bioRxiv preprint doi: https://doi.org/10.64898/2026.03.15.711440; this version posted March 18, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Uncertainty-aware breeding decisions: MCMC-based optimum contribution selection increases breeding decision robustness

Jon Ahlinder[1,*] and Patrik Waldmann[2]

[1]Department of Tree Breeding, The Forest Research Institute of Sweden (Skogforsk), Tomterna 1, Sävar, SE-91833, Sweden
[2]Mathematical Sciences, University of Oulu, Paavo Havaksen tie 3, Oulu, FI-90570, Finland

*Corresponding author: jon.ahlinder@skogforsk.se.

Abstract

Current optimum contribution selection (OCS) implementations use point estimates of estimated breeding values (EBVs), potentially leading to suboptimal selections when individuals have uncertain genetic evaluations. We developed a framework assessing how EBV uncertainty affects OCS decisions through MCMC-based approaches using the COSMO optimizer in Julia, evaluated on Norway spruce (Picea abies, n=5,525) and Loblolly pine (Pinus taeda, n=926) populations. Agreement between point estimate (MAP-OCS) and MCMC-OCS was surprisingly low: mean overlap of only 26.6 (4.8) individuals in Norway spruce genotyped subpopulation and 14.1 (3.6) in full pedigree, with Loblolly pine intermediate at 16.0 (9.6). Despite this low individual-level agreement, selection frequency across MCMC iterations corresponded well with EBV rankings (Spearman $\rho = 0.782$ for Norway spruce), confirming that higher-EBV individuals were preferentially selected under posterior uncertainty. To comprehensively quantify uncertainty impacts, we employed two complementary metrics: individual robustness scores measuring genetic gain stability upon candidate removal, and population-level contribution distribution metrics capturing concentration of genetic gain across selected individuals. Applying these metrics identified 25 high-risk individuals in Norway spruce and nine in Loblolly pine, and constrained exclusion of these individuals improved individual robustness by 16.5% in Loblolly pine (3.00% genetic gain loss) and 29.8% in Norway spruce (2.14% genetic gain loss). Our uncertainty-aware OCS framework successfully identifies unstable selections that may compromise long-term genetic gain, and we recommend assessing EBV uncertainty through posterior distributions and evaluating population-specific trade-offs when implementing uncertainty-aware selection strategies.

Keywords: Estimated breeding values; Posterior uncertainty; Genetic evaluation; Selection optimization; Norway spruce; Loblolly pine

Introduction

Maximizing genetic merit while restricting relatedness in a breeding population is a major challenge in many breeding programs (Robertson 1960; Lynch and Walsh 1998; Goddard 1994). If relatedness is not controlled, elevated levels of inbreeding will cause fitness depression and reduce the performance of the offspring in the population (Lynch and Walsh 1998; Charlesworth and Willis 2009). Typically, the genetic load will increase, that is, the presence of deleterious alleles in the population, leading to a higher incidence of genetic disorders (Derks and Steensma 2021; Tao *et al.* 2025). Therefore, much effort has been devoted to developing methods that optimize the balance between gain and relatedness of the selected set of individuals and their contribution to the next generation of breeding or deployment population (Goddard 1994; Grundy *et al.* 1998; Mullin and Belotti 2015; Hamilton 2020; Cowling *et al.* 2023; Tchounke *et al.* 2023).

One of the most popular approaches, the optimum contribution selection (OCS), optimizes a linear objective function with quadratic constraints on relatedness (inbreeding) (Meuwissen 1997; Grundy *et al.* 1998). To solve the optimization problem, the original approach uses Lagrangian multipliers in an iterative fashion where non-selected individuals are excluded one by one. Since then, many efforts have been made to further improve the original approach by introducing enhanced optimization techniques such as semidefinate programming (SDP) (Pong-Wong and Woolliams 2007), the branch-and-bound algorithm (Mullin and Belotti 2015), second-order cone programming (SOCP) (Yamashita *et al.* 2018), matrix algebra computational tricks (Hinrichs *et al.* 2006), comparing the effect of kinship matrices on genetic gain, inbreeding and diversity (Gautason *et al.* 2023), including mate allocation optimization (Waldmann 2025) and various alternative formulations of the optimization problem to fit a range of different applications (Grundy *et al.* 1998; Hallander and Waldmann 2009; Clark *et al.* 2013; Hamilton 2020; Hjortø *et al.* 2022; Fogg *et al.* 2024).

Recent advances in mathematical optimization theory and machine learning have paved the way for new tools that can handle large and complex optimization problems. The most common approach taken by modern convex solvers is the interior point method which is capable of solving both linear programs (LPs) and quadratic programs (QPs) in polynomial time (Nocedal and Wright 2006). The primal-dual interior point methods apply variants of Newton's second-order method to iteratively find a solution to a set of optimality conditions. At each iteration, the algorithm alternates between a Newton step that involves factoring a Jacobian matrix and a line search to determine the step size that ensures a feasible iterate. However, interior point methods typically do not scale well for very large problems since

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.15.711440; this version posted March 18, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

the Jacobian matrix has to be calculated and factored at each step. Another recent direction is first-order gradient-based optimization methods that provide faster computation and scale to much larger problems than second-order methods (Beck 2017). A popular first-order method is the Alternating Direction Method of Multipliers (ADMM) (Boyd *et al.* 2010) which is an optimization algorithm particularly well suited for large-scale optimization problems as it combines principles from dual decomposition and augmented Lagrangian methods (Lions and Mercier 1979). In short, ADMM decomposes a complex problem into smaller sub-problems that are easier to solve on their own. This makes ADMM effective for large-scale optimization tasks and suitable for applications of national breeding programs where pedigrees can consist of millions of individuals, such as the Breedbase system in plant breeding (Hershberger *et al.* 2021), and TreePlan in forest tree breeding (McRae *et al.* 2003).

The efficiency of OCS relies heavily on the accuracy of the estimated breeding values (EBVs), which are used to calculate the contributions of breeding candidates. Typically, the EBVs of the candidates are obtained as point estimates that are plugged into the OCS. As a consequence, the uncertainty of the EBVs is not taken into account, and few studies have investigated this aspect. Fogg *et al.* (2024) introduced the robust OCS method using concepts from robust optimization. Their idea is to use uncertainties of EBVs in the form of standard errors (SE) that can be obtained, for example, from frequentist inference methods or as summary statistics from Bayesian methods. They showed that individuals with high EBVs and SEs were penalized, and instead the robust OCS preferred individuals with low SEs even if they had high EBVs. Another probabilistic approach was proposed by Pocrnic *et al.* (2023) who analyzed a small simulated breeding population using posterior distributions of EBVs obtained from Bayesian MCMC analysis and then applied OCS on the posterior distribution of EBVs. Their results showed a considerable underestimate of the optimized contributions for standard OCS compared to their probabilistic OCS, which were closer to the true OCS. This approach is potentially interesting, but needs to be tested on real-world breeding pedigrees to enhance the understanding of probabilistic based OCS compared to point-estimate OCS (from here on MAP-OCS).

In Bayesian statistics, the highest probability model (HPM) is the model with the highest posterior probability, which means that it receives the most support from the observed data. However, in large model spaces, the HPM can be difficult to estimate accurately due to the large number of competing models. In addition, there might be competing models with very similar support, which means that ignoring the uncertainty of the model can result in overconfident or biased results. As an alternative to HPM, the median probability model (MPM) includes all variables that have a posterior inclusion probability of at least 0.5 (Barbieri and Berger 2004). MPM is considered optimal for prediction under a squared error loss function and therefore provides better predictive performance compared to HPM (Ghosh 2015). In the OCS framework, the MPM would be estimated as the median from the distribution of optimized contributions across all MCMC iterations. In summary, ignoring the impact of uncertainty of EBVs can lead to suboptimal selection decisions and potentially reduce genetic gain and increase inbreeding.

Despite the fundamental importance of breeding value uncertainty in genetic selection, current OCS frameworks treat all EBV point estimates equally regardless of their reliability — the squared correlation between true and estimated breeding values (Henderson 1975). This creates a practical problem: individuals with high but unreliable EBV estimates may be selected over candidates with moderately lower but more reliable estimates, potentially compromising long-term genetic progress (Bijma 2012). While Bayesian genetic evaluation naturally produces posterior distributions that reflect this uncertainty through their variance structure (Sorensen and Gianola 2002), these distributions are typically discarded when point estimates are extracted for use in OCS. A systematic framework that assesses EBV uncertainty through posterior distributions, identifies high-risk individuals, and replaces them with more reliable alternatives while maintaining OCS theoretical foundations represents an important unmet need in breeding and conservation programs.

The aim of our study is to first exploring the impact of uncertainty of the EBVs on the computed contributions by Monte Carlo sampling of the posterior distributions of EBVs. Then, we develop MCMC-based robustness scores metric that quantify the effects of the uncertainty of breeding value on individual importance, implementing these within a constrained exclusion framework to systematically replace high-risk selections and improve the stability of the OCS decision. The performance of the method is illustrated on two forest tree breeding pedigrees of Loblolly pine and Norway spruce and compared to traditional MAP-OCS.

Materials and methods

Optimum contribution

Let us assume that we have a recruitment or candidate population of n individuals and that genetic relationships between individuals are inferred from pedigree information or marker data (or both). For optimum contribution selection, the optimization problem can be formulated as (Pong-Wong and Woolliams 2007; Ahlinder *et al.* 2013; Mullin and Belotti 2015; Yamashita *et al.* 2018):

$$\begin{aligned}
\text{maximize} \quad & \mathbf{c}^T\mathbf{g} && (1)\\
\text{subject to} \quad & \mathbf{c}^T\mathbf{A}\mathbf{c}/2 \leq \Theta && (2)\\
& \mathbf{c}^T\mathbf{s} = 0.5 && (3)\\
& \mathbf{c}^T\mathbf{d} = 0.5 && (4)\\
& \mathbf{c}^T\mathbf{1} = 1 && (5)\\
& \mathbf{c} \leq \mathbf{m} && (6)\\
& \mathbf{c} \geq \mathbf{0} && (7)
\end{aligned}$$

where $\mathbf{c} \in R^n$ is a $1 \times n$ vector of individual contributions (i.e. the decision variable), $\mathbf{A} \in R^{n \times n}$ contains pair-wise relationships of the available candidates n, $\mathbf{m}$ is an upper bound on the contribution of any particular individual i within the cohort due to practical limitations, while the lowest possible contribution is 0. Note that the constraint on the gender contribution, expressed here as the vectors $\mathbf{s}$ and $\mathbf{d}$ in constraints (3) and (4), can be removed if the species of interest is monoicous (as long as the total contribution sums up to one). On the other hand, if the species is dioecious, the constraint (5) can be removed. Here, the object function is linear, with a quadratic constraint on average relatedness among the selected candidates. Note that the term $\mathbf{c}^T\mathbf{g}$ can be minimized if the negative expression is used instead (i.e. $-\mathbf{c}^T\mathbf{g}$), which formulation is used as a starting point for several software implementations (Ahlinder *et al.* 2013; Mullin and Belotti 2015; Yamashita *et al.* 2018).

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.15.711440; this version posted March 18, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Here, we used the COSMO solver available in the JumP Julia package. This solver handles convex optimization problems with quadratic objective functions and conic constraints using an operator splitting algorithm based on ADMM (Garstka *et al.* 2019). Please consult Supplementary information for further details on the tested optimization methods and the computational performance on the analyzed pedigrees.

Gibbs sampling of genetic mixed effect model analysis and uncertainty analysis

Following (Sorensen and Gianola 2002), we used the following multi-trait mixed effect model to infer breeding values:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e}, \tag{8}$$

where $\mathbf{y}$ is the vector containing each of m traits (i.e. responses) represented sequentially for n individuals in a single column vector of size $nm \times 1$, $\mathbf{X}$ and $\mathbf{Z}$ are the fixed and random effect design matrices of sizes $nm \times pm$ and $nm \times nm$, respectively associating predictors with the corresponding phenotypic records for p levels of fixed effect predictors. Random polygenic and residual effects are assumed to follow for two traits, one categorical and one continuous.

$$\begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} \mid \mathbf{G}, \mathbf{G}_0, \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \mathbf{G}_0 \times \mathbf{G} \right) \tag{9}$$

with $\mathbf{g}_1$ and $\mathbf{g}_2$ are breeding values or polygenic score vectors of continuous (subscript 1) and categorical traits (subscript 2) respectively, $\mathbf{G}$ is the genetic relationship matrix of size $n \times n$, $\mathbf{G}_0$ is the genetic covariance matrix between categorical and continuous traits.

The joint covariance model is defined as $\text{Cov}(\mathbf{y}) = \mathbf{Z}\mathbf{G}_0\mathbf{Z}^T + \mathbf{R}$, with random effects covariances defined as:

$$\mathbf{G}_0 = \begin{bmatrix} \sigma^2_{g_1} & \sigma_{g_1,g_2} \\ \sigma_{g_1,g_2} & \sigma^2_{g_2} \end{bmatrix} \tag{10}$$

where $\sigma^2_{g_1}$ and $\sigma^2_{g_2}$ are the additive genetic variances for traits 1 and 2, and σ_{g_1,g_2} is the additive genetic covariance between the traits. The multivariate responses are here a mixture of continuous and categorical traits, and the mixed response model in Equation 9 is used. For the categorical trait (trait 2), a threshold model is assumed on the underlying liability scale (Sorensen and Gianola 2002), where the observed categorical response is determined by an underlying continuous liability exceeding threshold values. The genetic covariance $\sigma^2_{g_1,g_2}$ and correlation $\rho_{g_1,g_2} = \sigma^2_{g_1,g_2} / \sqrt{(\sigma^2_{g_1} + \sigma^2_{g_2})}$ are defined on this liability scale. The residual covariance matrix $\mathbf{R}_0$ is defined similarly to $\mathbf{G}_0$ but with error (co)variances instead. After inferring breeding values $\mathbf{g}_1$ and $\mathbf{g}_2$, a joint selection index matrix $\mathbf{I}$, of size $n \times l$ was defined for all members of the n pedigree with l MCMC iterations by weighting the breeding values of the traits. For simplicity, we gave all included traits equal weights: therefore, no economic information was incorporated into the selection trait $\mathbf{I}$. Then, to accommodate uncertainty, we iterated through the l MCMC iterations and called the optimum contribution algorithm:

$$\mathbf{c}_j = f(\mathbf{I}_j = (\mathbf{g}_{1,j} + \mathbf{g}_{2,j})/2, \mathbf{G}), \quad j = 1, \ldots, l, \tag{11}$$

where $f(\cdot)$ is the OCS function call (performed in Julia) and $\mathbf{g}_1$ and $\mathbf{g}_2$ are the EBVs of the selection candidates for the continuous trait 1 and categorical trait 2 in any MCMC iteration j. Using this procedure, we acquire a matrix of contributions $\mathbf{C} = \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_l \end{bmatrix}$. We used $l = 1000$ for both case studies. The multi-trait mixed effect model implementation used here is freely available in the JWAS package (Chen *et al.* 2018) in Julia using default settings, bar the total number of iterations set to 1000.

Uncertainty Impact Analysis: MCMC-OCS versus MAP-OCS

To quantify the impact of breeding value uncertainty on optimum contribution selection decisions, we compared contributions obtained from MAP breeding values (MAP-OCS) with contributions derived from each MCMC iteration (MCMC-OCS). For each dataset, we performed OCS using both approaches under identical genetic constraints ($\Theta = 0.02$ for Norway spruce, $\Theta = 0.03$ for Loblolly pine) selecting the top 100 contributors.

MAP-OCS Reference Solution We first calculated the reference optimum contribution vector $\mathbf{c}_{MAP}$ using maximum a posteriori (MAP) breeding values:

$$\mathbf{c}_{MAP} = \arg\max_{\mathbf{c}} \left\{ \mathbf{g}_{MAP}^T \mathbf{c} \right\} \tag{12}$$

subject to:

$$\sum_{i=1}^{n} c_i = 1 \tag{13}$$

$$\frac{1}{2}\mathbf{c}^T\mathbf{G}\mathbf{c} \leq \Theta \tag{14}$$

$$c_i \geq 0 \quad \forall i, \tag{15}$$

where $\mathbf{g}_{MAP}$ is the vector of MAP breeding values, $\mathbf{G}$ is the genomic (or pedigree) relationship matrix, n is the number of selection candidates, and Θ is the genetic constraint parameter.

MCMC-OCS Iterative Analysis For each MCMC iteration j (where $j = 1, \ldots, N_{MCMC}$), we extracted the posterior sample of breeding values $\mathbf{g}_j$ and solved the identical OCS problem:

$$\mathbf{c}_j = \arg\max_{\mathbf{c}} \left\{ \mathbf{g}_j^T \mathbf{c} \right\}, \tag{16}$$

subject to the same constraints as MAP-OCS. This process yielded N_{MCMC} contribution vectors (989-1000 depending on the data set after filtering for convergence), representing the full posterior distribution of optimal contributions under breeding value uncertainty. To quantify the agreement between MAP-OCS and MCMC-OCS, multiple complementary metrics were used to capture different aspects of similarities. For each MCMC iteration j, we identified the top 100 contributors based on contribution values and calculated the number of individuals selected in both MAP-OCS and MCMC-OCS:

$$\text{Overlap}_j = |S_{MAP} \cap S_j|, \tag{17}$$

where S_{MAP} denotes the set of top 100 MAP-OCS contributors and S_j denotes the top 100 contributors from MCMC iteration j. To normalize for total unique selections, we calculated Jaccard similarity as:

$$J_j = \frac{|S_{MAP} \cap S_j|}{|S_{MAP} \cup S_j|}. \tag{18}$$

Furthermore, we computed Spearman rank correlation (ρ) between $\mathbf{c}_{MAP}$ and $\mathbf{c}_j$ across all individuals to assess ranking consistency beyond just the top 100 selections.

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.15.711440; this version posted March 18, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

EBV-Selection Frequency Analysis To validate that selection patterns remained logical despite high uncertainty (i.e., low degree of selection overlap at each MCMC iteration), we calculated selection frequency for each individual across all MCMC iterations:

$$f_i = \frac{1}{N_{MCMC}} \sum_{j=1}^{N_{MCMC}} \mathbb{1}(i \in S_j) \tag{19}$$

where $\mathbb{1}(i \in S_j)$ is an indicator function equal to 1 if individual i is in the top 100 contributors for iteration j. We then calculated Spearman correlations between selection frequencies $\mathbf{f}$ and MAP breeding values $\mathbf{g}_{MAP}$ to assess whether higher MAP-EBV individuals were preferentially selected despite the uncertainty of the contribution.

The relationship between MAP EBV and MCMC-OCS selection frequency was analyzed using Gaussian process (GP) regression (Rasmussen and Williams 2006) with a Matérn 3/2 covariance function. All hyperparameters (kernel lengthscale, signal variance, and observation noise variance) were estimated via maximum likelihood (ML) using the Nelder-Mead optimization algorithm. The GP model was fitted using AbstractGPs.jl (van der Wilk *et al.* 2020) and Optim.jl (Mogensen and Riseth 2018) in Julia 1.11.

Individual Importance: MCMC-Based Robustness Scores

To develop a framework for incorporating MCMC-based uncertainties in OCS. we need metrics for a direct comparison between uncertainty-naive MAP-OCS and uncertainty-aware MCMC-OCS selection strategies while maintaining the theoretical foundations underpinning OCS. The robustness score quantifies an individual's importance to the breeding program by measuring the impact on expected genetic gain when that individual is removed from the breeding population. For each individual i and MCMC iteration j:

1. **Baseline genetic gain:** Calculate expected genetic gain using the complete breeding population:

$$G_{\text{baseline},j} = \sum_{k=1}^{n} BV_{k,j} \times c_{k,j} \tag{20}$$

where $BV_{k,j}$ is the standardized breeding value of individual k in MCMC iteration j, and c_k is the contribution of individual k. We used standardized breeding values within each MCMC iteration before running OCS to avoid weighting differences during the MCMC procedure.

2. **Modified genetic gain:** Set individual i's breeding value to a value below the minimum (effectively removing them), re-run OCS, and calculate the new expected genetic gain:

$$G_{\text{modified},j} = \sum_{k=1}^{n} BV_{k,j} \times c'_k \tag{21}$$

where c'_k represents the new optimal contributions excluding individual i.

3. **Individual impact:** Calculate the genetic gain loss:

$$\Delta G_{i,j} = G_{\text{baseline},j} - G_{\text{modified},j} \tag{22}$$

4. **Mean robustness score:** Average across all MCMC iterations:

$$S_i = \frac{1}{N} \sum_{j=1}^{N} \Delta G_{i,j} \tag{23}$$

where N is the number of MCMC iterations.

We also calculated a stable robustness metric, which expresses robustness as percentage of baseline genetic gain:

$$P_i = \frac{S_i}{\bar{G}} \times 100 \tag{24}$$

where the standardized gain is $\bar{G} = \frac{1}{N} \sum_{j=1}^{N} G_{\text{baseline},j}$.

Higher robustness scores indicate individuals whose removal causes substantial loss of genetic gain, suggesting that they are crucial to the breeding program, while lower robustness scores indicate individuals whose contribution is less critical or easily replaceable. The robustness score thus provides an uncertainty-aware measure of individual importance that accounts for the variability in breeding value estimates across the posterior distribution.

To identify potential replacement candidates, we expanded the MCMC robustness analysis beyond the individuals initially selected. The analysis included the top 100 selected individuals (ranked by MAP contribution), and the top 100 unselected individuals (ranked by posterior mean breeding values), forming an expanded analysis set. For each individual in this expanded set, we calculated robustness scores by systematically removing them from the breeding population through multiple MCMC iterations and measuring the resulting impact on the expected genetic gain.

Selected individuals from the MAP-based OCS solution were classified into risk categories based on their robustness scores derived from MCMC-OCS analysis. We applied a dynamic threshold approach to classify individuals into risk categories based on their robustness score relative to the distribution: individuals in the bottom 25% were assigned as high-risk out of the available pool of individuals with calculated robust scores. Although this classification is arbitrary, we are interested in showing the principle rather than giving an absolute recommendation, as this limit varies with population history and breeding program specifics. To further speed up the computations, 100 (out of 1000) MCMC samples were randomly selected for robustness score calculations.

This individual importance metric is complemented by portfolio-level vulnerability assessment (Section 2.6) to provide a comprehensive evaluation of breeding program robustness.

Contribution Distribution Assessment

To complement individual robustness scores with population-level risk metrics, we quantified the distribution of genetic contributions across selected individuals. These metrics—conceptually similar to portfolio diversification measures in finance—assess whether expected genetic gain is concentrated in few individuals (vulnerable) or distributed evenly (robust). For each OCS solution (MAP-OCS and constrained), we calculated the contribution of individual i as:

$$C_i = c_i \times BV_i \tag{25}$$

where c_i is the optimal contribution and BV_i is the standardized breeding value of individual i. This represents the expected genetic gain attributable to each selected individual.

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.15.711440; this version posted March 18, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

We assessed contribution distribution structure using four complementary metrics:

1. **Mean contribution:** Average contribution across all selected individuals with non-zero contributions:

$$\bar{C} = \frac{1}{n_{\text{sel}}} \sum_{i \in S} C_i \tag{26}$$

where S is the set of selected individuals ($c_i > 10^{-4}$) and $n_{\text{sel}} = |S|$. Lower mean contributions indicate more evenly distributed breeding populations.

2. **Maximum contribution:** Largest single-individual dependency:

$$C_{\max} = \max_{i \in S} C_i \tag{27}$$

This quantifies vulnerability to the loss of the most critical individual. High maximum contributions indicate concentrated genetic gain dependence.

3. **Concentration ratio:** Proportion of total genetic gain from the top 10% of contributors:

$$\text{Conc}_{10} = \frac{\sum_{i \in T_{10}} C_i}{\sum_{i \in S} C_i} \times 100 \tag{28}$$

where T_{10} represents the top 10% of individuals ranked by contribution.

4. **Gini coefficient:** Inequality measure of contribution distribution. For n selected individuals with contributions sorted in ascending order ($C_1 \leq C_2 \leq \cdots \leq C_n$):

$$\text{Gini} = \frac{\sum_{i=1}^{n} (2i - n - 1) C_i}{n \sum_{i=1}^{n} C_i} \tag{29}$$

where Gini $\in [0, 1]$. Values near 0 indicate equal distribution (low concentration); values near 1 indicate high inequality (high concentration).

These contribution distribution metrics quantify whether genetic gain is concentrated in few individuals (vulnerable) or distributed evenly (robust). Lower values for mean contribution, maximum contribution, and concentration ratio indicate more resilient breeding population structures that are less dependent on any single individual or small subset of candidates.

Together with individual importance scores (Section 2.5), these dual metrics provide a comprehensive framework for evaluating breeding program robustness under uncertainty.

Norway spruce case study

Two field trials of Norway spruce, with Skogforsk IDs S23F8820476 located outside Vindeln, Västerbotten County (easting: 19.68 and northing: 64.3) and S23F8820477, Hädanberg, Västernorrland County, where utilized for showcasing the COSMO OC algorithm (Table 1). In both trials, approximately 6000 offspring from 150 families were planted in 1988 in single tree plot design for evaluating the performance of parent trees in the Hissjö seed orchard. The following phenotypic traits were scored: total tree height at year 7 and 17, early height growth at year 17, number of spike knots at year 17 and tree vitality at year 7. In total, 1218 offspring (568 and 650 in S23F8820476 and S23F8820477 respectively) where genotyped using exome capture technique resulting in 300 714 SNPs prior to filtration. For further details of the Norway spruce data, please consult Chen *et al.* (2018). The G matrix was calculated using the rrBLUP r package (Endelman 2011), with default settings. To combine genotyped and ungenotyped trees, a joint **H** matrix was calculated using the AGHmatrix r package (Amadeu *et al.* 2023), where the **G** was first tuned towards **A** following Christensen *et al.* (2012). The blending was performed using the Hmatrix function, with $\tau = 1$ and $\omega = 1$ (default) (Legarra *et al.* 2009). To ensure that **H** used in subsequent analyses was positive definite, we applied the cor_nearPD function from the Bigsimr.jl package (Qi and Sun 2006). Multi-trait analysis to obtain posterior estimates of breeding values (BV) were performed using the JWAS Julia package (Cheng *et al.* 2018) with $n = 1000$ MCMC iterations and otherwise default settings. The predictors used were a site factor and a nested plot within site term.

Loblolly pine case study

The data sets included highlight different organisms, pedigree sizes, and average relatedness (Table 1). The first analyzed pedigree was published as a benchmark data set to develop genomic selection methods and consists of a clonally replicated test of Loblolly pine in the southeastern United States (Resende *et al.* 2012). The population was derived by controlled crossing among 32 selected parents from the Lower Gulf Elite Population, consisting of 22 field selected F0 plus trees and 10 selected F1 progeny. The parents were crossed in a partial diallel mating design, and the progeny propagated for field testing as rooted cuttings. Here, we analyzed both F1 and F2 populations totalling 926 individuals using the EBVs of the 6-year tree height (HT6), Gall volume (GV6) and diameter at breast height (DBH6) and 4-year wood quality (WDN4) in a selection index of equal weight (negative for GV6) following Ahlinder *et al.* (2024). The 926 F1 and F2 members were distributed in 71 families in total with a average family size of 13.04 individuals. The candidate cohort had a average coancestry of $\Theta = 0.043$. Contributions of the selected set of ramets were constrained to a group coancestry of $\Theta \leq 0.03$. No further limitations were placed on the contributions at any given genotype. We calculated the **G** matrix using the R package rrBLUP with default settings, as relationship matrix in the mixed effect model analysis. Further analysis details are provided at https://github.com/jonhar97/MCMC_uncertainty_OCS.

Results

Genetic parameter estimates

Heritability estimates varied across relationship matrices and traits in both species (Table 2). For Norway spruce, the H matrix generally provided the highest estimates for growth traits, while the G matrix showed the most constrained estimates. For Loblolly pine, the A matrix provided higher estimates than the G matrix across all four traits. Genetic correlations were broadly consistent with expectations for multi-trait breeding indices, with strong positive correlations among growth traits and weak to moderate negative correlations between growth and wood quality traits (Table 3, Supplementary Fig. 1-2). Notably, the H matrix showed stronger genetic correlations among Norway spruce traits than either the G or A matrices. Posterior distributions of EBVs revealed substantial ranking uncertainty within families in both case studies (Fig. 1, Supplementary Fig. 3). In the high-performing Norway spruce family, ranking standard deviation reached 7.4, with the top-ranked individual having

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.15.711440; this version posted March 18, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Table 1 Description of included datasets. Average coancestry (Avg. coanc.) of the candidate population is calculated via the pedigree.

Study	Species	Pedigree	Candidates	Gen.	Parents	Families	Avg. coanc.	Propagation
Resende *et al.* (2012)	*P. taeda*	973	926	3	32	61	0.043	Clones
Chen *et al.* (2018)	*P. abies*	5,597	5,525	2	72	146	0.016	Seedlings
Chen *et al.* (2018)	*P. abies*	1,273	1,218	2	55	132	0.014	Seedlings

only 48% probability of ranking in the top 10% of its family. Similarly, in Loblolly pine, the top-ranked individual in the highest-performing family had only 52% probability of ranking first. These overlapping posterior distributions illustrate the core challenge motivating our uncertainty-aware framework.

Impact of EBV uncertainties on contributions

To investigate the impact of EBV uncertainty on selection decisions via optimum contribution selection (OCS), we performed OCS analysis for each MCMC iteration, thereby exploring the full posterior distribution of individual EBVs that is typically ignored in routine analyses. As a reference, we used maximum a posteriori (MAP) EBV values without uncertainty in the OCS analysis (MAP-OCS). The MAP-OCS results showed satisfactory genetic gains for the calculated EBV index across all datasets analyzed.

In the MCMC-OCS analysis, we compared the overlap among the top 100 contributing trees with the corresponding MAP-OCS highest contributors, along with Jaccard similarity between selected sets. For the Norway spruce **G** population (1,218 subset), there was considerable uncertainty in posterior EBV distributions, manifested as low mean overlap of 26.6 individuals (range: 10-46 individuals across MCMC iterations) and low mean Jaccard similarity of 15.4% (Table 4). Nearly all individuals were selected in the top 100 contributors in at least some MCMC iterations, with only one individual selected in more than half of the iterations (Figure 2).

Despite high uncertainty between individual MCMC iteration contributions and MAP contributions, there was a strong correlation between EBV values and selection frequency (Spearman $\rho = 0.782$; Figure 2A and B). Higher EBV individuals were selected more frequently, with a dramatic quartile difference where top 25% EBV individuals were selected 17× more often than the bottom 25% (Figure 2D). GP regression (Matérn 3/2 kernel) with ML hyperparameter estimation yielded a good fit to the selection frequency vs MAP-EBV results ($R^2 = 0.667, \rho_n^2 = 0.012$) while providing uncertainty quantification. Narrower confidence intervals at higher EBV values indicate greater selection stability for elite ranked individuals. OCS with $\Theta = 0.02$ severely constrains coancestry among the selected cohort, requiring representatives from most major families among the top 100 selections. EBV uncertainty causes substantial changes in within-family rankings, explaining much of the observed selection variability.

Using the **A** relationship matrix instead of the **G** matrix for the 1,218 subset reduced the agreement between MCMC-OCS and MAP-OCS across all metrics reported (Table 4), demonstrating the value of genomic information for selection consistency.

To examine sample size effects, we analyzed the complete 5,525-tree Norway spruce progeny test population. The obtained genetic gain of MAP-OCS was 2.38 when selecting the top 100 contributors. However, as the sample size increased, the mean overlap between MCMC-OCS and MAP-OCS decreased to 14.1 individuals (SD = 3.6), with a Jaccard similarity of 0.076, indicating poorer agreement and a greater impact of uncertainty. Using only pedigree information (**A**) further reduced the overlap to 10.4 individuals (SD = 4.0). MAP-EBV vs selection frequency across MCMC iterations resulted in reduced Spearman correlation of $\rho = 0.541$ as compare to the corresponding correlation obtained in the 1218 subpopulation analysis. The quartile analysis showed the expected frequency difference between top 25% and bottom 25% EBV individuals (ratio = 6.4). However, no individual was selected among the top 100 contributors in more than 50% of MCMC iterations, indicating that including more individuals increased difficulty in reliably selecting individuals within families.

For the Loblolly pine analysis, EBV uncertainties resulted in poor agreement between MCMC and MAP-based OCS (Table 4). Unlike Norway spruce, the **A** based MCMC-OCS analysis performed similarly to the **G** based analysis. MCMC-OCS revealed moderate uncertainty in posterior EBV distributions, where only 17.5 individuals on average received positive contributions per iteration across the MCMC samples. The relationship between MAP EBV and selection frequency showed a moderate positive correlation (Spearman $\rho = 0.456$; Supplementary Fig. 4A). Selection frequency exhibited high sparsity: the majority of individuals (approximately 70%) were never selected across any iteration, while a small subset of high-EBV individuals showed consistent selection (Supplementary Fig. 4B,4D). Quartile analysis demonstrated clear EBV-driven selection, with the top 25% EBV individuals selected 34.7× more frequently than the bottom 25% (mean frequencies: 0.048 vs 0.001; Supplementary Fig. 4C). However, even among the top 100 MAP EBV individuals, none were selected in more than half of the MCMC iterations, reflecting substantial within-family ranking uncertainty when EBV posterior distributions are incorporated.

In summary of all analyzes, there was substantial uncertainty between individual MCMC contributions and MAP contributions, manifested as low Jaccard similarity, few overlapping individuals, and low contribution-level correlations. However EBV-selection frequency correlations showed much better consistency, indicating that while we correctly identified which individuals should be selected (high EBV-frequency correlation), precise contribution allocation within iterations remained problematic (low overlap correlation).

Robustness analysis of contributions

We quantified breeding program robustness using two complementary approaches: individual importance scores (MCMC-based robustness analysis) and contribution distribution metrics (genetic gain allocation patterns across selected individuals). The goal is to identify key members of the breeding population that contribute greatly to the gain, but also to identify individuals that easily can be swapped with alternatives that might be more accessible and provide much more stable long-term contributors.

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.15.711440; this version posted March 18, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

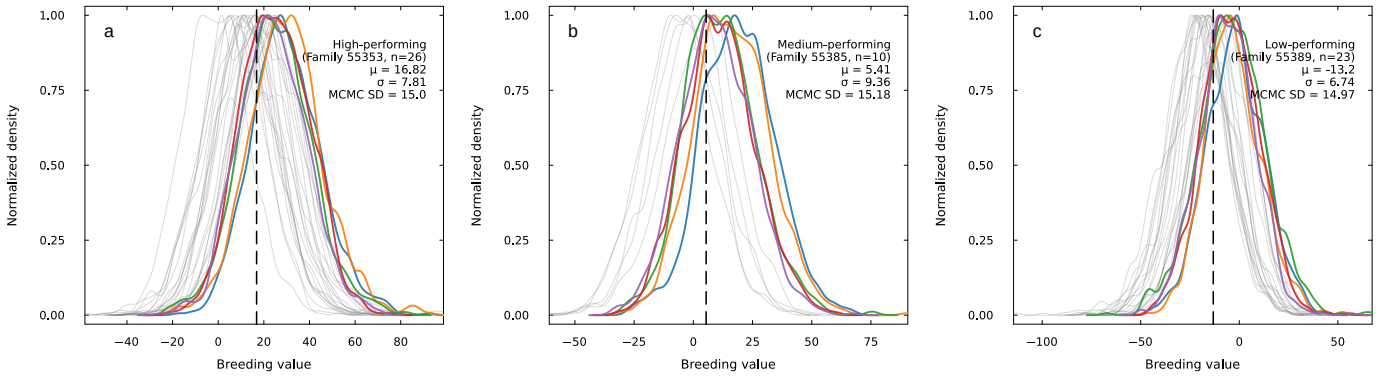


Figure 1 Posterior uncertainty in estimated breeding values reveals ranking ambiguity within Norway spruce families. Kernel density plots show the posterior distributions of selection index breeding values (Hjd17 + Htv17 - Sprant17) for individuals within three representative families: **a** high-performing (Family 55353, $n = 26$), **b** medium-performing (Family 55385, $n = 10$), and **c** low-performing (Family 55389, $n = 23$). Each thin curve represents the posterior distribution for one individual based on 1000 MCMC samples. Thick curves highlight the top 5 individuals within each family ranked by posterior mean. Dashed vertical lines indicate family means.

Table 2 Heritability estimates for Norway spruce and *Pinus taeda* traits across relationship matrices

Species	Trait	G Matrix	H Matrix	A Matrix
Norway spruce	Height growth, years 15–17 (Htv17)	0.251 ± 0.048	0.442 ± 0.045	0.376 ± 0.069
	Height at 17 years, July (Hjd17)	0.221 ± 0.056	0.442 ± 0.040	0.342 ± 0.049
	Height at 7 years (Hjd7)	0.199 ± 0.081	0.341 ± 0.053	0.262 ± 0.058
	Ramicorns/spike knots (Sprant17)	0.094 ± 0.040	0.108 ± 0.070	0.159 ± 0.091
	Survival/vitality (Lev17)	N/A[a]	0.074 ± 0.023	0.121 ± 0.044
P. taeda	Diameter at breast height, 6 years (DBH6)	0.378 ± 0.041	–	0.533 ± 0.110
	Tree height, 6 years (HT6)	0.284 ± 0.043	–	0.429 ± 0.115
	Wood density, 4 years (WDN4)	0.164 ± 0.038	–	0.181 ± 0.054
	Gall volume, 6 years (GV6)	0.168 ± 0.035	–	0.177 ± 0.060

G matrix: Genomic relationships; **H** matrix: Combined genomic and pedigree; **A** matrix: Pedigree only. Norway spruce: **G** matrix $n = 1218$, **H**/**A** matrix $n = 5525$; *P. taeda*: **G**/**A** matrix $n = 926$.

[a] No variation in survival among genotyped trees.

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.15.711440; this version posted March 18, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Table 3 Genetic correlation estimates between trait categories across species and relationship matrices

Species	Correlation Type	G Matrix	H Matrix	A Matrix
Norway spruce	Growth-Growth[a]	0.43 ± 0.22	0.78 ± 0.10	0.72 ± 0.13
	Growth-Ramicorns[b]	0.08 ± 0.07	−0.27 ± 0.09	−0.06 ± 0.05
	Growth-Survival[c]	–	−0.28 ± 0.08	−0.10 ± 0.06
	Ramicorns-Survival	–	0.85 ± 0.34	0.75 ± 0.42
P. taeda	Growth-Growth[d]	0.73 ± 0.15	–	0.79 ± 0.28
	Growth-Wood quality[e]	−0.21 ± 0.08	–	−0.19 ± 0.08
	Growth-Disease[f]	0.16 ± 0.01	–	0.07 ± 0.03
	Wood quality-Disease	−0.01 ± 0.14	–	−0.05 ± 0.14

[a] Height traits: Hjd7, Hjd17, Htv17 correlations.
[b] Height traits vs Sprant17 correlations.
[c] Height traits vs Lev17 correlations.
[d] HT6-DBH6 correlation.
[e] HT6/DBH6 vs WDN4 correlations.
[f] HT6/DBH6 vs GV6 correlations. –: Not applicable due to trait absence in analysis or no **H**-based analysis was performed.

Table 4 MCMC-OCS uncertainty analysis: summary statistics of primary overlap metrics by species and relationship matrix

Species	Matrix	Pop Size	Θ	Overlap Count	Jaccard Sim	Spearman ρ
Norway Spruce	A	1,218	0.02	24.9 (4.3)	0.143 (0.028)	0.015 (0.314)
Norway Spruce	G	1,218	0.02	26.6 (4.8)	0.154 (0.032)	0.065 (0.058)
Norway Spruce	A	5,525	0.02	10.4 (4.0)	0.056 (0.023)	0.073 (0.275)
Norway Spruce	H	5,525	0.02	14.1 (3.6)	0.076 (0.021)	−0.017 (0.183)
Loblolly Pine	A	926	0.03	18.7 (11.8)	0.108 (0.078)	−0.001 (0.252)
Loblolly Pine	G	926	0.03	16.0 (9.6)	0.090 (0.061)	0.013 (0.343)

Values in parentheses represent standard deviations calculated across MCMC iterations. **G**: genomic relationships; **A**: pedigree relationships; **H**: blended genomic and pedigree relationships; Θ: group coancestry. Overlap Count represents mean number of individuals selected in both MAP and MCMC solutions out of 100 total selections.

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.15.711440; this version posted March 18, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

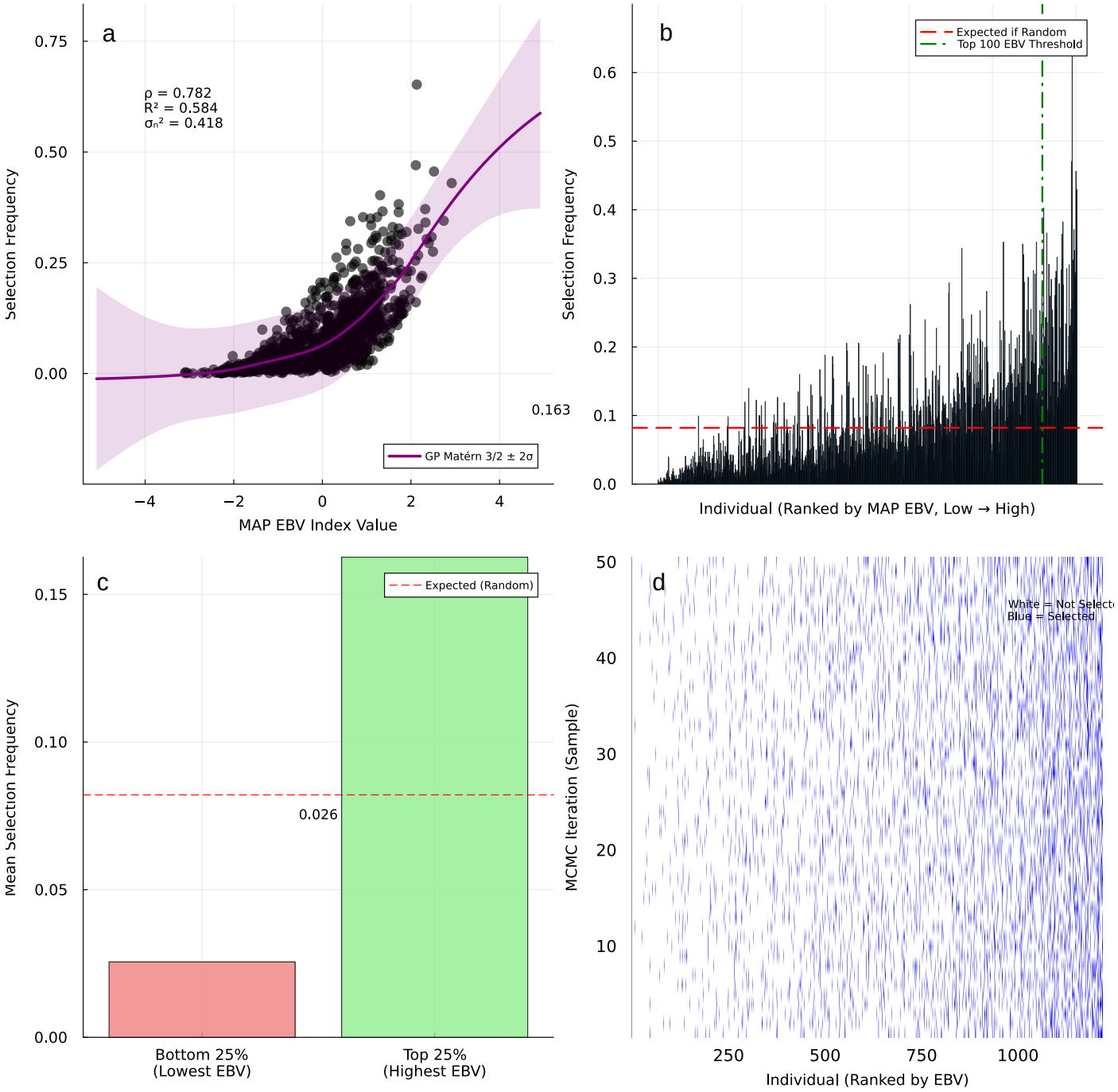


Figure 2 Detailed validation analysis of MCMC-based optimum contribution selection uncertainty for Norway spruce breeding population (G-matrix, 1,218 individuals). **a** Scatter plot showing relationship between MAP breeding values and selection frequency across MCMC iterations, with GP regression trend indicating positive correlation despite high uncertainty. The text corresponds to Spearman correlation (ρ), GP regression variance explained (R^2) and inferred noise variation (σ_n^2). **b** Selection frequency distribution for individual trees ranked by MAP breeding values, demonstrating clear gradient from low-EBV individuals (rarely selected) to high-EBV individuals (frequently selected), with reference lines showing expected random selection frequency (red dashed) and top 100 breeding value threshold (green dash-dot). **c** Quartile comparison showing mean selection frequencies for bottom 25% versus top 25% of breeding values, with expected random frequency reference (red dashed line). **D** Binary heatmap of selection patterns across a sample of MCMC iterations, where white pixels indicate non-selection and blue pixels indicate selection in top 100 contributors, with individuals sorted by breeding value rank. Despite substantial uncertainty (26.6% mean overlap), the analysis demonstrates that higher breeding value individuals are preferentially selected.

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.15.711440; this version posted March 18, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

We calculated individual importance robustness score by assessing the uncertainty of EBV via MCMC, by comparing the impact of gain by removing each selected individual each iteration. In this analysis, our objective is to evaluate the quality of the decision by identifying which of the top MAP-selected individuals are truly robust choices versus potentially risky ones.

For the 1,218-individual Norway spruce subset, robustness scores were calculated for the top 200 most relevant individuals (partly selected by the MAP-OCS first and the rest by average EBV). When applying the MAP-OCS, 154 individuals were selected with a expected genetic gain of $G_{MAP-OCS} = 1.46$ at an average coancestry of $\Theta = 0.02$ (Table 5, Fig. 3). Applying the dynamic threshold (bottom quartile of robustness score distribution), we identified 25 high-risk individuals with low robustness scores (mean $S_i = 0.013 \pm 0.008$) compared to the average selected individuals by MAP-OCS (mean $S_i = 0.044 \pm 0.027$), representing a 226% difference. Then we removed the high risk individuals and re-analyzed the remaining population, and this constrained OCS solution selected 58 individuals achieved an average robustness score of $S_i = 0.051$ and thus an increase with 16.5%. The loss in genetic gain was small, $G_{Constrained} = 1.42$, which corresponds to a loss in gain of 3.0%. There were 128 commonly selected individuals (81.1% of MAP-OCS) from both MAP-OCS and constrained OCS, while 26 and 33 were unique in the respective contribution solution obtained. Interestingly, out of the top 25 individuals in terms of contributions obtained by MAP-OCS, three were assigned as high risk individuals, and one of those individuals were in the top 10 selections.

Excluding high-risk individuals led to redistribution of genetic contributions across the selected population. Mean individual contribution decreased by 7.1% (0.00952 to 0.00885), reflecting the increased number of selected individuals (154 to 161). However, contribution variance increased by 6.9%, with the coefficient of variation rising from 122% to 135%. This pattern—lower mean but higher variance—indicates that replacement individuals took heterogeneous roles to compensate for excluded genotypes. The Gini coefficient increased from 0.586 to 0.619 (+5.6%), and top-10% concentration rose from 39.2% to 44.2%, reflecting slight concentration despite the broader selection base. This redistribution pattern likely reflects the constrained optimization adjusting contribution magnitudes to maintain genetic gain while satisfying relationship constraints with a modified candidate pool. Excluding high-risk individuals expanded selection from 154 to 161 individuals, reducing mean contributions by 7.1% while increasing contribution variance by 6.9%. Despite this increased heterogeneity, population robustness improved by 16.5% with minimal genetic gain cost (3.0%), demonstrating effective risk mitigation.

The robustness analysis of the Loblolly pine population resulted in 27 out of the original 37 individuals selected using MAP-OCS being retained after robustness scores were calculated using MCMC-OCS results (Supplementary Fig. 5). Nine individuals were removed due to high risk, while 15 new individuals were added to the constrained OCS selection, increasing the mean robustness score in the selected cohort by 29.8% from 0.016 ± 0.031 to 0.021 ± 0.028. The change in contributions resulted in a modest loss of genetic gain of 0.042 (2.14% reduction).

The exclusion of high-risk individuals in the selected Loblolly pine population impacted the contribution structure: mean contribution decreased 13.8%, maximum contribution decreased 3.9% (from 12.06% to 11.85% of total genetic gain), and the Gini coefficient increased 14.2% (from 0.517 to 0.590), indicating greater concentration among the constrained solution (Table 5). Despite increasing inequality metrics, the constrained solution achieved superior robustness while expanding the selected population from 37 to 42 individuals, suggesting that replacement individuals assumed heterogeneous roles to maintain genetic gain under more stringent restrictions ($\Theta = 0.03$).

Discussion

Optimum contribution selection (OCS) is a popular choice to select superior individuals and their optimal contribution to the next breeding generation (Dagnachew and Meuwissen 2016; Meuwissen 1997; Grundy *et al.* 1998; Hinrichs *et al.* 2006; Wellmann 2019; Mullin and Belotti 2015; Wang *et al.* 2017; Hjortø *et al.* 2022; El-Kassaby *et al.* 2024; Waldmann 2025). Despite its popularity, little effort have been devoted to investigate the impact of uncertainty in EBV on optimized contributions. Here, we use the JumP suite of optimizing tools in Julia to perform OCS on two forest tree breeding populations. The uncertainty was accounted for via repeated calls to the OCS algorithm across MCMC iterations, which was then compared to the MAP-OCS derived contributions (which ignore uncertainty). We show that the uncertainty in the EBV resulted in little average overlap between MCMC-OCS and MAP-OCS, highlighting the difficulties to rank individuals within families for realistic relatedness constraints. We introduce a robustness score metric that tells us whether the selection of an individual is low- or high risk. By replacing high risk individuals with suited low risk alternatives, the robustness score can be maximized among the selected cohort which helps decision making in breeding programs.

Given the relative harsh constraint on relatedness ($\Theta = 0.02$ in the Norway spruce case study), much of the gain obtained in the OCS was due to exploiting the within-family genetic component. However, when accessing the uncertainty of the EBV via the posterior distributions, there was a dramatic shift in the within-family rank across iterations compared to the MAP optimal solution which ignores the uncertainty in EBVs. These changes in within-family classifications are directly related to the importance of Mendelian segregation variance in OCS (Avendaño *et al.* 2004; Villanueva *et al.* 2006; Woolliams *et al.* 2015; Howard *et al.* 2018). Avendaño *et al.* (2004) showed that optimized contributions of selected candidates are allocated according to the selected individuals Mendelian sampling terms, or their superiority over their parental average. In other words, accurate estimation of the mendelian sampling advantage is of key importance if OCS are to be successful. Howard *et al.* (2018) confirmed this property of OCS using data from a commercial pig population. Thus, our results suggest that the emphasis on improving the estimation of the within-family additive genetic component should be prioritized when performing OCS in breeding and conservation programs. Our framework highlights how individuals with poorly estimated genetic component results in dramatic within-family rank changes when uncertainty is accounted for. To improve the precision of the within-family genetic component, a broad scale and scope of progeny testing have been suggested (Kerr *et al.* 2015), increasing family representation and size (Chu *et al.* 2019), and enhancing phenotyping protocols (Angidi *et al.* 2025).

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.15.711440; this version posted March 18, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

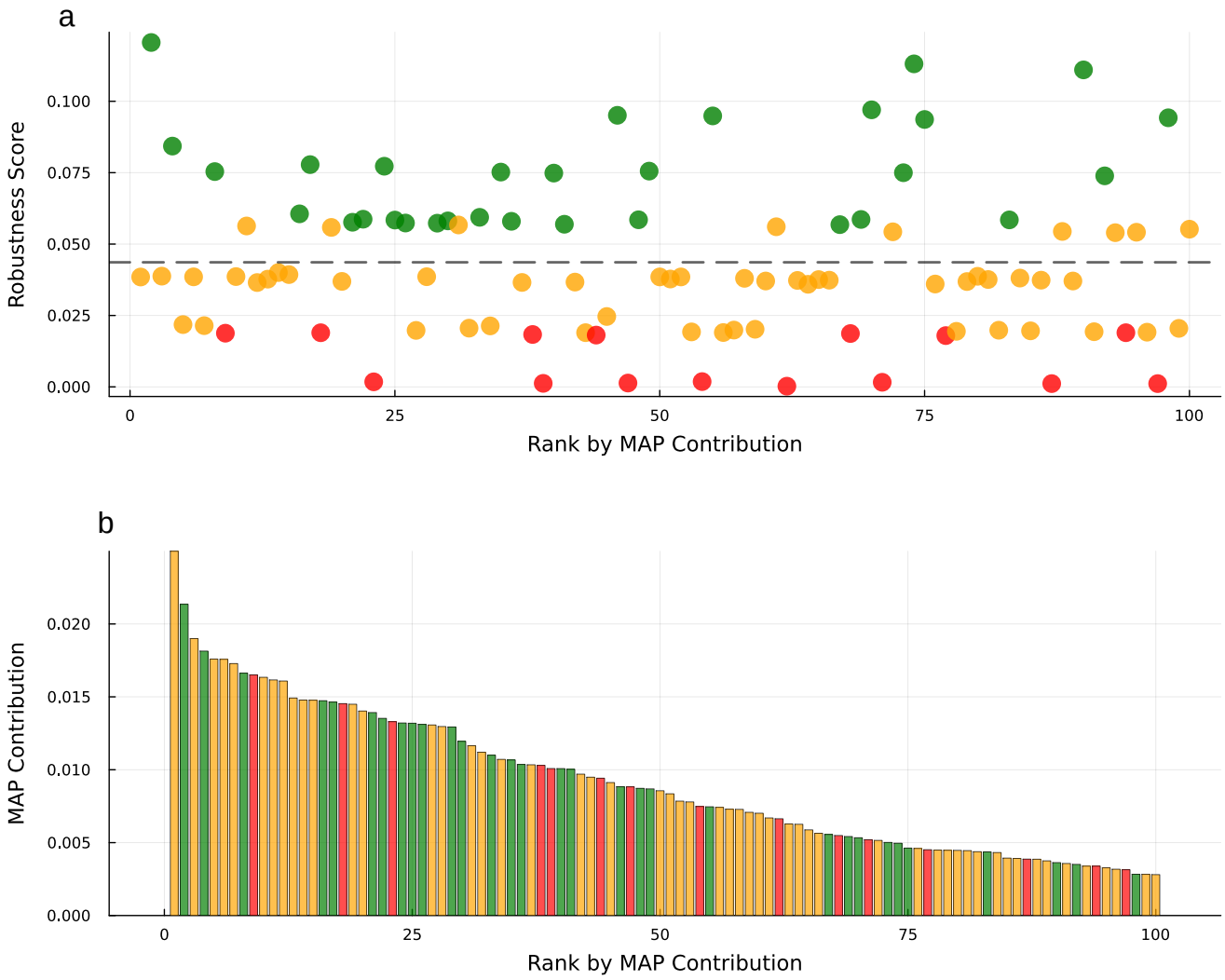


Figure 3 Risk assessment of top MAP-OCS selections based on MCMC robustness analysis in Norway spruce genotyped subpopulation. a Scatter plot showing the relationship between standardized breeding values (x-axis) and MCMC robustness scores (y-axis) for the top 100 individuals selected by MAP-based optimum contribution selection. Points are colored by risk category: green = Low Risk (robustness score $\geq 1.5 \times$ median), orange = Moderate Risk, red = High Risk (robustness score $\leq 0.85 \times$ median). Individual IDs are labeled above each point. Higher robustness scores indicate individuals whose removal causes greater genetic gain loss across MCMC iterations. **b** Bar plot showing the MAP contribution values for the same individuals, ranked by contribution magnitude, with bars colored according to their risk classification. The analysis reveals that some high-contributing individuals (left side of plot) have low robustness scores, indicating potential suboptimal selection decisions when breeding value uncertainty is considered. Dynamic thresholds classified 15 individuals as high-risk among the top 100 contributors.

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.15.711440; this version posted March 18, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Table 5 Dual metrics framework results for uncertainty-aware breeding program optimization. Comparison of MAP-OCS and Constrained solutions across Norway spruce ($n = 1,218$) and Loblolly pine ($n = 926$) populations. Constrained solutions exclude bottom 25% of selected individuals with robustness scores. Values show absolute measures and percentage changes from MAP-OCS baseline. Here, $\Theta = 0.02$ and $\Theta = 0.03$ for Norway spruce and Loblolly pine constraints in OCS, respectively.

Metric	Solution	Norway Spruce		Loblolly Pine	
		Value	Change (%)	Value	Change (%)
Genetic Gain & Selection					
Genetic Gain	MAP-OCS	1.464	—	1.943	—
	Constrained	1.420	−3.00	1.901	−2.14
Individuals Selected	MAP-OCS	154	—	37	—
	Constrained	161	+4.5	42	+13.5
Metric 1: Individual Importance (MCMC-Based)					
Analyzed Individuals	Total	200	—	137	—
Coverage	Selected (%)	65	—	100	—
	(n)	100/154	—	37/37	—
Mean Robustness	MAP-OCS	0.0436	—	0.0165	—
	Constrained	0.0508	+16.5	0.0214	+29.8
High-Risk Excluded	Count	25	—	9	—
	% of Selected	16.2	—	24.4	—
Metric 2: Contribution Distribution					
Mean Contribution	MAP-OCS	0.0095	—	0.0525	—
	Constrained	0.0089	−7.1	0.0453	−13.8
Max Contribution	MAP-OCS	0.0726	—	0.2344	—
	Constrained	0.0772	+6.3	0.2252	−3.9
Max (% of total)	MAP-OCS	4.95	—	12.06	—
	Constrained	5.42	+9.5	11.85	−1.7
Gini Coefficient	MAP-OCS	0.586	—	0.517	—
	Constrained	0.619	+5.6	0.590	+14.2
Top-10% Concentration	MAP-OCS	39.16	—	35.87	—
	Constrained	44.21	+12.9	43.26	+20.6

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.15.711440; this version posted March 18, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

The low overlap observed between MCMC-OCS and MAP-OCS solutions has important implications for within-family selection strategies commonly employed in forest and plant breeding programs (Danell 1993; Simmonds 1996). Traditional within-family selection—where a fixed number of individuals are selected from each family based on point estimates of breeding values—may frequently select suboptimal or high-risk individuals when uncertainty in EBVs is not accounted for. This is particularly relevant given that within-family selection is often implemented as a practical strategy to manage genetic diversity and maintain representation across families (i.e. equal founder contribution (Danell 1993; Avendaño *et al.* 2004)). Our robustness framework provides a mechanism to identify which within-family selections are stable and well-supported versus which are highly sensitive to estimation error, thereby improving the reliability of operational selection decisions in breeding programs that rely on within-family selection protocols.

The dual metrics framework introduced here provides complementary perspectives on breeding program stability (Woolliams *et al.* 2015). Individual importance scores (MCMC-based) quantify which candidates are critical contributors whose removal would substantially impact genetic gain, while contribution distribution metrics assess how genetic gain is allocated across the selected breeding population. This distinction is crucial: a breeding program may have high individual robustness scores (indicating important individuals are correctly identified) but poor contribution distribution structure (with genetic gain overly concentrated in few individuals). Conversely, an evenly distributed breeding population may still contain unreliable individuals with high breeding value uncertainty (Bijma 2012; Fogg *et al.* 2024).

Our results demonstrate that these metrics often reveal different aspects of selection quality. In Loblolly pine, the exclusion of high-uncertainty individuals improved both individual robustness (+29.9%) and contribution distribution (maximum contribution reduced by 3.9%, mean contribution by 13.8%), indicating that uncertainty was concentrated in structurally important individuals from key families. In Norway spruce, individual robustness improved by 16.5%, yet contribution distribution metrics moved in the opposite direction - mean contribution decreased modestly (-7.1%) while the Gini coefficient and top-10% concentration increased (+5.6% and +12.9%, respectively), suggesting that uncertainty was more broadly distributed across the candidate pool and that replacement individuals assumed heterogeneous roles to compensate for excluded genotypes. These contrasting case-specific patterns - where Loblolly pine case study showed concordant improvements across both metric types while Norway spruce case showed divergence between individual and portfolio-level outcomes - highlight the value of the comprehensive dual-metrics framework rather than relying on any single indicator of breeding program robustness (Bijma 2012; Jurcic *et al.* 2025).

Our approach is based on two important recent contributions to uncertainty-aware OCS. Pocrnic *et al.* (2023) demonstrated the value of incorporating posterior EBV distributions into OCS decisions using a simulated population, providing the conceptual foundation that motivated our work. Fogg *et al.* (2024) developed a robust OCS framework using standard errors of EBVs, offering a practical and computationally efficient solution applicable to frequentist inference settings. Our MCMC-OCS framework is complementary to both: it extends the posterior sampling approach of Pocrnic *et al.* (2023) to operational breeding pedigrees and addresses settings where full posterior distributions are available from Bayesian inference rather than summary uncertainty measures as in Fogg *et al.* (2024). Here, we have also provided a practical decision-making framework with robustness metrics and replacement strategies which can be utilized to boost long term stability of breeding programs. Together, these converging research efforts underscore the growing recognition that the incorporation of EBV uncertainty into OCS decisions is an important and timely direction for the field.

Data availability

Norway spruce phenotypic and genotypic data with recoded IDs are provided here: https://github.com/jonhar97/MCMC_uncertainty_OCS/.

Acknowledgments

This work was supported by grant from Föreningen Skogsträdsförädling project ID 23-511. The online version contains supplementary material here;

Funding

JA was supported by project ID 21414 (Föreningen ID 23-511) funded by Föreningen Skogsträdsförädling.

Conflicts of interest

The authors declare that they have no conflict of interest.

Literature cited

Ahlinder J, Hall D, Suontama M, Sillanpää MJ. 2024. Principal component analysis revisited : fast multi-trait genetic evaluations with smooth convergence principal component analysis revisited : fast multi-trait genetic evaluations with smooth convergence . G3: Genes, Genomes, Genetics. pp. 1–17.

Ahlinder J, Mullin TJ, Yamashita M. 2013. Using semidefinite programming to optimize unequal deployment of genotypes to a clonal seed orchard. Tree Genetics & Genomes. .

Amadeu RR, Garcia AAF, Munoz PR, Ferrão LFV. 2023. Aghmatrix: genetic relationship matrices in r. Bioinformatics. 39:btad445.

Angidi S, Madankar K, Tehseen MM, Bhatla A. 2025. Advanced high-throughput phenotyping techniques for managing abiotic stress in agricultural crops—a comprehensive review.

Avendaño S, Woolliams JA, Villanueva B. 2004. Mendelian sampling terms as a selective advantage in optimum breeding schemes with restrictions on the rate of inbreeding. Genetics Research. 83:55–64.

Barbieri MM, Berger JO. 2004. Optimal Predictive Model Selection. Annals of Statistics. 32:870–897.

Beck A. 2017. *First-Order Methods in Optimization*. volume 25 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics. Philadelphia, PA.

Bijma P. 2012. Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the accuracy of predicted phenotypes in selected populations. Journal of Animal Breeding and Genetics. 129:345–358.

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.15.711440; this version posted March 18, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. 2010. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning. 3:1–122.

Charlesworth D, Willis JH. 2009. The genetics of inbreeding depression. Nature Reviews Genetics. 10:783–796.

Chen ZQ, Baison J, Pan J, Karlsson B, Andersson B, Westin J, García-Gil MR, Wu HX. 2018. Accuracy of genomic selection for growth and wood quality traits in two control-pollinated progeny trials using exome capture as the genotyping platform in Norway spruce. BMC Genomics. 19:1–16.

Cheng H, Kizilkaya K, Zeng J, Garrick D, Fernando R. 2018. Genomic prediction from multiple-trait bayesian regression methods using mixture priors. Genetics. 209:89–103.

Christensen OF, Madsen P, Nielsen B, Ostersen T, Su G. 2012. Single-step methods for genomic evaluation in pigs. Animal. 6:1565–1571.

Chu TT, Bastiaansen JW, Berg P, Komen H. 2019. Optimized grouping to increase accuracy of prediction of breeding values based on group records in genomic selection breeding programs. Genetics Selection Evolution. 51.

Clark SA, Kinghorn BP, Hickey JM, van der Werf JH. 2013. The effect of genomic information on optimal contribution selection in livestock breeding programs. Genetics Selection Evolution. 45:44.

Cowling WA, Castro-Urrea FA, Stefanova KT, Li L, Banks RG, Saradadevi R, Sass O, Kinghorn BP, Siddique KHM. 2023. Optimal contribution selection improves the rate of genetic gain in grain yield and yield stability in spring canola in australia and canada. Plants. 12:383.

Dagnachew BS, Meuwissen THE. 2016. A fast Newton-Raphson based iterative algorithm for large scale optimal contribution selection. Genetics Selection Evolution. 48:70.

Danell O. 1993. Breeding programmes in sweden. 1. general approach. In: . pp. 80–94. Edinburgh, Scotland. Forestry Authority.

Derks MFL, Steensma M. 2021. Balancing selection for deleterious alleles in livestock. Frontiers in Genetics. 12:761728.

El-Kassaby YA, Cappa EP, Chen C, Ratcliffe B, Porth IM. 2024. Efficient genomics-based 'end-to-end' selective tree breeding framework. Heredity. 132:98–105.

Endelman JB. 2011. Ridge regression and other kernels for genomic selection with r package rrblup. Plant Genome. 4:250–255.

Fogg J, Ortiz J, Pocrnić I, Hall JAJ, Gorjanc G. 2024. Robust Optimal Contribution Selection. arXiv. pp. 1–10.

Garstka M, Fürst J, Schöps S, Clemens M. 2019. Cosmo: A conic operator splitting method for large convex problems. Optimization Methods and Software. 34:965–992.

Gautason E, Sahana G, Guldbrandtsen B, Berg P. 2023. Impact of kinship matrices on genetic gain and inbreeding with optimum contribution selection in a genomic dairy cattle breeding program. Genetics Selection Evolution. 55:48.

Ghosh J. 2015. Bayesian Model Selection Using the Median Probability Model. WIREs Computational Statistics. 7:185–193.

Goddard ME. 1994. Selection indices for non-linear profit functions. Theoretical and Applied Genetics. 64:339–344.

Grundy B, Villanueva B, Woolliams JA. 1998. Dynamic selection procedures for constrained inbreeding and their consequences for pedigree development. Genetical Research. 72:159–168.

Hallander J, Waldmann P. 2009. Optimum contribution selection in large general tree breeding populations with an application to Scots pine. Theoretical and Applied Genetics. 118:1133–1142.

Hamilton MG. 2020. Optimal Contribution Selection in Highly Fecund Species With Overlapping Generations. The Journal of Heredity. 111:646–651.

Henderson CR. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics. 31:423–447.

Hershberger J, Morales N, Simoes C *et al*. 2021. Making waves in breedbase: An integrated spectral data storage and analysis pipeline for plant breeding programs. Plant Phenome Journal. 4:e20012.

Hinrichs D, Wetten M, Meuwissen THE. 2006. An algorithm to compute optimal genetic contributions in selection programs with large numbers of candidates. Journal of Animal Science. 84:3212–3218.

Hjortø L, Andersen T, Kargo M, Sørensen AC. 2022. Breeding schemes with optimum-contribution selection or truncation selection for beef cattle destined for use on dairy females. Journal of Dairy Science. 105:4314–4323.

Howard DM, Pong-Wong R, Knap PW, Kremer VD, Woolliams JA. 2018. Selective advantage of implementing optimal contributions selection and timescales for the convergence of long-term genetic contributions. Genetics Selection Evolution. 50:24.

Jurcic EJ, Dutour J, Villalba PV, Centurión C, Cantet RJC, Munilla S, Cappa EP. 2025. Forest tree breeding using genomic markov causal models: a new approach to genomic tree breeding improvement. Heredity. 134:280–292.

Kerr RJ, Dutkowski GW, Jansson G, Persson T, Westin J. 2015. Connectedness among test series in mixed linear models of genetic evaluation for forest trees. Tree Genetics and Genomes. 11:1–13.

Legarra A, Aguilar I, Misztal I. 2009. A relationship matrix including full pedigree and genomic information. Journal of Dairy Science. 92:4656–4663.

Lions P, Mercier B. 1979. Splitting algorithms for the sum of two nonlinear operators. SIAM Journal on Numerical Analysis. 16:964–979.

Lynch M, Walsh B. 1998. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates. Sunderland, Massachusetts.

McRae T, Apiolaza L, Dutkowski G, Kerr R, Pilbeam D, Powell M, Tier B. 2003. Treeplan: A genetic evaluation system for forest trees. Impact Assessment. .

Meuwissen THE. 1997. Maximizing the response of selection with a predefined rate of inbreeding. Journal of Animal Science. 75:934–940.

Mogensen PK, Riseth AN. 2018. Optim: A mathematical optimization package for Julia. Journal of Open Source Software. 3:615.

Mullin TJ, Belotti P. 2015. Using branch-and-bound algorithms to optimize selection of a fixed-size breeding population under a relatedness constraint. Tree Genetics & Genomes. 12:4.

Nocedal J, Wright SJ. 2006. *Numerical Optimization*. Springer Science. New York, NY. second edition.

Pocrnic I, Ortiz J, Gorjanc G. 2023. Evaluating the impact of uncertainty in estimated breeding values on optimal contribution selection. In: . Lyon, France.

Pong-Wong R, Woolliams JA. 2007. Optimisation of contribution of candidate parents to maximise genetic gain and restricting inbreeding using semidefinite programming. Genetics Selection Evolution. 39:3–25.

Qi H, Sun D. 2006. A quadratically convergent newton method

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.15.711440; this version posted March 18, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

for computing the nearest correlation matrix. SIAM Journal on Matrix Analysis and Applications. 28:360–385.

Rasmussen CE, Williams CK. 2006. *Gaussian Processes for Machine Learning*. MIT Press. Cambridge, MA.

Resende JF, Muñoz P, Resende MD, Garrick DJ, Fernando RL, Davis JM, Jokela EJ, Martin TA, Peter GF, Kirst M. 2012. Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.). Genetics. 190:1503–1510.

Robertson A. 1960. A theory of limits in artificial selection. Proceedings of the Royal Society of London. Series B, Biological Sciences. 153:234–249.

Simmonds NW. 1996. Family selection in plant breeding. Euphytica. 90:201–208.

Sorensen D, Gianola D. 2002. *Likelihood, Bayesian, and MCMC methods in quantitative genetics*. Springer. New York.

Tao L, Liu H, Adeola AC, Xie HB, Feng ST, Zhang YP. 2025. The effects of runs-of-homozygosity on pig domestication and breeding. BMC Genomics. 26:6.

Tchounke B, Sanchez L, Bell JM, Cros D. 2023. Mate selection: A useful approach to maximize genetic gain and control inbreeding in genomic and conventional oil palm (Elaeis guineensis Jacq.) hybrid breeding. PLoS Computational Biology. 19:1–27.

van der Wilk M, Dutordoir V, John S, Artemev A, Adam V, Hensman J. 2020. A framework for interdomain and multioutput gaussian processes. arXiv preprint arXiv:2003.01115. .

Villanueva B, Avendaño S, Woolliams JA. 2006. Prediction of genetic gain from quadratic optimisation with constrained rates of inbreeding. Genetics Selection Evolution. 38:127–146.

Waldmann P. 2025. Genomic optimum contribution selection and mate allocation using jump. Bioinformatics Advances. 5.

Wang Y, Bennewitz J, Wellmann R. 2017. Novel optimum contribution selection methods accounting for conflicting objectives in breeding programs for livestock breeds with historical migration. Genetics Selection Evolution. 49.

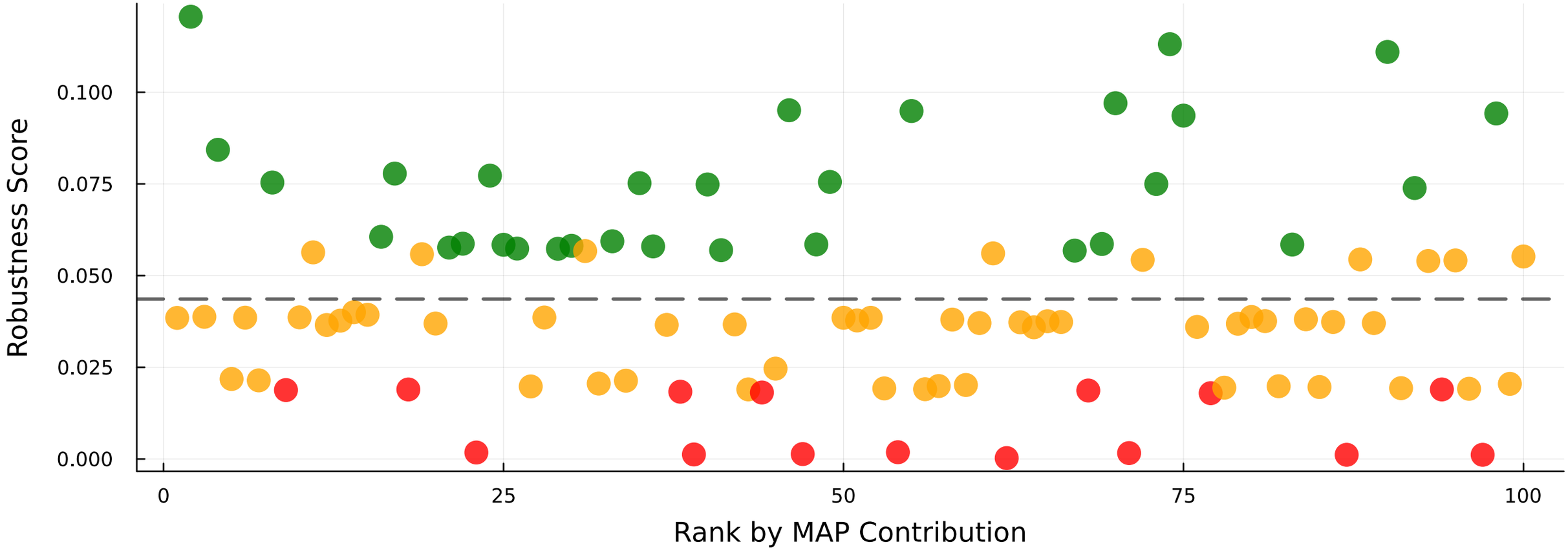
Wellmann R. 2019. Optimum contribution selection for animal breeding and conservation: The R package optiSel. BMC Bioinformatics. 20.

Woolliams JA, Berg P, Dagnachew BS, Meuwissen TH. 2015. Genetic contributions and their optimization. Journal of Animal Breeding and Genetics. 132:89–99.

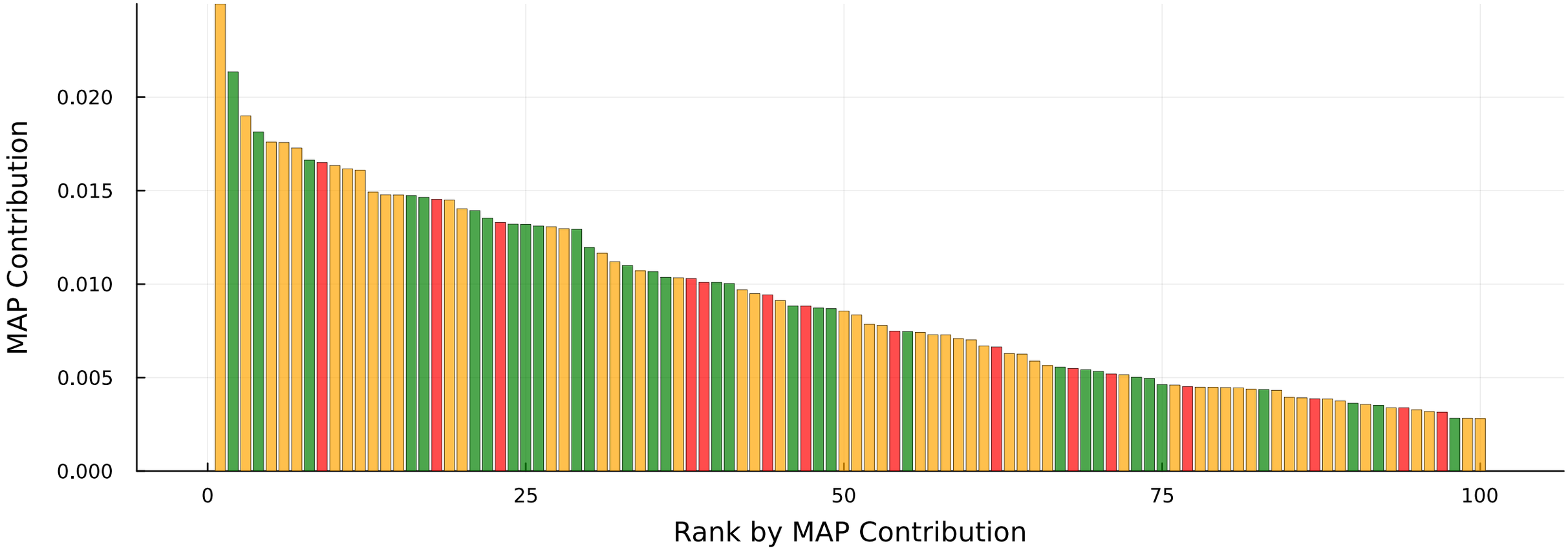
Yamashita M, Mullin TJ, Safarina S. 2018. An efficient second-order cone programming approach for optimal selection in tree breeding. Optimization Letters. 12:1683–1697.

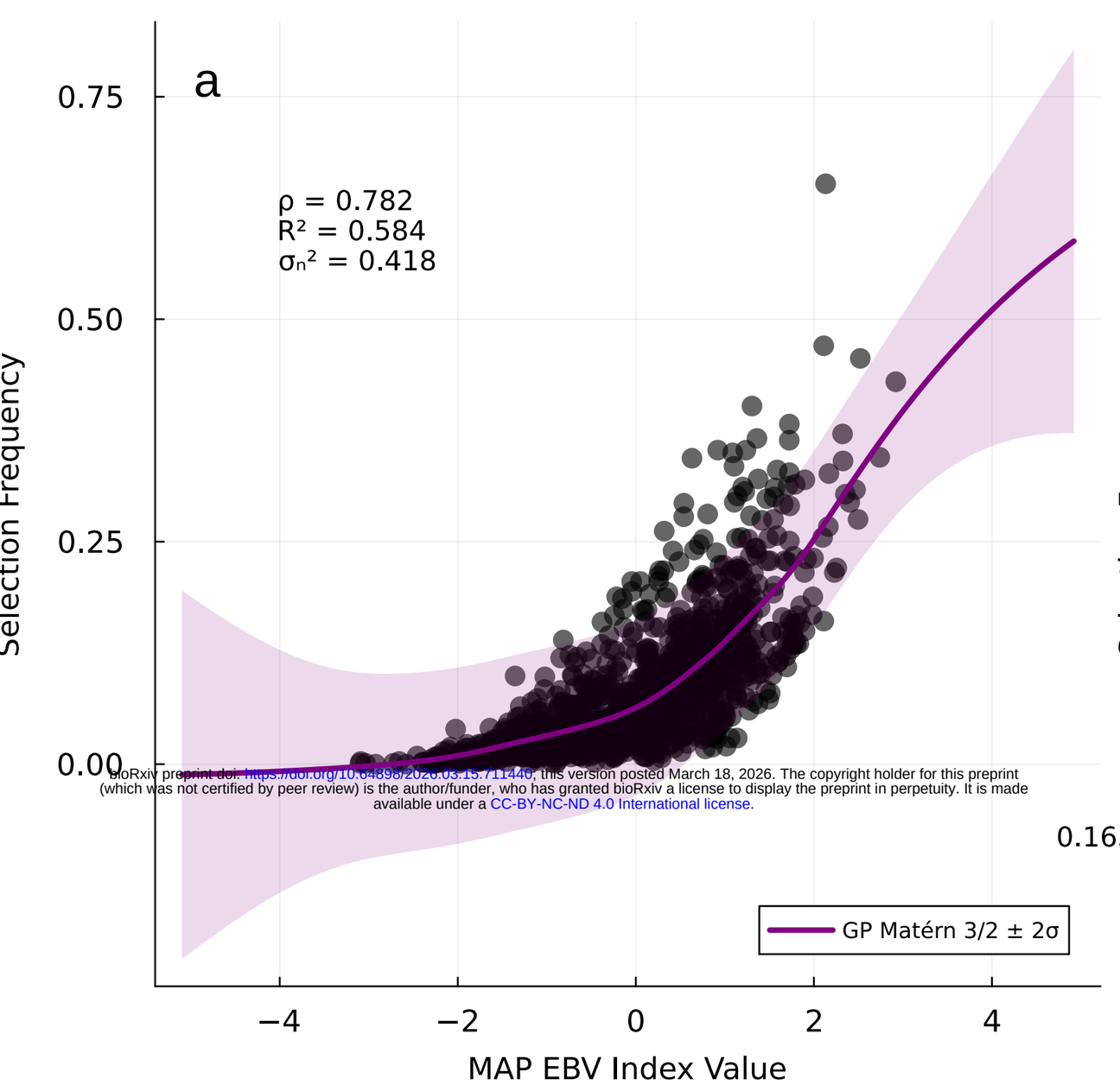
bioRxiv preprint doi: https://doi.org/10.64898/2026.03.15.711440; this version posted March 18, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

a

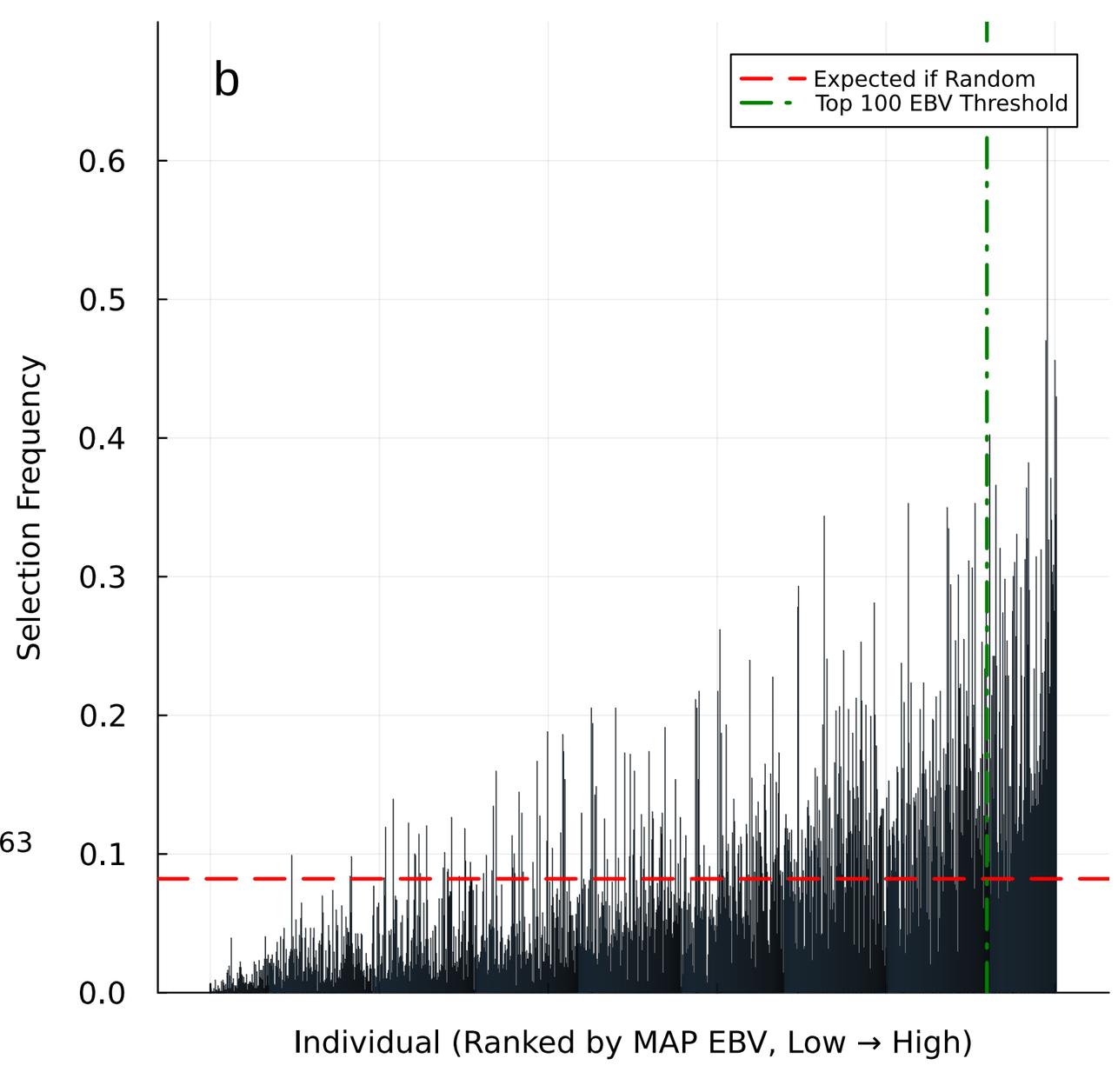


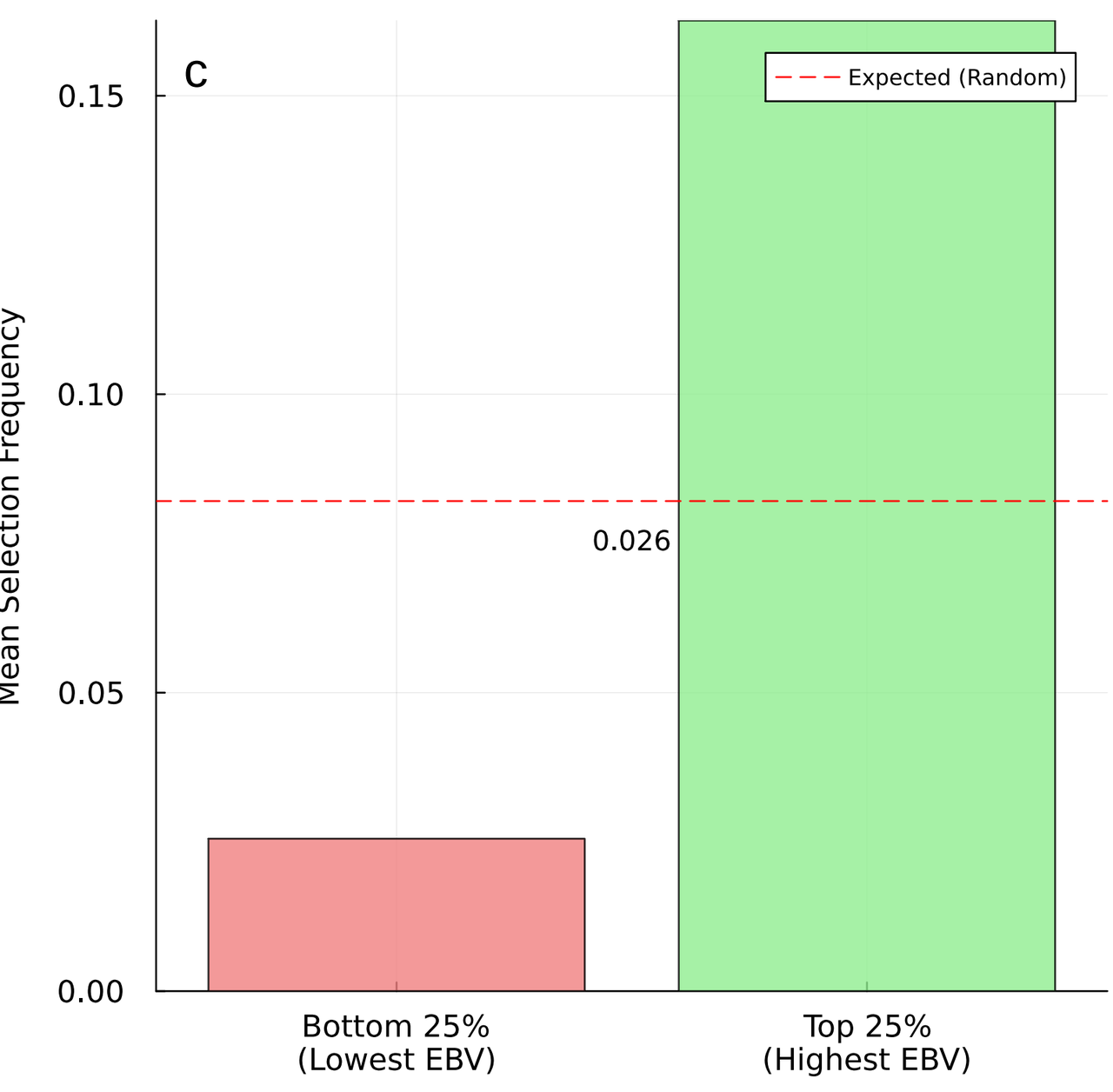
b

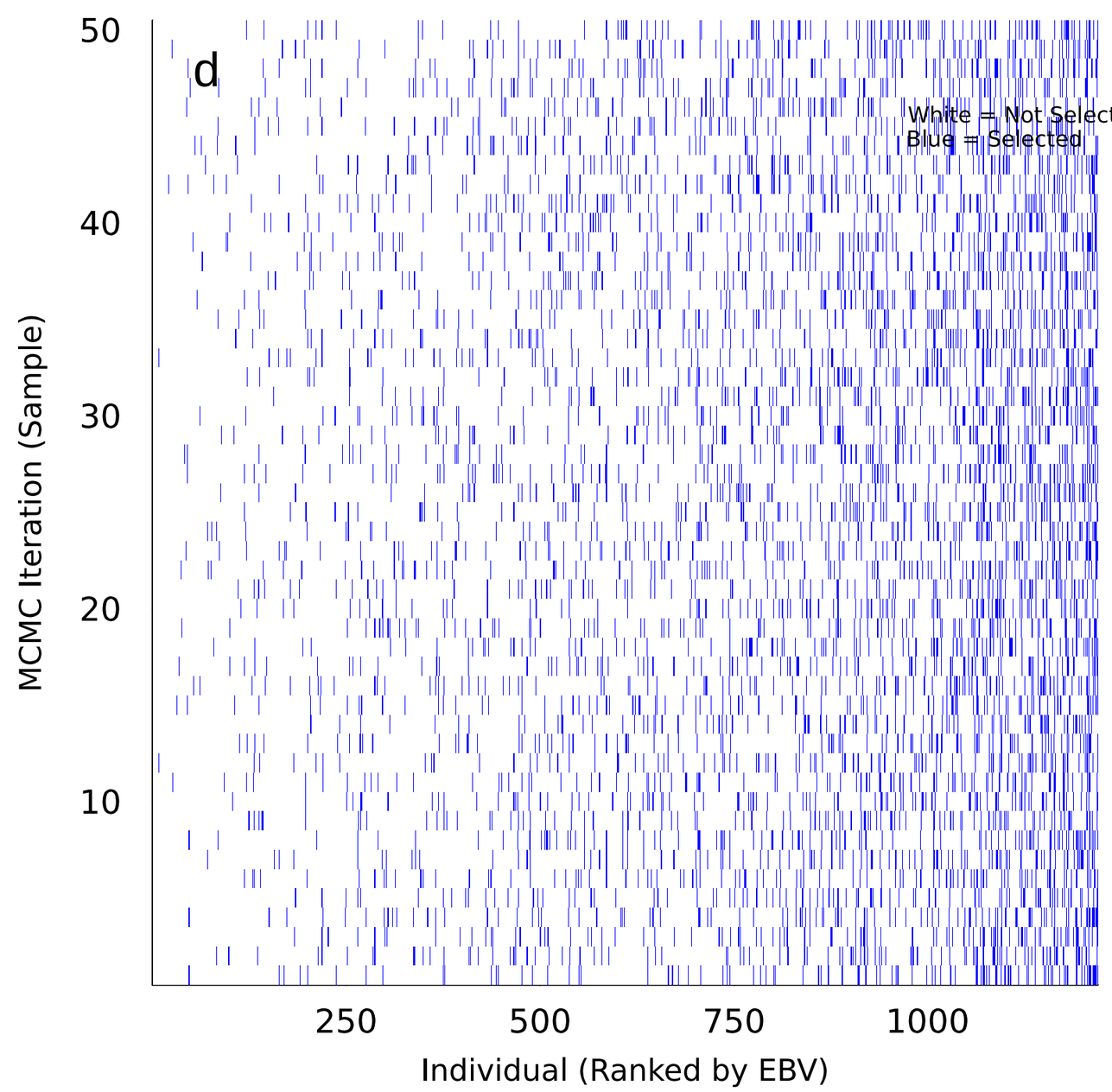


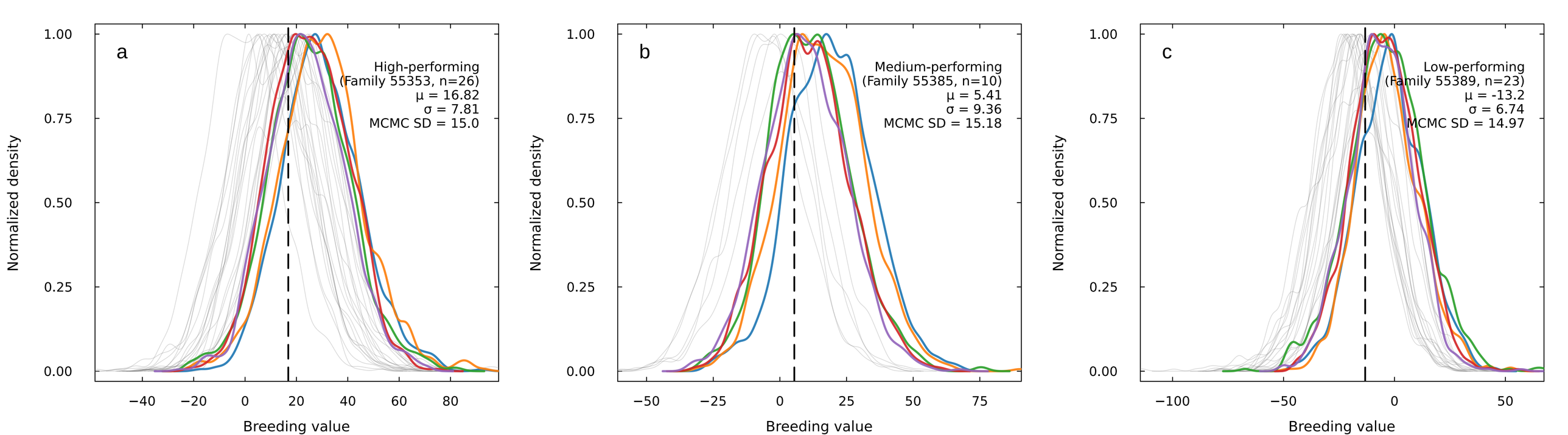


bioRxiv preprint doi: https://doi.org/10.64898/2026.03.15.711446; this version posted March 18, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.








a
High-performing
(Family 55353, n=26)
μ = 16.82
σ = 7.81
MCMC SD = 15.0
b
Medium-performing
(Family 55385, n=10)
μ = 5.41
σ = 9.36
MCMC SD = 15.18
c
Low-performing
(Family 55389, n=23)
μ = -13.2
σ = 6.74
MCMC SD = 14.97
Normalized density
Breeding value
1.00
0.75
0.50
0.25
0.00
−40
−20
0
20
40
60
80
−50
−25
0
25
50
75
−100
−50
0
50