Operations research applications in multi-allelic trait introgression in plant breeding

by

Ye Han

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Industrial and Manufacturing Systems Engineering

Program of Study Committee:

Lizhi Wang, Major Professor

William D Beavis

Mingyi Hong

Daniel John Nordman

Sarah M Ryan

Iowa State University

Ames, Iowa

2017

DEDICATION

I would like to dedicate this thesis to my wife and my parents without whose support I would not have been able to complete this work. I would also like to thank my friends for their loving guidance assistance during the writing of this work.

TABLE OF CONTENTS

LIST OF TABLES vi

LIST OF FIGURES vii

ACKNOWLEDGEMENTS x

ABSTRACT xi

CHAPTER 1. General Introduction 1

CHAPTER 2. The Predicted Cross Value for Genetic Introgression of Multiple Alleles 4

2.1 Introduction 5

2.2 Formulation 6

2.2.1 Multi-allelic introgression as an algorithmic process 6

2.2.2 Simplifying assumptions 8

2.2.3 Mathematical formulation of the multi-allelic introgression process . . . 9

2.2.4 Existing approaches for parental selection 11

2.3 PCV for parental selection 12

2.3.1 Definition of PCV 12

2.3.2 The water pipe algorithm for calculating PCV 12

2.3.3 Illustrative example 14

2.4 Conceptual distinctions of PCV, GBV, and OHV 17

2.4.1 Solving Example 3 using the GBV approach 18

2.4.2 Solving Example 3 using the OHV approach 19

2.4.3 Solving Example 3 using the PCV approach 19

2.5 Simulation experiments 21

2.5.1 Experiment description 21

2.5.2 Results for Example 1 22

2.5.3 Results for Example 2 23

2.6 Discussion 23

CHAPTER 3. Dynamic Programming for Resource Allocation in Multi-allelic Trait Introgressions 29

3.1 Introduction 30

3.1.1 Backgrounds 30

3.1.2 MATI Process and Resources Allocation Problem 32

3.2 Material and Methods 38

3.2.1 Model Definition 38

3.2.2 Solving the MDP model 41

3.2.3 Case Study 42

3.3 Results 44

3.3.1 MDP Model Performance Assessment 45

3.3.2 Budget Allocation and Optimal Total Budget 46

3.3.3 Comparison with Static Budget Allocation Strategies 48

3.4 Discussion 51

CHAPTER 4. The Predicted Cross Value for Multi-pair Parental Selection in Trait Introgression Process 53

4.1 Introduction 54

4.2 Formulations 55

4.2.1 Simplifying assumptions 55

4.2.2 Preliminary definitions 56

4.2.3 The NPCV definition 58

4.2.4 The water pipe algorithm for NPCV calculation 60

4.2.5 Illustrative example 61

4.3 Metric Optimization 64

4.3.1 A set cover model for GEBV selection 64

4.3.2 Heuristic NPCV optimization 65

4.4 Simulation Experiments 66

4.4.1 Experiment description 67

4.4.2 Results for Example 1 69

4.4.3 Results for Example 2 71

4.5 Discussion 73

CHAPTER 5. SUMMARY 75

APPENDIX A. Proof for Chapter 2 77

APPENDIX B. Proof for Chapter 4 82

B.1 Transition Matrix 82

B.2 Case I: $2M = 2^a, a \in \mathbb{I}^+$ 82

B.3 Case II: $2M \neq 2^a, a \in \mathbb{I}^+$ 84

B.3.1 Example 1: $2M = 6$ 84

B.3.2 Example 1: $2M = 10$ 86

B.3.3 Discussion 88

BIBLIOGRAPHY 89

LIST OF TABLES

Table 3.1 Parameters 42

Table 3.2 Generation 2 to generation 8 of one random simulation run with fixed budget allocation 48

Table 3.3 Generation 2 to generation 8 of one random simulation run with MDP based budget allocation 49

Table 3.4 Comparison under Optimal Budget 51

4.1 The recombination frequencies used in the simulation. 67

Table 4.2 A Random Simulation Run for Example 1 with 2 Pairs of Breeding Parents per Generation with the NPCV Approach. 68

Table 4.3 Simulation Result Summary of Example 1 70

Table 4.4 Simulation Result Summary of Example 2 73

LIST OF FIGURES

Figure 2.1 Flowchart of the multi-allelic introgression process. 7

Figure 2.2 Illustration of the plumbing system for Example 1. 15

Figure 2.3 Illustration of the plumbing system for Example 2. 16

Figure 2.4 An illustration of 10 loci in a population consisting of 50 individuals. . 18

Figure 2.5 The GEBVs for Example 3. 19

Figure 2.6 The OHVs for Example 3. 19

Figure 2.7 The PCV map for Example 3. 20

Figure 2.8 The recombination frequencies for the simulation between adjacent pairs of loci. 22

Figure 2.9 Performance of the GEBV, OHV, PCV-I, and PCV-II approaches in 1000 simulation runs of trait introgression of seven QTL. The vertical axis represents the proportion of desirable alleles in the genome. Histograms of the proportion of desirable alleles among 100 progeny from 1,000 simulation runs are plotted for each generation. The red curve shows the population mean. 25

Figure 2.10 Distributions of the terminal generation numbers of PCV-I and PCV-II approaches. 26

Figure 2.11 Performance of the GS, OHV, PCV-I, and PCV-II approaches in 1000 simulation runs of trait introgression of twenty QTL. 27

Figure 2.12 Distributions of the terminal generation numbers of PCV-I and PCV-II approaches. 28

Figure 3.1 Flowchart of the MATI process 34

Figure 3.2 Genotype indicator 39

Figure 3.3 BTP graph with $T = 8$. In the figure, the horizontal axis is different total budget scenarios of the breeding project and the vertical axis represents a stacked histogram of the probabilities of reaching success at different generations. In the figure, "GX" label means that the breeding process successfully finishes in X generations and "Failure" means no ideal individual is produced when the budget or the time is depleted. . 45

Figure 3.4 Budget allocation with $T = 8$. In the figure, the horizontal axis is different total budget scenarios of the breeding project and the vertical axis represents the proportion of budget allocated in different generations. Different gray scale are used for different generations. 46

Figure 3.5 Profits and Budgets. In the figure, the blue pentagrams represent the estimation results from simulations and the blue curve represents a nonlinear regression with model $y = a_1 + a_2 \times \exp(a_3 x)$ for the estimation. The red squares represent the difference between the adjacent estimations and the red cure represents the derivative of the expected total revenue curve. The red horizontal line is the marginal return equals to one unit increment of total budget, which is $1,000. 47

Figure 3.6 Comparison under Optimal Budget. The left 7 stacked bars represent the static budget allocation strategies with different progeny number per generation while the last barplot represents the MDP based strategy. 50

Figure 4.1 Scheme for NPCV 59

Figure 4.2 Illustration of the plumbing system for Example 4. 62

4.3 Illustration of the plumbing system for Example 5. 63

4.4 Example 1: Combined number of generations to terminate. Performance of the GEBV and the heuristic NPCV approaches in all the simulation runs of trait introgression of seven QTLs. The vertical axis represents the proportion of number of generations to terminate. The horizontal axis represents the possible number of generations to terminate. 69

4.5 Example 1: Proportion of desirable alleles in each generation with 6 pairs of breeding parents selected. The box plots for the average proportion of desriable alleles in the 6 pairs of selected breeding parents with the GEBV and the heuristic NPCV approaches in each generation. The vertical axis represents the proportion of desirable alleles in the 6 pairs of selected breeding parents. The horizontal axis represents different generations. 71

4.6 Example 2: Combined number of generations to terminate. Performance of the GEBV and the heuristic NPCV approaches in all the simulation runs of trait introgression of seven QTLs. The vertical axis represents the proportion of number of generations to terminate. The horizontal axis represents the possible number of generations to terminate. 72

4.7 Example 1: Proportion of desirable alleles in each generation with 6 pairs of breeding parents selected. The box plots for the average proportion of desriable alleles in the 6 pairs of selected breeding parents with the GEBV and the heuristic NPCV approaches in each generation. The vertical axis represents the proportion of desirable alleles in the 6 pairs of selected breeding parents. The horizontal axis represents different generations. 74

B.1 Conceptual Figure for the Breeding Scheme for 3 Pairs of Parents . . . 85

B.2 Conceptual Figure for the Breeding Scheme for 5 Pairs of Parents . . . 86

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis. First and foremost, Dr. Lizhi Wang for his guidance, patience and support throughout this research and the writing of this thesis. His insights and words of encouragement have often inspired me and renewed my hopes for completing my graduate education. I would also like to thank Dr. Beavis, for his guidance and support for helping me understand the agronomy knowledge. Also, I would like to thank my committee members for their efforts and contributions to this work: Dr. Hong, Dr. Nordman and Dr. Ryan. I would additionally like to thank my friends and colleagues for their help and encouragement throughout my graduate study.

ABSTRACT

Plant breeding has been defined as the art and science of producing desired characteristics through artificial selection. Practiced since the beginning of civilizations, plant breeders in the 20th Century made enormous changes to important agronomic traits. In the 21st Century, increasing demands for food, fiber and energy with less water, land, fuel and fertilizer will force plant breeding to become more efficient and effective. With the application of operations research, we frame the multi-allelic trait introgression project in plant breeding into an engineering system. We also discuss the major problems encountered and create new metrics or models to improve this process. We design the Predicted Cross Value (PCV) for one pair parental selection and demonstrate its advantages over the conventional metrics. Next, in order to optimize the resource allocation during the introgression, we propose the Markov Decision Process model to dynamically allocate resources. The results show that such approach outperforms the static breeding strategy. Finally, to make the PCV concept more practical to realistic breeding process, we extend the PCV to NPCV for multi-pair parental selection. We present the results and show that the NPCV makes the parental selection more efficient and effective. In general, this dissertation discusses applying operations research into the trait introgression process to improve the efficiency and effectiveness from several perspectives.

CHAPTER 1. General Introduction

Plant breeding has been defined as the art and science of producing desired characteristics through artificial selection (Poehlman, 2013). Practiced since the beginning of civilizations, plant breeders in the 20th century made enormous changes to important agronomic traits, e.g., grain yield and pest resistance, of cereal crops (Duvick, 1994; Rincker et al., 2014). This was accomplished through ad hoc adoptions of emerging technologies developed by agricultural, mechanical, electrical and information engineers. In the 21st century, demands for increasing production of food, fiber and energy with less water, fuel and fertilizer will force plant breeding to become more efficient and effective.

Discovery of genetic variants associated with crop phenotypic variants have been accelerating through use of forward and reverse genetics approaches. We now have databases cataloging thousands of genetic variants (alleles) associated with favorable phenotypic variants in large germplasm repositories (McCouch et al., 2012; Cavanagh et al., 2013). This information tells us that favorable alleles are distributed unevenly throughout germplasm collections and unevenly across crop genomes. These resources will provide desirable alleles for genetic improvement of crops in rapidly changing environments (Kumar et al., 2010; Leung et al., 2015).

Introgression of a single desirable allele from an inferior agronomic cultivar to an elite cultivar is routinely accomplished using marker assisted backcrossing strategies (Visscher et al., 1996; Frisch et al., 1999; Frisch and Melchinger, 2005; Peng et al., 2014a). Furthermore, as long as there are very few cultivars that are capable of maintenance and regeneration in tissue culture, creation of novel alleles through genome editing technologies will likewise depend on trait introgression for cultivar development. Introgression of multiple alleles is not as well studied, but genomic selection (Bernardo, 2009; Longin and Reif, 2014; Gorjanc et al., 2016) and marker assisted gene pyramiding (Servin et al., 2004; Xu et al., 2011; Canzar and El-Kebir, 2011; De Beukelaer et al., 2015) have been proposed as approaches for introgressing multiple alleles from unadapted landraces into elite cultivars.

The more complex challenge of aggregating sets consisting of multiple alleles into cultivars with predictable adaptive trait phenotypes will require transfer of knowledge from operations researchers and mathematicians to plant breeders. This dissertation discusses a series of research conducted from the perspective of operations research aiming at improving the efficiency and effectiveness of the trait introgression project. It explores two related fundamental topics in trait introgression, which are parental selection and resource allocation. The topic of parental selection is about how to efficiently select the breeding targets in the project based on genotypic and phenotypic information, while the resource allocation is about how to design more efficient breeding strategy to utilize the resources in the project. The remainder of the dissertation is organized as follows.

The first chapter is constructed from a paper published in the journal of Genetics (Han et al., 2017) on the topic about parental selection. This chapter formulates the multi-allelic trait introgression (MATI) as an engineering system and designs an algorithmic process with mathematical definitions. A new metric for parental selection, which is named as the "Predicted Cross Value" (PCV) with assistance of genetic markers is proposed in the chapter. Via the PCV metric, significant improvements and the great potential for further research on trait introgression projects are demonstrated.

The second chapter is constructed from a paper submitted to the journal of Frontiers of Genetics. The resources allocation in the introgression plays a crucial role as well as parental selection because well designed allocation plans can improve the efficiency of the breeding projects dramatically. In this chapter, we expand our discussion on designing more efficient strategy based on the Markov Decision Processes (MDP) model. In the chapter, we complete the process of multi-allelic trait introgression and propose an updated version of algorithmic simulating process for MATI process. At the same time, we formally state the resource allocation problem for introgression process and define the MDP model to solve the resource allocation problem. The results are demonstrated via computer simulation based case studies and comparisons with other breeding strategies according to the assessing criteria are proposed, as well. In the final part of this chapter we derive the conclusion that better resources allocation plans can accelerate the breeding project significantly.

The third chapter is constructed from a paper to be submitted to the journal of Genetics on the topic about parental selection for multiple breeding parents. In this chapter, we review the limitation of the predicted cross value (PCV) for one pair of breeding parents and propose the NPCV concept for multi-pair breeding parents selection. We update the plumbing system for calculation, as well. At the same time, we propose set cover models for the conventional approach and the new NPCV metric to select the optimal breeding parents. According to the simulation results, the NPCV metric is proved to outperform the conventional approach and improve the efficiency and effectiveness of trait introgression process significantly.

CHAPTER 2. The Predicted Cross Value for Genetic Introgression of Multiple Alleles

Abstract

We apply operations research approaches to optimize introgression of multiple alleles from a donor to a recipient genome. First, we frame the trait introgression project as an algorithmic process that can be mathematically formulated and optimized. We then introduce a novel metric for selecting breeding parents that we refer to as the Predicted Cross Value (PCV). Unlike the various forms of estimated breeding values, the PCV retains recombination as an essential parameter and calculates the probability that a pair of parents will produce a gamete with desirable alleles at all quantitative trait loci. We compared the PCV approach with existing approaches in two simulation experiments, in which seven and twenty desirable alleles were to be introgressed from a donor line into a recipient line. Results suggest that the PCV is more efficient and effective for multi-allelic trait introgression than existing approaches. We also discuss how the operations research framework can be used for other crop genetic improvement projects and several potential research directions.

2.1 Introduction

Discovery of genetic variants associated with crop phenotypic variants have been accelerating through use of forward and reverse genetics approaches. We now have databases cataloging thousands of genetic variants (alleles) associated with favorable phenotypic variants in large germplasm repositories (McCouch et al., 2012; Cavanagh et al., 2013). This information tells us that favorable alleles are distributed unevenly throughout germplasm collections and unevenly across crop genomes. These resources will provide desirable alleles for genetic improvement of crops in rapidly changing environments (Kumar et al., 2010; Leung et al., 2015).

Introgression of a single desirable allele from an inferior agronomic cultivar to an elite cultivar is routinely accomplished using marker assisted backcrossing strategies (Visscher et al., 1996; Frisch et al., 1999; Frisch and Melchinger, 2005; Peng et al., 2014a). Furthermore, as long as there are very few cultivars that are capable of maintenance and regeneration in tissue culture, creation of novel alleles through genome editing technologies will likewise depend on trait introgression for cultivar development. Introgression of multiple alleles is not as well studied, but genomic selection (Bernardo, 2009; Longin and Reif, 2014; Gorjanc et al., 2016) and marker assisted gene pyramiding (Servin et al., 2004; Xu et al., 2011; Canzar and El-Kebir, 2011; De Beukelaer et al., 2015) have been proposed as approaches for introgressing multiple alleles from unadapted landraces into elite cultivars.

The genomic estimated breeding value (GEBV) is a commonly used measure for parental selection in not only trait introgression but also genomic selection projects. More recently, the optimal haploid value (OHV) (Daetwyler et al., 2015) was proposed as an alternative breeding value, which measures the potential rather than the realized fitness. Herein, we propose a new metric, the Predicted Cross Value (PCV), for parental selection in introgression of multiple alleles. This metric calculates the probability that a cross will produce an ideal genotype in two generations. The main difference between this new metric and existing metrics for parental selection is that the PCV is defined for two breeding parents using recombination frequency information to measure their complementarity, whereas the GEBV and OHV are defined for a single individuals, assuming that the merit of a cross is an additive function of the two individuals' breeding values.

We compare selection using PCV with GEBV and OHV in two multi-allelic introgression projects: a) Introgression of seven independently segregating alleles. Such situations occur in self pollinated crops, e.g., sorghum and soybean, where the goal is to adapt a tropical line to high latitudes for purposes of evaluating other agronomic traits without confounding influences of maturity. b) Introgression of a larger number of alleles from an exotic into an elite cultivar for purposes of improving a polygenic trait.

2.2 Formulation

The general objective of multi-allelic introgression projects is to transfer a discrete set consisting of multiple desirable alleles, or haplotypes, from a donor to a recipient. The ultimate goal is to produce at least one individual genome consisting of homozygous desirable haplotypes and no other marker alleles from the donor. The introgression process begins by identifying the donor and recipient cultivars based on criteria defined by the breeder. The selected cultivars are then planted, grown to sexual maturity and crossed. The resulting seeds are harvested and planted along with the recipient parent. The progeny are evaluated to assure that they represent the F1 generation with half of their genomes inherited from each parent. In subsequent filial generations, breeding parents are selected from the current population to be crossed, and the progeny are evaluated to determine if any meet the goal. If not, the process of selection and reproduction will be repeated.

2.2.1 Multi-allelic introgression as an algorithmic process

We illustrate the major components of the introgression process in Figure 2.1 and explain each of the components as follows.

- **The** Start **point**

  The introgression process starts with identification of at least one recipient and one donor. In the case of most annual crops both recipient and donor are homozygous throughout

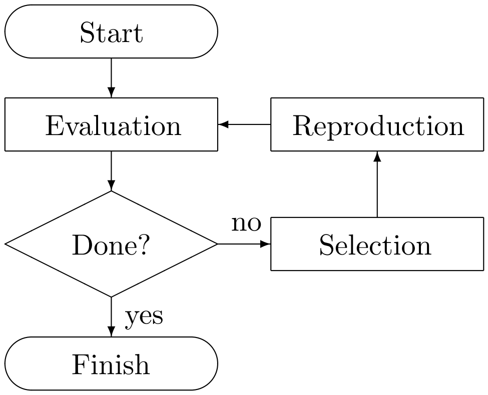


Figure 2.1: Flowchart of the multi-allelic introgression process.

their genomes. The majority of alleles in the donor are undesirable, but do have desirable versions of alleles that the recipient is lacking.

- **The** Evaluation **step**

  In this step, marker genotypes of individuals in the current generation are evaluated.

- **The** Done? **condition**

  The stopping condition is checked in this step, which is whether the newest generation of progeny contains an individual that is homozygous with only desirable alleles from both the recipient and donor.

- **The** Selection **step**

  In this step, breeding parents are selected from the current generation of individuals to produce the next generation of progeny. The current generation includes the recipient line and the newly produced generation of progeny but not individuals from previous generations. This is because the recipient is a replicable entity, whereas individual progeny from previous generations have lived through their life-cycle and were not replicable. If the cross involves the recipient cultivar, then it is referred to as a backcross. Another special case of selection is to select only one plant to cross through self-pollination.

- **The** Reproduction **step**

  In this step, the breeding parents selected from the Selection step are crossed to produce a new generation of progeny. The genotypes of this next generation of progeny are produced through the stochastic processes of transmission genetics.

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

  The goal of an introgression breeding project is to produce an ideal line that inherits only the desirable alleles from the recipient and the donor line. In other words, the ideal line is a homozygous one that does not contain undesirable alleles. The breeding process finishes when an ideal line has been produced. This line will then proceed to further stages of new seed variety development.

2.2.2 Simplifying assumptions

Several assumptions are made in order to simplify the formulation and illustrate the core elements of the process. In Section 6, we discuss relaxing these assumptions in future studies.

- Consider annual diploid and allopolyploid species such as corn, rice, soybean and wheat with subgenome specific loci. Extension to perennial and autopolyploid crops, such as alfalfa is deferred for future research.

- Consider a single multi-allelic trait, where all segregating loci associated with the trait are known. Results also apply to multiple traits where all traits are of equal value.

- All marker alleles are either desirable or undesirable. Values of alleles could be modeled as continuous from some distribution or in many cases, the value of an allele is unknown. We defer expansion to these situations for future research.

- To illustrate the principles, all desirable alleles missing in the recipient are carried by one donor line. Consideration of desirable alleles from multiple donors with each one carrying a subset is deferred for future research.
- One pair of parents is selected for crossing in each generation, with self-pollination as a special but feasible option. In actual breeding practice, multiple crosses are sometimes made to produce sufficient numbers of progeny for field trial evaluations. Our approach readily extends to these situations.
- During evaluation, a sufficient number of informative markers are distributed throughout the genome at sufficient density to allow estimation of recombination between all adjacent pairs of markers.
- Recombination events between pairs of adjacent loci are assumed to be independent (Haldane, 1919). Consideration of interference is deferred for future research.

2.2.3 Mathematical formulation of the multi-allelic introgression process

We use an N by 2 binary matrix, say $L \in \mathbb{B}^{N \times 2}$, to represent the genotype of an individual plant, where N is the total number of QTL in the genome. Each row represents a locus in the genome, and the two columns represent the paired chromosomes. The binary value $L_{i,j}$ indicates whether the allele in locus i of chromosome j is desirable ($L_{i,j} = 1$) or undesirable ($L_{i,j} = 0$).

Definition 1. *We define the* `Gamete` *function,* $g = \texttt{Gamete}(L, J)$*, as follows. Its input parameters include a binary matrix* $L \in \mathbb{B}^{N \times 2}$ *and a binary vector* $J \in \mathbb{B}^N$*. Its output is a binary vector* $g \in \mathbb{B}^N$*, which is determined as* $g_i = L_{i,J_i+1}, \forall i \in \{1, ..., N\}$*.*

In this definition, L represents the genotype of an individual plant, and the binary vector J indicates the sources of inheritance for the alleles in a gamete. If $J_i = 0$, then the g_i allele is inherited from $L_{i,1}$; otherwise it originates from $L_{i,2}$. To realistically represent the actual gamete formation process, the input binary vector J must be a random one following a special distribution, which is defined as follows.

Definition 2. *We say that the random binary vector $J \in \mathbb{B}^N$ follows an* inheritance distribution *with parameter $r \in [0, 0.5]^{N-1}$ if*

$$J_1 = \left\{ \begin{array}{lll} \mathit{0} & w.p. & 0.5 \\ \mathit{1} & w.p. & 0.5 \end{array} \right. , \tag{2.1}$$

$$J_i = \left\{ \begin{array}{ccc} J_{i-1} & w.p. & 1 - r_{i-1} \\ 1 - J_{i-1} & w.p. & r_{i-1} \end{array} \right. , \forall i \in \{2, ..., N\}. \tag{2.2}$$

According to Mendel's second law, $L_{1,1}$ and $L_{1,2}$ are equally likely to transmit g_1, hence Equation (4.1). Given the inheritance source of allele $(i-1)$ in the gamete, the probability that allele i comes from the same chromosome $(J_i = J_{i-1})$ is $1 - r_{i-1}$, which explains Equation (4.2).

Definition 3. *We define the* `Reproduce` *function, $X = \texttt{Reproduce}(L^1, L^2, r, K)$, as follows. Its input parameters include two binary matrices $L^1, L^2 \in \mathbb{B}^{N\times 2}$, a vector $r \in [0, 0.5]^{N-1}$, and a positive integer number K. Its output is a three-dimensional matrix $X \in \mathbb{B}^{N\times 2\times K}$, representing a population of K progeny, which is determined by first generating $2K$ independent and identically distributed random vectors from the inheritance distribution with parameter r, denoted as $J_p, \forall p \in \{1, ..., 2K\}$, and then setting $X_{i,j,k} = \texttt{Gamete}_i(L^j, J_{2k-2+j}), \forall i \in \{1, ..., N\}, j \in \{1, 2\}, k \in \{1, ..., K\}$.*

Definition 4. *The* `Select` *function, $[k_1, k_2] = \texttt{Select}(X, r)$, as follows. Its input parameters include a three-dimensional binary matrix, $X \in \mathbb{B}^{N\times 2\times K}$, and a vector $r \in [0, 0.5]^{N-1}$. Its output includes two integers, $k_1, k_2 \in \mathbb{Z}$*

Here, k_1 and k_2 are the indices of the selected parents in the breeding population X. If $k_1 = k_2$, then self-pollination is suggested as the breeding strategy.

Definition 5. *We define the* `Breed` *function as $G = \texttt{Breed}(P^0, r, K)$. Its input parameters include a three-dimensional binary matrix $P^0 \in \mathbb{B}^{N\times 2\times 2}$, a vector $r \in [0, 0.5]^{N-1}$, and a positive integer K. Its output, G, is the number of generations it takes to successfully finish the process, which is determined through the following steps.*

```
Step 0 (Initialization) Set $t = 0$ and go to Step 1.
Step 1 (Evaluation)
If $\max_k \left\{ \sum_{i=1}^{N} (P^t_{i,1,k} + P^t_{i,2,k}) \right\} = 2N$
    RETURN: $G = t$.
Else Go to Step 2.
Step 2 (Selection) Obtain $[k^t_1, k^t_2] = \texttt{Select}(P^t, r)$ and go to step 3.
Step 3 (Reproduction) Obtain $P^{t+1} =$
$\texttt{Reproduce}(P^t_{:,:,k^t_1}, P^t_{:,:,k^t_2}, r, K)$, update $t \leftarrow t + 1$, and go to Step 1.
```

The function $\texttt{Breed}(P^0, r, K)$ is a mathematical formulation of the multi-allelic introgression process, in which the selection step has the most significant influence on the efficiency of the process. In Section 2.4, we review existing approaches for parental selection, and then we propose a new approach in Section 3.

2.2.4 Existing approaches for parental selection

The genetic breeding value approach selects breeding parents based on the GEBV, which measures the fitness of individuals. In the context of multi-allelic introgression, if we assume uniform weight for all desirable alleles, then the GEBV of an individual L is equivalent to the number of desirable alleles:

$$\sum_{i=1}^{N} (L_{i,1} + L_{i,2}). \tag{2.3}$$

The two individuals with the highest GEBV will be selected according to the GEBV approach.

The optimal haploid value approach (Daetwyler et al., 2015) defined a different metric for parental selection. This approach recognizes that meiosis can produce gametes with recombined haplotype loci. The OHV of an individual can be defined as the fitness of the best doubled-haploid progeny that could possibly be produced by selfing such individual. As such, OHV measures the potential of fitness of the individual's progeny. In the context of multi-allelic introgression, the OHV of an individual L is defined as:

$$\sum_{i=1}^{N} 2 \max\{L_{i,1}, L_{i,2}\}. \tag{2.4}$$

The two individuals with the highest OHV will be selected according to the optimal haploid value approach.

2.3 PCV for parental selection

We propose a new parental selection approach using the predicted cross value, which is defined as follows.

2.3.1 Definition of PCV

Let $L^1, L^2 \in \mathbb{B}^{N\times 2}$ denote two breeding individuals, and let $[g^1, g^2]$ denote a random progeny of theirs, where $g^1 = \texttt{Gamete}(L^1, J^1)$ and $g^2 = \texttt{Gamete}(L^2, J^2)$ are random gametes produced by L^1 and L^2, respectively. When the progeny $[g^1, g^2]$ is crossed with another individual (or itself) in the next generation, it will produce a random gamete, which we denote as $g^3 = \texttt{Gamete}([g^1, g^2], J^3)$. Here J^1, J^2, and J^3 are three independent and identically distributed random vectors following the inheritance distribution with parameter r.

Definition 6. *For a given pair of individuals L^1 and L^2, the predicted cross value is defined as the probability that a random gamete, g^3, produced by a random progeny from crossing these two individuals will consist only of desirable alleles:*

$$\mathit{PCV}(L^1, L^2, r) = P(g_i^3 = 1, \forall i \in \{1, ...N\}).$$

Here, r is the recombination frequency vector.

The rationale for the PCV definition is to calculate the probability that none of the undesirable alleles survives two generations of meiosis. The essence of this approach is to select breeding parents based on their likelihood to produce an ideal gamete by combining their desirable alleles rather than the fitness of the two breeding parents themselves.

2.3.2 The water pipe algorithm for calculating PCV

We designed a polynomial time algorithm for calculating PCV, which draws an analogy between conditional probabilities and water flows through a plumbing system. The plumbing system consists of N rows and four columns of valves and a number of water pipes connecting them. The $4N$ valves correspond to the $4N$ alleles in the two breeding parents represented by the matrix $[L^1, L^2]$. For notational convenience, we will use $L \in \mathbb{B}^{N\times 4}$ to denote the matrix $[L^1, L^2]$, so $L_{i,1} = L^1_{i,1}$, $L_{i,2} = L^1_{i,2}$, $L_{i,3} = L^2_{i,1}$, and $L_{i,4} = L^2_{i,2}$ for all $i \in \{1, ..., N\}$. The intake on the top splits into four pipes with equal volumes leading to the four valves in the first row. Except for the four in the last row, each valve is connected by four pipes to the four valves in the next row. For all $i \in \{1, ..., N\}$ and $j \in \{1, 2, 3, 4\}$, if allele (i, j) is desirable, then the valve (i, j) is open, and all the water that flows into the valve from above gets redistributed into the immediate downstream pipes according to their relative volumes and goes down to the next row; but if the allele (i, j) is undesirable, then the valve (i, j) is closed, and no matter how much water flows into the valve from above, the water is retained there, neither passing further down nor going back up. For all $i \in \{1, ..., N-1\}$, $j \in \{1, 2, 3, 4\}$, and $k \in \{1, 2, 3, 4\}$, the volume of the pipe that connects valves (i, k) and $(i+1, j)$ is denoted as $T_{k,j,i}$, where T is a three-dimensional matrix, which is referred to as the *transition matrix* and defined as follows.

Definition 7. *For a given vector of recombination frequencies, $r \in [0, 0.5]^{N-1}$, the transition matrix $T \in [0, 0.5]^{4\times 4\times(N-1)}$ is defined as*

$$T_{:,:,i} = \begin{bmatrix} (1-r_i)^2 & r_i(1-r_i) & 0.5r_i & 0.5r_i \\ r_i(1-r_i) & (1-r_i)^2 & 0.5r_i & 0.5r_i \\ 0.5r_i & 0.5r_i & (1-r_i)^2 & r_i(1-r_i) \\ 0.5r_i & 0.5r_i & r_i(1-r_i) & (1-r_i)^2 \end{bmatrix}, \quad \forall i \in \{1, ..., N-1\}. \tag{2.5}$$

We define the water matrix $W \in [0, 1]^{N\times 4}$ to represent the amounts of water flowing inside the plumbing system. For all $i \in \{1, ..., N\}$ and $j \in \{1, 2, 3, 4\}$, $W_{i,j}$ represents the amount of water that flows out of the jth valve in the ith row. This value can be interpreted as the probability that the first i alleles in the random gamete g^3 are desirable and that the ith allele is inherited from the jth chromosome of the breeding parents.

Definition 8. *We define the* water matrix $W \in [0, 1]^{N\times 4}$ *as*

$$W_{i,j} = P(g_1 = ... = g_i = 1, g_i = L_{i,j}), \forall i \in \{1, ..., N\}, j \in \{1, 2, 3, 4\}. \tag{2.6}$$

Proposition 1. *The water matrix can be calculated as follows.*

$$W_{1,j} = \frac{1}{4} L_{1,j}, \forall j \in \{1,2,3,4\}; \tag{2.7}$$

$$W_{i,j} = L_{i,j} \sum_{k=1}^{4} T_{k,j,i-1} W_{i-1,k}, \forall i \in \{2,...,N\}, j \in \{1,2,3,4\}. \tag{2.8}$$

Proposition 2. *The PCV is the summation of the last row in the water matrix:*

$$\texttt{PCV}(L^1, L^2, r) = \sum_{j=1}^{4} W_{N,j}. \tag{2.9}$$

The proofs for Propositions 1 and 2 can be found in the appendix.

2.3.3 Illustrative example

We illustrate the plumbing system with the following example.

Example 1. *The two breeding parents are both ideal lines* $L^1 = L^2 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$ *and the recombination frequencies vector is* $r = \begin{bmatrix} 0.2 & 0.35 & 0.3 & 0.4 & 0.25 \end{bmatrix}^\top$.

The plumbing system corresponding to Example 1 is illustrated in Figure 2.2. The black rectangles are the valves, with binary numbers indicating whether they are open (1) or closed (0). The blue parallelograms are the water pipes, whose widths represent their relative volumes and not necessarily the actual amounts of water flowing through (they are equal only when both breeding parents are ideal lines, as in Example 1). Since both breeding parents are already ideal lines, their PCV is by definition equal to 1. Albeit trivial, this fact is verified by the plumbing system in Figure 2.2, where all the valves are open, and thus 100% of the water that is poured in will get its way out.

We now illustrate the water pipe algorithm for calculating the PCV of the following example.

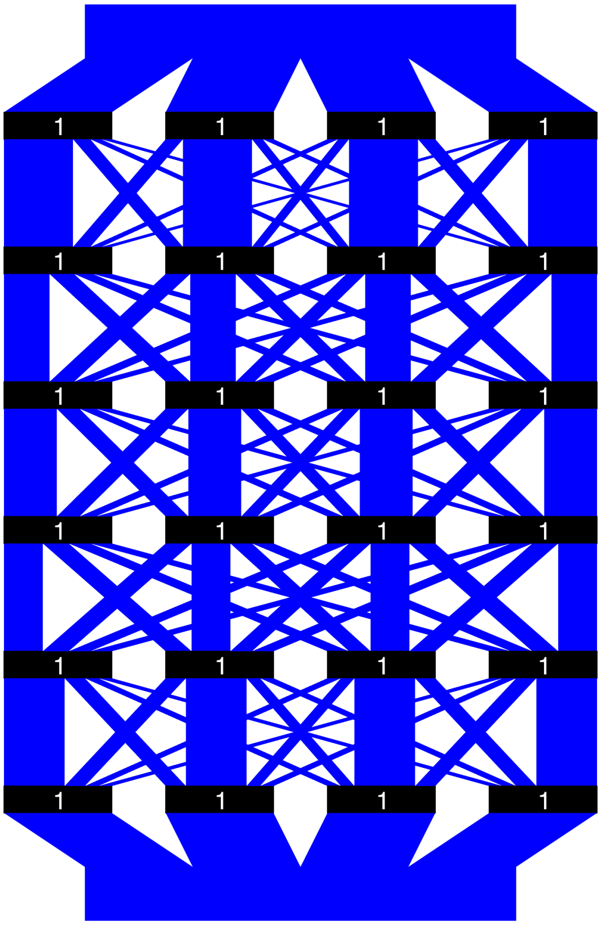


Figure 2.2: Illustration of the plumbing system for Example 1.

Example 2. *The two breeding parents are* $L^1 = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}$ *and* $L^2 = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$ *and the recombination frequencies vector is the same as in Example 1.*

The plumbing system corresponding to Example 2 is illustrated in Figure 2.3, in which we removed those water pipes whose immediate upstream valves are closed. The transition matrix

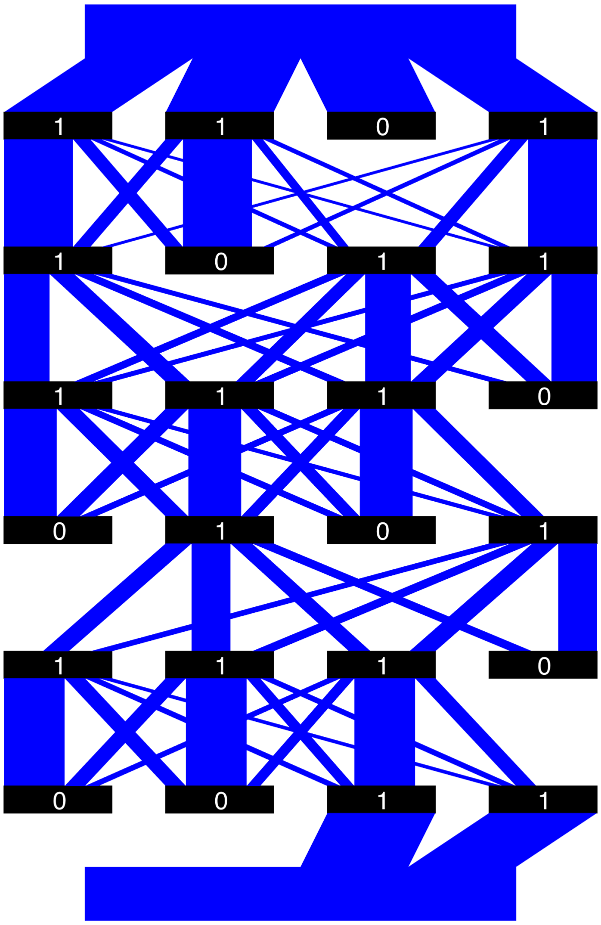


Figure 2.3: Illustration of the plumbing system for Example 2.

is $$T_{:,:,1} = \begin{bmatrix} 0.64 & 0.16 & 0.10 & 0.10 \\ 0.16 & 0.64 & 0.10 & 0.10 \\ 0.10 & 0.10 & 0.64 & 0.16 \\ 0.10 & 0.10 & 0.16 & 0.64 \end{bmatrix},\ T_{:,:,2} = \begin{bmatrix} 0.4225 & 0.2275 & 0.1750 & 0.1750 \\ 0.2275 & 0.4225 & 0.1750 & 0.1750 \\ 0.1750 & 0.1750 & 0.4225 & 0.2275 \\ 0.1750 & 0.1750 & 0.2275 & 0.4225 \end{bmatrix},\ T_{:,:,3} =$$

$$\begin{bmatrix} 0.49 & 0.21 & 0.15 & 0.15 \\ 0.21 & 0.49 & 0.15 & 0.15 \\ 0.15 & 0.15 & 0.49 & 0.21 \\ 0.15 & 0.15 & 0.21 & 0.49 \end{bmatrix}, T_{:,:,4} = \begin{bmatrix} 0.36 & 0.24 & 0.20 & 0.20 \\ 0.24 & 0.36 & 0.20 & 0.20 \\ 0.20 & 0.20 & 0.36 & 0.24 \\ 0.20 & 0.20 & 0.24 & 0.36 \end{bmatrix}, T_{:,:,5} = \begin{bmatrix} 0.5625 & 0.1875 & 0.1250 & 0.1250 \\ 0.1875 & 0.5625 & 0.1250 & 0.1250 \\ 0.1250 & 0.1250 & 0.5625 & 0.1875 \\ 0.1250 & 0.1250 & 0.1875 & 0.5625 \end{bmatrix}$$

and the water matrix is $W = \begin{bmatrix} 0.2500 & 0.2500 & 0 & 0.2500 \\ 0.2250 & 0 & 0.0900 & 0.2100 \\ 0.1476 & 0.1037 & 0.1252 & 0 \\ 0 & 0.1006 & 0 & 0.0640 \\ 0.0369 & 0.0490 & 0.0355 & 0 \\ 0 & 0 & 0.0307 & 0.0174 \end{bmatrix}$. Therefore, the PCV is $0 + 0 + 0.0307 + 0.0174 = 0.0481$.

2.4 Conceptual distinctions of PCV, GBV, and OHV

The fundamental difference between PCV and the two existing approaches, GEBV and OHV, lies on the rejection or acceptance of the assumption that the fitness of the progeny is an additive function of the fitness of the two parents. The GEBV and OHV approaches accept the additive assumption and select two individuals with the highest fitness measures as breeding parents. In contrast, the PCV approach rejects the additive assumption and proposes to select two individuals that have the highest probability to produce ideal offsprings that inherit desirable alleles from both parents.

We use the following simple example to demonstrate the conceptual distinctions of these three approaches.

Example 3. *Consider 10 loci of interest in a population of 50 individuals. Rather than describing the genotypes of this population using a three-dimensional binary matrix defined in Section 2.3, we illustrate the information in Figure 2.4. A black square is used to denote a "0" allele and a gray square for a "1". All the individuals in the sample of progeny are displayed abreast, so the figure contains a matrix of 10 by 100 black-and-gray squares. Two sightly different shades of gray are used to group together chromosome pairs that belong to the same individual. We will refer to the ith individual from the left as individual i. Then individuals 1,* *2, 3, 5, and 18 can be represented, respectively, by*

$$\begin{bmatrix} 0 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}, \textit{ and } \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

The recombination frequencies vector used in this example is

$$r = \begin{bmatrix} 0.1 & 0.2 & 0.1 & 0.2 & 0.1 & 0.2 & 0.1 & 0.2 & 0.1 \end{bmatrix}^\top.$$

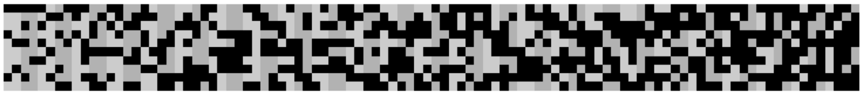


Figure 2.4: An illustration of 10 loci in a population consisting of 50 individuals.

The 50 individuals are ordered from left to right with a decreasing number of total desirable alleles, from 14 for individual 1 to 5 for individual 50.

2.4.1 Solving Example 3 using the GBV approach

The GEBVs of the 50 individuals are calculated using Equation (2.3) and plotted in Figure 2.5.

The GEBV approach would select two individuals with the largest GEBVs, i.e., individuals 1 and 2, both with 14 desirable alleles. A limitation of this approach is that it compromises long-term potential for short-term gains. In this example, crossing individuals 1 and 2 will produce a homozygous first locus with undesirable alleles, eliminating the possibility of accumulating desirable alleles at this locus in subsequent generations.

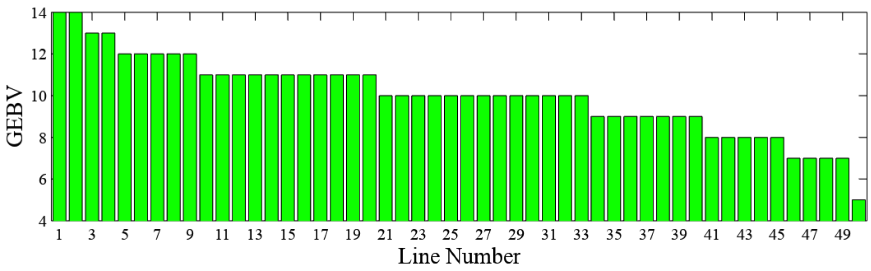


Figure 2.5: The GEBVs for Example 3.

2.4.2 Solving Example 3 using the OHV approach

The OHV of the 50 individuals are calculated using Equation (2.4) and plotted in Figure 2.6.

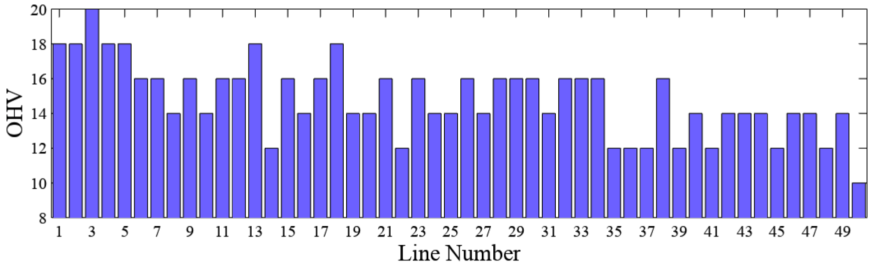


Figure 2.6: The OHVs for Example 3.

The OHV approach would select two individuals with the largest OHVs. Individual 3 has the largest OHV, whereas individuals 1, 2, 4, 5, 13, and 18 tie for the second place. A limitation of the OHV is the exclusive emphasis on the *possibility* without consideration of its *probability*. As such, the approach is unable to differentiate the six individuals with the same OHV based on their different likelihoods of combining nine desirable alleles into one gamete.

2.4.3 Solving Example 3 using the PCV approach

The PCVs of the 50 individuals are calculated using Equations (2.5) and (4.6)-(4.8) and plotted in Figure 2.7. The two subfigures at the top and left are the same population from Figure 2.4 with horizontal and vertical orientations, respectively. The largest subfigure is a PCV map. It consists of a 50 × 50 gray-shade matrix representing all PCV values for all possible pairs of breeding parents involving the 50 individuals in the population. As such, each

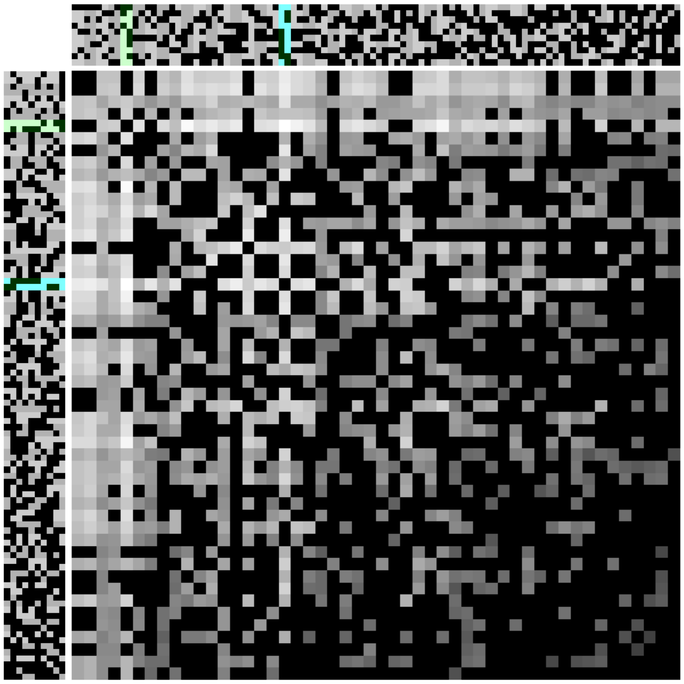


Figure 2.7: The PCV map for Example 3.

square representing a PCV value has an area four times as large as the one that represents an allele in the horizontal and vertical subfigures. The brightness indicates the PCV for the two individuals directly above and to the left. The brighter the color in the PCV map, the higher the PCV. We point out four observations. (1) The PCV is not an additive function of two individuals. (2) The PCV map is symmetric across the diagonal, since the order of the two parents does not matter in the definition of PCV. (3) The diagonal represents PCVs for self-pollination. (4) The highest PCV is achieved by individuals 5 and 18, which are respectively highlighted in light green and light blue on the margins. These two individuals should be selected according to the PCV approach. Appendix B discusses two approaches that can be used to select the pair of individuals with the highest PCV from a population.

Compared with the GEBV and OHV, PCV has two salient features. First, PCV evaluates each specific cross. In contrast, the GEBV and OHV calculate an estimated breeding value for each individual. In the context of mating designs, breeding values are analogous to general combining ability, whereas the PCV is analogous to specific combining ability. Second, the PCV integrates recombination frequencies to calculate conditional probabilities. In Example

3, out of the 1275 possible crosses, 711 have a zero PCV value because at least one locus will become homozygous for the undesirable allele. The remaining 564 combinations have a unique PCV. Therefore, the PCV map in Figure 2.7 has 565 different shades of gray. In contrast, there is a large number of tied GEBVs and OHVs in the example.

2.5 Simulation experiments

In this section, we describe and report results of simulated multi-allelic introgression experiments using the PCV, GEBV, and OHV approaches.

2.5.1 Experiment description

We simulated a polygenic trait consisting of 100 QTL that are responsible for genetic variability in the trait. The locations of the QTL are distributed as uniform random variable among ten simulated linkage groups. Each linkage group has from 8 to 12 QTL.

We considered two example trait introgression projects. In both examples the recipient and donor are homozygous at all QTL. In the first example, the recipient has favorable alleles at 93 of the QTL, while the donor has favorable alleles at the remaining 7. For reference the recipient has undesirable alleles at QTL numbered: 4, 6, 19, 20, 44, 53, and 58. In the second example the recipient has favorable alleles at 80 of the loci, while the donor has favorable alleles at the remaining 20. For reference the recipient has undesirable alleles at QTL numbered: 5, 10, 14, 19, 24, 29, 33, 38, 43, 48, 52, 57, 62, 67, 71, 76, 81, 86, 90, and 95. We use two binary matrices $L^1 \in \mathcal{B}^{100\times 2}$ and $L^2 \in \mathcal{B}^{100\times 2}$ to represent respective genomes of recipient and donor, respectively.

Recombination frequencies between linkage groups are set to be 0.5 while recombination between QTL within linkage groups were randomly generated. The aggregated vector of recombination frequencies is denoted as $r \in [0, 0.5]^{99\times 1}$ and plotted in Figure 2.8. Note that the figure shows recombination between linkage groups as terminal QTL on each linkage group: 10, 19, 30, 40, 50, 62, 73, 81, and 92.

We implemented the three iterative steps in the `Breed` function to simulate the introgression project, with the simulated genomes as the initial population for each example. In subsequent

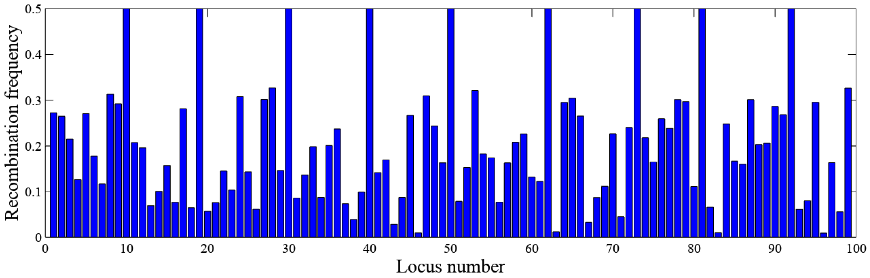


Figure 2.8: The recombination frequencies for the simulation between adjacent pairs of loci.

generations, 100 progeny were sampled from simulated crosses of two individuals selected from the previous generation. The recipient line is treated as a member of the sample so that backcrossing is always an option.

Within the Breed function we compared four selection approaches.

- The GEBV approach, which selects two different individuals with the highest GEBVs.
- The OHV approach, which selects two different individuals with the highest OHVs.
- The PCV-I approach, which selects two different individuals with the highest PCV.
- The PCV-II approach, which selects one (for self pollination) or two (for cross pollination) individuals with the highest PCV.

One thousand simulation runs were carried out, and the comparison was based on the time and probability of success, i.e., number of generations to completely introgress all desirable donor alleles. The simulation was implemented in Octave (Eaton et al., 2015).

2.5.2 Results for Example 1

Figure 2.9 plots the histograms of the normalized breeding values, which is the proportion of desirable alleles in the genome, of the populations over time. For all selection approaches the sample representing the first generation consists of the recipient line (with 93% of the desirable alleles), the donor line (with a 7% of the desirable alleles), and 98 F1 lines consisting of half of the alleles from the recipient and donor (with 50% of the desirable alleles). The histogram for progeny in generation 2 is the same for GEBV, PCV-I, and PCV-II because the recipient parent was crossed to the F1 for these selection approaches. On the other hand the histogram for the OHV approach is represented by a Doubled Haploid sample from the F1. If the GEBV or OHV approaches are used, the normalized breeding value will plateau at 95% and 93%, respectively. If either PCV-I or PCV-II approaches is used, an ideal progeny will have been produced in as early as the 7th generations and no later than the 11th.

Figure 2.10 compares the probability distributions of the terminal generation for PCV-I and PCV-II approaches. On average, the PCV-I approach takes 9.4 generations to produce an ideal progeny, whereas PCV-II takes 8.9 generations. Thus, allowing selfing during the breeding process increased the efficiency of the project by half a generation in this example.

2.5.3 Results for Example 2

Figures 2.11 and 2.12 reveal similar results of the four selection approaches as Figures 2.9 and 2.10, but as expected all approaches take more generations to plateau. Using the GEBV and OHV approaches, the normalized breeding values plateau at 86% and 90%, respectively. Out of the 1,000 simulation repetitions, both PCV-I and PCV-II approaches successfully produced an ideal progeny 996 times, taking an average of 14.8 and 14.7 generations, respectively. In the other four times, the trait introgression project failed by having at least one locus become homozygous with undesirable alleles for all individuals in the population. The GEBV and OHV approaches failed in the same way in all 1,000 simulation runs in both experiments.

2.6 Discussion

The formulation of the multi-allelic introgression problem captured the mathematical essence of the process and we hope that it will attract other operations researchers, applied mathematicians, and computational scientists to contribute to genetic improvement projects with more efficient algorithms. Using time (generations) and probability of success as criteria provides objective measurable criteria for comparing breeding strategies. Missing from these criteria is a consideration of cost. In general, the number of progeny evaluated every generation can serve as a surrogate for cost and in future research we will look at the relative impacts of sample size for each generation of evaluation. While these costs are relatively easy to quantify, considerable thought will be needed to formulate either social or commercial costs associated with slower introgression of alleles of economically important traits.

The PCV is a new metric for selection of parents. Rather than sticking to predetermined breeding strategies such as backcrossing, as widely used for trait introgression, PCV based selection identifies the pair of individuals whose complementary genotypes have the highest probability to yield an ideal gamete in two generations. The simulation results demonstrated that the PCV outperforms the existing approaches GEBV and OHV.

Applicability of our approach is limited by a number of simplifying assumptions summarized in Section 2.2. Relaxing these will provide potentially fruitful topics for future research. For example, a similar but more sophisticated definition of the PCV could be designed for autopolyploid perennial crops such as alfalfa. Also, if desirable alleles of interest are carried by multiple donors, then modifications are required to extend the PCV. Two approaches have been proposed for introgression of multiple alleles from multiple donors. One is to sequentially introgress alleles from each donor, and the other is to stack all their desirable alleles into single donor line (Peng et al., 2014a,b). A couple of optimization approaches have been proposed for the gene stacking problem (Canzar and El-Kebir, 2011; Xu et al., 2011), which has been proved to be NP-hard (Xu et al., 2011). It would be a challenging but useful extension to design PCV based breeding strategies for multiple donors. The selection of more than one pair of parent lines must be coordinated to not only produce enough seeds to allow for critical recombinations to occur but also expedite the integration of all desirable alleles into the recipient cultivar(s).

Another direction that deserves investigation in future research is the exploration of more optimal breeding strategies. The trait introgression breeding problem formulated in Section 2, even with the simplifying assumptions, is too complex to be readily solvable by existing optimization methodology. Although the PCV based multi-allelic introgression outperforms those based on breeding values, it is unclear to us how much further improvement could be made. A starting point could be to dynamically adjust the number of individuals evaluated every generation and the selection approach in each generation in response to the outcome of the previous cross.

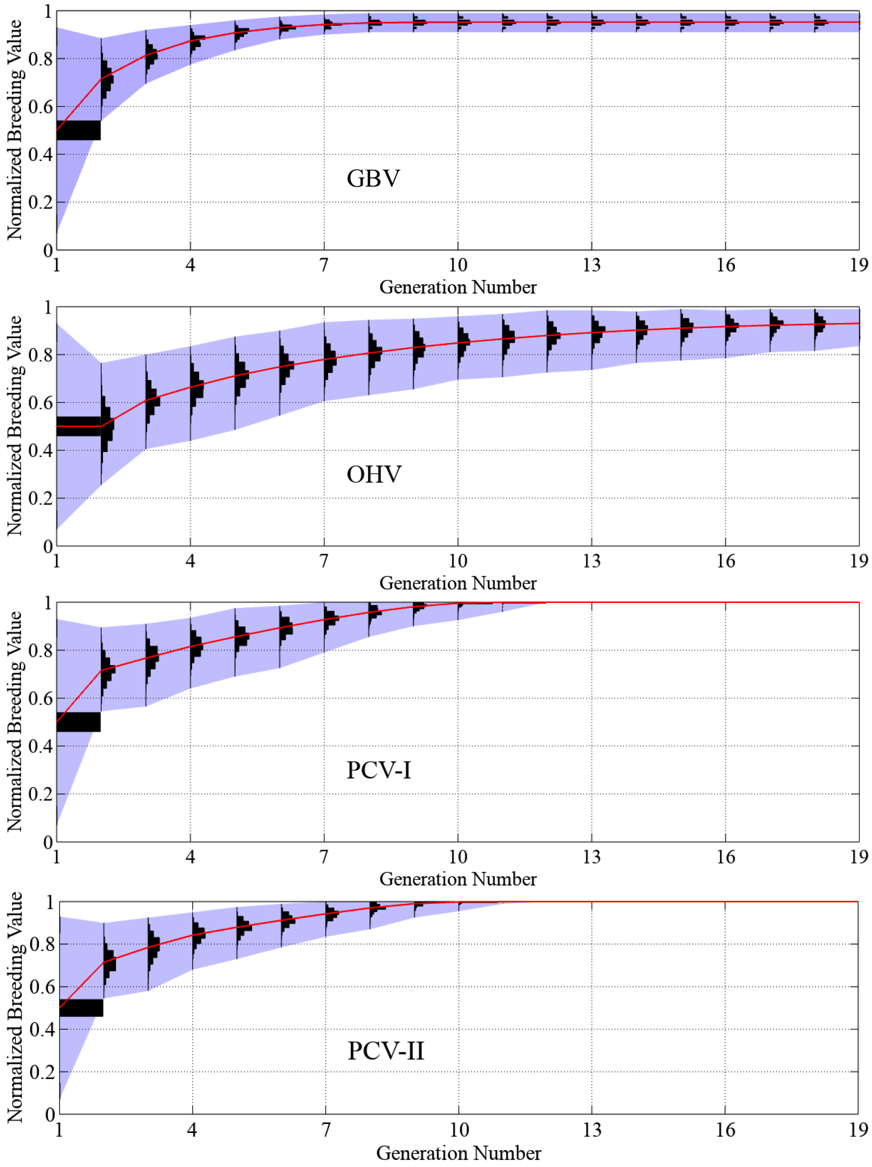


Figure 2.9: Performance of the GEBV, OHV, PCV-I, and PCV-II approaches in 1000 simulation runs of trait introgression of seven QTL. The vertical axis represents the proportion of desirable alleles in the genome. Histograms of the proportion of desirable alleles among 100 progeny from 1,000 simulation runs are plotted for each generation. The red curve shows the population mean.

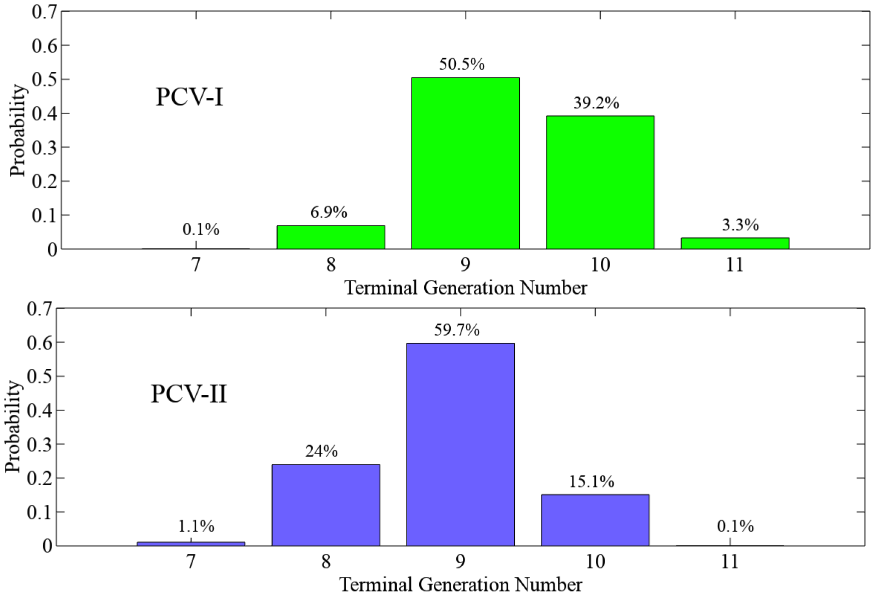


Figure 2.10: Distributions of the terminal generation numbers of PCV-I and PCV-II approaches.

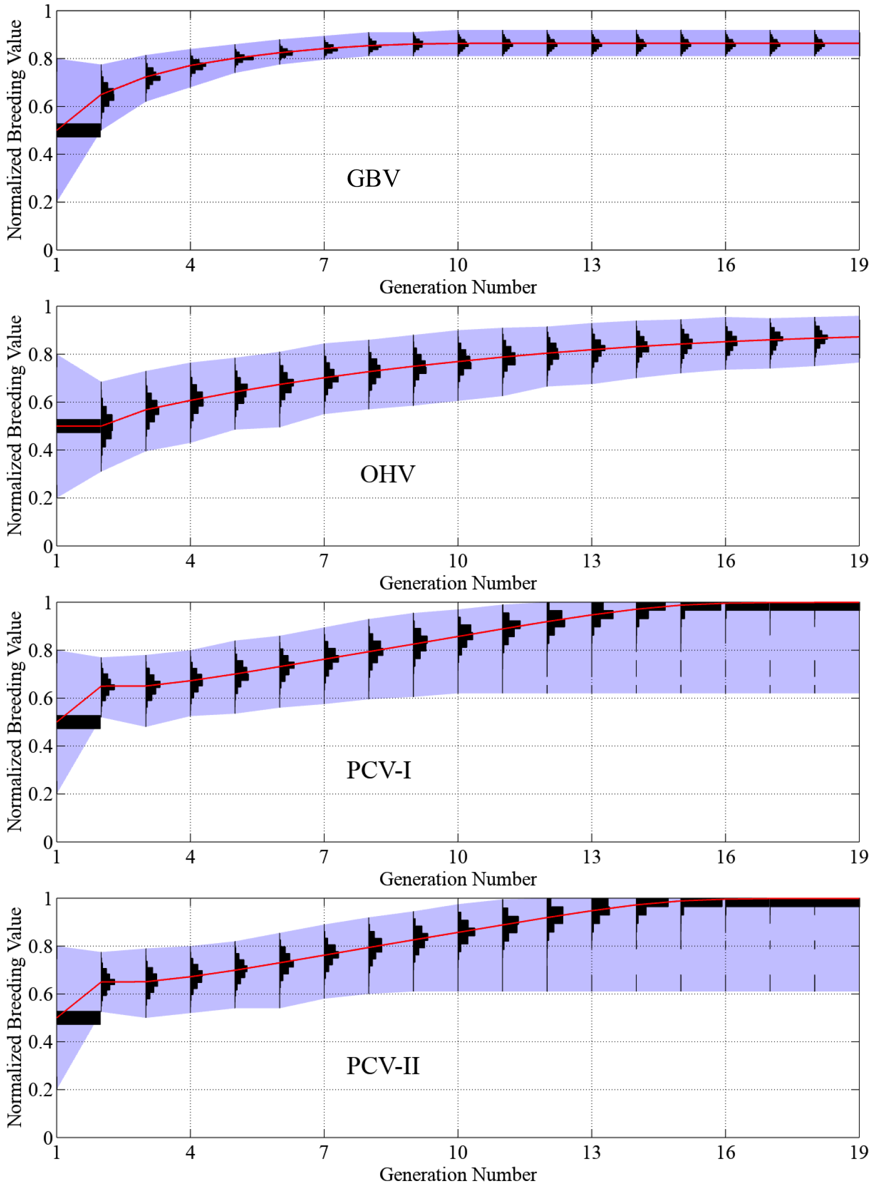


Figure 2.11: Performance of the GS, OHV, PCV-I, and PCV-II approaches in 1000 simulation runs of trait introgression of twenty QTL.

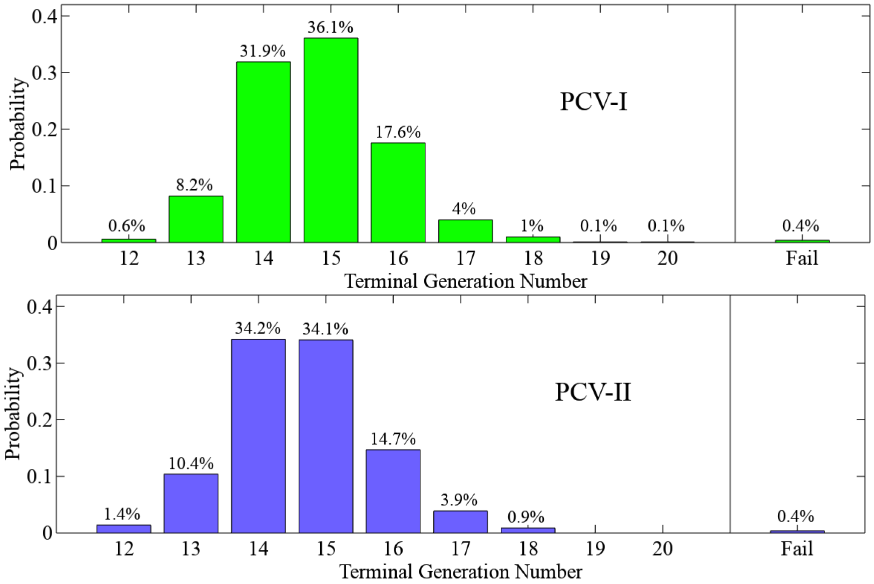


Figure 2.12: Distributions of the terminal generation numbers of PCV-I and PCV-II approaches.

CHAPTER 3. Dynamic Programming for Resource Allocation in Multi-allelic Trait Introgressions

Abstract

Introgression is a hybridization process that plant breeding companies in the agriculture industry use to transfer desirable alleles from one crop variety to another. This research addresses the resource allocation problem in this process, which is iterative and dynamic. Rather than using the conventional evenly allocating approach, we try to improve the resource allocation strategically based on the outcome from the previous generation. The methodology we use to solve the problem is widely used in the literature to solve many other resource allocation problems in different industries. The problem was motivated from collaborations with scientists in the seed companies, and the results present an alternative operational strategy that has been shown to be superior in silico with the potential to lead to significant savings to the industry. We formulated the resource allocation problem as a Markov decision processes (MDP) model and used backward induction to obtain optimal resource allocation strategies. We tested the methodology using the same data set from a case study in the literature, in which the conventional resources allocation approach was used. Simulation results suggest that the dynamic resource allocation approach from the MDP model significantly outperformed the conventional approach by reducing the average cost and time and improving the probability to successfully complete the introgression process. Results from intermediate generations in an introgression project, although requiring knowledge in plant genetics to interpret, contain valuable insight on how much resources are necessary to make expected progress in subsequent generations. Allocating more or less than the necessary amount would lead to either waste of resources or reduction in the probability of success. The proposed model provides plant breeders with a

more efficient approach to resource allocation to produce better seed varieties with the highest probability of success.

3.1 Introduction

3.1.1 Backgrounds

Plant breeding has been defined as the art and science of producing desired characteristics through artificial selection (Poehlman, 2013). Practiced since the beginning of civilizations, plant breeders in the 20th century made enormous changes to important agronomic traits, e.g., grain yield and pest resistance, of cereal crops (Duvick, 1994; Rincker et al., 2014). This was accomplished through ad hoc adoptions of emerging technologies developed by agricultural, mechanical, electrical and information engineers. In the 21st century, demands for increasing production of food, fiber and energy with less water, fuel and fertilizer will force plant breeding to become more efficient and effective. According to the USDA Long-Term Agricultural Projection Tables (USDA, 2017), the U.S. soybeans harvested area will decrease from 82.6 million acres in 2014/15 to 79.7 million acres in 2025/2026. However, despite the decreasing area for harvest, the U.S. soybeans export amount will need to increase from 50.2 million metric tons in 2014/15 to 52.4 million metric tons in 2025/26. In order to satisfy the increasing demand, the yield of soybeans has to increase from 47.5 bushels per harvested acre in 2014/15 to 51 bushels per harvested acre in 2025/26. Thus, more advanced and efficient plant breeding techniques are highly required for sustainable improvement and development.

Plant breeding in the future will require identification and rapid deployment of desirable physical attributes, also known as phenotypes, that will enable crops to predictably adapt to rapidly changing environments. Methods for discovery of genetic variants associated with phenotypic variants have been developed over the last 25 years and are now routinely applied using ‘omics’ technologies in forward and reverse genetics approaches. Because the genetic variants (alleles) associated with phenotypic variants are distributed unevenly throughout germplasm collections and breeding populations the challenge is to aggregate favorable alleles into improved cultivars. The transfer of a single desirable allele from an otherwise inferior cultivar to an otherwise superior cultivar is routinely accomplished using marker assisted breeding strategies (Visscher et al., 1996; Frisch et al., 1999; Frisch and Melchinger, 2005; Peng et al., 2014a) although recently Cameron et al. (2017) demonstrated that the efficiency of these routine processes can be doubled by reframing the objective using principles from operations research.

The more complex challenge of aggregating sets consisting of multiple alleles into cultivars with predictable adaptive trait phenotypes will require transfer of knowledge from operations researchers and mathematicians to plant breeders. This issue can be addressed by developing an improved breeding strategy to rapidly transfer multiple desirable genetic alleles from a donor individual to an elite recipient individual. In the vernacular of the plant breeder, this is known as trait introgression involving multiple alleles or multi-allelic trait introgression (MATI) process.

The MATI process can be regarded as a decision making system, of which the components are in uncertain states due to the stochastic nature of gene reconstruction during crop mating. In the process, the decision maker or plant breeder has the obligation to obtain the available genotypic and phenotypic information, decide parents to breed, allocate resources and fulfill goals of the breeding cycle. In order to accurately depict this decision making system and optimize the MATI process, mathematical transformations and formulations have been proposed to frame the MATI process as an engineering system (Han et al., 2017). An algorithmic process with mathematical definitions was designed for simulation, as well. In the paper, parental selection was addressed as a key procedure, which can affect the result dramatically. A new metric called the Predicted Cross Value (PCV) with the assistance of genetic markers for parental selection was proposed in the paper. The PCV was defined as a quantification metric for any pair of selected breeding parents. With the help of the PCV, significant improvements were demonstrated as well as the great potential for further research on MATI process.

As pointed out in Han et al. (2017), in addition to parental selection, the resource allocation also plays a crucial role in improving the efficiency of the MATI process. From such point of view, in this paper we expand our discussion on the decision making problem of resource allocation for MATI and improve the breeding strategy. Because of the dynamic and uncertain states of the system, we apply the Markov decision processes (MDP) model to frame MATI processes. The MDP model is a technique for solving stochastic sequential decision making problems (Puterman, 2014). The MDP model has been proved to make contributions to various practical decision making projects, such as optimal replacement policy for a motion picture exhibitor (Swami et al., 2001) or the vehicle mix decision in emergency medical service systems (Chong et al., 2015), which share many similarities with MATI processes.

The remainder of the paper is organized as follows. In section 2, a brief introduction of the MATI process via a flowchart is proposed and an algorithmic simulating process for the MATI is discussed. The statement of the resource allocation problem for the MATI process is captured, as well. In section 3, we define the Markov decision processes model to solve the resource allocation problem. In section 4, the results from computer simulations demonstrate the advantage on a hypothetical case study. We present the results on the tradeoffs among total budget, time and probability of success for the project. Under different total budget scenarios we analyze how the budget is allocated and determine the most cost-efficient total budget, as well. We also compare the MDP model with static resource allocation breeding strategies to demonstrate the improvement. In section 5, the conclusion and future work directions are discussed.

3.1.2 MATI Process and Resources Allocation Problem

In this section, we present a flowchart to describe the work flow for the MATI process, design a mathematical algorithmic simulating process and propose the problem statement. Recently, the work flow was proposed to describe the general MATI process (Han et al., 2017) without resource constraints such as deadline or budget limit, which will affect breeding projects . Thus, new components for resource allocation are brought into the work flow to make comprehensive introduction of the MATI process. In the last part of this section, with the necessary introduction, the problem statement for the resource allocation in the MATI process is proposed.

3.1.2.1 MATI Process Introduction

The work flow for the MATI process is presented in Figure 3.1. The MATI process begins with the **“Start”** step, in which at least one elite recipient individual and one donor individual are available. In most annual crops, both elite and donor individuals are homozygous throughout their genomes. The majority of alleles in the donor are undesirable, but it does have desirable versions of alleles that the elite individual is lacking at several loci. The goal of this process is to achieve an ideal individual inheriting all the desirable alleles from both donor and elite individuals within the provided resources. After the **"Start"** step, the process involves a loop of check boxes and steps: check box **"Genotype ideal ?"**, check box **"Resource enough?"**, step **"Resource allocation"**, step **"Selection"** and step **"Reproduction"**.

We briefly review these check boxes and steps here. In the **"Genotype ideal?"** check box, the genotypic information of current progeny is screened to check if the ideal individual is produced. If the ideal individual is obtained, the entire process is considered as a **"Success"**, otherwise, the process flows to the **"Resource enough?"** check box. This step involves the resources assessment and the process continues if the remaining resources are adequate. Usually, the resource consists of budget and time. A breeding process is associated with different terms of cost, such as genotyping assays, crossing, growing the crops, and labor. Some costs are fixed, while others are proportional to the number of crosses made or progeny produced. In practice, there may be a total budget constraint for the cost through the entire breeding project. In addition to the cost, the breeding project is often bounded by a deadline, which shall be regarded as a time resource limit. In the following step **"Resource allocation"**, the decision maker needs to observe the current status of the breeding project and allocate the resources based on policies. For commercial breeding projects, there is revenue associated with the ideal individual when delivered into the market. Hence, for resource allocation, the decision maker needs to consider revenue with the cost. When the process reaches the **"Selection"** step, two breeding parents are selected based on a provided selection metric. In the **"Reproduction"** step, the selected breeding parents are mated to produce a new generation of progeny and the process flows back to the check box **"Genotype ideal?"**. In this MATI process, we assume that the breeding parents would be retained for the next one generation.

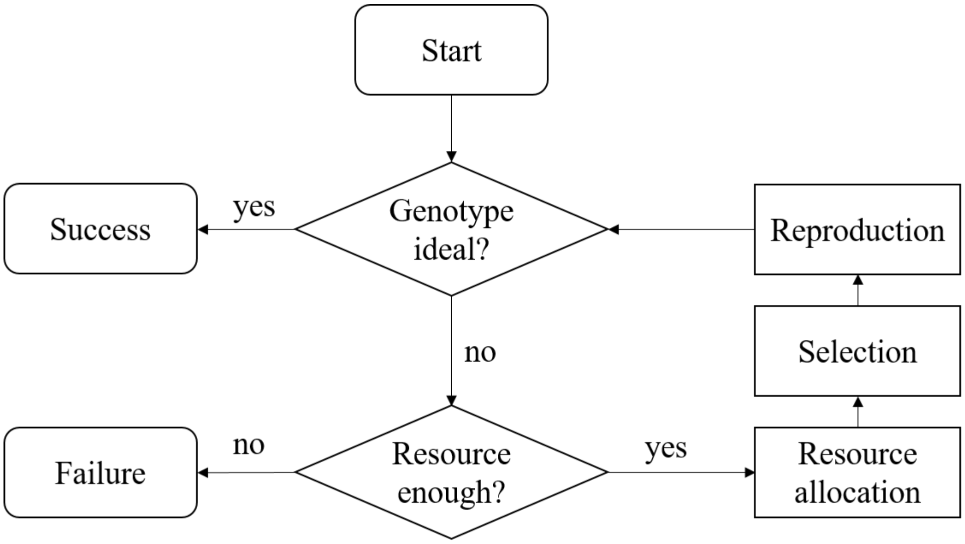


Figure 3.1: Flowchart of the MATI process

3.1.2.2 Mathematical Formulations for the MATI Process

According to the flowchart, we design a mathematical algorithmic engineering process for simulating the MATI process, in which some steps can be optimized such as **"Resource allocation"** and **"Selection"**. For the **"Selection"** step, random selection, genomic estimated breeding value (GEBV) (Meuwissen et al., 2001), optimal haploid value (OHV) (Daetwyler et al., 2015) and the newly designed predicted cross value (PCV) (Han et al., 2017) are possible metrics for determining the optimal breeding parents for the next generation. For the **"Resource allocation"** step, the remainder of the paper will discuss how to apply dynamic programming model to improve the efficiency. First, we define some major steps in the MATI process.

Definition 9. *(Han et al., 2017) "We define the* `Reproduce` *function,* $X = \texttt{Reproduce}(L^1, L^2, f, K)$*, as follows. Its input parameters include two binary matrices* $L^1, L^2 \in \mathbb{B}^{N\times 2}$*, a vector* $f \in [0, 0.5]^{N-1}$*, and a positive integer number* K*. Its output is a three-dimensional matrix* $X \in \mathbb{B}^{N\times 2\times K}$*, representing a random population of* K *progeny."*

The `Reproduce` function is defined the same way as the one in Han et al. (2017). In the definition, a binary matrix with dimension of $N \times 2$ is used to represent the genotype of a diploid individual with N loci where "0" represents undesirable alleles and "1" represents desirable alleles at each of the loci. In the function L_1 and L_2 are the selected breeding parents. The output X of the function represents the genotype of all the progeny produced by the breeding parents, whose element $X_{i,1,k}$ with $i \in \{1, 2, ..., N\}, k \in \{1, 2, ..., K\}$ represents the allele on the ith row (locus) of the first ('2' on the second dimension of X representing the second) chromosome set of the kth progeny in the population. The vector $f \in [0, 0.5]^{N-1}$ represents the recombination frequency, which reveals the inheritance characteristics of gene reconstruction. The parameter K in the function decides the number of progeny to produce.

Definition 10. *We define the* `Selection` *function,* $[k_1, k_2] = \texttt{Selection}(X)$*, as follows. Its input parameter includes a three-dimensional binary matrix* $X \in \mathbb{B}^{N \times 2 \times K}$ *representing a candidate population. Its output includes two integers,* $k_1, k_2 \in \mathbb{Z}$ *indicating the indexes of selected parents.*

Definition 11. *We define the* `Reward` *function,* $\texttt{Reward}(K, X, t, T) = \texttt{Revenue}(X, t, T) - \texttt{Cost}(K)$*, as follows. Its input parameters include a positive integer* K *representing the progeny number, a three-dimensional binary matrix* $X \in \mathbb{B}^{N \times 2 \times K}$ *representing a candidate population, a nonnegative integer* t *representing the current generation number and a nonnegative integer* T *representing a deadline. Its output is a reward consisting of the revenue from population* X *at generation* t *given deadline* T *and the cost for producing* K *progeny.*

Definition 12. *We define the* `Allocation` *function,*

$$K^t = \texttt{Allocation}(T, t, f, P^t, B^t, \texttt{Reward}),$$

as follows. Its input parameters include a positive integer T *representing the deadline, a nonnegative integer* t *representing the current generation number, a vector* $f \in [0, 0.5]^{N-1}$ *representing the recombination frequency, a three-dimensional binary matrix* $P^t \in \mathbb{B}^{N \times 2 \times K^{t-1}}$ *(*$t \geq 1$ *and* $K^0 = 2$*) representing the candidate breeding population for the current generation (produced by generation* $t - 1$*), a positive number* B^t *representing the current available budget and the* `Reward` *function. Its output* K^t *is a nonnegative integer representing the number of progeny to produce for generation* t*. Note that if* K^t *equals 0 with* $t \leq T$ *and* $B^t > 0$*, the project fails.*

With the definitions for three major steps in Flowchart 3.1, the definition for simulating the entire MATI process is proposed as follows.

Definition 13. *We define the* `MATI` *function,* $T_s = \texttt{MATI}(P^0, f, B, \texttt{Reward}, T)$*, as follows. Its input parameters include a three-dimensional binary matrix* $P^0 \in \mathbb{B}^{N\times 2\times 2}$ *representing the initial breeding population, a vector* $f \in [0, 0.5]^{N-1}$ *representing the recombination frequency, a positive integer* B *representing the total budget, a* `Reward` *function and a positive integer* T *representing the deadline. Its output* T_s*, is the number of generations the process takes to finish the breeding process, which is determined through the following steps.*

```
Step 0 (Initialization) Set $t = 0$ and go to Step 1.
Step 1 (Genotype check)
  if $\max_k \left\{ \sum_{i=1}^{N} (P^t_{i,1,k} + P^t_{i,2,k}) \right\} = 2N$ then
    return : $T_s = t$.
  end if
Step 2 (Resource check and resource allocation)
  $K^t = \texttt{Allocation}(T, t, f, P^t, B^t, \texttt{Reward})$
  if $K^t = 0$ or $t > T$ then
    return : $T_s = \infty$
  else
    Go to Step 3.
  end if
Step 3 (Selection) Obtain $[k^t_1, k^t_2] = \texttt{Selection}(P^t)$ and go to step 4.
Step 4 (Reproduction) Obtain $P^{t+1} = \texttt{Reproduction}(P^t_{:,:,k^t_1}, P^t_{:,:,k^t_2}, f, K^t)$, update $t \leftarrow t + 1$
and $B^{t+1} \leftarrow B^t - \texttt{Cost}(K^t)$, then go to Step 1.
```

3.1.2.3 Resource Allocation Problem in the MATI Process

In this section we propose the problem definition for the resource allocation step in the MATI process, which is related to designing the `Allocation` function in the `MATI` function. The resource allocation problem for the MATI process is a dynamic decision making problem. The plant breeder needs to determine how many progeny to produce according to the current generation number, the deadline, the funds remaining from the overall budget, the cost and revenue function and the available progeny at the beginning of each generation. This decision is a key factor affecting the MATI process because it affects the number of offspring produced in each generation as well as the cost and revenue. In each generation, producing more progeny can increase the cost but also the probability of obtaining a more promising genotype. The offspring's genotype and the amount of time spent on the process together determine the revenue of a project. Intuitively, the earlier a new genotypically designed product (i.e., offspring) can be delivered to the market, the more market share and revenue a company may attain. Hence, *designing the policy for resource allocation (i.e., how many progeny to produce at each generation) to maximize the expected net present value at the beginning of a breeding project* can be regarded as the general problem statement of the resource allocation problem in MATI process.

We frame the resource allocation problem as a dynamic programming problem. The state describing the status of a breeding project shall consist of genotypic indicators and the budget information. According to metrics such as GEBV or PCV, we can convert genotypic information into a number and use an interval to cover a group of progeny. Associated with the budget, the state is denoted as a combination of available budget and the metric interval for certain genotypes. By carefully designing the metric intervals, we can make the state space discrete and small enough to enumerate and cover all potential progeny genotypes.

The action that the decision maker needs to determine is the progeny number to produce at each state after selection of breeding parents has been made. This action determines the cost. Meanwhile, different actions affect the probabilities of transitioning among states, which are stored in the transition probabilites matrix. In addition, reaching a specific state at a certain generation will generate revenue. Based on the breeder's estimation, the revenue may not only be decided by the state, but also determined by the current generation number and deadline. There will be a decision policy describing a series of actions to optimize the expected revenue of the breeding project.

In such manners, with a discount factor, the objective of a breeding project can be formulated as determining the optimal policy to maximize the expected net present value in terms of rewards subjected to the deadline and budget. In mathematical formulations, the objective of this resource allocation problem can be stated as:

$$\max_\pi \mathbb{E}_s^\pi\{\textstyle\sum_{t=0}^T \lambda^t r_t(a, s, T)\}.$$

Herein, s represents the state; a represents the action; T represents the deadline; r represents the reward function; λ represents the discount factor and π represents the decision policy.

3.2 Material and Methods

The dynamic programming structure of the MATI process makes Markov decision processes (MDP) an appropriate approach for solving the stochastic decision making problem. In this section, we formulate an MDP model with finite horizon to identify the optimal resource allocation strategy, which is applied in the `Allocation` function of the described process.

3.2.1 Model Definition

In an MDP model, there are five major components including decision epochs, states, actions, trasition probabilities and rewards. The detailed notations for these components are as follows.

Decision epoches: We define the **decision epoch** as the beginning of each breeding generation, denoted as $\{1, 2, 3, ..., T\}$ and T is the deadline of a breeding project. Decisions like parental selection, resource allocation, etc., are made at each decision epoch. We assume the MATI process generally has a specified deadline, which implies that the MDP model has a finite horizon.

States: For any given sample of progeny P, we define a function $V(P)$ to measure the progress in the MATI process, which takes the values within the interval $[V(P^0), V(P^{\text{Ideal}})]$, with P^0 and P^{Ideal} denoting the original sample of progeny and a sample that includes an ideal individual (with all alleles being desirable). Various definitions of breeding values, such as GEBV, OHV and PCV, could be used for this function. Due to the enormous space of all possible samples of progenies, there is potentially a large number of possible values for V. For computational tractability, as illustrated in Figure 3.2, we group all possible V values into a small number of intervals $m_0, m_1, m_2, ..., m_{G-1}, m_G$, where G is a predetermined integer.

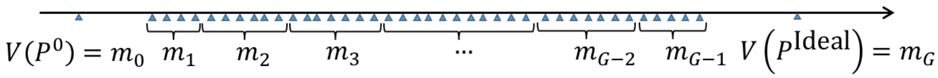


Figure 3.2: Genotype indicator

Next define the **state space** S as:

$$S = (m_g, b) \cup \{\text{failure}\} \cup \{\text{success}\}, g \in \{1, 2, ..., G-1\}, b \in \{1, 2, ..., B-1, B\},$$

where (m_g, b) is a 2-tuple. In the 2-tuple, m_g represents the metric interval indicating the genotype status and b represents the remaining budget for the breeding project. In the definition, B represents the total budget at the beginning of the process. The design of metric intervals is associated with the preference of the decision maker and shall not be fixed. We will propose one possible approach in the case study section for designing the metric interval.

Actions: The **action space** is denoted as $A = \bigcup_{s \in S} A_s = \{0, 1, 2, ..., a^{\max}\}$ representing the number of progeny to produce at each decision epoch. The maximum number of progeny that can be produced is set as $a^{\max}$ for each generation determined by reproductive biology of the plant species. In the remainder of this paper, action a is used to substitute K in the algorithmic process for `Allocation` function.

Transition Probabilities: In the MDP model, we use $W^a_{i,j}$ to denote the transition probability from interval m_i in one generation to m_j in the next generation under action a. One fact of our MDP model is that once the intervals are determined, W^a only depends on the action

a and is stationary at different epochs. According to the assumption that the breeding parents are retained to generate a new sample of progeny for the subsequent generation, the process either advances to the next interval or stays in the same one but never moves backwards, i.e., $W^a_{i,j} = 0$ if $j < i$. The matrix W^a could be estimated by simulations recording the information of action, the progeny produced at each generation and the hierarchical kinship information of mating. With the W^a matrix, we are ready to define the transition probabilities matrix, which consists of the probability of transitioning from one state s to another state s' under action a, i.e., $P_t(s'|s, a)$.

Definition 14. *Given action a, the **transition probabilities matrix** can be defined as a partitioned matrix M^a as follows:*

$$M^a =$$

$$\begin{array}{c|ccccccccccc|c}
 & S_B^\top & S_{B-1}^\top & \dots & S_{B-a}^\top & S_{B-a-1}^\top & S_{B-a-2}^\top & \dots & S_1^\top & failure & success & \\
S_B & 0 & 0 & \dots & \bar{W}^a & 0 & 0 & \dots & 0 & 0 & \hat{W}^a & \\
S_{B-1} & 0 & 0 & \dots & 0 & \bar{W}^a & 0 & \dots & 0 & 0 & \hat{W}^a & \\
S_{B-2} & 0 & 0 & \dots & 0 & 0 & \bar{W}^a & \dots & 0 & 0 & \hat{W}^a & \\
\vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \\
S_{a+1} & 0 & 0 & \dots & 0 & 0 & 0 & \dots & \bar{W}^a & 0 & \hat{W}^a & \\
S_a & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 1-\hat{W}^a & \hat{W}^a & \\
S_{a-1} & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 1 & 0 & \\
\vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \vdots \\
S_1 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 1 & 0 & \\
failure & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 1 & 0 & \\
success & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 1 &
\end{array}$$

where $\bar{W}^a = W^a_{1:G-1,1:G-1}$, $\hat{W}^a = W^a_{1:G-1,G}$ and $S_b = [(m_1, b), (m_2, b), ..., (m_{G-2}, b), (m_{G-1}, b)]^\top$ is a vector representing $G - 1$ states. Here, $P_t(s'|s, a) = M^a_{s,s'}, \forall t < T$.

Rewards: For an MDP model, the reward $r_t(s, a)$ received at epoch t is decided by the state $s \in S$ and action $a \in A_s$, which can be either positive or negative. In our MDP model for the MATI process, the **reward** is defined as $r_t(a, s, T) = -C(a) + R_t(s, T)$, where $C(a)$

is the cost function for producing a progeny and $R_t(s,T)$ is the revenue function at epoch t associated with state s and deadline T.

3.2.2 Solving the MDP model

Our finite horizon MDP model can be efficiently solved by the backwards induction method, which is introduced as follows.

The Backward Induction Algorithm: (Puterman, 2014)

Step 1. Set $t = T$ and $u_T^*(s) = r_T(s)$ for all $s \in S$.

Step 2. Set $t \leftarrow t-1$ for t and compute $u_t^*(s_t)$ for each $s_t \in S$ by

$$u_t^*(s_t) = \max_{a \in A_{s_t}} \{r_t(a, s_t, T) + \lambda \sum_{s' \in S} P_t(s'|s_t, a) u_{t+1}^*(s')\}. \tag{3.1}$$

and

$$A_{s_t,t}^* = \arg\max_{a \in A_{s_t}} \{r_t(a, s_t, T) + \lambda \sum_{s' \in S} P_t(s'|s_t, a) u_{t+1}^*(s')\}. \tag{3.2}$$

Step 3. If $t = 1$, stop. Otherwise return to step 2.

Herein, we use $\pi = (d_1, d_2, ..., d_{T-1})$ to denote a policy, where $d_t : S \to A_s$ is the decision rule prescribing the procedure for action selection in each state at epoch t. $r_t(a_t, s_t, T)$ denotes the random reward received at epoch $t < T$ and $r_T(s_T)$ denotes the terminal reward. $v_T^\pi(s_1)$ denotes the expected total reward over the decision making horizon if policy π is selected and the system is in state s_1 at the first decision epoch. With the discount factor $\lambda \in [0,1)$, the expected total discounted reward will be

$$v_T^\pi(s_1) = E_{s_1}^\pi\{\textstyle\sum_{t=1}^{T-1} \lambda^{t-1} r_t(a_t, s_t, T) + \lambda^{T-1} r_T(s_T)\}.$$

And the total expected reward obtained by using policy π at epochs $t, t+1, ..., T-1$ will be

$$u_t^\pi(s_t) = E_{s_t}^\pi\{\textstyle\sum_{n=t}^{T-1} \lambda^{n-1} r_n(a_n, s_n, T) + \lambda^{T-1} r_T(s_T)\},$$

and $u_T^\pi(s_T) = r_T(s_T)$.

Suppose u_t^*, $t = 1, ..., T$ and $A_{s_t,t}^*$, $t = 1, ..., T-1$ satisfy equation (1) and (2). Let $d_t^*(s_t) \in A_{s_t,t}^*$ for all $s_t \in S$, $t = 1, ..., T-1$ and let $\pi^* = (d_1^*, ..., d_{T-1}^*)$. Then, π^* is the optimal policy and satisfies

$$v_T^{\pi^*}(s) = \sup_\pi v_T^\pi(s),\ s \in S$$

and

$$u_t^{\pi^*}(s_t) = u_t^*(s_t),\ s_t \in S \text{ for } t = 1, ..., T.$$

3.2.3 Case Study

This section introduces a simulation-based case study for the MDP model to solve the resource allocation problem in MATI process. In this case study, we propose a budget, time and probability of success criteria to assess a breeding strategy. We also discuss how the budget is allocated through out the process and how to find the most cost-efficient total budget. For purposes of illustrations, we compare static budget allocation strategies and dynamic budget allocation strategy. All the simulations and case studies are implemented in MATLAB/Octave.

3.2.3.1 Breeding Project Setup

We consider a hypothetical multi-allelic trait introgression project for a case study with the same data structure as the simulation example 1 in Han et al. (2017). The detailed data is in the supplementary material. Table 3.1 contains all the parameters for the case study.

Table 3.1: Parameters

Parameter	Value	Interpretation
$a^{\max}$	1000	maximum progeny number for one generation
A	$\{0, 100, 200, ..., 900, 1000\}$	action space
$C(a)$	$10a$	cost function
$R_t(s, T)$	$2000000 - 100000t$	nominal market value (revenue) function
$r_t(s, T)$	$R_t(s, T)\mathcal{I}(s = success)\mathcal{I}(t \leq T)$	reward function
T	8	deadline (in number of generations)
B	\$11000, \$12000, ..., or \$80000	budget scenarios

3.2.3.2 Preliminary Simulation

Herein, we introduce one possible way to construct the intervals for state space. In order to estimate the intervals, we run 100 preliminary simulations for each possible non-zero action $a \in \{100, 200, ..., 1000\}$.

Preliminary Simulation:

Step 1 Let P^0 denote the initial population and L^{E}, L^{D} denote the elite recipient and donor individuals respectively, where $P^0_{:,:,1} = L^{\mathrm{E}}$ and $P^0_{:,:,2} = L^{\mathrm{D}}$.

Step 2 Set $G = 0$, which represents the largest terminal generation number.

Step 3 Set $m_0 = \texttt{PCV}(L^{\mathrm{E}}, L^{\mathrm{D}}, f)$, in which f represents the recombination frequency.

Step 4

```
for $a = 100 : 100 : 1000$ do
  for $n = 1 : 100$ do
    $g = 0$
    while $\max_k \left\{ \sum_{i=1}^{N} (P^g_{i,1,k} + P^g_{i,2,k}) \right\} < 2N$ do
      $[k^g_1, k^g_2] = \arg\max_{k_1,k_2} \{\texttt{PCV}(P^g_{:,:,k_1}, P^g_{:,:,k_2}, f)\}$
      $p^{n,a}_g = \texttt{PCV}(P^g_{:,:,k^g_1}, P^g_{:,:,k^g_2}, f)$
      $P^{g+1} = \texttt{reproduce}(P^g_{:,:,k^g_1}, P^g_{:,:,k^g_2}, f, a)$
      $g = g + 1$
    end while
    $G = \max(G, g)$
  end for
end for
```

We construct the state space based on parameter G and each $p^{n,a}_g$ estimated from the preliminary simulations. Since F1 will be the same deterministic individual after generation 1 for every simulation, we set $m_1 = p^{n,a}_1, \forall n, a$. On the other hand, m_G will be the deterministic PCV value of the ideal individual, which means $m_G = p^{n,a}_G = \texttt{PCV}(L^{\mathrm{Ideal}}, L^{\mathrm{Ideal}}, f)$. With the preliminary simulations, we define the interval m_g as $m_g = [\min_{n,a}(p^{n,a}_g), \min_{n,a}(p^{n,a}_{g+1})]$ where $2 \leq g \leq G-1, n \in \{1, ..., 100\}, a \in \{100, 200, ..., 1000\}$. The state space construction will be trivial based on the definition.

After the construction of the state space, we need to estimate matrix W^a for the transition probabilities matrix. First, we define function $m_k = \texttt{Interval}(p)$, as the unique interval to which p belongs, i.e., $p \in m_k$. Also, we need to define another matrix $N^a \in \mathbb{I}^{G \times G}$ saving the number of simulations, which is related to the transition between two intervals under action a.

```
for $a = 100 : 100 : 1000$ do
  for $n = 1 : 100$ do
    $g = 1$
    while $p_g^{n,a} < m_G$ do
      $m_{k_1} = \texttt{Interval}(p_g^{n,a})$
      $m_{k_2} = \texttt{Interval}(p_{g+1}^{n,a})$
      $N_{k_1,k_2}^a = N_{k_1,k_2}^a + 1$
      $g = g + 1$
    end while
  end for
end for
$W_{i,j}^a = \frac{N_{i,j}^a}{\sum_j N_{i,j}^a}$
```

3.3 Results

This section covers the results and analysis on the performance about the MDP model in the MATI process. Tradeoffs among the criteria of budget, time and probability of success are shown in Figure 3.3. The detailed allocation policy through the process based on the MDP model is shown in Figure 3.4 and the most cost-efficient total budget for our case study is found in Figure 3.5. Last, we make comparison between static budget allocation strategies and the MDP model based dynamic budget allocation strategy. We illustrate the randomly picked simulation runs for visualization in Table 3.2 and Table 3.3 and show the comparison result in Figure 3.6 to demonstrate the advantage of dynamic resource allocation strategy.

3.3.1 MDP Model Performance Assessment

In order to assess the performance of a strategy for the MATI process, we propose the budget, time and probability of success quantitative criteria. Such criteria are essential for evaluation, improvement, and optimization of the process. In this section, with the simulated state space and estimated transition probabilities, we solved the MDP model for all the budget scenarios. Under the derived optimal policy, we calculate the probability of reaching each state at each decision epoch.

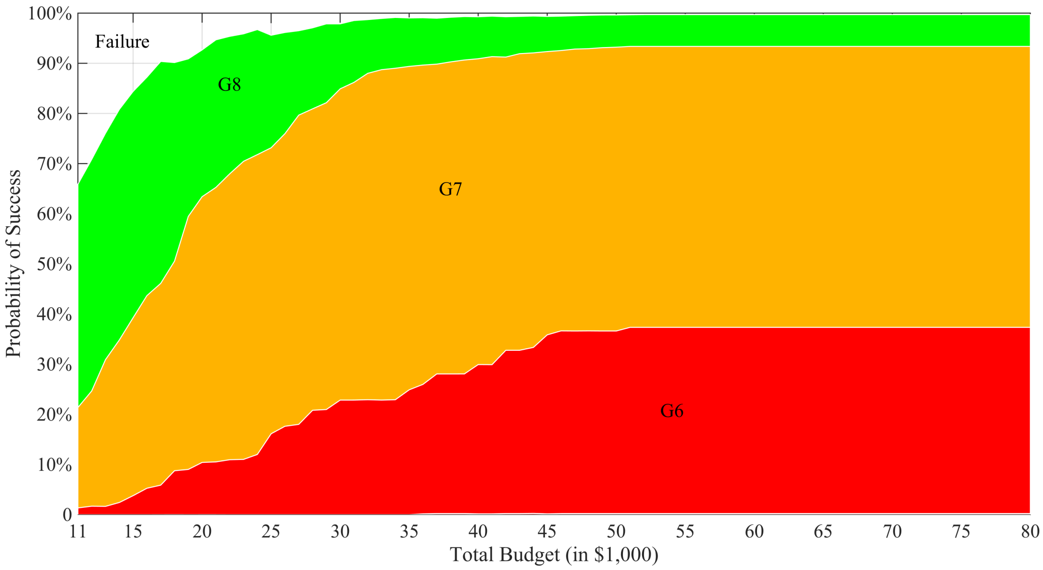


Figure 3.3: BTP graph with $T = 8$. In the figure, the horizontal axis is different total budget scenarios of the breeding project and the vertical axis represents a stacked histogram of the probabilities of reaching success at different generations. In the figure, "GX" label means that the breeding process successfully finishes in X generations and "Failure" means no ideal individual is produced when the budget or the time is depleted.

According to the probabilities of reaching success at each discrete epoch (generation) under all budget scenarios, Figure 3.3 illustrates the tradeoff among total budget, time and probability of success for the MATI process. For example, when the total budget is $11,000, the project can successfully finish in 6,7 or 8 generations with probability about 2%, 20% or 44%, respectively.

The project also has about 34% probability to fail. At the beginning of all the total budget scenarios, increasing total budget can increase the probability of success or decrease the number of generations that the project requires to finish successfully. When the budget is beyond a certain point, e.g., more than $50,000, increasing the budget further cannot make the project finish earlier or increase the probability of success. The performance of the model is irrelevant to the total budget when the budget exceeds this point.

3.3.2 Budget Allocation and Optimal Total Budget

The results from the MDP model provide the plant breeders with intuition about how to efficiently allocate the budget through the entire breeding process and how to identify the most cost-efficient total budget.

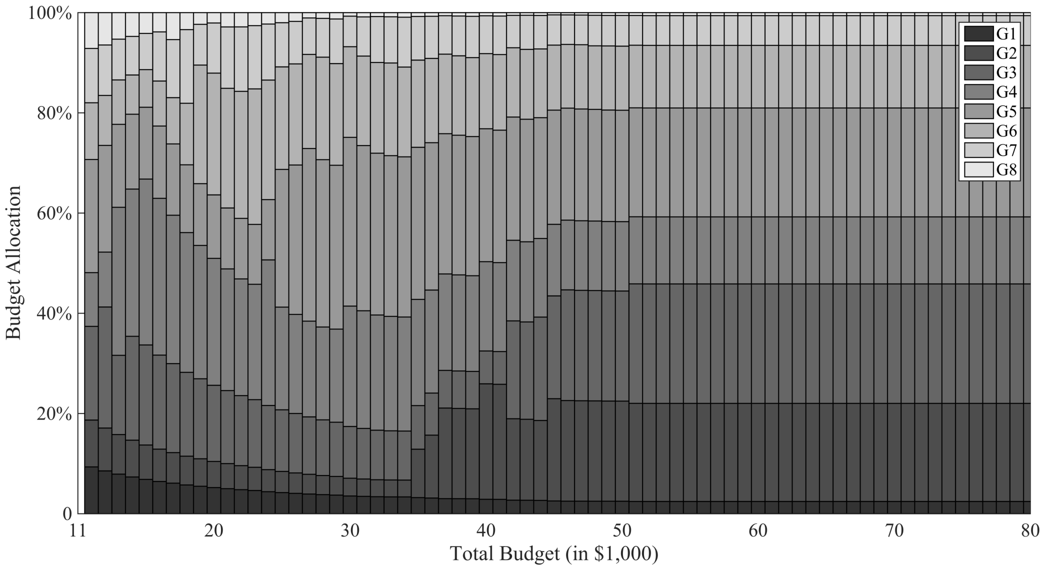


Figure 3.4: Budget allocation with $T = 8$. In the figure, the horizontal axis is different total budget scenarios of the breeding project and the vertical axis represents the proportion of budget allocated in different generations. Different gray scale are used for different generations.

In Figure 3.4, we observe that the model prefers to allocate more budget on the early stage with inadequate budget. Along with increasing the budget, more resources are allocated to the middle generations of the project. When the budget is over abundant, the model sticks to a fixed policy of allocating the budgets. The reason for this pattern is that the model prefers to allocate more resource at the beginning of project to produce progeny inheriting as many desirable alleles as possible to increase the probability of success when the budget is inadequate. When the budget increases, the model tends to save resources on the early stage of the project and efficiently allocates the budget through the entire process. When the budget is more than adequate, the proportion of budget allocated in each generation is fixed.

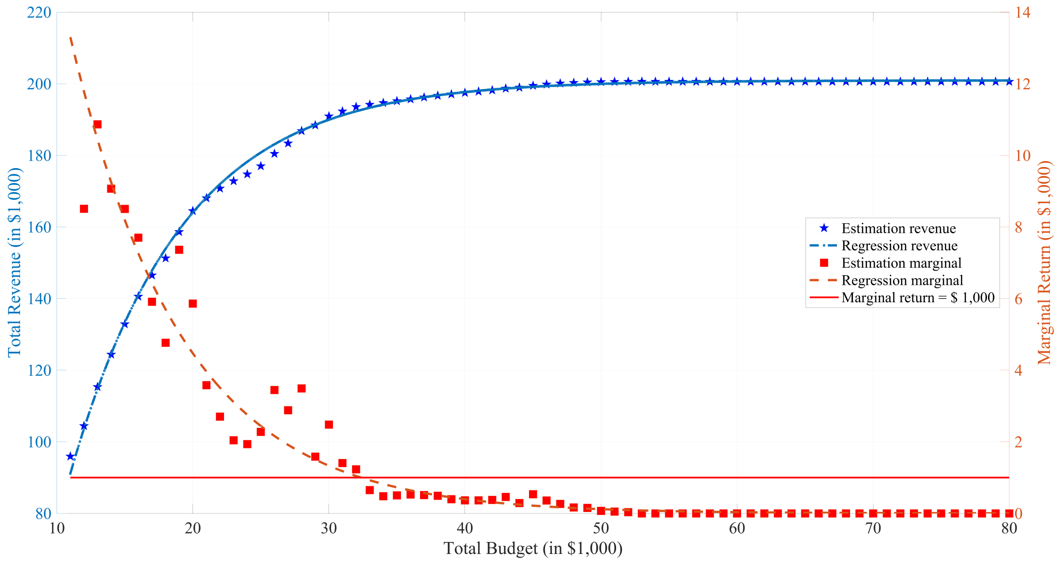
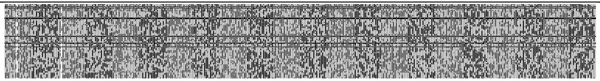
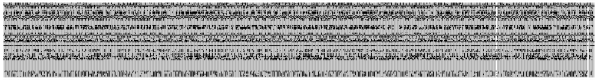
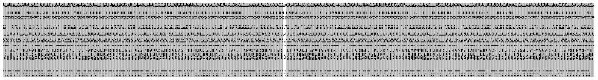
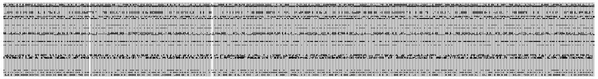
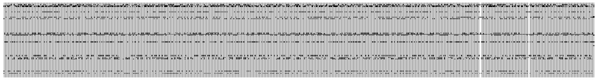
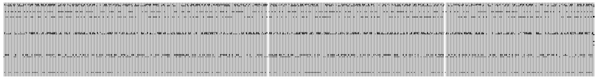
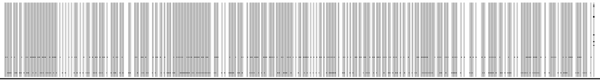


Figure 3.5: Profits and Budgets. In the figure, the blue pentagrams represent the estimation results from simulations and the blue curve represents a nonlinear regression with model $y = a_1 + a_2 \times \exp(a_3 x)$ for the estimation. The red squares represent the difference between the adjacent estimations and the red cure represents the derivative of the expected total revenue curve. The red horizontal line is the marginal return equals to one unit increment of total budget, which is $1,000.

Figure 3.5 illustrates the relation between expected total revenue and budget as well as the marginal return and the budget. Generally speaking, the expected total revenue will increase as the total budget is increased. When the budget is more than sufficient, revenue increments will diminish to zero. The intersection of unit marginal return and the derivative curve indicates

Table 3.2: Generation 2 to generation 8 of one random simulation run with fixed budget allocation

G	Breeding population	$\log_{10}$ PCV
2		–11.69
3		–6.89
4		–4.98
5		–2.76
6		–1.38
7		–0.25
8		0

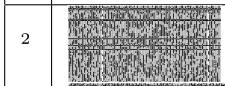
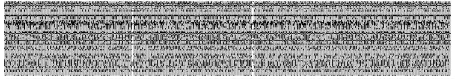
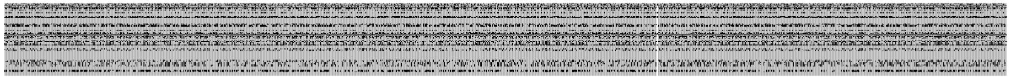
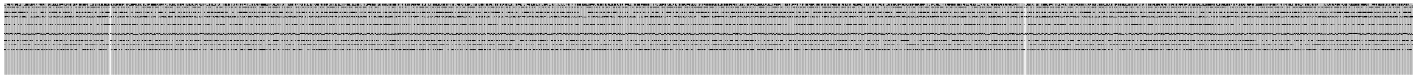
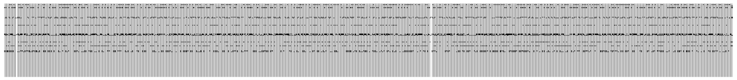
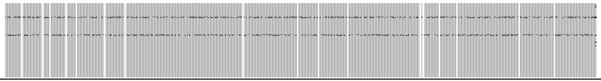
the optimal total budget is about $32,000. When the total budget exceeds the optimal point, the ratio between the revenue increments and budget increments is less than one.

3.3.3 Comparison with Static Budget Allocation Strategies

We demonstrate the different outcomes of the breeding process using static and dynamic resource allocation strategies in this section. For the static strategies, the K^t (or action a) in each generation t is fixed in the range from 100 to 700 progeny per generation with increments of 100. For the dynamic strategy, the optimal policy from the MDP model is applied to allocate resources. The comparison is conducted under the optimal total budget, i.e., $32,000 and the deadline is set to be large enough that the project either succeeds or terminates with budget depleted.

Table 3.2 and Table 3.3 are two examples of randomly selected simulations. Table 3.2 shows the result simulated from the static strategy with $K^t = 400$ for each generation t, whereas Table 3.3 shows the result from the MDP model based strategy. In both tables, the first column is the generation number. In the second column, at each generation, all the progeny produced in the simulation are put abreast to each other to form a wide rectangle and the width of the rectangle

Table 3.3: Generation 2 to generation 8 of one random simulation run with MDP based budget allocation

G	Breeding population	$\log_{10}$ PCV
2		−13.45
3		−9.27
4		−4.94
5		−1.77
6		−0.55
7		0

reflects the sample size. Here we use gray pixels to represent the desirable alleles whereas black pixels to represent the undesirable alleles. Those individuals highlighted by white are the selected breeding parents and several ideal individuals are produced at the last generations. The third column of each table is the base 10 logarithm of PCV value of the selected breeding parents. The fundamental difference between these two resource allocation strategies is that the MDP model allows the decision maker to dynamically allocate the resources based on the outcomes from the previous generation. Note that the dynamic approach produced 17 desirable progeny in seven generations, whereas the fixed plan produced about twice as many desirable progeny, but required an extra generation.

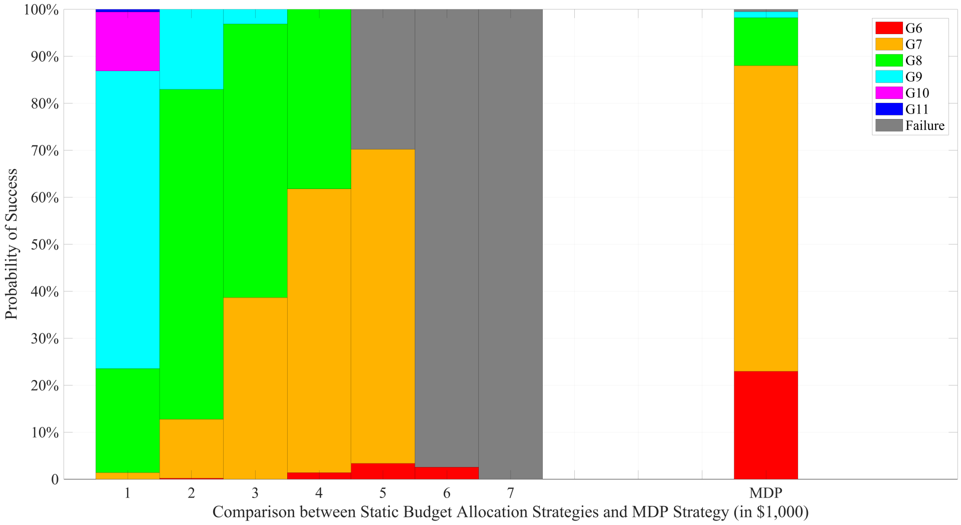


Figure 3.6: Comparison under Optimal Budget. The left 7 stacked bars represent the static budget allocation strategies with different progeny number per generation while the last barplot represents the MDP based strategy.

The 500 repetitions of simulations for each strategy (100 to 700 progeney per generation for static strategy and the MDP based strategy) under a $32,000 total budget reveals the advantage of the MDP based strategy over the static budget allocation strategies. Among static allocation strategies, $4,000 per generation ($K^t = 400$) appears to be the best. Spending less for each generation requires more generations to succeed while spending more for each generation brings a higher risk of failure due to the insufficient total budget. Compared with the strategy of $4,000 per generation, the MDP based strategy has over 20% probability to finish in 6 generations and about 65% probability to finish in 7 generations, which is higher than the probabilities of success in the same generations with $K^t = 400$. The numerical values of the comparison can be found in the following Table 3.4.

Table 3.4: Comparison under Optimal Budget

Progeny/generation	G6	G7	G8	G9	G10	G11	Failure
100	0	1.37%	22.16%	63.33%	12.55%	0.59%	0
200	0.20%	12.55%	70.20%	17.06%	0	0	0
300	0	38.63%	58.24%	3.14%	0	0	0
400	1.37%	60.39%	38.24%	0	0	0	0
500	3.33%	66.86%	0	0	0	0	29.80%
600	2.55%	0	0	0	0	0	97.45%
700	0	0	0	0	0	0	100%
MDP	22.92%	65.07%	10.20%	1.29%	0	0	0.52%

3.4 Discussion

In this paper, we discussed the MATI process and the resource allocation problem. We updated the work flow of MATI process by adding the components of resource allocation. A mathematical algorithmic simulating process was then formulated for the MATI process and we addressed the resource allocation challenge. We used Markov decision processes with backwards induction method due to the dynamic programming characteristics of the problem. The performance of the MDP model was examined for a hypothetical breeding project with realistic estimated parameters. We proposed three quantitative assessing criteria and corresponding figures for visualization. We analyzed the pattern in resource allocation through the process under different budget scenarios and found the optimal total budget for this hypothetical project, as well. Under the optimal budget, the MDP model based strategy was compared with the static budget allocation strategies and we demonstrated the improvement brought by dynamically allocating the resources.

Even with the outstanding improvement in terms of time and probability of success, one unresolved issue is that we cannot guarantee the global optimality of the policy because of the complexities in the problem. As pointed out, the construction of the state space is highly related to the preference of the decision maker. On the other hand, due to assumptions made for our model, the result for continuous actions or budgets is still unknown. From such points of view, we shall see the need for many further research studies on the MATI process.

Also, estimating the cost and revenue function is another possible economical research topic for further discussion. Plant breeding organizations have their own forecasting models about the market value of a certain genotype as well as its revenue associated with time when it is delivered to the market. Thus, the research on the discussion about cost and revenue functions may reveal more economic discoveries about the trait introgression problem and bring more inspirations.

CHAPTER 4. The Predicted Cross Value for Multi-pair Parental Selection in Trait Introgression Process

Abstract

The Predicted Cross Value (PCV) designed for one donor and one recipient has been proved to outperform the conventional selection metrics for parental selection in the trait introgression process. As per the request of practical introgression projects, we extend the PCV for multi-pair parental selection in this chapter, which we refer to as the NPCV metric. The updated metric takes the estimates of recombination frequencies as input parameters and calculates the probability that a gamete with desirable alleles at all specified loci being produced by a number of pairs of breeding parents after a certain number of generations. We design set cover models to select the optimal set of breeding parents based on different metrics. With the simulation based case studies, we compare this NPCV metric with the conventional GEBV metric, and the results demonstrate the advantage of this new metric on the efficiency and effectiveness for parental selection. Also, we recommend some future work directions for improving the trait introgression process.

4.1 Introduction

In Chapter 2, we introduced the multi-allelic trait introgression process and designed the predicted cross value (PCV) for one pair of breeding parents. According to the comparisons between the PCV and the genetic estimated breeding value (GEBV) or the optimal haploid value (OHV) (Daetwyler et al., 2015), the PCV offered significant advantages by taking the recombination into consideration. However, as pointed out in the discussion of Chapter 2, because of the simplifying assumptions, the PCV metric has some limitations, one of which is that the PCV is limited to one pair parental selection. In practical breeding projects, one pair of breeding parents is not enough sometimes because the desirable alleles may not be carried by one pair of individuals or one pair of breeding parents may not produce sufficient seeds like soybeans.

Some research has been conducted to deal with such problems. For the case that the desirable alleles are carried by multiple individuals, plant breeders are used to combine different alleles from several individuals together to create one promising donor for the following breeding process, which is referred to as the gene stacking or gene pyramiding procedure. Different approaches have been proposed for gene stacking or gene pyramiding. Sequentially introgressing alleles from each individual is one possible approach. Another approach is to pairwise stack all the desirable alleles into one single individual (Peng et al., 2014a,b). Servin et al. (2004) described another efficient method to design the pyramiding scheme. Meanwhile, some optimization based methods have been proposed for gene stacking problem (Canzar and El-Kebir, 2011; Xu et al., 2011), which has been proved as a NP-hard problem (Xu et al., 2011). However, for such gene stacking or gene pyramiding problem, one of the major issues is the instability of recovering all the favorable background genes (Halpin, 2005). During stacking different desirable genes from different individuals, because of the recombination, there are possibilities that the undesirable alleles are brought into the offspring gene pool, as well. Thus, even the final offspring successfully carries all the desirable alleles, some of the desirable background genes may be lost. Hence, a method for combining targeted desirable genes as well as preserving desirable background genes is demanded. On the other hand, for the case that one pair of breeding parents cannot produce sufficient number of progeny, breeders have to select multiple pairs of breeding parents to produce enough progeny for the following breeding process. Based on the existing problems, an efficient and effective strategy to select multiple pairs of breeding parents is required not only to combine desirable alleles from different individuals but also to preserve desirable background genes. The PCV designed in Chapter 2 is able to quantify the efficiency of selections based on the entire genetic marker set information including the recombination frequency, which guarantees that the desirable background alleles are maintained during the parental selection. Inspired by the advantages of the PCV, we design a new NPCV metric for trait introgression process with multi-pair parental selection.

The remainder of this chapter is organized as follows. First, we propose the definition of the NPCV. Meanwhile, we update the plumbing calculation system for the NPCV. In addition to the NPCV, we build a set cover model to select multiple pairs of breeding parents according to the GEBV for simulating the conventional selection approach. We also build a set cover model for the NPCV metric and propose a heuristic approach for the optimal parental selection. We next conduct simulation experiments to compare the efficiency and effectiveness of the GEBV and the NPCV based approaches. We demonstrate the results via the criteria of cost, time and probability of success for different approaches and make the discussion for our research and future work at the end of this chapter.

4.2 Formulations

4.2.1 Simplifying assumptions

Several assumptions made in Chapter 2 to simplify the formulation and illustrate the core elements of the introgression process are still necessary for proposing the NPCV, which are:

- Consider annual diploid and allopolyploid species such as corn, rice, soybean and wheat with subgenome specific loci. Extension to perennial and autopolyploid crops, such as alfalfa is deferred for future research.

- Consider a single multi-allelic trait, where all segregating loci associated with the trait are known. Results also apply to multiple traits where all traits are of equal value.

- All marker alleles are either desirable or undesirable. Values of alleles could be modeled as continuous from some distribution or in many cases, the value of an allele is unknown. We defer expansion to these situations for future research.

- During evaluation, a sufficient number of informative markers are distributed throughout the genome at sufficient density to allow estimation of recombination between all adjacent pairs of markers.

- Recombination events between pairs of adjacent loci are assumed to be independent (Haldane, 1919). Consideration of interference is deferred for future research.

At the same time, we relax two of the assumptions in Chapter 2 for the NPCV design, which are stated as follows:

- To illustrate the principles, all desirable alleles missing in the recipient **can be carried by one or multiple individuals**.

- **One or multiple pairs of breeding parents** are selected for crossing in each generation.

4.2.2 Preliminary definitions

We use the same notations as Chapter 2 to quantitatively describe the essence of trait introgression. The N by 2 binary matrix, say $L \in \mathbb{B}^{N\times 2}$, denotes the genotype of an individual plant, where N is the total number of QTL in the genome and the binary value $L_{i,j}$ indicates whether the allele at locus i of chromosome j is desirable ($L_{i,j} = 1$) or undesirable ($L_{i,j} = 0$). As defined in the Definition 1 in Chapter 2, the Gamete function, $g = \texttt{Gamete}(L, J)$ represents the actual gamete formation process and the input binary vector J must be a random one following an *inheritance distribution* with parameter $r \in [0, 0.5]^{N-1}$, which is:

$$J_1 = \begin{cases} 0 & \text{w.p.} \quad 0.5 \\ 1 & \text{w.p.} \quad 0.5 \end{cases}, \tag{4.1}$$

$$J_i = \begin{cases} J_{i-1} & \text{w.p.} \quad 1 - r_{i-1} \\ 1 - J_{i-1} & \text{w.p.} \quad r_{i-1} \end{cases}, \forall i \in \{2, ..., N\}. \tag{4.2}$$

The Reproduce function, $X = \texttt{Reproduce}(L^1, L^2, r, K)$, was defined to represent the reproduction of selected breeding parents. Its input parameters include two binary matrices $L^1, L^2 \in \mathbb{B}^{N\times 2}$, a vector $r \in [0, 0.5]^{N-1}$, and a positive integer number K. Its output is a three-dimensional matrix $X \in \mathbb{B}^{N\times 2\times K}$, representing a population of K progeny, which is determined by first generating $2K$ independent and identically distributed random vectors from the inheritance distribution with parameter r, denoted as $J_p, \forall p \in \{1, ..., 2K\}$, and then setting $X_{i,j,k} = \texttt{Gamete}_i(L^j, J_{2k-2+j}), \forall i \in \{1, ..., N\}, j \in \{1, 2\}, k \in \{1, ..., K\}$.

The targets for parental selection in this chapter has been switched to multiple pairs of breeding parents, say M pairs. We update the core Breed function to mathematically formulate the breeding process, which is defined as follows.

Definition 15. *We define the Breed_new function as $G = \texttt{Breed\_new}(P^0, r, K, M)$. Its input parameters include a three-dimensional binary matrix $P^0 \in \mathbb{B}^{N\times 2\times 2}$, a vector $r \in [0, 0.5]^{N-1}$, and positive integers K and M. Its output, G, is the number of generations it takes to successfully finish the process or to terminate with no improvement can be achieved, which is determined through the following steps.*

```
Step 0 (Initialization) Set $t = 0$ and go to Step 1.
Step 1 (Evaluation)
If $\max_k \left\{ \sum_{i=1}^{N} (P^t_{i,1,k} + P^t_{i,2,k}) \right\} = 2N$ or $\texttt{Metric}(P^t) \leq \texttt{Metric}(P^{t-1})$
    RETURN: $G = t$.
Else Go to Step 2.
Step 2 (Selection) Obtain $[k^t_1, k^t_2, ..., k^t_{2M}] = \texttt{Select}(P^t, r, M)$ and go to step 3.
Step 3 (Reproduction) Obtain $P^{t+1}_{:,:,(m-1)K+1:mK} =$
$\texttt{Reproduce}(P^t_{:,:,k^t_{2m-1}}, P^t_{:,:,k^t_{2m}}, r, K), m \in \{1, ..., M\}$, update $t \leftarrow t + 1$, and go to Step 1.
```

Herein, $\texttt{Metric}(\cdot)$ returns the value of a breeding population based on the given metric.

4.2.3 The NPCV definition

We propose the definition of the NPCV in this section for more realistic introgression problems. The NPCV is defined as a metric for quantifying the selection of multiple, say $2M$ individuals $\{L^1, L^2, L^3, ..., L^{2M}\}$ as M pairs of breeding parents. For one pair of individuals, the PCV is defined as the conditional probability of producing a gamete inheriting all the desirable alleles after 2 generations. For $2M$ breeding individuals, the NPCV calculates the conditional probability of producing a gamete inheriting all the desirable after $\lceil \log_2 2M \rceil + 1$ generations. The number $\lceil \log_2 2M \rceil + 1$ is derived based on the logic as follows. If there are $2M$ desirable alleles and each of them is carried by one individual, at least $\lceil \log_2 2M \rceil + 1$ generations are required to stack all the desirable alleles into one gamete (Servin et al., 2004). An example with 16 desirable alleles carried by 16 breeding parents is illustrated by the following Figure 4.1.

The concept of the NPCV is described as follows. Let $L^1, L^2, ..., L^{2M} \in \mathbb{B}^{N\times 2}$ denote $2M$ breeding individuals, and let $[g_1^{i,1}, g_1^{i,2}]$ denote a random progeny produced by the ith pair of breeding parents L^{2i-1} and L^{2i} after generation 1, where $g_1^{i,1} = \texttt{Gamete}(L^{2i-1}, J)$ and $g_1^{i,2} = \texttt{Gamete}(L^{2i}, J)$, respectively. In the second generation, $[g_1^{i,1}, g_1^{i,2}]$ is randomly crossed with $[g_1^{j,1}, g_1^{j,2}]$, which is produced by the jth pair of breeding parents after generation 1, to produce another random progeny denoted as $[g_2^{m,1}, g_2^{m,2}]$, where $g_2^{m,1} = \texttt{Gamete}([g_1^{i,1}, g_1^{i,2}], J)$ and $g_2^{m,2} = \texttt{Gamete}([g_1^{j,1}, g_1^{j,2}], J)$, respectively. In generation 3, this random progeny will be randomly mated to another random progeny produced after generation 2 to generate new offspring. In each of the consecutive generation, the available number of randomly progeny decreases by half. If the number of available random progeny is odd in a certain generation, one progeny will be left to the next generation as the available progeny for the random mating process. The NPCV calculates the conditional probability of producing a gamete $g_{\lceil \log_2 2M \rceil+1}^{m^*,1} = \texttt{Gamete}([g_{\lceil \log_2 2M \rceil}^{m',1}, g_{\lceil \log_2 2M \rceil}^{m',2}], J)$ inheriting all the desirable alleles by $[g_{\lceil \log_2 2M \rceil}^{m',1}, g_{\lceil \log_2 2M \rceil}^{m',2}]$ after the $\lceil \log_2 2M \rceil + 1$ generation.

Definition 16. *For given M pairs of individuals $\{L^1, L^2, ..., L^{2M}\}$, the NPCV is defined as the probability that in generation $\lceil \log_2 2M \rceil + 1$ a random gamete, $g_{\lceil \log_2 2M \rceil+1}^{m^*,1}$, produced by a*

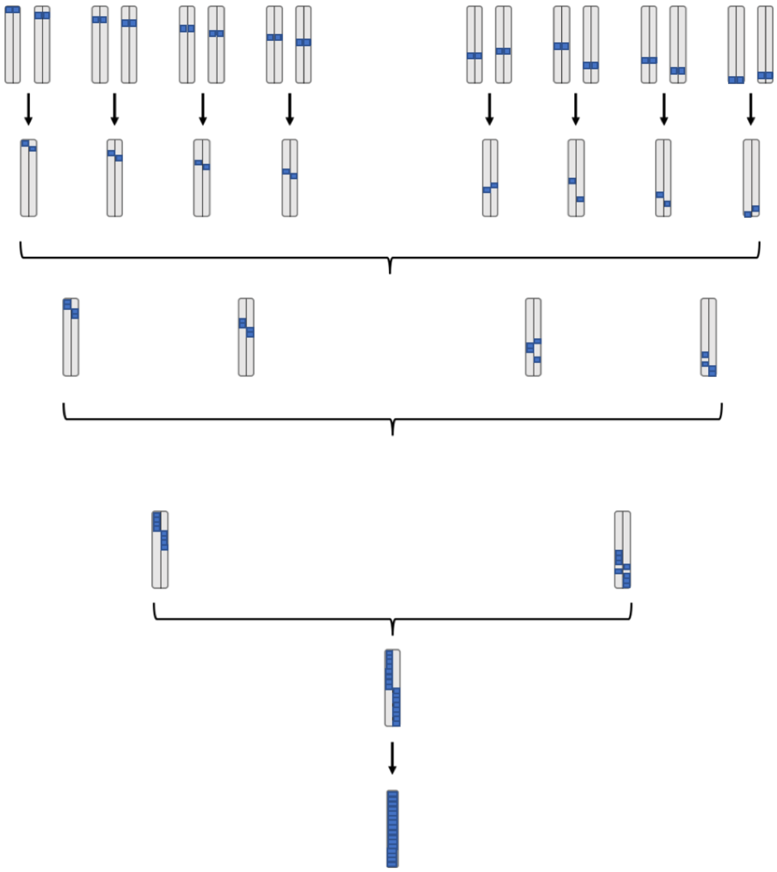


Figure 4.1: Scheme for NPCV

random progeny $[g^{m',1}_{\lceil \log_2 2M \rceil}, g^{m',2}_{\lceil \log_2 2M \rceil}]$ *will consist only of desirable alleles:*

$$\mathtt{NPCV}(\{L^1, L^2, ..., L^{2M}\}, r) = P(g^{m^*,1}_{\lceil \log_2 2M \rceil + 1} = 1, \forall i \in \{1, ...N\}).$$

Here, r *is the recombination frequency vector.*

The rationale for the NPCV definition is to calculate the probability that none of the undesirable alleles survives $\lceil \log_2 2M \rceil + 1$ generations of meiosis. Similar to the PCV, the essence of this approach is to select breeding parents based on their likelihood to produce an ideal gamete by combining their desirable alleles rather than the fitness of the breeding parents themselves. We shall notice that the design of this random breeding scheme for the NPCV is not unique and herein, we just provide with a feasible breeding scheme to derive a new metric for parental selection. Thus, we can not claim that the NPCV is the optimal approach to select multiple pairs of breeding parents. Based on such metric, we will design optimization approaches to select breeding parents that can lead to the largest metric value in the following sections.

4.2.4 The water pipe algorithm for NPCV calculation

We update the polynomial time plumbing algorithm for calculating the NPCV. For $2M$ breeding parents, the plumbing system for the NPCV consists of N rows and $4M$ columns of valves and a number of water pipes connecting them. The $4MN$ valves correspond to the $4MN$ alleles in the $2M$ breeding parents represented by a set of matrices $\{L^1, L^2, ..., L^{2M}\}$. For notational convenience, we will use $L \in \mathbb{B}^{N\times 4M}$ to denote the set of matrices $\{L^1, L^2, ..., L^{2M}\}$, so $L_{i,j} = L^{\lceil j/2\rceil}_{i,2-j\%2}, \forall i \in \{1, ..., N\}, j \in \{1, ..., 4M\}$, and % represents taking the remainder. Similarly, the algorithm draws an analogy between the conditional probability and the water flows through a plumbing system as the algorithm in Chapter 2. For all $i \in \{1, ..., N-1\}$, $j \in \{1, ..., 4M\}$, and $k \in \{1, ..., 4M\}$, the volume of the pipe that connects valves (i, k) and $(i+1, j)$ is denoted as $T_{k,j,i}$, where T is a three-dimensional matrix, which is referred to as the *transition matrix* and defined as follows.

Definition 17. *For a given vector of recombination frequencies, $r \in [0, 0.5]^{N-1}$, the transition matrix $T \in [0, 0.5]^{4M\times 4M\times(N-1)}$ is defined as*

$$T_{4j-3:4j,4j-3:4j,i} \approx \begin{bmatrix} (1-r_i)^t & r_i(1-r_i)^{t-1} & 0.5r_i(1-r_i)^{t-2} & 0.5r_i(1-r_i)^{t-2} \\ r_i(1-r_i)^{t-1} & (1-r_i)^t & 0.5r_i(1-r_i)^{t-2} & 0.5r_i(1-r_i)^{t-2} \\ 0.5r_i(1-r_i)^{t-2} & 0.5r_i(1-r_i)^{t-2} & (1-r_i)^t & r_i(1-r_i)^{t-1} \\ 0.5r_i(1-r_i)^{t-2} & 0.5r_i(1-r_i)^{t-2} & r_i(1-r_i)^{t-1} & (1-r_i)^t \end{bmatrix},$$
$$\forall i \in \{1, ..., N-1\}, j \in \{1, ..., M\} \quad (4.3)$$

where $t = \lceil \log_2(2M)\rceil + 1$ and for the rest of elements $T_{p,q,i} \notin T_{4j-3:4j,4j-3:4j,i}$ of T,

$$T_{p,q,i} \approx \frac{1-(1-r_i)^{t-2}}{4M-4}, \quad (4.4)$$

In the definition, the reason for the approximately equality sign is that when $2M$ equals to 2 raised to a certain power, we can achieve equality. When $2M$ is even but not equals to 2 raised to a certain power, in some generation during the breeding scheme, a random progeny will be left for the following generations. Thus, the matrix is just a close approximation. The derivation of the transition matrix is in the Appendix B. Here, although the matrix is an approximation for some scenarios, the following simulation based on this matrix exactly simulates the breeding process and the results derived shall be valid.

We define the water matrix $W \in [0,1]^{N\times 4M}$ to represent the amounts of water flowing inside the plumbing system. For all $i \in \{1,...,N\}$ and $j \in \{1,...,4M\}$, $W_{i,j}$ represents the amount of water that flows out of the jth valve in the ith row. This value can be interpreted as the probability that the first i alleles in the gamete $g^{m^*,1}_{\lceil \log_2 2M \rceil+1}$ are desirable and that the ith allele is inherited from the jth chromosome of the breeding parents.

Definition 18. *We define the* water matrix $W \in [0,1]^{N\times 4M}$ *as*

$$W_{i,j} = P(g_1 = ... = g_i = 1, g_i = L_{i,j}), \forall i \in \{1,...,N\}, j \in \{1,...,4M\}. \tag{4.5}$$

Proposition 3. *The water matrix can be calculated as follows.*

$$W_{1,j} = \frac{1}{4M} L_{1,j}, \forall j \in \{1,...,4M\}; \tag{4.6}$$

$$W_{i,j} = L_{i,j} \sum_{k=1}^{4M} T_{k,j,i-1} W_{i-1,k}, \forall i \in \{2,...,N\}, j \in \{1,...,4M\}. \tag{4.7}$$

Proposition 4. *The NPCV is the summation of the last row in the water matrix:*

$$\texttt{NPCV}(\{L^1, L^2, ..., L^{2M}\}, r) = \sum_{j=1}^{4M} W_{N,j}. \tag{4.8}$$

The proofs for these propositions are similar to the ones in Chapter 2.

4.2.5 Illustrative example

We illustrate the plumbing system for the NPCV calculation with the following examples.

Example 4. *The 6 breeding parents are both ideal lines* $L^1 = L^2 = ... = L^5 = L^6 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$ *and* $r = \begin{bmatrix} 0.2 & 0.1 & 0.4 & 0.5 \end{bmatrix}^\top$.

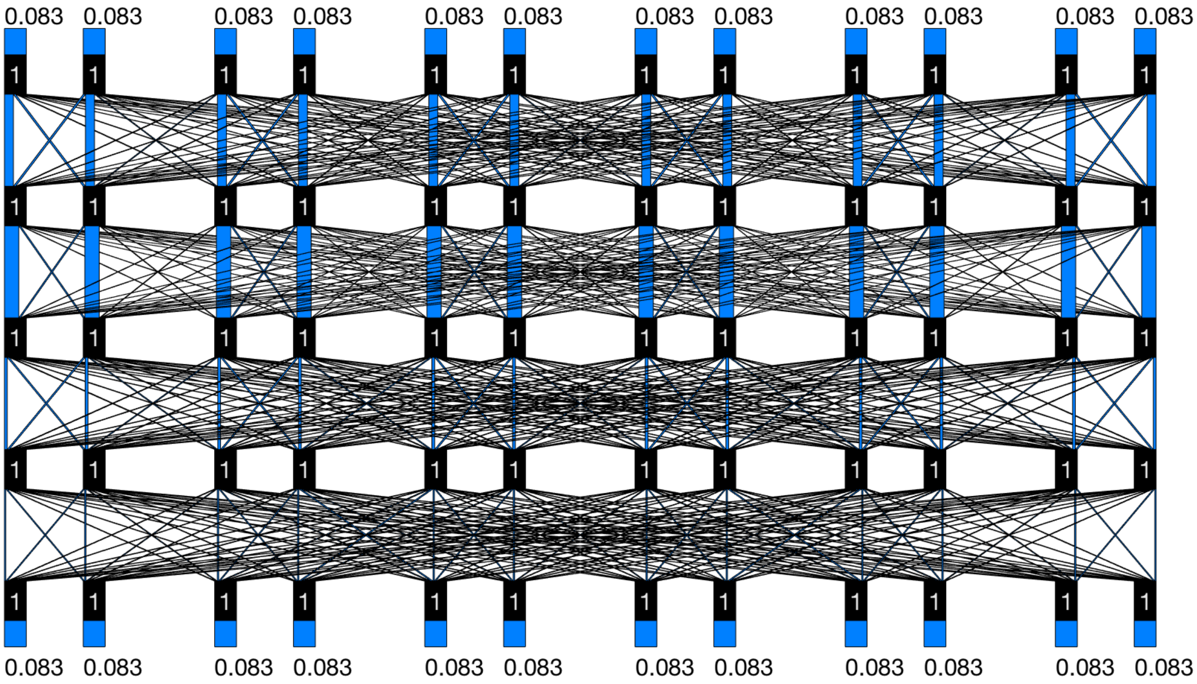


Figure 4.2: Illustration of the plumbing system for Example 4.

The plumbing system corresponding to Example 1 is illustrated in Figure 4.2. Since all breeding parents are already ideal lines, their NPCV equals to 1. Albeit trivial, this fact is verified by the plumbing system in Figure 4.2, where all the valves are open, and thus 100% of the water that is poured in will get its way out.

We now illustrate the water pipe algorithm for calculating the NPCV of the following example.

Example 5. *The 6 breeding parents are* $L^1 = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, L^2 = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, L^3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, L^4 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}, L^5 = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, L^6 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}$ *and the recombination frequencies vector is the same as in Example 4.*

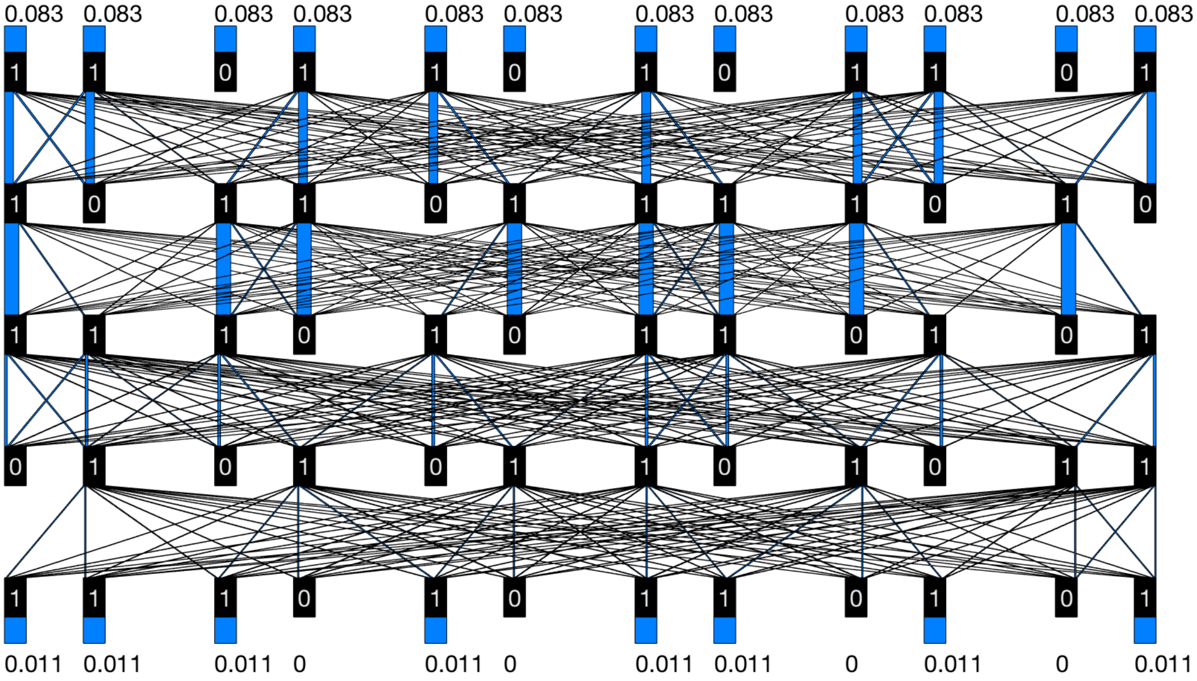


Figure 4.3: Illustration of the plumbing system for Example 5.

The plumbing system corresponding to Example 5 is illustrated in Figure 4.3, in which we removed those water pipes whose immediate upstream valves are closed.

4.3 Metric Optimization

The GEBV and NPCV are designed as the metrics to quantify the selection of breeding parents, based on which we can optimize the parental selection in trait introgression. For the GEBV and NPCV, we propose set cover models to choose the optimal breeding parents. For the NPCV, because of the complexity of the model, we propose a heuristic algorithm to derive the solution.

4.3.1 A set cover model for GEBV selection

Given a population of candidate individuals and required number of breeding parents in each generation, we present a set cover model to select a group of breeding parents with the maximum overall GEBV. The conventional GEBV based selection approach selects breeding parents based on the GEBV as the equation 4.9, which measures the fitness of each individual. This metric quantifies the single merit of each individual and those with the largest GEBV shall be selected as the breeding parents according to the metric.

$$\sum_{i=1}^{N}(L_{i,1} + L_{i,2}). \tag{4.9}$$

The set cover model takes two parameters as input: the set of progeny of individuals $P \in \mathbb{B}^{N\times 2\times K}$ with K being the total number of progeny in this population and the required number of pairs of breeding parents $M, M \in \mathbb{I}^{+}$. That is to say, $2M$ individuals will be selected. There is one set of decision variables: $x = (x_1, x_2, ..., x_K) \in \mathbb{B}^{K\times 1}$ with x_k indicating whether $(x_k = 1)$ or not $(x_k = 0)$ individual k is selected as a breeding parent, for all $k \in \{1, ..., K\}$.

The optimization model is presented in (4.10)-(4.13), which is an integer linear program (ILP). The objective function (4.10) calculates the overall summation of GEBVs of the selected breeding parents, which is to be maximized. Constraint (4.11) requires that exactly $2M$ breeding parents are selected from the population. Constraints (4.12) requires that at least one desirable allele is present at each locus (each row in the matrix) among all the selected breeding parents. This ILP model can be efficiently solved to optimality by existing algorithms and software.

$$
\begin{aligned}
\max_{x} \quad & \sum_{i=1}^{N}\sum_{j=1}^{2}\sum_{k=1}^{K} P_{i,j,k}x_k && (4.10)\\
\text{s. t.} \quad & \sum_{k=1}^{K} x_k = 2M && (4.11)\\
& \sum_{j=1}^{2}\sum_{k=1}^{K} P_{i,j,k}x_k \geq 1 \quad \forall i \in \{1, ..., N\} && (4.12)\\
& x \text{ binary.} && (4.13)
\end{aligned}
$$

The output x^* of the model indicates which individuals are selected as breeding parents achieving the largest overall GEBV. We group those selected individuals as $\{L^1, L^2, ..., L^{2M}\}^*$ in which $L^m = P_{:,:,S^m}$, $x^*_{S^m} = 1, \forall m \in \{1, 2, ..., 2M\}$, $S^m \in \{1, ..., K\}$ and when $m_i \neq m_j$, we have $S^{m_i} \neq S^{m_j}$. The binary constraint on variable x in the model indicates that self pollination is not considered in this ILP. Self pollination for multiple pairs parental selection as a special case increases the complexity of the problem. For instance, we need to consider how many pairs of self pollination is allowed for one progeny and we need to determine if one progeny is allowed to conduct self pollination and cross with other progeny simultaneously. Thus, this special case provides with more potential future research topics and it could be discussed in the future work. At the same time, there could be multiple optimal solutions for one candidate population P. Also, our model only conducts the selection but not pairs the selected parents up, which means the groups consist of $\{L^1, L^2, ..., L^{2M}\}^*$ with different orders are considered as the same result.

4.3.2 Heuristic NPCV optimization

Given a population of candidate individuals and required number of breeding parents, we present another set cover model as (4.14)-(4.19) to select a group of breeding parents with the maximum NPCV.

$$\begin{array}{lll}
\max\limits_{x} & \texttt{NPCV}(\{L^1, L^2, ..., L^{2M}\}, r) & \quad (4.14)\\
\text{s.t.} & \sum\limits_{k=1}^{K} x_k = 2M & \quad (4.15)\\
& x_{S^m} = 1 \qquad m \in \{1, 2, ..., 2M\}, S^m \in \{1, ..., K\} & \quad (4.16)\\
& L^m = P_{:,:,S^m} \qquad m \in \{1, 2, ..., 2M\}, S^m \in \{1, ..., K\} & \quad (4.17)\\
& \sum\limits_{j=1}^{2} \sum\limits_{k=1}^{K} P_{i,j,k} x_k \geq 1 \qquad \forall i \in \{1, ..., N\} & \quad (4.18)\\
& x \text{ binary.} & \quad (4.19)
\end{array}$$

Herein, when $m_i \neq m_j$, we have $S^{m_i} \neq S^{m_j}$

The set cover model for the NPCV metric is similar to the one for the GEBV metric. However, the nonlinear definition of the NPCV makes the model too complex to find the exact optimal solution. Thus, we propose the following heuristic algorithm for finding a solution.

Step 0 Derive x^0 by solving model (4.10)-(4.13) and get $\{L^1, L^2, ..., L^{2M}\}^0$ where $L^m = P_{:,:,S^m}$ and $x^0_{S^m} = 1, m \in \{1, ..., 2M\}, S^m \in \{1, ..., K\}$.

Step 1 Set $\texttt{NPCV}^{\max} = \texttt{NPCV}(\{L^1, L^2, ..., L^{2M}\}^0, r)$ and $\{L^1, L^2, ..., L^{2M}\}^{\max} = \{L^1, L^2, ..., L^{2M}\}^0$.

Step 2 (Iteratively Update)

```
for $i = 1 : 2M$ do
    $\{L^1, L^2, ...L^{2M}\}^i = \arg\max_{L^i \in \{L^1,...,L^K\}/\{L^1,...,L^{2M}\}^{\max}} \texttt{NPCV}(\{L^1, ..., L^i, ...L^{2M}\}^{\max}, r)$
    if $\texttt{NPCV}(\{L^1, L^2, ...L^{2M}\}^i, r) > \texttt{NPCV}^{\max}$ then
        $\texttt{NPCV}^{\max} = \texttt{NPCV}(\{L1, ..., L^i, ...L^{2M}\}^i, r)$ and $\{L^1, ..., L^{2M}\}^{\max} = \{L1, ..., L^i, ...L^{2M}\}^i$
    end if
end for
```

4.4 Simulation Experiments

In this section, we describe the case study on two simulated experiments using both of the GEBV and NPCV approaches on the same data set in Chapter 2 and we report the results, as well.

4.4.1 Experiment description

We consider a hypothetical genotype consisting of 100 QTLs. The locations of QTLs are uniformly and randomly distributed among ten simulated linkage groups with each linkage group having from 8 to 12 QTLs. Two example trait introgression projects are considered. The setup of the parameters is the same as Chapter 2, which is reviewed as follows:

The recipient and donor individuals are homozygous at all QTLs in both projects. The recipient in the first example has desirable alleles at 93 of the QTLs and the donor carries the remaining 7 desirable alleles. In the recipient, the undesirable alleles are at C1Q4, C1Q6, C2Q9, C3Q1, C5Q4, C6Q3, and C6Q8, where CiQj denotes the jth QTL in chromosome i. The recipient in the second example has desirable alleles at 80 of the QTLs and the donor carries the remaining 20 desirable alleles. In the recipient, the undesirable alleles are at C1Q5, C1Q10, C2Q4, C2Q9, C3Q5, C3Q10, C4Q3, C4Q8, C5Q3, C5Q8, C6Q2, C6Q7, C6Q12, C7Q5, C7Q9, C8Q3, C8Q8, C9Q5, C9Q9, and C10Q3. A Recombination frequencies vector is given in Table 4.1 for the simulation.

Table 4.1: The recombination frequencies used in the simulation.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Q1	0.2725	0.2075	0.0569	0.0860	0.1414	0.0791	0.0126	0.2179	0.0659	0.0610
Q2	0.2649	0.1957	0.0759	0.1362	0.1693	0.1529	0.2951	0.1647	0.0102	0.0800
Q3	0.2148	0.0692	0.1452	0.1983	0.0285	0.3210	0.3044	0.2597	0.2480	0.2955
Q4	0.1262	0.1004	0.1037	0.0874	0.0875	0.1823	0.2654	0.2383	0.1667	0.0096
Q5	0.2705	0.1570	0.3078	0.2009	0.2670	0.1737	0.0329	0.3012	0.1600	0.1633
Q6	0.1776	0.0768	0.1434	0.2371	0.0097	0.0772	0.0873	0.2970	0.3016	0.0560
Q7	0.1169	0.2814	0.0616	0.0739	0.3096	0.1630	0.1118	0.1114	0.2033	0.3262
Q8	0.3130	0.0649	0.3016	0.0391	0.2434	0.2080	0.2266	0.5000	0.2059	0.5000
Q9	0.2920	0.5000	0.3266	0.0989	0.1629	0.2264	0.0455	–	0.2865	–
Q10	0.5000	–	0.1463	0.5000	0.5000	0.1318	0.2404	–	0.2685	–
Q11	–	–	0.5000	–	–	0.1225	0.5000	–	0.5000	–
Q12	–	–	–	–	–	0.5000	–	–	–	–

We implemented the `Breed_new` function to simulate the introgression project, with the simulated genomes as the initial population for each example. In subsequent generations, 2 to 10 pairs of breeding parents or equivalently, 200 to 1000 progeny with 100 progeny per pair were sampled from simulated crosses. The case for one pair breeding parents has been addressed in the chapter for PCV. Two versions of the `Select` function were compared:

- The GEBV approach, which selects individuals with the largest overall GEBVs.

- The NPCV approach, which selects individuals with the largest NPCV.

3600 simulation runs in total were carried out for all possible number of pairs ranging from 2 to 10, and the comparison was based on the time and probability of success, i.e., number of generations to terminate and also different choice of number of pairs of breeding parents. The simulation was implemented and results were generated using GNU Octave (Eaton et al., 2015). One random run of simulation for example 1 with selecting 2 pairs of breeding parents in each generation based on the heuristic NPCV approach is presented in table 4.2.

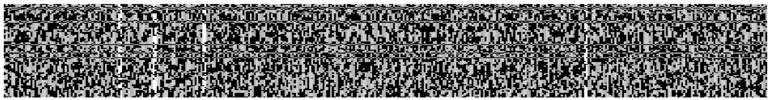
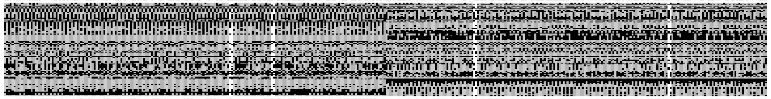
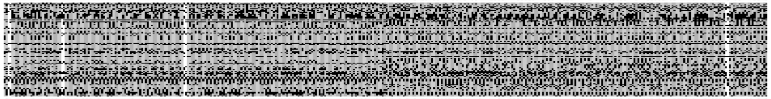
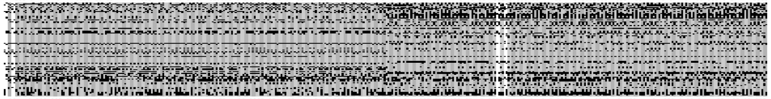
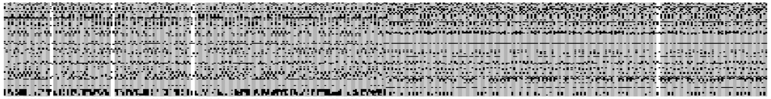
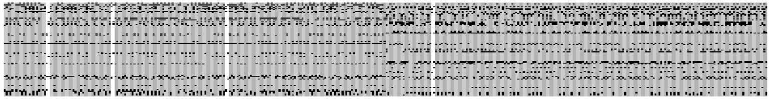
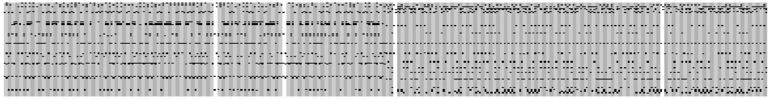
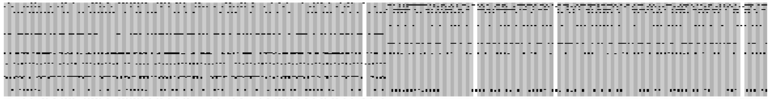
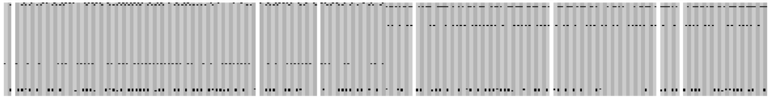
Generation	Population	$\log_{10}$ `PCV`
2		–21.76
3		–16.98
4		–13.56
5		–10.22
6		–7.49
7		–5.26
8		–2.77
9		–0.55
10		0

Table 4.2: A Random Simulation Run for Example 1 with 2 Pairs of Breeding Parents per Generation with the NPCV Approach.

4.4.2 Results for Example 1

Figure 4.4 plots the bar plots of number of generations to terminate across all the simulations for Example 1. Observed from the figure, we can derive the following conclusions that the NPCV metric based approach has a greater chance to terminate 2 to 3 generations earlier compared with the GEBV metric based approach. This result demonstrates that the NPCV approach outperforms the GEBV approach in terms of the efficiency of the breeding process. We also compare the effectiveness between two approaches in the following Table 4.3.

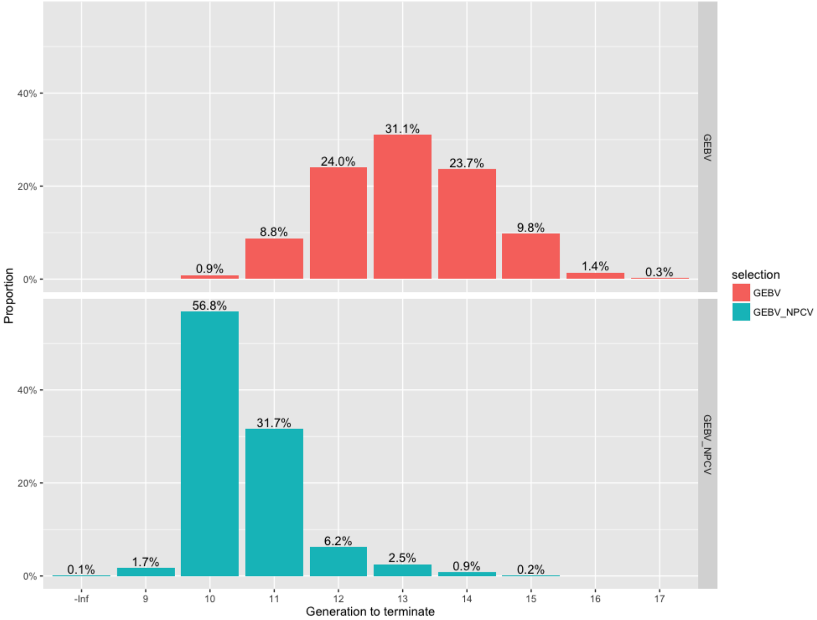


Figure 4.4: Example 1: Combined number of generations to terminate. Performance of the GEBV and the heuristic NPCV approaches in all the simulation runs of trait introgression of seven QTLs. The vertical axis represents the proportion of number of generations to terminate. The horizontal axis represents the possible number of generations to terminate.

Table 4.3 presents the average number of generations to succeed for each choice of number of pairs of breeding parents. Also, the table presents the proportion of simulations to successfully achieve all the desirable alleles in the last generation, respectively. Reading from the table, we can draw the following two major conclusions. First, the more pairs of breeding parents are selected in each generation, the earlier the introgression process succeeds. Second, both methods could achieve a relative high probability of success. We shall notice that when we select only 2 or 3 pairs of breeding parents, the performance of the GEBV approach is better.

The reason is that the heuristic approach of NPCV is determined by the initial result from the set cover model of the GEBV approach. In the iterative updates, we only search a small portion of all the possible combinations and it is possible that the initial starting point can not guarantee a promising candidate. However, generally speaking, the NPCV method is better compared with the GEBV approach and with sufficient number of pairs of breeding parents selected in each generation, the NPCV approach could lead the probability of success very close to 1.

Result Summary				
Pairs	G_{GEBV}	R_{GEBV}	G_{GEBV_NPCV}	R_{GEBV_NPCV}
2	13.45	0.98	11.28	0.73
3	13.35	0.96	10.98	0.91
4	13.35	0.97	10.79	0.96
5	13.19	0.94	10.46	0.99
6	13.15	0.94	10.34	1.00
7	12.93	0.98	10.22	0.99
8	12.75	0.94	10.24	1.00
9	12.64	0.94	10.07	1.00
10	12.69	0.93	10.07	1.00

Table 4.3: Simulation Result Summary of Example 1

Figure 4.5 presents the box plots of the proportion of desirable alleles among the selected breeding parents in each generation, which could be considered as the approximation of the distributions of the proportion of desirable alleles. This figure takes selecting 6 pairs of breeding parents in each generation as an example. We can observe that before generation 7, the performance of the GEBV approach is better compared with the NPCV approach. After generation 7, the improvement in each generation of the NPCV approach is greater than the GEBV approach and the performance of the NPCV approach is better. In general, the improvements of the GEBV approach has a exponential trend while the trend of the NPCV approach seems to be linear.

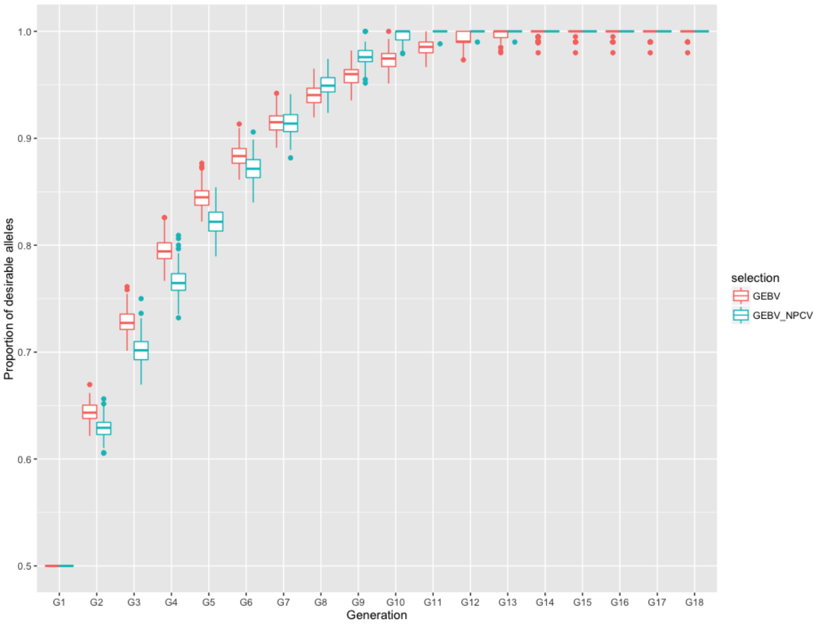


Figure 4.5: Example 1: Proportion of desirable alleles in each generation with 6 pairs of breeding parents selected. The box plots for the average proportion of desriable alleles in the 6 pairs of selected breeding parents with the GEBV and the heuristic NPCV approaches in each generation. The vertical axis represents the proportion of desirable alleles in the 6 pairs of selected breeding parents. The horizontal axis represents different generations.

4.4.3 Results for Example 2

Figure 4.6 plots the bar plots of number of generations to terminate across all the simulations for Example 2. Observed from the figure, we can derive the following conclusions that the NPCV metric based approach has a greater chance to terminate 4 to 5 generations earlier compared with the GEBV metric based approach. For 20 alleles to introgress, the advantage of the NPCV approach is more significant compared with the GEBV approach. This result demonstrates that the NPCV approach outperforms the GEBV approach in terms of the efficiency of the breeding process. We also compare the effectiveness between two approaches in the following Table 4.4.

Table 4.4 presents the average number of generations to succeed for each choice of number of pairs of breeding parents. Also, the table presents the proportion of simulations to successfully achieve all the desirable alleles in the last generation, respectively. Reading from the table, we can draw similar conclusions. First, the more pairs of breeding parents are selected in each generation, the earlier the introgression process succeeds. Second, both methods could achieve a relative high probability of success. We shall notice that only in the case of 2 pairs of

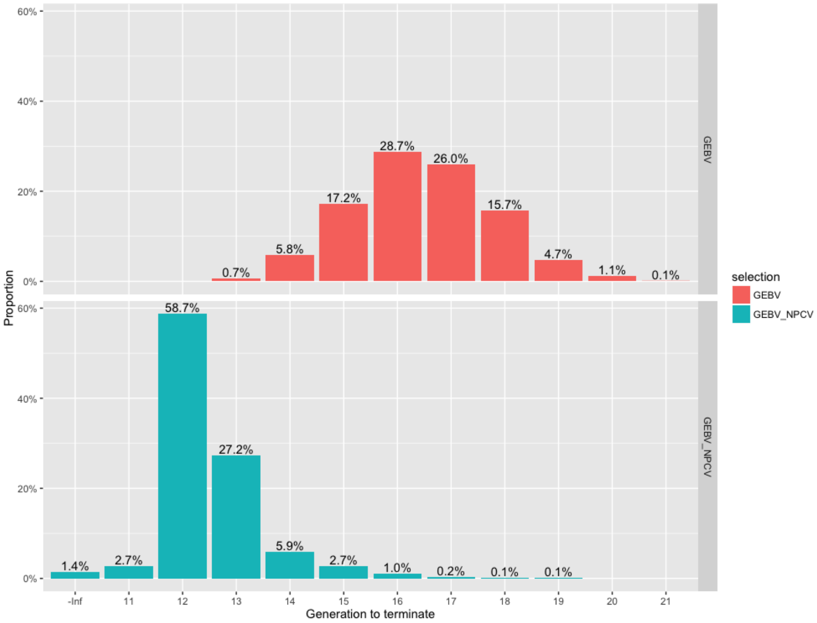


Figure 4.6: Example 2: Combined number of generations to terminate. Performance of the GEBV and the heuristic NPCV approaches in all the simulation runs of trait introgression of seven QTLs. The vertical axis represents the proportion of number of generations to terminate. The horizontal axis represents the possible number of generations to terminate.

breeding parents per generation, the performance of the GEBV approach is better. Generally speaking, the NPCV method is better compared with the GEBV approach and with sufficient number of pairs of breeding parents selected in each generation, the NPCV approach could lead the probability of success very close to 1. In general, we can observe that when we need to introgress more alleles, the advantage of the NPCV approach is more significant in terms of the effectiveness and efficiency.

Figure 4.7 presents the box plots of the proportion of desirable alleles among the selected breeding parents in each generation, which could be considered as the approximation of the distributions of the proportion of desirable alleles. This figure takes selecting 6 pairs of breeding parents in each generation as an example. We can observe that before generation 7, the performance of the GEBV approach is better compared with the NPCV approach. After generation 7, the improvement in each generation of the NPCV approach is greater than the GEBV approach and the performance of the NPCV approach is better. In general, the improvements of the GEBV approach has a exponential trend while the trend of the NPCV approach seems

Result Summary				
Pairs	G_{GEBV}	R_{GEBV}	G_{GEBV_NPCV}	R_{GEBV_NPCV}
2	16.96	0.98	13.73	0.70
3	16.73	0.95	13.02	0.99
4	16.53	0.96	12.83	0.96
5	16.40	0.96	12.31	1.00
6	16.36	0.96	12.31	1.00
7	16.39	0.94	12.18	1.00
8	16.17	0.95	12.17	1.00
9	16.31	0.96	12.00	1.00
10	16.24	0.95	11.97	1.00

Table 4.4: Simulation Result Summary of Example 2

to be linear.

4.5 Discussion

In order to make the PCV metric designed in Chapter 2 more applicable to practical breeding problems, we extend the definition of the PCV to NPCV in this chapter. The NPCV is designed for multi-pair breeding parents selection. With such metric, we are able to deal with the problems such as the desirables are carried by multiple individuals or one pair of breeding parents cannot produce enough offspring progeny. Sharing similar concepts with the PCV, the NPCV takes the recombination frequencies into consideration and selects the pairs of individuals that have the highest probability to yield an ideal gamete in several generations later. With the simulation experiments, we compare the NPCV metric with the conventional GEBV approach. The results from the simulations demonstrate that the NPCV has advantages over the conventional selection approach.

We compare different metrics in terms of time (generations) and the proportion of desirable alleles in the final result. Missing from these criteria is a consideration of cost or the resource allocation through the breeding process with the NPCV. In general, the number of pairs of parents selected in each generation can serve as a surrogate for cost and in future research we will look at the relative impacts of number of pairs of parents for each generation of evaluation. Another potentially fruitful topic for future research is compete the design of NPCV. In this

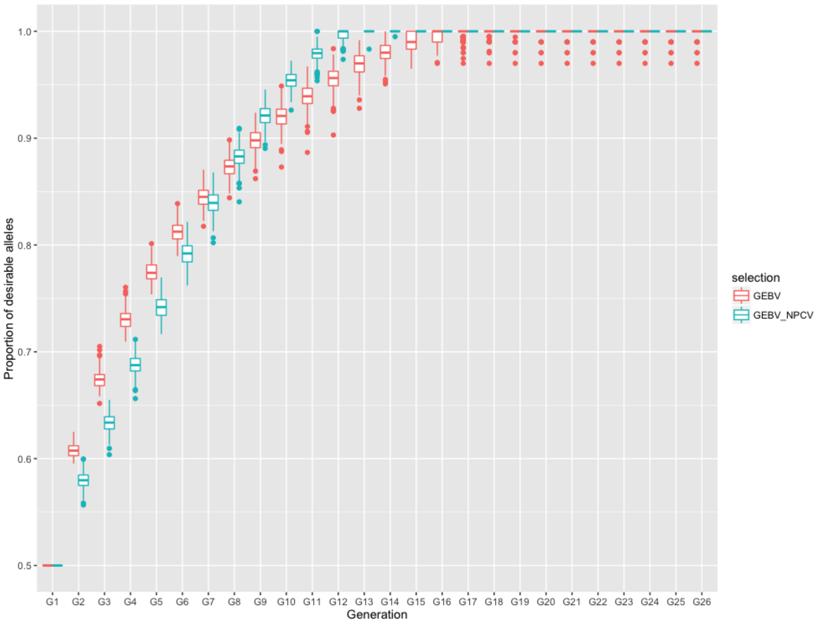


Figure 4.7: Example 1: Proportion of desirable alleles in each generation with 6 pairs of breeding parents selected. The box plots for the average proportion of desriable alleles in the 6 pairs of selected breeding parents with the GEBV and the heuristic NPCV approaches in each generation. The vertical axis represents the proportion of desirable alleles in the 6 pairs of selected breeding parents. The horizontal axis represents different generations.

chapter, we used an approximate transition matrix for calculation and we did not optimize the order of selected breeding parents. Also, since we are selecting multiple pairs of breeding parents, the research on the weight of each pair and the pair up policy design could lead to more promising result. Meanwhile, the special mating procedure like self pollination or back crossing can be possible research topics, as well. The research on those topics will make the definition of NPCV more comprehensive and rigorous.

Another direction that deserves investigation in future research is the exploration of more optimal breeding strategies. As mentioned before, the trait introgression breeding problem formulated in this chapter, even with the simplifying assumptions, is too complex to be readily global solvable by existing optimization methodology. Although the PCV and the NPCV based introgression outperforms those approaches based on conventional metrics, it is still unclear to us how much further improvement could be approached. Thus, we shall see the great potential on applying operations research techniques to trait introgression project to improve plant breeding.

CHAPTER 5. SUMMARY

This dissertation focuses on applying operations research approaches to solve the problems in multi-allelic trait introgression process in plant breeding and improve its efficiency and effectiveness. It consists of three major parts, which are the "Predicted Cross Value" design for parental selection problem, Markov decision processes based breeding strategy design for dynamically allocating resources and the extended NPCV and set cover models for multi-pair parental selection problem.

The first part proposed the formulation of the multi-allelic introgression problem capturing the mathematical essence of the process. Using time (generations) and probability of success as criteria provides objective measurable criteria for comparing breeding strategies. At the same time, this chapter proposed the PCV, which is a new metric for selection of parents. Rather than sticking to predetermined breeding strategies such as backcrossing, as widely used for trait introgression, PCV based selection identifies the pair of individuals whose complementary genotypes have the highest probability to yield an ideal gamete in two generations. The simulations demonstrated that the PCV outperforms the existing approaches GBV and OHV.

In the second part, we continued the discussion on the Multi-allelic Trait Introgression process and we completed the work flow of MATI process by adding the components of resource allocation. Then, we formulated the mathematical algorithm to simulate the MATI process and addressed the resource allocation problem in the MATI process. To solve this problem, we used the Markov decision processes model and we demonstrated the improvement brought by dynamically allocating the budgets.

In the third part, we continued the discussion on parental selection of multi-allelic introgression process, in which the previous PCV metric may not work very well. We modified our existing PCV metric in order to quantify the parental selection for multiple pairs of breeding parents. We defined the NPCV metric and updated the plumbing system for calculation. We built set cover models for different metrics to select multiple the optimal pairs of breeding parents to cover all the desirable alleles and proposed a heuristic approach for the NPCV metric. The simulation based case studies demonstrated the advantage of the NPCV over the conventional metric.

In general, the trait introgression problem in plant breeding is more and more complex as well as attractive when we deep dive into this topic. We believe that solving such problems will make evolutionary contributions to plant breeding and this topic deserves more attention for designing better solutions.

APPENDIX A. Proof for Chapter 2

Part 1: Lemmas, propositions, and proofs

The following lemma is a straightforward derivation from the definitions in Section 2.1.

Lemma 1. *For all $i \in \{1, ..., N\}$, we have:*

$$g_i = \begin{cases} L_{i,1} & \textit{if } J_i^1 = 0 \textit{ and } J_i^3 = 0; \\ L_{i,2} & \textit{if } J_i^1 = 1 \textit{ and } J_i^3 = 0; \\ L_{i,3} & \textit{if } J_i^2 = 0 \textit{ and } J_i^3 = 1; \\ L_{i,4} & \textit{if } J_i^2 = 1 \textit{ and } J_i^3 = 1. \end{cases} \tag{A.1}$$

The following lemma reveals the rationale behind the definition for the transition matrix.

Lemma 2. *For all $i \in \{1, ..., N-1\}$, $j \in \{1, 2, 3, 4\}$, and $k \in \{1, 2, 3, 4\}$, we have*

$$P(g_{i+1} = L_{i+1,j} | g_i = L_{i,k}) = T_{k,j,i}.$$

Proof. For all $i \in \{1, ..., N-1\}$, we prove the equation for $j = 1$ and $k \in \{1, 2, 3\}$. The proof for the other cases is similar.

$$\begin{aligned} & P(g_{i+1} = L_{i+1,1} | g_i = L_{i,1}) \\ = \; & P(J_{i+1}^1 = 0, J_{i+1}^3 = 0 | J_i^1 = 0, J_i^3 = 0) \\ = \; & P(J_{i+1}^1 = 0 | J_i^1 = 0, J_i^3 = 0) P(J_{i+1}^3 = 0 | J_i^1 = 0, J_i^3 = 0) \\ = \; & P(J_{i+1}^1 = 0 | J_i^1 = 0) P(J_{i+1}^3 = 0 | J_i^3 = 0) \\ = \; & (1 - r_i)^2 \\ = \; & T_{1,1,i}. \end{aligned}$$

$$\begin{aligned}
& P(g_{i+1} = L_{i+1,2} | g_i = L_{i,1}) \\
= \quad & P(J_{i+1}^1 = 1, J_{i+1}^3 = 0 | J_i^1 = 0, J_i^3 = 0) \\
= \quad & P(J_{i+1}^1 = 1 | J_i^1 = 0, J_i^3 = 0) P(J_{i+1}^3 = 0 | J_i^1 = 0, J_i^3 = 0) \\
= \quad & P(J_{i+1}^1 = 1 | J_i^1 = 0) P(J_{i+1}^3 = 0 | J_i^3 = 0) \\
= \quad & r_i(1 - r_i) \\
= \quad & T_{1,2,i}.
\end{aligned}$$

$$\begin{aligned}
& P(g_{i+1} = L_{i+1,3} | g_i = L_{i,1}) \\
= \quad & P(J_{i+1}^2 = 1, J_{i+1}^3 = 0 | J_i^1 = 0, J_i^3 = 0) \\
= \quad & P(J_{i+1}^2 = 1 | J_i^1 = 0, J_i^3 = 0) P(J_{i+1}^3 = 0 | J_i^1 = 0, J_i^3 = 0) \\
= \quad & P(J_{i+1}^2 = 1) P(J_{i+1}^3 = 1 | J_i^3 = 0) \\
= \quad & 0.5 r_i \\
= \quad & T_{1,3,i}.
\end{aligned}$$

□

Proof for Proposition 1:

Proof. We establish the respective equivalence between Equation (4.5) and Equations (4.6)-(4.7) as follows.

Equation (4.5) for $i = 1$ and Equation (4.6) are equivalent because for all $j \in \{1, 2\}$, we have

$$\begin{aligned} W_{1,j} &= P(g_1 = 1, g_1 = L^1_{1,j}) \\ &= P(L^1_{1,j} = 1, g_1 = L^1_{1,j}) \\ &= L^1_{1,j} P(g_1 = L^1_{1,j}) \\ &= L^1_{1,j} P(J^1_1 = j-1, J^3_1 = 0) \\ &= L^1_{1,j} P(J^1_1 = j-1) P(J^3_1 = 0) \\ &= \frac{1}{4} L^1_{1,j}. \end{aligned}$$

The case for $j \in \{3,4\}$ is similar.

Equation (4.5) for $i \in \{2, ..., N\}$ and Equation (4.7) are equivalent because for all $i \in \{2, ..., N\}$ and $j \in \{1,2,3,4\}$, we have

$$\begin{aligned} W_{i,j} &= P(g_1 = ... = g_i = 1, g_i = L_{i,j}) \\ &= P(g_1 = ... = g_{i-1} = 1, g_i = L_{i,j}, L_{i,j} = 1) \\ &= L_{i,j} P(g_1 = ... = g_{i-1} = 1, g_i = L_{i,j}) \\ &= L_{i,j} \sum_{k=1}^{4} P(g_1 = ... = g_{i-1} = 1, g_{i-1} = L_{i-1,k}, g_i = L_{i,j}) \\ &= L_{i,j} \sum_{k=1}^{4} P(g_i = L_{i,j} | g_1 = ... = g_{i-1} = 1, g_{i-1} = L_{i-1,k}) \\ & \quad \times P(g_1 = ... = g_{i-1} = 1, g_{i-1} = L_{i-1,k}) \\ &= L_{i,j} \sum_{k=1}^{4} P(g_i = L_{i,j} | g_{i-1} = L_{i-1,k}) W_{i-1,k} \\ &= L_{i,j} \sum_{k=1}^{4} T_{k,j,i-1} W_{i-1,k}. \end{aligned}$$

□

Proof for Proposition 2:

Proof.

$$\begin{aligned}
\texttt{PCV}(L^1, L^2, r) &= P(g_1 = ... = g_N = 1) \\
&= \sum_{j=1}^{4} P(g_1 = ... = g_N = 1, g_N = L_{N,j}) \\
&= \sum_{j=1}^{4} W_{N,j}.
\end{aligned}$$

□

Part 2: Optimization of PCV

We present an optimization model that can be used to select the optimal pair of individuals with the highest PCV from a given population.

The model takes two parameters as input: the set of progeny of lines $P \in \mathbb{B}^{N\times 2\times K}$ with K being the number of lines and the recombination frequencies vector $r \in [0, 0.5]^{N-1}$. There are three sets of decision variables:

- $t \in \mathbb{B}^{2\times K}$ is a binary variable, indicating whether ($t_{m,k} = 1$) or not ($t_{m,k} = 0$) line k is selected as the mth parent, for all $m \in \{1, 2\}$ and $k \in \{1, ..., K\}$.
- $x \in \mathbb{B}^{N\times 4}$ represents the genotypes of the two selected parents. If $t_{1,k^1} = t_{2,k^2} = 1$, then $x_{:,1:2} = P_{:,:,k^1}$ and $x_{:,3:4} = P_{:,:,k^2}$.
- $w \in \mathbb{B}^{N\times 4}$ is the water matrix of x.

The optimization model is presented in (A.3)-(A.10), which is a mixed integer linear program (MILP). The objective function (A.3) calculates the PCV of the two selected parent lines, which is to be maximized. Constraint (A.4) requires that exactly two breeding parents are selected from the population, which could possibly be the same line. Constraints (A.6) and (A.7) assign the genotypes of the selected lines from the breeding population to the x matrix. Constraints (A.7), (A.8), and (A.9) calculate the water matrix for x. Constraint (A.7) is equivalent to Equation (4.6), and the two linear inequalities (A.8) and (A.9) are equivalence to

$$w_{i,j} \leq x_{i,j} \sum_{k=1}^{4} T_{k,j,i-1} w_{i-1,j}. \tag{A.2}$$

Due to the objective function, inequality (A.2) will hold at equality when the model (A.3)-(A.10) is solved to optimality, which is equivalent to Equation (4.7). Constraint (A.10) defines the types and ranges of the decision variables. This MILP model can be solved to optimality by existing algorithms and software.

$$
\begin{aligned}
\max_{w,x,t} \quad & \sum_{k=1}^{4} w_{N,k} && && \text{(A.3)}\\
\text{s. t.} \quad & \sum_{k=1}^{K} t_{m,k} = 1 && \forall m = 1, 2 && \text{(A.4)}\\
& x_{i,j} = \sum_{k=1}^{K} t_{1,k} P_{i,j,k} && \forall i \in \{1, ..., N\}; \forall j \in \{1, 2\} && \text{(A.5)}\\
& x_{i,j} = \sum_{k=1}^{K} t_{2,k} P_{i,j-2,k} && \forall i \in \{1, ..., N\}; \forall j \in \{3, 4\} && \text{(A.6)}\\
& w_{1,j} = 0.25 x_{1,j} && \forall j \in \{1, 2, 3, 4\} && \text{(A.7)}\\
& w_{i,j} \leq x_{i,j} && \forall i \in \{2, ..., N\}, \forall j \in \{1, 2, 3, 4\} && \text{(A.8)}\\
& w_{i,j} \leq \sum_{k=1}^{4} T_{k,j,i-1} w_{i-1,j}, && \forall i \in \{2, ..., N\}, \forall j \in \{1, 2, 3, 4\} && \text{(A.9)}\\
& 0 \leq w \leq 1; x, t \text{ binary.} && && \text{(A.10)}
\end{aligned}
$$

Alternatively, the optimal selection of breeding parents can be achieved via a brute force enumeration of all possible $\frac{1}{2}n(n+1)$ combinations (excluding symmetric duplications).

APPENDIX B. Proof for Chapter 4

B.1 Transition Matrix

Here, we present the brief derivation of the transition matrix in Chapter 4. For a given vector of recombination frequencies, $r \in [0, 0.5]^{N-1}$, the transition matrix $T \in [0, 0.5]^{4M \times 4M \times (N-1)}$ is defined as

$$T_{4j-3:4j,4j-3:4j,i} \approx \begin{bmatrix} (1-r_i)^t & r_i(1-r_i)^{t-1} & 0.5r_i(1-r_i)^{t-2} & 0.5r_i(1-r_i)^{t-2} \\ r_i(1-r_i)^{t-1} & (1-r_i)^t & 0.5r_i(1-r_i)^{t-2} & 0.5r_i(1-r_i)^{t-2} \\ 0.5r_i(1-r_i)^{t-2} & 0.5r_i(1-r_i)^{t-2} & (1-r_i)^t & r_i(1-r_i)^{t-1} \\ 0.5r_i(1-r_i)^{t-2} & 0.5r_i(1-r_i)^{t-2} & r_i(1-r_i)^{t-1} & (1-r_i)^t \end{bmatrix}, \quad \forall i \in \{1, ..., N-1\}, j \in \{1, ..., M\}. \tag{B.1}$$

where $t = \lceil \log_2(2M) \rceil + 1$ and for the rest elements of $T_{p,q,i} \notin T_{4j-3:4j,4j-3:4j,i}$ in T,

$$T_{p,q,i} \approx \frac{1-(1-r_i)^{t-2}}{4M-4}, \tag{B.2}$$

B.2 Case I: $2M = 2^a, a \in \mathbb{I}^+$

Lemma 3. *Let* $t = \lceil \log_2 2M \rceil + 1$*, for* $2M = 2^a, a \in \mathbb{I}^+$*, we have*

$$P(g_{i+1}^t = L_{i+1,q} | g_i^t = L_{i,p}) = T_{p,q,i}$$

.

Proof. **When** $2M = 2^a$**:**

Without loss of generality, we first prove the equality when $j = 1$ and $T_{p,q,i} \in T_{4j-3:4j,4j-3:4j,i} = T_{1:4,1:4,i}$.

For all $i \in \{1, ..., N-1\}$, we first prove the equation for $p = 1$, i.e., $T_{1,1,i}$,$T_{1,2,i}$, $T_{1,3,i}$ and $T_{1,4,i}$. The proof for the rest elements in $T_{1:4,1:4,i}$ will be similar.

$$\begin{aligned}
& P(g_{i+1}^t = L_{i+1,1} | g_i^t = L_{i,1}) \\
= \; & P(\text{no recombination, generation } 1) P(\text{no recombination, generation } 2) ... \\
& P(\text{no recombination, generation } t-1) P(\text{no recombination,generation } t) \\
= \; & (1 - r_i)^t \\
= \; & T_{1,1,i}.
\end{aligned}$$

$$\begin{aligned}
& P(g_{i+1}^t = L_{i+1,2} | g_i^t = L_{i,1}) \\
= \; & P(\text{recombination, generation } 1) P(\text{no recombination, generation } 2) ... \\
& P(\text{no recombination, generation } t-1) P(\text{no recombination,generation } t) \\
= \; & r_i (1 - r_i)^{t-1} \\
= \; & T_{1,2,i}.
\end{aligned}$$

$$\begin{aligned}
& P(g_{i+1}^t = L_{i+1,3} | g_i^t = L_{i,1}) \\
= \; & P(\text{generation } 1) P(\text{inheriting } L_{i+1,3}) P(\text{recombination, generation } 2) ... \\
& P(\text{no recombination, generation } t-1) P(\text{no recombination, generation } t) \\
= \; & 1 \cdot 0.5 r_i (1 - r_i)^{t-2} \\
= \; & T_{1,3,i}.
\end{aligned}$$

$$\begin{aligned} & P(g_{i+1}^t = L_{i+1,3} | g_i^t = L_{i,1}) \\ = \; & P(\text{generation } 1) P(\text{inheriting } L_{i+1,4}) P(\text{recombination, generation } 2) ... \\ & P(\text{no recombination, generation } t-1) P(\text{no recombination, generation } t) \\ = \; & 1 \cdot 0.5 r_i (1-r_i)^{t-2} \\ = \; & T_{1,4,i}. \end{aligned}$$

For the rest elements of $T_{1,q,i}, q > 4$, because of the random mating design after the first generation, $T_{1,q,i} = \frac{1-(\sum_{q'=1}^{4} T1,q',i)}{4M-4} = \frac{1-(1-r_i)^{G-2}}{4M-4}, q > 4$

□

B.3 Case II: $2M \neq 2^a, a \in \mathbb{I}^+$

For $2M \neq 2^a$, we claim that

$$P(g_{i+1}^t = L_{i+1,q} | g_i^t = L_{i,p}) \approx T_{p,q,i}$$

, for all $i \in \{1, ..., N-1\}$, $p \in \{1, ..., 4M\}$, and $q \in \{1, ..., 4M\}$. Here we give the calculation of two examples as $2M = 6$ and $2M = 10$.

B.3.1 Example 1: $2M = 6$

In this example, the breeding scheme is present in Figure B.1. Based on the scheme, we can derive the following relations.

$$\begin{aligned} & P(g_{i+1}^t = L_{i+1,1} | g_i^t = L_{i,1}) \\ = \; & P(\text{NR, G1})[P(\text{Mating}) P(\text{NR, G2}) P(\text{NR, G3}) P(\text{NR, G4}) + P(\text{Keeping}) P(\text{NR, G3}) P(\text{NR, G4})] \\ = \; & (1-r_i)[\frac{2}{3}(1-r_i)^3 + \frac{1}{3}(1-r_i)^2] \\ = \; & (1-r_i)^3 (1 - \frac{2}{3} r_i) \\ \approx \; & T_{1,1,i}. \end{aligned}$$

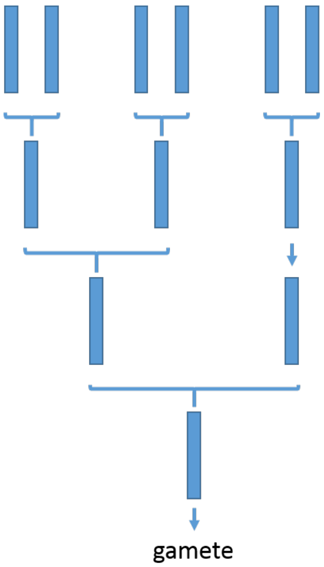


Figure B.1: Conceptual Figure for the Breeding Scheme for 3 Pairs of Parents

In the equation, NR represents no recombination and G represents Generation. Similarly, we can derive the relations for other elements in the transition matrix.

$$
\begin{aligned}
& P(g_{i+1}^t = L_{i+1,2} | g_i^t = L_{i,1}) \\
= \;& P(\text{R, G1})[P(\text{Mating})P(\text{NR, G2})P(\text{NR, G3})P(\text{NR, G4}) + P(\text{Keeping})P(\text{NR, G3})P(\text{NR, G4})] \\
= \;& r_i[\frac{2}{3}(1-r_i)^3 + \frac{1}{3}(1-r_i)^2] \\
= \;& r_i(1-r_i)^2(1-\frac{2}{3}r_i) \\
\approx \;& T_{1,2,i}.
\end{aligned}
$$

$$
\begin{aligned}
& P(g_{i+1}^t = L_{i+1,3} | g_i^t = L_{i,1}) \\
= \;& P(\text{G1})P(\text{Inheriting}L_{i+1,3}) \\
& [P(\text{Mating})P(\text{R, G2})P(\text{NR, G3})P(\text{NR, G4}) + P(\text{Keeping})P(\text{R, G3})P(\text{NR, G4})] \\
= \;& 1 \cdot 0.5[\frac{2}{3}r_i(1-r_i)^2 + \frac{1}{3}r_i(1-r_i)] \\
= \;& 0.5r_i(1-r_i)(1-\frac{2}{3}r_i) \\
\approx \;& T_{1,3,i}.
\end{aligned}
$$

The derivation for $P(g^t_{i+1} = L_{i+1,4}|g^t_i = L_{i,1}) \approx T_{1,4,i}$ is similar as the equations above.

For the rest elements of $T_{1,q,i}, q > 4$, because of the random mating design after the first generation, $T_{1,q,i} \approx \frac{1-(\sum_{q'=1}^{4} T1,q',i)}{4M-4} = \frac{1-(1-r_i)^{G-2}}{4M-4}, q > 4$

B.3.2 Example 1: $2M = 10$

In this example, the breeding scheme is present in Figure B.2. Based on the scheme, we can derive the following relations.

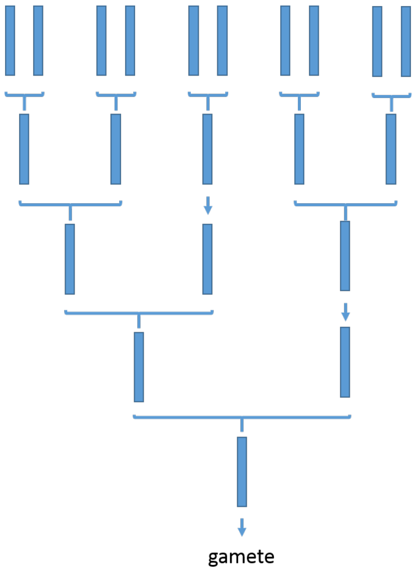


Figure B.2: Conceptual Figure for the Breeding Scheme for 5 Pairs of Parents

$$\begin{aligned}
& P(g^t_{i+1} = L_{i+1,1}|g^t_i = L_{i,1}) \\
= \; & P(\text{NR, G1}) \\
& \{P(Mating)P(NR, G2)[P(\text{Mating})P(\text{NR, G3})P(\text{NR, G4})P(\text{NR, G5}) + P(\text{Keeping})P(\text{NR, G4})P(\text{NR, G5})] + \\
& P(Keeping)[P(\text{Mating})P(\text{NR, G3})P(\text{NR, G4})P(\text{NR, G5}) + P(\text{Keeping})P(\text{NR, G4})P(\text{NR, G5})]\} \\
= \; & (1-r_i)\{\frac{4}{5}(1-r_i)[\frac{2}{3}(1-r_i)^3 + \frac{1}{3}(1-r_i)^2] + \frac{1}{5}[\frac{2}{3}(1-r_i)^3 + \frac{1}{3}(1-r_i)^2]\} \\
= \; & (1-r_i)^3(1-\frac{2}{3}r_i)(1-\frac{4}{5}r_i) \\
\approx \; & T_{1,1,i}.
\end{aligned}$$

In the equation, NR represents no recombination and G represents Generation. Similarly, we can derive the relations for other elements in the transition matrix.

$$\begin{aligned}
& P(g^t_{i+1} = L_{i+1,1}|g^t_i = L_{i,1}) \\
= \; & P(\text{NR, G1}) \\
& \{P(\text{Mating})P(NR, G2)[P(\text{Mating})P(\text{NR, G3})P(\text{NR, G4})P(\text{NR, G5}) + P(\text{Keeping})P(\text{NR, G4})P(\text{NR, G5})] + \\
& P(\text{Keeping})[P(\text{Mating})P(\text{NR, G3})P(\text{NR, G4})P(\text{NR, G5}) + P(\text{Keeping})P(\text{NR, G4})P(\text{NR, G5})]\} \\
= \; & (1-r_i)\{\frac{4}{5}(1-r_i)[\frac{2}{3}(1-r_i)^3 + \frac{1}{3}(1-r_i)^2] + \frac{1}{5}[\frac{2}{3}(1-r_i)^3 + \frac{1}{3}(1-r_i)^2]\} \\
= \; & (1-r_i)^3(1-\frac{2}{3}r_i)(1-\frac{4}{5}r_i) \\
\approx \; & T_{1,1,i}.
\end{aligned}$$

$$\begin{aligned}
& P(g^t_{i+1} = L_{i+1,2}|g^t_i = L_{i,1}) \\
= \; & P(\text{R, G1}) \\
& \{P(\text{Mating})P(NR, G2)[P(\text{Mating})P(\text{NR, G3})P(\text{NR, G4})P(\text{NR, G5}) + P(\text{Keeping})P(\text{NR, G4})P(\text{NR, G5})] + \\
& P(\text{Keeping})[P(\text{Mating})P(\text{NR, G3})P(\text{NR, G4})P(\text{NR, G5}) + P(\text{Keeping})P(\text{NR, G4})P(\text{NR, G5})]\} \\
= \; & r_i\{\frac{4}{5}(1-r_i)[\frac{2}{3}(1-r_i)^3 + \frac{1}{3}(1-r_i)^2] + \frac{1}{5}[\frac{2}{3}(1-r_i)^3 + \frac{1}{3}(1-r_i)^2]\} \\
= \; & r_i(1-r_i)^2(1-\frac{2}{3}r_i)(1-\frac{4}{5}r_i) \\
\approx \; & T_{1,1,i}.
\end{aligned}$$

$$\begin{aligned}
& P(g^t_{i+1} = L_{i+1,3}|g^t_i = L_{i,1}) \\
= \; & P(\text{G1})P(\text{Inheriting}L_{i+1,3}) \\
& \{P(\text{Mating})P(R, G2)[P(\text{Mating})P(\text{NR, G3})P(\text{NR, G4})P(\text{NR, G5}) + P(\text{Keeping})P(\text{NR, G4})P(\text{NR, G5})] + \\
& P(\text{Keeping})[P(\text{Mating})P(\text{R, G3})P(\text{NR, G4})P(\text{NR, G5}) + P(\text{Keeping})P(\text{R, G4})P(\text{NR, G5})]\} \\
= \; & 1 \cdot 0.5\{\frac{4}{5}r_i[\frac{2}{3}(1-r_i)^3 + \frac{1}{3}(1-r_i)^2] + \frac{1}{5}[\frac{2}{3}r_i(1-r_i)^2 + \frac{1}{3}r_i(1-r_i)]\} \\
= \; & 0.5r_i(1-r_i)(1-\frac{2}{3}r_i)(1-\frac{4}{5}r_i) \\
\approx \; & T_{1,3,i}.
\end{aligned}$$

The derivation for $P(g_{i+1}^{t} = L_{i+1,4} | g_{i}^{t} = L_{i,1}) \approx T_{1,4,i}$ is similar as the equations above.

For the rest elements of $T_{1,q,i}, q > 4$, because of the random mating design after the first generation, $T_{1,q,i} \approx \frac{1-(\sum_{q'=1}^{4} T1,q',i)}{4M-4} = \frac{1-(1-r_i)^{G-2}}{4M-4}, q > 4$

B.3.3 Discussion

From the two examples above, we can observe that when $2M \neq 2^a, a \in \mathbb{I}^+$, the equation for $2M = 2^a$ could be a good estimation. Also, for the elements $T_{p,q,i} \in T_{4j-3:4j,4j-3:4j,i}$ in T, the estimation will be a lower bound and for the elements $T_{p,q,i} \notin T_{4j-3:4j,4j-3:4j,i}$ in T, the estimation will be a upper bound. Meanwhile, even based on the estimation, the simulation exactly simulates the process of trait introgression and the results from the simulation are valid.

The exact derivation for this general case could be a future research topic.

BIBLIOGRAPHY

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

Cameron, J. N., Han, Y., Wang, L., and Beavis, W. D. (2017). Systematic design for trait introgression projects. *Theoretical and Applied Genetics*, pages 1993–2004.

Canzar, S. and El-Kebir, M. (2011). A mathematical programming approach to marker-assisted gene pyramiding. *Algorithms in Bioinformatics*, pages 26–38.

Cavanagh, C. R., Chao, S., Wang, S., Huang, B. E., Stephen, S., Kiani, S., Forrest, K., Saintenac, C., Brown-Guedira, G. L., Akhunova, A., et al. (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proceedings of the National Academy of Sciences*, 110(20):8057–8062.

Chong, K. C., Henderson, S. G., and Lewis, M. E. (2015). The vehicle mix decision in emergency medical service systems. *Manufacturing & Service Operations Management*.

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

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

Duvick, D. (1994). volume 20, chapter Maize breeding: Past, present and future, pages 170–179.

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

Frisch, M., Bohn, M., and Melchinger, A. E. (1999). Comparison of selection strategies for marker-assisted backcrossing of a gene. *Crop Science*, 39(5):1295–1301.

Frisch, M. and Melchinger, A. E. (2005). Selection theory for marker-assisted backcrossing. *Genetics*, 170(2):909–917.

Gorjanc, G., Jenko, J., Hearne, S. J., and Hickey, J. M. (2016). Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. *BMC Genomics*, 17(1):1.

Haldane, J. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics*, 8(29):299–309.

Halpin, C. (2005). Gene stacking in transgenic plants–the challenge for 21st century plant biotechnology. *Plant biotechnology journal*, 3(2):141–155.

Han, Y., Cameron, J. N., Wang, L., and Beavis, W. D. (2017). The predicted cross value for genetic introgression of multiple alleles. *Genetics*, 205(4):1409–1423.

Kumar, G. R., Sakthivel, K., Sundaram, R. M., Neeraja, C. N., Balachandran, S., Rani, N. S., Viraktamath, B., and Madhav, M. (2010). Allele mining in crops: prospects and potentials. *Biotechnology Advances*, 28(4):451–461.

Leung, H., Raghavan, C., Zhou, B., Oliva, R., Choi, I. R., Lacorte, V., Jubay, M. L., Cruz, C. V., Gregorio, G., Singh, R. K., et al. (2015). Allele mining and enhanced genetic recombination for rice breeding. *Rice*, 8(1):1.

Longin, C. F. H. and Reif, J. C. (2014). Redesigning the exploitation of wheat genetic resources. *Trends in Plant Science*, 19(10):631–636.

McCouch, S. R., McNally, K. L., Wang, W., and Hamilton, R. S. (2012). Genomics of gene banks: A case study in rice. *American Journal of Botany*, 99(2):407–423.

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

Peng, T., Sun, X., and Mumm, R. H. (2014a). Optimized breeding strategies for multiple trait integration: I. Minimizing linkage drag in single event introgression. *Molecular Breeding*, 33(1):89–104.

Peng, T., Sun, X., and Mumm, R. H. (2014b). Optimized breeding strategies for multiple trait integration: II. rocess efficiency in event pyramiding and trait fixation. *Molecular Breeding*, 33(1):105–115.

Poehlman, J. M. (2013). *Breeding field crops.* Springer Science & Business Media.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons.

Rincker, K., Nelson, R., Specht, J., Sleper, D., Cary, T., Cianzio, S. R., Casteel, S., Conley, S., Chen, P., Davis, V., et al. (2014). Genetic improvement of US soybean in maturity groups II, III, and IV. *Crop Science*, 54(4):1419–1432.

Servin, B., Martin, O. C., Mézard, M., et al. (2004). Toward a theory of marker-assisted gene pyramiding. *Genetics*, 168(1):513–523.

Swami, S., Puterman, M. L., and Weinberg, C. B. (2001). Play it again, sam? optimal replacement policies for a motion picture exhibitor. *Manufacturing & Service Operations Management*, 3(4):369–386.

USDA (2017). Long-term agricultural projections. *https://www.usda.gov/oce/commodity/projections/*.

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

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