ABSTRACT We consider the plant genetic improvement challenge of introgressing multiple alleles from a homozygous donor to a recipient. First, we frame the project as an algorithmic process that can be mathematically formulated. We then introduce a novel metric for selecting breeding parents that we refer to as the Predicted Cross Value (PCV). Unlike estimated breeding values, which represent predictions of general combining ability, the PCV predicts specific combining ability. The PCV takes estimates of recombination frequencies as an input vector and calculates the probability that a pair of parents will produce a gamete with desirable alleles at all specified loci. We compared the PCV approach with existing estimated breeding value 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. We also discuss how operations research can be used for other crop genetic improvement projects and suggest several future research directions.

KEYWORDS predicted cross value; trait introgression; gene stacking; parental selection; operations research

1. Introduction

Discoveries 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 desirable phenotypes in large germplasm repositories (McCouch et al. 2012; Cavanagh et al. 2013). This information tells us that desirable alleles are distributed unevenly throughout germplasm collections and unevenly across crop genomes. None-the-less 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) (Meuwissen et al. 2001) based on large sets of genomic markers is commonly used for parental selection and has been proposed for trait introgression projects in crop species (Bernardo 2009). The optimal haploid value (OHV) (Daetwyler et al. 2015) was proposed as an alternative breeding value metric to evaluate potential rather than a realized breeding value among progeny. Both of these metrics may be thought of as predictors of general combining ability for an inference space consisting of a breeding population. Prior to development of the GEBV, van Berloo and Stam (1998) proposed a combination index (CI) to identify pairs of recombinant inbred lines (RILs) derived from a single mating of two homozygous lines for purposes of accumulating desirable QTL in subsequent generations of breeding. Their CI metric may be thought of as a predictor of specific combining ability among sets of RILs within the family derived from the initial cross. While the CI was evaluated in simulated genomes with various degrees of linkage disequilibrium (LD) among QTL, none...
of these metrics explicitly use recombination or linkage phase information. Recently, Bonk et al. (2016) developed breeding value models for dairy herds that utilize linkage map information to identify matings that will produce progeny with large variances of breeding values. The premise of this approach was that a higher Mendelian sampling variance for a given pair of parents would lead to a higher chance to generate outstanding progeny with many desirable alleles.

Herein, we build on the concepts introduced by van Berloo and Stam (1998) as well as those of Bonk et al. (2016) and propose a metric, the Predicted Cross Value (PCV) for selecting specific crosses. We apply the PCV to the process of introgressing multiple alleles from a single homozygous donor. However, rather than applying the metric to RIL's derived from a single cross, we apply it to any arbitrary set of progeny derived from crosses in multiple generations required for introgressing multiple alleles from a donor. The PCV is a predictor of specific combining ability for an inference space consisting of future sets of progeny derived from possible crosses in a breeding population. Explicitly, the PCV calculates the probability that a cross will produce an ideal genotype in two future generations of matings.

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 when the goal is to adapt a tropical cultivar to high latitudes for purposes of evaluating other agronomic traits without confounding influences of maturity. b) Introgression of twenty alleles from an exotic accession into an elite cultivar for purposes of improving a polygenic trait.

2. Formulation

The general objective of multi-allelic introgression projects is to transfer multiple desirable alleles, or haplotypes, from a donor to a recipient. The ultimate goal is to produce at least one individual with a genome consisting of homozygous desirable haplotypes and no undesirable 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.

2.1. Multi-allelic introgression as an algorithmic process

We illustrate the major components of the introgression process in Figure 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 their genomes. The majority of alleles in the recipient are desirable but some are undesirable. The donor carries the desirable versions of alleles that the recipient is lacking, but the rest of the alleles of the donor are undesirable.

- **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 determines whether the current 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 cultivar development.
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, sorghum, sunflower, rice, and wheat with subgenome specific loci.
- Consider a single multi-allelic trait, where all segregating loci associated with the trait are known. Results also apply to multiple traits such as gene stacks 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.
- To illustrate the principles, all desirable alleles missing in the recipient are carried by one donor line.
- One pair of parents is selected for crossing in each generation, with self-pollination as a special but feasible option.
- To realistically represent the actual gamete formation process, the input binary vector \( J \) for the alleles in a gamete. If \( J \) must be a random one following a special distribution, \( \mathbb{N} \) is equivalent to the inheritance distribution with parameter vector \( r \in [0,0.5]^{N-1} \) if \( J_1 = (J_{i-1}) \) is 1 − \( r_{i-1} \), which explains Equation (2).

According to Mendel’s second law, \( L_{1,1} \) and \( L_{1,2} \) are equally likely to transmit \( 1 \), hence Equation (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 (2).

Definition 0.3. We define the \( \text{Reproduce} \) function, \( X = \text{reproduce}(L^1, L^2, r, K) \), 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 vector \( r \), denoted as \( J_{p}, \forall p \in \{1,..,2\} \), and then setting \( X_{i,j,k} = \text{Genome}(L_{j,1+2k}), \forall i \in \{1,..,N\}, j \in \{1,2\}, k \in \{1,..,K\} \).

Definition 0.4. The \( \text{Select} \) function, \( [k_1, k_2] = \text{Select}(X, r) \), 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 0.5. We define the \( \text{Breed} \) function as \( G = \text{breed}(P^0, r, K) \). 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 \( \max \left\{ \sum_{i=1}^{N} (P_{i,1,k} + P_{i,2,k}) \right\} = 2N \)

RETURN: \( G = t \).

Else go to Step 2.

Step 2 (Selection) Obtain \( [k_1, k_2] = \text{Select}(P^t, r) \) and go to Step 3.

Step 3 (Reproduction) Obtain \( P^{t+1} = \text{Reproduce}(P_{i,k_1,1}, P_{i,k_2,2}, r, K) \), update \( t \leftarrow t + 1 \), and go to Step 1.

The function \( \text{Breed}(P^0, r, K) \) 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.

2.4. Existing approaches for parental selection

The genomic estimated breeding value approach selects breeding parents based on the GEBV. 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:

\[
\sum_{i=1}^{N} (L_{i,1} + L_{i,2}).
\]

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 interpreted as the best doubled-haplod progeny that could possibly be produced by selfing such individual. As such, OHV measures the potential breeding value of the individual’s progeny. In the context of multi-allelic introgression, the OHV of an individual \( L \) is defined as:

\[
\sum_{i=1}^{N} 2 \max\{L_{i,1}, L_{i,2}\}. \tag{4}
\]

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

### 3. PCV for parental selection

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

#### 3.1. Definition of PCV

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

**Definition 0.6.** For a given pair of individuals \( L_{1} \) and \( L_{2} \), the predicted cross value is defined as the probability that a gamete, \( g^{3} \), produced by a random progeny from crossing these two individuals will consist only of desirable alleles:

\[
\text{PCV}(L_{1}, L_{2}, r) = P(g^{3} = 1, \forall i \in \{1, ..., N\}).
\]

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.

#### 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 \( L_{i,1} = L_{i,1}', L_{i,2} = L_{i,2}', L_{i,3} = L_{i,3}', \) and \( L_{i,4} = L_{i,2}' \) 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 0.7.** 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

\[
T_{k,j,i} = \begin{bmatrix}
(1 - r_{i}) & r_{i} & 0.5r_{i} & 0.5r_{i} \\
0.5r_{i} & 0.5r_{i} & (1 - r_{i}) & r_{i} \\
0.5r_{i} & 0.5r_{i} & (1 - r_{i}) & (1 - r_{i}) \\
0.5r_{i} & 0.5r_{i} & 0.5r_{i} & 0.5r_{i}
\end{bmatrix},
\]

where \( r_{i} \) is the recombination frequency of the \( i \)-th valve, and \( T_{k,j,i} \) is the volume of the pipe that connects valves \((i, k)\) and \((i + 1, j)\) 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 0.8.** We define the water matrix \( W \in [0, 1]^{N \times 4} \) as

\[
W_{ij} = P(g_{1} = ... = g_{i} = 1, g_{i + 1} = L_{ij}), \forall i \in \{1, ..., N\}, j \in \{1, 2, 3, 4\}.
\]

**Proposition 0.1.** The water matrix can be calculated as follows.

\[
W_{ij} = \frac{1}{4} L_{i,j}, \forall j \in \{1, 2, 3, 4\};
\]

\[
W_{ij} = L_{ij} \sum_{k=1}^{4} T_{k,j,i} W_{i-1,k}, \forall i \in \{2, ..., N\}, j \in \{1, 2, 3, 4\}.
\]

**Proposition 0.2.** The PCV is the summation of the last row in the water matrix:

\[
\text{PCV}(L_{1}, L_{2}, r) = \sum_{j=1}^{4} W_{N,j}.
\]

The proofs for Propositions 0.1 and 0.2 can be found in Appendix A.

#### 3.3. Illustrative example

We illustrate the plumbing system with the following example.

**Example 0.1.** The two breeding parents are both ideal lines

\[
L_{1} = L_{2} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}
\]

and the recombination frequencies vector

\[
r = \begin{bmatrix} 0.2 & 0.35 & 0.3 & 0.4 & 0.25 \end{bmatrix}^{	op}.
\]
The plumbing system corresponding to Example 0.1 is illustrated in Figure 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 0.1). 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, where all the valves are open, and thus 100% of the water that is poured in will find its way out.

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

Example 0.2. The two breeding parents are $L^1 = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}$ and $L^2 = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$ and the recombination frequencies vector is the same as in Example 0.1.

The plumbing system corresponding to Example 0.2 is illustrated in Figure 3, in which we removed those water pipes whose immediate upstream valves are closed. The transition matrix is $T_{1,1} = \begin{bmatrix} 0.64 & 0.16 & 0.10 & 0.10 \\ 0.16 & 0.64 & 0.10 & 0.10 \\ 0.10 & 0.10 & 0.64 & 0.16 \\ 0.10 & 0.10 & 0.16 & 0.64 \end{bmatrix}$, $T_{1,2} = \begin{bmatrix} 0.4225 & 0.2275 & 0.1750 & 0.1750 \\ 0.2275 & 0.4225 & 0.1750 & 0.1750 \\ 0.1750 & 0.1750 & 0.4225 & 0.2275 \\ 0.1750 & 0.1750 & 0.2275 & 0.4225 \end{bmatrix}$, $T_{1,3} = \begin{bmatrix} 0.49 & 0.21 & 0.15 & 0.15 \\ 0.21 & 0.49 & 0.15 & 0.15 \\ 0.15 & 0.15 & 0.49 & 0.21 \\ 0.15 & 0.15 & 0.21 & 0.49 \end{bmatrix}$, $T_{1,4} = \begin{bmatrix} 0.36 & 0.24 & 0.20 & 0.20 \\ 0.24 & 0.36 & 0.20 & 0.20 \\ 0.20 & 0.20 & 0.36 & 0.24 \\ 0.20 & 0.20 & 0.24 & 0.36 \end{bmatrix}$, and $T_{1,5} = \begin{bmatrix} 0.5625 & 0.1875 & 0.1250 & 0.1250 \\ 0.1875 & 0.5625 & 0.1250 & 0.1250 \\ 0.1250 & 0.1250 & 0.5625 & 0.1875 \\ 0.1250 & 0.1250 & 0.1875 & 0.5625 \end{bmatrix}$, and the water ma-
All the individuals in the sample of progeny are displayed in two generations. We use the following simple example to demonstrate the conceptual distinctions of PCV, GEBV, and OHV.

Example 0.3. 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 4. A purple square is used to denote a “0” allele and a yellow 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 purple and yellow squares. 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

\[
\begin{bmatrix}
0 & 0 \\
1 & 1 \\
1 & 1 \\
1 & 0 \\
1 & 1 \\
1 & 1 \\
0 & 1 \\
0 & 0 \\
1 & 1 \\
1 & 1 \\
\end{bmatrix}
\]

The recombination frequencies vector used in this example is

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

The PCV is

\[
W = \begin{bmatrix}
0.2500 & 0.2500 & 0 & 0.2500 \\
0.2250 & 0 & 0.0900 & 0.2100 \\
0.1476 & 0.1037 & 0.1252 & 0 \\
0.0369 & 0.0490 & 0.0355 & 0 \\
0 & 0 & 0.0307 & 0.0174 \\
\end{bmatrix}.
\]

Therefore, the PCV is \(0 + 0 + 0.0307 + 0.0174 = 0.0481\).

4. Conceptual distinctions of PCV, GEBV, and OHV

The fundamental difference between these three metrics is that GEBV and OHV assess the merit of two breeding parents as the summation of their separate breeding values, whereas PCV treats two breeding parents as a unique pair and selects the best pair that has the highest probability to produce an ideal gamete in two generations. We use the following simple example to demonstrate the conceptual distinctions of the three metrics.

The GEBVs of the 50 individuals are calculated using Equation (3) and plotted in Figure 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.

4.2. Solving Example 0.3 using the OHV approach

The OHV of the 50 individuals are calculated using Equation (4) and plotted in Figure 6. 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.

4.3. Solving Example 0.3 using the PCV approach

The PCVs of the 50 individuals are calculated using Equations (5) and (7)-(9) and plotted in Figure 7. The two subfigures at the top and left are the same population from Figure 4 with horizontal and vertical orientations, respectively. The largest subfigure is a PCV map. It consists of a 50 × 50 grey-shade matrix representing all PCV values for all possible pairs of breeding parents involving the 50 individuals in the population. As such, each 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)
Figure 7 The PCV map for Example 0.3.

The highest PCV is achieved by individuals 5 and 18, which are respectively highlighted in green and pink 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 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 0.3, 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 all have a unique PCV. Therefore, the PCV map in Figure 7 has 565 different shades of grey. In contrast, there is a large number of tied GEBVs and OHVs in the example.

5. Simulation experiments

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

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 desirable alleles at 93 of the QTL, while the donor has desirable alleles at the remaining 7. For reference the recipient has undesirable alleles at C1Q4, C1Q6, C2Q9, C3Q1, C5Q4, C6Q3, and C6Q8, where CiQj denotes the jth QTL in chromosome i. In the second example the recipient has desirable alleles at 80 of the loci, while the donor has desirable alleles at the remaining 20. For reference the recipient has undesirable alleles at C1Q5, C1Q10, C2Q4, C2Q9, C3Q5, C4Q3, C4Q8, C5Q3, C5Q8, C6Q2, C6Q7, C6Q12, C7Q5, C7Q9, C8Q3, C8Q8, C9Q5, C9Q9, and C10Q3.

Recombination frequencies used in the simulation are given in Table 1. The value shown for column Ci and row Qj is the recombination frequency between the QTL pairs CiQj and C(i + 1)Qj. Such value for the last QTL in a chromosome is 0.5, in accordance with the principle of independent assortment of chromosomes.

We implemented the Breed function to simulate the introgression project, with the simulated genomes as the initial population for each example. In subsequent 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. Four versions of the Select function were compared:

- 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 and results were generated using GNU Octave (Eaton et al. 2015).

5.2 Results for Example 1

Figure 8 plots the population histograms of the proportion of desirable alleles in the genome 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. When GEBV and OHV are used, the average proportions of desirable alleles reach 95% and 93% in the 19th generation, respectively. When PCV-I and PCV-II are used, an ideal progeny will be produced in as early as 7 generations and no later than the 11th.

Figure 9 compares the probability distributions of the terminal generation numbers of 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 self pollination during the breeding process increased the efficiency of the project by half a generation in this example.

5.3. Results for Experiment 2.

Figures 10 and 11 reveal similar results of the four selection approaches as Figures 8 and 9, but as expected all approaches
take more generations to plateau. Using the GEBV and OHV approaches, the proportions of desirable alleles reach 86% and 90% in the 19th generation, 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.

Figure 10 Performance of the GS, OHV, PCV-I, and PCV-II approaches in 1000 simulation runs of trait introgression of twenty QTL.

6. Discussion

With a few exceptions (Johnson et al. 1988; Xu et al. 2011; Canzar and El-Kebir 2011; De Beukelaer et al. 2015; Akdemir and Sanchez 2016) genetic improvement projects in crop species have not been approached using the tools of operations research. On the other hand, the optimization challenge of maximizing genetic gain while minimizing inbreeding in animal breeding has been approached extensively using linear programming, genetic algorithms, integer programming, and semidefinite programming (Woolliams et al. 2015). We formulated the multi-allelic introgression challenge as a mathematical process and we hope that it will attract other operations researchers, applied mathematicians, and computational scientists to contribute to genetic improvement projects.

Figure 11 Distributions of the terminal generation numbers of PCV-I and PCV-II approaches.

We framed the challenge of introgressing multiple alleles from a homozygous donor to a homozygous recipient using time (number of generations) and proportions of desirable alleles in each generation of progeny as quantitatively 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.

A meaningful future work would be to conduct a comprehensive comparison with more recently invented selection approaches that appear to be very promising. For example, Woolliams et al. (2015) proposed the optimal contribution selection, which attempts to maximize genetic gain in the next generation. Akdemir and Sanchez (2016) proposes genomic mating as an alternative to genomic selection, which “uses concepts of estimated breeding values, risk (usefulness) and coefficient of ancestry to optimize mating between parents”. They also discussed alternative risk measures that include linkage disequilibrium information.

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 stack-
The selection approach demonstrated in Section 5 applies to species that can produce a large number of progeny from a cross (such as corn, sunflower, and sorghum). Two strategies can be used to apply this approach to less productive species (such as soybean, wheat, and peanut). One is to compensate the small progeny size with a large number of crosses by selecting multiple pairs of parents using the PCV metric repeatedly. The other is to extend the PCV definition to select multiple pairs of parents and explicitly include the numbers of crosses and expected progeny as parameters, which is one of our future research topics.

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.

Appendix A: Lemmas and proofs

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

Lemma 0.1. For all \( i \in \{1, \ldots, N\} \), we have:

\[
\gamma_i = \begin{cases} 
L_{i,1} & \text{if } f^1_i = 0 \text{ and } f^3_i = 0; \\
L_{i,2} & \text{if } f^1_i = 1 \text{ and } f^3_i = 0; \\
L_{i,3} & \text{if } f^2_i = 0 \text{ and } f^3_i = 1; \\
L_{i,4} & \text{if } f^2_i = 1 \text{ and } f^3_i = 1.
\end{cases} \tag{10}
\]

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

Lemma 0.2. For all \( i \in \{1, \ldots, N-1\} \), \( j \in \{1,2,3,4\} \), 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 \( i \in \{1, \ldots, N-1\} \), we prove the equation for \( j = 1 \) and \( k \in \{1,2,3\} \). The proof for the other cases is similar.

\[
P(g_{i+1} = L_{i+1,1} | g_i = L_{i,1}) = P(f^1_{i+1} = 0, f^2_{i+1} = 0 | f^1_i = 0, f^3_i = 0)
= P(f^1_i = 0 | f^3_i = 0) P(f^2_{i+1} = 0 | f^1_i = 0, f^3_i = 0)
= P(f^1_i = 0 | f^3_i = 0) P(f^3_{i+1} = 0 | f^3_i = 0)
= (1 - r_i)^2
= T_{1,1,i}.
\]

Proof for Proposition 0.1:

Proof. We establish the respective equivalence between Equation (6) and Equations (7)-(8) as follows.

Equation (6) for \( i = 1 \) and Equation (7) are equivalent because for all \( j \in \{1,2\} \), we have

\[
W_{i,j} = P(\gamma_i = 1, \gamma_i = L_{1,j}) = P(L_{i,j} = 1, \gamma_i = L_{1,j}) = L_{1,j} P(\gamma_i = 1, \gamma_i = L_{1,j}) = L_{1,j}^2 = P(\gamma_i = 1, \gamma_i = L_{1,j}) = L_{1,j} P(\gamma_i = 1, \gamma_i = L_{1,j} - 1) = L_{1,j} P(\gamma_i = 1, \gamma_i = L_{1,j}) = L_{1,j} \sum_{k=1}^{4} P(\gamma_i = 1, \gamma_i = L_{1,j} - 1, \gamma_i = L_{1,k}) = L_{1,j} P(\gamma_i = 1, \gamma_i = L_{1,j} - 1) = L_{1,j} P(\gamma_i = 1, \gamma_i = L_{1,j} - 1) = L_{1,j} P(\gamma_i = 1, \gamma_i = L_{1,j} - 1) = L_{1,j} \sum_{k=1}^{4} T_{k,j,i} W_{i,k} = L_{1,j} \sum_{k=1}^{4} T_{k,j,i} W_{i,k}.
\]

Proof for Proposition 0.2:
Proof.

\[
\text{PCV}(L_1, L_2, r) = P(g_1 = \ldots = g_N = 1) = \sum_{j=1}^4 P(g_1 = \ldots = g_N = 1, g_N = L_{N,j}) = \sum_{j=1}^4 W_{N,j}.
\]

\[\square\]

Appendix B: 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 lines \(P \in \mathbb{B}^{N \times 2 \times K}\) with \(K\) being the number of lines and the recombination frequencies vector \(r \in [0, 0.5]^N\). 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 \(m\)th parent, for all \(m \in \{1, 2\}\) and \(k \in \{1, \ldots, K\}\).
- \(x \in \mathbb{B}^{N \times 4}\) represents the genotypes of the two selected parents. If \(x_{j,1,2} = p_{j,1,2} \) and \(x_{j,3,4} = p_{j,3,4}\).
- \(w \in \mathbb{B}^{N \times 4}\) is the water matrix of \(x\).

The optimization model is presented in (12)-(19), which is a mixed integer linear program (MILP). The objective function (12) calculates the PCV of the two selected parent lines, which possibly be the same line. Constraints (15) and (16) assign the population.

\[\sum_{j=1}^4 P(g_1 = \ldots = g_N = 1, g_N = L_{N,j}) = \sum_{j=1}^4 W_{N,j}.
\]

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).

Literature Cited


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


