & Dekkers, J. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics, 177* (4), 2389-2397.

Han, Y., Wang, L., Beavis, W. D., & Cameron, J. N. (2016). *A systems engineering approach to defining and improving plant breeding*. Manuscript submitted for publication.

Hayes, B., Bowman, P., Chamberlain, A., & Goddard, M. (2009). Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science, 92* (2), 433-443.

Heffner, E. L., Jannink, J.-l., & Sorrells, M. E. (2011). Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *The Plant Genome, 4* (1), 65-75.

Heffner, E. L., Jannink, J.-L., Iwata, H., Souza, E., & Sorrells, M. E. (2011). Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Science, 51* (6), 2597-2606.

Heffner, E. L., Lorenz, A. J., Jannink, J.-L., & Sorrells, M. E. (2010). Plant breeding with genomic selection: Gain per unit time and cost. *Crop Science, 50* (5), 1681-1690.

Heffner, E., Sorrells, M., & Jannink, J. (2009). Genomic selection for crop improvement. *Crop Science, 49* (1), 1-12.

Hill, W., Goddard, M., & Visscher, P. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics, 4* (2).

www.manaraa.com

Isidro, J., Jannink, J.-L., Akdemir, D., Poland, J., Heslot, N., & Sorrells, M. (2014). Training set optimization under population structure in genomic selection. *Theoretical and Applied Genetics,* (128), 145-158.

Jannink, J., Lorenz, A. J., & Iwata, H. (2010). Genomic selection in plant breeding: From theory to practice. *Briefings in Functional Genomics, 9* (2), 166-177.

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

Kärkkäinen, H., & Sillanpää, M. (2012). Back to basics for Bayesian model building in genomic selection. *Genetics, 191*, 969-987.

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

Lande, R., & Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics, 124* (3), 743-756.

Leaver, J. (2011). Global food supply: A challenge for sustainable agriculture. *Nutrition Bulletin, 36* (4), 416-421.

Leiboff, S., Li, X., Hu, H.-C., Todt, N., Yang, J., Li, X., … Scanlon, M. (2015). Genetic control of morphometric diversity in the maize shoot apical meristem. *Nature Communications, 6*, 8974.

Li, H., Wang, J., & Bao, Z. (2015). A novel genomic selection method combining GBLUP and LASSO. *Genetica, 143* (3), 299-304.

Liu, H., Meuwissen, T., Sørensen, A., & Berg, P. (2015). Upweighting rare favourable alleles increases long-term genetic gain in genomic selection programs. *Genetics Selection Evolution : GSE, 47* (1), 19.

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., … Jannink, J.-L. (2011). Genomic selection in plant breeding. Knowledge and Prospects. *Advances in Agronomy, 110* (C), 77-123.

Lorenz, A., Smith, K., & Jannink, J. (2012). Potential and optimization of genomic selection for fusarium head blight resistance in six-row barley. *Crop Science, 52* (4), 1609-1621.

Lorenzana, R., & Bernardo, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theoretical and Applied Genetics, 120* (1), 151-161.

www.manaraa.com

Lund, M., de Ross, S., de Vries, A., Druet, T., Ducrocq, V., Fritz, S., … Su, G. (2011). A common reference population from four European Holstein populations increases reliability of genomic predictions. *Genetics Selection Evolution : GSE, 43* (1), 43.

MATLAB (2015a). Statistics and Machine Learning Toolbox [computer software]. Natick, Massachusetts, United States: The MathWorks, Inc.

Meuwissen, T., & Goddard, M. (2010). Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics, 185* (2), 623-631.

Meuwissen, T., & Goddard, M. (1996). The use of marker haplotypes in animal breeding schemes. *Genetics, Selection, Evolution : GSE, 28* (2), 161-176.

Meuwissen, T., Hayes, B., & Goddard, M. (2013). Accelerating improvement of livestock with genomic selection. *Annual Review of Animal Biosciences, 1* (1), 221-237.

Meuwissen, T., Hayes, B., & Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics, 157* (4), 1819-1829.

Muir, W. M. (2007). Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *Journal of Animal Breeding and Genetics, 124* (6), 342-355.

Ogutu, J. O., Schulz-Streeck, T., & Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proceedings, 6* (Suppl 2), S10.

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., … Moreau, L. (2012). Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: Comparison of methods in two diverse groups of maize inbreds (Zea mays L.). *Genetics, 192* (2), 715-728.

Romay, M., Millard, M., Glaubitz, J., Peiffer, J., Swarts, K., Casstevens, T., et al. (2013). Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biology, 14* (6), R55.

Rutkoski, J., Singh, R., Huerta-Espino, J., Bhavani, S., Poland, J., Jannink, J.-L., … Sorrells, M. (2015). Efficient use of historical data for genomic selection: A case study of stem rust resistance in wheat. *The Plant Genome, 8* (1), 1-10.

Schnable, P., Liu, S., & Wu, W. (2013). *Patent No. 13/739,874.* United States of America.

www.manaraa.com

Solberg, T. R., Sonesson, A. K., Woolliams, J., & Meuwissen, T. (2009). Reducing dimensionality for prediction of genome-wide breeding values. *Genetics, Selection, Evolution : GSE, 41*, 29.

The Royal Society of London. (2009). *Reaping the Benefits: Science and the Sustainable Intensification of Global Agriculture.* London: The Royal Society.

VanRaden, P. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science, 91* (11), 4414-4423.

What is Operations Research? (2016). Retrieved June 25, 2016, from INFORMS: https://www.informs.org/About-INFORMS/What-is-Operations-Research

What is Plant Breeding? (2016). Retrieved June 25, 2016, from National Association of Plant Breeders: https://www.plantbreeding.org/content/what-is-plant-breeding

Whittaker, J., Thompson, R., & Denham, M. (2000). Marker-assisted selection using ridge regression. *Genetical Research, 75* (2), 249-252.

Wong, C., & Bernardo, R. (2008). Genomewide selection in oil palm: Increasing selection gain per unit time and cost with small populations. *Theoretical and Applied Genetics, 116* (6), 815-824.

Xu, P., Wang, L., & Beavis, W. (2011). An optimization approach to gene stacking. *European Journal of Operational Research, 214* (1), 168-178.

Yu, J., Holland, J., McMullen, M., & Buckler, E. (2008). Genetic design and statistical power of nested association mapping in maize. *Genetics, 178* (1), 539-551.

Zhong, S., Dekkers, J. C., Fernando, R. L., & Jannink, J.-L. (2009). Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. *Genetics, 182* (1), 355-364.

www.manaraa.com

APPENDIX A

PSEUDO-CODE FOR THE IMPLEMENTATION OF THE BASIC BREEDING PROCESS

Simulation Framework: Basic Breeding Process

1. **Read** Experiment Parameters
2. **For each** Generation
3. **Update** Data
4. **Select** Breeding Population
5. **Pair** Breeding Population
6. **Cross** Pairs
7. **End**
8. **Return** Data

www.manaraa.com

APPENDIX B

PSEUDO-CODE FOR THE IMPLEMENTATION OF THE EVALUATION OF TRUE BREEDING VALUE

Function: TBV Evaluation

$V = \textbf{evaluate\_Phenotype}(A^i, x)\{$

1. **Read** A^i and x
2. **Calculate** $V_n = \sum_{lm} A^i_{lmn}\, x_l$ for all n
3. **Return** V

www.manaraa.com

APPENDIX C

PSEUDO-CODE FOR THE IMPLEMENTATION OF THE EVALUATION OF THE UPPER SELECTION LIMIT

Function: OPV Evaluation

$OPV = \textbf{evaluate\_OPV}(A^i, x)$ {

1. **Read** A^i and x
2. **Set** $lociOPV_l = max\ (max\ (A_{l11}x_l, A_{l21}x_l), \ldots, \max\ (A_{l1N}x_l, A_{l2N}x_l))$ for all $l = 1, 2, \ldots, L$
3 **Set** $OPV = 0$
4. **For** $l = 1\ to\ L$
5. $OPV = OPV + lociOPV_l$
6. **End**
7. $OPV = 2 * OPV$
8. **Return** OPV }

www.manaraa.com

APPENDIX D

PSEUDO-CODE FOR THE IMPLEMENTATION OF THE EVALUATION OF THE TOTAL ADDITVE GENETIC VARIANCE

Function: Total Variance Evaluation

$TotVariation =$ **evaluate_TotalVariation**(A^i, x) {

1. **Read** A^i and x
2. **Set** $y_l = 0, z_l = 0$ for all $l = 1\ to\ L$
3. **For** $l = 1\ to\ L$
4. **For** $m = 1\ to\ 2$
5. **For** $n = 1\ to\ N$
6. **If** $A_{lmn} = 1$
7. $y_l = y_l + 1$
8. **Else**
9. $z_l = z_l + 1$
10. **End**
11. **End**
12. **End**
13. **Set** $y_l = y_l/(2 * N), z_l = z_l/(2 * N)$ for all $l = 1\ to\ L$
14. **Set** $varAdd_l = 2 * (y_l * z_l) * x_l^2$ for all $l = 1\ to\ L$
15. **Set** $TotVariation = 0$
16. **For** $l = 1\ to\ L$
17. $TotVariation = TotVariation + varAdd_l$
18. **End**
19. **Return** $TotVariation$ }

www.manaraa.com

APPENDIX E

PSEUDO-CODE FOR THE IMPLEMENTATION OF GENOMIC SELECTION

Function: Genomic Selection

$A^i = \mathbf{select\_Genomic}(A^i, x, q)\{$

1. **Read** A^i, x, and q
2. **Calculate** $V_n = \sum_{lm} A^i_{lmn}\, x_l$ for all n
3. **Order** A^i according to descending order of V
4. **Return** A^i_{lmn} for all $n = 1, 2, \ldots, q$ $\}$

www.manaraa.com

APPENDIX F

PSEUDO-CODE FOR THE IMPLEMENTATION OF WEIGHTED GENOMIC SELECTION

Function: Weighted Genomic Selection

$A^i = \mathbf{select\_WeightedGenomic}(A^i, x, q)\{$

1. **Read** A^i, x, and q
2. **Set** $y_l = 0, z_l = 0$ for all $l = 1\ to\ L$
3. **For** $l = 1\ to\ L$
4. **For** $m = 1\ to\ 2$
5. **For** $n = 1\ to\ N$
6. **If** $A_{lmn} = 1$
7. $y_l = y_l + 1$
8. **Else**
9. $z_l = z_l + 1$
10. **End**
11. **End**
12. **End**
13. **Set** $y_l = y_l/(2 * N), z_l = z_l/(2 * N)$ for all $l = 1\ to\ L$
14. **For** $l = 1\ to\ L$
15. **If** $sgn(x_l) = 1$
16. $weight_l = y_l$
17. **Else**
18. $weight_l = z_l$
19. **End**
20. **If** $weight_l = 0$
21. $weight_l = 1$
22. **End**
20. **End**
21. **Calculate** $V_n = \sum_{lm} A^i_{lmn}\, x_l (weight_l^{-0.5})$ for all n
22. **Order** A^i according to descending order of V
23. **Return** A^i_{lmn} for all $n = 1, 2, \dots, q$ }

www.manaraa.com

APPENDIX G

PSEUDO-CODE FOR THE IMPLEMENTATION OF OHV SELECTION

Function: OHV Selection

$A^i = \textbf{select\_OHV}(A^i, x, q, loci)$ {

1. **Read** A^i, x, q, and $loci$
2. **Calculate** $Y^i_{lmn} = A^i_{lmn} x_l$ for all l, m, and n
3. **For** $j = 1$ *to* $J - 1$
4. **For** $n = 1$ *to* N
5. **Set** $sum1 = 0$ and $sum2 = 0$
6. **For** $l = loci_j + 1$ *to* $loci_{j+1}$
7. **Calculate** $sum1 = Y^i_{l1n} + sum1$
8. **Calculate** $sum2 = Y^i_{l2n} + sum2$
9. **End**
10. **Set** $haploid_{j1n} = sum1$
11. **Set** $haploid_{j2n} = sum2$
12. **End**
13. **Set** $OHV_{j1n} = max\,(haploid_{j1n}, haploid_{j2n})$
14. **Set** $totOHV_{11n} = totOHV_{11n} + OHV_{j1n}$
15. **End**
16. **Set** $totOHV_{11n} = 2 * totOHV_{11n}$
17. **Order** A^i according to descending order of $totOHV$
18. **Return** A^i_{lmn} for all $n = 1, 2, \ldots, q$ }

www.manaraa.com

APPENDIX H

PSEUDO-CODE FOR IMPLEMENTATION OF OPV SELECTION

Function: OPV Selection

$A^i = \textbf{select\_OPV}(A^i, x, q, loci)$ {

1. **Read** A^i, x, q, and $loci$
2. **Calculate** $Y^i_{lmn} = A^i_{lmn} x_l$ for all l, m, and n
3. **For** $j = 1$ *to* $J - 1$
4. **For** $n = 1$ *to* N
5. **Set** $sum1 = 0$ and $sum2 = 0$
6. **For** $l = loci(j) + 1$ *to* $loci(j + 1)$
7. **Calculate** $sum1 = Y^i_{l1n} + sum1$
8. **Calculate** $sum2 = Y^i_{l2n} + sum2$
9. **End**
10. **Set** $haploid_{j1n} = sum1$
11. **Set** $haploid_{j2n} = sum2$
12. **End**
13. **Set** $OHV_{j1n} = max\ (haploid_{j1n}, haploid_{j2n})$
14. **Set** $totOHV_{11n} = totOHV_{11n} + OHV_{j1n}$
15. **End**
16. **Set** $totOHV_{11n} = 2 * totOHV_{11n}$
17. **Order** A^i according to descending order of $totOHV$
18. **For** $j = 1$ *to* q
19. **Set** $indexA_j = j$
20. **Set** $indexB_j = indexA_j$
20. **End**
21. **Set** $check = 0$
22. **While** $check = 0$
23. **For** $k = 1$ *to* q
24. **Set** $index = indexA_j$
25. **Set** $index_p = index_p$ for all $p \neq k$
26. **For Each** $n \in \{1,2, \dots, N\} \notin index$
27. **If** $k = 1$
28. **Set** $candidate_t = concatenate(n, index)$
29. **Else If** $k = q$
30. **Set** $candidate_t = concatenate(index, n)$
31. **Else**
32. **Set** $candidate_t = concatenate(index_{1:k-1}, n, index_{k:q-1})$
33. **End**
34. **End**
35. **For Each** $candidate_t$

www.manaraa.com

36. **Set** $OPV_j = max\ (max\ (haploid_{j1candidate_{t1}}, haploid_{j2candidate_{t1}})$ $, \dots, max\ (haploid_{j1candidate_{tq}}, haploid_{j2candidate_{tq}}))$
37. **Set** $sum = 0$
38. **For** $j = 1\ to\ J - 1$
39. **Calculate** $sum = OPV_j + sum$
40. **End**
41. **Set** $candidateOPV_{\mathrm{t}} = 2 * sum$
42. **End**
43. **Set** $indexA = candidate_{s(1:q)}$ where s is the index of the largest $candidateOPV_t$
44. **End**
45. **Set** check=1
46. **For** $j = 1\ to\ q$
47. **If** $indexA_j \neq indexB_j$
48. **Set** $check = 0$
49. **End**
50. **End**
51. **Set** $indexB_j = indexA_j$
52. **End**
53. **Set** $A^i_{lmj} = A^i_{lmn}$ for all $j = 1, \dots, q$ where $n = indexA_j$
54. **Return** A^i }

www.manaraa.com

APPENDIX I

PSEUDO-CODE FOR THE IMPLEMENTATION OF RANDOM PAIRING

Function: Pair Randomly
$A^i = \mathbf{pair\_rand1}(A^i)$ {
1. **Read** A^i
2. **Order** A^i randomly
3. **Return** A^i }

www.manaraa.com

APPENDIX J

PSEUDO-CODE FOR THE IMPLEMENTATION OF REPRODUCTION

Function: Cross

$A^i = \mathbf{cross}(A^i, r, N)$ {

1. **Read** A^i, r, and N
2. **For** $n = 1$ *to* $q/2$
3. **For** $k = 1$ *to* $N/(\frac{q}{2})$
4. **Generate** $rand1 \sim Uniform(0,1)$ and $rand2 \sim Uniform(0,1)$
5. **For** $l = 1$ *to* L
6. **If** $rand1_l \leq r_l$
7. **Set** $rand1_l = 1$
8. **Else**
9. **Set** $rand1_l = 0$
10. **End**
6. **If** $rand2_l \leq r_l$
7. **Set** $rand2_l = 1$
8. **Else**
9. **Set** $rand2_l = 0$
10. **End**
11. **End**
12. **Set** $sum1 = 0$
13. **Set** $sum2 = 0$
14. **For** $l = 1$ *to* L
15. **Calculate** $sum1 = rand1_l + sum1$
16. **Calculate** $sum2 = rand2_l + sum2$
17. **Calculate** $rand1_l = sum1 \ mod \ 2$
18. **Calculate** $rand2_l = sum2 \ mod \ 2$
19. **End**
20. **For** $l = 1$ *to* L
21. **If** $rand1_l = 0$
22. **Set** $gamete1_l = A^i_{l1n}$
23. **Else**
24. **Set** $gamete1_l = A^i_{l2n}$
25. **End**
26. **If** $rand2_l = 0$
27. **Set** $gamete2_l = A^i_{l1n}$
28. **Else**
29. **Set** $gamete2_l = A^i_{l2n}$
30. **End**
31. **End**

www.manaraa.com

32. Calculate $j = \left((n-1) * \frac{N}{\frac{q}{2}}\right) + k$
33. **Set** $Atemp_{lmj}^{i} = concatenate(gamete1, gamete2)$
34. **End**
35. **End**
36. **Set** $A^i = Atemp^i$
37. **Return** A^i }

www.manaraa.com

APPENDIX K

MEAN AND STANDARD ERROR OF EACH SELECTION METHOD'S TOTAL RESPONSE

	Population 1			Population 2		
Order	Selection Method	Mean Response (x 10^6)	Standard Error (x 10^6)	Selection Method	Mean Response (x 10^6)	Standard Error (x 10^6)
1	OPV 30 2/CHR	2.135	0.005	OPV 30 2/CHR	1.066	0.003
2	OPV 30 3/CHR	2.134	0.004	OPV 10 6/CHR	1.064	0.003
3	OPV 50 3/CHR	2.125	0.005	OPV 30 CHR	1.061	0.002
4	OPV 50 2/CHR	2.120	0.004	OPV 50 2/CHR	1.059	0.003
5	OPV 30 6/CHR	2.114	0.006	OPV 30 3/CHR	1.059	0.002
6	OPV 30 CHR	2.112	0.005	OPV 50 CHR	1.058	0.003
7	OPV 50 CHR	2.102	0.005	GS	1.058	0.004
8	OHV 2/CHR	2.045	0.009	OPV 10 CHR	1.056	0.004
9	OHV 3/CHR	2.043	0.011	OPV 10 2/CHR	1.054	0.005
10	OPV 50 6/CHR	2.030	0.005	OPV 10 3/CHR	1.051	0.005
11	OPV 30 12/CHR	2.026	0.005	OPV 10 12/CHR	1.050	0.004
12	OHV CHR	2.009	0.010	OPV 50 3/CHR	1.041	0.003
13	OPV 10 6/CHR	1.986	0.015	OHV 2/CHR	1.033	0.003
14	OPV 10 12/CHR	1.986	0.017	WGS	1.021	0.002
15	OPV 10 CHR	1.985	0.016	OPV 30 6/CHR	1.019	0.003
16	OPV 10 3/CHR	1.983	0.016	OHV 3/CHR	1.017	0.003
17	OPV 10 2/CHR	1.982	0.016	OHV 6/CHR	1.004	0.004
18	OHV 6/CHR	1.981	0.012	OHV 12/CHR	0.987	0.004
19	GS	1.971	0.016	OHV CHR	0.976	0.005
20	OHV 12/CHR	1.953	0.012	OPV 30 12/CHR	0.919	0.004
21	WGS	1.911	0.009	OPV 50 6/CHR	0.891	0.004
22	OPV 50 12/CHR	1.877	0.005	OPV 50 12/CHR	0.774	0.004

www.manaraa.com

APPENDIX L

MEAN RESPONSE OF OPV 30 2/CHR, OHV 2/CHR, AND GS IN ALL GENERATIONS

		Generation Mean Response (x 10^6)										
Population	Selection Method	0	1	2	3	4	5	6	7	8	9	10
1	OPV 30 2/Chr	0	0.555	1.228	1.499	1.644	1.792	1.916	1.998	2.054	2.098	2.135
	OHV 2/Chr	0	0.613	0.930	1.253	1.488	1.642	1.754	1.848	1.924	1.996	2.045
	GS	0	0.612	1.339	1.467	1.609	1.729	1.800	1.847	1.883	1.914	1.941
2	OPV 30 2/Chr	0	0.335	0.535	0.661	0.762	0.855	0.922	0.968	1.005	1.038	1.066
	OHV 2/Chr	0	0.325	0.499	0.605	0.709	0.789	0.856	0.910	0.957	0.997	1.033
	GS	0	0.336	0.513	0.640	0.760	0.844	0.901	0.945	0.982	1.013	1.040

www.manaraa.com