OPTIMIZING EXPECTED CROSS VALUE FOR GENETIC INTROGRESSION

By
POUYA AHADI

Bachelor of Science in Industrial Engineering
Sharif University of Technology
Tehran, Iran
2018

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
May, 2021

OPTIMIZING EXPECTED CROSS VALUE FOR GENETIC INTROGRESSION

Thesis Approved:

Dr. Baski Balasundaram

Thesis Advisor

Dr. Juan Borrero

Dr. Charles Chen

Dr. Farzad Yousefian

ACKNOWLEDGMENTS

I would like to express my gratitude with sincere respect to my advisor Dr. Baski Balasundaram for his support and encouragement through all my study and research. His extensive discussions around my work and his interesting explorations are of great value to this thesis. His enthusiasm and encouragement made me eager to succeed.

I am very grateful to my MS committee members, Dr. Juan Borrero, Dr. Charles Chen, and Dr. Farzad Yousefian, for dedicating their precious time for my thesis and for their valuable comments. Without their help, my journey would not have been possible.

I also want to thank my parents and my sister for all of their support. They have always been a source of inspiration and motivation to me.

Acknowledgments reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: POUYA AHADI

Date of Degree: MAY, 2021

Title of Study: OPTIMIZING EXPECTED CROSS VALUE FOR GENETIC

INTROGRESSION

Major Field: INDUSTRIAL ENGINEERING AND MANAGEMENT

Abstract: In this study, we consider a combinatorial optimization problem that arises in plant breeding that involves selecting parent plants for crossing based on their genomic characteristics. We wish to ensure that individuals with the most desirable genomic characteristics are selected to increase the likelihood that desirable genetic materials will be passed on to the progeny. Unlike most of the approaches that use phenotypic values for parental selection and evaluate individuals separately, we use a criterion that relies on population genotypic information and evaluates the combination of a pair of individuals. Thus, we introduce the expected cross value (ECV) criterion that takes the vector of recombination frequencies between genes as an input and returns the expected number of desirable alleles for a gamete produced by two individuals of the population as selected parents. We use the ECV criterion to develop a mathematical optimization formulation for the parental selection problem. We target a single phenotypic trait for the genetic improvement program and optimally solving the mathematical formulation to find the best parental pair with maximum ECV. We propose a procedure to obtain multiple parental pairs by finding multiple pairs of (near) optimal solutions. Finally, we discuss how the ECV criterion can improve the genetic introgression process based on computational experiments.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Cells are biological units of any organism. Each cell includes a nucleus that contains the chromosomes of the organism. A chromosome is a long DNA molecule consisting of genes. Genes are basic units of DNA molecules that include the genetic information of an organism. Figure 1 illustrates these basic concepts.

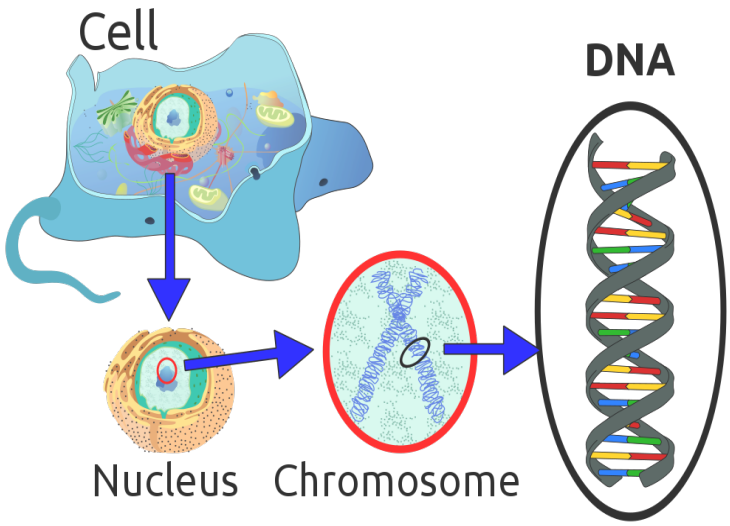

Figure 1: Cell structure [Source: Wikipedia]

Genes are located in fixed positions on the chromosome called loci. A phenotype trait is an observable trait or characteristic of an organism. For example, for the wheat plant, protein content and grain yield are two phenotypic traits. Eye color in humans is another example of a phenotypic trait. A quantitative trait locus (QTL) is a locus that is associated with a trait. A phenotypic trait is determined by many QTL on the chromosome. At each locus, a variation of a gene associated with a phenotypic trait that is present is called an

allele. For a QTL, each allele can be either desirable or undesirable for the associated trait.

The concept of inheritance is also crucial for this study. Inheritance is the process of passing alleles from parents to progeny. This process happens using gamete cells. A gamete is a reproductive cell that contains only half of the genetic information. Meiosis is a special type of cell division that produces the gametes. Two gametes from each of the parents will form the child's genetic material. This process happens by crossing two individuals as parents and the crossing is the process of mating two individuals to obtain a new progeny.

1.1 Breeding program and parental selection problem

The ultimate goal of a plant breeding program is to improve some phenotypic traits over multiple generations. Breeders can achieve this goal by selecting the best individuals from the population and crossing them to create a new generation of progeny, which improves the target phenotypic traits. In plant breeding, this is called *the parental selection problem*. Improving phenotypic traits in the breeding program can be achieved by transferring desirable alleles from parents to progeny and repeating this process for multiple generations. This is called the *genetic introgression process*. The goal of the introgression process is to generate a progeny with as many desirable alleles as possible. Thus, the parental selection problem plays a vital role in the genetic introgression process. We aim to develop a procedure that can lead us to find the best parental pairs from populations such that the proportion of desirable alleles will be increasing through multiple generations in a breeding program.

There are two essential steps for the parental selection problem. First, we need to define a criteria for evaluating individuals or crosses. Second, we need to specify the selection method and how we use the criterion to obtain the best parents out of a population. The flowchart in Figure 2 represents the general idea of the breeding program. We assume that the breeders consider T generations of progeny for their genetic improvement plan. Extending the breeding program for one more generation requires a large amount of time and financial resources, and therefore, breeders intend to make improvements in a specific

number of generations.

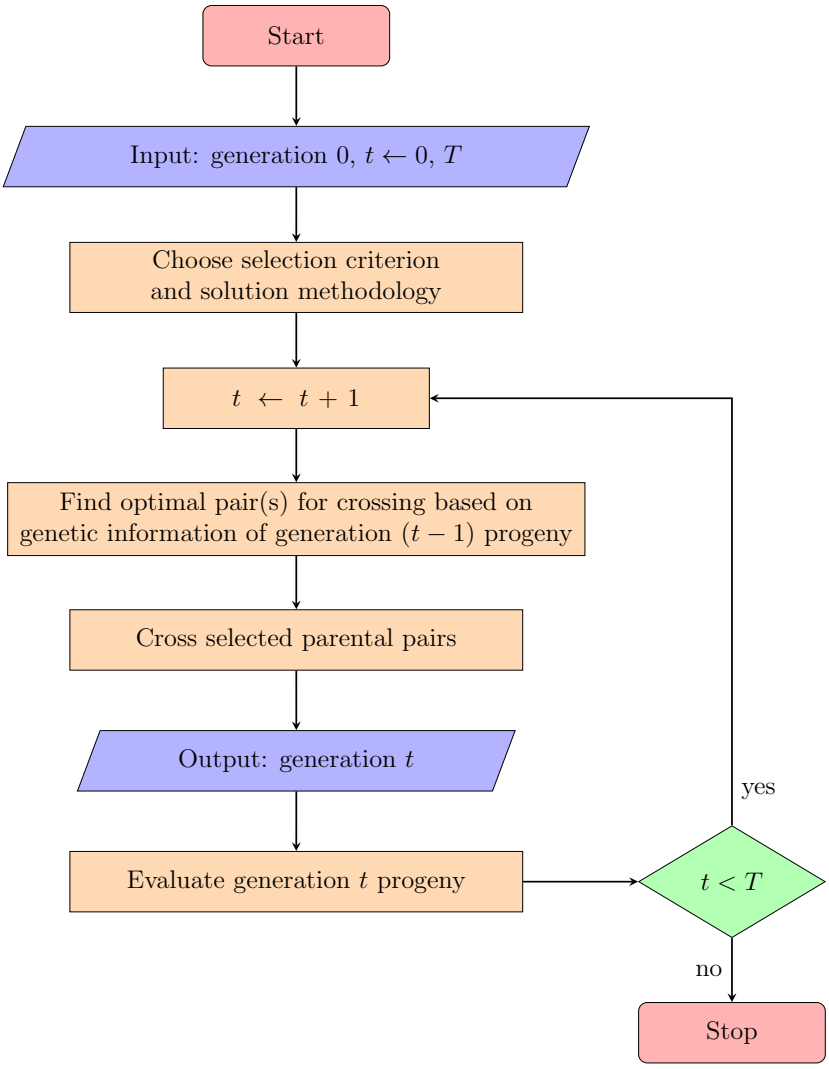

Figure 2: Flowchart for breeding program and parental selection

We can observe in Figure 2 that the program starts with an initial population of individuals (generation 0) and uses the population's genetic information to select optimal pair(s) for crossing. At the end of each generation, the new progeny is evaluated in terms of measures like the proportion of desirable alleles and each target trait's phenotypic value. The process will be repeated until generation T is obtained. Different parental selection criteria can be compared by evaluating generation T progeny.

1.2 Problems of interest

For solving the parental selection problem, having a fitness criterion is a crucial requirement. Any criterion that can help us with improving the target phenotypic trait based on genetic values, can be a viable choice for a breeding program. For a selected pair of individuals, the criterion can be represented as a function of the genetic information of individuals. Then, the problem is to find the best pair of individuals that optimizes this fitness function.

We can propose an extended form of the previous problem by selecting more than one pair from a population for crossing. The single parental pair provides only one new family after crossing for the next generation. In a breeding program, having multiple families from different crosses increases the likelihood of transferring more desirable alleles to the next generation. Moreover, high genomic relationships between individuals of the same family will increase the inbreeding through generations, which is usually not favorable in a breeding program. Thus, breeders prefer to choose more than one cross and obtain multiple families for the next generation. To summarize, we address two types of parental selection problems where the second is an extension of the first:

1. Single parental pair selection problem;

2. Multiple parental pair selection problem.

Recall that there are two main challenges to both single and multiple pair parental selection problems. We need to define a viable criterion that can provide a quantitative measure of fitness to serve as a surrogate for phenotypic performance for any arbitrary pair of individuals in a population. The methodology must be capable of selecting multiple parental pairs out of a population. In the following chapters, we addressed these challenges by proposing a new criterion for parental selection and developing a procedure for selecting multiple pairs of parents.

CHAPTER II

LITERATURE REVIEW

Marker-assisted back-crossing is commonly used for introgression of a single desirable allele [Visscher et al., 1996]. On the other hand, genomic selection was first described by Meuwissen et al. [2001] to estimate individuals' breeding values inside a population for introgressing multiple alleles. Later, genomic selection was used for rapid improvement of traits in breeding [Bernardo, 2009].

Different criteria have been studied for selecting parents in a parental selection problem. The genomic estimated breeding value (GEBV) [Meuwissen et al., 2001] and later the weighted genetic estimated breeding value (WGEBV) [Goddard, 2009, Jannink, 2010], are approaches that use genetic value predictions based on markers. A gene marker is a location on a chromosome that is used for identifying individuals. Daetwyler et al. [2015] proposed the optimal haploid value (OHV) approach, which evaluates the potential genetic value of the population. Wang et al. [2018] show that the OHV approach outperforms the other two approaches in long-term breeding programs. These approaches propose measures that evaluate each individual in a population and return a quantitative value as a breeding value. Finally, individuals with higher breeding values are selected as parental pairs in the breeding program.

These aforementioned criteria evaluate individuals independently, and therefore there is no assessment of a cross. In the genetic introgression problem, we favor finding a viable criterion that can evaluate a pair of individuals and return a quantitative value for each cross. Thus, we can compare all possible crosses and choose the best. Some studies propose

measures defined for a set of individuals as parents. Goiffon et al. [2017] proposed optimal population value (OPV) approach, and a subset of the population with maximum possible haploid value will be selected as the set of parents. This approach can be applied to multi-parental selection programs. Recently, Allier et al. [2019] expanded the concept of usefulness criterion (UC) proposed by Schnell and Utz [1976] to obtain a new measure. The value of UC is used for determining the gain of a cross for a given trait, but the newly developed approach, usefulness criterion parental contribution (UCPC), determines the genetic gain of a multi-parental selection and therefore can be obtained as a measure to quantify the transfer of genetic gain to the next generations.

Some works have studied the selection of multiple parental pairs. Gene pyramiding or stacking is the process of crossing multiple parents and has been used for introgressing multiple desirable alleles. In most cases, gene pyramiding aims to minimize the number of generations to achieve the ideal line, which is the generation of progeny with the highest possible number of desirable alleles. Canzar and El-Kebir [2011] addressed this problem as crossing schedule optimization by adding one more objective. They also considered minimization of the number of crosses required for creating the ideal line. The problem is proved to be NP-hard by Canzar and El-Kebir [2011].

Genetic improvement has been studied in the operations research literature as well. Johnson et al. [1988] used a linear optimization model for deciding the weights of multiple traits in a multi-trait selection problem. Canzar and El-Kebir [2011] and Xu et al. [2011] proposed multi-objective mathematical models to deal with the gene pyramiding problem. More specifically, Canzar and El-Kebir [2011] introduced a mixed-integer linear program (MIP) formulation for crossing schedule problem. De Beukelaer et al. [2015] proposed an improved model heuristic approach for solving the problem. Woolliams et al. [2015] used semidefinite programming for maximizing genetic gain by controlling the inbreeding effect. Akdemir and Sánchez [2016] used mathematical programming to extend the genomic selection to introduce the genomic mating approach. This approach is shown to have better performance in

the long term in comparison to genomic selection approaches.

The concept of predicted cross value (PCV) was introduced by Han et al. [2017]. Unlike most of the existing approaches like GEBV and OHV that evaluate individuals independent of each other, PCV evaluates pairs of individuals and returns the best pair of individuals based on the PCV criterion. For any two arbitrary individuals, the PCV finds the probability that the cross will generate an ideal gamete consisting of only desirable alleles after two generations. The pair of individuals with the highest PCV value will be selected. The so-called "water-pipe algorithm" is suggested for calculating PCV for an arbitrary pair, and an integer programming model is also proposed that maximizes the PCV to find the optimal pair based on genotypic information of the population. The PCV approach is compared to GEBV and OHV approaches in [Han et al., 2017] and it is shown that PCV outperforms other approaches in terms of number of generations it takes to transfer all desirable alleles to create the ideal progeny.

There are two main drawbacks to the PCV criteria. Firstly, a pair of individuals with undesirable alleles in a specific QTL will not be selected based on PCV criteria because the PCV of that selected pair is zero. However, this pair may transfer more desirable alleles coming from different loci and outperform the PCV selection in terms of genotypic or phenotypic values for the next generations. This can be problematic when a breeder looks for some significant improvements in a fewer number of generations. The PCV approach requires a long-term breeding program to show its performance. In most breeding programs, obtaining a new generation requires a vast amount of resources, including time and financial costs. Thus, breeders try to reduce the number of generations that are required for a desired genetic improvement.

The second drawback appears when the genotypic information of the population includes a large number of QTL. In the next chapter, we show that for a large number of QTL, the value of PCV for any selected pair of individuals will be reduced to zero with high probability. This is a conceptual limitation of the PCV criterion, and it makes it virtually impossible to

choose a pair for crossing.

Moreover, there are other aspects of the PCV approach studied by Han et al. [2017] that are noteworthy. The approach is designed for selecting a single parental pair and it ignores the inbreeding effect between parents, as there is no procedure for controlling that effect. Inbreeding happens when two genetically similar individuals are crossed. By crossing parents with high inbreeding for multiple generations, some undesired excessive traits might appear in the future generation's progeny. Also, inbreeding reduces genetic diversity through generations. Therefore, controlling inbreeding between mates is an important target in breeding programs.

In this study, we propose a new criterion called the *Expected Cross Value* (ECV) that returns a cross value for a specific pair of individuals based on the population's genotypic information. We develop an integer programming model to select a pair of parents for introgressing desirable alleles to the next generation with the ECV criterion as the objective while controlling the inbreeding effect between selected parents. Then we propose a procedure for finding multiple pairs of individuals for the multi-parental selection problem. The use of ECV as a new criterion for parental selection will address the limitations of the PCV approach.

CHAPTER III

EXPECTED CROSS VALUE FOR THE PARENTAL SELECTION PROBLEM

In this chapter, we propose and define the new *Expected Cross Value* (ECV) criterion for use in the parental selection problem. In the following, we use the index set notation $[a] \coloneqq \{1, 2. \ldots, a\}$ for any positive integer a.

Definition 3.0.1 *Assume that the target trait is affected by N different QTL in the genome. For each individual, we define an $N \times 2$ binary matrix in which each row represents the pair of alleles in the corresponding QTL. Thus, the genotype matrix L^k associated with the k-th individual is as follows:*

$$L^k_{i,j} = \begin{cases} 1 & \text{if the allele in row } i, \text{ column } j \text{ is desirable,} \\ 0 & \text{otherwise.} \end{cases} \qquad \forall i \in [N], j \in [2] \tag{3.0.1}$$

Genotype matrix information of all individuals is an input for the ECV selection approach. In order to define the ECV criterion, we need to understand how alleles transfer from parents to children, i.e., how a gamete inherits alleles from an individual as a parent. This can be modeled using the concept of inheritance distribution defined in the following.

Definition 3.0.2 (Han et al. [2017]) *Let us suppose $r \in [0, 0.5]^{N-1}$ represents the vector of recombination frequencies between the genes where N is the number of QTL. Consider a random N-dimensional binary vector J that is defined with respect to the vector of recombination frequencies. The inheritance distribution gives the probability that two consecutive*

alleles will be transferred to the gamete from the same chromosome:

$$J_i = \begin{cases} 0 & \textit{if allele in the } i\textit{-th QTL is transferred from first chromosome to the gamete,} \\ & \forall i \in [N], \\ 1 & \textit{otherwise.} \end{cases} \tag{3.0.2}$$

Let us suppose that the first component of the vector J will take a value of 0 *with probability* α_0 *and value of* 1 *with probability* α_1*, thus :*

$$\Pr(J_1 = 0) = \alpha_0,\ \Pr(J_1 = 0) = \alpha_1,\ \alpha_0 + \alpha_1 = 1. \tag{3.0.3}$$

Based on the definition of recombination frequency, if the allele $(i-1)$ *in the gamete is transferred from first or second chromosome of the individual, the probability that the* i*-th allele transfers from the same chromosome is* $1 - r_{i-1}$*. In other words for any* $i \in \{2, \ldots, N\}$*:*

$$\Pr(J_i = J_{i-1}) = 1 - r_{i-1} \tag{3.0.4}$$

$$\Pr(J_i = 1 - J_{i-1}) = r_{i-1} \tag{3.0.5}$$

3.1 Inheritance distribution

The following proposition illustrates an important property related to the inheritance distribution. Consider the following function defined as follows for each $i \in \{2, \ldots, N\}$:

$$\begin{aligned} \phi_i(r) =& (r_1 + r_2 + \cdots + r_{i-1}) + (-2)^1(r_1 r_2 + r_1 r_3 + \cdots + r_{i-1} r_i) + \\ & (-2)^2(r_1 r_2 r_3 + r_1 r_2 r_4 + \cdots + r_{i-2} r_{i-1} r_i) + \cdots + (-2)^{i-2}(r_1 r_2 \ldots r_{i-1}). \end{aligned} \tag{3.1.1}$$

More specifically, for any $i \in \{2, \ldots, N\}$, we can define the following functions:

$$\gamma_i^1(r_j) = r_j, \ \forall j \in [i-1]. \tag{3.1.2}$$

If $i > 2$, for any $m \in \{2, \ldots, i-1\}$, we define the following functions:

$$\gamma_i^m(r_j) = r_j \left(\sum_{k=j+1}^{i-m+1} \gamma_i^{m-1}(r_k) \right), \ \forall j \in [i-m]. \tag{3.1.3}$$

Using these functions, we define the formula for the function $\phi_i(r)$ as follows:

$$\phi_i(r) = \sum_{m=1}^{i-1} \left((-2)^{m-1} \sum_{j=1}^{i-m} \gamma_i^m(r_j) \right), \ \forall i \in \{2, \ldots, N\}. \tag{3.1.4}$$

Proposition 3.1.1 *Suppose J follows inheritance distribution with respect to a vector of recombination frequencies $r \in [0, 0.5]^{N-1}$. If we define α_0 and α_1 based on Equation* (3.0.3)*, for any $i \in \{2, \ldots, N\}$, following equations hold:*

$$\Pr(J_i = 0) = \alpha_0 + (\alpha_1 - \alpha_0)\phi_i(r), \tag{3.1.5}$$

$$\Pr(J_i = 1) = \alpha_1 + (\alpha_0 - \alpha_1)\phi_i(r). \tag{3.1.6}$$

Proof. We model the random vector J as a stochastic process. Let us assume a discrete time Markov chain (DTMC) $J = \{J_n : n \geq 0\}$ where J_n represents the state of the process at n-th step, i.e., the value of the random vector J in the n-th position. The state space is defined as $S = \{0, 1\}$. This process is not a time-homogeneous DTMC. According to Equations (3.0.4) and (3.0.5), the transition probability matrix from step k to step $k+1$ is as follows:

$$P_{k:k+1} = \begin{bmatrix} 1 - r_k & r_k \\ r_k & 1 - r_k \end{bmatrix}, \ \forall k \in [i-1]. \tag{3.1.7}$$

The transition probability matrix from first step to step i is then obtained as follows:

$$P_{1:i} = \begin{bmatrix} 1-r_1 & r_1 \\ r_1 & 1-r_1 \end{bmatrix} \begin{bmatrix} 1-r_2 & r_2 \\ r_2 & 1-r_2 \end{bmatrix} \dots \begin{bmatrix} 1-r_{i-1} & r_{i-1} \\ r_{i-1} & 1-r_{i-1} \end{bmatrix},$$

in other words:

$$P_{1:i} = \prod_{k=1}^{i-1} P_{k:k+1}.$$

We claim that:

$$P_{1:i} = \begin{bmatrix} 1-\phi_i(r) & \phi_i(r) \\ \phi_i(r) & 1-\phi_i(r) \end{bmatrix}, \tag{3.1.8}$$

where $\phi_i(r)$ is defined in Equation (3.1.4). We prove this claim by induction on i.

Base Case: For $i = 2$, Equation 3.1.4 implies that $\phi_2(r) = r_1$, therefore:

$$P_{1:2} = \begin{bmatrix} 1-r_1 & r_1 \\ r_1 & 1-r_1 \end{bmatrix}.$$

This result is identical to the definition of transition probability matrix $P_{1:2}$.

Induction Step: Let us suppose (3.1.8) holds for step $i = n$. We want to prove that it holds for $i = n + 1$ as well. By induction hypothesis, we know that:

$$P_{1:n} = \begin{bmatrix} 1-\phi_n(r) & \phi_n(r) \\ \phi_n(r) & 1-\phi_n(r) \end{bmatrix}.$$

As we know that $P_{1:n+1} = P_{1:n}P_{n:n+1}$, therefore,

$$\begin{aligned} P_{1:n+1} &= \begin{bmatrix} 1-\phi_n(r) & \phi_n(r) \\ \phi_n(r) & 1-\phi_n(r) \end{bmatrix} \begin{bmatrix} 1-r_n & r_n \\ r_n & 1-r_n \end{bmatrix} \\ &= \begin{bmatrix} 1-r_n-\phi_n(r)+2r_n\phi_n(r) & r_n-2r_n\phi_n(r)+\phi_n(r) \\ r_n-2r_n\phi_n(r)+\phi_n(r) & 1-r_n-\phi_n(r)+2r_n\phi_n(r) \end{bmatrix}. \end{aligned} \tag{3.1.9}$$

From Equation (3.1.1) and (3.1.4) we can conclude the following:

$$\phi_{n+1}(r) - \phi_n(r) = r_n - 2r_n\phi_n(r). \tag{3.1.10}$$

By (3.1.9) and (3.1.10), we conclude the following:

$$P_{1:n+1} = \begin{bmatrix} 1-\phi_{n+1}(r) & \phi_{n+1}(r) \\ \phi_{n+1}(r) & 1-\phi_{n+1}(r) \end{bmatrix}, \tag{3.1.11}$$

which means that the claim holds for $i = n+1$. For the DTMC J we have the following property,

$$\Pr(J_i = j) = \left(\alpha^T P_{1:i}\right)_j, \forall i \in \{2,\ldots,N\}, j \in \{0,1\}, \tag{3.1.12}$$

where $\alpha^T = [\alpha_0, \alpha_1]^T$ is the vector of initial probabilities and $\left(\alpha^T P_{1:i}\right)_j$ represents the j-th component of the vector $\left(\alpha^T P_{1:i}\right)$. Thus, for every $i \in \{2,\ldots,N\}$,

$$\begin{bmatrix} \Pr(J_i=0) \\ \Pr(J_i=1) \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix}^T \begin{bmatrix} 1-\phi_i(r) & \phi_i(r) \\ \phi_i(r) & 1-\phi_i(r) \end{bmatrix} = \begin{bmatrix} \alpha_0 + (\alpha_1-\alpha_0)\phi_i(r) \\ \alpha_1 + (\alpha_0-\alpha_1)\phi_i(r) \end{bmatrix},$$

which implies the correctness of Proposition 3.1.1. ■

In this study, we consider diploid parents. A diploid cell has a paired chromosomes, one

from each parent. In this case, we can take advantage of Mendel's second law to illustrate the inheritance of alleles in the first QTL. The following corollary represents this case.

Corollary 3.1.1 *According to Mendel's second law of inheritance, the law of segregation states that for a diploid parent, the two alleles in the first QTL segregate randomly during meiosis; therefore, each allele will transmit to the gamete with equal probability. This implies that* $\alpha_0 = \alpha_1 = 0.5$ *and based on the Proposition 3.1.1 we can conclude the following:*

$$\Pr(J_i = 0) = 0.5,\ \Pr(J_i = 1) = 0.5,\ \forall i \in [N]. \tag{3.1.13}$$

Furthermore,

$$\mathbb{E}(J_i) = 0 \times \Pr(J_i = 0) + 1 \times \Pr(J_i = 1) = 0.5,\ \forall i \in [N], \tag{3.1.14}$$

where $\mathbb{E}(\cdot)$ *represents the expected value.*

3.2 The gamete function and a loss function

Inheritance distribution determines the source of alleles transmitted from a parent to the gamete. Therefore, we can define a function based on inheritance distribution that can specify the alleles in the gamete. This leads to the concept of *gamete function*. In this section, we define a closed form expression for the ECV of a pair of individuals.

Definition 3.2.1 (Han et al. [2017]) *For an individual with genotype matrix* L *and a random binary vector* J *following the inheritance distribution defined in Definition 3.0.2, we denote the gamete function as* $gamete(L, J)$*. The output of this function, the gamete vector*

$g = gamete(L, J)$, is a binary vector defined as follows:

$$g_i = \begin{cases} L_{i,1} & \text{if } J_i = 0, \\ & \forall i \in [N] \\ L_{i,2} & \text{if } J_i = 1. \end{cases} \tag{3.2.1}$$

By the definition of inheritance distribution, we can conclude that:

$$g_i = L_{i,1}(1 - J_i) + L_{i,2}J_i, \ \forall i \in [N]. \tag{3.2.2}$$

Let us suppose we have two individuals with genotype matrices L^1 and L^2 and two samples from inheritance distribution J^1 and J^2. By crossing these two individuals, the genotype matrix for a child in the progeny is represented by matrix $[g^1, g^2]$ where $g^1 = gamete(L^1, J^1)$ and $g^2 = gamete(L^2, J^2)$. We are interested in defining the gamete that is produced by a child of this progeny for the next generation. This gamete is defined as follows where J^3 is another sample from inheritance distribution:

$$g^3 = gamete\big([g^1, g^2], J^3\big). \tag{3.2.3}$$

The gamete g^3 can be used for deriving a criterion for the parental selection problem. To this end, we define a "loss function" in terms of gamete g^3 vector.

Definition 3.2.2 *Using the gamete g^3 from Definition 3.2.1, we define a stochastic loss function as follows:*

$$loss(L^1, L^2, r) = \sum_{i=1}^{N}(1 - g_i^3) = N - \sum_{i=1}^{N} g_i^3. \tag{3.2.4}$$

The loss function is stochastic because the output of the gamete function depends on the random vector J that follows inheritance distribution defined based on the vector of recombination frequencies r. This function counts the number of undesirable alleles in the

gamete g^3, and it can be used for parental selection. Next, we present the definition of PCV introduced by Han et al. [2017].

Definition 3.2.3 (Han et al. [2017]) *For a pair of individuals with genotype matrices L^1 and L^2, we define the gamete g^3 as* (3.2.3)*. The PCV is the probability that the gamete $g3$ contains only desirable alleles.*

In other words, PCV finds the probability that the stochastic loss function is equal to zero. That is,

$$PCV(L^1, L^2, r) = \Pr\big(loss(L^1, L^2, r) = 0\big). \tag{3.2.5}$$

3.3 The drawbacks of PCV

Recall from Chapter II that the PCV is a criterion for parental selection problem that evaluates any possible cross in the population. A pair of individuals with the highest PCV value will be selected for crossing. Han et al. [2017] proposed a polynomial-time algorithm for calculating PCV between any pair of individuals. We now address two drawbacks of the PCV approach. The first issue arises when a specific QTL contains only undesirable alleles for a pair of individuals.

Let us assume an arbitrary pair of individuals from the population with genotype matrices L^k and $L^{k'}$. We are interested in the event where all alleles in both individuals are undesirable for a specific QTL. In other words, there exists an $i \in \{1, \ldots, N\}$ such that,

$$L^k_{i,1} = L^k_{i,2} = L^{k'}_{i,1} = L^{k'}_{i,2} = 0. \tag{3.3.1}$$

For a specific QTL, we represent this event as a failure, denoted as F. We have $P(F) = \alpha^4$ where $\alpha > 0$ is the probability that an allele at any QTL is undesirable. When a failure event occurs in i-th QTL, it implies that the i-th component of gamete g^3 is zero and

hence $PCV(L^k, L^{k'}, r) = 0$. In this case, two individuals will not be selected based on the PCV criterion. However, matrices L^k and $L^{k'}$ might include many desirable alleles in other QTLs, and the cross between individuals k and k' could be a good candidate for genetic improvement.

Proposition 3.3.1 *In the hypothetical setting when $N \to +\infty$, the PCV approaches zero for any pair of individuals.*

Proof. We define the random variable X, which counts the number of QTL at which the aforementioned failure occurs during transmission of alleles from selected individuals to g^3. Obviously, X follows Binomial distribution as $X \sim B(N, \alpha^4)$.

As one failure is enough for making PCV equal to zero, we are interested in finding the probability that at least one failure event occurs.

$$\Pr(X \geq 1) = 1 - \Pr(X = 0) = 1 - (1 - \alpha^4)^N \tag{3.3.2}$$

As $N \to +\infty$, the probability in (3.3.2) converges to the value of one as long as $\alpha > 0$. In other words,

$$\lim_{N \to +\infty} \Pr(X \geq 1) = 1, \tag{3.3.3}$$

which implies the correctness of Proposition 3.3.1. ■

The hypothetical case of $N \to +\infty$ is applicable in practice when N is very large. Our numerical experiments in Chapter V also demonstrate this issue.

3.4 The expected cross value criterion

In this section, we use Definition 3.2.2 to develop a new criterion based on allelic information of individuals. The measure depends on the gamete g^3 defined in Equation (3.2.3) and can evaluate a pair of individuals as a parental cross.

Definition 3.4.1 *For a selected pair of individuals with genotype matrices L^1 and L^2, the ECV is the expected number of desirable alleles in gamete g^3 defined as Equation* (3.2.3)*. As the stochastic loss function represents the number of undesirable alleles in g^3, the ECV can be stated as follows,*

$$ECV(L^1, L^2, r) = N - \mathbb{E}\big(loss(L^1, L^2, r)\big) = \mathbb{E}\bigg(\sum_{i=1}^{N} g_i^3\bigg). \tag{3.4.1}$$

A pair of individuals with the highest ECV value will be selected as parents for crossing. The following theorem provides a closed-form expression for calculating ECV for a pair of parents.

Theorem 3.4.1 *For a selected pair of individuals L^k and $L^{k'}$, the ECV corresponding to the target phenotypic trait can be computed using the following equation:*

$$\mathbb{E}\bigg(\sum_{i=1}^{N} g_i^3\bigg) = 0.25 \sum_{i=1}^{N} (L_{i,1}^k + L_{i,2}^k + L_{i,1}^{k'} + L_{i,2}^{k'}). \tag{3.4.2}$$

Proof. We extend the definition in Equation (3.4.1) to find a closed-form expression for ECV function. For a selected pair of individuals L^k and $L^{k'}$ and three independent samples from inheritance distribution J^1 ,J^2 and J^3, we know $g^3 = gamete\left(\left[g^1, g^2\right], J^3\right)$ where $g^1 = gamete\big(L^k, J^1\big)$ and $g^2 = gamete\big(L^{k'}, J^2\big)$. Based on the definition of inheritance distribution in Equation (3.2.2), we have.

$$g_i^1 = L_{i,1}^k(1 - J_i^1) + L_{i,2}^k J_i^1, \ \forall i \in [N], \tag{3.4.3}$$

$$g_i^2 = L_{i,1}^{k'}(1 - J_i^2) + L_{i,2}^{k'} J_i^2, \ \forall i \in [N], \tag{3.4.4}$$

and,

$$g_i^3 = g_i^1(1 - J_i^3) + g_i^2 J_i^3, \ \forall i \in [N]. \tag{3.4.5}$$

Using Equations (3.4.3) and (3.4.4) in Equation (3.4.5), the expected cross value for the target trait is:

$$\begin{aligned}
\mathbb{E}\Big(\sum_{i=1}^{N} g_i^3\Big) &= \mathbb{E}\Big(\sum_{i=1}^{N} \big(L_{i,1}^k(1-J_i^1)+L_{i,2}^k J_i^1\big)(1-J_i^3)+\big(L_{i,1}^{k'}(1-J_i^2)+L_{i,2}^{k'}J_i^2\big)J_i^3\Big)\\
&= \mathbb{E}\Big(\sum_{i=1}^{N} L_{i,1}^k+\big(L_{i,2}^k-L_{i,1}^k\big)J_i^1-L_{i,1}^k J_i^3-\big(L_{i,2}^{k,}-L_{i,1}^k\big)J_i^1J_i^3+\\
&\qquad L_{i,1}^{k'}J_i^3+\big(L_{i,2}^{k'}-L_{i,1}^{k'}\big)J_i^2J_i^3\Big)\\
&= \sum_{i=1}^{N}\Big(L_{i,1}^k+\big(L_{i,2}^k-L_{i,1}^k\big)\mathbb{E}(J_i^1)-L_{i,1}^k\mathbb{E}(J_i^3)-\big(L_{i,2}^k-L_{i,1}^k\big)\mathbb{E}(J_i^1J_i^3)+\\
&\qquad L_{i,1}^{k'}\mathbb{E}(J_i^3)+\big(L_{i,2}^{k'}-L_{i,1}^{k'}\big)\mathbb{E}(J_i^2J_i^3)\Big). \qquad (3.4.6)
\end{aligned}$$

From Proposition 3.1.1 we know that,

$$\mathbb{E}(J_i^1)=\mathbb{E}(J_i^2)=\mathbb{E}(J_i^3)=\alpha_1+(\alpha_0-\alpha_1)\phi_i(r),\ \forall i\in[N]. \qquad (3.4.7)$$

As J^1, J^2 and J^3 are independent, we know that,

$$\mathbb{E}(J_i^1J_i^3)=\mathbb{E}(J_i^1)\mathbb{E}(J_i^3)=(\alpha_1+(\alpha_0-\alpha_1)\phi_i(r))^2,\ \forall i\in[N], \qquad (3.4.8)$$

$$\mathbb{E}(J_i^2J_i^3)=\mathbb{E}(J_i^2)\mathbb{E}(J_i^3)=(\alpha_1+(\alpha_0-\alpha_1)\phi_i(r))^2,\ \forall i\in[N]. \qquad (3.4.9)$$

Thus,

$$\begin{aligned}
\mathbb{E}\Big(\sum_{i=1}^{N} g_i^3\Big) = \sum_{i=1}^{N}\Big(&L_{i,1}^k+(\alpha_1+(\alpha_0-\alpha_1)\phi_i(r))(L_{i,2}^k-2L_{i,1}^k+L_{i,1}^{k'})\\
&+(\alpha_1+(\alpha_0-\alpha_1)\phi_i(r))^2(L_{i,2}^{k'}+L_{i,1}^k-L_{i,2}^k-L_{i,1}^{k'})\Big). \qquad (3.4.10)
\end{aligned}$$

Based on Mendel's second law, $\alpha_0=\alpha_1=0.5$ and the Equation (3.4.10) reduces to Equation (3.4.2). ■

Remark 3.4.1 *In a case where Mendel's second law does not hold, the ECV criterion is still applicable. In this case,* $\alpha_0 \neq \alpha_1$*, and we can directly use Equation* (3.4.10) *to find the expected cross value between a pair of individuals.*

The ECV approach can be applied for genotypic information, including a large number of QTL, and will not converge to zero as the QTL number increases. This approach can also overcome the other limitation of the PCV approach, as any two individuals with undesirable alleles in a specific QTL will be measured based on the expected number of desirable alleles that they can transfer to the next generation. Theorem 3.4.1 provides a closed-form expression for the ECV criterion that enables us to formulate the parental selection problem as an integer programming problem in the next chapter.

CHAPTER IV

MATHEMATICAL FORMULATION FOR PARENTAL SELECTION

In this chapter, we develop an integer programming (IP) formulation of the parental selection problem using the ECV criterion as the optimization objective. The formulation is based on the mixed-integer programming formulation for the PCV approach introduced by Han et al. [2017]. We also restrict the inbreeding between selected individuals. Using marker genotype information we can find the matrix G that specifies the genomic relationship between any pair of individuals in the population [VanRaden, 2008]. We use matrix G for controlling inbreeding in parental selection. We use following notations in our IP formulation.

Parameters:

- $K \in \mathbb{Z}_{\geq 0}$: Number of individuals in the population
- $N \in \mathbb{Z}_{\geq 0}$: Number of QTL for the target trait
- G: $K \times K$ genomic matrix of inbreeding values with elements $g_{k,k'}$ for $k, k' \in [K]$
- $\epsilon \in \mathbb{R}$: A tolerance parameter for inbreeding relationship between a pair of selected individual
- $P \in \mathbb{B}^{N \times 2 \times K}$: The population matrix, where:

$$P_{i,j,k} = \begin{cases} 0, & \text{if } L^k_{i,j} = 0, \\ & \forall i \in [N],\ j \in [2],\ k \in [K], \\ 1, & \text{otherwise.} \end{cases}$$

Decision Variables:

- $t \in \mathbb{B}^{2\times K}$ representing the parental selection decision such that:

$$t_{m,k} = \begin{cases} 1, & \text{if } k\text{-th individual is selected as } m\text{-th parent,} \\ 0, & \text{otherwise.} \end{cases} \qquad \forall m \in [2],\ k \in [K],$$

- $x \in \mathbb{B}^{N\times 4}$ representing genotypes of selected individuals for all traits. If we suppose k-th and k'-th individuals are selected as first and second parents respectively, so $t_{1,k} = 1$ and $t_{2,k'} = 1$, then:

$$x_{i,j} = L^k_{i,j}, \qquad \forall i \in [N],\ j \in \{1,2\},$$
$$x_{i,j} = L^{k'}_{i,j}, \qquad \forall i \in [N],\ j \in \{3,4\}.$$

Objective Function:

Using the Equation (3.4.2) from Theorem 3.4.1, we formulate the ECV as a function of decision variables as follows:

$$f(t,x) = 0.25 \sum_{i=1}^{N} \big(x_{i,1} + x_{i,2} + x_{i,3} + x_{i,4}\big). \tag{4.0.1}$$

Following is the integer programming formulation for the single parental pair selection prob-

lem.

$$\max \ f(t,x), \tag{4.0.2a}$$

$$\text{s.t.} \ \sum_{k=1}^{K} t_{m,k} = 1, \qquad \forall m \in [2], \tag{4.0.2b}$$

$$x_{i,j} = \sum_{k=1}^{K} t_{1,k} P_{i,j,k}, \qquad \forall i \in [N], \ j \in \{1,2\}, \tag{4.0.2c}$$

$$x_{i,j} = \sum_{k=1}^{K} t_{2,k} P_{i,j-2,k}, \qquad \forall i \in [N], \ j \in \{3,4\}, \tag{4.0.2d}$$

$$t_{1,k} + t_{2,k'} \leq 1, \qquad \forall k, k' \in [K] \mid g_{k,k'} \geq \epsilon, \tag{4.0.2e}$$

$$t_{m,k} \in \{0,1\}, \qquad \forall m \in [2], \ k \in [K], \tag{4.0.2f}$$

$$x_{i,j} \in \{0,1\}, \qquad \forall i \in [N], \ j \in [4]. \tag{4.0.2g}$$

The objective function (4.0.2a) maximizes the ECV. Constraint (4.0.2b) ensures that only two individuals will be selected for the crossing. Constraints (4.0.2c) and (4.0.2d) will assign genotypic information in genotype matrices of selected individuals to $x_{i,j}$ variables. Constraint (4.0.2e) implies that any two individuals with genomic relationship coefficient greater than the tolerance ϵ, will not be selected as parents. As the genomic relationship coefficient between any individual and itself has the highest value equal to one, for any value of ϵ less than one, this set of constraints will prevent self-crossing between individuals. Finally, constraints (4.0.2f) and (4.0.2g) are forcing decision variables to take binary values.

4.1 Extension to multi-parental pair selection

In this section, we propose a method for finding more than one pair of individuals for the multi-parental selection problem. Formulation (4.0.2) will return a pair of individuals as the optimal solution for the optimization model. In a breeding program, we may require to find a set of parental pairs for crossing. Suppose we are interested in finding n_c different parental pairs from the population. The value of n_c must be smaller than the number of initial feasible

crosses. Assuming that the self-crossing is not allowed, we represent the number of feasible solutions (crosses) with n_f which is bounded above by $\binom{K}{2}$. As we impose constraint (4.0.2e) for controlling inbreeding, the number of feasible crosses might be less than $\binom{K}{2}$. In this case, the number of feasible solutions (feasible crosses) would be half the number of elements in matrix G smaller than ϵ.

If there is no element in matrix G that is smaller than ϵ, it means $n_f = 0$ and optimization model in 4.0.2 is infeasible. In this case, we need to increase the value of tolerance ϵ such that there might be at least n_c possible crosses for the selection. Any positive integer value for n_c such that $n_c \leq n_f$ would be suitable for our approach.

Let us assume after solving the model in (4.0.2), we find the optimal solutions where $t^*_{1,i} = t^*_{2,j} = 1$. This means that i-th and j-th individuals are optimal pairs. To obtain another pair from the model, we can add the following "conflict constraints" to the previous optimization model:

$$t_{1,i} + t_{2,j} \leq 1, \tag{4.1.1}$$

$$t_{1,j} + t_{2,i} \leq 1. \tag{4.1.2}$$

These two constraints force the model to choose a different pair of individuals for finding the next optimal solution. We can repeat this procedure to find n_c pairs. The procedure is summarized in Algorithm 1.

Algorithm 1 Finding multiple pairs for the parental selection problem

Input: $S \leftarrow \emptyset$, n_c
Output: Set of selected parental pairs
while $|S| < n_c$ **do**
    Solve Formulation 4.0.2 and obtain optimal solutions $t^*_{1,i} = t^*_{2,j} = 1$.
    Add $\{i, j\}$ to set S.
    Update Formulation 4.0.2 by adding following constraints: $t_{1,i} + t_{2,j} \leq 1$, $t_{1,j} + t_{2,i} \leq 1$.
end while
return S

CHAPTER V

COMPUTATIONAL EXPERIMENTS

In this chapter, we conduct computational studies to assess the ECV criterion. We used the QU-GENE engine and QuLinePlus proposed by Ali et al. [2020] to simulate initial populations and other generations. The QU-GENE engine receives gene information, including recombination frequencies and the number of desired individuals, and returns the simulated initial population. We consider a "wild" initial population assuming that the population is not affected by any selections or diseases, so the proportion of desirable allele for this population should be close to 0.5. QuLinePlus receives genotypic information of a population and a list of selected pairs as well as other parameters and simulates the next progeny by crossing the selected pairs. This provides a tool for performing parental selection for multiple generations and the breeding program represented in Figure 2 in Chapter I. This software also provides genotypic and phenotypic information for all individuals at each generation, and we can use this information to define metrics for assessments in our experiments. We used the Gurobi Optimization Solver [Gurobi Optimization, LLC, 2020] for solving the integer programming formulations that were implemented in the Python programming language.

We report results from two numerical experiments in this section. The first experiment will demonstrate the drawback of the PCV criterion identified in Chapter III. In the second experiment, we compare the ECV approach with the phenotypic selection method to assess the phenotypic performance of the progeny resulting from each selection criterion.

5.1 Limitation of the PCV criterion

Based on Proposition 3.3.1, the PCV approach may not be reliable for a phenotypic trait with a large number of QTL. We simulated five populations; each includes 500 individuals. Based on Definition 3.2.3 if the maximum PCV value in a population is zero, it implies that the PCV value for any pair of individuals equals zero. In this case, the parental selection is impossible because the criterion returns the same value (zero) for any feasible cross in the population. Figure 3 illustrates the results from the experiment. The vertical axis shows the maximum PCV value among all pairs of individuals in a population. We can observe that for $N = 5$, the maximum PCV value in the population equals one, which shows that there is at least one parental pair with the highest possible PCV value. By increasing N, the maximum PCV value decreases, and the criterion returns zero value for any $N \geq 20$ in our experiments.

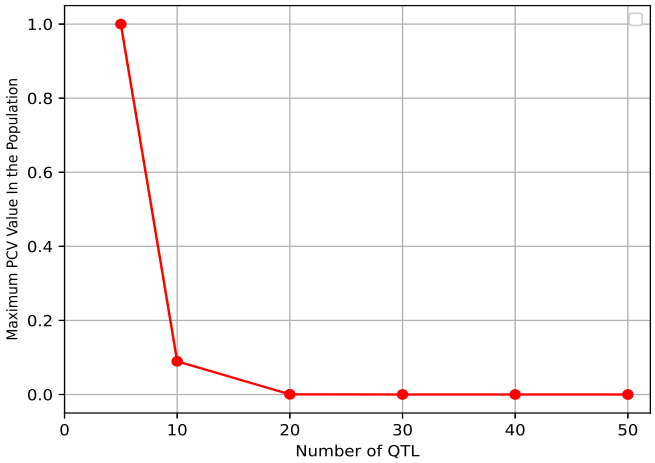

Figure 3: Effect of number of QTL on the PCV approach

5.2 Comparison of ECV vs phenotypic parental selection

This experiment compares the ECV approach with the phenotypic selection method. Phenotypic selection is the most common approach for parental selection. In this approach, for a specific target trait, breeders obtain the phenotypic trait values of all individuals and

choose two individuals with the highest phenotypic values for crossing. We can develop a mathematical optimization model for the phenotypic selection as follows.

Parameters:

- $K \in \mathbb{Z}_{\geq 0}$: Number of individuals in the population
- $p_k \in \mathbb{R}^+$: Phenotypic trait value for k-th individual

Decision Variables:

- $t \in \mathbb{B}^{2\times K}$ representing the parental selection decision such that:

$$t_{m,k} = \begin{cases} 1 & \text{if } k\text{-th individual is selected as } m\text{-th parent,} \\ 0 & \text{otherwise.} \end{cases} \qquad \forall m \in [2],\ k \in [K]$$

Objective Function: Returns the summation of phenotypic values for all selected individuals:

$$g(t) = \sum_{k=1}^{K} \sum_{m=1}^{2} p_k t_{m,k}. \tag{5.2.1}$$

The optimization model for the phenotypic selection is as follows:

$$\max \ g(t) = \sum_{k=1}^{K} \sum_{m=1}^{2} p_k t_{m,k} \tag{5.2.2a}$$

$$\text{s.t.} \ \sum_{k=1}^{K} t_{m,k} = 1, \qquad \forall m \in [2], \tag{5.2.2b}$$

$$\sum_{m=1}^{2} t_{m,k} \leq 1, \qquad \forall k \in [K], \tag{5.2.2c}$$

$$t_{m,k} \in \{0, 1\}. \tag{5.2.2d}$$

Constraint (5.2.2b) ensures that the model will select two individuals. Thus, the objective function (5.2.2a) returns the maximum value for the summation of phenotypic values among

all possible pairs in the population. The inbreeding between pairs is generally disregarded in the phenotypic selection approach and hence, there is no constraint regarding the inbreeding control in the model. Constraint (5.2.2c) removes self-crosses from the set of feasible crosses.To obtain multiple parental pairs, we can use the same approach mentioned in Section 4.1. In this case, $n_f = \binom{K}{2}$, and any positive integer value for n_c that is less than or equal to $\binom{K}{2}$ will be an acceptable choice for the number of crosses.

We conduct this experiment for five generations, i.e., $T = 5$ in the breeding program, as described in Figure 2. We simulated $10,000$ individuals for the first population with 50 QTL for the target trait and 100 individuals per cross for each parental pair.

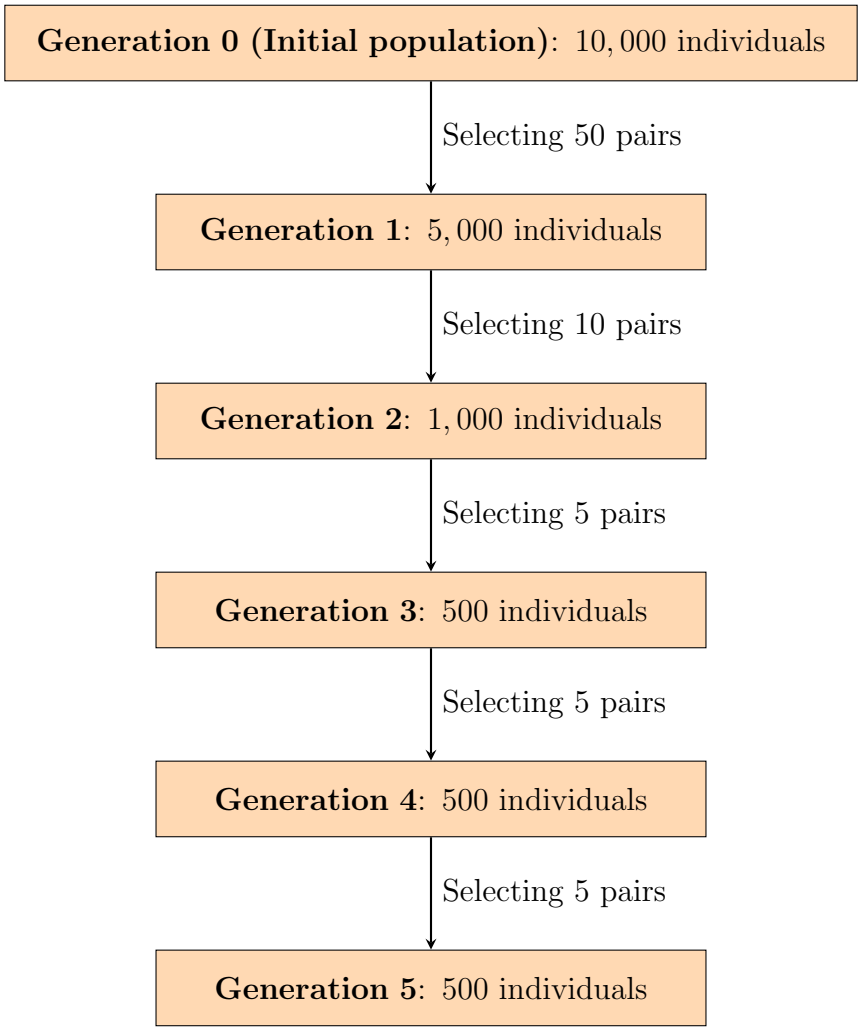

Figure 4: Structure of simulation study

Figure 4 represents the simulation framework for the five generations. We used this

framework to compare the phenotypic selection approach against the ECV approach. The tolerance for the inbreeding coefficient in ECV approach is 0.25, i.e., $\epsilon = 0.25$ in Formulation (4.0.2). The simulation was replicated five times for each selection approach.

We compare the approaches using two different measures: average phenotypic trait values and the average proportion of desirable alleles for each generation of progeny. For a specific individual with genotype matrix L, the proportion of desirable allele is computed as follows:

$$\frac{\sum_{i=1}^{N}\sum_{j=1}^{2} L_{i,j}}{2N}, \tag{5.2.3}$$

where N represents the number of QTL. Figure 5 represents density plots for the five generations. As we observe, by moving from generation 0 to generation 1, the proportion of desirable alleles makes a significant improvement based on the ECV approach. For the final generation, the proportion of desirable alleles for the ECV selection approach takes a value close to one for all replications. By extending the breeding program for more generations, the phenotypic selection approach may achieve similar status, but achieving a significant improvement in fewer generations is the ultimate goal in the breeding program and extending the program for one more generation requires a significant amount of time and resources. Figure 6 represents the outputs for all replications in a box plot. A significant difference for all generations is observable, and the gap between the results of the two approaches is increasing when moving from one generation to the next.

Figure 7 show the density plot for the phenotypic values. We observe that when moving from one generation to the next, both approaches maintain improvements. The difference between approaches is detectable for all generations, especially for the last generation which the progeny obtained by using ECV selection is significantly better. This difference is also noticeable in the box plot for phenotypic values shown in Figure 8. Tables 1 and 2 in the Appendices, illustrate statistical information of the simulation outputs that were used to obtain the plots and representing the 95% confidence intervals for the proportion of desirable alleles and phenotypic trait values for the five replications.

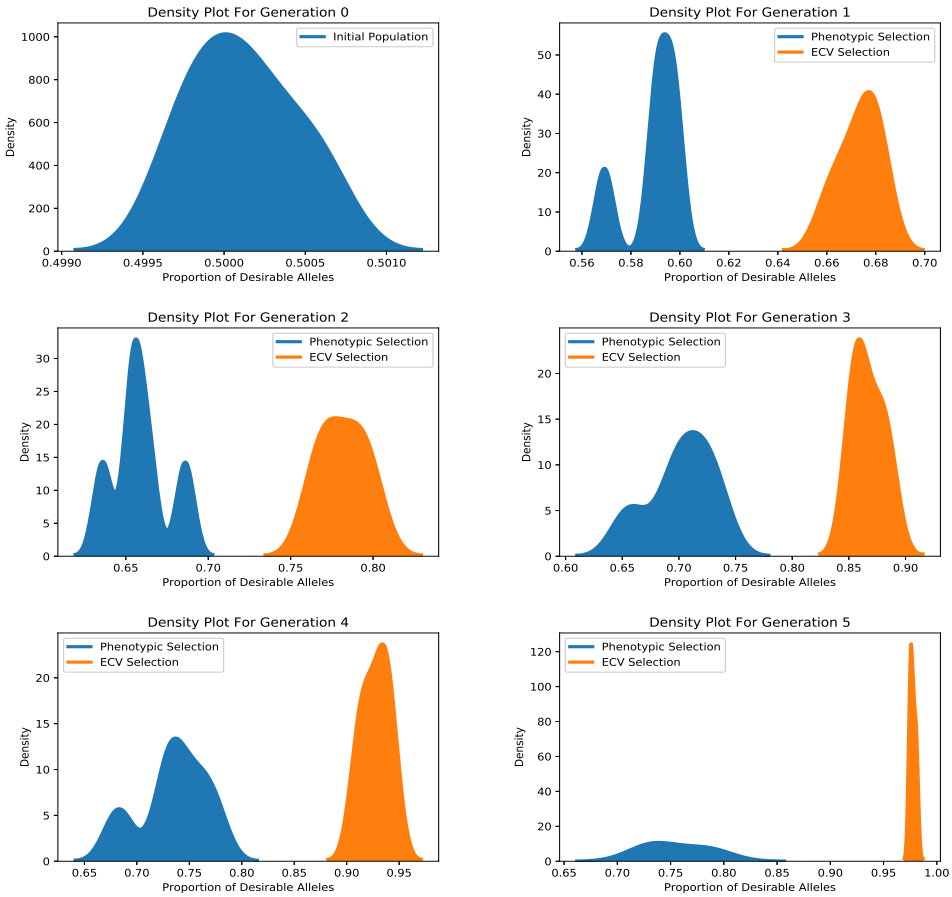

Figure 5: Density plots for the proportion of desirable alleles

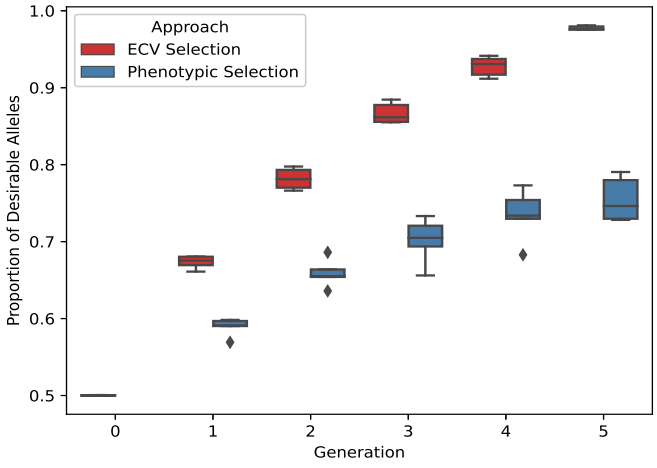

Figure 6: Box plot for the proportion of desirable alleles

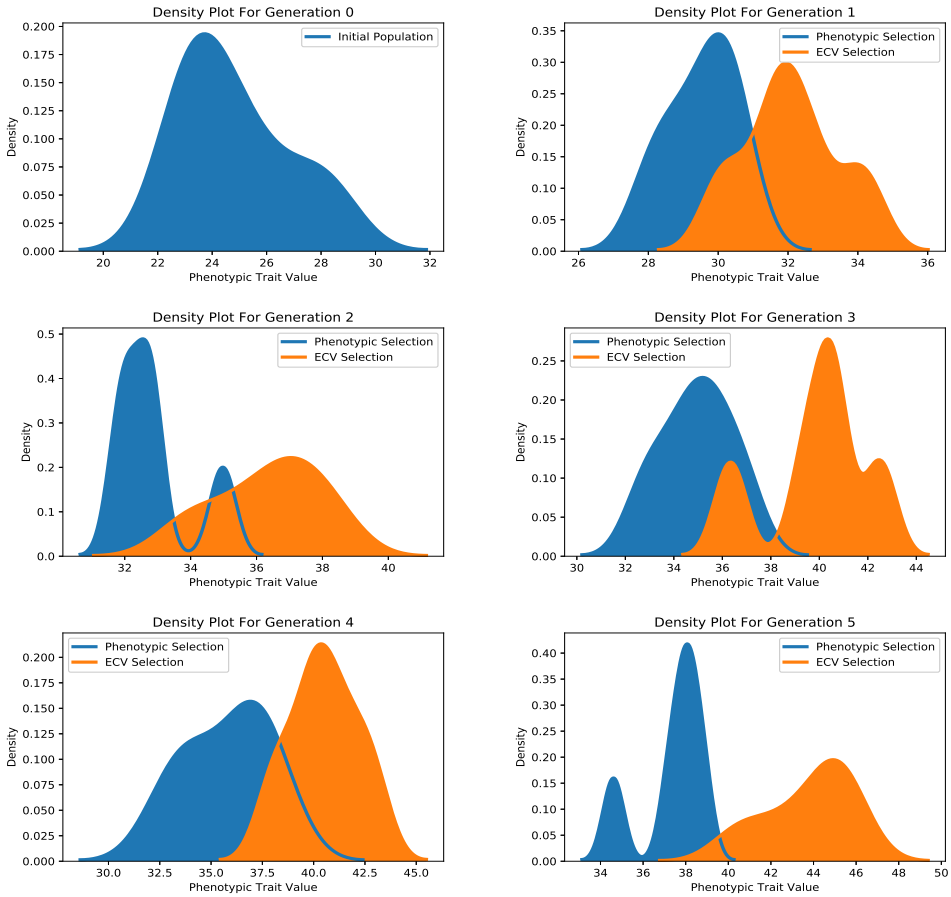

Figure 7: Density plots for the phenotypic values

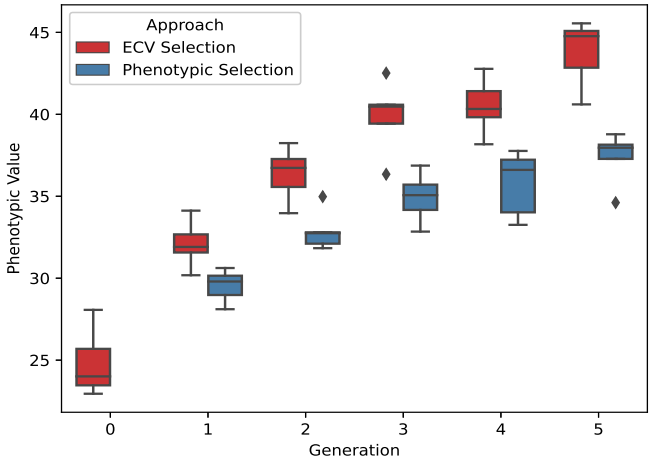

Figure 8: Box plot for the phenotypic values

CHAPTER VI

CONCLUSION AND FUTURE WORK

Developing a robust procedure for parental selection is the key to improvement in any breeding programs. This requires a proper fitness criterion for assessing crosses and a proper mathematical model that can optimize the criterion and return the best pair(s) for crossing.

This thesis introduces a new criterion for the parental selection problem. Unlike some other criteria, ECV evaluates a pair of individuals rather than assessing individuals separately. The criterion is functional for any genotypic data size and any number of QTL. This thesis then proposed a mathematical optimization model to choose an optimal parental pair based on the ECV criterion, with an extension for choosing multiple parental pairs. Based on the results of this study, the ECV approach offers a significant improvement in the breeding program in a few generations and simultaneously limits the inbreeding level between selected parents.

There are several directions for future research on this topic. In particular, considering multi-trait improvement is one of the directions. In a breeding program, breeders might be interested in improving some phenotypic traits simultaneously. In this case, we need to extend the criterion for multiple traits. We assume we have M target traits in the breeding program and N_ℓ, for each $\ell \in [M]$ shows the number of QTL for the ℓ-th trait. We can define the ECV for the ℓ-th trait denoted by ECV^ℓ using the Definition 3.4.1. We tend to choose a parental pair(s) that optimizes M ECV functions simultaneously. In the scope of the breeding program, this might be challenging because some phenotypic traits might have a negative correlation with each other, and that means improving one trait might cause a

decrease in the other.

This problem can be modeled using multi-objective optimization where each objective represents the criterion defined for a specific phenotypic trait. We developed an IP formulation for a single target trait in Chapter IV. We can extend the formulation by considering a vector of objective functions $F(t,x) = \langle f_1(t,x), f_2(t,x), \ldots, f_M(t,x)\rangle$ where $f_\ell(t,x)$, for each $\ell \in [M]$ denotes the ECV functions corresponding to ℓ-th trait (ECV^ℓ). We use the following notations in the multi-objective formulation.

Parameters:

- $K \in \mathbb{Z}_{\geq 0}$: Number of individuals in the population
- $M \in \mathbb{Z}_{\geq 0}$: Number of target traits for the breeding program
- $N^\ell \in \mathbb{Z}_{\geq 0}$: Number of QTL for the ℓ-th trait $\forall \ell \in [M]$
- G: $K \times K$ genomic matrix of inbreeding values with elements $g_{k,k'}$ for $k, k' \in [K]$
- $\epsilon \in \mathbb{R}$: A tolerance parameter for inbreeding relationship between a pair of selected individual
- $P \in \mathbb{B}^{N\times 2\times K\times M}$: Representing the population matrix, where:

$$P_{i,j,k,\ell} = \begin{cases} 0, & \text{if } L^{k,\ell}_{i,j} = 0, \\ & \\ 1, & \text{otherwise,} \end{cases} \qquad \forall i \in [N],\ j \in [2],\ k \in [K],\ \ell \in [M],$$

  where $L^{k,\ell}$ represents the genotype matrix for k-th individual corresponding to the ℓ-th trait.

Decision Variables:

- $t \in \mathbb{B}^{2\times K}$ representing the parental selection decision such that:

$$t_{m,k} = \begin{cases} 1, & \text{if } k\text{-th individual is selected as } m\text{-th parent,} \\ 0, & \text{otherwise.} \end{cases} \qquad \forall m \in [2],\ k \in [K]$$

- $x \in \mathbb{B}^{N\times 4\times M}$ representing genotypes of selected individuals for all traits. If we suppose k-th and k'-th individuals are selected as first and second parents, so $t_{1,k} = 1$ and $t_{2,k'} = 1$, then:

$$x_{i,j,\ell} = L^{k,\ell}_{i,j},\ \forall i \in [N], \forall j \in \{1,2\}, \forall \ell \in [M],$$
$$x_{i,j,\ell} = L^{k',\ell}_{i,j},\ \forall i \in [N], \forall j \in \{3,4\}, \forall \ell \in [M].$$

Objective Function:

We define the ECV corresponding to ℓ-th trait as a function of decision variables:

$$f_\ell(t,x) = 0.25 \sum_{i=1}^{N} \big(x_{i,1,\ell} + x_{i,2,\ell} + x_{i,3,\ell} + x_{i,4,\ell}\big).$$

The multi-objective formulation (6.0.1) for multi-trait single-pair parental selection problem is presented next. The objective (6.0.1a) maximizes the vector of objective functions. Constraint (6.0.1b) states that only two individuals will be selected for crossing. Constraints (6.0.1c) and (6.0.1d) will assign genotypes of selected individuals to $x_{i,j,\ell}$ variables. Constraint (6.0.1e) implies that any two individuals with an inbreeding coefficient greater than tolerance ϵ can not be selected as parents for the crossing program. Note that since the inbreeding coefficient between any individual and itself has the highest value (which equals 0.5), for any value of ϵ less than 0.5, this set of constraints will prevent self-crossing between individuals. Finally, constraints (6.0.1f) and (6.0.1g) are enforcing decision variables to take binary values. The framework mentioned in Chapter IV can be applied here to obtain

multiple parental selections.

$$\max \; F(t,x) = \langle f_1(t,x), f_2(t,x), \ldots, f_M(t,x)\rangle, \tag{6.0.1a}$$

$$\text{s.t.} \; \sum_{k=1}^{K} t_{m,k} = 1, \qquad \forall m \in \{1,2\}, \tag{6.0.1b}$$

$$x_{i,j,\ell} = \sum_{k=1}^{K} t_{1,k} P_{i,j,k,\ell}, \qquad \forall \ell \in [M], \forall i \in [N^\ell], \forall j \in \{1,2\}, \tag{6.0.1c}$$

$$x_{i,j,\ell} = \sum_{k=1}^{K} t_{2,k} P_{i,j-2,k,\ell}, \qquad \forall \ell \in [M], \forall i \in [N^\ell], \forall j \in \{3,4\}, \tag{6.0.1d}$$

$$t_{1,k} + t_{2,k'} \leq 1, \qquad \forall k, k' \in [K] \mid g_{k,k'} \geq \epsilon, \tag{6.0.1e}$$

$$t_{m,k} \in \{0,1\}, \qquad \forall m \in \{1,2\}, \forall k \in [K], \tag{6.0.1f}$$

$$x_{i,j,\ell} \in \{0,1\}, \qquad \forall i \in [N], \forall j \in [4], \forall \ell \in [M]. \tag{6.0.1g}$$

There are several approaches for solving a multi-objective optimization problem. The most common is the *weighted sum approach* where we blend the objectives using weights. This approach requires a vector of weights that capture the importance of each phenotypic trait in the breeding program. In practice, it is difficult to identify a weights vector as there are many external factors that might affect the importance of traits. However, it is possible to order the traits based on their importance. We can assume that the vector of objective functions defined in Equation (6.0.1a) includes the ordering where $f_1(t,x)$ is the most important objective in the breeding program. In this case, we can use the *lexicographic approach* to solve the multi-objective optimization problem. Gurobi solver [Gurobi Optimization, LLC, 2020] provides the tool for using the lexicographic approach with degradation tolerances for objectives. The method starts by optimizing the first objective function and with an allowed tolerance of degradation, optimizes the next objective function. This process is repeated until the last objective is optimized and the optimal solution will be returned as a result of this framework. Investigating this multi-objective framework using ECV is an important direction for future research.

REFERENCES

Deniz Akdemir and Julio I Sánchez. Efficient breeding by genomic mating. *Frontiers in Genetics*, 7:210, 2016.

Mohsin Ali, Luyan Zhang, Ian DeLacy, Vivi Arief, Mark Dieters, Wolfgang H Pfeiffer, Jiankang Wang, and Huihui Li. Modeling and simulation of recurrent phenotypic and genomic selections in plant breeding under the presence of epistasis. *The Crop Journal*, 2020.

Antoine Allier, Laurence Moreau, Alain Charcosset, Simon Teyssèdre, and Christina Lehermeier. Usefulness criterion and post-selection parental contributions in multi-parental crosses: application to polygenic trait introgression. *G3: Genes, Genomes, Genetics*, 9(5):1469–1479, 2019.

Rex Bernardo. Genomewide selection for rapid introgression of exotic germplasm in maize. *Crop Science*, 49(2):419–425, 2009.

Stefan Canzar and Mohammed El-Kebir. A mathematical programming approach to marker-assisted gene pyramiding. In *International Workshop on Algorithms in Bioinformatics*, pages 26–38. Springer, 2011.

Hans D Daetwyler, Matthew J Hayden, German C Spangenberg, and Ben J Hayes. Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. *Genetics*, 200(4):1341–1348, 2015.

Herman De Beukelaer, Geert De Meyer, and Veerle Fack. Heuristic exploitation of genetic structure in marker-assisted gene pyramiding problems. *BMC Genetics*, 16(1):2, 2015.

Mike Goddard. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, 136(2):245–257, 2009.

Matthew Goiffon, Aaron Kusmec, Lizhi Wang, Guiping Hu, and Patrick S Schnable. Improving response in genomic selection with a population-based selection strategy: optimal population value selection. *Genetics*, 206(3):1675–1682, 2017.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2020.

Ye Han, John N Cameron, Lizhi Wang, and William D Beavis. The predicted cross value for genetic introgression of multiple alleles. *Genetics*, 205(4):1409–1423, 2017.

Jean-Luc Jannink. Dynamics of long-term genomic selection. *Genetics Selection Evolution*, 42(1):35, 2010.

Blaine E Johnson, Jerald P Dauer, and Charles O Gardner. A model for determining weights of traits in simultaneous multitrait selection. *Applied Mathematical Modelling*, 12(6):556–564, 1988.

Theo HE Meuwissen, Ben J Hayes, and Michael E Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829, 2001.

Wolfgang Schnell and Friedrich Utz. F1-leistung und elternwahl in der züchtung von selbstbefruchtern. *Bericht über die Arbeitstagung der Vereinigung österreichischer Pflanzenzüchter*, 1976.

Paul M VanRaden. Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11):4414–4423, 2008.

Peter M Visscher, Chris S Haley, and Robin Thompson. Marker-assisted introgression in backcross breeding programs. *Genetics*, 144(4):1923–1932, 1996.

Lizhi Wang, Guodong Zhu, Will Johnson, and Mriga Kher. Three new approaches to genomic selection. *Plant Breeding*, 137(5):673–681, 2018.

John A Woolliams, Peer Berg, Binyam S Dagnachew, and Theo Meuwissen. Genetic contributions and their optimization. *Journal of Animal Breeding and Genetics*, 132(2):89–99, 2015.

Pan Xu, Lizhi Wang, and William D Beavis. An optimization approach to gene stacking. *European Journal of Operational Research*, 214(1):168–178, 2011.

APPENDICES

Comparison Between ECV Approach and Phenotypic Selection

We provide two tables reporting the data used for the plots in Chapter V. Table 1 reports the results based on the proportion of desirable alleles and Table 2 shows the results based on the phenotypic values in progeny. The information shows the confidence intervals for the difference in means obtained by the two approaches and it represents the value of ECV selection approach minus the value of phenotypic selection method.

Method	Generation 1		Generation 2		Generation 3		Generation 4		Generation 5	
	mean	half width	mean	half width	mean	half width	mean	half width	mean	half width
ECV Selection Approach	0.673	0.010	0.782	0.017	0.867	0.017	0.928	0.016	0.977	0.003
Phenotypic Selection Approach	0.589	0.015	0.659	0.023	0.702	0.037	0.735	0.042	0.755	0.036
Difference in means	0.084	0.021	0.123	0.025	0.165	0.038	0.193	0.048	0.223	0.037

Table 1: Confidence intervals for 95% confidence level based on the proportion of desirable alleles for five generation progeny. The results represent confidence intervals for five replications of simulation study.

Method	Generation 1		Generation 2		Generation 3		Generation 4		Generation 5	
	mean	half width	mean	half width	mean	half width	mean	half width	mean	half width
ECV Selection Approach	32.091	1.801	36.356	2.046	39.869	2.811	40.502	2.143	43.770	2.542
Phenotypic Selection Approach	29.531	1.239	32.892	1.536	34.931	1.896	35.776	2.499	37.356	2.018
Difference in means	2.560	1.193	3.464	1.994	4.937	1.355	4.726	1.664	6.414	1.840

Table 2: Confidence intervals for 95% confidence level based on the phenotypic values for five generation progeny. The results represent confidence intervals for five replications of simulation study.

VITA

Pouya Ahadi

Candidate for the Degree of

Master of Science

Thesis: OPTIMIZING EXPECTED CROSS VALUE FOR GENETIC INTROGRESSION

Major Field: Industrial Engineering and Management

Biographical:

Education:

Completed the requirements for the Master of Science in Industrial Engineering and Management at Oklahoma State University, Stillwater, Oklahoma in May 2021.

Completed the requirements for the Bachelor of Science in Industrial Engineering at Sharif University of Technology, Tehran, Iran in 2018.

Experience:

Graduate Teaching Assistant, School of Industrial Engineering and Management, Oklahoma State University (August 2019 - May 2021)

Vice president of the INFORMS Student Chapter, School of Industrial Engineering and Management, Oklahoma State University (August 2020 - May 2021)