On the equivalence between marker effect models and breeding value models and direct genomic values with the Algorithm for Proven and Young

Matias Bermann, Daniela Lourenco, Natalia S Forneris, Andres Legarra, Ignacy Misztal

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

On the equivalence between marker effect models and breeding value models and direct genomic values with the Algorithm for Proven and Young

Matias Bermann[1*], Daniela Lourenco[1], Natalia S. Forneris[2,3], Andres Legarra[4] and Ignacy Misztal[1]

Abstract

Background: Single-step genomic predictions obtained from a breeding value model require calculating the inverse of the genomic relationship matrix ($\mathbf{G}^{-1}$). The Algorithm for Proven and Young (APY) creates a sparse representation of $\mathbf{G}^{-1}$ with a low computational cost. APY consists of selecting a group of core animals and expressing the breeding values of the remaining animals as a linear combination of those from the core animals plus an error term. The objectives of this study were to: (1) extend APY to marker effects models; (2) derive equations for marker effect estimates when APY is used for breeding value models, and (3) show the implication of selecting a specific group of core animals in terms of a marker effects model.

Results: We derived a family of marker effects models called APY-SNP-BLUP. It differs from the classic marker effects model in that the row space of the genotype matrix is reduced and an error term is fitted for non-core animals. We derived formulas for marker effect estimates that take this error term in account. The prediction error variance (PEV) of the marker effect estimates depends on the PEV for core animals but not directly on the PEV of the non-core animals. We extended the APY-SNP-BLUP to include a residual polygenic effect and accommodate non-genotyped animals. We show that selecting a specific group of core animals is equivalent to select a subspace of the row space of the genotype matrix. As the number of core animals increases, subspaces corresponding to different sets of core animals tend to overlap, showing that random selection of core animals is algebraically justified.

Conclusions: The APY-(ss)GBLUP models can be expressed in terms of marker effect models. When the number of core animals is equal to the rank of the genotype matrix, APY-SNP-BLUP is identical to the classic marker effects model. If the number of core animals is less than the rank of the genotype matrix, genotypes for non-core animals are imputed as a linear combination of the genotypes of the core animals. For estimating SNP effects, only relationships and estimated breeding values for core animals are needed.

Background

Genomic predictions can be obtained from either a breeding value model or a marker effects model [1]. The equivalence between these two models [2] facilitates interpretation in terms of either marker effects or genomic relationships between individuals. On the one hand, from a computational point of view, breeding value models fit a random effect, with the number of levels

*Correspondence: mbermann@uga.edu

[1] Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602, USA
Full list of author information is available at the end of the article

equal to the number of genotyped animals (n_{gt}). On the other hand, marker effects models fit a random effect, with the number of levels equal to the number of markers (m). To obtain predictions for the first type of models using Henderson's mixed model equations (MME), the inverse of the genomic relationship matrix ($\mathbf{G}$) is needed [3]. However, direct inversion of $\mathbf{G}$ when n_{gt} is large is very expensive or even not feasible—for instance, genomic evaluation data for US dairy cattle data contain millions of genotyped animals. In contrast, marker effect models do not suffer from this constraint because the size of the block of equations concerning marker effects remains constant. Besides this advantage, marker effect models require multiplications of dense matrices and the condition number (i.e. the quotient between the largest and smallest eigenvalue of the MME) of the system is larger than for the breeding value models [4]. To handle these issues, complex solvers [5] and refined convergence criteria [6] are needed.

Increasing numbers of animals are genotyped every year but the number of markers (m) remains constant at around 50K [7], and therefore the rank of $\mathbf{G}$ is (at most) 50K. To make the breeding value models feasible with large n_{gt}, several strategies have been proposed. Misztal et al. [8] and Misztal [9] developed the Algorithm for Proven and Young (APY), which uses a sparse representation of $\mathbf{G}^{-1}$. The APY consists of selecting a group of genotyped animals, known as *core* animals, which are selected to span roughly 99% of the eigenvalue spectra of $\mathbf{G}$, and then expressing the breeding values of the remaining genotyped animals, known as *non-core* animals, as a linear function of the breeding values of the core animals plus an error term. Mäntysaari et al. [10] proposed the GT best linear unbiased predictor (GTBLUP) model, which uses the Woodbury Identity [11] to obtain the inverse of $\mathbf{G}$ plus a regularization matrix without explicitly computing $\mathbf{G}^{-1}$. The GTBLUP model requires more operations than APY when the number of core animals is less than m [10]. Fernando et al. [12] reviewed and developed different alternatives to the APY. In their study, the most practical implementation (Strategy IV) consisted of fitting a model with a design matrix that results from orthonormalization of the rows of the genotype matrix, and then obtaining the estimated breeding values as the product between the design matrix and the vector of solutions. APY, GTBLUP, and Strategy IV were compared in several scenarios and gave similar predictions when the spectrum of $\mathbf{G}$ was well covered by core animals [10, 13].

The lack of clear-cut, deterministic selection criteria for core animals in APY has been criticized [10, 12]; however, there are empirical studies on which [14] and how many [15] animals to include in the core group. However, the implications of using APY in terms of indirect predictions (direct genomic values) or single nucleotide polymorphism (SNP) effects are also unclear.

Although there are empirical results showing the reliable behavior of APY for prediction, an analytical framework to justify and examine its properties is lacking. Thus, the objective of this study was to derive a marker effects model that is equivalent to the APY. Departing from such a model, the secondary objectives of this study were to derive appropriate formulae for SNP and indirect predictions for genotyped animals without progeny and records, and to show the theoretical implications of selecting a specific group of core animals in APY.

Theory

Let $\mathbf{u}' = \left[\begin{array}{cc}\mathbf{u}_c^{'} & \mathbf{u}_n^{'}\end{array}\right]'$ be the vector of breeding values, where $\mathbf{u}_c$ represents the sub-vector of breeding values for the core animals and $\mathbf{u}_n$ is the sub-vector of breeding values for the non-core animals. Hereafter, the sub or superscript c will denote an object belonging to the core animals, whereas the sub or superscript n will indicate that an object belongs to the non-core animals. For a conventional GBLUP model [1], the covariance matrix of $\mathbf{u}$, assuming a genetic variance equal to unity, is equal to:

$$\text{Var}\left[\begin{array}{c}\mathbf{u}_c\\ \mathbf{u}_n\end{array}\right] = \mathbf{G} = \left[\begin{array}{cc}\mathbf{G}_{cc} & \mathbf{G}_{cn}\\ \mathbf{G}_{nc} & \mathbf{G}_{nn}\end{array}\right] = \left[\begin{array}{cc}k\mathbf{Z}_c\mathbf{Z}_c^{'} & k\mathbf{Z}_c\mathbf{Z}_n^{'}\\ k\mathbf{Z}_n\mathbf{Z}_c^{'} & k\mathbf{Z}_n\mathbf{Z}_n^{'}\end{array}\right] = k\mathbf{Z}\mathbf{Z}, \quad (1)$$

where k is a scaling factor based on the level of heterozygosity (for instance: $1/2\sum p_i q_i$) or on the number of markers, and $\mathbf{Z} = \left[\begin{array}{c}\mathbf{Z}_c\\ \mathbf{Z}_n\end{array}\right]$ is the matrix of genotypes for all the genotyped animals. Here it is assumed that $\mathbf{Z}$ was derived after genotype quality control [16], centering [2], and/or additional scaling [17]. Furthermore, it is assumed that the number of core animals is smaller than or equal to the number of markers and that the core animals are chosen such that $\mathbf{G}_{cc}$ is non-singular.

The APY approach is based on the following recursion [9]:

$$\left[\begin{array}{c}\mathbf{u}_c\\ \mathbf{u}_n\end{array}\right] = \left[\begin{array}{cc}\mathbf{I} & \mathbf{0}\\ \mathbf{P}_{nc} & \mathbf{I}\end{array}\right]\left[\begin{array}{c}\mathbf{u}_c\\ \boldsymbol{\xi}\end{array}\right], \quad (2)$$

where $\mathbf{P}_{nc} = \mathbf{G}_{nc}\mathbf{G}_{cc}^{-1}$, and $\boldsymbol{\xi}$ is an error term that is assumed to be independent of $\mathbf{u}_c$. Taking the variance of both sides in Eq. (2), with their respective inverses, leads to:

$$\begin{aligned}\text{Var}\left[\begin{array}{c}\mathbf{u}_c\\ \mathbf{u}_n\end{array}\right] = \mathbf{G}_{\text{APY}} &= \left[\begin{array}{cc}\mathbf{I} & \mathbf{0}\\ \mathbf{P}_{nc} & \mathbf{I}\end{array}\right]\left[\begin{array}{cc}\mathbf{G}_{cc} & \mathbf{0}\\ \mathbf{0} & \mathbf{M}_{nn}\end{array}\right]\left[\begin{array}{cc}\mathbf{I} & \mathbf{P}_{cn}\\ \mathbf{0} & \mathbf{I}\end{array}\right]\\ &= \left[\begin{array}{cc}\mathbf{G}_{cc} & \mathbf{G}_{cn}\\ \mathbf{G}_{nc} & \mathbf{M}_{nn} + \mathbf{G}_{nc}\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn}\end{array}\right],\end{aligned}$$

$$\text{Var}\begin{bmatrix}\mathbf{u}_c\\ \mathbf{u}_n\end{bmatrix}^{-1} = \mathbf{G}_{\text{APY}}^{-1} = \begin{bmatrix}\mathbf{I} & -\mathbf{P}_{cn}\\ \mathbf{0} & \mathbf{I}\end{bmatrix} \begin{bmatrix}\mathbf{G}_{cc}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{M}_{nn}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ -\mathbf{P}_{nc} & \mathbf{I}\end{bmatrix} = \begin{bmatrix}\mathbf{G}_{cc}^{-1}+\mathbf{P}_{cn}\mathbf{M}_{nn}^{-1}\mathbf{P}_{nc} & -\mathbf{P}_{cn}\mathbf{M}_{nn}^{-1}\\ -\mathbf{M}_{nn}^{-1}\mathbf{P}_{nc} & \mathbf{M}_{nn}^{-1}\end{bmatrix} \quad (3)$$

where $\text{Var}(\boldsymbol{\xi}) = \mathbf{M}_{nn} = \mathbf{G}_{nn} - \mathbf{G}_{nc}\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn}$. In practice, and hereafter, $\mathbf{M}_{nn}$ is assumed to be a diagonal matrix, that is, $\mathbf{M}_{nn} = \text{diag}\left(\mathbf{G}_{nn} - \mathbf{G}_{nc}\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn}\right)$ [9]. This implies that, conditional on core animals, the breeding values of non-core animals are conditionally independent, which is more explicitly shown in [8], and is a variant of the so-called "approximate kernel methods" [18, 19]. It has to be noted that, even when $\mathbf{G}$ is full rank and invertible, $\mathbf{G}_{\text{APY}}^{-1}$ is not equal to $\mathbf{G}^{-1}$.

APY-GBLUP

Based on the matrices from Eq. (3), the APY-GBLUP model is defined as:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{W}\mathbf{u} + \mathbf{e},$$

$$\text{E}[\mathbf{y}] = \mathbf{X}\mathbf{b},$$

$$\text{Var}\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{G}_{\text{APY}}\sigma_u^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\sigma_e^2\end{bmatrix}, \quad (4)$$

where $\mathbf{y}$ is the vector of phenotypes, $\mathbf{b}$ is the vector of fixed effects, $\mathbf{e}$ is the vector of error terms, $\mathbf{X}$ and $\mathbf{W}$ are design matrices, and σ_u^2 and σ_e^2 are the genetic and residual variances, respectively. Hereafter, it will be assumed that $\text{E}[\mathbf{y}] = \mathbf{X}\mathbf{b}$. Assuming multivariate normality for $\mathbf{u}$ and $\mathbf{e}$, the MME for the APY-GBLUP model are:

$$\begin{bmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W}\\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \mathbf{G}_{\text{APY}}^{-1}\alpha\end{bmatrix}\begin{bmatrix}\widehat{\mathbf{b}}\\ \widehat{\mathbf{u}}\end{bmatrix} = \begin{bmatrix}\mathbf{X}'\mathbf{y}\\ \mathbf{W}'\mathbf{y}\end{bmatrix}, \quad (5)$$

where $\alpha = \frac{\sigma_e^2}{\sigma_u^2}$.

Holding its expectation and covariance structure, the model of Eq. (4) can be partitioned into core and non-core animals as follows:

$$\begin{bmatrix}\mathbf{y}_c\\ \mathbf{y}_n\end{bmatrix} = \begin{bmatrix}\mathbf{X}_c\\ \mathbf{X}_n\end{bmatrix}\mathbf{b} + \begin{bmatrix}\mathbf{W}_c & \mathbf{0}\\ \mathbf{0} & \mathbf{W}_n\end{bmatrix}\begin{bmatrix}\mathbf{u}_c\\ \mathbf{u}_n\end{bmatrix} + \mathbf{e}, \quad (6)$$

resulting in the following MME:

$$\begin{bmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}_c'\mathbf{W}_c & \mathbf{X}_n'\mathbf{W}_n\\ \mathbf{W}_c'\mathbf{X}_c & \mathbf{W}_c'\mathbf{W}_c + \mathbf{G}_{\text{APY}}^{cc}\alpha & \mathbf{G}_{\text{APY}}^{cn}\alpha\\ \mathbf{W}_n'\mathbf{X}_n & \mathbf{G}_{\text{APY}}^{nc}\alpha & \mathbf{W}_n'\mathbf{W}_n + \mathbf{G}_{\text{APY}}^{nn}\alpha\end{bmatrix}\begin{bmatrix}\widehat{\mathbf{b}}\\ \widehat{\mathbf{u}}_c\\ \widehat{\mathbf{u}}_n\end{bmatrix} = \begin{bmatrix}\mathbf{X}'\mathbf{y}\\ \mathbf{W}_c'\mathbf{y}_c\\ \mathbf{W}_n'\mathbf{y}_n\end{bmatrix}, \quad (7)$$

where the superscripts in $\mathbf{G}_{\text{APY}}$ denote the blocks of $\mathbf{G}_{\text{APY}}^{-1}$ corresponding to a specific combination of core and non-core animals. Substituting $\mathbf{u}_n = \mathbf{P}_{nc}\mathbf{u}_c + \boldsymbol{\xi}$ in Eq. (6) leads to the following model:

$$\begin{bmatrix}\mathbf{y}_c\\ \mathbf{y}_n\end{bmatrix} = \begin{bmatrix}\mathbf{X}_c\\ \mathbf{X}_n\end{bmatrix}\mathbf{b} + \begin{bmatrix}\mathbf{W}_c\\ \mathbf{W}_n\mathbf{P}_{nc}\end{bmatrix}\mathbf{u}_c + \begin{bmatrix}\mathbf{0}\\ \mathbf{W}_n\end{bmatrix}\boldsymbol{\xi} + \mathbf{e},$$

$$\text{Var}\begin{bmatrix}\mathbf{u}_c\\ \boldsymbol{\xi}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{G}_{cc}\sigma_u^2 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{M}_{nn}\sigma_u^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2\end{bmatrix}, \quad (8)$$

and the following MME:

$$\begin{bmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}_c'\mathbf{W}_c & \mathbf{X}_n'\mathbf{W}_n\\ \mathbf{W}_c'\mathbf{X}_c & \mathbf{W}_c'\mathbf{W}_c + \mathbf{P}_{cn}\mathbf{W}_n'\mathbf{W}_n\mathbf{P}_{nc} + \mathbf{G}_{cc}^{-1}\alpha & \mathbf{P}_{cn}\mathbf{W}_n'\mathbf{W}_n\\ \mathbf{W}_n'\mathbf{X}_n & \mathbf{W}_n'\mathbf{W}_n\mathbf{P}_{nc} & \mathbf{W}_n'\mathbf{W}_n + \mathbf{M}_{nn}^{-1}\alpha\end{bmatrix}\begin{bmatrix}\widehat{\mathbf{b}}\\ \widehat{\mathbf{u}}_l\\ \widehat{\boldsymbol{\xi}}\end{bmatrix} = \begin{bmatrix}\mathbf{X}'\mathbf{y}\\ \mathbf{W}_c'\mathbf{y}_c\\ \mathbf{W}_n'\mathbf{y}_n\end{bmatrix}. \quad (9)$$

Then, the BLUP of $\mathbf{u}_n$ is obtained as $\widehat{\mathbf{u}}_n = \mathbf{P}_{nc}\widehat{\mathbf{u}}_c + \widehat{\boldsymbol{\xi}}$. These MME are similar to those shown in Eq. (13) in [12] but differ in the error term $\boldsymbol{\xi}$.

An equivalent model: APY-SNP-BLUP

Assuming that the genetic value of the core animals is fully explained by the markers, then:

$\mathbf{u}_c = \mathbf{Z}_c\mathbf{a}$ and

$$\mathbf{u}_n = \mathbf{Z}_n\mathbf{Z}_c'\left(\mathbf{Z}_c\mathbf{Z}_c'\right)^{-1}\mathbf{Z}_c\mathbf{a} + \boldsymbol{\xi} = \mathbf{Z}_n\boldsymbol{\mathcal{P}}\mathbf{a} + \boldsymbol{\xi}, \quad (10)$$

where $\mathbf{a}$ is the vector of marker effects, and $\boldsymbol{\mathcal{P}} = \mathbf{Z}_c'\left(\mathbf{Z}_c\mathbf{Z}_c'\right)^{-1}\mathbf{Z}_c$, that is, $\boldsymbol{\mathcal{P}}$ is the perpendicular projection operator [11] to $\mathcal{C}\left(\mathbf{Z}_c'\right)$ (i.e., the vector space spanned by the genotypes of the core animals). Assuming that $k\sigma_u^2$ is the variance of marker effects, i.e. $\text{Var}(\mathbf{a}) = \mathbf{I}k\sigma_u^2$, $\text{Var}(\boldsymbol{\xi}) = \mathbf{M}_{nn}\sigma_u^2$, and $\text{cov}(\mathbf{a}, \boldsymbol{\xi}) = \mathbf{0}$. Then:

$\text{Var}(\mathbf{u}_c) = \mathbf{Z}_c\text{Var}(\mathbf{a})\mathbf{Z}_c' = \sigma_u^2 k\mathbf{Z}_c\mathbf{Z}_c'$ and

$$\begin{aligned}\text{Var}(\mathbf{u}_n) &= \mathbf{Z}_n\boldsymbol{\mathcal{P}}\text{Var}(\mathbf{a})\boldsymbol{\mathcal{P}}'\mathbf{Z}_n' + \text{Var}(\boldsymbol{\xi}) \\ &= \sigma_u^2 k\mathbf{Z}_n\boldsymbol{\mathcal{P}}\mathbf{Z}_n' + \sigma_u^2\mathbf{M}_{nn}.\end{aligned} \tag{11}$$

Letting $\mathbf{Z}^\dagger = \begin{bmatrix}\mathbf{Z}_c \\ \mathbf{Z}_n\boldsymbol{\mathcal{P}}\end{bmatrix}$ and $\mathbf{Q} = \begin{bmatrix}\mathbf{0} \\ \mathbf{W}_n\end{bmatrix}$, the following APY-SNP-BLUP model is equivalent to the APY-GBLUP model presented in Eq. (4):

$\mathbf{y} = \mathbf{Xb} + \mathbf{WZ}^\dagger\mathbf{a} + \mathbf{Q}\boldsymbol{\xi} + \mathbf{e}$, with

$$\text{Var}\begin{bmatrix}\mathbf{a} \\ \boldsymbol{\xi} \\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}k\mathbf{I}\sigma_u^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{nn}\sigma_u^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2\end{bmatrix}. \tag{12}$$

Assuming multivariate normality for $\mathbf{a}$, $\boldsymbol{\xi}$, and $\mathbf{e}$, the MME for the APY-SNP-BLUP model are:

$$\begin{bmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{WZ}^\dagger & \mathbf{X}_n'\mathbf{W}_n \\ \mathbf{Z}^{\dagger'}\mathbf{W}'\mathbf{X} & \mathbf{Z}^{\dagger'}\mathbf{W}'\mathbf{WZ}^\dagger + \mathbf{I}\gamma & \boldsymbol{\mathcal{P}}\mathbf{Z}_n'\mathbf{W}_n'\mathbf{W}_n \\ \mathbf{W}_n'\mathbf{X}_n & \mathbf{W}_n'\mathbf{W}_n\mathbf{Z}_n\boldsymbol{\mathcal{P}} & \mathbf{W}_n'\mathbf{W}_n + \mathbf{M}_{nn}^{-1}\alpha\end{bmatrix}$$
$$\begin{bmatrix}\widehat{\mathbf{b}} \\ \widehat{\mathbf{a}} \\ \widehat{\boldsymbol{\xi}}\end{bmatrix} = \begin{bmatrix}\mathbf{X}'\mathbf{y} \\ \mathbf{Z}^{\dagger'}\mathbf{W}'\mathbf{y} \\ \mathbf{W}_n'\mathbf{y}_n\end{bmatrix}, \tag{13}$$

where $\gamma = \frac{\sigma_e^2}{k\sigma_u^2}$.

If $rank(\mathbf{Z}_c) = rank(\mathbf{Z})$, which is true when the number of core animals is equal to the number of markers and given a non-singular $\mathbf{G}_{cc}$, $\boldsymbol{\mathcal{P}} = \mathbf{I}$. Further, $\mathbf{M}_{nn} = \mathbf{0}$, and since $\boldsymbol{\xi}$ has null expectation, $\boldsymbol{\xi} = \mathbf{0}$. Consequently, for a sufficiently large number of core animals:

$\mathbf{u}_c = \mathbf{Z}_c\mathbf{a}$ and

$$\mathbf{u}_n = \mathbf{Z}_n\boldsymbol{\mathcal{P}}\mathbf{a} + \boldsymbol{\xi} = \mathbf{Z}_n\mathbf{a} + \boldsymbol{\xi} = \mathbf{Z}_n\mathbf{a}. \tag{14}$$

Then, the model in Eq. (12) reduces to:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{WZa} + \mathbf{e},$$

$$\text{with Var}\begin{bmatrix}\mathbf{a} \\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}k\mathbf{I}\sigma_u^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_e^2\end{bmatrix}, \tag{15}$$

and the MME are:

$$\begin{bmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{Z}'\mathbf{W}'\mathbf{WZ} + \mathbf{I}\gamma\end{bmatrix}\begin{bmatrix}\widehat{\mathbf{b}} \\ \widehat{\mathbf{a}}\end{bmatrix} = \begin{bmatrix}\mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{W}'\mathbf{y}\end{bmatrix}, \tag{16}$$

which are the MME for a typical SNP-BLUP model [1]. Thus, APY-SNP-BLUP converges to the regular SNP-BLUP when the number of core animals increases.

Distribution of breeding values and marker effects with APY

By defining the breeding values as in Eq. (10) and keeping the distributional assumptions of $\mathbf{a}$ and $\boldsymbol{\xi}$ from Eq. (12), the covariance matrix of the joint distribution of $\mathbf{u}$ and $\mathbf{a}$ is:

$$\text{Var}\begin{bmatrix}\mathbf{u} \\ \mathbf{a}\end{bmatrix} = \begin{bmatrix}k\mathbf{Z}^\dagger\mathbf{Z}^{\dagger'} + \mathbf{V} & k\mathbf{Z}^\dagger \\ k\mathbf{Z}^{\dagger'} & k\mathbf{I}\end{bmatrix}\sigma_u^2, \tag{17}$$

where $\mathbf{V} = \begin{bmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{nn}\end{bmatrix}$. Then, assuming normality for $\mathbf{a}$ and $\boldsymbol{\xi}$, the conditional distribution of $\mathbf{u}$ on $\mathbf{a}$ is:

$$p(\mathbf{u}|\mathbf{a}) = N\left(\mathbf{Z}^\dagger\mathbf{a}, \mathbf{V}\sigma_u^2\right). \tag{18}$$

Note that this is a degenerate normal distribution because its covariance matrix is non-positive definite. In the classical GBLUP and SNP-BLUP models [1], the variance of $\mathbf{u}$ conditional on $\mathbf{a}$ is zero (i.e., if marker effects are known, the breeding value is fully explained), whereas Eq. (18) shows that in APY, the variance of $\mathbf{u}$ conditional on $\mathbf{a}$ is nonzero for the non-core animals.

Conversely, the conditional distribution of $\mathbf{a}$ on $\mathbf{u}$ is:

$$\begin{aligned}p(\mathbf{a}|\mathbf{u}) = N\Bigg(&k\mathbf{Z}^{\dagger'}\left(k\mathbf{Z}^\dagger\mathbf{Z}^{\dagger'} + \mathbf{V}\right)^{-1}\mathbf{u}, \\ &\left(k\mathbf{I} - k\mathbf{Z}^{\dagger'}\left(k\mathbf{Z}^\dagger\mathbf{Z}^{\dagger'} + \mathbf{V}\right)^{-1}\mathbf{Z}^\dagger k\right)\sigma_u^2\Bigg).\end{aligned} \tag{19}$$

The BLUP of $\mathbf{a}$ given $\mathbf{u} = \widehat{\mathbf{u}}$ can be obtained from Eq. (19) as [1]:

$$\widehat{\mathbf{a}}|\widehat{\mathbf{u}} = \text{E}\left[\mathbf{a}|\mathbf{u} = \widehat{\mathbf{u}}\right] = k\mathbf{Z}^{\dagger'}\left(k\mathbf{Z}^\dagger\mathbf{Z}^{\dagger'} + \mathbf{V}\right)^{-1}\widehat{\mathbf{u}} = k\mathbf{Z}^{\dagger'}\mathbf{G}_{APY}^{-1}\widehat{\mathbf{u}}, \tag{20}$$

which after algebra reduces to [20]:

$$\widehat{\mathbf{a}}|\widehat{\mathbf{u}} = k\mathbf{Z}^{\dagger'}\mathbf{G}_{APY}^{-1}\widehat{\mathbf{u}} = k\mathbf{Z}_c'\mathbf{G}_{cc}^{-1}\widehat{\mathbf{u}}_c, \tag{21}$$

with variance equal to:

$$\text{Var}\left(\widehat{\mathbf{a}}|\widehat{\mathbf{u}}\right) = k^2\mathbf{Z}_c'\mathbf{G}_{cc}^{-1}(\mathbf{G}_{cc} - \text{PEV}_{core})\mathbf{G}_{cc}^{-1}\mathbf{Z}_c. \tag{22}$$

Thus, in order to obtain predictions of marker effects from the APY-GBLUP model, only the estimated breeding values of the core animals and the corresponding blocks of matrices are needed.

Indirect predictions with APY

Indirect predictions $\left(\widehat{\mathbf{u}}_{ip}\right)$ are estimated breeding values for animals without own records or progeny with records, based on estimated marker effects from the genetic evaluation. Therefore, an indirect-predicted animal is by definition a non-core animal. Using Eq. (18), indirect predictions with APY are calculated as:

$$\widehat{\mathbf{u}}_{ip}|\widehat{\mathbf{a}} = \mathbf{Z}_{ip}\mathcal{P}\widehat{\mathbf{a}} = \mathbf{G}_{ip,c}\mathbf{G}_{cc}^{-1}\widehat{\mathbf{u}}_c, \tag{23}$$

where $\mathbf{Z}_{ip}$ is the centered and scaled genotype matrix for the indirect-predicted animals, and $\mathbf{G}_{ip,c}$ is the genomic relationship matrix between indirect-predicted and core animals. Note that the genotype matrix $\mathbf{Z}_{ip}$ has to be centered and scaled in exactly the same manner as the matrices $\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_c \\ \mathbf{Z}_n \end{bmatrix}$ that are used for genomic evaluation by any method; otherwise, the algebra (i.e. Eq. (17)) used to derive the joint distribution of breeding values and markers is an approximation. Changing the centering of $\mathbf{Z}_{ip}$ (i.e., using a different set of allele frequencies) results in a shift of the mean and, in addition, a different implicit $\boldsymbol{\xi}$ in the APY approximation. Changing the scale of $\mathbf{Z}_{ip}$ (i.e., multiplying by a different k) changes the scale of marker effects and of the indirect predictions.

Note that the leftmost expression in Eq. (23) is, as expected, the selection index formulation for estimation of breeding values. Two measures of uncertainty are associated with estimation of $\widehat{\mathbf{u}}_{ip}$. First, the uncertainty associated with the variance in Eq. (18) that arises from the error term in Eq. (10), i.e. due to the approximation in APY. This uncertainty is calculated similarly to a reliability:

$$\rho_i = 1 - \frac{\mathbf{m}_{ii}}{\mathbf{g}_{ii}}, \tag{24}$$

where $\mathbf{g}_{ii}$ and $\mathbf{m}_{ii}$ are the diagonal element of $\mathbf{G}$ and the element of $\mathbf{M}_{nn}$, respectively, corresponding to the i^{th} animal. Equation (24) converges to one when $\boldsymbol{\xi} \to 0$, which occurs when the size of the core increases. The second measure of uncertainty is the usual reliability associated with the prediction error variance of $\widehat{\mathbf{u}}_{ip}|\widehat{\mathbf{a}}$, which is:

$$\mathrm{Var}\left(\widehat{\mathbf{u}}_{ip}|\widehat{\mathbf{a}} - \mathbf{u}_{ip}\right)_i = \mathbf{m}_{ii} + \mathbf{g}_{i,c}\mathbf{G}_{cc}^{-1}\mathrm{PEV}_{core}\mathbf{G}_{cc}^{-1}\mathbf{g}_{c,i}, \tag{25}$$

where $\mathbf{g}_{i,c}$ is the block of $\mathbf{G}$ that relates the i^{th} individual with the core animals, and $\mathbf{g}_{c,i} = \mathbf{g}_{i,c}'$. The reliability associated with Eq. (25) is:

$$\begin{aligned} \mathrm{rel}_i &= 1 - \frac{\mathbf{m}_{ii} + \mathbf{g}_{i,c}\mathbf{G}_{cc}^{-1}\mathrm{PEV}_{core}\mathbf{G}_{cc}^{-1}\mathbf{g}_{c,i}}{\mathbf{g}_{ii}} \\ &= \rho_i - \frac{\mathbf{g}_{i,c}\mathbf{G}_{cc}^{-1}\mathrm{PEV}_{core}\mathbf{G}_{cc}^{-1}\mathbf{g}_{c,i}}{\mathbf{g}_{ii}}. \end{aligned} \tag{26}$$

APY-GBLUP and APY-SNP-BLUP models with a residual polygenic effect

To consider an extra polygenic effect based on pedigree, $\mathbf{G}_{APY}^{-1}$ in Eq. (5) can be calculated based on the following $\mathbf{G}^*$:

$$\begin{aligned} \mathbf{G}^* = \mathbf{Z}^*\mathbf{Z}^{*'} &= \left[\sqrt{(1-\beta)k}\mathbf{Z} \;\; \sqrt{\beta}\mathbf{L}\right]\begin{bmatrix} \sqrt{(1-\beta)k}\mathbf{Z}' \\ \sqrt{\beta}\mathbf{L}' \end{bmatrix} \\ &= (1-\beta)\left(k\mathbf{Z}\mathbf{Z}'\right) + \beta\mathbf{A}_{22}, \end{aligned} \tag{27}$$

where $\mathbf{A}_{22}$ is the block of the numerator relationship matrix corresponding to the genotyped animals, $\mathbf{L}$ is the Cholesky factor or its approximation of $\mathbf{A}_{22}$ [21], and β is the proportion of the residual polygenic effect. When the matrices from model Eq. (4) are constructed based on Eq. (27), the resulting model is designated as APY-GBLUP with a residual polygenic effect. In this case, the breeding values are not fully explained by the markers. Therefore:

$$\mathbf{u}_c = \mathbf{Z}_c\mathbf{a} + \boldsymbol{\varepsilon},$$

$$\mathbf{u}_n = \mathbf{P}_{nc}^*(\mathbf{Z}_c\mathbf{a} + \boldsymbol{\varepsilon}) + \boldsymbol{\xi} = \mathbf{Z}_n^*\mathcal{P}^*\mathbf{S}\mathbf{a} + \mathbf{Z}_n^*\mathbf{Z}_c^{*'}\left(\mathbf{Z}_c^*\mathbf{Z}_c^{*'}\right)^{-1}\boldsymbol{\varepsilon} + \boldsymbol{\xi}, \tag{28}$$

where $\boldsymbol{\varepsilon} \sim \mathrm{MVN}\left(\mathbf{0}, \beta\mathbf{A}_{cc}\sigma_u^2\right)$ is the vector of residual polygenic effects [22], $\mathbf{A}_{cc}$ is the block of the numerator relationship matrix corresponding to the core animals, $\mathbf{P}_{nc}^* = \mathbf{G}_{nc}^*\mathbf{G}_{cc}^{*^{-1}}$, $\mathcal{P}^* = \mathbf{Z}_c^{*'}\left(\mathbf{Z}_c^*\mathbf{Z}_c^{*'}\right)^{-1}\mathbf{Z}_c^*$, and $\mathbf{S} = \begin{bmatrix} \mathbf{I}\frac{1}{\sqrt{(1-\beta)k}} \\ \mathbf{0} \end{bmatrix}$. Then, a marker effects model equivalent to APY-GBLUP with a residual polygenic effect is:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{W}\mathbf{Z}^{\dagger *}\mathbf{a} + \mathbf{W}\mathbf{R}\boldsymbol{\varepsilon} + \mathbf{Q}\boldsymbol{\xi} + \mathbf{e},$$

$$\mathrm{Var}\begin{bmatrix} \mathbf{a} \\ \boldsymbol{\varepsilon} \\ \boldsymbol{\xi} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} (1-\beta)k\mathbf{I}\sigma_u^2 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \beta\mathbf{A}_{cc}\sigma_u^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{M}_{nn}\sigma_u^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2 \end{bmatrix}, \tag{29}$$

where $\mathbf{Z}^{\dagger *} = \begin{bmatrix} \mathbf{Z}_c \\ \mathbf{Z}_n^*\mathcal{P}^*\mathbf{S} \end{bmatrix}$ and $\mathbf{R} = \begin{bmatrix} \mathbf{I} \\ \mathbf{Z}_n^*\mathbf{Z}_c^{*'}\left(\mathbf{Z}_c^*\mathbf{Z}_c^{*'}\right)^{-1} \end{bmatrix}$.

The MME for the model in Eq. (29) are:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W}\mathbf{Z}^{\dagger *} & \mathbf{X}'\mathbf{W}\mathbf{T} & \mathbf{X}_n'\mathbf{W}_n \\ \mathbf{Z}^{\dagger *'}\mathbf{W}'\mathbf{X} & \mathbf{Z}^{\dagger *'}\mathbf{W}'\mathbf{W}\mathbf{Z}^{\dagger *} + \mathbf{I}\delta & \mathbf{Z}^{\dagger *'}\mathbf{W}'\mathbf{W}\mathbf{T} & \mathbf{S}'\mathcal{P}^*\mathbf{Z}_n^{*'}\mathbf{W}_n'\mathbf{W}_n \\ \mathbf{T}'\mathbf{W}'\mathbf{X} & \mathbf{T}'\mathbf{W}'\mathbf{W}\mathbf{Z}^{\dagger *} & \mathbf{T}'\mathbf{W}'\mathbf{W}\mathbf{T} + \mathbf{A}_{cc}^{-1}\zeta & \left(\mathbf{Z}_c^*\mathbf{Z}_c^{*'}\right)^{-1}\mathbf{Z}_c^*\mathbf{Z}_n^{*'}\mathbf{W}_n'\mathbf{W}_n \\ \mathbf{W}_n'\mathbf{X}_n & \mathbf{W}_n'\mathbf{W}_n\mathbf{Z}_n^*\mathcal{P}^*\mathbf{S} & \mathbf{W}_n'\mathbf{W}_n\mathbf{Z}_n^*\mathbf{Z}_c^{*'}\left(\mathbf{Z}_c^*\mathbf{Z}_c^{*'}\right)^{-1} & \mathbf{W}_n'\mathbf{W}_n + \mathbf{M}_{nn}^{-1}\alpha \end{bmatrix}\begin{bmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{a}} \\ \widehat{\boldsymbol{\varepsilon}} \\ \widehat{\boldsymbol{\xi}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}^{\dagger *'}\mathbf{W}'\mathbf{y} \\ \mathbf{R}'\mathbf{W}'\mathbf{y} \\ \mathbf{W}_n'\mathbf{y}_n \end{bmatrix}. \tag{30}$$

where $\delta = \frac{\sigma_e^2}{(1-\beta)k\sigma_u^2}$ and $\zeta = \frac{\sigma_e^2}{\beta\sigma_u^2}$.

APY single-step GBLUP (APY-ssGBLUP) and APY single-step SNP-BLUP (APY-ssSNP-BLUP)

For a general ssGBLUP model [23] with APY [14], assuming a genetic variance equal to 1, the covariance matrix of the breeding values is equal to:

$$\begin{aligned}&\mathrm{Var}\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{bmatrix}\\ &= \mathbf{H}_{\mathrm{APY}} = \begin{bmatrix}\mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G}_{\mathrm{APY}} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}_{\mathrm{APY}}\\ \mathbf{G}_{\mathrm{APY}}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G}_{\mathrm{APY}}\end{bmatrix},\end{aligned} \tag{31}$$

where the subscripts 1 and 2 denote the non-genotyped and genotyped animals, respectively. Letting $\mathbf{A}_{22}^{-1} = \begin{bmatrix}\mathbf{A}_{22}^{\mathrm{cc}} & \mathbf{A}_{22}^{\mathrm{cn}}\\ \mathbf{A}_{22}^{\mathrm{nc}} & \mathbf{A}_{22}^{\mathrm{nn}}\end{bmatrix}$, the inverse of $\mathbf{H}_{\mathrm{APY}}$ is equal to [24]:

$$\begin{aligned}\mathbf{H}_{\mathrm{APY}}^{-1} &= \mathbf{A}^{-1} + \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}_{\mathrm{APY}}^{-1} - \mathbf{A}_{22}^{-1}\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{A}^{11} & \mathbf{A}^{1c} & \mathbf{A}^{1n}\\ \mathbf{A}^{c1} & \mathbf{A}^{cc} & \mathbf{A}^{cn}\\ \mathbf{A}^{n1} & \mathbf{A}^{nc} & \mathbf{A}^{nn}\end{bmatrix} + \begin{bmatrix}\mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}_{cc}^{-1} + \mathbf{P}_{cn}\mathbf{M}_{nn}^{-1}\mathbf{P}_{nc} - \mathbf{A}_{22}^{cc} & -\mathbf{P}_{cn}\mathbf{M}_{nn}^{-1} - \mathbf{A}_{22}^{cn}\\ \mathbf{0} & -\mathbf{M}_{nn}^{-1}\mathbf{P}_{nc} - \mathbf{A}_{22}^{nc} & \mathbf{M}_{nn}^{-1} - \mathbf{A}_{22}^{nn}\end{bmatrix}.\end{aligned} \tag{32}$$

Then, the APY-ssGBLUP model is defined as:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{W}\mathbf{u} + \mathbf{e},$$

$$\mathrm{Var}\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{H}_{\mathrm{APY}}\sigma_u^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\sigma_e^2\end{bmatrix}. \tag{33}$$

The MME for the model in Eq. (33) are:

$$\begin{bmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}_1'\mathbf{W}_1 & \mathbf{X}_2'\mathbf{W}_2\\ \mathbf{W}_1'\mathbf{X}_1 & \mathbf{W}_1'\mathbf{W}_1 + \mathbf{A}^{11}\alpha & \mathbf{A}^{12}\alpha\\ \mathbf{W}_c'\mathbf{X}_c & \mathbf{A}^{21}\alpha & \mathbf{W}_2'\mathbf{W}_2 + \left(\mathbf{A}^{22} + \mathbf{G}_{\mathrm{APY}}^{-1} - \mathbf{A}_{22}^{-1}\right)\alpha\end{bmatrix}$$
$$\begin{bmatrix}\widehat{\mathbf{b}}\\ \widehat{\mathbf{u}}_1\\ \widehat{\mathbf{u}}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}'\mathbf{y}\\ \mathbf{W}_1'\mathbf{y}_1\\ \mathbf{W}_2'\mathbf{y}_2\end{bmatrix}. \tag{34}$$

Following [25], the breeding values of non-genotyped animals can be written as: $\mathbf{u}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2 + \boldsymbol{\eta}$, where $\boldsymbol{\eta} \sim \mathrm{MVN}\left(0, \left(\mathbf{A}^{11}\right)^{-1}\right)$ and represents the imputation error [23, 25]. Then, the breeding values of all animals can be expressed as:

$$\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_c\\ \mathbf{u}_n\end{bmatrix} = \begin{bmatrix}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\begin{bmatrix}\mathbf{Z}_c\mathbf{a}\\ \mathbf{Z}_n\boldsymbol{\mathcal{P}}\mathbf{a} + \boldsymbol{\xi}\end{bmatrix} + \boldsymbol{\eta}\\ \mathbf{Z}_c\mathbf{a}\\ \mathbf{Z}_n\boldsymbol{\mathcal{P}}\mathbf{a} + \boldsymbol{\xi}\end{bmatrix}. \tag{35}$$

Letting $\mathbf{W}^\dagger = \begin{bmatrix}\mathbf{W}_1\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{Z}^\dagger\\ \mathbf{W}_2\mathbf{Z}^\dagger\end{bmatrix}$, $\mathbf{T} = \begin{bmatrix}\mathbf{W}_1\\ \mathbf{0}\end{bmatrix}$, and $\mathbf{Q}^\dagger = \begin{bmatrix}\mathbf{W}_1\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{Q}\\ \mathbf{Q}\end{bmatrix}$, an equivalent model to that presented in Eq. (33) is:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{W}^\dagger\mathbf{a} + \mathbf{T}\boldsymbol{\eta} + \mathbf{Q}^\dagger\boldsymbol{\xi} + \mathbf{e},$$

$$\mathrm{Var}\begin{bmatrix}\mathbf{a}\\ \boldsymbol{\eta}\\ \boldsymbol{\xi}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}k\mathbf{I}\sigma_u^2 & \mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \left(\mathbf{A}^{11}\right)^{-1}\sigma_u^2 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{M}_{\mathrm{nn}}\sigma_u^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2\end{bmatrix} \tag{36}$$

The MME corresponding to this model are:

$$\begin{bmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W}^\dagger & \mathbf{X}_1'\mathbf{W}_1 & \mathbf{X}'\mathbf{Q}^\dagger\\ \mathbf{W}^{\dagger'}\mathbf{X} & \mathbf{W}^{\dagger'}\mathbf{W}^\dagger + \mathbf{I}\gamma & \mathbf{W}^{\dagger'}\mathbf{T} & \mathbf{W}^{\dagger'}\mathbf{Q}^\dagger\\ \mathbf{W}_1'\mathbf{X}_1 & \mathbf{T}'\mathbf{W}^\dagger & \mathbf{W}_1'\mathbf{W}_1 + \mathbf{A}^{11}\alpha & \mathbf{T}'\mathbf{Q}^\dagger\\ \mathbf{Q}^{\dagger'}\mathbf{X} & \mathbf{Q}^{\dagger'}\mathbf{W}^\dagger & \mathbf{Q}^{\dagger'}\mathbf{T} & \mathbf{Q}^{\dagger'}\mathbf{Q}^\dagger + \mathbf{M}_{\mathrm{nn}}^{-1}\alpha\end{bmatrix}$$
$$\begin{bmatrix}\widehat{\mathbf{b}}\\ \widehat{\mathbf{a}}\\ \widehat{\boldsymbol{\eta}}\\ \widehat{\boldsymbol{\xi}}\end{bmatrix} = \begin{bmatrix}\mathbf{X}'\mathbf{y}\\ \mathbf{W}^{\dagger'}\mathbf{y}\\ \mathbf{T}'\mathbf{y}\\ \mathbf{Q}^{\dagger'}\mathbf{y}\end{bmatrix} \tag{37}$$

APY-ssGBLUP and APY-ssSNP-BLUP with a residual polygenic effect

If $\mathbf{G}_{\mathrm{APY}}^{-1}$ in $\mathbf{H}_{\mathrm{APY}}^{-1}$ is built using $\mathbf{G}^*$ from Eq. (27), the resulting model is called APY-ssGBLUP with a residual polygenic effect. In such a case, Eq. (35) is modified to:

$$\begin{aligned}\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_c\\ \mathbf{u}_n\end{bmatrix} &= \begin{bmatrix}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\begin{bmatrix}\mathbf{Z}_c\mathbf{a} + \boldsymbol{\varepsilon}\\ \mathbf{P}_{\mathrm{nc}}^*(\mathbf{Z}_c\mathbf{a} + \boldsymbol{\varepsilon}) + \boldsymbol{\xi}\end{bmatrix} + \boldsymbol{\eta}\\ \mathbf{Z}_c\mathbf{a} + \boldsymbol{\varepsilon}\\ \mathbf{P}_{\mathrm{nc}}^*(\mathbf{Z}_c\mathbf{a} + \boldsymbol{\varepsilon}) + \boldsymbol{\xi}\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\left(\mathbf{Z}^{\dagger *}\mathbf{a} + \mathbf{R}\boldsymbol{\varepsilon} + \mathbf{Q}\boldsymbol{\xi}\right) + \boldsymbol{\eta}\\ \mathbf{Z}_c\mathbf{a} + \boldsymbol{\varepsilon}\\ \mathbf{Z}_n^*\boldsymbol{\mathcal{P}}^*\mathbf{S}\mathbf{a} + \mathbf{Z}_n^*\mathbf{Z}_c^{*'}\left(\mathbf{Z}_c^*\mathbf{Z}_c^{*'}\right)^{-1}\boldsymbol{\varepsilon} + \boldsymbol{\xi}\end{bmatrix}.\end{aligned} \tag{38}$$

Letting $\mathbf{W}^{\dagger *} = \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{Z}^{\dagger *} \\ \mathbf{Z}^{\dagger *} \end{bmatrix}$, $\mathbf{R}^{\dagger *} = \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{R} \\ \mathbf{R} \end{bmatrix}$, and $\mathbf{Q}^{\dagger *} = \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{Q} \\ \mathbf{Q} \end{bmatrix}$, Eq. (33) leads to the following model:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{W}^{\dagger *}\mathbf{a} + \mathbf{R}^{\dagger *}\boldsymbol{\varepsilon} + \mathbf{T}\boldsymbol{\eta} + \mathbf{Q}^{\dagger *}\boldsymbol{\xi} + \mathbf{e},$$

$$\text{Var}\begin{bmatrix} \mathbf{a} \\ \boldsymbol{\varepsilon} \\ \boldsymbol{\eta} \\ \boldsymbol{\xi} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} (1-\beta)k\mathbf{I}\siga_u^2 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \beta\mathbf{A}_{cc}\sigma_u^2 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \left(\mathbf{A}^{11}\right)^{-1}\sigma_u^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{M}_{nn}\sigma_u^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2 \end{bmatrix}, \tag{39}$$

with the following MME:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W}^{\dagger *} & \mathbf{X}'\mathbf{R}^{\dagger *} & \mathbf{X}_1'\mathbf{W}_1 & \mathbf{X}'\mathbf{Q}^{\dagger *} \\ \mathbf{W}^{\dagger *'}\mathbf{X} & \mathbf{W}^{\dagger *'}\mathbf{W}^{\dagger *} + \mathbf{I}\delta & \mathbf{W}^{\dagger *'}\mathbf{R}^{\dagger *} & \mathbf{W}^{\dagger *'}\mathbf{T} & \mathbf{W}^{\dagger *'}\mathbf{Q}^{\dagger *} \\ \mathbf{R}^{\dagger *'} & \mathbf{R}^{\dagger *'} & \mathbf{R}^{\dagger *'}\mathbf{R}^{\dagger *} + \mathbf{A}_{cc}^{-1}\zeta & \mathbf{R}^{\dagger *'}\mathbf{T} & \mathbf{R}^{\dagger *'}\mathbf{Q}^{\dagger *} \\ \mathbf{W}_1'\mathbf{X}_1 & \mathbf{T}'\mathbf{W}^{\dagger *} & \mathbf{T}'\mathbf{R}^{\dagger *} & \mathbf{W}_1'\mathbf{W}_1 + \mathbf{A}^{11}\alpha & \mathbf{T}'\mathbf{Q}^{\dagger *} \\ \mathbf{Q}^{\dagger *'}\mathbf{X} & \mathbf{Q}^{\dagger *'}\mathbf{W}^{\dagger *} & \mathbf{Q}^{\dagger *'}\mathbf{R}^{\dagger *} & \mathbf{Q}^{\dagger *'}\mathbf{T} & \mathbf{Q}^{\dagger *'}\mathbf{Q}^{\dagger *} + \mathbf{M}_{nn}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{a}} \\ \widehat{\boldsymbol{\varepsilon}} \\ \widehat{\boldsymbol{\eta}} \\ \widehat{\boldsymbol{\xi}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}^{\dagger *'}\mathbf{y} \\ \mathbf{R}^{\dagger *'}\mathbf{y} \\ \mathbf{T}'\mathbf{y} \\ \mathbf{Q}^{\dagger *'}\mathbf{y} \end{bmatrix} \tag{40}$$

Computational complexity of the APY-based models

In terms of pre-processing steps, i.e., preparation of the required matrices to set up the MME, the APY-GBLUP model has a cubic cost in the number of core animals, and a linear cost in both the number of markers and the number of non-core animals. In contrast, the APY-SNP-BLUP model has a quadratic cost in both the number of core animals and the number of markers, and a linear cost in the number of non-core animals. Per iteration, the computational cost of APY-GBLUP is quadratic in the number of core and linear in the number of non-core animals, while APY-SNP-BLUP is quadratic in the number of markers and linear in the number of non-core animals. Thus, as long as the number of core animals is smaller than the number of markers, APY-GBLUP will be computationally less demanding than APY-SNP-BLUP. If a residual polygenic effect is assumed, the choice of the core animals plays a major role in computational efficiency because of the computation of $\mathbf{A}_{cc}^{-1}$. For example, if unrelated core animals are chosen, then computations that involve $\mathbf{A}_{cc}^{-1}$ are trivial. Regardless, the APY-SNP-BLUP model may not be of great practical interest, as it can be replaced by either GBLUP or SNP-BLUP. However, the APY-SNP-BLUP model is useful to derive analytical properties of the APY algorithm.

Under single-step models, computational requirements for genetic evaluation of genotyped animals depend on how $\mathbf{A}_{22}^{-1}$ is included in the MME for both breeding value and marker-based models. That is, $\mathbf{A}_{22}^{-1}$ is included either as the product of $\mathbf{A}_{22}^{-1}$ times a vector for breeding value models [26] or as the solution of the sparse system $\mathbf{A}^{11}\widehat{\mathbf{Z}}^{\dagger} = -\mathbf{A}^{12}\mathbf{Z}^{\dagger}$ for marker based models [25], where $\widehat{\mathbf{Z}}^{\dagger}$ is the matrix of imputed genotypes.

Discussion

In this study, we derived a marker effects model that is equivalent to the APY-GBLUP model. On the one hand, when the number of core animals is equal to the rank of the genotype matrix, $\mathbf{G}_{APY}$ is singular, as noted by [12]. This can be interpreted as core animals covering all the genetic variation in the population that is captured by the genotyped markers. In that case, APY-GBLUP is equivalent to a regular GBLUP or SNP-BLUP, which has been analytically proven in this study (for further information, a numerical illustration is provided in Additional file 1). On the other hand, when the number of core animals is lower than the rank of the genotype matrix, $\mathbf{G}_{APY}$ is non-singular and a marker effects model, named APY-SNP-BLUP, can be constructed that is equivalent to APY-GBLUP. The APY-SNP-BLUP differs from the regular SNP-BLUP model in the following manner: (i) it has a reduction in the row and column spaces of $\mathbf{Z}$ (from its replacement with $\mathbf{Z}^{\dagger}$), and (ii) it has an additional error term ($\boldsymbol{\xi}$) for non-core animals. The former indicates that the number of possible genotypes and haplotypes that can be formed by a linear combination of the rows or columns of the genotype matrix is reduced. The degree of reduction of both spaces is controlled by the number of core animals. With a fixed number of core animals, selecting different sets of core animals is equivalent to selecting different subspaces of the row and column spaces of $\mathbf{Z}$. When the number of core animals is such that a large portion of the spectrum of $\mathbf{G}$ (or the set of singular values of $\mathbf{Z}$) is covered, those sub-spaces that correspond to different sets of core animals will increasingly overlap, which means that their intersection will not be equal to the null vector space. In that case, many genotypes in the population can be generated as linear combinations of the rows of $\mathbf{Z}_c$. If a certain genotype, say $\mathbf{z}_i$, cannot be formed, the distance between $\mathbf{z}_i$ and its projection to the row space of $\mathbf{Z}_c(d(\mathbf{z}_i - \mathcal{P}\mathbf{z}_i))$ is in general expected to be small. The particular choice of core animals is not important, as long as it covers the spectrum of $\mathbf{G}$. Although the

heuristics of how core animals should be chosen needs to be refined, a wide random choice appears to be adequate [10, 13, 27, 28]. However, as the population evolves and new haplotype combinations are created, the chosen core may become less representative, although this is expected to be a slow process. Methods to choose and update a set of core animals that spans most variability in **G**, e.g., [29], is an open area of research.

The large changes in estimated breeding values for some non-core animals when the core animals are changed, while keeping its number constant, as reported by [30], can be explained by a large distance between the projected and the real genotype of those individuals. This is the case, for instance, when a mislabeled animal is evaluated within one breed but actually belongs to another breed.

Note also that core animals do not need to be individuals with, a priori, high reliability. Marker effects are back-solved from core animals, but these animals gather information from the entire population through the genomic relationships. Thus, the accuracy of the estimated breeding values of these core animals will be very high. For instance, in dairy cattle, core animals could be based on cows. If these cows have a strong relationship with the whole population, they are very accurately estimated, and so will be the SNP effects and indirect predictions.

We also derived the distribution of breeding values conditional on the marker effects and vice versa when using APY. For the first case, breeding values of the non-core animals are only partially explained by the marker effects because of the error term $\boldsymbol{\xi}$ in Eq. (10). For the second case, we showed that the BLUP of the marker effects conditional on the breeding values of animals, only requires the matrices and estimated breeding values corresponding to the core animals—this result was not known before. Although non-core animals do not appear in the explicit calculation of the marker effects, their information is used to estimate breeding values of the core animals. When Eq. (21) is not used to obtain estimates for the marker effects from an APY-GBLUP model, those estimates will not have minimum variance.

In genetic evaluations, indirect predictions for animals without own records or progeny with records can be calculated from estimates of marker effects that are obtained by back-solving from the estimated breeding values [20]. If an APY-(ss)GBLUP model is used for genetic evaluations then the proper way to calculate indirect predictions is based on the distribution in Eq. (18). For core animals, the formula to calculate indirect predictions is equal to the Eq. (10) in [20]. However, for genetic evaluation, animals for which indirect predictions are calculated are, by definition, non-core animals. Therefore, in that case, their indirect predictions must be obtained from Eq. (23).

Two measures of uncertainty associated with indirect predictions were derived. The first measure of uncertainty, Eq. (24), quantifies how different the indirect prediction would be from that calculated with a regular GBLUP or SNP-BLUP model. If the number of core animals is large enough, this measure will be close to 1. A value of 1 indicates that the individual is in the space of genetic variation described by the core animals and indirect predictions based on APY and SNP-BLUP are identical. The second measure of uncertainty, based on Eq. (26), is a classical reliability of estimated breeding value and is a function of the prediction error variance of the indirect prediction in Eq. (25). This reliability can be expressed in terms of prediction error variances of marker effects or in terms of prediction error variances of core animals. The latter results in higher computational efficiency because the number of core animals is smaller than the number of markers.

In the same way that an equivalent SNP-BLUP model exists for the APY-BLUP model, we showed that when genotyped and non-genotyped animals are combined in the evaluation using a single-step approach, there is an APY-ssSNP-BLUP model that is equivalent to the APY-ssGBLUP model. In the APY-ssGBLUP model, breeding values are jointly estimated for non-genotyped, core, and non-core animals, while the APY-ssSNP-BLUP model estimates SNP effects based on the core animals, an error term for non-core animals, and the genotype imputation error for non-genotyped animals. Estimates of breeding values for all the animals can then be obtained by a linear combination of the corresponding design matrices and the vector of solutions. As in Fernando et al. [31], estimated breeding values for non-genotyped animals can be obtained directly to improve computational efficiency. Adding the polygenic effect does not change the MME for APY-ssGBLUP but adds an extra term ($\widehat{\boldsymbol{\varepsilon}}$) to the MME for APY-ssSNP-BLUP. In that case, the number of unknowns is equal to that in the model presented by [22]. Therefore, APY-ssSNP-BLUP is more complex and involves more convoluted matrix multiplications. Whenever the MME are augmented, there is a question on the feasibility and convergence of the model with real datasets. Vandenplas et al. [5] proposed a second-level preconditioner to ease convergence problems in ssSNP-BLUP, and Vandenplas et al. [6] presented a different termination criterion to determine convergence of such models. Although APY-ssSNP-BLUP is more flexible regarding the use of different priors for marker effects [25], its convergence will need to be monitored carefully.

Conclusions

The APY-GBLUP model is equivalent to a family of marker effect models that are here described as APY-SNP-BLUP. We show that when the choice of core animals covers the rank of the genotype matrix, which is generally equal to the number of markers, APY-GBLUP is equivalent to a conventional SNP-BLUP model and is, therefore, just as accurate. If the choice of core animals does not cover the spectrum of the genomic relationship matrix, the genotypes for the non-core animals are imputed as a linear combination of the genotypes of the core animals. Marker effect estimates for the APY-SNP-BLUP model can then be calculated by a linear transformation of the estimated breeding values of the core animals. Thus, all the matrices involved have a size equal to the number of core animals. Indirect predictions for non-core animals with APY can be calculated from estimated marker effects from APY-SNP-BLUP, or from estimated breeding values for core animals, without a need to consider non-core animals, which simplifies the calculations. The reliability of indirect predictions is solely a function of the prediction error variance of the estimated breeding values of core animals. Therefore, choosing core animals that cover correctly the spectra of the genotype matrix will give indirect predictions with high reliability. Both the APY-GBLUP and the APY-SNP-BLUP models can fit a residual polygenic effect, and when the APY-SNP-BLUP is used, only residual polygenic effects for core animals are fitted.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12711-022-00741-7.

Additional file 1. Weighted marker effects in APY-SNP-BLUP and numerical illustration of the equivalence between APY-GBLUP and APY-SNP-BLUP with and without a residual polygenic effect.

Acknowledgements
In the year of his retirement, the authors would like to thank Dr. Rohan Fernando for his extensive and valuable contribution to the field of statistics and animal breeding and genetics.

Author contributions
MB conceived the initial idea for the study, derived the formulas, and tested the algebra in the simulated datasets. AL suggested to include the topic "indirect predictions" in the study. IM, DL, and AL mentored MB and took part in the discussions. MB and DL co-wrote the manuscript. IM, AL, and NSF edited the manuscript. All authors read and approved the final manuscript.

Funding
This study was partially funded by Agriculture and Food Research Initiative Competitive Grant No. 2020-67015-31030 from the US Department of Agriculture's National Institute of Food and Agriculture. This project has received funding from the European Unions' Horizon 2020 Research & Innovation program under grant agreement No 772787 -SMARTER.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602, USA. [2]Facultad de Agronomía, Universidad de Buenos Aires, C1417DSQ Buenos Aires, Argentina. [3]Instituto de Investigaciones en Producción Animal (INPA), CONICET - Universidad de Buenos Aires, C1427CWO Buenos Aires, Argentina. [4]INRA, UMR1388 GenPhySE, 31326 Castanet-Tolosan, France.

Received: 22 October 2021 Accepted: 29 June 2022
Published online: 16 July 2022

References

1. Strandén I, Garrick DJ. Technical note: derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J Dairy Sci. 2009;92:2971–5.
2. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
3. Henderson CR. Applications of linear models in animal breeding. Guelph: University of Guelph; 1984.
4. Vandenplas J, Eding H, Calus MPL, Vuik C. Deflated preconditioned conjugate gradient method for solving single-step BLUP models efficiently. Genet Sel Evol. 2018;50:51.
5. Vandenplas J, Calus MPL, Eding H, Vuik C. A second-level diagonal preconditioner for single-step SNPBLUP. Genet Sel Evol. 2019;51:30.
6. Vandenplas J, Calus MPL, Eding H, van Pelt M, Bergsma R, Vuik C. Convergence behavior of single-step GBLUP and SNPBLUP for different termination criteria. Genet Sel Evol. 2021;53:34.
7. Misztal I, Lourenco D, Legarra A. Current status of genomic evaluation. J Anim Sci. 2020;98:skaa101.
8. Misztal I, Legarra A, Aguilar I. Using recursion to compute the inverse of the genomic relationship matrix. J Dairy Sci. 2014;97:3943–52.
9. Misztal I. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. Genetics. 2016;202:401–9.
10. Mäntysaari EA, Evans RD, Strandén I. Efficient single-step genomic evaluation for a multibreed beef cattle population having many genotyped animals. J Anim Sci. 2017;95:4728–37.
11. Harville DA. Matrix algebra from a statistician's perspective. New York: Springer; 2008.
12. Ødegård J, Indahl U, Strandén I, Meuwissen THE. Large-scale genomic prediction using singular value decomposition of the genotype matrix. Genet Sel Evol. 2018;50:6.
13. Fernando RL, Cheng H, Garrick DJ. An efficient exact method to obtain GBLUP and single-step GBLUP when the genomic relationship matrix is singular. Genet Sel Evol. 2016;48:80.
14. Fragomeni BO, Lourenco DA, Tsuruta S, Masuda Y, Aguilar I, Legarra A, et al. Hot topic: Use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes. J Dairy Sci. 2015;98:4090–4.
15. Pocrnic I, Lourenco DA, Masuda Y, Legarra A, Misztal I. The dimensionality of genomic information and its effect on genomic prediction. Genetics. 2016;203:573–81.

16. Wiggans GR, Sonstegard TS, VanRaden PM, Matukumalli LK, Schnabel RD, Taylor JF, et al. Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. J Dairy Sci. 2009;92:3431–6.
17. Vitezica ZG, Aguilar I, Misztal I, Legarra A. Bias in genomic predictions for populations under selection. Genet Res. 2011;93:357–66.
18. Snelson E, Ghahramani Z. Local and global sparse Gaussian process approximations. In Proceedings of the 11th International Conference on Artificial Intelligence and Statistics: 21–24 March 2007; San Juan. 2007. https://proceedings.mlr.press/v2/snelson07a.html
19. Cuevas J, Montesinos-López OA, Martini JWR, Pérez-Rodríguez P, Lillemo M, Crossa J. Approximate genome-based Kernel models for large data sets including main effects and interactions. Front Genet. 2020;11: 567757.
20. Garcia ALS, Masuda Y, Tsuruta S, Miller S, Misztal I, Lourenco D. Indirect predictions with a large number of genotyped animals using the algorithm for proven and young. J Anim Sci. 2020;98:skaa154.
21. Ben Zaabza H, Mäntysaari EA, Strandén I. Using Monte Carlo method to include polygenic effects in calculation of SNP-BLUP model reliability. J Dairy Sci. 2020;103:5170–82.
22. Liu Z, Goddard ME, Reinhardt F, Reents R. A single-step genomic model with direct estimation of marker effects. J Dairy Sci. 2014;97:5833–50.
23. Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic information. J Dairy Sci. 2009;92:4656–63.
24. Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J Dairy Sci. 2010;93:743–52.
25. Fernando RL, Dekkers JC, Garrick DJ. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genet Sel Evol. 2014;46:50.
26. Masuda Y, Misztal I, Legarra A, Tsuruta S, Lourenco DA, Fragomeni BO, et al. Technical note: avoiding the direct inversion of the numerator relationship matrix for genotyped animals in single-step genomic best linear unbiased prediction solved with the preconditioned conjugate gradient. J Anim Sci. 2017;95:49–52.
27. Pocrnic I, Lourenco DA, Masuda Y, Misztal I. Dimensionality of genomic information and performance of the Algorithm for Proven and Young for different livestock species. Genet Sel Evol. 2016;48:82.
28. Vandenplas J, Calus MPL, Ten Napel J. Sparse single-step genomic BLUP in crossbreeding schemes. J Anim Sci. 2018;96:2060–73.
29. Pocrnic I, Lindgren F, Gorjanc G. Optimised core subset construction for the APY model. In Proceedings of the 72nd Annual Meeting of the European Federation of Animal Science: 30 August-3 September 2021; Davos.
30. Misztal I, Tsuruta S, Pocrnic I, Lourenco D. Core-dependent changes in genomic predictions using the Algorithm for Proven and Young in single-step genomic best linear unbiased prediction. J Anim Sci. 2020;98:skaa374.
31. Fernando RL, Cheng H, Golden BL, Garrick DJ. Computational strategies for alternative single-step Bayesian regression models with large numbers of genotyped and non-genotyped animals. Genet Sel Evol. 2016;48:96.

Publisher's Note