Development and optimization of expected cross value for mate selection problems

Pouya Ahadi[1], Balabhaskar Balasundaram[2], Juan S. Borrero[2], and Charles Chen*[3]

[1]*H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA.* [2]*School of Industrial Engineering and Management, Oklahoma State University, Stillwater, Oklahoma, USA.* [3]*Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, Oklahoma, USA.*

*Corresponding author: Department of Biochemistry and Molecular Biology, Oklahoma State University, 246 Noble Research Center, Stillwater, OK 74078, Tel (office): +14057444025, Tel (Lab): +14057446194, email:charles.chen@okstate.edu

Abstract

In this study, we address the mate selection problem in the hybridization stage of a breeding pipeline, which constitutes the multi-objective breeding goal key to the performance of a variety development program. The solution framework we formulate seeks to ensure that individuals with the most desirable genomic characteristics are selected to cross in order to maximize the likelihood of the inheritance of desirable genetic materials to the progeny. Unlike approaches that use phenotypic values for parental selection and evaluate individuals separately, we use a criterion that relies on the genetic architecture of traits and evaluates combinations of genomic information of the pairs of individuals. We introduce the *expected cross value* (ECV) criterion that measures the expected number of desirable alleles for gametes produced by pairs of individuals sampled from a population of potential parents. We use the ECV criterion to develop an integer linear programming formulation for the parental selection problem. The formulation is capable of controlling the inbreeding level between selected mates. We evaluate the approach or two applications: (i) improving multiple target traits simultaneously, and (ii) finding a multi-parental solution to design crossing blocks. We evaluate the performance of the ECV criterion using a simulation study. Finally, we discuss how the ECV criterion and the proposed integer linear programming techniques can be applied to improve breeding efficiency while maintaining genetic diversity in a breeding program.

INTRODUCTION

Plant and animal breeding consists of methodologies for the creation, selection, and fixation of superior phenotypes to fulfill the breeding goals of increasing productivity and financial returns, improving welfare, and reducing environmental impact Oldenbroek and van der Waaij (2015). Traditionally, breeders achieve these goals by identifying the individuals with desirable phenotypes and crossing them to produce the segregation of phenotypes in a new generation that allows further selection for advancement. This breeding strategy is perpetuated because high-volume crossing and evaluation led to the identification of the iconic *Green Revolution* varieties that successfully doubled rice and wheat yields from the 1960s to 1990s (Hesser 2006), despite the inevitable inefficiency of producing a high number of failed crosses. However, the future of food security and livestock will be driven not only by the demand but also by severe competition with other uses of land and water resource (Cassandro 2020). Therefore, more efficient breeding strategies ought to be considered because making many crosses with the knowledge that most fail is not justified either by theory or comparative experiments, and is also socially unacceptable.

Ultimately, the overall objective of a breeding program is to produce lines and varieties that are genetically homogeneous and perform at a high level, with end-use quality supportive of the intended market class. A wheat breeding pipeline, for instance, would begin with assembling parental stocks with a careful examination of available germplasm and donor traits. In principle, this is to construct and partition parental stocks respective to a specific goal or goals, to create the genetic variability needed for producing an adapted, high-yielding pure-line variety with perceived quality demands in the future marketplace.

With the continued advancement of genomic technologies and steady decline in genotyping costs, breeders are now able to take full advantage of the availability of genetic information embedded in the genome (Heffner et al. 2010; Hayes et al. 2009). Nevertheless, except

for the potential application of a higher selection intensity with GEBVs (genomic estimated breeding values) (Meuwissen et al. 2001), experimental data for the optimal number of crosses as well as the optimal numbers of progeny to sample from each cross required for selection as the initial investment to fulfill the breeding goal have not been reported in the literature (Donald 1968). This is unsurprising given that the number of individuals a breeding program can phenotypically evaluate is resource-limited (Rincent et al. 2017). For example, consider a single cross of two genetically distinct parental lines with 100 QTLs associated with variability among multiple desirable traits. Assuming independent assortment and co-dominance, the complete population from this single pair of founders will consist of $3^{100} \approx 5.1 \times 10^{47}$ genotypic combinations. Even considering a moderate number of 200 wheat lines in the parental stocks, the number of combinations to be evaluated in the field is astronomically high (Beans 2020). Consequently, analytical approaches based on operations research, mathematical optimization, and statistical learning to optimize breeding decision-making have gained prominence over the years (Johnson et al. 1988; Byrum 2015; Byrum et al. 2016, 2017; Kusmec et al. 2021).

There are two essential steps to addressing this problem using mathematical optimization. The first is to define a fitness criterion to evaluate individuals or crosses based on genetic information. The second step is to devise a mathematical optimization framework that incorporates the fitness criterion along with other essential requirements of the breeding program, and whose objective is to find the individuals or crosses that maximize the fitness criterion. The mathematical optimization framework, while faithfully capturing the various breeding requirements and objectives, must also be computationally viable in order for it to be useful in practice.

In contrast to addressing these breeding challenges in a traditional phenotype-centric paradigm, genetic improvement can also be more efficiently achieved by transferring desirable alleles from parents to progeny as a genetic process while avoiding alleles that show antagonistic

pleiotropic effects. Therefore, our aim is to devise a multi-objective mathematical optimization framework that targets more than one phenotype and generates multiple crosses that identify a set of best parental pairs from populations to address multiple breeding goals simultaneously.

As it is to be expected in any non-trivial multi-criteria decision-making setting, the criteria (breeding objectives) can be mutually conflicting, making it challenging to design an effective multi-objective optimization framework. For example, yield production in wheat has been found to be negatively correlated with grain protein content (Simmonds 1995), which is an essential factor for its commercial demand (Visscher et al. 1996). This makes concurrently fulfilling breeding goals of high yielding and protein content difficult. The negative correlation between the mass of beef cows and various measures of fertility and stayability could have attributed to the increasing concerns about compromised reproductive efficiency as a result of selection for growth (Berry and Evans 2014; Mwansa et al. 2002).

In this study, we propose a new fitness criterion called the *expected cross value* (ECV) that is inspired by a related fitness criterion called *predicted cross value* (PCV) introduced by Han et al. (2017). ECV returns a probabilistic measure of the fitness of the progeny of a specific pair of individuals based on the genetic architecture of trait variation. We consider the complexity of genetic architecture that underlies agronomic performance characteristics and develop an integer linear programming formulation of the parental pair selection problem that optimizes the ECV criterion. We further extend its capability to select multiple pairs of parents. Our optimization framework is based on the genetic transmission of all detectable genetic loci and can mitigate the potential impact of crossing within highly related individuals. Based on simulation studies, we demonstrate that using ECV as a fitness criterion would address the limitations of other related approaches for mate selection problems, and our multi-objective methodology can simultaneously improve a group of target phenotypic traits.

METHODS

We begin with some preliminaries needed to formally define the *expected cross value* (ECV) as a new fitness criterion for the mate selection problem. Considering all diploid and polyploid species that may behave as diploids cytologically, e.g., bread wheat (Riley and Chapman 1958), we assume that the variability of target traits is governed by segregating alleles at N different loci of all chromosomes in the genome. We use the index set notation $[a] \coloneqq \{1, 2, \ldots, a\}$ for a positive integer a, and define the genotype matrix next.

Definition 1. *Given an individual k, we define its genotype matrix L^k as an $N \times 2$ binary matrix with the i, j-th entry for every $i \in [N]$ and $j \in [2]$ given by:*

$$L^k_{i,j} = \begin{cases} 1, & \textit{if the allele in locus } i \textit{ of gamete from parent } j \textit{ is desirable,} \\ 0, & \textit{otherwise.} \end{cases} \tag{1}$$

Genotype matrix information of all individuals is an input for the ECV. Hereafter, we refer to the allele in QTL i as the i-th allele for ease of discussion. In our simulations, alleles are desirable when they enhance the trait value, assuming larger the positive value, the better. Observe that each column of L^k represents a gamete from one of the parents of individual k.

We model how alleles transfer from parents to children, i.e., how a gamete inherits alleles from the parent, by using a random N-dimensional binary vector J, with each component being a random variable J_i for each $i \in [N]$ defined as follows:

$$J_i = \begin{cases} 0, & \text{if the } i\text{-th allele is transferred from the first column of } L^k \text{ to the gamete,} \\ 1, & \text{otherwise.} \end{cases} \tag{2}$$

For a given individual k and QTL i, random variable J_i determines which of the gametes that comprise the genome of individual k will transfer the allele in the i-th locus.

Definition 2 (Han et al. (2017))**.** *We say that the random vector* $J \in \{0,1\}^N$ *follows an inheritance distribution with parameters* $r \in [0, 0.5]^{N-1}$ *and* α_0*, denoted by* $J \sim \mathcal{I}(r, \alpha_0)$*, if and only if*

$$\Pr(J_1 = 0) = \alpha_0,\ \Pr(J_1 = 1) = 1 - \alpha_0, \tag{3}$$

$$\Pr(J_i = J_{i-1}) = 1 - r_{i-1},\ \Pr(J_i = 1 - J_{i-1}) = r_{i-1}, \quad \forall i \in [N] \text{ and } i \geq 2. \tag{4}$$

In Definition 2, the $(N-1)$-dimensional vector $r \in [0, 0.5]^{N-1}$ represents the recombination frequencies between the consecutive pairs of loci. The value of r_i is the probability that the i-th and $(i+1)$-th alleles come from different gametes that comprise the genome of individual k. Note that if $r_i = 0$ for all $i \in [N-1]$, then the gamete produced by individual k is identical to exactly one of the parental gametes, while the maximum possible recombination between gametes is expected to be observed when $r_i = 0.5$ for all $i \in [N-1]$.

Deriving the closed-form marginal inheritance distributions

Given $J \sim \mathcal{I}(r, \alpha_0)$, we now derive the marginal distribution of J_i for each $i \in [N]$. The closed-form expressions so obtained then allow us to compute the expectations required to obtain a general closed-form expression for the ECV. For each $i \in [N]$, define the recursive function $\phi_i : \mathbb{R}^{N-1} \longrightarrow \mathbb{R}$ as follows:

$$\phi_1(r) = 0, \quad \phi_2(r) = r_1, \tag{5}$$

$$\phi_i(r) = r_{i-1} + (1 - 2r_{i-1})\,\phi_{i-1}(r),\ \forall i \in \{3, 4, \ldots, N\}. \tag{6}$$

Proposition 1. *Suppose that* $J \sim \mathcal{I}(r, \alpha_0)$*. Then, for each* $i \in [N]$*, the marginal distribution*

of J_i satisfies the following equations:

$$\Pr(J_i = 0) = \alpha_0 + (1 - 2\alpha_0)\phi_i(r), \tag{7a}$$

$$\Pr(J_i = 1) = 1 - \alpha_0 + (2\alpha_0 - 1)\phi_i(r). \tag{7b}$$

Proposition 1 (proved in the Supplement) establishes the marginal distributions through a recursion, which can be used to obtain a closed-form expression. This result can be further simplified using the laws of inheritance that the allele pairs of a locus segregate randomly during meiosis, and each allele transmits to the gamete with equal probability. Specifically, Proposition 1 then implies the following corollary.

Corollary 1. *Assume that Mendel's second law holds and $\alpha_0 = 0.5$. Then,*

$$\Pr(J_i = 0) = \Pr(J_i = 1) = 0.5, \ \forall i \in [N]. \tag{8}$$

Furthermore,

$$\mathbb{E}(J_i) = 0 \times \Pr(J_i = 0) + 1 \times \Pr(J_i = 1) = 0.5, \ \forall i \in [N], \tag{9}$$

where $\mathbb{E}(\cdot)$ represents the expectation operator.

The gamete and loss functions

The inheritance distribution characterizes the source of alleles transmitted from a parent to its gametes. Therefore, we can define a so-called *gamete function* to specify the alleles in the gamete according to the inheritance distribution. Given this gamete function, we derive a closed-form expression for the ECV of a pair of individuals.

Definition 3 (Han et al. (2017))**.** *Given an individual with genotype matrix L and a vector*

$J \sim \mathcal{I}(r, \alpha_0)$, the vector-valued gamete function $\text{gam} : (L, J) \mapsto g$ *outputs the binary gamete vector g defined as follows for each $i \in [N]$:*

$$g_i = \begin{cases} L_{i,1}, & \text{if } J_i = 0, \\ L_{i,2}, & \text{if } J_i = 1. \end{cases} \tag{10}$$

Equivalently, $g_i = L_{i,1}(1 - J_i) + L_{i,2}J_i$.

Suppose we have two individuals with genotype matrices L^1 and L^2, and two independent random vectors $J^1, J^2 \sim \mathcal{I}(r, \alpha_0)$. By crossing these two individuals, the genotype matrix for a child in the progeny is given by matrix $[g^1,\ g^2]$ where $g^1 = \text{gam}(L^1, J^1)$ and $g^2 = \text{gam}(L^2, J^2)$. Then, the gamete that is produced by a child of this progeny for the next generation is given by:

$$g^3 = \text{gam}\left(\left[g^1, g^2\right], J^3\right), \tag{11}$$

where $J^3 \sim \mathcal{I}(r, \alpha_0)$ is independent of J^1 and J^2. Below, we define a *loss function* in terms of the g^3 gamete vector that will lead us to the ECV criterion.

Definition 4. *Suppose L^1 and L^2 are the genotype matrices of two individuals and let J^k, $k = 1, 2, 3$, be independent random vectors following the distribution $\mathcal{I}(r, \alpha_0)$ for some given r and α_0. Let $g^k = \text{gam}(L^k, J^k)$ for $k = 1, 2$ and $g^3 = \text{gam}([g^1, g^2], J^3)$. We define the loss function associated with L^1, L^2, r, and α_0 as the following random variable:*

$$\text{loss}(L^1, L^2, r, \alpha_0) = \sum_{i=1}^{N}(1 - g_i^3) = N - \sum_{i=1}^{N} g_i^3. \tag{12}$$

The loss function counts the number of undesirable alleles in the gamete g^3. If the loss function is equal to 0, then all alleles in g^3 are desirable, while the opposite is true if it is equal to N. Before deriving our ECV criterion, we introduce the related PCV criterion of Han

et al. (2017).

Definition 5 (Han et al. (2017))**.** *Let L^1 and L^2 be the genotype matrices of two individuals, and let r and α_0 be given. Define the gamete g^3 using Equation* (11)*. Then, the PCV associated with L^1, L^2, r, and α_0 is the probability that the gamete g^3 contains only desirable alleles. That is,*

$$\text{PCV}(L^1, L^2, r, \alpha_0) = \Pr\big(\text{loss}(L^1, L^2, r, \alpha_0) = 0\big). \tag{13}$$

The expected cross value criterion

Next, we use the loss function in Definition 4 to define the ECV, an alternative criterion to PCV, based on allelic information of individuals. The measure depends on the gamete g^3 defined in Equation (11) and can evaluate a pair of individuals that could be mated.

Definition 6. *For a selected pair of individuals with genotype matrices L^1 and L^2, the ECV is the expected number of desirable alleles in gamete g^3 defined in Equation* (11)*. As the loss function represents the number of undesirable alleles in g^3, the ECV can be computed as:*

$$\text{ECV}(L^1, L^2, r, \alpha_0) = N - \mathbb{E}\big(loss(L^1, L^2, r, \alpha_0)\big) = \mathbb{E}\bigg(\sum_{i=1}^{N} g_i^3\bigg). \tag{14}$$

A pair of individuals with the highest ECV value could be selected as parents for crossing. Theorem 1 (proved in the supplement) constitutes our main result that provides a closed-form expression for calculating ECV for a pair of parents.

Theorem 1. *Assume Mendel's second law holds true and let L^1 and L^2 be the genotype matrices of two individuals. The ECV corresponding to the desired phenotypic trait can be*

computed using the following equation:

$$\text{ECV}(L^1, L^2, r, 0.5) = 0.25 \sum_{i=1}^{N} (L^1_{i,1} + L^1_{i,2} + L^2_{i,1} + L^2_{i,2}). \tag{15}$$

Remark 1. *Without relying on Mendel's second law, the ECV can still be computed in closed-form more generally as:*

$$\begin{aligned}\text{ECV}(L^1, L^2, r, \alpha_0) = \sum_{i=1}^{N} \Big(& L^1_{i,1} + [1 - \alpha_0 + (2\alpha_0 - 1)\phi_i(r)](L^1_{i,2} - 2L^1_{i,1} + L^2_{i,1}) \\ & + [1 - \alpha_0 + (2\alpha_0 - 1)\phi_i(r)]^2 (L^2_{i,2} + L^1_{i,1} - L^1_{i,2} - L^2_{i,1}) \Big).\end{aligned}$$

Theorem 1 provides a closed-form expression for the ECV criterion that enables us to formulate the parental selection problem as an integer linear program.

Single-trait parental selection problem

We develop an integer programming (IP) formulation for the parental selection problem using the ECV criterion as the single optimization objective (see Supplementary Formulation (27)) and the constraint system (and decision variables) from the mixed-integer programming formulation for the PCV introduced by Han et al. (2017). The formulation finds the best pair of individuals maximizing the ECV criterion based on a desired phenotypic trait. In addition, we restrict the inbreeding between selected individuals by preventing pairs of individuals with a sufficiently large inbreeding value from being selected as parents. By using the marker genotype information we can construct the genomic matrix G that quantifies the genomic relationship between any pair of individuals in the population (VanRaden 2008). Any pair in the population that has a genomic relationship (i.e., inbreeding value) higher than a pre-determined parameter ϵ, will be excluded from the set of feasible pairs using a family of constraints we include in the formulation.

In a breeding program, we may also seek to find multiple pairs for crossing, rather than just a single pair. In order to do so, we introduce Supplementary Algorithm 1 that iteratively solves our IP formulation for the single-trait parental pair selection problem. Note that solving the Supplementary Formulation (27) will identify a pair of individuals as the optimal solution for the problem. By adding "conflict constraints" corresponding to this optimal pair to the formulation, we can exclude *just* this optimal solution from the set of feasible solutions and reoptimize to find the next optimal pair. We can repeat this process until the required number of pairs have been chosen for the crossing program (assuming that many solutions exist).

The flowchart in Figure 1 illustrates the workflow of the proposed ECV approach for mate selection problems for a single trait. The process begins with an initial population where genetic marker and QTL information are available for the selection of parental lines to assemble the crossing block to advance specific breeding targets (Velu and Singh 2013). The ECV criteria can be optimized over several generations (denoted by T in Figure 1). In each generation, genomic information related to QTLs and genetic markers, along with a genetic relationship matrix (G) is used for constructing the optimization model detailed in Supplementary Formulation (27), and solving it to find an optimal set of mating pairs for crossing. The workflow for solving the multi-trait parental selection by optimizing the ECV metric mirrors the process in in Figure 1 for single-trait ECV optimization. The key difference is that we solve the Supplementary Formulation (29) via lexicographic optimization with user-specified degradation tolerances as described in detail in the Supplement.

Multi-trait parental selection problem

In general, breeders may be interested in improving several phenotypic traits simultaneously. In this case, we need to extend the ECV criterion to account for multiple traits. We assume there are M target traits in the breeding program and that the ℓ-th desired trait for every

$\ell \in [M]$ is affected by N_ℓ different QTL in the genome. For each individual we define M genotype matrices, one for each trait. Each such matrix is an $N_\ell \times 2$ binary matrix in which each row represents the pair of alleles in the corresponding genetic locus. Thus, we extend the previous definitions as follows.

Definition 7. *For $k \in [K]$ and $\ell \in [M]$, the genotype matrix $L^{k,\ell}$ associated with the k-th individual and the ℓ-th trait is defined as:*

$$L^{k,\ell}_{i,j} = \begin{cases} 1 & \textit{if } i\textit{-th allele of gamete from parent } j \textit{ is desirable for trait } \ell, \\ 0 & \textit{otherwise.} \end{cases} \qquad \forall i \in [N_\ell], j \in [2]. \tag{16}$$

Consider two individuals with genotype matrices $L^{1,\ell}$ and $L^{2,\ell}$ for target trait $\ell \in [M]$, and suppose that we have three independent random vectors J^1, J^2 and J^3 following an inheritance distribution $\mathcal{I}(r, \alpha_0)$. Using the definition of gamete function (10), the genotype matrix corresponding to the ℓ-th trait for a child in the progeny is represented by matrix $[g^{1,\ell}, g^{2,\ell}]$ where $g^{1,\ell} = \mathrm{gam}(L^{1,\ell}, J^1)$ and $g^{2,\ell} = \mathrm{gam}(L^{2,\ell}, J^2)$. The gamete that is produced by this progeny for the next generation is then given by:

$$g^{3,\ell} = \mathrm{gam}\left([g^{1,\ell}, g^{2,\ell}], J^3\right). \tag{17}$$

Definition 8. *For the ℓ-th target trait and a selected pair of individuals with genotype matrices $L^{1,\ell}$ and $L^{2,\ell}$, the ECV^ℓ is the expected number of desirable alleles of trait ℓ in gamete $g^{3,\ell}$. Following Equation* (17)*, ECV^ℓ, for each $\ell \in [M]$ is defined as:*

$$ECV^\ell(L^{1,\ell}, L^{2,\ell}, r, \alpha_0) = N_\ell - \mathbb{E}\big(loss(L^{1\ell}, L^{2,\ell}, r, \alpha_0)\big) = \mathbb{E}\Big(\sum_{i=1}^{N_\ell} g_i^{3,\ell}\Big). \tag{18}$$

Following Theorem 1, we can obtain a closed-form expression for ECV^ℓ function.

Theorem 2. *Assume Mendel's second law holds true and let $\ell \in [M]$. Then, for a selected pair of individuals with genotype matrices $L^{1,\ell}$ and $L^{2,\ell}$, the ECV corresponding to the ℓ-th target phenotypic trait can be computed as:*

$$\text{ECV}^{\ell}(L^{1,\ell}, L^{2,\ell}, r, 0.5) = 0.25 \sum_{i=1}^{N_\ell} (L_{i,1}^{1,\ell} + L_{i,2}^{1,\ell} + L_{i,1}^{2,\ell} + L_{i,2}^{2,\ell}). \tag{19}$$

Ideally, a breeding program would like to select parental pair(s) that simultaneously optimize all the ECV^{ℓ} functions. Such an optimum is not likely to exist in practice because some phenotypic traits are negatively correlated. Therefore, improving one trait might worsen others. In order to achieve a reasonable trade-off, one turns to the theory of multi-objective optimization.

Consider a vector of objective functions $F(t, x) = \langle f_1(t, x^1), f_2(t, x^2), \ldots, f_M(t, x^M) \rangle$ where $f_\ell(t, x^\ell)$ denotes the ECV function (19) corresponding to ℓ-th trait. Supplementary Formulation (29) for the multi-trait parental selection problem seeks to find a pair of individuals that will "maximize" the vector-valued objective function. Similar to the single-trait optimization model, this formulation also excludes pairs of individuals with genomic relationship exceeding the tolerance threshold from the set of feasible solutions. Furthermore, as explained in the previous section, this approach can also be extended to select multiple parental pairs for the breeding program by iteratively adding conflict constraints. The differences lie in the handling of multiple traits, especially in the vector objective function.

Multi-objective or vector optimization problems are commonly handled by scalarization—converting the vector optimization problem into one or more scalar optimization problems (Miettinen 2012; Sawaragi et al. 1985); see survey by Miettinen et al. (2016) for interactive and other methods. One approach is to use a weighted combination of the individual objective functions to produce an optimization problem with a scalar objective. The weights, which are predetermined by the user, need to be carefully chosen to ensure they reflect the rela-

tive importance of the individual objectives and also scale them appropriately as necessary. Another approach, *lexicographic optimization*, prioritizes the objective functions based on their importance and optimizes them sequentially, starting with the most important. While optimizing lower priority objectives, we restrict the feasible region to only those solutions that will not degrade the higher priority objectives, or limit their degradation by user-specified tolerances.

The weighted sum approach, where we aggregate the individual objectives into a single objective using user-defined weights, requires a vector of weights that capture the importance of each phenotypic trait in the breeding program. In practice, it is difficult to identify a precise and meaningful weight for each trait as there are many factors of the breeding program (some of them potentially unknown) that might play a role in defining it. By contrast, it might be simpler for a breeding program to order the traits based on their importance.

The lexicographic optimization approach is not without drawbacks, as it could degenerate into single-objective optimization with the highest priority objective if we subsequently allow no degradation of higher priority objectives. In the worst case, if the first objective has a unique optimal solution and we tolerate no degradation on the first objective, the subsequent objectives are irrelevant. The use of tolerance is therefore important as it allows limited degradation of a higher priority objective when optimizing a lower priority objective, but allows for a larger feasible solution space for the lower priority objective (when compared against using zero tolerance). Hence, we will be using lexicographic maximization with positive tolerances in solving Supplementary Formulation (29).

Assume without loss of generality that the vector of objective functions, $F(t, x) = \langle f_\ell(t, x^\ell)\rangle_{\ell=1}^{M}$, is already in decreasing order of importance. Thus trait ℓ is more important than trait $\ell + 1$, for each $\ell \in [M - 1]$. The solver we use in our computational studies is capable of lexicographic optimization with degradation tolerances for objectives specified by the decision maker. Let us denote these tolerances by $\tau = (\tau_1, \tau_2, \ldots, \tau_M)$, where $\tau_\ell \in [0, 1]$

for each trait $\ell \in [M]$. The solver optimizes the first objective function $f_1(t, x)$ and then, among those feasible solutions within a factor $(1 - \tau_1)$ of the optimal objective value of the first objective function, optimizes the second objective function. This process is repeated until the last objective is optimized. In particular, this method assures that the optimal solution for the ℓ-th objective, for $\ell = 2, \ldots, M$, is within a factor $(1 - \tau_i)$ of the optimal value of the i-th objective, for every $i \in [\ell - 1]$. As $f_M(t, x^M)$ is the last objective function to optimize, there is no need for a tolerance τ_M, and hence we set $\tau_M = 0$ for all our experiments.

Simulation study

Simulations were conducted to evaluate the performance of ECV compared to other parent selection approaches using phenotypes and breeding values (GEBV). Two simulation experiments were considered in this study. First, we considered a single-trait optimization problem to solely improve Trait 1, simulated as a mixture of traits with oligogenic and polygenic genetic architectures. Next, we examined a multiple-trait parent selection problem where the breeding program was tasked to simultaneously improve all traits of interest. In this experiment, we simulated a polygenic architecture for Trait 3, representing a trait such as yielding capacity that is usually governed by a large number of loci where each allele has a small impact on the expression of the trait and in a negative genetic correlation with Trait 1, in addition to an oligogenic phenotype (Trait 2) that may imitate the genetic architecture underlying disease resistance.

For all experiments, two metrics were reported from the simulations, average desirable allele frequency and average phenotypic trait values of the progeny, to compare the performance of the methods in each generation. We also recorded the average genomic relationship for the selected individuals for all three approaches. In the case of multi-parental pair selection, we sorted pairs of individuals based on the summation of their trait values or GEBVs and made

selection decisions based on the summations of trait values. Moreover, by default, there was no control over the genomic relationship between selected parent pairs for the phenotypic selection and GEBV selection approaches; however, we assumed that self-crossing is not a feasible choice in these approaches.

The QU-GENE engine and QuLinePlus proposed by Ali et al. (2020) were used to simulate initial populations and the progeny in the subsequent generations. The QU-GENE engine establishes the initial population with inputs of genetic effects for segregating alleles, recombination frequencies and the number of desired individuals. We considered an initial population such that the allele frequency at all loci was set at 0.5. In our experiments, QuLinePlus took the genotypic information of a population and a list of selected pairs, simulated the progeny by crossing the selected parental pairs, and output genotypic and phenotypic information for all individuals in the subsequent generation. The GEBVs were calculated using the "rrBLUP" package (Endelman 2011). The Gurobi Optimization Solver (Gurobi Optimization, LLC 2024) was used to solve the integer linear programming formulations that were implemented in the Python programming language.

The initial population consisted of 10,000 individuals, with 200 biallelic genetic loci and 100 markers. Of these, 40 genetic loci had effects on Trait 1, 10 on Trait 2, and 70 on Trait 3. The markers had no genetic effects on any of the traits. Trait 3 and Trait 1 share 20 common loci with pleiotropic effects, which resulted in a negative correlation between those phenotypic traits. We conducted all of the experiments for four generations and for each cross we simulated 100 progeny for the next generation. We performed two sets of simulation studies, assuming a consistent growing environment across generations. In the first simulation study, the number of parental pairs that we chose from the initial population, generations one, two, and three, was 50, 10, 5, and 5, respectively. Thus, the population size in the simulation studies for generations one, two, three and four were, respectively, 5,000, 1,000, 500, and 500, respectively.

To further investigate the effectiveness of our methodology, we explored the impact of selection intensity in our second simulation study. Scenario A imposed a higher selection intensity with 50 crosses made from the initial population (generation 0), and 10, 3, and 3, respectively, for generations one, two, and three. For intermediate selection intensity (Scenario B, same as the first simulation study), from the initial population, generations one, two, and three, we chose 50, 10, 5, and 5 parental pairs, respectively. Finally, in Scenario C, representing a case of reduced selection intensity, the numbers of parental pairs selected in generation one was 25, and 5 parental pairs for the generations two and three.

RESULTS

The single-trait simulation results over five generations are summarized in Figure 2. For all traits considered, ECV significantly increased the proportion of desirable alleles (see Figure 2a) while showing the capacity to regulate the relatedness within the breeding population by avoiding crossing closely related individuals (see Figure 2c). Further, although statistically insignificant, genetic crosses done by phenotypically superior individuals returned the lowest means of the progeny in all traits, compared to genetics-informed approaches, like GEBVs and ECV (see Figure 2b). However, populations generated by ECV provided a greater potential for advancing individuals with larger phenotypic values.

Single-trait optimization does not guarantee improvement for phenotypes other than the target trait. Figure 3 shows boxplots for Trait 2 and Trait 3 when we optimize Trait 1 in a single-trait ECV optimization framework. The frequency for the desirable allele (Figure 3a) as well as the phenotypic values (Figure 3b) remained unimproved for Trait 2. The scenario could be worse if target traits are determined by QTLs with antagonistic pleiotropic effects. This can be seen in Figures 3c and 3d, which depict a significant decrease in the proportion of desirable alleles and phenotypic values of Trait 3 as a result of optimizing for Trait 1.

For multi-trait parental selection based on ECV, we employed the lexicographic multi-objective optimization approach described earlier. The tolerances were chosen based on preliminary experiments as follows: let $\tau_{i,c}$ denote the degradation tolerance for optimization objective i in generation c, then we used $\tau_{1,0} = 0.17, \tau_{1,1} = 0.05, \tau_{1,2} = 0.05, \tau_{1,3} = 0.05$ and $\tau_{2,0} = 0.00, \tau_{2,1} = 0.00, \tau_{2,2} = 0.00, \tau_{2,3} = 0.05$. In general, the tolerance parameters can be calibrated to have the desired impact on the model. The results in Figure 4 show the advantage of using ECV. Despite the negative genetic correlation, ECV was able to increase the desirable allele frequency to 0.70 (± 0.02), 0.65 (± 0.08), and 0.72 (± 0.01), for Trait 1, Trait 2, and Trait 3, respectively. In contrast, the impact of negative correlation between Trait 3 and Trait 1 was most obvious when the phenotypic selection was used, leading to a significant loss of desirable allele for Trait 1 (see Figure 4a). Similarly, ECV improved phenotypic values of the progeny for all traits simultaneously, whereas no improvement for Trait 1 and Trait 2 was found using phenotypic selection in our simulations when the tolerances were set slightly favoring Trait 3. It is noteworthy that a genomics-informed selection method, GEBV, returned comparable results to ECV for Trait 1. This benefit of GEBV, however, is at the expense of genetic diversity, as shown in Figure 5. Genomic relatedness (VanRaden 2008) has increased noticeably over the four generations using the GEBV selection method.

Figure 6 displays the boxplots for the three selection intensity scenarios A, B, and C introduced in the Simulation study, focusing on the proportion of desirable alleles as the performance metric. In the early generations, particularly generation 2, our proposed ECV approach outperformed other selection strategies, most evidently for Traits 1 and 3 in the implementation of multiple trait selection. As the generations advanced, the ECV approach continued to excel in Scenarios B and C, resulting in higher proportion of desirable alleles for Traits 1 and 3. In Scenario A, where selection intensity was higher and ECV selection method was not dominant, the method still yielded replications with a greater proportion of desirable alleles compared to other strategies, despite the genomic relationship constraints

inherent in the ECV method. Furthermore, in the last generation under scenario A, the mean ($\pm$ standard deviation) of genomic relatedness over all replications for the ECV, GEBV, and Phenotypic selection approaches are 0.15 ($\pm$0.04), 0.42 ($\pm$0.10), and 0.25 ($\pm$0.10), respectively. These results illustrate the effectiveness of our ECV optimization framework, with its explicit constraints limiting genomic relatedness, in managing genomic relatedness over generations when selection intensity is higher, while improving desirable breeding traits. The impact of higher selection intensity, however, led to greater variability in the proportion of desirable alleles in Trait 3 of Scenario A, which could be due to the drift effect of the smaller breeding population size (Turner-Hissong et al. 2020).

DISCUSSION

The principal objective of breeding is to combine as many desirable traits as possible into a genotype that can be distributed to farmers, producers or breeders. For example, in plant breeding the breeders are tasked with developing elite genotypes that display desirable use characteristics including high yields, disease resistance, and are also well-adapted to a range of environmental conditions (Breseghello and Coelho 2013). These desirable characteristics are typically possessed by multiple founders. By mixing and recombining founder genomes, the distribution of these desirable phenotypes observed in the offspring, owing to the segregation of alleles often distributed throughout the genome, allow breeders to identify superior individuals for subsequent breeding, widespread evaluations, and sales.

However, when these desired characteristics differ in variability, heritability, economic importance, and are correlated with other phenotypes and genotypes, effective mating designs capable of improving multiple traits simultaneously can be challenging to identify (Johnson et al. 1988). This breeding process is also ineffective as breeders tend to make hundreds or thousands of crosses, of which only a few are advanced in the subsequent years (Witcombe

et al. 2013). Traditionally, these objectives are achieved by breeding from the "best"—the best being determined by their own phenotypic values (Allard 1999; Akdemir et al. 2019). More advanced techniques, such as pedigree-based (Henderson 1984; Gianola and Fernando 1986), marker-based genetic value predictions (Lande and Thompson 1990; Hospital and Charcosset 1997; Bernardo and Charcosset 2006), and mating designs by genomic information (Akdemir and Sánchez 2016) are also available.

Beginning with the work of Johnson et al. (1988), mathematical programming approaches have facilitated the improvement of genetic traits through the use of mathematical optimization models that aid breeders in making better decisions in selecting mating parents. Toro et al. (1991) solved mate selection problems using linear programming techniques and demonstrated the effectiveness of their approach within multiple ovulation and embryo transfer (MOET) breeding schemes for dairy cattle with the help of simulation studies. Jansen and Wilton (1985) addressed the issue of factorial growth in the number of combinations to cross by formulating and solving an integer programming model to improve the overall progeny merit.

Moeinizade et al. (2019) recently proposed a single-trait optimization of a "look ahead" metric that focuses on a predetermined terminal generation to optimize mating decisions for maximizing expected GEBVs in the terminal generation without explicitly considering the impact of genetic erosion. Amini et al. (2021) further improved this look-ahead framework by prioritizing best individuals for crossing and using multiple prediction algorithms to improve prediction accuracy. These approaches are also complemented by Zhang and Wang (2022) who proposed a "net present value" inspired mechanism for discounting future gains, which values early-term genetic gains more than those anticipated in the future. This was done to overcome a drawback of the original look-ahead scheme by Moeinizade et al. (2019), which can produce slow genetic gains in the early generations and accelerating more rapidly as we approach the terminal generation.

Byrum et al. (2016) report on their long-term development and quantification of an unbiased

genetic gain performance metric, and pioneered its use in evaluating breeding projects as varieties were developed. Byrum et al. (2016) and Byrum et al. (2017) demonstrate the successful commercial use of advanced analytics and operations research tools such as integer linear programming, Monte Carlo simulation, and stochastic optimization by the agriculture industry, which has served to further motivate its broader use in many areas of crop and animal sciences; see also (Byrum 2015, 2016).

Furthermore, when the breeding objective involves more than one trait, a selection index of progeny merit was considered as a linear function of estimated breeding values for each trait by Allaire (1980). In animal breeding, for instance, the genetic merit of calves is estimated as half of the sire's and half of the dam's breeding value. An optimization-based procedure for mate selection in animal breeding is introduced by Kinghorn (1998, 2011) based on a mate selection criterion proposed by Kinghorn and Shepherd (1999).

In the genomics era, the parental selection problem has been increasingly addressed with the use of genomic relationships (Sun et al. 2013), heuristic searches for gene pyramiding (De Beukelaer et al. 2015), and by modeling the recombination of desirable alleles as a result of crossing (Han et al. 2017; Moeinizade et al. 2019). For the purpose of introgressing a small number of desirable alleles from a donor to a recipient, Han et al. (2017) proposed an efficient algorithm for calculating the PCV defined as the probability that a gamete of a random progeny from crossing two genetic individuals would consist only of desirable alleles. In a specific case where the desirable allele for the i-th locus is not present in both parents (denoted by k and k'), such that $L^k_{i,1} = L^k_{i,2} = L^{k'}_{i,1} = L^{k'}_{i,2} = 0$, PCV will conclude that the i-th component of gamete g^3 is zero with probability one and hence $PCV(L^k, L^{k'}, r, \alpha_0) = 0$. In this case, the individuals k and k' will not be selected, regardless whether or not there maybe be desirable alleles present in the rest of the genome. While such a result is desirable for the goals of introgressing a small number of alleles for traits like herbicide, disease, or insect resistance, it would be inappropriate for identifying crosses that will have the best

opportunity to combine a large number of genetic alleles. Considering the polygenic inheritance of agronomical performance traits (Lynch and Walsh 1998; Scott et al. 2021), the PCV approaches zero for all breeding parents as the number of loci with desirable alleles increases. Consider the following probabilistic inequality (Fréchet inequality):

$$\mathrm{PCV}(L^k, L^{k'}, r, \alpha_0 = 0.5) = \Pr\left(g_i^3 = 1 \ \forall i \in [N]\right) \leq \min_{i \in [N]} \Pr(g_i^3 = 1). \tag{20}$$

Hence, the larger the value of N, the greater the chance $L_{i,1}^k = L_{i,2}^k = L_{i,1}^{k'} = L_{i,2}^{k'} = 0$ for some QTL i. The PCV method could therefore lead to indiscriminate mate selection for traits that have hundreds or thousands of loci with desirable alleles because the PCV value is (nearly) zero for essentially any choice of mates. This observation motivated us to introduce our ECV criterion, especially for breeding targets governed by a large number of genetic loci and for non-introgression projects.

As Figure 2a shows, our results demonstrated a significantly greater capacity to increase desirable allele frequency compared to the conventional phenotypic selection and the selection done by the genomics-derived GEBV; and, the benefit of using ECV can be realized in as short as two generations. Moreover, the greater range of trait value distribution presents additional opportunities for breeders to identify the superiors for population advancement (see Figure 2b).

Based on our simulations, we observe that the breeding population has gone from unrelated to essentially full-sibs in three generations of selecting breeding parents based on GEBVs (see Figure 2c). Compared to the phenotypic selection, GEBV selection might have manifested a rapid increase of relatedness by crossing individuals closely related to the training population (Bassi et al. 2016; Forutan et al. 2018). Though GEBV selection might show a capacity to provide short-term genetic gain, selecting breeding parents solely by GEBVs would lead to undesirable consequences such as loss of genetic diversity, further diminishing long-term

genetic gain (Jannink 2010; Doekes et al. 2018).

To ensure the capacity to preserve multiple genetic lineages, ECV allows for the selection of more than one pair of individuals, and while self-crossing was not allowed in this study, our method permitted the same individual to be crossed with multiple breeding parents as long as the genomic relationship of the parents was not greater than ϵ, a parameter that breeders can use to control how much inbreeding is acceptable.

Fundamental to all variety improvement programs is the identification of the most efficient path to reach breeding objectives (Bernardo 2002; Akdemir et al. 2019). However, breeders are usually tasked with combining a suite of traits in addition to yield and growth components. The negative genetic correlations caused by the non-random association of alleles underlying these breeding objectives impose additional challenges, as selecting based on one trait may adversely impact another (Lynch and Walsh 1998). To simultaneously improve multiple traits, phenotype-based selection indices have been widely considered (Hazel and Lush 1942; Hazel et al. 1994; Villanueva and Woolliams 1997; Jannink et al. 2000; Moeinizade et al. 2020). Selecting breeding parents based on a selection index does not necessarily choose the best genetics to recombine; further, since the selection index applied is a weight assignment of target phenotypes, such decisions could result in the loss of beneficial alleles.

In this study, the proposed ECV framework is based on an allele transmission process. Rather than relying on the phenotypes of breeding parents, ECV identifies the crosses with the highest likelihood of transmitting desirable alleles from pairs of parents to the progeny. In the case where multiple traits need to be considered simultaneously, ECV seeks the optimal combination of alleles for all target phenotypes ordered by their importance, while maintaining a customizable tolerance such that QTLs with antagonistic pleiotropic gene action could remain in consideration before the final breeding recommendation is made. Figure 4 and 6 showed that despite the negative correlation between Traits 1 and 3, ECV was able to increase desirable allele frequency for all traits in our simulation studies. In addition, as

seen from Figure 5, the inbreeding coefficient in the progeny was regulated as ECV was optimized with the tolerance constraint on the genomic relatedness between breeding parents. As genotyping has become routine in breeding programs (Hayes and Goddard 2010; Bentley et al. 2022), the application of this constraint ought to be considered to mitigate the multiple trait scenario in Figure 4, where the gain might be built at the expense of genetic diversity (Figure 5), a phenomenon also found in index selection methods (Akdemir et al. 2019). If practical considerations favor breeding parents to be selected from a narrow genetic pool, the constraint could be moved to the objective as a penalty term.

Breeding programs develop elite genotypes that often demonstrate similar essential genomic profiles of desirable end-use characteristics, agronomical attributes, disease resistance packages, as well as adaptation to the target environment. Breeding among the elites can produce new variability as the source of new cultivars with minimal risk of introducing undesirable features. This variation may eventually be exhausted, and new genes and alleles must be introduced. Identifying beneficial alleles from un-adapted material itself has been described as searching for a needle in a haystack (Pixley et al. 2014). Introgressing these novel alleles can also be risky because the unwanted alleles in exotic germplasm may distrupt essential allele combinations (Willcox et al. 2022); and, it requires a higher institutional cost due to a greater number of crosses and longer breeding cycle needed to achieve the breeding objectives (Snelling et al. 2019; Neyhart et al. 2019). Based on our simulations, we reckon that ECV can be an option.

Beyond animal and cereal crop breeding, we suspect that implementing optimization-based methods like ECV could be advantageous to breeding of genetically diverse, long-generation, and slow reaction, cross-pollinated species, such as conifers. Tree breeders generally establish open-pollinated seed orchards for selection (White et al. 2007), and several mating designs have been proposed (Namkoong 1976; Zobel and Talbert 1984), among which the polycross is considered as one of the most cost-effective (Kumar et al. 2007; Lenz et al. 2020). The

ability to design the pollen pool while managing inbreeding with ECV will provide the capacity to rapidly increase desirable allele frequencies and, at the same time, avoid severe inbreeding depression for conifer species (Berry and Evans 2014; Mwansa et al. 2002; Snelling et al. 2019).

The conceptual framework we have introduced and our results show that adopting multi-objective optimization tools from operations research to solve breeding problems is highly advantageous (Cameron et al. 2017; Beans 2020; Kusmec et al. 2021). Several improvements or extensions suggested next should also be considered. When the pool of genetic diversity increases, solving integer programming problems for ECV will require further development to account for different distributions of crossover events in different crosses (Stapley et al. 2017; Nachman 2002; Jabbari et al. 2019; Dreissig et al. 2019). Furthermore, as multi-parent populations like MAGIC (multi-parent advanced generation inter-cross) have become a means to provide germplasm for breeding programs (Scott et al. 2020), there is a need to expand optimization frameworks such as ECV to consider multiple parental lineages, which might also help guide the polycross mating design in forestry (Frandsen 1940; Lambeth et al. 2001).

Our proposed methodology relies on the the underlying genetic information of the breeding population, such as QTLs and genetic association of desirable traits. While affordable large-scale genotyping and phenotyping technologies are becoming accessible to breeding programs (Reynolds et al. 2020; Bassi et al. 2024), large breeding populations necessitate extensive genomic information, which can be computationally demanding. Moreover, integer linear programming is NP-hard in general, making it challenging to solve very large-scale problems to optimality. In the case of mate selection, the size of the population and the number of genes directly influence the computational time, which implies that massive datasets could make obtaining optimal solutions unrealistic for practical applications. In such circumstances, we may consider modifying our approach to solving the integer linear program by employing decomposition techniques to address the large-scale instances and likely settle

for sub-optimal (but good quality) feasible solutions.

Our simulations also indicate that the ECV selection framework may result in higher performance variability when the selection intensifies in earlier generations. To alleviate this issue, we could either relax the genomic relatedness constraint (smaller ϵ) in earlier generations or intensify selection only in advanced generations. Further, care must also be taken in choosing the degradation tolerances τ_ℓ for each trait $\ell \in [M]$ in the lexicographic multi-trait ECV optimization framework, which will entail computational expenditure in terms of preliminary computational experiments, which could become challenging at larger scales.

ACKNOWLEDGEMENTS

Funding for this work was supported by grants from the Oklahoma Wheat Research Foundation (for CC), Oklahoma Center for the Advancement of Science and Technology (OCAST) award number PS15-011-2 and PS19-004 for CC. This work was completed utilizing the High-Performance Computing Center facilities of Oklahoma State University at Stillwater, and also in part by the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC) under the resource allocation MCB-180177. The authors are grateful to the anonymous reviewers for their careful reading of our original manuscript, their constructive criticism, and providing detailed and thoughtful comments that helped us improve this manuscript.

AUTHOR CONTRIBUTIONS

BB, JB, and CC were responsible for the conceptualization of the study. PA developed the theoretical results and computer implementations as part of his thesis. PA and CC performed the analysis and wrote the original draft. All authors contributed to interpreting results, providing feedback, and editing and approving the manuscript.

COMPETING INTERESTS

The authors have no competing financial interests to declare.

RESEARCH ETHICS STATEMENT

No approval of research ethics committees was required because no experimental work was conducted; only computer simulations were used in this study.

DATA AVAILABILITY

The data and codes are available at: `https://github.com/transgenomicsosu/ECV`.

References

Akdemir, D., Beavis, W., Fritsche-Neto, R., Singh, A. K., and Isidro-Sánchez, J. (2019). Multi-objective optimized genomic breeding strategies for sustainable food improvement. *Heredity*, 122(5):672–683.

Akdemir, D. and Sánchez, J. I. (2016). Efficient breeding by genomic mating. *Frontiers in Genetics*, 7:210.

Ali, M., Zhang, L., DeLacy, I., Arief, V., Dieters, M., Pfeiffer, W. H., et al. (2020). Modeling and simulation of recurrent phenotypic and genomic selections in plant breeding under the presence of epistasis. *The Crop Journal*, 8(5):866–877.

Allaire, F. (1980). Mate selection by selection index theory. *Theoretical and Applied Genetics*, 57(6):267–272.

Allard, R. W. (1999). *Principles of plant breeding*. John Wiley & Sons.

Amini, F., Franco, F. R., Hu, G., and Wang, L. (2021). The look ahead trace back optimizer for genomic selection under transparent and opaque simulators. *Scientific Reports*, 11(1):4124.

Bassi, F. M., Bentley, A. R., Charmet, G., Ortiz, R., and Crossa, J. (2016). Breeding schemes for the implementation of genomic selection in wheat (triticum spp.). *Plant Science*, 242:23–36.

Bassi, F. M., Sanchez-Garcia, M., and Ortiz, R. (2024). What plant breeding may (and may not) look like in 2050? *The Plant Genome*, 17(1):e20368.

Beans, C. (2020). Inner workings: Crop researchers harness artificial intelligence to breed crops for the changing climate. *Proceedings of the National Academy of Sciences*, 117(44):27066–27069.

Bentley, A., Chen, C., and D'Agostino, N. (2022). Editorial: Genome wide association studies and genomic selection for crop improvement in the era of big data. *Frontiers in Genetics*.

Bernardo, R. (2002). *Breeding for quantitative traits in plants*. Stemma Press, Woodbury, Minnesota, USA.

Bernardo, R. and Charcosset, A. (2006). Usefulness of gene information in marker-assisted recurrent selection: A simulation appraisal. *Crop Science*, 46(2):614–621.

Berry, D. P. and Evans, R. (2014). Genetics of reproductive performance in seasonal calving beef cows and its association with performance traits. *Journal of Animal Science*, 92(4):1412–1422.

Breseghello, F. and Coelho, A. S. G. (2013). Traditional and modern plant breeding methods with examples in rice (oryza sativa l.). *Journal of Agricultural and Food Chemistry*, 61(35):8277–8286.

Byrum, J. (2015). Agriculture: Fertile ground for analytics and innovation. *OR/MS Today*, 42(6):28–32.

Byrum, J. (2016). Optimizing crop management: "Smart" application of fertilizer illustrates payoff in using analytical tools to enhance crop yields and improve the environment. *OR/MS Today*, 43(3):26–30.

Byrum, J., Beavis, B., Davis, C., Doonan, G., Doubler, T., Kaster, V., et al. (2017). Genetic gain performance metric accelerates agricultural productivity. *Interfaces*, 47(5):442–453.

Byrum, J., Davis, C., Doonan, G., Doubler, T., Foster, D., Luzzi, B., et al. (2016). Advanced analytics for agricultural product development. *Interfaces*, 46(1):5–17.

Cameron, J. N., Han, Y., Wang, L., and Beavis, W. D. (2017). Systematic design for trait introgression projects. *Theoretical and Applied Genetics*, 130(10):1993–2004.

Cassandro, M. (2020). Animal breeding and climate change, mitigation and adaptation. *Journal of Animal Breeding and Genetics*, 137(2):121–122.

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):1–16.

Doekes, H. P., Veerkamp, R. F., Bijma, P., Hiemstra, S. J., and Windig, J. J. (2018). Trends in genome-wide and region-specific genetic diversity in the Dutch-Flemish Holstein–Friesian breeding program from 1986 to 2015. *Genetics Selection Evolution*, 50(1):1–16.

Donald, C. T. (1968). The breeding of crop ideotypes. *Euphytica*, 17(3):385–403.

Dreissig, S., Mascher, M., and Heckmann, S. (2019). Variation in recombination rate is shaped by domestication and environmental conditions in barley. *Molecular Biology and Evolution*, 36(9):2029–2039.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome*, 4(3):250–255.

Forutan, M., Ansari Mahyari, S., Baes, C., Melzer, N., Schenkel, F. S., and Sargolzaei, M. (2018). Inbreeding and runs of homozygosity before and after genomic selection in north american holstein cattle. *BMC Genomics*, 19(1):1–12.

Frandsen, H. (1940). Some breeding experiments with timothy. *Imp Agr Bur Jt Pub*, 3:80–92.

Gianola, D. and Fernando, R. L. (1986). Bayesian methods in animal breeding theory. *Journal of Animal Science*, 63(1):217–244.

Gurobi Optimization, LLC (2024). Gurobi optimizer reference manual. `https://www.gurobi.com`. Accessed 26 May 2024.

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.

Hayes, B. and Goddard, M. (2010). Genome-wide association and genomic selection in animal breeding. *Genome*, 53(11):876–883.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited review: genomic selection in dairy cattle: progress and challenges. *Journal of Dairy Science*, 92(2):433–443.

Hazel, L., Dickerson, G., and Freeman, A. (1994). The selection index—then, now, and for the future. *Journal of Dairy Science*, 77(10):3236–3251.

Hazel, L. and Lush, J. L. (1942). The efficiency of three methods of selection. *Journal of Heredity*, 33(11):393–399.

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

Henderson, C. R. (1984). *Applications of linear models in animal breeding*. University of Guelph, Guelph, ON, Canada.

Hesser, L. F. (2006). *The man who fed the world: Nobel Peace Prize laureate Norman Borlaug and his battle to end world hunger: An authorized biography*. Leon Hesser.

Hospital, F. and Charcosset, A. (1997). Marker-assisted introgression of quantitative trait loci. *Genetics*, 147(3):1469–1485.

Jabbari, K., Wirtz, J., Rauscher, M., and Wiehe, T. (2019). A common genomic code for chromatin architecture and recombination landscape. *PLoS One*, 14(3):e0213278.

Jannink, J.-L. (2010). Dynamics of long-term genomic selection. *Genetics Selection Evolution*, 42(1):1–11.

Jannink, J.-L., Orf, J., Jordan, N., and Shaw, R. (2000). Index selection for weed suppressive ability in soybean. *Crop Science*, 40(4):1087–1094.

Jansen, G. and Wilton, J. (1985). Selecting mating pairs with linear programming techniques. *Journal of Dairy Science*, 68(5):1302–1305.

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

Kinghorn, B. P. (1998). Mate selection by groups. *Journal of Dairy Science*, 81:55–63.

Kinghorn, B. P. (2011). An algorithm for efficient constrained mate selection. *Genetics Selection Evolution*, 43(1):1–9.

Kinghorn, B. P. and Shepherd, R. K. (1999). Mate selection for the tactical implementation of breeding programs. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics*, 13:130–133.

Kumar, S., Gerber, S., Richardson, T., and Gea, L. (2007). Testing for unequal paternal contributions using nuclear and chloroplast ssr markers in polycross families of radiata pine. *Tree Genetics & Genomes*, 3(3):207–214.

Kusmec, A., Zheng, Z., Archontoulis, S., Ganapathysubramanian, B., Hu, G., Wang, L., et al. (2021). Interdisciplinary strategies to enable data-driven plant breeding in a changing climate. *One Earth*, 4(3):372–383.

Lambeth, C., Lee, B.-C., O'Malley, D., and Wheeler, N. (2001). Polymix breeding with parental analysis of progeny: an alternative to full-sib breeding and testing. *Theoretical and Applied Genetics*, 103(6):930–943.

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

Lenz, P., Nadeau, S., Azaiez, A., Gérardi, S., Deslauriers, M., Perron, M., et al. (2020). Genomic prediction for hastening and improving efficiency of forward selection in conifer polycross mating designs: an example from white spruce. *Heredity*, 124(4):562–578.

Lynch, M. and Walsh, B. (1998). *Genetics and analysis of quantitative traits*. Sinauer Sunderland, MA.

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

Miettinen, K. (2012). *Nonlinear multiobjective optimization*, volume 12. Springer Science & Business Media.

Miettinen, K., Hakanen, J., and Podkopaev, D. (2016). Interactive nonlinear multiobjective optimization methods. In Greco, S., Ehrgott, M., and Figueira, J. R., editors, *Multiple criteria decision analysis: State of the art surveys*, pages 927–976. Springer New York, New York, NY.

Moeinizade, S., Hu, G., Wang, L., and Schnable, P. S. (2019). Optimizing selection and mating in genomic selection with a look-ahead approach: An operations research framework. *G3: Genes, Genomes, Genetics*, 9(7):2123–2133.

Moeinizade, S., Kusmec, A., Hu, G., Wang, L., and Schnable, P. S. (2020). Multi-trait genomic selection methods for crop improvement. *Genetics*, 215(4):931–945.

Mwansa, P., Crews Jr, D., Wilton, J., and Kemp, R. (2002). Multiple trait selection for maternal productivity in beef cattle. *Journal of Animal Breeding and Genetics*, 119(6):391–399.

Nachman, M. W. (2002). Variation in recombination rate across the genome: evidence and implications. *Current opinion in genetics & development*, 12(6):657–663.

Namkoong, G. (1976). A multiple-index selection strategy. *Silvae Genetica*, 25:5–6.

Neyhart, J. L., Lorenz, A. J., and Smith, K. P. (2019). Multi-trait improvement by predicting genetic correlations in breeding crosses. *G3: Genes, Genomes, Genetics*, 9(10):3153–3165.

Oldenbroek, K. and van der Waaij, L. (2015). Textbook animal breeding and genetics for bsc students. *Centre for Genetic Resources The Netherlands and Animal Breeding and Genomics Centre*, page 245.

Pixley, K., Hearne, S., Willcox, M., Chen, C., Burgueño, J., Li, H., et al. (2014). Seeds of discovery: characterizing and utilizing maize genetic resources for germplasm diversification. *Maize for Food, Feed, Nutrition and Environmental Security*, page 61.

Reynolds, M., Chapman, S., Crespo-Herrera, L., Molero, G., Mondal, S., Pequeno, D. N., et al. (2020). Breeder friendly phenotyping. *Plant Science*, 295:110396.

Riley, R. and Chapman, V. (1958). Genetic control of the cytologically diploid behaviour of hexaploid wheat. *Nature*, 182(4637):713–715.

Rincent, R., Charcosset, A., and Moreau, L. (2017). Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. *Theoretical and Applied Genetics*, 130(11):2231–2247.

Sawaragi, Y., Nakayama, H., and Tanino, T. (1985). *Theory of multiobjective optimization*. Elsevier.

Scott, M. F., Fradgley, N., Bentley, A. R., Brabbs, T., Corke, F., Gardner, K. A., et al. (2021). Limited haplotype diversity underlies polygenic trait architecture across 70 years of wheat breeding. *Genome Biology*, 22(1):1–30.

Scott, M. F., Ladejobi, O., Amer, S., Bentley, A. R., Biernaskie, J., Boden, S. A., et al. (2020). Multi-parent populations in crops: a toolbox integrating genomics and genetic mapping with breeding. *Heredity*, 125(6):396–416.

Simmonds, N. W. (1995). The relation between yield and protein in cereal grain. *Journal of the Science of Food and Agriculture*, 67(3):309–315.

Snelling, W. M., Kuehn, L. A., Thallman, R. M., Bennett, G. L., and Golden, B. L. (2019). Genetic correlations among weight and cumulative productivity of crossbred beef cows. *Journal of Animal Science*, 97(1):63–77.

Stapley, J., Feulner, P. G., Johnston, S. E., Santure, A. W., and Smadja, C. M. (2017). Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1736):20160455.

Sun, C., VanRaden, P., O'Connell, J., Weigel, K., and Gianola, D. (2013). Mating programs including genomic relationships and dominance effects. *Journal of Dairy Science*, 96(12):8014–8023.

Toro, M., Silió, L., and Pérez-Enciso, M. (1991). A note on the use of mate selection in closed moet breeding schemes. *Animal Science*, 53(3):403–406.

Turner-Hissong, S. D., Mabry, M. E., Beissinger, T. M., Ross-Ibarra, J., and Pires, J. C. (2020). Evolutionary insights into plant breeding. *Current Opinion in Plant Biology*, 54:93–100.

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

Velu, G. and Singh, R. P. (2013). Phenotyping in wheat breeding. *Phenotyping for plant breeding: applications of phenotyping methods for crop improvement*, pages 41–71.

Villanueva, B. and Woolliams, J. (1997). Optimization of breeding programmes under index selection and constrained inbreeding. *Genetics Research*, 69(2):145–158.

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

White, T. L., Adams, W. T., and Neale, D. B. (2007). *Forest genetics*. CABI.

Willcox, M. C., Burgueño, J. A., Jeffers, D., Rodriguez-Chanona, E., Guadarrama-Espinoza, A., Kehel, Z., et al. (2022). Mining alleles for tar spot complex resistance from cimmyt's maize germplasm bank. *Frontiers in Sustainable Food Systems*, page 297.

Witcombe, J. R., Gyawali, S., Subedi, M., Virk, D. S., and Joshi, K. D. (2013). Plant breeding can be made more efficient by having fewer, better crosses. *BMC Plant Biology*, 13(1):1–12.

Zhang, Z. and Wang, L. (2022). A look-ahead approach to maximizing present value of genetic gains in genomic selection. *G3: Genes, Genomes, Genetics*, 12(8):jkac136.

Zobel, B. and Talbert, J. (1984). *Applied forest tree improvement*. Wiley New York.

Supplementary information: Development and optimization of expected cross value for mate selection problems

Pouya Ahadi[1], Balabhaskar Balasundaram[2], Juan S. Borrero[2], and Charles Chen*[3]

[1]*H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA.* [2]*School of Industrial Engineering and Management, Oklahoma State University, Stillwater, Oklahoma, USA.* [3]*Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, Oklahoma, USA.*

*Corresponding author: charles.chen@okstate.edu

PROOFS

Proof of Proposition 1

We model the random vector J that follows an inheritance distribution as a discrete time Markov chain (DTMC) with $J = \{J_n \colon n \geq 0\}$ where J_n represents the state of the process at n-th step, i.e., the value of the random vector J in the n-th position, with the state space $\{0, 1\}$. This process is not a time-homogeneous DTMC. According to Equation (4) in the main article, the transition probability matrix from step k to step $k+1$ is as follows:

$$P_{k:k+1} = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 1 - r_k & r_k \\ r_k & 1 - r_k \end{pmatrix} \end{array} \qquad \forall k \in [N-1].$$

The transition probability matrix from the first step 1 to step $i \in [N-1]$ is then given by:

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

We claim that:

$$P_{1:i} = \begin{bmatrix} 1 - \phi_i(r) & \phi_i(r) \\ \phi_i(r) & 1 - \phi_i(r) \end{bmatrix}, \tag{21}$$

where $\phi_i(r)$ is defined in Equations (5) and (6) in the main article. We prove this claim by induction on i. The claim holds for the base case $i = 2$ by definition, because according to Equation (5) in the main article, $\phi_2(r) = r_1$. Let us suppose Equation (21) holds for step

$i = n$. By induction hypothesis, we know that:

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

As $P_{1:n+1} = P_{1:n} P_{n:n+1}$, we obtain the following:

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

establishing the claim in Equation (21).

The DTMC J satisfies the following property (Kulkarni 2016):

$$\Pr(J_i = j) = \left(\alpha^\top P_{1:i}\right)_j \qquad \forall i \in \{2, 3, \ldots, N\}, j \in \{0, 1\}, \tag{22}$$

where $\alpha^\top = [\alpha_0, \alpha_1]$ is the vector of initial probabilities and $\left(\alpha^\top P_{1:i}\right)_j$ denotes the $(j+1)$-th component of the row vector $\alpha^\top P_{1:i}$. Thus, for every $i \in \{2, 3, \ldots, N\}$,

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

Proposition 1 follows by noting that $\alpha_0 + \alpha_1 = 1$. □

Proof of Theorem 1

We use the definition in Equation (14) in the main article to find a closed-form expression for the ECV. Let L^1 and L^2 be the genotype matrices for the selected pair of individuals, and let J^1 ,J^2 and J^3 be three independent samples from the inheritance distribution. We know that $g^3 = \text{gam}\left(\left[g^1, g^2\right], J^3\right)$ where $g^1 = \text{gam}\left(L^1, J^1\right)$ and $g^2 = \text{gam}\left(L^2, J^2\right)$. Based on the definition of inheritance distribution in Definition 3 in the main article, we have,

$$g_i^1 = L_{i,1}^1(1 - J_i^1) + L_{i,2}^1 J_i^1 \qquad \forall i \in [N], \tag{23}$$

$$g_i^2 = L_{i,1}^2(1 - J_i^2) + L_{i,2}^2 J_i^2 \qquad \forall i \in [N], \text{ and,} \tag{24}$$

$$g_i^3 = g_i^1(1 - J_i^3) + g_i^2 J_i^3 \qquad \forall i \in [N]. \tag{25}$$

Substitutions in Equation (25) using Equations (23) and (24) yields the expected cross value for the target trait as:

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{N} g_i^3\right) &= \mathbb{E}\left(\sum_{i=1}^{N} \left[L_{i,1}^1(1 - J_i^1) + L_{i,2}^1 J_i^1\right](1 - J_i^3) + \left[L_{i,1}^2(1 - J_i^2) + L_{i,2}^2 J_i^2\right] J_i^3\right) \\
&= \mathbb{E}\left(\sum_{i=1}^{N} L_{i,1}^1 + \left(L_{i,2}^1 - L_{i,1}^1\right) J_i^1 - L_{i,1}^1 J_i^3 - \left(L_{i,2}^{k,} - L_{i,1}^1\right) J_i^1 J_i^3 + \right. \\
&\qquad \left. L_{i,1}^2 J_i^3 + \left(L_{i,2}^2 - L_{i,1}^2\right) J_i^2 J_i^3\right) \\
&= \sum_{i=1}^{N} \left[L_{i,1}^1 + \left(L_{i,2}^1 - L_{i,1}^1\right)\mathbb{E}(J_i^1) - L_{i,1}^1 \mathbb{E}(J_i^3) - \left(L_{i,2}^1 - L_{i,1}^1\right)\mathbb{E}(J_i^1 J_i^3) + \right. \\
&\qquad \left. L_{i,1}^2 \mathbb{E}(J_i^3) + \left(L_{i,2}^2 - L_{i,1}^2\right)\mathbb{E}(J_i^2 J_i^3)\right].
\end{aligned}$$

From Proposition 1 we know that,

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

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

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

Thus,

$$\mathbb{E}\left(\sum_{i=1}^{N} g_i^3\right) = \sum_{i=1}^{N}\Big(L_{i,1}^1 + [\alpha_1 + (\alpha_0 - \alpha_1)\phi_i(r)](L_{i,2}^1 - 2L_{i,1}^1 + L_{i,1}^2) + [\alpha_1 + (\alpha_0 - \alpha_1)\phi_i(r)]^2(L_{i,2}^2 + L_{i,1}^1 - L_{i,2}^1 - L_{i,1}^2)\Big). \tag{26}$$

Assuming $\alpha_0 = \alpha_1 = 0.5$ based on Mendel's second law, Equation (26) reduces to Equation (15) claimed in the main article. □

FORMULATIONS

Mathematical formulation for single-trait parental selection

Following Han et al. (2017), we use the following notations in our integer programming (IP) formulation (27). We use ECV as our objective function and add constraints to restrict inbreeding.

Parameters:

- $K \in \mathbb{Z}_{\geq 0}$: Number of individuals in the population
- $N \in \mathbb{Z}_{\geq 0}$: Number of QTL for the target trait
- G: $K \times K$ genomic matrix of inbreeding values with elements $g_{k,k'}$ for $k, k' \in [K]$
- $\epsilon \in \mathbb{R}_+$: Inbreeding tolerance on a pair of selected individuals

Decision variables:

- $t \in \mathbb{B}^{2\times K}$ representing the parental selection decision where,

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

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

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

Objective function: Using Equation (15) in the main article, the ECV can be expressed as a function of the decision variables as: $f(t,x) = 0.25\sum_{i=1}^{N}\sum_{j=1}^{4} x_{i,j}$.

Formulation:

$$\max\ 0.25 \sum_{i=1}^{N} \sum_{j=1}^{4} x_{i,j} \tag{27a}$$

$$s.t. \quad \sum_{k=1}^{K} t_{m,k} = 1 \qquad \forall m \in [2] \tag{27b}$$

$$x_{i,j} = \sum_{k=1}^{K} t_{1,k} L^{k}_{i,j} \qquad \forall i \in [N], j \in \{1,2\} \tag{27c}$$

$$x_{i,j} = \sum_{k=1}^{K} t_{2,k} L^{k}_{i,j-2} \qquad \forall i \in [N], j \in \{3,4\} \tag{27d}$$

$$t_{1,k} + t_{2,k'} \leq 1 \qquad \forall k, k' \in [K] \text{ such that } g_{k,k'} \geq \epsilon \tag{27e}$$

$$t_{m,k} \in \{0,1\} \qquad \forall m \in [2], k \in [K] \tag{27f}$$

$$x_{i,j} \in \{0,1\} \qquad \forall i \in [N], j \in [4] \tag{27g}$$

The objective function (27a) maximizes the ECV. Constraint (27b) ensures that exactly two individuals will be selected for the crossing. Constraints (27c) and (27d) assign genotypic information in genotype matrices of the selected individuals to the $x_{i,j}$ variables. Constraint (27e) ensures that two individuals with genomic relationship coefficient greater than the tolerance ϵ will not be selected simultaneously as parents. As the genomic relationship coefficient between any individual with itself has the highest value of one, this set of constraints will prevent self-crossing between individuals for any value of ϵ less than one. Finally, constraints (27f) and (27g) force decision variables to take binary values.

Algorithm for selecting multiple parental pairs

Suppose we are interested in finding n_c different parental pairs from the population. Assuming that self-crossing is not allowed, we denote the number of feasible solutions (crosses) by n_f, which is bounded above by $\binom{K}{2}$. As we impose a constraint for controlling inbreeding, the

number of feasible crosses might be strictly less than $\binom{K}{2}$. Specifically, the number of feasible solutions (feasible crosses) is precisely half the number of off-diagonal elements in matrix G that are smaller than ϵ.

If there is no element in matrix G that is smaller than ϵ, then $n_f = 0$ and formulation (27) is infeasible. In this case, we need to increase the value of tolerance ϵ such that there might be at least n_f possible crosses for the selection. Then, any positive integer value for n_c such that $n_c \leq n_f$ is suitable for our approach.

Assume that after solving the single-trait formulation (27), we find that in the optimal solution we have $t^*_{1,k} = t^*_{2,k'} = 1$. This solution means that k-th and k'-th individuals are optimal parents that should be crossed. To obtain another pair of parents from the model, we can add the following "conflict constraints" to the single-parent single-trait formulation (27):

$$t_{1,k} + t_{2,k'} \leq 1 \text{ and } t_{1,k'} + t_{2,k} \leq 1. \tag{28}$$

These two constraints will exclude this pair of individuals, k, k', from being selected if we reoptimize formulation (27) with these additional conflict constraints. We can repeat this procedure to find n_c pairs by accumulating the appropriate set of conflict constraints corresponding to individuals selected in the previous iteration. The procedure is summarized in Algorithm 1.

Algorithm 1 Finding multiple pairs for the parental selection problem

1: **Input:** Appropriate n_c (assumed to be no larger than n_f), G, P, ϵ
2: **Output:** Set S of selected parental pairs
3: $S \leftarrow \emptyset$
4: **while** $|S| < n_c$ **do**
5: Solve formulation (27) and obtain optimal solutions $t^*_{1,k} = t^*_{2,k'} = 1$.
6: Add the pair $\{k, k'\}$ to set S.
7: Update the formulation by adding the constraints: $t_{1,k} + t_{2,k'} \leq 1$, $t_{1,k'} + t_{2,k} \leq 1$.
8: **end while**
9: **return** S

Mathematical formulation for multi-trait parental selection

Additional parameters:

- $M \in \mathbb{Z}_{\geq 0}$: Number of target traits for the breeding program
- $N_\ell \in \mathbb{Z}_{\geq 0}$: Number of QTL for the ℓ-th trait $\forall \ell \in [M]$

Additional decision variables:

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

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

Objective function: We define the ECV corresponding to the ℓ-th trait as a function of the decision variables as: $f_\ell(t, x^\ell) = 0.25 \sum_{i=1}^{N_\ell} \sum_{j=1}^{4} x^\ell_{i,j}$. The components of the objective function vector $F(t, x) = \langle f_1(t, x^1), \ldots, f_M(t, x^M) \rangle$ are in decreasing order of importance. Thus, trait ℓ is more important than trait $\ell + 1$, for each $\ell \in [M - 1]$. Note that we denote the collection of variables $\langle x^1, \ldots, x^M \rangle$ succinctly as x.

Formulation:

$$\text{lexmax } F(t,x) = \langle f_1(t,x^1), \ldots, f_M(t,x^M)\rangle, \tag{29a}$$

$$s.t. \quad \sum_{k=1}^{K} t_{m,k} = 1 \qquad \forall m \in \{1,2\} \tag{29b}$$

$$x_{i,j}^{\ell} = \sum_{k=1}^{K} t_{1,k} L_{i,j}^{k,\ell} \qquad \forall i \in [N_\ell], j \in \{1,2\}, \ell \in [M] \tag{29c}$$

$$x_{i,j}^{\ell} = \sum_{k=1}^{K} t_{2,k} L_{i,j-2}^{k,\ell} \qquad \forall i \in [N_\ell], j \in \{3,4\}, \ell \in [M] \tag{29d}$$

$$t_{1,k} + t_{2,k'} \leq 1 \qquad \forall k, k' \in [K] \text{ such that } g_{k,k'} \geq \epsilon \tag{29e}$$

$$t_{m,k} \in \{0,1\} \qquad \forall m \in \{1,2\}, k \in [K] \tag{29f}$$

$$x_{i,j}^{\ell} \in \{0,1\} \qquad \forall i \in [N_\ell], j \in [4], \ell \in [M] \tag{29g}$$

The multi-objective optimization formulation (29) for the multi-trait parental selection problem lexicographically maximizes the vector of ECV functions corresponding to each trait. We describe this approach in greater detail in the next section. Constraint (29b) states that exactly two individuals will be selected for crossing. Constraints (29c) and (29d) will assign genotypes of selected individuals to $x_{i,j}^{\ell}$ variables. Constraint (29e) implies that any two individuals with an inbreeding coefficient greater than tolerance ϵ can not be selected as parents for the crossing program. Note that since the inbreeding coefficient between any individual and itself has the highest value (which equals one), for any value of ϵ less than one, this set of constraints will prevent self-crossing between individuals. Finally, constraints (29f) and (29g) enforce decision variables to take binary values.

Lexicographic multi-objective optimization with degradation tolerances

Define a vector of tolerances $\tau = (\tau_1, \tau_2, \ldots, \tau_M)$ such that $\tau_\ell \in [0, 1]$ for all $\ell \in [M]$. Since we do not need degradation for the last objective, we set $\tau_M = 0$. Tolerance τ_ℓ represents the allowable degradation for the ℓ-th objective function. Let us assume that χ^1 is the set of feasible solutions based on the constraints of formulation (29). Let z_1^* be the optimal objective value for the first objective function $f_1(t, x^1)$ over all feasible solutions in set χ^1. That is,

$$z_1^* = \max\{f_1(t, x^1) \mid (t, x) \in \chi^1\}. \tag{30}$$

As the tolerance for the first objective is τ_1, the set of feasible solutions for the second objective is given by:

$$\chi^2 = \{(t, x) \in \chi^1 \mid f_1(t, x^1) \geq (1 - \tau_1) z_1^*\}, \tag{31}$$

and the best objective value for the second objective function is:

$$z_2^* = \max\{f_2(t, x^2) \mid (t, x) \in \chi^2\}. \tag{32}$$

Generally, the set of feasible solutions for the $\ell + 1$-th objective function and its best objective value are as follows:

$$\chi^{\ell+1} = \{(t, x) \in \chi^\ell \mid f_\ell(t, x^\ell) \geq (1 - \tau_\ell) z_\ell^*\} \qquad \forall \ell \in [M - 1], \tag{33}$$

$$z_{\ell+1}^* = \max\{f_{\ell+1}(t, x^{\ell+1}) \mid (t, x) \in \chi^{\ell+1}\} \qquad \forall \ell \in [M - 1]. \tag{34}$$

The set of "tolerance-optimal" solutions for the problem is given by:

$$\chi^* = \underset{(t,x)\in\chi^M}{\operatorname{argmax}} f_M(t, x^M). \tag{35}$$

By construction, the feasible sets satisfy the following relationship:

$$\chi^* \subseteq \chi^M \subseteq \chi^{M-1} \subseteq \cdots \subseteq \chi^2 \subseteq \chi^1. \tag{36}$$

An optimal solution in multi-objective optimization is called an *efficient* or *non-dominated* solution based on some domination structure chosen by the decision maker (Sawaragi et al. 1985). Arguably, the most well-known notion of efficiency is Pareto optimality. We say that the solution $(\hat{t}, \hat{x})$ is Pareto optimal (or non-dominated) if there is no feasible solution (t, x) to Formulation (29) such that $F(t, x) \geq F(\hat{t}, \hat{x})$ and $f_\ell(t, x) > f_\ell(\hat{t}, \hat{x})$ for at least one trait ℓ (Miettinen et al. 2016).

As we show using the next result, the set of tolerance-optimal solutions χ^* is guaranteed to contain Pareto optimal solutions. Furthermore, if $(t, x) \in \chi^*$ then, either (t, x) is Pareto optimal or it is dominated by a Pareto optimal solution in χ^*.

Proposition 1. *Every solution $(t^*, x^*) \in \chi^*$ is either Pareto optimal, or it is dominated by a Pareto optimal solution in χ^*.*

Proof. If $(t^*, x^*) \in \chi^*$ is not Pareto optimal, then there exists a Pareto optimal solution $(t', x') \in \chi^1$ that dominates (t^*, x^*). Note that the existence of such a solution follows from the finiteness of χ^1. We prove that $(t', x') \in \chi^*$ by contradiction.

Suppose $(t', x') \notin \chi^*$. Then, there exists $i \in [M]$ such that $(t', x') \in \chi^i$ and $(t', x') \notin \chi^{i+1}$ (where $\chi^{M+1} = \chi^*$). Therefore,

$$f_i(t', x'^i) < (1 - \tau_i) z_i^*. \tag{37}$$

If $i = M$, we arrive at a contradiction because inequality (37) implies that (t', x') does not dominate (t^*, x^*). (Recall that $\tau_M = 0$.)

Now suppose, $i \in [M-1]$. By Equation (36), we know that $(t^*, x^*) \in \chi^* \subseteq \chi^{i+1}$, and that,

$$(1 - \tau_i) z_i^* \leq f_i(t^*, x^{*i}). \tag{38}$$

Inequalities (37) and (38) imply that,

$$f_i(t', x'^i) < f_i(t^*, x^{*i}). \tag{39}$$

Again, contradicting the assumption that (t', x') dominates (t^*, x^*). This implies that *every* Pareto optimal solution (t', x') that dominates (t^*, x^*) belongs to χ^*. □

Based on Proposition 1, if we seek a Pareto optimal solution, we can guarantee the identification of one by carrying out an additional step and solving one more optimization problem given by:

$$\max \left\{ \sum_{\ell \in [M]} f_\ell(t, x^\ell) \mid (t, x) \in \chi^* \right\}.$$

References

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.

Kulkarni, V. G. (2016). *Modeling and analysis of stochastic systems*. Chapman and Hall/CRC, 3 edition.

Miettinen, K., Hakanen, J., and Podkopaev, D. (2016). Interactive nonlinear multiobjective optimization methods. In Greco, S., Ehrgott, M., and Figueira, J. R., editors, *Multiple*

criteria decision analysis: State of the art surveys, pages 927–976. Springer New York, New York, NY.

Sawaragi, Y., Nakayama, H., and Tanino, T. (1985). *Theory of multiobjective optimization*. Elsevier.