Title:

How to reach a higher selection plateau with optimal contribution selection and compensatory mating?

Short title:

Optimal contribution selection and compensatory mating.

Authors:

Tiret Mathieu[1,2]*, Pégard Marie[1], Sánchez-Rodriguez Leopoldo[1]

1. INRAE-Orléans, BioForA, ONF, Orléans, France.

2. Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden.

* : Corresponding author: mathieu.tiret@ebc.uu.se.

Abstract

In breeding programs, balancing short-term genetic gain and loss of diversity per generation is essential to sustain a long-term genetic response. Depending on the dynamic of the species, the acceptable trade-off will be different. One of the most common and successful tools to achieve this management is the Optimal Contribution Selection (OCS), which readily mathematically formulate the trade-off between genetic gain and coancestry. However, OCS only accounts for the next generation gain and diversity, which can lead to suboptimality given the uncertainties of random mating and segregation. In this paper, we have extended the OCS by conveniently integrating a way to promote certain parental pairs, so that this method can account for the next t+2 generation. In the study case of Populus *nigra*, fully phenotyped and SNP array genotyped, we have shown that (i) a non negligible part of the long-term success of a breeding strategy depends on the implemented mating strategy, and (ii) favoring a compensatory mating can accelerate the selection without compromising the future diversity.

keywords : genomic selection, quantitative genetics, mating selection, simulation.

Introduction

Genomic selection (Meuwissen *et al.*, 2001), in spite of using more precise Mendelian sampling terms compared to pedigree-based selection, and drastically increasing genetic response doing so, it also accelerates the loss of genetic diversity per unit of time (Goddard, 2009; Hayes *et al.*, 2009; Rutkoski *et al.*, 2015). Yet, over the past decades, a growing number of authors have emphasized the importance of balancing loss of diversity and short-term genetic gain to sustain a long-term genetic response (Brisbane & Gibson, 1995; Woolliams *et al.*, 2002; Jannink, 2010). This is especially true for perennial trees, for which minimal husbandry is provided while they can face harsh environments over a long period.

Optimal contribution selection (or OCS) was introduced by Meuwissen (1997) as a complement to the widely used Best Linear Unbiased Predictor (BLUP). BLUP conveniently integrated family information for increased accuracy but also led to rapid co-selection of relatives. OCS maximizes the genetic gain while maintaining the inbreeding rate to a predefined level. It was originally formulated with the pedigree-based relationship matrix (A), but nowadays it can be easily adapted to the genomic relationship matrix (G). Its simplicity has made it today one of the most successful strategies to address the problem of selection-induced loss of diversity (Meuwissen *et al.*, 2020). Furthermore, its flexibility has facilitated countless extensions. As a non-exhaustive illustration of its potential, De Beukelaer (2017) extended OCS for different measures, such as heterozygosity, or the criterion of Li *et al*. (2008); Gebregiwergis *et al.* (2020) incorporated an alternative formulation of the genomic relationship matrix via QTL and markers; and some authors even attempted to combine it with mating (Varona & Misztal, 1999; Toro & Varona, 2010; Vitezica *et al.*, 2013; Akdemir & Sánchez, 2016).

Mating selection also addresses the long-term dynamics, but on a different timescale than OCS, by controlling homozygosity in the next generation, or equivalently inbreeding at generation t+2. Softening selection on this expanded timescale is of particular interest in breeding in order to reduce the so-called and undesired Bulmer effect (Bulmer, 1971). Therefore, by their different timescales, OCS and mating selection, although incompatible in their formulations, might be complementary (Akdemir & Sánchez, 2016; Yoshida *et al.*, 2020). Some sort of mating selection could be useful for perennials where production plantings consist of elite clones resulting from selected crosses. We have therefore attempted to combine OCS and mating selection with an alternative formulation of the genomic relationship matrix. However, planning mating strategies raises some operational issues, and we need simple solutions that impose mating without the operational constraints.

The twofold challenge of this article, as an extension of the classic OCS, is to (1) devise a way to make OCS promote the likelihood of useful matings, while (2) avoiding the need to explicitly

optimize the mating allocation. To do so, following the approach of Gebregiwergis *et al.* (2020), we explored alternative ways of constructing the genomic relationship matrices to help OCS account for coancestry at the t+2 or higher generation. We believe that studying these questions, and finding a way to incorporate mating into OCS could lead to a methodology that allows for higher selection plateaus (Jannink, 2010) in a long term breeding program. First, we have developed a deterministic algorithm that can solve both classic OCS and any of its extensions. We have then applied our method to the case study of selection in a population of *Populus nigra*, to show by simulations that a multi-generation breeding program can achieve higher performances than with a classic OCS, for a certain range of parameters. Finally, we have shown that, among all the possible ways of constructing the genomic relationship matrix, the best overall strategy in our comparison is to favor mating similar to the so-called compensatory mating (Santiago & Caballero, 1995; Sánchez *et al.*, 2006).

Methods

Optimal contribution selection

Genetic contributions are the cornerstone of OCS. The two opposing items in the optimization of selection in OCS, gain and diversity, can be formulated according to the same decision variable, genetic contribution. The genetic contribution can be defined as the proportion of genes from a given ancestor that are present in a given cohort of descendants. More generally, the genetic contribution of an individual is its proportional contribution to the gene pool of the descendant population (Woolliams & Mäntysaari, 1995). In our case, when considering non-overlapping generations, the genetic contribution of a given parent would simply be its proportional contribution to the offspring of the next generation. We denote **c** the vector of N genetic contributions, with N being the size of the parental population. Defined as such, $\sum c_i = 1$.

With the knowledge of Y the vector of parental breeding values and A the numerator relationship matrix between parents (or G the realized genomic relationship matrix), it is possible to formulate the expected future performance or inbreeding coefficient of the population as c^TY (cY for

simplicity) or as ½ c^TAc respectively (or c^TGc when using the genomic relationship matrix; cAc and cGc for simplicity). Deriving optimal selection decisions simultaneously accounting for future performance and future inbreeding is then possible through OCS, using genetic contributions as a decision variable. One of the simplest formulations of such a problem is to solve min. $cY - \lambda\, cGc$ (see Woolliams *et al.*, 2002 and references therein). It is important to note that the weighted average of c^TAc (or c^TGc) represents the expected inbreeding assuming *panmixia* or, more precisely, uniting in a full diallelic way all parents while respecting each parental *c*. This corresponds to the best expectation for inbreeding when the mating regime is unknown or not under the control of the breeder, which is often the case when, for example, matings are allowed to follow an open pollination regime as in a seed orchard.

As can be demonstrated, cGc is proportional to $\| f - 0.5 \|$, where f is the vector of allelic frequencies in the next generation, and $\|.\|$ the l2-norm. Therefore, minimizing cGc makes the resulting frequencies tend toward 0.5, which is the furthest point from fixation. As selection over cycles tends to accelerate the change of frequencies when they are intermediate, i.e. when the variances are at the highest levels, the risks of loss of favorable alleles by hitchhiking the alternative detrimentals would also increase (Sánchez *et al.*, 2006). By maintaining frequencies at intermediate levels, OCS would reduce this risk with potentially a slower fixation of favorable alleles, but overall benefits over the long-term genetic gain.

OCS and convexe optimisation

Originally, Meuwissen *et al.* (2001) formulated OCS as the maximization of genetic gain (cY) subject to a constrained inbreeding coefficient (cAc). Likewise, it is also possible to minimize the inbreeding rate while constraining the genetic gain (Akdemir & Sánchez, 2016). Choosing the adequate constraint is critical for populations never confronted to OCS. The methodology was primarily devised with long-term domesticated populations in mind, where records of change in inbreeding and genetic gain are typically known over several cycles, facilitating the choice of constraints. In the absence of historical references, for novel species, a gradient of constraints

would need to be evaluated *a priori*. In this sense, a holistic approach allowing visualization of the optimized function over a wide range of scenarios would be preferable as a start.

Following the approach of Akdemir *et al.* (2019), we considered OCS as a multi-objective optimization, where gain and diversity are improved simultaneously, i.e. maximizing gain and minimizing coancestry, by pondering weights that set the balance between the two items. The solutions of a multi-objective optimization, called Pareto optima (Figure S1), delineate a two-dimensional curve.

Optimizing in two dimensions (genetic gain and coancestry) does not have the same meaning as optimizing in one dimension with constraints. Mathematically, multi-objective optimization (or minimize (cGc, -cY)) is equivalent to the scalarized version of the problem: minimize λ cGc - cY for any $\lambda > 0$. The scalarized problem has a unique global solution since the objective function is strictly convex (G is positive definite, as shown below). In other words, the curve of Pareto optima is the parametric curve, as a function of λ, of solutions minimizing λ cGc - cY. It is therefore the set of (c*Gc*, c*Y), where c* is the optimal contribution vector for a given λ. Without loss of generality, we consider the scalarized problem parameterized by α such that the problem becomes:

$$\text{min. } \alpha\, cGc - (1 - \alpha)\, cY \qquad (1)$$

where $\alpha \in [0;1]$ can be interpreted as the tradeoff value between coancestry and genetic gain (or the weight of coancestry compared to that of genetic gain).

Each OCS solution given a constraint (as in Meuwissen *et al.*, 2001) is a particular Pareto optimum, or, in other words, is the solution of the scalarized problem for a particular α (if the constraint is not ill-formulated, i.e. not out of range). Choosing a value for α is as arbitrary as choosing a value for a constraint when there is no *a priori*, but at least the range of α is known, and

its interpretation is biologically intuitive : α is the quantification of the tradeoff between genetic gain and diversity. Different selection scenarios can then be compared by varying a single parameter.

We have furthermore added some operational constraints to the multi-objective problem: the contributions must be larger than 0 to have a biological meaning but smaller than 0.5 to avoid selfing ($0 \leq c \leq 0.5$), and the sum of all contributions must be equal to 1 ($\mathbf{1}^T.c = 1$). The constrained scalarized problem for a particular value of α is a constrained quadratic programming and can therefore be solved deterministically with an interior point method, adequate to solve constrained convex optimization (Boyd *et al.*, 2004). The algorithm used here (coded in R − version 3.6.3) is available on github (https://github.com/mathieutiret/ocs_solve.git).

Different genomic matrices

Let X be the matrix describing the genotypes of the population, with L rows (number of markers) and N columns (number of individuals). The two homozygous states are encoded as −1 and 1, and the heterozygous state as 0 (as in VanRaden, 2008). The realized genomic relationship matrix is then formulated as $G = X^TX$ (N x N matrix). Note that the elements of X are not corrected by minor allele frequencies, nor should the resulting G be scaled by the expected heterozygosity, as is usually done for genomic estimation. Therefore, the diagonal elements of G provide information on the number of homozygous loci per individual, while the off-diagonal elements reflect the number of homozygous states shared by individuals across loci. Thus, for off-diagonals, the same homozygous state at a given locus adds one unit to the count, while one unit is subtracted for opposite homozygous states (not accounting for heterozygote by heterozygote, pointed out by Gao & Martin, 2009), producing overall large values for pairs with resembling parents and small values for pairs with genetically distinct parents. For illustration, the matrix multiplication for one locus was summarized in Table S1.

We can see G as a penalty matrix when used to minimize the future population coancestry (or cGc). Some parental combinations will be penalized, such as the pairs comprising parents with the

same homozygous state (1 in Table S1). Likewise, some pairs will be promoted, such as those with distinct homozygous states (-1 in Table S1), as they will produce heterozygous offspring It is possible to generalize how different pairs contribute differentially to future diversity, as summarized in Table 1. In this generalization, β and γ range from -1 to 1, respectively weighting heterozygous x heterozygous (He x Ho) pairs, or heterozygous x heterozygous (He x He) pairs. He x Ho pairs were called compensatory matings in previous studies when involving a selected QTL (Caballero *et al.*, 1996; Sánchez *et al.*, 2006). A high value of γ penalizes pairs of heterozygous candidates such that they would not be promoted among the selected parents and their likelihood of mating be consequently reduced. Playing with γ and β makes it possible to influence the mating strategy by pondering different states among the contributing parents according to their future impact on diversity. Hereafter, we denote G* the matrix constructed according to Table 1, with G* = G + β Q + γ W, Q being the matrix accounting for heterozygous x homozygous loci and W accounting for heterozygous x heterozygous loci, as G account for homozygote loci (details in Supplementary materials).

Table 1. Generalization of the genomic relationship matrix for one locus, with parameters β and γ, both ranging from -1 to 1, weighting respectively heterozygous x homozygous pairs, or heterozygous x heterozygous pairs.

parents	− 1	0	1
− 1	1	β	− 1
0	β	γ	β
1	− 1	β	1

One might notice that, for one single locus, He x He pairs are different from He x Ho in terms of resulting diversity at the generation t+2. G does not distinguish between these matings, although G* could. An alternative construction to correct this drawback would involve weightings resulting in relatedness among the offspring equal to the Jacquard distance, which would correspond to $\beta = 0.25$ and $\gamma = 0$ in G*. The allele sharing distance of Gao & Martin (2009) offers another alternative construction for G*.

In addition, we can derive that (details in Supplementary materials) :

$$c^T G^* c = c^T G c + \beta\, c^T Q c + \gamma\, c^T W c = c^T G c + \beta\, 2He^T Ho + \gamma\, He^T He, \qquad (2)$$

where He is the vector (of size L) of the proportion of heterozygous individuals contributing to the next generation for each locus, and Ho the equivalent for homozygous individuals. We can see from this equation that controlling β and γ allows us to change the frequency of the carriers of favorable genotypic states, which in turn would favor the occurrence of certain crosses increasing the segregation of diversity, for instance by promoting double heterozygous pairs over homozygous x heterozygous pairs. Such extra segregation could intuitively allow for a more sustainable genetic progress over the long term, without the risk of hitchhiking. It is important to note that using alternative G* in OCS is not intended as a mating selection *per se* since OCS focuses on contributions per individual, but rather as a strategy for managing sustained heterozygosity.

For the objective function to be strictly convex, G* must be positive definite. Therefore, we performed a spectral projection of G* on the set of positive definite matrices, ensuring that the projected matrix is the closest positive definite matrix to G* (according to the spectral norm; Boyd *et al.*, 2004). We will from here onward denote the projection of G* as G* itself to ease the reading.

Other approaches to handle mating such as Minimum Coancestry Mating (Sonesson & Meuwissen, 2000) were proposed, but additional steps on top of OCS makes it computationally or mathematically demanding, whereas modifying the genomic relationship matrix directly and easily modifies how OCS behaves and impacts matings.

Genomic data

The population used in the study included 1009 individuals from the French breeding population of *Populus nigra* (Pégard *et al.*, 2020). All of them were genotyped with a 12k Infinium array (Faivre-Rampant *et al.*, 2016) resulting in 5253 usable SNP markers after quality and frequency filtering (minor allele frequency higher than 0.05). The resulting genotypes were phased, imputed and a consensus recombination map derived (Pégard *et al.*, 2019) by using FImpute software (Sargolzaei *et al.*, 2014). The allelic effects were estimated from a genomic multitrait evaluation using breedR (Muñoz & Sanchez, 2020). In this study, we considered the trunk circumference as a focal trait.

Poplar is a dioecious species. Sex, however, cannot be determined before seven years of age, nor can it be predicted from the genomic profile yet. For the prospections in the study, and to overcome the missing sex of unsexed candidates, we assumed a monoecious population. However, our method could be easily extended to dioecious populations by adding one simple constraint : $d^T c = 0.5$, where d is a design vector indicating the female/male individuals.

Due to the arbitrary sampling of poplar individuals required for breeding purposes, the dataset used here was strongly structured (Figure S2). In order to check for possible effects of the structure, in addition to the original dataset, we considered another dataset in which we shuffled the alleles at each locus so that the allele frequencies remained unchanged, while linkage disequilibrium could be removed.

Simulation pipeline

We considered different simulation scenarios, assuming different values of α, β and γ. We simulated multi-generation breeding schemes of a constant population size N at each generation,

and the parameters (α, β and γ) remained constant across generations within a given scenario. No introgression of external genetic material was considered here. As mentioned above, we estimated the allelic effects from the genotype data with a real phenotype (trunk circumference). The resulting allelic effects were considered “true” and constant across generations, and used to obtain the True Breeding Value (TBV) of the newly simulated candidates.

When simulating multi-generation breeding programs, the question arises as to how genomic estimated breeding values (GEBVs) should be assessed at each new virtual generation. Due to the strong demand for computing resources, we mainly considered the case of heritability $h^2 = 1$, that is GEBVs equal to TBVs. However, we considered a smaller set of simulations for the case of $h^2 = 0.66$, with only α varying (β and γ fixed to 0). In this case, the newly generated phenotypes underwent a normally distributed perturbation of mean 0 and variance $(1 - h^2)\sigma_g^2$, where σ_g^2 was the TBV variance. At each generation, as was done in Jannink (2010) and De Beukelaer *et al.* (2017), GEBV was assessed with a ridge regression model (Searle 2006), without including all previous generations in the genomic evaluation (which would have been done in a real evaluation). The precision therefore became gradually smaller over generations.

OCS was then applied at each generation on the simulated GEBVs, and the resulting contributions were converted into a random mating plan that fits the OCS solution. Finally, with the mating plan, we simulated the next generation with an ad hoc program (written in C++17 and available on github at https://github.com/mathieutiret/meiosis.git).

Every simulated replicate proceeded for 20 (non-overlapping) generations, and was devised without mutation. The simulation followed a grid of parameters: α ranged from 0.1 to 0.9 (with steps of 0.1), β ranged from − 1 to 1 (with steps of 0.5), and γ ranged from − 1 to 1 (with steps of 0.5). Each parametric setting was simulated 100 times. In order to compare the results from different β and γ, we also computed for each simulation the “true” coancestry, which is the coancestry the population would have if β and γ were equal to 0.

Statistical analyses

Finally, in order to determine the relative importance of the factors α, β and γ on the genetic gain and coancestry (through an analysis of variance) at a given generation, we considered the following statistical model:

$$Y \sim \alpha + \beta + \gamma + \alpha^2 + e,$$

where Y is the output variable (either genetic gain or true coancestry), α, β and γ the simulation parameters treated as fixed effects, α^2 the squared parameter of α, and e the residual fitted to a normal distribution. We consider here a polynomial regression (quadratic term α^2) as there could be some symmetrical boundary effects of α upon Y. We will hereafter refer to this as model (1). In some cases however, we focused on a given value of α, in which case we considered the following model:

$$Y \sim \beta + \gamma + e,$$

with equal notations and assumptions as model (1). We will hereafter refer to this as model (2). These two models were analyzed using R - 3.6.3 (functions *lm* and *anova* from the base R package *stats*). We considered a Type I analysis of variance for polynomial regression, and the statistical significance of each factor was assessed with a Fisher's *F*-test.

Results

Overall evolution of genetic gain and coancestry met common expectations

In this section, we only studied the effect of α, considering β and γ both equal to 0, corresponding then to the classical formulation of OCS. As expected, the average value of both genetic gain and

coancestry increased over time (Table 2). Moreover, and also as expected, the increases in genetic gain and coancestry were more pronounced for lower α. Different heritabilities led to a different average genetic gain after twenty generations: 161.8 ± 4.2 when $h^2 = 1$, and 148.7 ± 4.9 when $h^2 = 0.66$. When looking at the ratios between changes in coancestry and changes in genetic gain over the studied period (in other words a measure of the cost of diversity per unit of genetic gain), it is clear that increases in α reduced the impact of genetic gain on diversity accordingly. In the extreme with the highest α, producing genetic gain has no perceptible cost in coancestry. Such efficiency was most evident when raising the heritability levels. This result is again expected, corroborating that selection is more efficient when the heritability of the trait is higher.

Table 2. Average values, standard deviations and coefficients of variation (CV, in percentages) of genetic gain or coancestry depending on the heritability, at generation 1 and 20, with β = γ = 0.

		genetic gain				**coancestry**			
		h^2 = 0.66		**h^2 = 1**		**h^2 = 0.66**		**h^2 = 1**	
		t = 1	**t = 20**	**t = 1**	**t = 20**	**t = 1**	**t = 20**	**t = 1**	**t = 20**
α = 0.1	**μ**	42.1	148.7	44.6	161.8	0.3	1.0	0.3	1.0
	σ	1.2	4.9	0.1	4.2	0.03	0.01	0.0001	0.01
	CV	3.0	3.3	0.2	2.6	8.5	1.0	0.1	0.9
α = 0.9	**μ**	34.2	103.0	35.3	118.3	0.2	0.2	0.2	0.2
	σ	0.3	1.8	0.11	2.1	0.0004	0.002	0.0001	0.003
	CV	1.0	1.8	0.32	1.75	0.2	1.36	0.07	1.70

There is, however, no expectation regarding the genetic gain and coancestry variances, since OCS formulates its objective function and constraints in terms of expected values, not of variances. Nevertheless, simulations showed a consistent increase in the variance and coefficient of variation (CV) for genetic gain over time (Table 2). As for coancestry, the variances also increased, except for the combination $h^2 = 0.66$ and $\alpha = 0.1$, where both the variance and CV decreased. The increases in variance for both genetic gain and coancestry were less pronounced for $\alpha = 0.9$ than for $\alpha = 0.1$ (Table 2), suggesting that the risk associated with a targeted genetic gain was better controlled when restricting coancestry. Finally, in the long run, variances and CV were similar between different values of heritabilities, suggesting that even with a perfect heritability, the variance eventually increases. It can be noted, however, that the variances and CV remained overall low (CV < 10%).

Trading genetic gain for diversity decreases short-term genetic gain, but increases long-term gain

In this section, we studied the evolution pattern of genetic gain and coancestry in the case of β and γ equal to 0, on the shuffled dataset, as it is the case for “classic” OCS.

Short-term vs long-term genetic gain. Different long-term horizons of genetic gain were reached depending on the value of α when $h^2 = 1$ (Figure 1). In the relatively short-term, less than 6 generations, the achieved genetic gain decreased with increasing α. Eventually, in the longer term, the optimum α for genetic gain shifted from lower values to intermediate and then higher values of α. At the highest horizon, the lower the value of α, the lower the plateau of genetic gain, or the sooner the bend marking the start of the plateau. Such a plateau has already been described in the literature for OCS (Jannink, 2010; De Beukelaer *et al.*, 2017), indicating a trade-off between the short and long-term horizons when setting the importance of gain versus diversity. When $h^2 = 0.66$ (not shown), the increase in genetic gain over time was slower than when $h^2 = 1$, with no view of the bend preceding the plateau: as stated above, the increase in genetic gain is slower for smaller heritabilities.

Figure 1. Evolution over time of coancestry (left panel) or genetic gain (right panel), for different values of α (dark blue for α = 0.1, light blue for α = 0.9), with β = γ = 0, and h^2 = 1.

Short-term vs long-term coancestry. The parameter α also had a strong effect on the pattern of evolution of coancestry when h^2 = 1 (Figure 1). The optimum α for coancestry (i.e. α = 1) did not change across generations, i.e. the same ranking of α according to coancestries occurred for all generations. The most relevant feature, however, is that the change in coancestry between extreme values of α was larger than those observed for gain. In the long term (at generation twenty), coancestry increased by 448% when α shifted from 0.9 to 0.1 (Table 2). In the same time horizon and over the same shift of α, genetic gain had only increased by 37%. Such a difference in scale of response between gain and coancestry is clear even at the first generation of application of OCS. When observing the Pareto function (Figure S1), depicting the best trade-offs between gain and coancestry across the range of α, it is clear that there is potentially a substantial reduction in coancestry with minimum costs in gain whenever α is set between zero and intermediate values (α < 0.5). The difference between h^2 = 1 and h^2 = 0.66 was significant for genetic gain (all p < 0.05/180, Student's t-tests with Bonferroni correction for 180 tests), while it was not for coancestry (only 58 Student's t-tests showed a p-value < 0.05/180).

The effect of α as the most explanatory factor after a long run

After one generation of OCS. Model (1) showed strong adjusted coefficients of determination for both genetic gain (r^2 = 0.86) and coancestry (r^2 = 0.95) (Table S2 and S3, respectively), thus suggesting a strong explanatory power of the parameters α, β and γ. Likewise, in the small dataset with h^2 = 0.66 (β and γ fixed to 0), model (1) showed strong coefficients for both genetic gain (r^2 = 0.96) and coancestry (r^2 = 0.96).

The proportion of the sum of squares varied across parameters, ranking α as the most explanatory factor for both genetic gain (76.8%) and coancestry (78.2%). β had a weak effect on both genetic

gain (4.0%) and coancestry (0.07%), showing that weighting pairs according to their impact on segregation is not efficient for controlling the population in the very short-term. Comparatively, effects of γ were small for both genetic gain (1.4%) and coancestry (0.07%). Knowing that controlling mating only has an effect on a population at generation t+2, β and γ were expected to have small effects, as confirmed by these simulations: their effect is expected to be relevant in the subsequent generations. Quadratic coefficient of α did not have a strong sum of squares for genetic gain (4.0%), but rather large for coancestry (17.1%), making the coefficient α explain 95.3% of the total variance: in the short-term, α is the main predictive and explanatory factor.

After twenty generations of OCS. Model (1) still showed strong adjusted coefficients of determination for both genetic gain (r^2 = 0.79) and coancestry (r^2 = 0.93). For genetic gain, quadratic α was the most explanatory factor, followed by α causing the joint sum of squares of α to be equal to 63% (21% + 42%; Table S2). Likewise, for coancestry, the effect of α was the most explanatory factor, with the joint sum of squares equal to 91.5.3% (91.3% + 0.2%; Table S3), suggesting that the overall population dynamics (genetic gain and coancestry) is mostly determined by α. From generation one to twenty, the effect of β increased from 4.0% to 15.9% for genetic gain, and increased from 0.07% to 1.4% for coancestry: its effect converged towards intermediate values for genetic gain and remained low for coancestry, showing that He x Ho pairs had a lesser but still not negligible role in the population dynamics. The effect of γ decreased to an even smaller value for both genetic gain (0.007%) and coancestry (0.01%), corroborating the fact that He x He pairs are negligible for the population dynamics. Overall, in the long run, α was the predominant effect, with β coming second, and a negligible effect for γ. As for the small dataset with h^2 = 0.66, model (1) showed strong adjusted coefficients of determination for both the genetic gain (r^2_{adj} = 0.99) and coancestry (r^2_{adj} > 0.99), showing that even with a weaker heritability, α was still a strong explanatory factor of the long-term population dynamics. Similar results have been shown in previous studies (see Clark *et al.*, 2013).

Compensatory matings as the most efficient mating strategy

In this section, we finally studied the influence of all the parameters, α, β and γ, on a multi-generation OCS, on the shuffled dataset, with a single value of heritability ($h^2 = 1$). Each parametric combination was kept constant across generations for a given set of parameters. It is important to remember that β weights the importance of He x Ho pairs in G* (higher β, lower importance), while γ weights the importance of He x He pairs (higher γ, lower importance).

Negligibility of He x He pairs. Model (2) showed a much stronger effect of β on both genetic gain and coancestry than γ (Figures S3 and S5). For high values of α (α > 0.4), the proportion of genetic gain variance explained was much higher for β (> 30%) than for γ (< 5%), again suggesting a stronger effect of He x Ho pairs on genetic gain than He x He pairs (Figure S4). Indeed, the estimated values of β were increasingly negative as α increased, corroborating that favoring He x Ho pairs (negative β) improves genetic gain. On the other hand, the estimates of γ remained close to 0, showing its negligibility. Likewise, for coancestry, β had a higher explanatory power (> 70%) than γ (< 10%, Figure S6). The effects of β increased as α increased, showing that favoring He x Ho pairs (β < 0) would enable to maintain a lower coancestry.

Advantage of semi compensatory mating. After twenty generations, the performances in terms of genetic gain or (true) coancestry were different according to α and β (Figure 2). For low values of α (< 0.5, i.e. promoting the maximization of genetic gain over the minimization of coancestry), both genetic gain and coancestry were strongly influenced by α, but only negligibly by β, suggesting that when genetic gain is the main driver of OCS, information about preferential pairs does not allow further improvement in performance in the breeding system. However, for larger α (> 0.5, i.e. promoting the minimization of coancestry over the maximization of gain), favoring He x Ho pairs (β < 0) or no mating (β = 0) resulted in best performances, i.e. a higher genetic gain for a given value of (true) coancestry. More precisely, the apparent Pareto function is upwardly bounded by β = 0 and β = -0.5, and this is even more clear when α tends towards its maximum value, suggesting that optimality is between the "*no mating*" strategy and the semi-compensatory mating.

It can be noted that full compensatory mating (β = -1) is suboptimal, detached from the Pareto limit, compared to the previous two strategies. The results thus suggest that β has a significant role when coancestry is the main driver of OCS. In this case, promoting the He x Ho pairs could be as favorable as none, both situations being on the Pareto curve.

Figure 2. Coancestry versus genetic gain, at generation 20, for different values of α (dark blue for α = 0.1, light blue for α = 0.9) and β (circle for β = -1; triangle for β = -0.5; square for β = 0; cross for β = 0.5; and square crossed for β = 1), with γ = 0 and h^2 = 1. Coancestry on the x-axis is the true coancestry, meaning that it is computed assuming that β = γ = 0.

However, considering an alternative way of assessing the potential of a breeding program by measuring the GEBV of the population from favorable alleles that are not fixed yet, the two alternatives with a "no mating" strategy (β = 0) and the semi-compensatory mating (β = -0.5) can be ultimately distinguished (Figure 3). For α < 0.5, β = 0 is the best strategy with a slightly higher genetic potential over counterparts. As α increases beyond 0.5, however, the average GEBV gets progressively higher for β = -0.5, meaning that, while the two alternatives are equally on the Pareto curve, semi-compensatory mating keeps a higher potential of genetic gain, potentially ensuring a higher selection plateau.

Figure 3. Boxplots of GEBV, at generation 20, for different values of α and β, when only considering alleles that are not yet fixed, with γ = 0 and h^2 = 1.

Discussion

Long-term strategy in breeding programs

In a multi-generation breeding program, being able to select favorable alleles with little losses due would be the most desirable feature. Drift often occurs through unwanted genetic hitchhiking when favorable and unfavorable alleles are *trapped* by limited sampling in continuous segments in

linkage disequilibrium. There is therefore always a risk of loss as recombination might not be able to cope with the pace of selection and sampling generating the unfavorable linkage. One way to render recombination more efficient without slowing down the selection process would be to favor the pairing of candidates with a high potential for segregation in the offspring. It can be done more or less explicitly. One of the classical approaches, as shown in previous works (Jannink, 2010; De Beukelaer *et al.*, 2017; Allier *et al.*, 2019), consists in accounting for diversity through the tradeoff parameter α, where diversity among candidates is modeled through relatedness or coancestry. This weighted approach, or the constrained formulation, can enhance the selection plateau, but often at the cost of slowing down the rate of progress. When such a perspective is applied to late maturing perennials, the cost in time for slow progress becomes a heavy drawback.

Another alternative to accelerate breeding without the drawback of drift is to minimize the uncertainties concerning the consequences of selection decisions, so that decisions can be based on sounding predictions of the impact of selection on future generations. A way of doing this is to account for the mating regime between candidates, which is typically set at random for OCS. Explicit consideration of mating has proven to be a successful alternative (see Akdemir & Sánchez, 2016), although often computationally demanding and not deterministically solvable, except for regular, very specific schemes based on a “round-robin” handling of reproductive roles and genetic contributions (Sánchez *et al.*, 2003). We tried a different, simpler approach here, by tweaking the OCS to implicitly reflect the choice of pairs: we obtained better long-term performances and the problem is deterministically solvable. We have shown that modifying the genomic relationship matrix could be an easy solution with a satisfying compromise between high performance and doability.

When focusing on genetic gain (low α), the best modification of the genomic relationship matrix was undoubtedly none (β = γ = 0), meaning that no mating plan can counterbalance the effect of drift. However, when considering higher α, the best modification was when we favored He x Ho pairs (lower β). Such pairs can be seen as a kind of compensatory mating, and they have already

been described as performing alternatives when it comes to minimize hitchhiking effects in selected populations (Caballero *et al.*, 1996; Sánchez *et al.*, 2006). Our work suggests that benefits of compensatory mating come from maintaining a high selective potential among not yet fixed alleles (Figure 3).

Mating (β and γ) as a nested effect of diversity (α)

In the OCS formulation, α remains the most important factor determining gain and diversity in the population dynamics over generations. In its classical formulation, i.e. β and γ equal to zero, OCS has become one of the most efficient strategies with a combination of genetic gain and coancestry making the very edge of the Pareto optimum, as expected. When allowing for different weightings for the breeding pairs through the extra parameters β and γ, it was clear that pair choice had also a non-negligible role that increased in importance over generations (Tables S2 and S3), both for genetic gain and coancestry. Mating, seen through the implicit formulation derived here, was then the second most important factor in the optimal outputs. We could safely conjecture that through an explicit formulation such importance could be further enhanced, although it would probably remain second to the effect of α, as already pinpointed by Toro & Varona (2010).

For a given α, and notably for the highest values, β modulated the genetic gain over generations in a direct and conspicuous manner (Figure 2): promoting He x Ho pairs (i.e., $\beta < 0$) increased genetic gain, while penalizing He x Ho pairs ($\beta > 0$) decreased genetic gain. Increasing genetic gain usually means: (i) a lower genetic (and genic) variance after selection, (ii) a higher level of fixation of favorable alleles, which constitutes the matter making up genetic gain, and (iii) a higher level of negative linkage disequilibrium covariance due to the Bulmer effect (Bulmer, 1974).

Therefore, the extra gain obtained from $\beta < 0$ could come from using more efficiently the genic variance, resulting in more depletion compared to that of higher β levels (Figure S7), and thus converting this available variation into favorable allele fixation, or likewise unfavorable allele elimination. Compared to $\beta = 0$, $\beta = -0.5$ was always on the side of the Pareto curve with higher coancestry (and gain). Promoting He x Ho pairs, which could theoretically boost segregation in

descendants more than any other combination, could be the source of the extra genetic variation that is made available favorably for selection under $\beta < 0$. On the other hand, promoting segregation across loci could be a source of linkage disequilibrium generating negative covariation of effects, as the chances of uniting alleles of opposing effects across neighboring loci could be increased and further boosted by selection. Such a tendency is clearly shown in Figure S8.

When looking at the opposite extreme, when $\beta > 0$ and notably at $\beta = 1$, genetic gain is lowest while the levels of genotypic covariance due to linkage disequilibrium are soon recovered from initial Bulmer effect depression and reached the highest values (Figure S8). We can confidently assume that selection was weak with $\beta = 1$, leaving available variance unused. A substantial part of the genotypic variance is inflated due to positive linkage disequilibrium, which means that the underlying genic variance is not that high. This is confirmed partially by Figure S7 on genic variance, notably for the extreme case of $\beta = 1$.

If He x Ho are penalized with $\beta = 1$, we could suppose that other pairs like Ho x Ho get promoted instead, and the fact that positive linkage disequilibrium is present would mean that genotypes concentrate either favorable homozygotes or unfavorable homozygotes. This situation, or when the genotypic variance is higher than the genic variance, is potentially favorable for selection to discriminate efficiently between favorable and unfavorable alleles, the extreme case being that of sublines or lineages with selection proceeding between them. This potentiality, however, did not materialize in better gains for $\beta = 1$, suggesting that there is a cause of inefficiency preventing selection from using that advantage. One such cause could be the fact of having unfavorable alleles segregating in the population with a risk of fixation, dragging gain downwards. The analysis on effect of selection and drift − that is, the accumulation of favorable allele and loss of unfavorable allele, or respectively the accumulation of unfavorable allele and loss of favorable alleles (Figure S9) − was not conclusive. They are surprisingly equivalent, almost symmetrical, suggesting that selection has no advantages over drift, and whenever favorable alleles get fixed, unfavorable counterparts get fixed on the way.

Variance in OCS equation

The main challenge of OCS lies in the management of stochasticity: the objective function, as stated above, is formulated with expected values, and not with variances, thus neglecting the variability caused by the uncertainties of random mating and the Mendelian sampling. This causes stochasticity around predicted Pareto optima, even more pronounced when considering a low heritability. In that latter case, selection precision is expected to be low, giving higher chances to drift to lower the selection plateau.

Introducing variability in the parameter could also be desirable in the case of multi-generation breeding, such as considering different values of α, β and γ for each generation, depending on the current state of the population. For instance, considering a high value of α (high diversity) could be important at the very short term to prevent losses of favorable alleles in low frequency, but once the Bulmer effect is absorbed by recombination, it could be safely to switch to a more aggressive strategy such as with lower values of α, and semi compensatory mating (β = -0.5). Preliminary works showed indeed that a decreasing α along the generations brings better long-term performances. Differential selection over generation is therefore a field worth investigating, and warrants further studies.

One potential source of uncertainty in our study could be the limitation of the evaluation population to the current and the previous parental generation in order to simplify and speed up the simulations. To some extent, such a constraint could have favored the scenarios with higher α, which were less affected by evaluation errors (data not shown). Indeed, they provided a wider diversity on which to evaluate more precisely the allelic effects. However, it would not be expected that such an effect distorts the comparison across neighboring values of β.

Conclusions

In this paper, we have shown an intuitive way to extend the classic OCS by (1) reformulating the concept with the tradeoff α (and providing a deterministically solvable algorithm) and (2) accounting

implicitly for mating by extending the formulation of the genomic relationship matrix. As shown above, in multi-generation breeding programs it is important to account for diversity to reach a higher selection plateau, even though the speed at which it is reached can be slow. Accounting for mating, even implicitly as proposed here by favoring some kind of semi compensatory mating, could minimize the speed problem, by “accelerating” the breeding while maintaining a high level of diversity and selective potential for future generations. Finally, this study could open up new possibilities to make OCS a more adaptive tool to handle the dynamics of genetic diversity in selected populations.

Funding

This study was funded by the following sources: the INRA SELGEN funding program (project BreedToLast) providing sequencing, genotyping data and PhD grant for MP; and from the European Union’s Horizon 2020 research and innovation program under grant agreement No 676876 for basic functioning and postdoc grant for MT.

Acknowledgments

The authors acknowledge the GIS peuplier for access to data.

Author Contribution

TM and SL conceptualized the analyses. PM worked on the data curation. SL supervised the work, administrated the project and acquired the funding and resources. TM developed the software and analyzed the data. TM, PM and SL wrote the manuscript.

References

Akdemir D, Sánchez JI. Efficient breeding by genomic mating. Frontiers in genetics. 2016; 7: 210.

Akdemir D, Beavis W, Fritsche-Neto R, Singh AK, Isidro-Sánchez J. Multi-objective optimized genomic breeding strategies for sustainable food improvement. Heredity. 2019; 122: 672-683.

Allier A, Lehermeier C, Charcosset A, Moreau L, Teyssèdre S. Improving short-and long-term genetic gain by accounting for within-family variance in optimal cross-selection. Frontiers in genetics. 2019; 10: 1006.

Boyd S, Boyd SP, Vandenberghe, L. Convex optimization. Cambridge university press; 2004.

Brisbane JR, Gibson JP. Balancing selection response and rate of inbreeding by including genetic relationships in selection decisions. Theoretical and Applied Genetics. 1995; 91: 421-431.

Bulmer M. The effect of selection on genetic variability. The American Naturalist. 1971; 105: 201-211.

Caballero A, Santiago E, Toro MA. Systems of mating to reduce inbreeding in selected populations. Animal science. 1996; 62: 431-442.

Caballero A, Toro MA. Interrelations between effective population size and other pedigree tools for the management of conserved populations. Genetical Research. 2000; 75: 331–43.

Clark SA, Kinghorn BP, Hickey JM, Van der Werf JH. The effect of genomic information on optimal contribution selection in livestock breeding programs. Genetics selection evolution. 2013; 45: 1-8.

De Beukelaer H, Badke Y, Fack V, De Meyer G. Moving beyond managing realized genomic relationship in long-term genomic selection. Genetics. 2017; 206: 1127-1138.

Faivre-Rampant P, Zaina G, Jorge V, Giacomello S, Segura V, et al. New resources for genetic studies in Populus nigra: genome-wide SNP discovery and development of a 12k Infinium array. Molecular Ecology Resources. 2016; 16: 1023–1036.

Gao X, Martin ER. Using allele sharing distance for detecting human population stratification. Human heredity. 2009; 68: 182-191.

Garcia-Cortes LA, Legarra A, Chevalet C, Toro MA. Variance and Covariance of Actual Relationships between Relatives at One Locus. PLOS ONE. 2013; 8: e57003.

Gebregiwergis GT, Sørensen AC, Henryon M, Meuwissen T. Controlling Coancestry and Thereby Future Inbreeding by Optimum-Contribution Selection Using Alternative Genomic-Relationship Matrices. Frontiers in Genetics. 2020; 11: 345.

Goddard M. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 2009; 136: 245-257.

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of dairy science. 2009; 92: 433-443.

Jannink JL. Dynamics of long-term genomic selection. Genetics Selection Evolution. 2010; 42: 1-11.

Li Y, Kadarmideen H, Dekkers J. Selection on multiple qtl with control of gene diversity and inbreeding for long-term benefit. J. Anim. Breed. Genet. 2008; 125: 320–329.

Meuwissen THE. Maximizing the response of selection with a predefined rate of inbreeding. Journal of animal science. 1997; 75: 934-940.

Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001; 157: 1819-1829.

Meuwissen THE, Sonesson AK, Gebregiwergis G, Woolliams JA. Management of Genetic Diversity in the Era of Genomics. Front. Genet. 2020; 11: 880.

Muñoz F, Sánchez L. breedR: Statistical Methods for Forest Genetic Resources Analysts. R package version 0.12-5. 2020; https://github.com/famuvie/breedR

Pégard M, Rogier O, Bérard A, Faivre-Rampant P, Paslier MCL, et al. Sequence imputation from low density single nucleotide polymorphism panel in a black poplar breeding population. BMC Genomics. 2019; 20: 302.

Pégard M, Segura V, Muñoz F, Bastien C, Jorge V, Sanchez L. Favorable conditions for genomic evaluation to outperform classical pedigree evaluation highlighted by a proof-of-concept study in poplar. Frontiers in plant science. 2020; 11.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2020; URL https://www.R-project.org/.

Rutkoski J, Singh RP, Huerta-Espino J, Bhavani S, Poland J, Jannink JL, Sorrells ME. Efficient use of historical data for genomic selection: a case study of stem rust resistance in wheat. The Plant Genome. 2015; 8: plantgenome2014-09.

Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC GENOMICS. 2014; 15.

Sánchez L, Bijma P, Woolliams JA. Minimizing inbreeding by managing genetic contributions across generations. Genetics. 2003; 164.

Sánchez L, Caballero A, Santiago E. Palliating the impact of fixation of a major gene on the genetic variation of artificially selected polygenes. Genetical Research. 2006; 88.

Santiago E, Caballero A. Effective size of populations under selection. Genetics. 1995; 139: 1013-1030.

Sonesson AK, Meuwissen THE. Mating schemes for optimum contribution selection with constrained rates of inbreeding. Genetics Selection Evolution. 2000; 32: 231-248.

Toro MA, Varona L. A note on mate allocation for dominance handling in genomic selection. Genetics Selection Evolution. 2010; 42: 33.

VanRaden PM. Efficient methods to compute genomic predictions. Journal of dairy science. 2008; 91: 4414-4423.

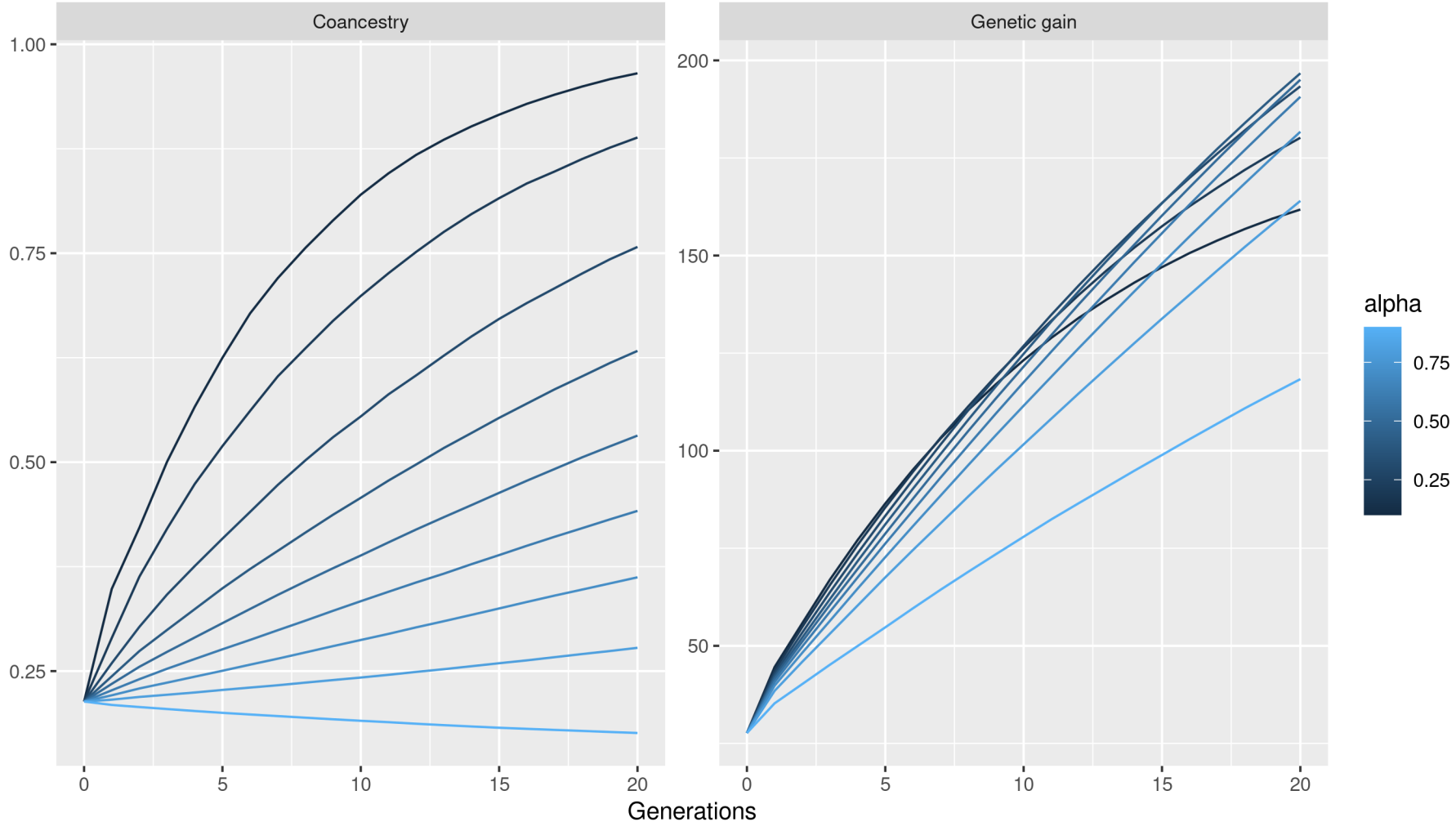
Varona L, Misztal I. Prediction of parental dominance combinations for planned matings, methodology, and simulation results. Journal of dairy science. 1999; 82: 2186-2191.

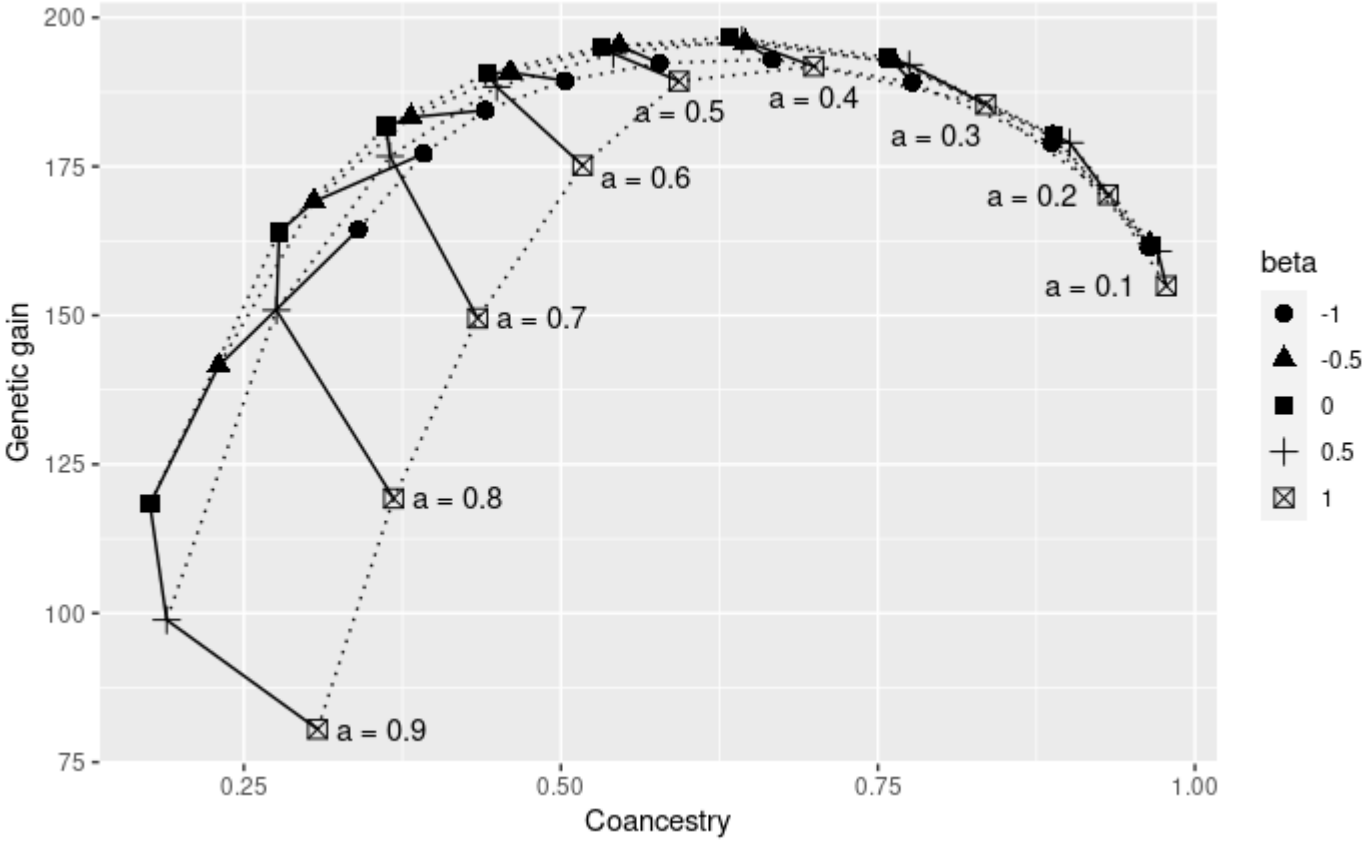
Vitezica ZG, Varona L, Legarra A. On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics. 2013; 195: 1223-1230.

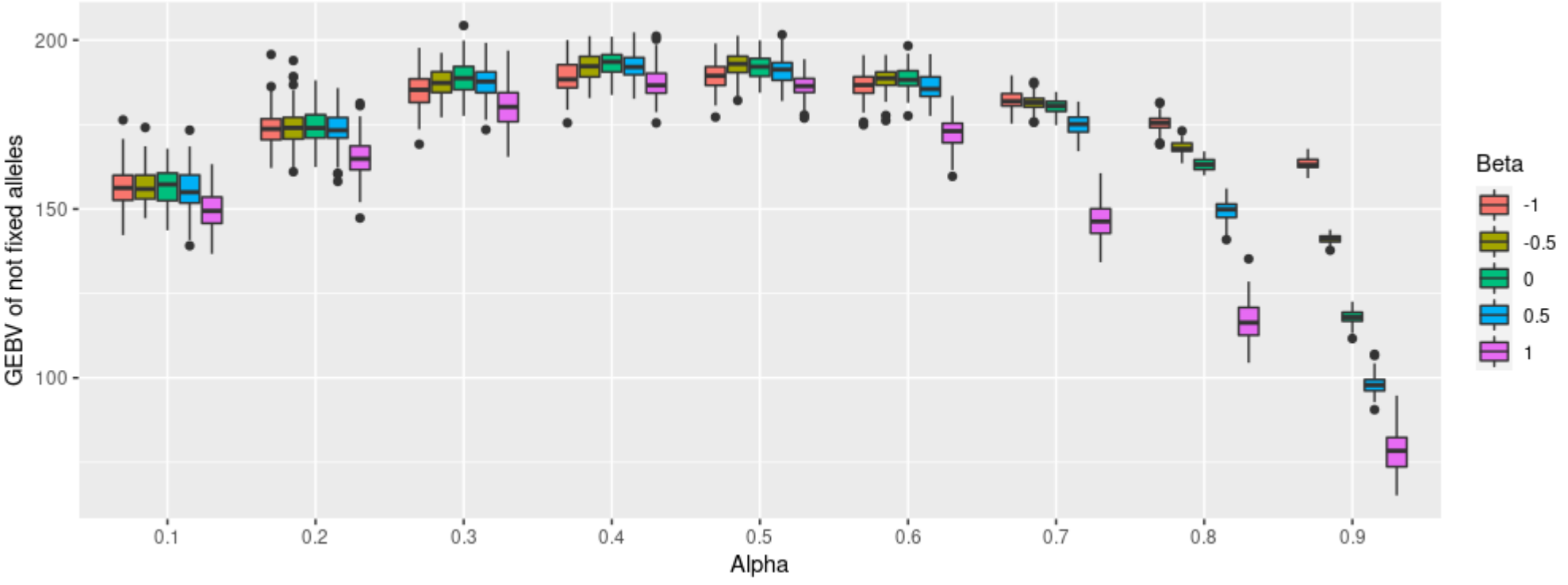
Woolliams JA, Mäntysaari EA. Genetic contributions of Finnish Ayrshire bulls over four generations. Animal Science. 1995; 61: 177–187.

Woolliams JA, Pong-Wong R, Villanueva, B. Strategic optimisation of short-and long-term gain and inbreeding in MAS and non-MAS schemes. In Proceedings of the 7th World Congress on Genetics Applied to Livestock Production (Vol. 33, pp. 155-162). 2002; Montpellier.

Yoshida GM, Yáñez JM, de Queiroz SA, Carvalheiro R. Mate selection provides similar genetic progress and average inbreeding than optimum contribution selection in the long-term. Aquaculture. 2020; 526: 735376.


Coancestry
Genetic gain
1.00
0.75
0.50
0.25
200
150
100
50
0
5
10
15
20
Generations
alpha
0.75
0.50
0.25


Genetic gain
Coancestry
200
175
150
125
100
75
0.25
0.50
0.75
1.00
a = 0.1
a = 0.2
a = 0.3
a = 0.4
a = 0.5
a = 0.6
a = 0.7
a = 0.8
a = 0.9
beta
-1
-0.5
0
0.5
1


GEBV of not fixed alleles
200
150
100
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Alpha
Beta
-1
-0.5
0
0.5
1