frontiers in Genetics

TYPE Methods
PUBLISHED 03 June 2026
DOI 10.3389/fgene.2026.1792190

OPEN ACCESS

EDITED BY
Yongjun Shu,
Harbin Normal University, China

REVIEWED BY
Hailan Liu,
Maize Research Institute of Sichuan Agricultural University, China
Zhixu Pang,
Shanxi Agriculture University, China

*CORRESPONDENCE
Vinícius Silva Junqueira,
viniciussilva.junqueira@bayer.com

RECEIVED 20 January 2026
REVISED 22 April 2026
ACCEPTED 12 May 2026
PUBLISHED 03 June 2026

CITATION
Junqueira VS, Yokoo MJ-I and Cardoso FF (2026) Derivation of prediction error variance for non-genotyped individuals in genomic selection.
Front. Genet. 17:1792190.
doi: 10.3389/fgene.2026.1792190

Derivation of prediction error variance for non-genotyped individuals in genomic selection

Vinícius Silva Junqueira[1]*, Marcos Jun-Iti Yokoo[2] and Fernando Flores Cardoso[3]

[1]Bayer R&D, Bayer Crop Science, Uberlândia, Brazil, [2]Embrapa Southeastern Livestock, São Carlos, Brazil, [3]Embrapa South Livestock, Bagé, Brazil

Genomic selection has transformed plant and animal breeding by enabling accurate prediction of genetic merit using DNA markers; however, comprehensive genotyping of all selection candidates remains economically prohibitive for most breeding programs. While breeding programs must decide which subset of individuals to genotype within budget constraints, current approaches rely primarily on experience-based decisions rather than quantitative frameworks. We present explicit mathematical derivations for prediction error variance (PEV) in non-genotyped individuals under mixed model equations, providing a theoretical foundation for evaluating genotyping strategies prospectively. The approach derives PEV expressions for non-genotyped selection candidates under different relationship matrix structures, including pedigree-based, genomic, and hybrid single-step methodologies that combine both information sources. The derivations accommodate complex breeding program structures with historical training populations containing both genotypes and phenotypes alongside contemporary selection candidates with only pedigree information. Using Schur complement methods applied to partitioned mixed model equations, the framework enables calculation of prediction uncertainty without requiring actual phenotypic data from selection candidates. The expressions simplify under different information scenarios, from cases with complete phenotypic data to situations where only relationship information is available. The method was validated through simulations across six scenarios with populations ranging from 180 to 15,500 individuals, confirming numerical equivalence with direct matrix inversion while demonstrating computational and memory advantages that increase with population size. Although genomic relationship matrix operations dominate the complexity, matrix decomposition techniques, including Cholesky factorization and APY methodology, can improve efficiency. The mathematical framework provides quantitative tools for transitioning from experience-based to mathematically-informed genotyping decisions, with applications extending to any field requiring prospective quantification of prediction uncertainty under resource constraints.

KEYWORDS
budget constrained, matrix decomposition, mixed model equations, prediction uncertainty, schur complement

1 Introduction

Genomic selection has revolutionized plant and animal breeding by enabling the prediction of genetic merit using DNA markers (Meuwissen et al., 2001), fundamentally changing how breeding programs allocate resources and make selection decisions. Since its introduction 2 decades ago, genomic selection has delivered genetic gains across diverse species, from cattle (Garrick, 2011; Wiggans et al., 2017; Guinan et al., 2023) and pigs (Sharif-Islam et al., 2024) to wheat (He et al., 2016) and maize (Crossa et al., 2017), by allowing breeders to identify superior individuals earlier in their development and with higher accuracy than traditional phenotypic selection methods. However, despite dramatic reductions in genotyping costs, comprehensive genotyping of all selection candidates remains economically prohibitive for some breeding programs, creating a fundamental challenge: determining which individuals should be genotyped to maximize genetic improvement within budgetary constraints. This challenge is especially acute in commercial breeding programs, which generate thousands of selection candidates annually, while genotyping budgets typically cover only a small fraction of them. For example, in animal breeding (e.g., cattle, poultry, and pigs), all individuals have pedigree-based breeding values regardless of phenotyping status, allowing for initial pre-selection based on performance. Pre-selected individuals are subsequently phenotyped, but genotyping strategies differ by species and company budget. While large poultry and pig breeding companies may typically genotype a large number of pre-selected individuals, smaller companies and all cattle breeding programs cannot afford this approach. Consequently, many individuals that ultimately contribute phenotypes to the training set remain ungenotyped. In these situations, prioritization rules must be employed to define which individuals to genotype under budget constraints.

Practical examples highlight the extent of this resource constraint. In Brazilian beef cattle breeding, the PROMEBO Angus program has genotyped only 9.4% of phenotyped animals, while the Brangus + program has genotyped 25.5%. In contrast, poultry breeding programs typically genotype all layers but only a selected subset of males from elite families (personal communication). Similar resource allocation decisions arise with emerging sequencing technologies, where programs must determine which individuals to sequence at higher densities or depths. Across all these scenarios, the question remains: which individuals should be prioritized for genotyping to maximize their contribution to prediction accuracy and genetic gain under budget constraints?

Genotyping decisions directly impact selection accuracy and overall operational efficiency, with suboptimal choices reducing genetic progress and return on investment in genomic technologies (Lopez-Cruz and De Los Campos, 2021). Traditional approaches to this problem have relied heavily on breeder experience or simple rules of thumb, such as genotyping a fixed proportion of individuals from each family (Isidro et al., 2015). While these methods have proven workable, they lack the mathematical foundation needed to quantitatively evaluate competing genotyping strategies or to balance objectives such as maximizing prediction accuracy while efficiently utilizing limited resources across complex breeding population structures.

The prediction error variance (PEV) from mixed model theory provides a natural mathematical framework for quantifying prediction uncertainty, as it captures the expected accuracy of genetic predictions (Henderson et al., 1984) before genotyping decisions are implemented. PEV represents the uncertainty associated with predicted breeding values, with lower values indicating higher prediction accuracy (Misztal and Wiggans, 1988). The prediction accuracy is proportional to the amount of information used to compute breeding values, with sources of information being pedigree, phenotypes, and genotypes.

Modern breeding programs often involve complex population structures that encompass historical training populations with both genotypic and phenotypic data, alongside contemporary young candidates that may only have pedigree information. This configuration enables the construction of joint relationship matrices that integrate genomic and pedigree information through single-step methodologies (Aguilar et al., 2010; Christensen et al., 2012). The mathematical framework of this study derives explicit PEV expressions under different relationship matrix structures, providing the theoretical foundation for the quantitative evaluation of genotyping strategies. The practical value of these expressions extends to fundamental questions about breeding program design, including training population composition (Rincent et al., 2012; Akdemir and Isidro-Sánchez, 2019), genetic diversity maintenance, and the integration of phenotypic and genomic data collection strategies.

Although several studies have previously addressed training set design for genomic selection, they assume genotypic information is available for all individuals (Rincent et al., 2012; Akdemir and Isidro-Sánchez, 2019). However, in many breeding programs, phenotyped individuals in the training set lack genotypes due to budget constraints. To our knowledge, no previous study has developed a quantitative method to guide the selection of which individuals to genotype under such resource limitations. Here, we develop explicit mathematical expressions for PEV in non-genotyped individuals that will eventually contribute phenotypes to the training population.

2 Methods

Consider a breeding program using genomic selection where candidates are first preselected based on pedigree-based breeding values. These individuals are then phenotyped and added to the training population. However, budget constraints typically prevent genotyping all preselected candidates, requiring the breeding program to prioritize which individuals will be genotyped.

This study addresses this optimization problem by deriving PEV equations for ungenotyped individuals. These equations provide a quantitative basis for prioritizing genotyping decisions and can be integrated into optimization algorithms such as differential evolution, integer programming, or simulated annealing with cost-related constraints, enabling breeders to maximize genetic gains within budget limitations. The following sections establish the mixed model equations, derive explicit expressions for prediction error variance in non-genotyped individuals, and examine how PEV calculations simplify under different information scenarios. While the practical implementation of

these equations in an optimization algorithm will be presented in a future publication, we briefly describe the rationale for such implementation in this paper.

2.1 Linear mixed model framework

Consider the standard mixed model equation (MME) for genetic evaluation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where $\mathbf{y}$ is the vector of phenotypic observations, $\mathbf{X}$ is the design matrix for fixed effects $\boldsymbol{\beta}$, $\mathbf{Z}$ is the design matrix relating observations to individuals, $\mathbf{u}$ is the vector of random additive effects and $\mathbf{e}$ is the vector of residual effects.

The variance structure is defined as:

$$\mathrm{Var}\begin{pmatrix}\mathbf{u}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{pmatrix}$$

where $\mathbf{G} = \mathbf{H}\sigma_u$, σ_u is the variance of additive effects, $\mathbf{H}$ the joint relationship matrix that combines pedigree and genomic information (Aguilar et al., 2010; Christensen and Lund, 2010). The residual term of the equations is $\mathbf{R} = \mathbf{I}\sigma_e$, σ_e is the residual variance.

2.2 Partitioned system for training and young individuals

As our objective is to decide which young individuals should be genotyped, partitioning the coefficient matrix provides a convenient way to simplify the derivations for this group, while still accounting for information from the entire training population. Accordingly, individuals are partitioned into two groups: a training set (with historical phenotypic and genotypic data) and a young set (available only through pedigree). Initially, a general matrix $\mathbf{G}$ can be partitioned into training (t) and young (y) components, along with their cross-components, as follows:

$$\mathbf{G} = \begin{bmatrix}\mathbf{G}_{tt} & \mathbf{G}_{ty}\\ \mathbf{G}_{yt} & \mathbf{G}_{yy}\end{bmatrix}$$

where $\mathbf{G}_{tt}$ is the genetic covariance matrix among training individuals, $\mathbf{G}_{yy}$ among young candidates, and $\mathbf{G}_{ty} = \mathbf{G}_{yt}^{\top}$ the cross-covariance between the two groups.

The inverse of the partitioned matrix $\mathbf{G}$ can be written as:

$$\mathbf{G}^{-1} = \begin{bmatrix}\mathbf{G}^{tt} & \mathbf{G}^{ty}\\ \mathbf{G}^{yt} & \mathbf{G}^{yy}\end{bmatrix}$$

The following presents the partitioned form of the MME, with explicit separation of training and young candidate blocks:

$$\begin{bmatrix}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}_t & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}_y\\ \mathbf{Z}_t^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_t^T\mathbf{R}^{-1}\mathbf{Z}_t + \mathbf{G}^{tt} & \mathbf{Z}_t^T\mathbf{R}^{-1}\mathbf{Z}_y + \mathbf{G}^{ty}\\ \mathbf{Z}_y^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_y^T\mathbf{R}^{-1}\mathbf{Z}_t + \mathbf{G}^{yt} & \mathbf{Z}_y^T\mathbf{R}^{-1}\mathbf{Z}_y + \mathbf{G}^{yy}\end{bmatrix} \times \begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{u}}_t\\ \hat{\mathbf{u}}_y\end{bmatrix} = \begin{bmatrix}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{y}\\ \mathbf{Z}_t^T\mathbf{R}^{-1}\mathbf{y}\\ \mathbf{Z}_y^T\mathbf{R}^{-1}\mathbf{y}\end{bmatrix}$$

However, as the young candidates do not have phenotypes, the MME can be simplified as follows:

$$\begin{bmatrix}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}_t & \mathbf{0}_{n_t}\\ \mathbf{Z}_t^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_t^T\mathbf{R}^{-1}\mathbf{Z}_t + \mathbf{G}^{tt} & \mathbf{G}^{ty}\\ \mathbf{0}_{n_y} & \mathbf{G}^{yt} & \mathbf{G}^{yy}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{u}}_t\\ \hat{\mathbf{u}}_y\end{bmatrix} = \begin{bmatrix}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{y}\\ \mathbf{Z}_t^T\mathbf{R}^{-1}\mathbf{y}_t\\ \mathbf{0}_{n_y}\end{bmatrix}$$

An important structural insight of the partitioned mixed model equations is that young individuals without phenotypic records contribute nothing to the cross-product terms $\mathbf{Z}_y^{\top}\mathbf{R}^{-1}\mathbf{Z}_y$, $\mathbf{Z}_t^{\top}\mathbf{R}^{-1}\mathbf{Z}_y$, and $\mathbf{X}^{\top}\mathbf{R}^{-1}\mathbf{Z}_y$ in the coefficient matrix. Consequently, the corresponding block of the right-hand side reduces to a null vector with the length as the number of unphenotyped individuals ($\mathbf{0}_{n_y}$).

2.3 Prediction error variance

Following Henderson et al. (1984), the prediction error variance is defined as

$$\mathrm{PEV} = \mathrm{Var}\,(\mathbf{u} - \hat{\mathbf{u}})$$

Expanding the variance of the prediction error:

$$\begin{aligned}\mathrm{PEV} &= \mathrm{Var}\,(\mathbf{u} - \hat{\mathbf{u}})\\ &= \mathrm{Var}\,(\mathbf{u}) + \mathrm{Var}\,(\hat{\mathbf{u}}) - 2\mathrm{Cov}\,(\mathbf{u}, \hat{\mathbf{u}})\end{aligned}$$

Since $\hat{\mathbf{u}}$ is the BLUP of $\mathbf{u}$, and using the property that $\mathrm{Cov}\,(\mathbf{u}, \hat{\mathbf{u}}) = \mathrm{Var}\,(\hat{\mathbf{u}})$ for best linear unbiased predictors:

$$\begin{aligned}\mathrm{PEV} &= \mathrm{Var}\,(\mathbf{u}) - \mathrm{Var}\,(\hat{\mathbf{u}})\\ &= \mathrm{Var}\,(\mathbf{u}) - \mathrm{Cov}\,(\mathbf{u}, \hat{\mathbf{u}})\\ &= \mathbf{G} - \mathrm{Cov}\,(\mathbf{u}, \hat{\mathbf{u}})\\ &= \mathbf{C}^{-1}\end{aligned}$$

where $\mathbf{C}$ is the coefficient matrix (i.e., left-hand side block) of the mixed model equations and $(\mathbf{C}^{-1})_{uu}$ is the block corresponding to the random effects.

Therefore, the prediction error variance equals:

$$\mathrm{PEV} = \mathbf{C}^{-1}_{uu}$$

In practice, the interest lies only in the prediction error variance of the young individuals. Therefore, the direct inversion of the full coefficient matrix would not be necessary. By partitioning the MME and applying the Schur complement (Ouellette, 1981) to the $\mathbf{u}_y$-block, it is possible to obtain the relevant block of the inverse directly. This approach not only simplifies the algebra but also has a clear interpretation: the Schur complement represents the information available for the young individuals after accounting for the training population and fixed effects. Thus, applying the Schur complement to the $\mathbf{u}_y$-block, the PEV of $\mathbf{u}_y$ can be expressed as

$$\begin{aligned}\mathbf{B} &= \begin{bmatrix}\mathbf{X}^{\top}\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^{\top}\mathbf{R}^{-1}\mathbf{Z}_t\\ \mathbf{Z}_t^{\top}\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_t^{\top}\mathbf{R}^{-1}\mathbf{Z}_t + \mathbf{G}^{tt}\end{bmatrix},\\ \mathbf{E} &= \begin{bmatrix}\mathbf{0}\\ \mathbf{G}^{ty}\end{bmatrix},\\ \mathbf{F} &= \mathbf{G}^{yy}.\end{aligned}$$

Then the full coefficient matrix can be written as

$$\mathbf{D} = \begin{bmatrix}\mathbf{B} & \mathbf{E}\\ \mathbf{E}^{\top} & \mathbf{F}\end{bmatrix}$$

Then, following the rules for the inverse of Schur Complement, the block corresponding to the $\mathbf{u}_y$ parameters is given by the following block-inverse formula:

$$\mathbf{S}^{-1} = \left(\mathbf{F} - \mathbf{E}^{\top}\mathbf{B}^{-1}\mathbf{E}\right)^{-1}$$

where $\mathbf{S} = \mathbf{F} - \mathbf{E}^{\top}\mathbf{B}^{-1}\mathbf{E}$ is the Schur complement of $\mathbf{B}$-block in $\mathbf{D}$.

Therefore:

$$\mathrm{PEV}\left(\mathbf{u}_y\right) = \mathbf{S}^{-1}$$

This Schur complement captures the conditional information about the young individuals after accounting for the training population and fixed effects through the term $\mathbf{E}^{\top}\mathbf{B}^{-1}\mathbf{E}$.

2.4 PEV for ungenotyped individuals

Using the Schur complement of $\mathbf{B}$ in the partitioned system and substituting the blocks for young and training individuals, the information matrix for $\mathbf{u}_y$ is

$$\mathbf{S} = \mathbf{G}^{yy} - \mathbf{G}^{yt}\mathbf{B}^{22}\mathbf{G}^{ty},$$

where $\mathbf{B}^{22}$ denotes the (u_t, u_t) block of $\mathbf{B}^{-1}$. Using the block matrix inversion formula (see Appendix 1 for derivation), $\mathbf{B}^{22}$ equals:

$$\mathbf{B}^{22} = \left[\left(\mathbf{Z}_t^{\top}\mathbf{R}^{-1}\mathbf{Z}_t + \mathbf{G}^{tt}\right) - \mathbf{Z}_t^{\top}\mathbf{R}^{-1}\mathbf{X}\left(\mathbf{X}^{\top}\mathbf{R}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{R}^{-1}\mathbf{Z}_t\right]^{-1}$$

Hence, the prediction error variance of the young individuals is

$$\mathrm{PEV}\left(\mathbf{u}_y\right) = \mathbf{S}^{-1} = \left[\mathbf{G}^{yy} - \mathbf{G}^{yt}\mathbf{B}^{22}\mathbf{G}^{ty}\right]^{-1} \quad (1)$$

2.4.1 Simplified PEV expressions

When phenotypic information from the training population is not utilized or is limited, simplified versions of Equation 1 can be derived. Three such cases are presented below, corresponding to scenarios where $\mathbf{X}$, $\mathbf{Z}$, and their cross-products from MME can be disregarded.

2.4.1.1 Case 1: no fixed effects

In many practical situations, fixed effect terms may be omitted from the model or their contribution may be negligible. When fixed effects are omitted, the $\mathbf{X}$-related blocks vanish from the mixed model equations. The coefficient matrix for $[\mathbf{u}_t^{\top}, \mathbf{u}_y^{\top}]^{\top}$ reduces to

$$\mathbf{B} = \mathbf{Z}_t^{\top}\mathbf{R}^{-1}\mathbf{Z}_t + \mathbf{G}^{tt}$$

Assuming $\mathbf{B}$ is nonsingular, the Schur complement of $\mathbf{B}$ in $\mathbf{D}$ is

$$\mathbf{S} = \mathbf{F} - \mathbf{E}^{\top}\mathbf{B}^{-1}\mathbf{E} = \mathbf{G}^{yy} - \mathbf{G}^{yt}\left(\mathbf{Z}_t^{\top}\mathbf{R}^{-1}\mathbf{Z}_t + \mathbf{G}^{tt}\right)^{-1}\mathbf{G}^{ty}.$$

Using Equation 1 as reference, the PEV for the young candidates is

$$\mathrm{PEV}\left(\mathbf{u}_y\right) = \mathbf{S}^{-1} = \left[\mathbf{G}^{yy} - \mathbf{G}^{yt}\left(\mathbf{Z}_t^{\top}\mathbf{R}^{-1}\mathbf{Z}_t + \mathbf{G}^{tt}\right)^{-1}\mathbf{G}^{ty}\right]^{-1}$$

2.4.1.2 Case 2: no phenotypic data

When phenotypic information is ignored, $\mathbf{Z}_t^{\top}\mathbf{R}^{-1}\mathbf{Z}_t$, $\mathbf{Z}_t^{\top}\mathbf{R}^{-1}\mathbf{Z}_y$, $\mathbf{Z}_t^{\top}\mathbf{R}^{-1}\mathbf{X}$, and $\mathbf{X}^{\top}\mathbf{R}^{-1}\mathbf{X}$ of $\mathbf{B}$ are zero. In this limiting case the $\mathbf{X}$ and $\mathbf{Z}$-related blocks disappear from the MME, and only the relationship block remains. Consequently, $\mathbf{B}$ reduces to $\mathbf{G}^{tt}$, and $\mathbf{B}^{22}$ simplifies to:

$$\mathbf{B}^{22} = \left(\mathbf{G}^{tt}\right)^{-1}$$

Substituting this into the PEV expression:

$$\mathrm{PEV}\left(\mathbf{u}_y\right) = \left[\mathbf{G}^{yy} - \mathbf{G}^{yt}\left(\mathbf{G}^{tt}\right)^{-1}\mathbf{G}^{ty}\right]^{-1}$$

2.4.1.3 Case 3: weak phenotypic information

When phenotypic information is weak relative to the genetic information (i.e., $\mathrm{tr}\left(\mathbf{Z}_t^{\top}\mathbf{R}^{-1}\mathbf{Z}_t\right)/\mathrm{tr}\left(\mathbf{G}^{tt}\right) \ll 1$) (Misztal and Wiggans, 1988; Meyer, 1989), the matrix $\mathbf{B}^{22}$ can be approximated as

$$\mathrm{PEV}\left(u_y\right) \approx \left[\mathbf{G}^{yy} - \mathbf{G}^{yt}\left(\mathbf{G}^{tt}\right)^{-1}\mathbf{G}^{ty}\right]^{-1}$$

This approximation applies when: (i) traits have very low heritability, or (ii) the training population has substantial missing phenotypes.

These simplified cases illustrate successive reductions in information available to the mixed model equations, with computational complexity decreasing as phenotypic data and fixed effects are removed.

2.4.1.4 Estimation of variance components

Across all three cases, the genetic covariance matrix requires both a relationship matrix and an estimate of the additive and residual variances. The relationship matrix $\mathbf{H}$ can be constructed using pedigree information alone (Henderson, 1976), genomic markers (VanRaden, 2008), or a combination of both through single-step methods (Aguilar et al., 2010; Christensen et al., 2012). The variance components σ_u^2 and σ_e^2 are typically estimated using restricted maximum likelihood (REML) (Patterson and Thompson, 1971) or Bayesian methods (Gianola et al., 2009) from available phenotypic and relationship data in the training population.

2.5 Simulation design for validation

To evaluate the similarity between the equations derived in this study and the direct inversion of the coefficient matrix of the MME, we designed a simulation scenario with six problem sizes to assess computational scalability. The simulation created populations ranging from 180 to 15,500 individuals, with varying proportions

TABLE 1 Population structure for each simulation scenario.

Component	S1	S2	S3	S4	S5	S6
Founders	30	50	100	150	300	500
Progeny	150	500	1,000	2,000	8,000	15,000
Training (genotyped)	100	300	600	1,200	5,000	8,000
Young candidates	30	100	200	400	1,500	3,000
Total individuals	180	550	1,100	2,150	8,300	15,500
SNP markers	5,000	8,000	10,000	12,000	15,000	20,000

Values represent the number of individuals in each category.

of genotyped individuals to reflect realistic breeding program structures (Table 1). Each population consisted of founders and their progeny, with young selection candidates representing the most recent generation. Founders ranged from 30 to 500 individuals across problem sizes, with 33% designated as sires and the remaining as dams. Progeny generations (150–15,000 individuals) were simulated with diverse family structures: 25% as full-sibs (sharing both parents), 35% as paternal half-sibs (sharing only sires), 25% as maternal half-sibs (sharing only dams), and 15% with only one known parent. Pedigree records were constructed by recording the sire and dam assignments for each progeny individual. Founders were assigned as unrelated base animals (i.e., with unknown parents). For each progeny individual, a sire was randomly sampled from the pool of male founders and a dam from the pool of female founders, according to the family structure proportions described above. Full-sibs shared both the same sire and dam, paternal half-sibs shared only the sire, and maternal half-sibs shared only the dam. Individuals with only one known parent had the other parent recorded as missing. The pedigree-based relationship matrix ($\mathbf{A}$) was then constructed using Henderson, (1976) method, and the $\mathbf{A}_{22}^{-1}$ submatrix corresponding to genotyped individuals was extracted for use in the single-step blending procedure. Young selection candidates (30–3,000 individuals) represented the final cohort and had only pedigree information available, whereas training populations (100–8,000 genotyped individuals) included founders and selected progeny characterized by genotypes, pedigree records, and phenotypes. Genotypes were simulated for 5,000 to 20,000 bi-allelic SNP markers with allele frequencies drawn from a uniform distribution. Marker genotypes for founders were randomly sampled assuming the Hardy-Weinberg equilibrium, while progeny genotypes followed Mendelian inheritance patterns based on parental genotypes when available. The genomic relationship matrix was constructed using VanRaden's method (VanRaden, 2008). Single-step GBLUP (Aguilar et al., 2010) was implemented by blending genomic matrix with the pedigree-based relationship matrix using weights of 0.95 and 0.05, respectively, to construct the $\mathbf{H}^{-1}$ matrix. Variance components were set to $\sigma_e^2 = 1.0$ and $\sigma_u^2 = 0.5$, yielding a heritability of 0.33. Each scenario was replicated five times with different random seeds to ensure robust timing estimates and numerical precision assessments. Further details on the data are available in the R code available on GitHub. Peak memory usage was measured for each method using R's garbage collection diagnostics, recording memory allocation before and after each computation to quantify the memory footprint of each approach. To validate the derived expressions, we compared the direct inversion of the full coefficient matrix of the MME versus the Schur complement approach derived in this study. All analyses were performed on an Apple M2 processor with 8 cores and 16 GB of RAM, running macOS 26.4.1 and R version 4.5.2.

3 Results and discussion

The approach presented in this study applies to breeding programs practicing genomic selection that must decide which individuals to genotype under budget constraints. While genotyping all selection candidates would be ideal, this remains economically unfeasible for many species and companies, as evidenced by animal breeding programs that routinely genotype only a fraction of top-performance individuals. Genotyping strategies in commercial breeding programs are often closely guarded as part of competitive business plans. Although animal and crop commercial breeding programs continue to allocate increasing budgets to genotyping, strategic resource allocation decisions remain a key aspect when evaluated from a return on investment perspective, particularly as breeding programs scale and must balance genotyping comprehensiveness against other operational priorities.

Although several studies have addressed training set optimization using quantitative genetics approaches (Akdemir et al., 2015; Fernández-González et al., 2023), to our knowledge, this is the first study to address genotyping decisions under resource constraints in practical breeding programs. The derivations provide quantitative genetic tools to address these resource-allocation challenges by placing the genetic evaluation model at the center of genotyping decisions. The method exploits genetic relationships between training and ungenotyped individuals to calculate PEV using only relationship matrices and variance components, thereby eliminating the need for actual phenotypic data from ungenotyped individuals. The explicit PEV expressions enable prospective evaluation of prediction uncertainty under different genotyping scenarios, establishing the mathematical foundation for transitioning from experience-based to quantitatively informed genotyping decisions.

Genotyping decisions directly affect selection accuracy and genetic gain (Hayes et al., 2009; Rocha et al., 2025), with important implications for the long-term sustainability and effectiveness of genomic selection programs. The prospective evaluation framework developed here naturally extends beyond traditional genotyping to emerging sequencing technologies (Sthapit et al., 2025), where breeding programs face analogous resource allocation decisions about optimal sequencing depth and coverage. From a crop breeding lens, the approach is valuable in modern crop breeding programs that have adopted schemes with shorter recombination cycles (Gaynor et al., 2017; Gorjanc et al., 2018), where outbred parents have become the norm rather than traditional inbred lines. As newer genotyping platforms become increasingly cost-effective, the method can inform strategic decisions about which individuals to sequence at higher densities, thereby improving imputation accuracy across selection candidates and optimizing resource allocation within breeding pipelines.

As expected, PEV estimated using Equation 1 in the simulations produced results numerically identical to those obtained by direct inversion of the MME across all six scenarios (Table 1), confirming the mathematical equivalence of the derivations. Figure 1 presents the computational performance of all methods evaluated across scenarios ranging from 180 to 15,500 individuals. The Schur complement method consistently outperformed direct inversion of the full MME coefficient matrix, with the computational advantage increasing at larger population sizes from approximately 20% reduction in computation time at S3 ($n = 1,100$) to 44% at S6 ($n = 15,500$; 55.5 versus 31.3 s). Among the simplified cases, progressive removal of model components yielded corresponding reductions in computation

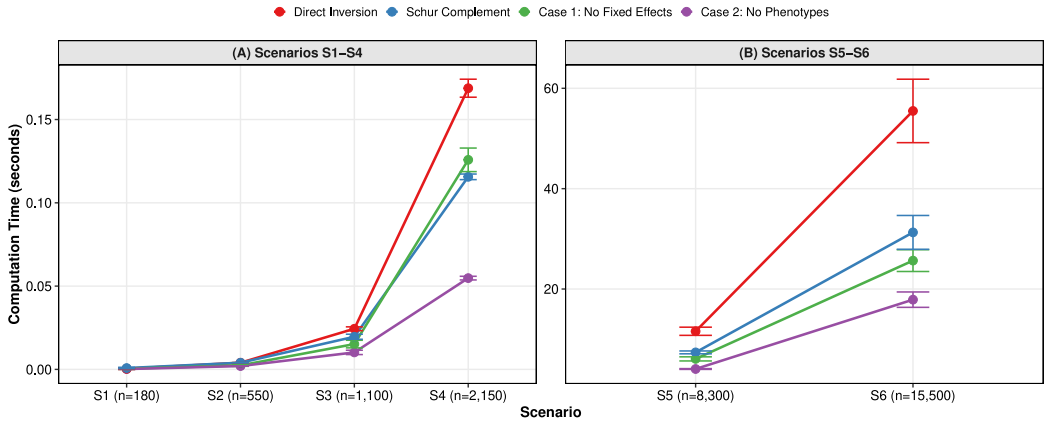

FIGURE 1
Computation time (in seconds) for prediction error variance (PEV) estimation across scenarios of increasing population size. **(A)** shows scenarios S1–S4 and **(B)** shows scenarios S5–S6, with independent y-axis scales to visualize differences across all methods. Results are shown for four methods: direct inversion of the full mixed model equations (MME) coefficient matrix, the Schur complement method derived in this study, and simplified Cases 1 and 2 (Case 1: no fixed effects; Case 2: no phenotypic data). Error bars represent standard deviations across five replicates.

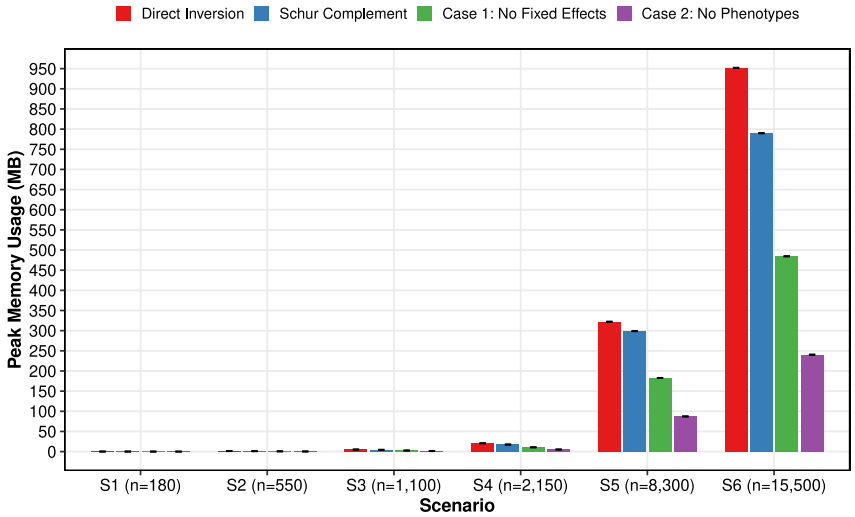

FIGURE 2
Peak memory usage (in MB) for prediction error variance (PEV) estimation across scenarios of increasing population size. Results are shown for four methods: direct inversion of the full MME coefficient matrix, the Schur complement method, and simplified Cases 1 and 2.

time. Case 1 (no fixed effects) reduced computation by eliminating the absorption of fixed effect equations, and Case 2 (no phenotypic data) further reduced the burden by removing all data-dependent cross-product terms. At the largest scale (S6), Case 2 required only 17.9 s compared to 55.5 s for direct inversion. In addition to computation time, peak memory usage was monitored across all scenarios (Figure 2). The Schur complement method required less memory than direct MME inversion across all scenarios, with the difference becoming more pronounced at larger population sizes. At S6, direct inversion required approximately 952 MB compared to 790 MB for the Schur complement, while Case 2 required only 240 MB. Furthermore, the integration of dimensionality reduction techniques such as the Algorithm for Proven and Young (APY) (Misztal et al., 2015) or eigen decomposition methods could yield additional improvements in computational efficiency, particularly for problems involving dense genomic relationship matrices where matrix operations, especially inversion (Junqueira et al., 2022), dominate the computational burden.

3.1 Application to breeding programs

Although practical implementation was beyond the scope of this study, we briefly outline the implementation of the method. A comprehensive evaluation of the method's performance will be presented in a subsequent publication.

Prediction error variance has been well established since the 1970s (Henderson, 1975; Henderson, 1976), and the formulas derived in this study follow this same theoretical foundation.

Estimating PEV for ungenotyped and unphenotyped individuals is straightforward as it is a function only of parents' PEV. The benefits of applying the equations presented in this study come when integrating it with simulating progeny genotypes based on parental genomic information, which is a process extensively documented in the literature (Bernardo, 2014; Mohammadi et al., 2015; Pook et al., 2020; Gaynor et al., 2021). These simulated genotypes can then be used to construct the relationship matrices $\mathbf{G}_{yt}$ and $\mathbf{G}_{yy}$ to capture Mendelian sampling variation and genomic similarity to the training population, moving beyond simple pedigree-based relationships. These simulated genotypes are then incorporated into Equation 1 to calculate PEV values for virtual progeny. Although the actual genotypes of future progeny remain unknown, simulating multiple potential progeny from each mating provides insight into the range of genotypes that could arise. Consequently, the average PEV across simulated progeny offers a reliable approximation of expected prediction uncertainty for offspring from specific matings, enabling prospective evaluation of genotyping strategies before progeny are generated.

The selection of which individuals to genotype can be formulated as a combinatorial linear or non-linear optimization problem (Chopard and Tomassini, 2018) with an appropriate objective function subject to budgetary and operational constraints. Several alternative objective functions can be formulated for optimizing genotyping decisions. For example, one straightforward approach is to minimize the average PEV across all ungenotyped individuals estimated using Equation 1. Although computationally feasible, this objective may not represent the optimal strategy for practical breeding programs, as it tends to prioritize genotyping individuals most closely related to the training population. While such individuals benefit from lower prediction uncertainty, this strategy creates a genetically narrow training population that inadequately represents diversity across family structures. This leads to biased predictions that favor closely related individuals at the expense of prediction accuracy for underrepresented families, ultimately compromising both short-term selection decisions and long-term genetic progress. This concern is particularly relevant for breeding programs with larger effective population sizes (Ne), where greater genetic diversity and more complex family structures necessitate broader representation across genetic backgrounds in the training population (Pocrnic et al., 2016). Alternative objective functions could address these limitations by incorporating diversity constraints, such as penalizing excessive relatedness among selected individuals or explicitly balancing prediction error variance against maintenance of genetic diversity within the genotyped subset. Such multi-objective optimization problems can be solved using various algorithms, including differential evolution, integer programming, or simulated annealing.

While this study focuses on genotyping decisions, the framework extends naturally to other resource allocation problems in breeding programs, as Pocrnic et al. (Pocrnic et al., 2022) have discussed. Beyond core subset optimization for APY, Pocrnic et al. (2022) identified several applications following the same underlying rationale. These include prioritizing individuals for high-density genotyping or sequencing, designing selective phenotyping strategies where phenotyping resources are limited, and managing genetic diversity in genebanks. The present study provides an analogous framework for the upstream genotyping decision problem. Whether the objective is selecting candidates for genotyping, constructing a computational core, designing a phenotyping panel, or managing genetic collections, the method demonstrates that systematic evaluation of prediction uncertainty is broadly applicable across diverse breeding program objectives where resource constraints necessitate strategic allocation decisions.

The PEV expressions derived in this study enable breeding programs to quantify prediction uncertainty for ungenotyped individuals before making genotyping decisions. The method calculates explicit PEV values for each individual based on their relationships to the training population, providing a mathematical measure of how much uncertainty would remain in their predicted breeding values if left ungenotyped. For example, individuals with stronger genetic connections to the training population through genetic relationships will exhibit lower PEV values, indicating higher expected prediction accuracy. Conversely, candidates from novel genetic backgrounds or with weak training population connectivity will show higher PEV values, signaling greater prediction uncertainty that could potentially be reduced through genotyping.

The three simplified cases derived in this study provide practical tools for different breeding scenarios. Case 1 (no fixed effects) simplifies computations when fixed effects are negligible or when working within homogeneous populations, reducing computational burden without sacrificing accuracy. Case 2 (no phenotypic data) establishes the baseline prediction accuracy achievable from relationships alone, particularly useful for pre-phenotyping decisions and understanding the fundamental contribution of pedigree structure to prediction accuracy. Case 3 (weak phenotypic information) identifies scenarios where phenotypic data collection provides minimal improvement, particularly relevant for low-heritability traits or when training individuals have limited phenotypic records. Together, these cases enable rapid evaluation of prediction uncertainty across different data availability scenarios, helping breeders make informed decisions about resource allocation between genotyping and phenotyping strategies while maintaining computational efficiency when evaluating large numbers of selection candidates.

The reliability of PEV calculations depends mainly on the quality and representativeness of the training population data. Well-established programs with extensive training populations spanning multiple generations provide comprehensive phenotypic and genomic data, enabling accurate variance component estimation that underlies reliable PEV calculations (Pszczola et al., 2012; Misztal et al., 2014). Programs with limited training data or poor variance component estimates may experience reduced accuracy in PEV calculations, potentially leading to suboptimal inferences about prediction uncertainty (Clark et al., 2012). The method accommodates complex population structures and diverse relationship matrix configurations, but the accuracy of the resulting PEV estimates remains fundamentally dependent on the underlying genetic and statistical assumptions being met in practice.

Ultimately, the success of a breeding program depends on its profitability. Genetic progress must be aligned with costs and investments. Integrating this method into an optimization model can enhance the long-term sustainability of genetic evaluation and guide strategic genotyping decisions, directing investments toward

areas that deliver the greatest value to breeding programs while considering operational and cost constraints.

3.2 Computational considerations

The computational requirements for evaluating the derived PEV expressions are dominated by operations involving the genomic relationship matrix, which is dense and computationally demanding, unlike the sparse design matrices $\mathbf{X}$ and $\mathbf{Z}$ that characterize fixed effects and phenotypic data structure (Junqueira et al., 2022). The core computational bottleneck lies in calculating the Schur complement $\mathbf{G}^{yy} - \mathbf{G}^{yt}\mathbf{B}^{22}\mathbf{G}^{ty}$, where the dense genomic blocks require $O(n_t^3 + n_y n_t^2 + n_y^2 n_t)$ operations per evaluation. While design matrix operations benefit from sparsity and can be computed efficiently, the genomic relationship matrix operations cannot exploit such structural advantages (Misztal et al., 2015).

Matrix decomposition techniques offer computational advantages by exploiting the structure of the genetic covariance matrix and avoiding repeated expensive matrix operations (Strandén and Garrick, 2009; Misztal et al., 2015; Misztal, 2016). For large-scale applications, approximate methods that further exploit matrix structure are required for computational tractability. Two primary approaches can address these challenges: dimensionality reduction and hierarchical approximation.

A common dimensionality-reduction strategy is to approximate the genomic relationship matrix using a principal component (eigen) decomposition (Akdemir et al., 2015; Ødegård et al., 2018), retaining only the leading components that capture most of the variance. This low-rank approximation can then substitute the full matrix, substantially reducing computational cost while preserving the majority of the information content.

Another approach to reduce computational complexity in PEV calculations involves the Algorithm for Proven and Young (APY) methodology (Misztal et al., 2015). APY maintains a core subset of the most informative training individuals while approximating relationships for the remaining population, thereby reducing matrix dimensions in the genetic covariance structure without substantial loss of information. The selection of the core population is a critical step that directly affects the quality of the approximation. Several strategies have been proposed for core selection, including random sampling (Misztal et al., 2015), selection based on maximizing genetic diversity using algorithms such as those described by Pocrnic et al. (2016), and choosing individuals that maximize the number of independent chromosome segments represented in the core (Pocrnic et al., 2016). The optimal core size is related to the effective population size (N_e) of the breed or population, with the number of independent chromosome segments approximated as $4N_eL$, where L is the genome length in Morgans (Strandberg, 2006; Pocrnic et al., 2016). In practice, random selection of core animals has been shown to perform well when the core size is sufficiently large relative to the number of independent chromosome segments (Misztal et al., 2015; Bradford et al., 2017). For the PEV framework presented here, the core population would ideally be selected from the training set to maximize representation of the genetic diversity present in the breeding population, ensuring that the approximated relationship matrix adequately captures the covariance structure between training and young individuals. When applied to PEV derivations, APY enables the decomposition of the relationship matrix into a computationally manageable core component and an approximated remainder. This hierarchical matrix structure allows the Schur complement calculations central to PEV estimation to operate on reduced dimensions. The APY core captures the primary genetic relationships within the training population, while the PEV expressions evaluate prediction uncertainty for young candidates based on their connections to this genetic foundation rather than the full training matrix.

3.3 Limitations and assumptions

The prediction error variance method developed in this study relies on several key assumptions that, when violated, may compromise the accuracy and reliability of the derived expressions. Understanding these limitations is important for the appropriate application of the methodology in breeding programs.

3.3.1 Variance component estimation

The PEV calculations assume that variance components are known with certainty. In practice, variance components are estimated from data with associated sampling uncertainty (Junqueira et al., 2017), using either frequentist or Bayesian methods. Misspecification of variance components (Junqueira et al., 2022) directly propagates through the PEV expressions (Meyer, 1989), potentially leading to systematic over- or underestimation of prediction uncertainty. Consequently, theoretical PEV values and realized prediction errors may diverge substantially when variance component estimates are inaccurate. The method presented in this study does not account for this estimation uncertainty, treating variance components as known quantities.

3.3.2 Relationship matrix accuracy

The derivations assume that the relationship matrix $\mathbf{G}$ accurately captures genetic covariances among individuals. However, both pedigree-based and genomic relationship matrices are subject to errors. Pedigree errors, including misidentified parentage, unknown parents, or incomplete genealogies, directly distort the covariance structure (Junqueira et al., 2017) and therefore affect PEV calculations. For genomic relationship matrices, genotyping errors, poor marker coverage, or inadequate representation of causal variants can lead to misestimation of realized relationships (Christensen and Lund, 2010). Single-step methods that combine pedigree and genomic information inherit errors from both sources. Since PEV calculations depend fundamentally on the accuracy of genetic covariances between training and young populations through the term $\mathbf{G}^{yt}\mathbf{B}^{22}\mathbf{G}^{ty}$, relationship matrix errors can impact the reliability of prediction uncertainty estimates.

3.3.3 Model specifications

The derivations of this study explicitly model only additive effects. Although non-additive genetic effects (i.e., dominance and epistasis), are not captured in the presented derivation, this approach can be extended to incorporate non-additive relationship matrices using the same Schur complement logic. For traits where non-additive effects

contribute substantially to genetic variance, the PEV expressions will underestimate total prediction uncertainty by failing to account for these unmodeled variance components (Vitezica et al., 2013). The magnitude of this underestimation depends on the proportion of genetic variance attributable to non-additive effects and the extent to which non-additive relationships between training and ungenotyped individuals differ from additive patterns (Muñoz et al., 2014). While single-step genomic BLUP can accommodate dominance or epistatic relationship matrices (Vitezica et al., 2017; Varona et al., 2018), implementing such extensions requires constructing appropriate non-additive relationship matrices and partitioning their contributions within the mixed model equations. Breeding programs selecting for traits with substantial non-additive genetic effect, such as fitness-related traits in livestock or heterosis-dependent traits in crops, should recognize that additive-only PEV calculations may provide optimistic estimates of prediction accuracy.

3.3.4 Absence of selection bias

When relationship matrices incorporate properly defined founder groups or genetic groups to account for population stratification and temporal trends in genetic merit, many systematic biases from historical selection can be mitigated (Misztal et al., 2013; Junqueira et al., 2020). However, within-cohort selection effects and reductions in genetic variance due to linkage disequilibrium (Bulmer effect) remain unaccounted for in standard relationship matrix formulations (Bulmer, 1971). Additionally, when young selection candidates experience substantially different selection intensities or breeding objectives than the training population, the genetic covariances in $\mathbf{G}^{ty}$ may not fully represent prediction relationships, potentially affecting PEV estimates (Habier et al., 2007; Patry and Ducrocq, 2011). While the method accommodates genetic groups within the relationship matrix structure, residual selection bias in variance component estimates could still impact PEV calculations in populations under intensive selection.

3.3.5 Static genetic architecture

The derivations assume a constant genetic architecture across environments and over time. Genotype-by-environment interactions (Jarquín et al., 2014), changes in allele effects across generations, or evolution of the genetic background can alter prediction accuracy in ways not captured by static PEV calculations. Breeding programs operating across diverse environments or planning long-term selection strategies should recognize that PEV values calculated under current conditions may not accurately reflect future prediction uncertainty.

4 Conclusion

This study presents the equations for computing prediction error variance in ungenotyped individuals under mixed model equations, providing breeding programs with a quantitative foundation for genotyping allocation decisions. The derivation of explicit PEV expressions under different relationship matrix structures enables prospective evaluation of genotyping strategies without requiring actual phenotypic or genomic data ungenotyped individuals. By applying Schur complement methods to partitioned populations, the method accommodates complex breeding program structures while maintaining computational tractability.

The presented mathematical foundation provides the theoretical basis to transition genotyping decisions from experience-based approaches to optimization problems with well-defined objective functions. The PEV expressions reveal how genetic relationships between training populations and ungenotyped individuals directly influence prediction accuracy, providing mathematical indicators to address resource allocation challenges of modern breeding programs where comprehensive genotyping remains economically prohibitive. While implementation requires consideration of computational trade-offs and potential limitations, the framework provides a rigorous mathematical basis for the optimization of limited genotyping resources across diverse plant and animal breeding programs. Future research can build on the presented PEV expressions to implement optimization algorithms and further evaluate the practical benefits and limitations of the approach.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

VJ: Conceptualization, Writing – original draft, Investigation, Resources, Software, Supervision, Formal Analysis, Validation, Methodology, Writing – review and editing. MJ-IY: Writing – original draft, Writing – review and editing. FC: Writing – review and editing, Methodology, Writing – original draft.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This publication fee of this study was supported by Bayer Crop Science. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

Acknowledgements

The authors thank Daniela Lourenço, Christina Lehermeier, Paul Nelson, Simon Teyssèdre, and R. Chris Gaynor for their valuable suggestions during the preparation of the manuscript.

Conflict of interest

Author VJ was employed by Bayer R&D, Bayer Crop Science.

The remaining author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. Generative AI was used for language and grammar checking.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., and Lawlor, T. (2010). Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of holstein final score. *J. Dairy Sci.* 93, 743–752. doi:10.3168/jds.2009-2730

Akdemir, D., and Isidro-Sánchez, J. (2019). Design of training populations for selective phenotyping in genomic prediction. *Sci. Rep.* 9, 1446. doi:10.1038/s41598-018-38081-6

Akdemir, D., Sanchez, J. I., and Jannink, J.-L. (2015). Optimization of genomic selection training populations with a genetic algorithm. *Genet. Sel. Evol.* 47, 38. doi:10.1186/s12711-015-0116-6

Bernardo, R. (2014). Genomewide selection of parental inbreds: classes of loci and virtual biparental populations. *Crop Sci.* 54, 2586–2595. doi:10.2135/cropsci2014.01.0088

Bradford, H., Pocrnić, I., Fragomeni, B., Lourenco, D., and Misztal, I. (2017). Selection of core animals in the algorithm for proven and young using a simulation model. *J. Animal Breed. Genet.* 134, 545–552. doi:10.1111/jbg.12276

Bulmer, M. (1971). The effect of selection on genetic variability. *Am. Nat.* 105, 201–211. doi:10.1086/282718

Chopard, B., and Tomassini, M. (2018). *An Introduction to Metaheuristics for Optimization*, 226. Springer.

Christensen, O. F., and Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol.* 42, 2. doi:10.1186/1297-9686-42-2

Christensen, O. F., Madsen, P., Nielsen, B., Ostersen, T., and Su, G. (2012). Single-step methods for genomic evaluation in pigs. *Animal* 6, 1565–1571. doi:10.1017/S1751731112000742

Clark, S. A., Hickey, J. M., Daetwyler, H. D., and van der Werf, J. H. (2012). The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genet. Sel. Evol.* 44, 4. doi:10.1186/1297-9686-44-4

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., De Los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* 22, 961–975. doi:10.1016/j.tplants.2017.08.011

Fernández-González, J., Akdemir, D., and Isidro y Sánchez, J. (2023). A comparison of methods for training population optimization in genomic selection. *Theor. Appl. Genet.* 136, 30. doi:10.1007/s00122-023-04265-6

Garrick, D. J. (2011). The nature, scope and impact of genomic prediction in beef cattle in the United States. *Genet. Sel. Evol.* 43, 17. doi:10.1186/1297-9686-43-17

Gaynor, R. C., Gorjanc, G., Bentley, A. R., Ober, E. S., Howell, P., Jackson, R., et al. (2017). A two-part strategy for using genomic selection to develop inbred lines. *Crop Sci.* 57, 2372–2386. doi:10.2135/cropsci2016.09.0742

Gaynor, R. C., Gorjanc, G., and Hickey, J. M. (2021). Alphasimr: an r package for breeding program simulations. *G3* 11, jkaa017. doi:10.1093/g3journal/jkaa017

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E., and Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics* 183, 347–363. doi:10.1534/genetics.109.103952

Gorjanc, G., Gaynor, R. C., and Hickey, J. M. (2018). Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. *Theor. Applied Genetics* 131, 1953–1966. doi:10.1007/s00122-018-3125-3

Guinan, F., Wiggans, G., Norman, H., Dürr, J., Cole, J., Van Tassell, C., et al. (2023). Changes in genetic trends in us dairy cattle since the implementation of genomic selection. *J. Dairy Sci.* 106, 1110–1129. doi:10.3168/jds.2022-22205

Habier, D., Fernando, R. L., and Dekkers, J. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177, 2389–2397. doi:10.1534/genetics.107.081190

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited review: genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* 92, 433–443. doi:10.3168/jds.2008-1646

He, S., Schulthess, A. W., Mirdita, V., Zhao, Y., Korzun, V., Bothe, R., et al. (2016). Genomic selection in a commercial winter wheat population. *Theor. Appl. Genet.* 129, 641–651. doi:10.1007/s00122-015-2655-1

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31, 423–447. doi:10.2307/2529430

Henderson, C. R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* 32, 69–83. doi:10.2307/2529339

Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding*, 462. Guelph, Canada: University of Guelph.

Isidro, J., Jannink, J.-L., Akdemir, D., Poland, J., Heslot, N., and Sorrells, M. E. (2015). Training set optimization under population structure in genomic selection. *Theor. Appl. Genet.* 128, 145–158. doi:10.1007/s00122-014-2418-4

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., et al. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor. Appl. Genet.* 127, 595–607. doi:10.1007/s00122-013-2243-1

Junqueira, V. S., Cardoso, F. F., Oliveira, M. M., Sollero, B. P., Silva, F. F., and Lopes, P. S. (2017). Use of molecular markers to improve relationship information in the genetic evaluation of beef cattle tick resistance under pedigree-based models. *J. Animal Breed. Genet.* 134, 14–26. doi:10.1111/jbg.12239

Junqueira, V. S., Lopes, P. S., Lourenco, D., Silva, F. F., and Cardoso, F. F. (2020). Applying the metafounders approach for genomic evaluation in a multibreed beef cattle population. *Front. Genet.* 11, 556399. doi:10.3389/fgene.2020.556399

Junqueira, V. S., Lourenco, D., Masuda, Y., Cardoso, F. F., Lopes, P. S., Silva, F. F., et al. (2022). Is single-step genomic reml with the algorithm for proven and young more computationally efficient when less generations of data are present? *J. Animal Sci.* 100, skac082. doi:10.1093/jas/skac082

Lopez-Cruz, M., and De Los Campos, G. (2021). Optimal breeding-value prediction using a sparse selection index. *Genetics* 218, iyab030. doi:10.1093/genetics/iyab030

Meuwissen, T. H., Hayes, B. J., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi:10.1093/genetics/157.4.1819

Meyer, K. (1989). Approximate accuracy of genetic evaluation under an animal model. *Livest. Prod. Sci.* 21, 87–100. doi:10.1016/0301-6226(89)90041-9

Misztal, I. (2016). Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. *Genetics* 202, 401–409. doi:10.1534/genetics.115.182089

Misztal, I., and Wiggans, G. (1988). Approximation of prediction error variance in large-scale animal models. *J. Dairy Sci.* 71, 27–32. doi:10.1016/s0022-0302(88)79976-2

Misztal, I., Vitezica, Z.-G., Legarra, A., Aguilar, I., and Swan, A. (2013). Unknown-parent groups in single-step genomic evaluation. *J. Animal Breed. Genet.* 130, 252–258. doi:10.1111/jbg.12025

Misztal, I., Legarra, A., and Aguilar, I. (2014). Using recursion to compute the inverse of the genomic relationship matrix. *J. Dairy Sci.* 97, 3943–3952. doi:10.3168/jds.2013-7752

Misztal, I., Fragomeni, B. O., Lourenco, D. A., Tsuruta, S., Masuda, Y., Aguilar, I., et al. (2015). *Efficient Inversion of Genomic Relationship Matrix by the Algorithm for Proven and Young (APY)*. Orlando, Unites States: Interbull Bulletin.

Mohammadi, M., Tiede, T., and Smith, K. P. (2015). Popvar: a genome-wide procedure for predicting genetic variance and correlated response in biparental breeding populations. *Crop Sci.* 55, 2068–2077. doi:10.2135/cropsci2015.01.0030

Muñoz, P. R., Resende Jr, M. F., Gezan, S. A., Resende, M. D. V., de Los Campos, G., Kirst, M., et al. (2014). Unraveling additive from nonadditive effects using genomic relationship matrices. *Genetics* 198, 1759–1768. doi:10.1534/genetics.114.171322

Ødegård, J., Indahl, U., Strandén, I., and Meuwissen, T. H. (2018). Large-scale genomic prediction using singular value decomposition of the genotype matrix. *Genet. Sel. Evol.* 50, 6. doi:10.1186/s12711-018-0373-2

Ouellette, D. V. (1981). Schur complements and statistics. *Linear Algebra Its Appl.* 36, 187–295. doi:10.1016/0024-3795(81)90232-9

Patry, C., and Ducrocq, V. (2011). Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. *J. Dairy Sci.* 94, 1011–1020. doi:10.3168/jds.2010-3804

Patterson, H. D., and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* 58, 545–554. doi:10.1093/biomet/58.3.545

Pocrnic, I., Lourenco, D. A., Masuda, Y., Legarra, A., and Misztal, I. (2016). The dimensionality of genomic information and its effect on genomic prediction. *Genetics* 203, 573–581. doi:10.1534/genetics.116.187013

Pocrnic, I., Lindgren, F., Tolhurst, D., Herring, W. O., and Gorjanc, G. (2022). Optimisation of the core subset for the apy approximation of genomic relationships. *Genet. Sel. Evol.* 54, 76. doi:10.1186/s12711-022-00767-x

Pook, T., Schlather, M., and Simianer, H. (2020). Mobps-modular breeding program simulator. *G3 Genes, Genomes, Genet.* 10, 1915–1918. doi:10.1534/g3.120.401193

Pszczola, M., Strabel, T., Mulder, H., and Calus, M. (2012). Reliability of direct genomic values for animals with different relationships within and to the reference population. *J. Dairy Sci.* 95, 389–400. doi:10.3168/jds.2011-4338

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., et al. (2012). Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays l.). *Genetics* 192, 715–728. doi:10.1534/genetics.112.141473

Rocha, A. O., Gloria, L. S., Araujo, A. C., Wen, H., Wilson, C. S., Freking, B. A., et al. (2025). Genotyping strategies for single-step genomic predictions in a simulated sheep population under different scenarios of pedigree error types. *Front. Genet.* 16, 1697103. doi:10.3389/fgene.2025.1697103

Sharif-Islam, M., van Der Werf, J. H., Wood, B. J., and Hermesch, S. (2024). The predicted benefits of genomic selection on pig breeding objectives. *J. Animal Breed. Genet.* 141, 685–701. doi:10.1111/jbg.12873

Sthapit, S. R., Crain, J., Larson, S., Anderson, J. A., Bajgain, P., DeHaan, L. R., et al. (2025). A low-coverage skim-sequencing and imputation pipeline for genomic selection. *Plant Genome* 18, e70139. doi:10.1002/tpg2.70139

Strandberg, E. (2006). "Estimation of the number of independent chromosome segments," in Proceedings of the 8th World Congress on Genetics Applied to Livestock Production.

Strandén, I., and Garrick, D. (2009). Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *J. Dairy Sci.* 92, 2971–2975. doi:10.3168/jds.2008-1929

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi:10.3168/jds.2007-0980

Varona, L., Legarra, A., Toro, M. A., and Vitezica, Z. G. (2018). Non-additive effects in genomic selection. *Front. Genet.* 9, 78. doi:10.3389/fgene.2018.00078

Vitezica, Z. G., Varona, L., and Legarra, A. (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics* 195, 1223–1230. doi:10.1534/genetics.113.155176

Vitezica, Z. G., Legarra, A., Toro, M. A., and Varona, L. (2017). Orthogonal estimates of variances for additive, dominance, and epistatic effects in populations. *Genetics* 206, 1297–1307. doi:10.1534/genetics.116.199406

Wiggans, G. R., Cole, J. B., Hubbard, S. M., and Sonstegard, T. S. (2017). Genomic selection in dairy cattle: the usda experience. *Annu. Rev. Animal Biosci.* 5, 309–327. doi:10.1146/annurev-animal-021815-111422

Appendix 1

Derivation of $\mathbf{B}^{22}$

To derive the explicit expression for $\mathbf{B}^{22}$, we apply the block matrix inversion formula to a general 2×2 block matrix following Ouellette (1981).

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}.$$

Assuming $\mathbf{A}$ is nonsingular and $\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ is invertible, let $\mathbf{S} = \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$. The inverse is

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{S}^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{S}^{-1} \\ -\mathbf{S}^{-1}\mathbf{C}\mathbf{A}^{-1} & \mathbf{S}^{-1} \end{bmatrix},$$

so the (2,2) block is

$$\mathbf{M}^{22} = \mathbf{S}^{-1} = \left(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}.$$

Application to the block $\mathbf{B}$

In the main text, the block $\mathbf{B}$ (corresponding to $(\boldsymbol{\beta}, \mathbf{u}_t)$) is

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^\top\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^\top\mathbf{R}^{-1}\mathbf{Z}_t \\ \mathbf{Z}_t^\top\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_t^\top\mathbf{R}^{-1}\mathbf{Z}_t + \mathbf{G}^{tt} \end{bmatrix}.$$

$$\begin{aligned} \mathbf{B}_{11} &= \mathbf{X}^\top\mathbf{R}^{-1}\mathbf{X} \\ \mathbf{B}_{12} &= \mathbf{X}^\top\mathbf{R}^{-1}\mathbf{Z}_t \\ \mathbf{B}_{21} &= \mathbf{Z}_t^\top\mathbf{R}^{-1}\mathbf{X} \\ \mathbf{B}_{22} &= \mathbf{Z}_t^\top\mathbf{R}^{-1}\mathbf{Z}_t + \mathbf{G}^{tt} \end{aligned}$$

Step-by-step derivation

Step 1: Compute $\mathbf{B}_{11}^{-1}$

$$\mathbf{B}_{11}^{-1} = \left(\mathbf{X}^\top\mathbf{R}^{-1}\mathbf{X}\right)^{-1}.$$

Step 2: Compute $\mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}$

$$\mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{B}_{12} = \mathbf{Z}_t^\top\mathbf{R}^{-1}\mathbf{X}\left(\mathbf{X}^\top\mathbf{R}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{R}^{-1}\mathbf{Z}_t$$

Step 3: Apply the Schur Complement formula for $\mathbf{B}^{22}$

Using the block-inverse formula,

$$\mathbf{B}^{22} = \left(\mathbf{B}_{22} - \mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\right)^{-1},$$

and substituting the expressions above yields

$$\mathbf{B}^{22} = \left(\left(\mathbf{Z}_t^\top\mathbf{R}^{-1}\mathbf{Z}_t + \mathbf{G}^{tt}\right) - \mathbf{Z}_t^\top\mathbf{R}^{-1}\mathbf{X}\left(\mathbf{X}^\top\mathbf{R}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{R}^{-1}\mathbf{Z}_t\right)^{-1}$$