A method for partitioning trends in genetic mean and variance to understand breeding practices

Oliveira, T. P. [1*], Obšteter, J. [2], Pocrnic, I. [1], Heslot, N. [3], Gorjanc, G. [1*]

1 The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, UK
2 Agricultural Institute of Slovenia, Ljubljana, Slovenia
3 Limagrain, Saint-Beauzire, France

*** thiago.oliveira@ed.ac.uk**
*** gregor.gorjanc@roslin.ed.ac.uk**

Abstract

Background: In breeding programmes, the observed genetic change is a sum of the contributions of different groups of individuals. Quantifying these sources of genetic change is essential for identifying the key breeding actions and optimizing breeding programmes. However, it is difficult to disentangle the contribution of individual groups due to the inherent complexity of breeding programmes. Here we extend the previously developed method for partitioning genetic mean by paths of selection to work both with the mean and variance of breeding values.

Methods: We first extended the partitioning method to quantify the contribution of different groups to genetic variance assuming breeding values are known. Second, we combined the partitioning method with the Markov Chain Monte Carlo approach to draw samples from the posterior distribution of breeding values and use these samples for computing the point and interval estimates of partitions for the genetic mean and variance. We implemented the method in the `R` package `AlphaPart`. We demonstrated the method with a simulated cattle breeding programme.

Results: We showed how to quantify the contribution of different groups of individuals to genetic mean and variance. We showed that the contributions of different selection paths to genetic variance are not necessarily independent. Finally, we observed some limitations of the partitioning method under a misspecified model, suggesting the need for a genomic partitioning method.

Conclusion: We presented a partitioning method to quantify sources of change in genetic mean and variance in breeding programmes. The method can help breeders and researchers understand the dynamics in genetic mean and variance in a breeding programme. The developed method for partitioning genetic mean and variance is a powerful method for understanding how different paths of selection interact within a breeding programme and how they can be optimised.

Keywords: Genetic trend, Genetic mean, Genetic variance, Partitioning, Bayesian, MCMC, `AlphaPart`

1 Background

Analysing genetic trends is essential for identifying key breeding actions and optimizing breeding programmes. The observed genetic change is a sum of contributions from different groups of individuals, often referred to as selection pathways or paths. However, these contributions are difficult to quantify due to the inherent complexity of breeding programmes. The contributions of different groups vary due to different selection intensity, accuracy, genetic variation, generation interval, and dissemination. To quantify the contributions, GarciaCortes et al. [1] developed a method for analysing the change in the genetic mean by partitioning the breeding values into the contributions of several paths. The method uses the standard partitioning of an individual's breeding value a_k into parent average $\frac{1}{2}a_{f(k)} + \frac{1}{2}a_{m(k)}$ and Mendelian sampling term w_k. Further, the method assigns Mendelian sampling terms to analyst-defined groups, such as sex, origin, selection path, etc. By aggregating these partitions by other variables, such as year, the method summarises the contributions of different groups of individuals to the overall genetic trend. This approach has been used to quantify the contributions of different countries to the overall genetic trend in the global Brown Swiss population [2], global and local Holstein populations [2], as well as Croatian Simmental cattle [3], and Croatian Landrace and Large-White pigs [4]. More recently, the method was used to estimate the starting point of adopting genomic selection by quantifying differences in genetic trends estimated with pedigree-based and single-step genomic BLUP [5].

In addition to the contribution of paths to changes in genetic mean, breeding programmes should also consider analysing changes in genetic variance to fully understand the drivers of genetic change in a population [6, 7]. Managing the change in genetic mean and variance in breeding programmes is essential to ensure a long-term genetic gain [8, 9]. Therefore, it is crucial that we quantify the contribution of different groups of individuals in a breeding programme to the genetic mean and variance. For example, in several economically important species, male selection and dissemination represent a key lever that has the largest impact on the genetic mean and variance in a population.

The aim of this paper is to extend the method of GarciaCortes et al. [1] to i) partition the genetic mean and variance, ii) implement the method in `AlphaPart R` package, and iii) apply the partitioning method to estimated breeding values following the work of [6] and [7]. We used simulation to demonstrate the methodology and give insights on how to use the `AlphaPart R` package [10] to analyse real data. The remainder of the paper is organised as follows. Section 2 presents the theory of the partitioning method, the extension to partition the genetic variance, and an empirical approach when breeding values are estimated from data. Results are presented in Section 3, where it is shown that the developed approach works well and covers the true values. Generalizations and limitations of the approach are discussed in Section 4, and the conclusions in Section 5.

2 Methods

Subsection 2.1 describes the theory of partitioning genetic mean and variance. Estimation of breeding values and partitions from phenotype data with a statistical model and a Markov Chain Monte Carlo (MCMC) method are presented in the subsection 2.2. Subsection 2.3 describes the measures of agreement and bias used to analyse the breeding values and partitions. Subsection 2.4 shows `AlphaPart` implementation working with MCMC samples of breeding values. Next, subsection 2.5 describes simulation of a cattle breeding programme used to demonstrate an `AlphaPart` analysis. Finally, a summary of the used software is described in 2.6.

2.1 Partitioning theory

Let $\boldsymbol{a}$ be a vector of breeding values following a normal distribution with mean $\mathbf{0}$ and pedigree-based covariance $\boldsymbol{A}\sigma_a^2$. We can write $\boldsymbol{a}$ as a linear combination of the individual's ancestor breeding values and individual's deviation from the ancestors $\boldsymbol{a} = \boldsymbol{T}\boldsymbol{w}$, where $\boldsymbol{T}$ is a lower-triangular matrix of expected gene flow between ancestors and individuals following a pedigree, and $\boldsymbol{w} \sim N\left(\mathbf{0}, \boldsymbol{W}\sigma_a^2\right)$ are Mendelian sampling terms representing the deviations, with $\boldsymbol{W}$ being a diagonal matrix of variance coefficients and σ_a^2 the base population (additive) genetic variance [11, 12, 13, 14].

Assuming a factor with p levels, representing our paths (groups) of interest, and for any set $\sum_{j=1}^{p} \boldsymbol{P}_j = \boldsymbol{I}$, Garcia-Cortes et al. [1] partitioned the gene flow matrix into contributions of each path by defining $\boldsymbol{T}_j = \boldsymbol{T}\boldsymbol{P}_j$, $j = 1, 2, \ldots, p$, and further partitioned the contribution of each path to breeding values *a priori* using the equality:

$$\boldsymbol{a} = (\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_p)\, \boldsymbol{w} = \boldsymbol{a}_1 + \boldsymbol{a}_2 + \ldots + \boldsymbol{a}_p. \tag{1}$$

Garcia-Cortes et al. [1] further showed that these contributions can be estimated from data collected in breeding programmes (*a posteriori*). They first calculated the conditional expectation of breeding values given phenotype data ($\boldsymbol{y}$), $EBV = \widehat{\boldsymbol{a}} = E\left(\boldsymbol{a}|\boldsymbol{y}\right)$. Then they plugged the EBV ($\widehat{\boldsymbol{a}}$) and estimated Mendelian sampling terms ($\widehat{\boldsymbol{w}}$) into the equation (1). This approach enabled them to estimate the conditional expectation of partitions, that is, $\widehat{\boldsymbol{a}} = E\left(\boldsymbol{a}|\boldsymbol{y}\right)$:

$$\widehat{\boldsymbol{a}} = (\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_p)\, \widehat{\boldsymbol{w}} = \widehat{\boldsymbol{a}}_1 + \widehat{\boldsymbol{a}}_2 + \ldots + \widehat{\boldsymbol{a}}_p. \tag{2}$$

By summarising the breeding values partitions over time, Garcia-Cortes et al. [1] quantified the contribution of each path (for example, males vs. females, different countries, etc.) to the change in individual breeding values and their average over time. Technically this is achieved by sub-setting the $\widehat{\boldsymbol{a}}_1, \widehat{\boldsymbol{a}}_2, \ldots, \widehat{\boldsymbol{a}}_p$ and averaging each subset. This method has been implemented in the `AlphaPart` `R` package [10, 15]. The `AlphaPart` `R` package efficiently calculates the partitions by leveraging the sparse $\boldsymbol{T}^{-1}$ [11, 12, 13, 14], and enables a straightforward summary by one variable, such as year, or combination of variables (interaction), such as year and sex. The levels of this variable are used to compute the conditional expectation and we refer to it as x_t^*, with $t = 1, 2, \ldots, m$, and m representing the number of distinct categories. Importantly, `AlphaPart` enables use of any function to summarise the partitions of breeding values, that is, $f\left(\boldsymbol{a}_j\right)$.

To enable the use of variance as one of the summary functions in `AlphaPart`, we are here extending the partitioning method to analyse the contribution of paths to genetic variance. Variance of breeding values is, *a priori*, $Var\left(\boldsymbol{a}\right) = Var\left(\boldsymbol{T}\boldsymbol{w}\right) = \boldsymbol{T}\boldsymbol{W}\boldsymbol{T}^T\sigma_a^2$. Using the equation (1), we can further partition the genetic

variance by paths as:

$$
\begin{aligned}
Var\left(\boldsymbol{a}\right) &= Var\left[\left(\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_p\right)\boldsymbol{w}\right], \\
&= \sum_{j=1}^{p} Var\left(\boldsymbol{T}_j\boldsymbol{w}\right) + 2\sum_{j=1}^{p-1}\sum_{j'=j+1}^{p} Cov\left(\boldsymbol{T}_j\boldsymbol{w}, \boldsymbol{T}_{j'}\boldsymbol{w}\right), \\
&= \sum_{j=1}^{p} \boldsymbol{T}_j\boldsymbol{W}\boldsymbol{T}_j^T\sigma_a^2 + 2\sum_{j=1}^{p-1}\sum_{j'=j+1}^{p} \boldsymbol{T}_j\boldsymbol{W}\boldsymbol{T}_{j'}^T\sigma_a^2, \\
&= \left(\sum_{j=1}^{p}\boldsymbol{A}_j + 2\sum_{j=1}^{p-1}\sum_{j'=j+1}^{p}\boldsymbol{A}_{j,j'}\right)\sigma_a^2,
\end{aligned} \tag{3}
$$

where $\boldsymbol{A}_j = \boldsymbol{T}_j\boldsymbol{W}\boldsymbol{T}_j^T$ and $\boldsymbol{A}_{j,j'} = \boldsymbol{T}_j\boldsymbol{W}\boldsymbol{T}_{j'}^T$. While this partitioning by paths involves dense matrices such as $\boldsymbol{A}_j$ and $\boldsymbol{A}_{j,j'}$, we can efficiently calculate the partitions $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_p$ by working with the sparse $\boldsymbol{T}^{-1}$ [11, 12, 13, 14, 16]. We can again assume x_t^* as a variable with m distinct categories, $t = 1, 2, \ldots, m$, used to summarise the paths. Thus, we can define the genetic variance for the partition j given category t that has $n_k \leq nI$ individuals, $k^* = 1, 2, \ldots, n_k$, as:

$$
\begin{aligned}
Var\left(\boldsymbol{a}_{j_t}\right) &= E\left(\boldsymbol{a}_{j_t}^2\right) - E^2\left(\boldsymbol{a}_{j_t}\right), \\
&= \frac{1}{n_k}\sum_{k^*=1}^{n_k}\left(a_{j_t,k^*} - \bar{a}_{j_t}\right)^2,
\end{aligned} \tag{4}
$$

where $\boldsymbol{a}_{j_t}$ is a column for partition j, but only considering individuals in category t, n_k is the number of individuals in category t, and $\bar{a}_{j_t} = \dfrac{1}{n_k}\displaystyle\sum_{k^*=1}^{n_k} a_{j_t,k^*}$. Similarly, genetic covariance between the partitions j and j', $j \neq j'$, given category t is then:

$$
\begin{aligned}
Cov\left(\boldsymbol{a}_{j_t}, \boldsymbol{a}_{j'_t}\right) &= E\left(\boldsymbol{a}_{j_t}\boldsymbol{a}_{j'_t}\right) - E\left(\boldsymbol{a}_{j_t}\right)E\left(\boldsymbol{a}_{j'_t}\right), \\
&= \frac{1}{n_k}\sum_{k^*=1}^{n_k}\left(a_{j_t,k^*} - \bar{a}_{j_t}\right)\left(a_{j'_t,k^*} - \bar{a}_{j'_t}\right).
\end{aligned} \tag{5}
$$

By sub-setting the partitions $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_p$ by additional variable x_t^*, such as year, we can calculate (4) and (5) for each level of the additional variable.

It is worth noting that there is a difference between (3) and (4) or (5). The σ_a^2 in (3) represents the genetic variance in a base population, while the expression $\left(\sum_{j=1}^{p}\boldsymbol{A}_j + 2\sum_{j=1}^{p-1}\sum_{j'=j+1}^{p}\boldsymbol{A}_{j,j'}\right)\sigma_a^2$ describes how that variance *a priori* changes through a given pedigree and how that variance can be partitioned by paths. Equations (4) and (5) represent the variance and covariance of breeding value partitions $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_p$, that contribute to the total genetic variance, but conditioned on the category t of x_t^*, that is, calculated just for individuals in the category t. Therefore, we can partition the genetic variance of a population into path contributions, which can be summarised in the same ways as genetic mean [1, 17]. Such analyses can quantify the contribution of different paths to changes in genetic mean and variance over time.

The presented partitioning of genetic variance holds for *a priori* breeding values or true breeding values. However, when EBV are available, we can not substitute $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_p$ with their expectations $\widehat{\boldsymbol{a}}_1, \widehat{\boldsymbol{a}}_2, \ldots, \widehat{\boldsymbol{a}}_p$ in (4) and (5) as Garcia-Cortes et al. [1] could do it for the partitioning of genetic mean. To see this, imagine a situation where EBV are based on very limited phenotype information. Such EBV will be shrunken strongly

towards zero and will have low accuracy [14]. As such, these EBV will not be a good representation of true breeding values, and their variance, $Var(EBV) = Var\left(E\left(\boldsymbol{a}|\boldsymbol{y}\right)\right)$ will be much smaller than variance of true breeding values $Var\left(\boldsymbol{a}\right)$. To address this issue, we use the approach from Sorensen et al. [6] and Lara et al. [7] that involves three steps. First, sample breeding values from their posterior distribution [16]. Second, for every sample of breeding values calculate desired quantities, in our case breeding value partitions and their means and variances. Multiple samples of these quantities describe their posterior distribution. Third, summarise the samples to describe the posterior distributions of interest, in our case partitions of genetic mean:

$$\begin{aligned} E\left(\boldsymbol{a}|\boldsymbol{y}\right) &= E\left(\boldsymbol{a}_1 + \boldsymbol{a}_2 + \ldots + \boldsymbol{a}_p|\boldsymbol{y}\right) \\ &= E\left(\boldsymbol{a}_1|\boldsymbol{y}\right) + E\left(\boldsymbol{a}_2|\boldsymbol{y}\right) + \ldots + E\left(\boldsymbol{a}_p|\boldsymbol{y}\right) \end{aligned}$$

and partitions of genetic variance:

$$\begin{aligned} Var\left(\boldsymbol{a}|\boldsymbol{y}\right) &= Var\left(\boldsymbol{a}_1 + \boldsymbol{a}_2 + \ldots + \boldsymbol{a}_p|\boldsymbol{y}\right) \\ &= Var\left(\boldsymbol{a}_1|\boldsymbol{y}\right) + Var\left(\boldsymbol{a}_2|\boldsymbol{y}\right) + \ldots + \\ &\quad + Var\left(\boldsymbol{a}_p|\boldsymbol{y}\right) + 2\left[Cov\left(\boldsymbol{a}_1, \boldsymbol{a}_2|\boldsymbol{y}\right) + \right. \\ &\quad \left. + \ldots + Cov\left(\boldsymbol{a}_{p-1}, \boldsymbol{a}_p|\boldsymbol{y}\right)\right] \end{aligned}$$

or their variants when we only considering individuals in a specific category.

2.2 Statistical model and computational approaches

In the previous subsection, we assumed we knew the true breeding values. However, in reality, we infer breeding values from phenotype data. To this end, we fitted the standard pedigree-based model to observed data:

$$\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{a} \sim N\left(\boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{a}, \boldsymbol{I}\sigma_e^2\right), \tag{6}$$

$$\boldsymbol{a} \sim N\left(\boldsymbol{0}, \boldsymbol{A}\sigma_a^2\right), \tag{7}$$

where $\boldsymbol{y}$ is a vector of observed phenotypes, $\boldsymbol{b}$ is a vector of fixed effects with the design matrix $\mathbf{X}$, $\mathbf{a}$ is a vector of breeding values with the design matrix $\mathbf{Z}$, σ_e^2 is a residual variance, $\boldsymbol{A}$ is pedigree-based relationship matrix and σ_a^2 is genetic variance in the base population.

The directed acyclic graph (DAG) representing the model (7) considering only intercept as the fixed effect is illustrated in Figure 1, where pedigree and phenotypic records are displayed in separate plates as a generalization of the case where animals might have a phenotypic record, and the dotted lines indicate a possibly missing parent in the pedigree. In the pedigree plate we have nI individuals represented by founders and non-founders, where founder's breeding values are `a priori` distributed as $p\left(a_k|\sigma_a^2\right) \sim N\left(0, \sigma_a^2\right)$. Breeding value of non-founders given their parents are then represented by $a_k = \frac{1}{2}a_{f(k)} + \frac{1}{2}a_{m(k)} + w_k$, where $a_{f(k)}$ and $a_{m(k)}$ are parent's breeding values and w_k is the Mendelian sampling term $p\left(w_k|\boldsymbol{W}_{k,k}\right) \sim N\left(0, \boldsymbol{W}_{k,k}\sigma_a^2\right)$.

Following equation (3) matrix $\boldsymbol{A}$ can be decomposed as $\boldsymbol{A} = \boldsymbol{T}\boldsymbol{W}\boldsymbol{T}^T$ using the generalised Cholesky (LDL) decomposition [11, 12, 13, 14]. The diagonal elements of $\boldsymbol{W}$ can be computed according to specific scenarios described by [11, 12, 13, 14] as i) $\boldsymbol{W}_{k,k} = \frac{1}{2} - \frac{1}{4}\left(F_{f(k)} + F_{m(k)}\right)$ when both parents are known; ii) $\boldsymbol{W}_{k,k} = \frac{3}{4} - \frac{1}{4}F_{m(k)}$ or $\boldsymbol{W}_{k,k} = \frac{3}{4} - \frac{1}{4}F_{f(k)}$ when one parent is known; and iii) $\boldsymbol{W}_{k,k} = 1$ when both parents

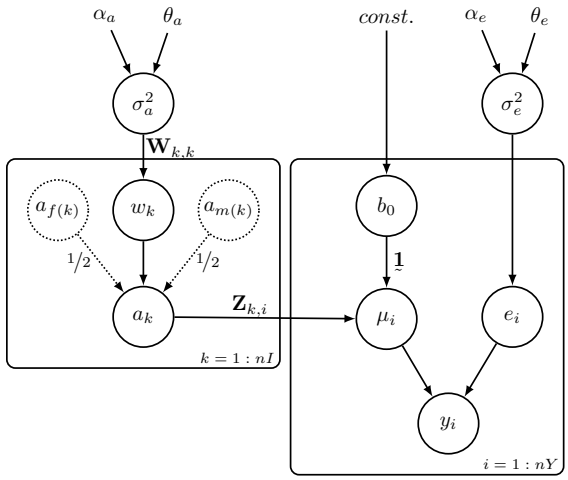

Figure 1. Directed acyclic graph of the pedgiree-based model with nI individuals and nY phenotypic records (y_i) with explicit representation of Mendelian sampling terms (w_k) and error term (e_i), where σ_a^2 is the additive genetic variance, $a_{f(k)}$ and $a_{m(k)}$ are parent's breeding value, $\mathbf{1}$ represents a vector of ones, μ_i the linear predictor, and σ_e^2 the residual variance

are unknown, where $F_{f(k)}$ and $F_{m(k)}$ are respectively the coefficients of inbreeding of the father and mother of the individual k [11, 12, 13, 14]. While this theory is standard, some quantitative genetics software still ignore parental inbreeding coefficients in setting up the $\boldsymbol{A}^{-1}$, which can have significant impact on the analysis of genetic variance as we will show in results.

We used the full Bayesian approach to infer breeding values using (7). To this end we specified prior distribution for all model parameters, as shown in Figure 1. Thus, $\boldsymbol{b}$, σ_a^2, and σ_e^2 were assumed to have a joint prior density of the form $p\left(\boldsymbol{b}, \sigma_a^2, \sigma_e^2\right) = p\left(\boldsymbol{b}\right) p\left(\sigma_a^2\right) p\left(\sigma_e^2\right)$ with a flat distribution for b and a conjugate inverse-gamma(α, θ) distribution for variances (or gamma distribution for precisions = 1/variance), where α and θ are set to a value such as 0.1[3]:

$$
\begin{aligned}
p\left(\boldsymbol{b}\right) &\propto const., \\
p\left(\tau_a = 1/\sigma_a^2 | \alpha_a, \theta_a\right) &\propto \tau_a^{\alpha_a - 1} \exp\left(-\theta_a \tau_a\right), \\
p\left(\tau_e = 1/\sigma_e^2 | \alpha_e, \theta_e\right) &\propto \tau_e^{\alpha_e - 1} \exp\left(-\theta_e \tau_e\right),
\end{aligned}
$$

with $\tau_a > 0$, $\alpha_a \geq 0$, $\theta_a \geq 0$, $\tau_e > 0$, $\alpha_e \geq 0$ and $\theta_e \geq 0$. The posterior distribution is then obtained by applying

$$
\begin{aligned}
p\left(\boldsymbol{b}, \boldsymbol{a}, \sigma_a^2, \sigma_e^2 | \boldsymbol{y}\right) \propto & p\left(\boldsymbol{y} | \boldsymbol{b}, \boldsymbol{a}, \sigma_e^2\right) p\left(\boldsymbol{b}\right) p\left(\boldsymbol{a} | \boldsymbol{A}, \sigma_a^2\right) \times \\
& p\left(\sigma_a^2 | \alpha_a, \theta_a\right) p\left(\sigma_e^2 | \alpha_e, \theta_e\right).
\end{aligned}
$$

the Bayes' theorem conditional on the data:

We sampled from the posterior distribution with a Markov Chain Monte Carlo (MCMC) algorithm (the Gibbs algorithm) as implemented in [18]. We constructed one chain with 80,000 samples, from which 20,000 were considered as burn-in, while the remaining 60,000 were stored and thinned by saving every 40-th sample. We assessed the MCMC convergence by inspecting the trace and auto-correlation plots. Consequently, 1,500 samples of breeding values were stored representing the posterior distribution $p\left(\boldsymbol{a}|\boldsymbol{y}\right)$. These samples were passed as input to the `AlphaPart` `R` package.

It is imperative to note that the proposed partitioning method requires samples from the posterior

distribution $p\left(\boldsymbol{a}|\boldsymbol{A},\sigma_a^2\right)$ to enable inference of the path contributions to both genetic mean and variance. While we have used the full Bayesian approach with MCMC [16], an alternative is to use the empirical Bayesian approach, that is, estimating variance components with REML and sampling breeding values assuming that variance components are known (see [7]). Obviously, the full Bayesian approach is recommended to account for uncertainty in estimating all model parameters.

2.3 Measures of agreement and bias

The partitioning methodology depends on unbiased estimates of breeding values. If the used model (7) does not adequately describe the data, we can end up with biased inference of $\boldsymbol{a}$. Let t be the index for generation with $t = 1, 2, \ldots, m$. In simulations we can calculate $E\left(\boldsymbol{a}_t\right)$, $E\left(\boldsymbol{a}_t|\boldsymbol{y}\right)$, $Var\left(\boldsymbol{a}_t\right)$, and $Var\left(\boldsymbol{a}_t|\boldsymbol{y}\right)$, where $\boldsymbol{a}_t$ and $\boldsymbol{a}_t|\boldsymbol{y}$ are respectively vectors containing the true and inferred breeding values of all individuals from generation t. To evaluate unbiasedness of breeding values, we can evaluate the agreement between $\boldsymbol{Y}_1^* = \left(E\left(\boldsymbol{a}_1\right), E\left(\boldsymbol{a}_2\right), \ldots, E\left(\boldsymbol{a}_t\right)\right)$ and $\boldsymbol{Y}_2^* = \left(E\left(\boldsymbol{a}_1|\boldsymbol{y}\right), E\left(\boldsymbol{a}_2|\boldsymbol{y}\right), \ldots, E\left(\boldsymbol{a}_t|\boldsymbol{y}\right)\right)$. Note that the same holds if we are interested in the agreement between variances, we only need to replace $E(.)$ by $Var(.)$. The same holds for breeding value partitions.

Let us assume that the pairs of $\left(Y_{1,t}^*, Y_{2,t}^*\right)$ are independent draws from a bi-variate population with means μ_1 and μ_2 and a covariance matrix:

$$Cov(Y_{1t}^*, Y_{2t}^*) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

We can evaluate the agreement between $\boldsymbol{Y}_1^*$ and $\boldsymbol{Y}_2^*$ with the concordance correlation coefficient [19], which is an agreement index that lies between -1 and 1, and is given by:

$$\rho_c = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + \left(\mu_1 - \mu_2\right)^2}$$

where $\mu_1 = \mathrm{E}\left(\boldsymbol{Y}_1^*\right)$, $\mu_2 = \mathrm{E}\left(\boldsymbol{Y}_2^*\right)$, $\sigma_1^2 = \mathrm{Var}\left(\boldsymbol{Y}_1^*\right)$, $\sigma_2^2 = \mathrm{Var}\left(\boldsymbol{Y}_2^*\right)$, and $\sigma_{12} = \mathrm{Cov}\left(\boldsymbol{Y}_1^*, \boldsymbol{Y}_2^*\right)$. It can be shown that $\rho_c = \rho \times C_b$, where ρ is the Pearson correlation coefficient (a measure of precision), and C_b is the bias correction factor (a measure of accuracy). Here, ρ measures how far each observation deviates from the best-fit line, and $C_b \in [0, 1]$ measures how far the best-fit line deviates from the identity line $y = x$ and is defined as $C_b = 2\left(v + v^{-1} + u^2\right)^{-1}$, where $v = \sigma_1^2/\sigma_2^2$ is a scale shift and $u = (\mu_1 - \mu_2)/\sqrt{\sigma_1\sigma_2}$ is a location shift relative to the scale. When $C_b = 1$, there is no deviation from the identity line. In this sense, $1 - C_b$ can be seen as a measure of bias.

We can define bias as a measure of the systematic error of an estimator by evaluating the difference between the expected value of a statistic and the true value of its corresponding parameter. In this study, we also want to investigate the distribution of the difference between true and inferred quantities of interest. Our quantities of interest were partitions of genetic mean and variance over various categories. Let n_r be the number of simulation replicates, $r = 1, 2, ..., n_r$, and $\boldsymbol{a}_{j_t,r}$ and $\hat{\boldsymbol{a}}_{j_t,r}$, respectively, the j-th true and inferred partitions of breeding values for the category t at replicate r. As we are using a Bayesian approach, we can obtain the posterior distribution of our quantities of interest, $E\left(\boldsymbol{a}_{j_t,r}\right)$, $Var\left(\boldsymbol{a}_{j_t,r}\right)$, and $Cov\left(\boldsymbol{a}_{j_t,r}, \boldsymbol{a}_{j'_t,r}\right)$. We summarised these posterior distributions with the posterior mean and calculated the difference $\boldsymbol{z}_{j_t,r}$ between this posterior mean and the true value for means, variances, or covariances. Thus, we can assume

the difference $\boldsymbol{z}_{j_t,r}$ follows a normal distribution for a given partition j and category t:

$$Z_{j_t} \sim N\left(\mu_{j_t}, \sigma^2_{j_t}\right),$$

where μ_{j_t} and $\sigma^2_{j_t}$ represent the mean and variance of Z_{j_t}. Specifically, we inspected the distribution of the difference Z_{j_t} for each partition j and generations as categories indexed by t.

2.4 `AlphaPart` Implementation

We implemented the partitioning method in the `AlphaPart` `R` package [10, 17] with the summary method:

```
R> summary(object = part, by = "time", FUN = var)
```

where `object` is an object of class `AlphaPart` with path-partitioned breeding values, `by` represents the name of a column by which summary function `FUN` should be applied; and `FUN` is the summarising function. We also included an extra argument called `cov` that controls how the covariances will be displayed in the output. If `cov = FALSE`, the default, all covariances will be returned in a single column represented by $2\sum_{j=1}^{p-1}\sum_{j'=j+1}^{p} Cov(a_j, a_{j'})$, otherwise, if `cov = TRUE`, the `summary` method returns $p\left(p-1\right)/2$ columns, where each column represents covariances in the form of $2Cov(a_j, a_{j'})$.

We further describe how to use the posterior samples of breeding values from subsection 2.2 in `AlphaPart`. Let T be the number of traits and S be the number of samples of breeding values. Suppose `data` is a data frame containing columns for individual (`id`), father (`Fid`), mother (`Mid`), path (`path`), and generation (`Gen`). Now suppose a more general case where `bv_samples` represents a data frame containing a column for the individual (`id`) identification and S columns for the samples of breeding values, as shown in Figure 2. To prepare the input data for `AlphaPart`, we can merge the data frames called `data` and `bv_samples` into a new data frame called `newData` (Figure 2). We can then use the function `AlphaPart()` to calculate breeding value partitions with the difference that we now should pass the samples names to the argument `colBV` (Figure 2). Afterwards, the `summary()` function can be called to summarize the partitions using an explanatory variable, such as generation (`Gen`). Since we work with posterior samples of breeding values we obtain posterior samples for the summaries of the partitions (see the accompanying code). Finally, in the case with more than one trait, we suggest a `for` loop (possibly parallelized) to create one output per trait, as shown in Figure 2. In an extreme case with more traits than samples an alternative approach would be to save one sample of breeding values for multiple traits in one data frame and loop over the samples.

2.5 Simulation

To evaluate the method and `AlphaPart` implementation, we used `AlphaSimR` to simulate a simple cattle breeding programme over 40 generations with 1,000 individuals per generation. The first 20 years represented a burn-in phase, where we selected the best 5 males (out of 500) as sires based on their phenotype and mated them with all 500 females from the previous generation and all 500 females from the current generation. These matings produced 1,000 selection candidates for the next generation. After the burn-in phase, we tested two selection scenarios over further 20 generations: we selected 5 best males from 500 male candidates based on i) their phenotypes ('medium-accuracy' scenario, $r = 0.3$) or ii) true breeding values ('high-accuracy' scenario, $r = 1$), as shown in Figure 3. We replicated the simulation 30 times with the same founding genomes.

The simulation was done with the `AlphaSimR` `R` package version 1.0 [20]. We simulated a cattle genome with 30 chromosomes, 30,000 quantitative trait loci (QTL) with additive effects sampled from a normal distribution,

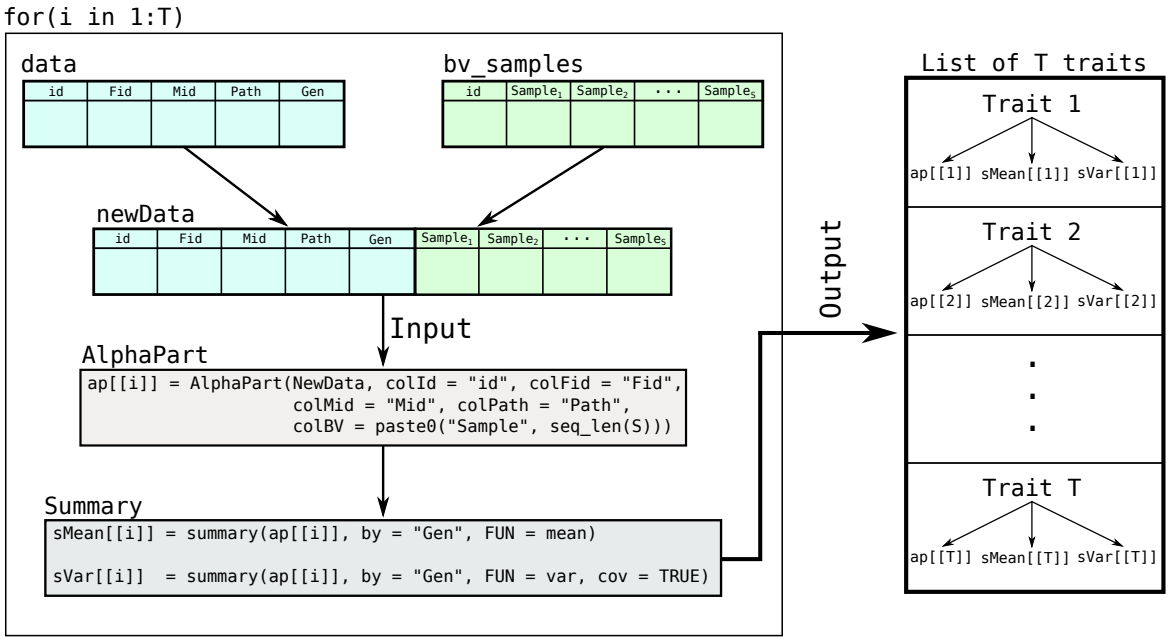

Figure 2. Flowchart representing a possible algorithm to evaluate contributions of paths to genetic mean and variance using samples of breeding values with the `AlphaPart` `R` package in a multi-trait case

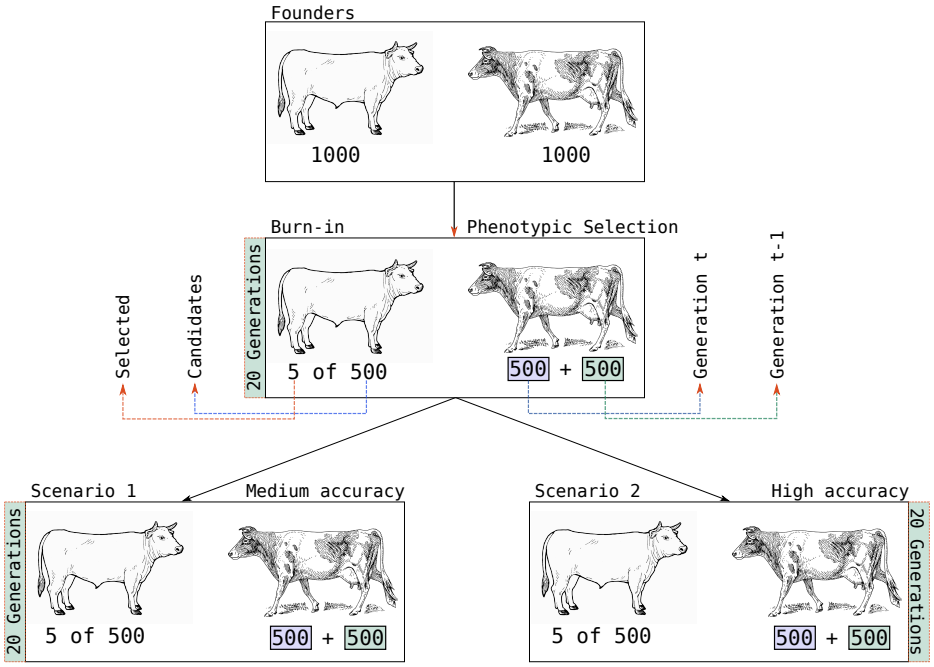

Figure 3. Simulation scheme illustrating an overview of the medium- and high-accuracy scenarios

single-trait phenotype with a heritability of 0.3. The above described breeding programme has a low effective population size (for example, $N_e \sim (4 \times nSires \times nDams)/(nSires + nDams) = (4 \times 5 \times 1000)/(1005) < 20$, because we aimed to generate an intense selection situation that would show changes in genetic mean and variance. We split the gene-flow matrix $\boldsymbol{T}$ by specifying male and female paths ($\boldsymbol{P}_m + \boldsymbol{P}_f = \boldsymbol{I}$) and further splitting the male path into selected and non-selected path ($\boldsymbol{P}_m^s + \boldsymbol{P}_m^n = \boldsymbol{P}_m$), where $\boldsymbol{P}_m$ is a diagonal matrix with ones in rows for males and zeros otherwise; $\boldsymbol{P}_f = \boldsymbol{I} - \boldsymbol{P}_m$; $\boldsymbol{P}_m^s$ is a diagonal matrix with ones in rows for

selected males, and $\boldsymbol{P}_m^n = \boldsymbol{P}_m - \boldsymbol{P}_m^s$ is a diagonal matrix with ones in rows for non-selected males.

2.6 Software implementation

We have simulated the cattle breeding programme using `AlphaSimR` `R` package [20]. We have fitted the model (7) using `BLUPF90` family of programs [18] while all post-processing operations were done in `R` [15]. To compute and summarise the partitions we used the `AlphaPart` `R` package [10], to prepare data and present results we used the collection of `tidyverse` `R` packages [21] and `patchwork` `R` package [22]. The simulation and analysis code is fully available at the GitHub repository https://github.com/HighlanderLab/toliveira_alphapart_variance.

3 Results

We used a simulated cattle breeding programme to illustrate the method and software implementation and divided the results into two subsections. Subsection 3.1 describes the genetic mean and variance over generations for both medium- and high-accuracy scenarios and summarises the partitioning results using true breeding values. In the subsection 3.2 we show partitioning results when breeding values are estimated from phenotype data, comparing the estimated and true partitions.

3.1 Partitioning of true breeding values

Analysing true breeding values is a crucial way to demonstrate how the partitioning of breeding values and their means and variances works without the uncertainty due to estimating breeding values. Figure 4A shows partitions of genetic mean and variance by sex (males and females), while Figure 4B represents partitions by sex and selection status (selected males, non-selected males, and females) over generations. We scaled the genetic mean and variance of the base population respectively to zero and one to facilitate interpretation. We note that while we partitioned true breeding values, simulation was driven by selection with medium or high accuracy. This means that the accuracy impacted true trends in genetic mean and variance, and we analysed these simulation outputs.

Figure 4A shows that males contributed almost twice as much to the genetic mean as females. Interestingly, females were not selected at all (all females contributed progeny for two generations), so they contributed to genetic gain through the dissemination of the past male selections as we will show below. The higher accuracy expectedly drove larger changes in genetic mean and variance than the medium accuracy. If we would use only sex to analyze the drivers of change in genetic variance, we would have concluded that both male and female partitions contributed similarly to genetic variance for both selection scenarios. This observation raises a question: "How can males and females contribute a similar proportion to the genetic variance over time, if we are selecting among males, but not among females?". The answer to this question is shown in Figure 4B where we partitioned breeding values by sex and selection status. Although non-selected males did not contribute to the change in the genetic mean, their partition still contributed to the genetic variance in their generation. As expected 5 selected males contributed the least to genetic variance, followed by the 495 non-selected males and 500 females (born in a specific generation). Furthermore, since the male partition encompasses selected and non-selected males, this partition spans a wide distribution of partitioned breeding values and this wide distribution has a large variance. This is shown in the Supplementary Figures S1 and S2 with distributions of breeding value partitions for all 40 generations.

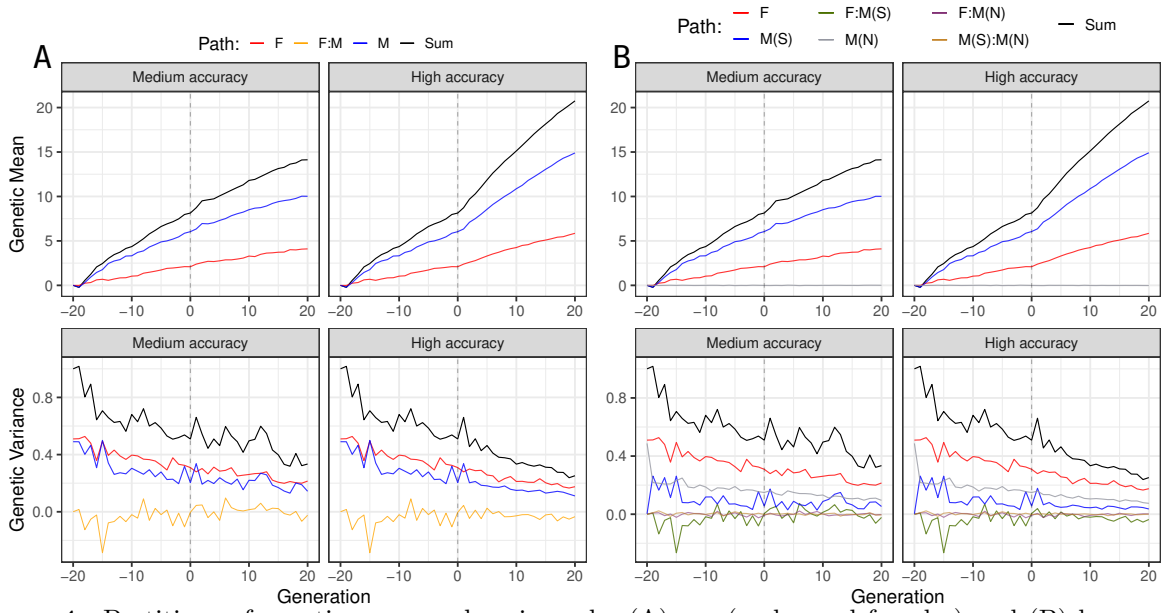

Figure 4. Partitions of genetic mean and variance by (A) sex (males and females) and (B) by sex and selection status (selected males (M(S)), non-selected males (M(N)), and females (F)) using true breeding values for one simulation replicate

Comparing the medium- and high-accuracy scenarios showed that the contribution of paths to genetic variance is a function of selection accuracy, with higher accuracy driving more changes in genetic variance. Notably, with medium accuracy, we saw a smaller difference between partitions of genetic variance for selected and non-selected males. The main reason for this is that the medium accuracy likely did not enable the selection of the top males from the tail of the distribution, which would have much lower variance.

Splitting the male path into selected and non-selected paths also showed that the negative covariance between male and female partitions in Figure 4A was driven by the covariance between female and selected male partitions (F:M(S), Figure 4B). This covariance was consistently negative from generation 8 to 20 in the high-accuracy scenario (Figure 4B), resulting in a mean correlation of -0.33 (± 0.15) for that generation interval. As a result, the total genetic variance in a generation t can be smaller than the sum of genetic variances for partitions. This non-independence of partitions of genetic variance is more evident in the high-accuracy scenario from generations 8 to 20, where the correlation decreased even more than in the medium-accuracy scenario (Supplementary Figure S3). The non-independence of partitions of genetic variance is yet another reason why individual partitions of genetic variance have to be interpreted with caution.

To further clarify why female partition had a non-zero contribution to the genetic gain, despite the absence of selection among females, the Figure 5 shows the distributions of breeding value partitions and Mendelian sampling terms by the group in generation 39 for medium-accuracy (Figure 5A) and high-accuracy (Figure 5B) scenarios. We can see that the female partition contributed significantly to genetic gain (red distribution), though clearly less than the selected males partition (blue distribution) and clearly more than the non-selected males partition (gray distribution). The inspection of Mendelian sampling terms for females, and for non-selected males, expectedly showed no selection - their Mendelian sampling terms were distributed around zero, while selected males had consistently positive Mendelian sampling terms. However, females were the progeny of previously selected males, and their sons were subject to selection, which creates a non-zero contribution to the female partition - through the dissemination of genes selected in their sires and through

their (dam's) sons.

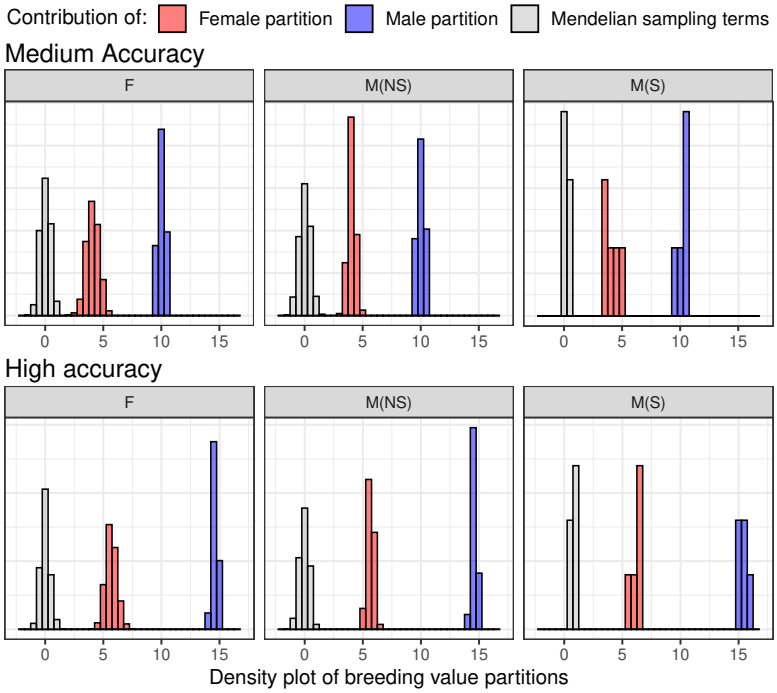

Figure 5. Distribution of breeding value partitions and Mendelian sampling terms by group in generation 39 for medium-accuracy (A) and high accuracy (B) selection scenarios

The presented results showed one replicate of the simulation. In the supplementary Figure S4 we show the partitioning analysis for each of the 30 replicates that all used the same founding genomes. Our aim was to show that the above results are consistently observed across many replicates, but to also show the magnitude of variation between replicates. The Supplementary Figure S4 shows the 30 replicates with the solid line representing the median and the ribbon represents the distribution of true partitions of genetic mean and variance as well as correlation between selected male and female partitions.

3.2 Estimating the partitions of genetic mean and variance

3.2.1 Model fit

The data were analysed with the model (7) using the complete pedigree that enabled accurate estimation of residual and base population genetic variance, though we slightly overestimated base population genetic variance in the high-accuracy scenario (Table 1). Evaluating the model further in the terms of estimating the

Table 1. Variance components (VC) true values, point estimates, and their 95% highest posterior density (HPD) interval

				95% HPD	
Scenario	VC	True	Estimate	Lower	Upper
Medium accuracy	σ^2_a	0.3	0.27	0.25	0.30
	σ^2_e	0.7	0.69	0.68	0.71
High accuracy	σ^2_a	0.3	0.35	0.33	0.38
	σ^2_e	0.7	0.66	0.64	0.67

quantities of interest well, we observed good agreement and unbiased estimates under the medium accuracy

for genetic means over generations and slightly lower agreement and some bias for the genetic variance over generations (Table 2). Under the high accuracy, the genetic mean over generations was also well estimated, but there was considerable disagreement and bias for the genetic variance over generations (Table 2). The estimated and true genetic means and variances over 40 generations are shown in Supplementary Figures S5 and S6. One reason for a worse performance of the model (7) under the high accuracy, was that this scenario generated significant genetic change both in mean and variance (Figure 4, which was also manifested in a higher level of inbreeding than the medium-accuracy scenario (Supplementary Figure S7). As inbreeding increases over generations, it generates variation between individuals that is challenging to represent using only pedigree-based relationships and better approaches are needed, such as, genomic relationships.

Table 2. Estimate and 95% confidence interval for the concordance correlation coefficient ($\widehat{\rho}_c$) between the estimated and true statistic, and point estimates for the Pearson correlation coefficient ($\widehat{\rho}$) and bias correction factor ($\widehat{C}$) between the variables in each case within scenario

		Est. and 95% CI				
Scenario	Statistic	$\widehat{\rho}_c$	Lower	Upper	$\widehat{\rho}$	$\widehat{C}$
Medium accuracy	$E(\boldsymbol{a}\|\boldsymbol{y})$ vs. μ_{a_t}	1.00	1.00	1.00	1.00	1.00
	$Var(\boldsymbol{a}\|\boldsymbol{y})$ vs. $\sigma^2_{a_t}$	0.95	0.93	0.96	0.96	0.99
High accuracy	$E(\boldsymbol{a}\|\boldsymbol{y})$ vs. μ_{a_t}	1.00	1.00	1.00	1.00	1.00
	$Var(\boldsymbol{a}\|\boldsymbol{y})$ vs. $\sigma^2_{a_t}$	0.87	0.83	0.90	0.97	0.89

3.2.2 Genetic means and its partitions

Now that the adequacy of the model (7) has been assessed and its impact on the estimates of genetic means and variances over generations has been evaluated, we show the partitioning results when breeding values are estimated from phenotypes. First, we illustrate partitioning results from a single replicate, then we extend it by showing results from 30 replicates. Figure 6 shows the true and estimated genetic mean over 40 generations for the medium- and high-accuracy scenarios considering the total genetic mean (Sum), the path for selected males (M(S)), non-selected males (M(NS)), and females (F). For the medium-accuracy scenario, although the point estimate for the mean of selected males partition showed underestimation), the true means of partition of each path was inside the 95% credible interval. For the high-accuracy scenario, we observed underestimation for females and selected males partition. Consequently, underestimation for the total genetic mean was even higher because it is the sum of those two contributions, while non-selected males had a zero contribution. Figure 7 confirms this result by showing the difference between true and estimated means of partition over 30 replicates. The Supplementary Figure S8 shows that the observed deviations in both scenarios do not come from inadequately estimated Mendelian sampling terms, hence the source of error must be due to the inadequate estimation of the parent average terms.

3.2.3 Genetic variance and its partitions

The partitioning of genetic variance by paths in the medium- and high-accuracy scenarios in a single replicate are shown in Figure 8. While we correctly estimated the overall trends in the total genetic variance and its partitions, we observed a slight overestimation for the females and non-selected males paths and the total in either the medium- or high-accuracy scenarios. However, from generation 1 to 20 in the high-accuracy scenario, the overestimation increased compared to the medium-accuracy scenario. These observations were confirmed also across 30 replicates for both scenarios (Figure 9). Importantly, distribution over 30 replicates did not include zero in later generations indicating significant departure of the estimates from the true values.

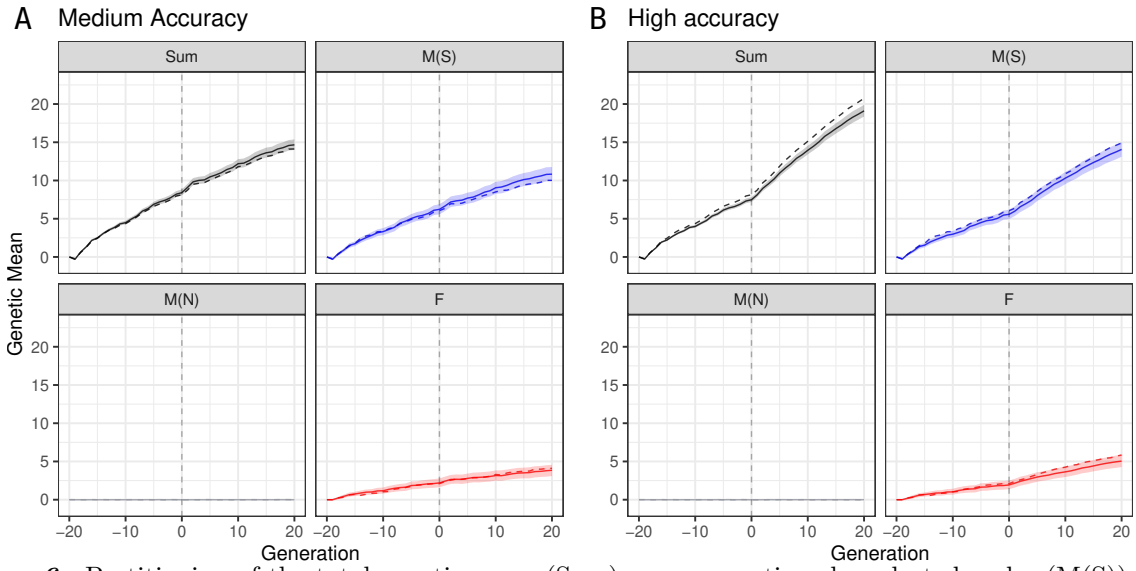

Figure 6. Partitioning of the total genetic mean (Sum) over generations by selected males (M(S)), non-selected males (M(N)), and females (F) paths in the medium-accuracy (A) and high-accuracy (B) selection scenario, considering one replicate (true value is denoted with a dashed line and posterior mean is denoted with a solid line and 95% credible interval is denoted with a ribbon)

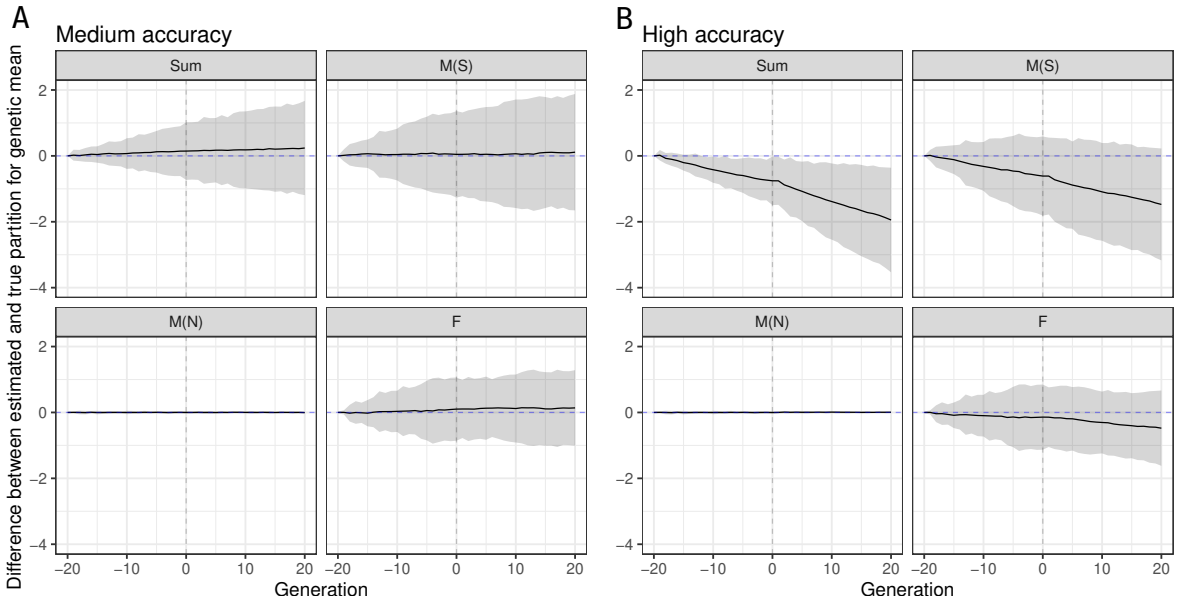

Figure 7. Distribution of the difference between true and estimated genetic means over generations for the total (Sum) partitioned by selected males (M(S)), non-selected males (M(N)), and females (F) paths in the medium-accuracy (A) and high-accuracy (B) selection scenario, considering 30 replicates (zero value is denoted with a dashed line and mean difference over replicates is denoted with a solid line and 95% quantile of differences over replicates is denoted with a ribbon)

Figure 9 also shows an underestimation of genetic variance for the selected males path in early generations (-19 to 2), which lead to the underestimation of the total genetic variance in the high-accuracy scenario.

We have initially observed even larger biases but have addressed these by adequately accounting for inbreeding in setting up the $\boldsymbol{A}^{-1}$. Ignoring inbreeding significantly impacted the estimates of genetic means and variances and their partitions (Supplementary Figures S9, S10, and S11).

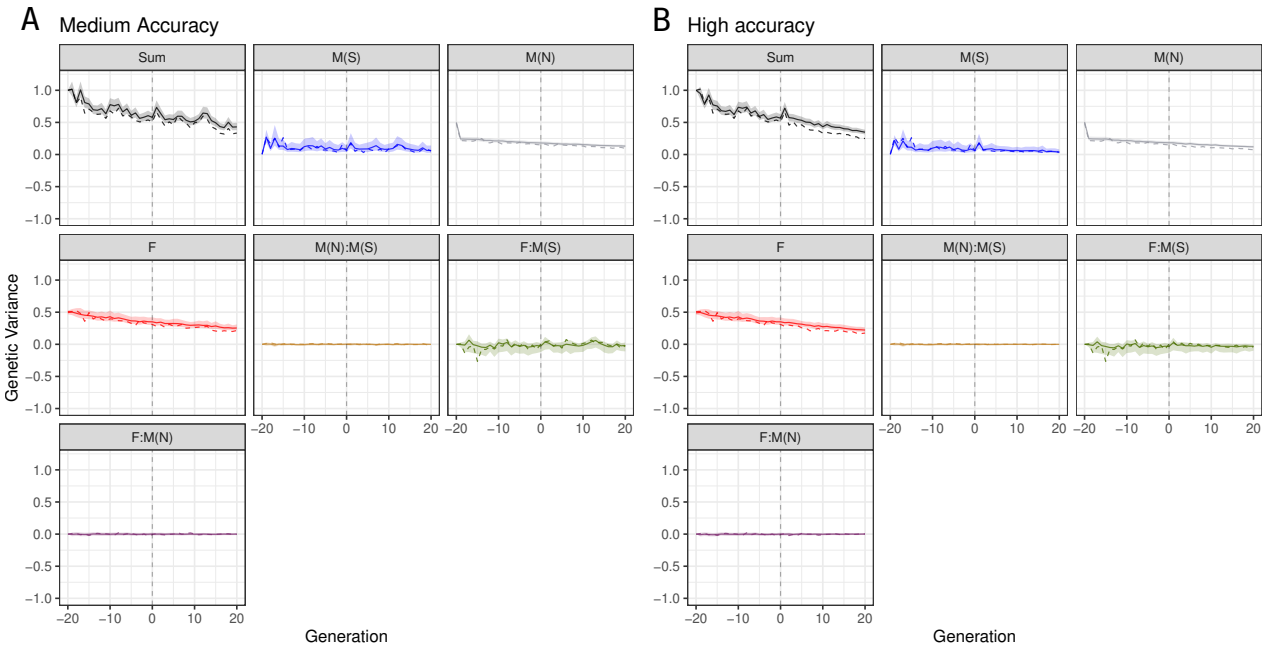

Figure 8. Partitioning of the total genetic variance (Sum) over a generation by selected males (M(S)), non-selected males (M(N)), and females (F) path in the medium-accuracy (A) and high-accuracy (B) selection scenario, considering one replicate (true value is denoted with a dashed line and posterior mean denoted with a solid line and 95% credible interval is denoted with a ribbon)

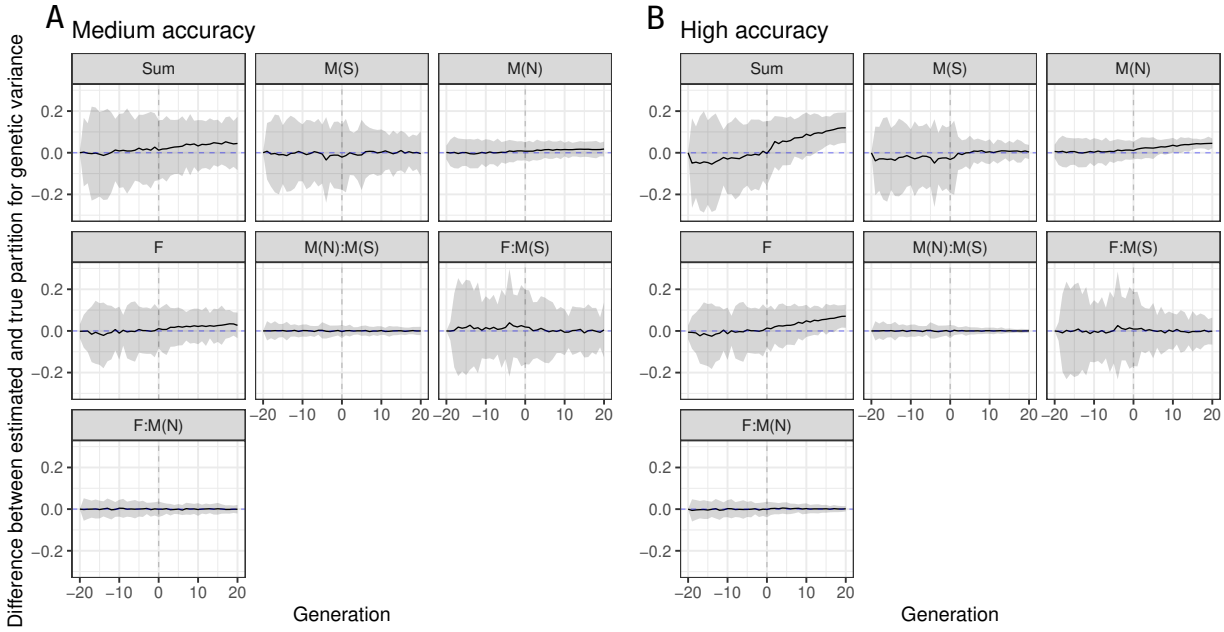

Figure 9. Distribution of the difference between true and estimated genetic variances over generations for the total (Sum) partitioned by selected males (M(S)), non-selected males (M(N)), and females (F) paths in the medium-accuracy (A) and high-accuracy (B) selection scenario, considering 30 replicates (zero value is denoted with a dashed line and mean difference over replicates is denoted with a solid line and 95% quantile of differences over replicates is denoted with a ribbon)

4 Discussion

We developed a method for partitioning the trends in genetic mean and variance into contributions of different paths as an extension of the work of Garcia-Cortes et al. [1] and Obsteter et al. [17]. The method used to infer the path contributions over generations is illustrated using a single-trait model; however, extension to multiple traits is straightforward and already implemented in `AlphaPart`. The extension presented here allows researchers to quantify the drivers of genetic variance in their breeding programmes in addition to the drivers of the genetic mean. Consequently, it could help to understand the dynamics between genetic mean and variance in global animal breeding [2, 4], test how different breeding schemes may impact their long-term sustainability [23, 24], and optimise investment to bring more variability to early stages of the breeding programme [25, 26]. Therefore, it is a powerful and valuable method to understand how different groups of breeding individuals contribute to change in genetic mean and variance, a hot topic that has been discussed in the last few years [5, 7, 27]. For this reason, we implemented this method in the `AlphaPart R` package. The extension has been available since version 0.9.3, freely available from CRAN.

The simulated cattle breeding programme with the medium- and high-accuracy scenarios illustrated the power of the partitioning method to summarise genetic trends in mean and variance. However, some care is needed when using the proposed method. We have shown that the path variable must be considered carefully because a specific choice can lead to a misinterpretation of the contributions, especially regarding the partition of genetic variance. To this end, we recommend plotting the distribution of partitioned breeding values, where partitions can be done with different variables of interest, like sex and selection status in our study.

By partitioning the genetic mean and variance, we showed that in the high-accuracy scenario, the covariance between females and selected males plays an important role when partitioning the genetic variance. Consequently, in this case $Var\left(\boldsymbol{a}\right) < Var\left(\boldsymbol{a}_F\right) + Var\left(\boldsymbol{a}_M\right)$, where F and M represent the female and male paths. Furthermore, most of the (additive) genetic variance in the breeding programme pertained to females and non-selected males, which were not the most relevant individuals for disseminating genetic gain. A negative correlation between female and selected males partitions in Figure 4 and Supplementary Figure S3 means that the partitions of genetic variance may not be independent. Since the male partition contributes more and more over generations, the female partition has to contribute less, hence the negative correlation between them. In this sense, we demonstrated that the choice of paths is essential and that partitions are not necessarily independent; therefore, they should not be analysed in isolation from each other.

We observed some bias in our estimates of partitions, but we showed that this bias came from the model specification. Hence, the bias is closely related to estimating variance components in populations under selection and low effective population size [28, 29]. Furthermore, the impact of variance component estimation on estimating temporal genetic variances from the model for the high-accuracy scenario was demonstrated by showing a lower concordance correlation between true ($Var\left(\boldsymbol{a}\right)$) and estimated ($Var\left(\boldsymbol{a}|\boldsymbol{y}\right)$) compared to the medium-accuracy scenario. However, this extreme example is essential to show how the bias in variance components affects the proposed partitioning methodology.

This overestimation of the genetic variance and its partitions in the high-accuracy scenario is likely impacted by the "miss-specified" pedigree-based model for such an intense selection and low effective population size (Ne < 20) simulated in our study [12]. Namely, we have observed significant changes in the genetic variance of up to 75% over 40 generations. While the pedigree-based model can account for selection [28, 29], it can not account appropriately for such a significant change in genetic variance [6, 7, 12]. Therefore, our next step is to develop an extension of the partitioning method considering genomic data to overcome the issue of

working with the expected probability of identity by descent from pedigrees by using the realised identity by descent or state from genomic data [30, 31]. We have recently already extended the Sorensen et al. [6] method for temporal estimation of genetic variance with a pedigree-based model to work with genomic data. Extending the partitioning method of Garcia-Cortes et al. [1] and current work is a natural next step.

5 Conclusions

We developed a method to quantify the sources of genetic variance in breeding programmes by partitioning the genetic variance by analyst-defined paths. The developed method can help breeders and researchers better understand changes in genetic gain and variance in specific breeding programmes. Furthermore, the method can be easily applied to real data by leveraging established software to draw posterior breeding values samples given the observed phenotype data. Working with the posterior sample of breeding values also enables straightforward uncertainty quantification in evaluated genetic variance and its partitions.

By partitioning the genetic variance in a simulated cattle breeding programme, we showed the contribution of different groups of individuals (paths) to the variance and that the covariance between paths can also have a substantial contribution. Hence, to comprehend and manage changes in the genetic variance in a breeding programme, we should not consider the contribution of different paths in isolation but should perform a holistic analysis of partitions of the genetic variance.

We observed some overestimating genetic variance and its partitions, but this was caused by the extreme selection in our simulation study and the pedigree-based model, which could not cope with the genetic change both in mean and genetic variance. Our future research will extend the proposed method using genomic data to overcome the model miss-specification under such extreme selection settings.

The developed method for partitioning genetic variance is a powerful and valuable method for understanding how different paths contribute to changes in genetic variance over time, how they interact within a breeding programme, and how they can be optimised.

Funding

The authors acknowledge support from the BBSRC Institute Strategic Programme Grant to The Roslin Institute (BBS/E/D/30002275). TPO acknowledges funding from Limagrain and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 801215 and the University of Edinburgh Data-Driven Innovation programme part of the Edinburgh and South East Scotland City Region Deal. JO acknowledges funding from the Slovenian Research Agency (grant P4-0133). For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising from this submission.

Abbreviations

CCC - concordance correlation coefficient

CI - confidence interval

HPD - highest posterior density MCMC - Markov chain Monte Carlo

REML - residual maximum likelihood

TBV - true breeding values

Availability of data and materials

Project name: `AlphaPart`.

Project home page: https://cran.r-project.org/package=AlphaPart.

Operating system(s): Windows, MacOS, Linux.

Programming language: R & C++.

Licence: GPL-3.

Data and Code: https://github.com/HighlanderLab/toliveira_alphapart_variance

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

All authors read and approved the publication of the final manuscript.

Authors' contributions

GG initiated and supervised the project. TO extended AlphaPart, simulated and analysed the data, and drafted the manuscript. JO, IP, NH, and GG contributed to the discussion of results and revised the manuscript.

References

1. L.A. García-Cortés, J.C. Martínez-Ávila, and M.A. Toro. Partition of the genetic trend to validate multiple selection decisions. *Animal*, 2(6):821–824, 2008. ISSN 17517311. doi: 10.1017/S175173110800205X. URL https://linkinghub.elsevier.com/retrieve/pii/S175173110800205X.

2. G Gorjanc, L A García-Cortés, J Jakobsen, and J Dürr. Partitioning of International Genetic Trends by Origin in Brown Swiss Bulls. (44):6, 2011.

3. M. Špehar, Z. Ivkic, V. Bulic, Z Barac, and G. Gorjanc. Partitioning of Genetic Trends by Origin in Croatian Simmental Cattle. *Agriculturae Conspectus Scientificus*, 76(4):4, 2011.

4. D. Škorput, G. Gorjanc, A. Kasap, and Z. Luković. Partition of genetic trends by origin in Landrace and Large-White pigs. *Animal*, 9(10):1605–1609, 2015. ISSN 17517311. doi: 10.1017/S1751731115001056. URL https://linkinghub.elsevier.com/retrieve/pii/S1751731115001056.

5. Rostam Abdollahi-Arpanahi, Daniela Lourenco, Andres Legarra, and Ignacy Misztal. Dissecting genetic trends to understand breeding practices in livestock: a maternal pig line example. *Genetics*

Selection Evolution, 53(1):89, December 2021. ISSN 1297-9686. doi: 10.1186/s12711-021-00683-6. URL https://gsejournal.biomedcentral.com/articles/10.1186/s12711-021-00683-6.

6. D. Sorensen, R. Fernando, and D. Gianola. Inferring the trajectory of genetic variance in the course of artificial selection. *Genetical Research*, 77(1):83–94, February 2001. ISSN 0016-6723, 1469-5073. doi: 10.1017/S0016672300004845. URL https://www.cambridge.org/core/product/identifier/S0016672300004845/type/journal_article.

7. Letícia A. de C. Lara, Ivan Pocrnic, Thiago de P. Oliveira, R. Chris Gaynor, and Gregor Gorjanc. Temporal and genomic analysis of additive genetic variance in breeding programmes. *Heredity*, December 2021. ISSN 0018-067X, 1365-2540. doi: 10.1038/s41437-021-00485-y. URL https://www.nature.com/articles/s41437-021-00485-y.

8. J.A. Woolliams, P. Berg, B.S. Dagnachew, and T.H.E. Meuwissen. Genetic contributions and their optimization. *Journal of Animal Breeding and Genetics*, 132(2):89–99, April 2015. ISSN 09312668. doi: 10.1111/jbg.12148. URL https://onlinelibrary.wiley.com/doi/10.1111/jbg.12148.

9. Gregor Gorjanc and John M Hickey. AlphaMate: a program for optimizing selection, maintenance of diversity and mate allocation in breeding programs. *Bioinformatics*, 34(19):3408–3411, October 2018. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bty375. URL https://academic.oup.com/bioinformatics/article/34/19/3408/4990830.

10. G. Gorjanc, Jana Obšteter, and T. P. Oliveira. AlphaPart: Partition/Decomposition of Breeding Values by Paths of Information, 2022. URL https://CRAN.R-project.org/package=AlphaPart.

11. C. R. Henderson. A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. *Biometrics*, 32(1):69, March 1976. ISSN 0006341X. doi: 10.2307/2529339. URL https://www.jstor.org/stable/2529339?origin=crossref.

12. B W Kennedy, L R Schaeffer, and D A Sorensen. Genetic Properties of Animal Models. *Journal of Dairy Science*, 71:17–26, 1988. ISSN 0022-0302. doi: 10.1016/S0022-0302(88)79975-0. Publisher: Elsevier.

13. R. L. Quaas. Additive genetic model with groups and relationships. *Journal of Dairy Science*, 71 (5):1338–1345, 1988. URL http://jds.fass.org/cgi/content/abstract/71/5/1338. tex.checked: [2009-06-14] tex.vnos: SM.

14. R. A. Mrode. *Linear models for the prediction of animal breeding values*. CAB International, Wallingford, 2 edition, 2005.

15. R Core Team. R: A Language and Environment for Statistical Computing, 2021.

16. D. Sorensen and D. Gianola. *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*, volume 1. Springer, New York, 1 edition, 2007. ISBN 0-387-95440-6.

17. Jana Obšteter, Justin Holl, John M. Hickey, and Gregor Gorjanc. AlphaPart—R implementation of the method for partitioning genetic trends. *Genetics Selection Evolution*, 53(1):30, December 2021. ISSN 1297-9686. doi: 10.1186/s12711-021-00600-x. URL https://gsejournal.biomedcentral.com/articles/10.1186/s12711-021-00600-x.

18. I. Misztal, S. Tsuruta, D. A. L. Lourenco, Y. Masuda, I. Aguilar, A. Legarra, and Z. Vitezica. Manual for BLUPF90 family programs, 2018. URL http://nce.ads.uga.edu/wiki/doku.php?id=documentation.

19. Lawrence I-Kuei Lin. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, 45(1):255, March 1989. ISSN 0006341X. doi: 10.2307/2532051. URL https://www.jstor.org/stable/2532051?origin=crossref.

20. R Chris Gaynor, Gregor Gorjanc, and John M Hickey. AlphaSimR: An R-package for Breeding Program Simulations. *bioRxiv*, page 21, 2020. doi: https://doi.org/10.1101/2020.08.10.245167.

21. Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Pedersen, Evan Miller, Stephan Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43):1686, November 2019. ISSN 2475-9066. doi: 10.21105/joss.01686. URL https://joss.theoj.org/papers/10.21105/joss.01686.

22. Thomas L. Pedersen. patchwork: The Composer of Plots, 2020. URL https://CRAN.R-project.org/package=patchwork.

23. Gregor Gorjanc, R. Chris Gaynor, and John M. Hickey. Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. *Theoretical and Applied Genetics*, 131(9):1953–1966, September 2018. ISSN 0040-5752, 1432-2242. doi: 10.1007/s00122-018-3125-3. URL http://link.springer.com/10.1007/s00122-018-3125-3.

24. Giovanny Covarrubias-Pazaran, Zelalem Gebeyehu, Dorcus Gemenet, Christian Werner, Marlee Labroo, Solomon Sirak, Peter Coaldrake, Ismail Rabbi, Siraj Ismail Kayondo, Elizabeth Parkes, Edward Kanju, Edwige Gaby Nkouaya Mbanjo, Afolabi Agbona, Peter Kulakow, Michael Quinn, and Jan Debaene. Breeding Schemes: What Are They, How to Formalize Them, and How to Improve Them? *Frontiers in Plant Science*, 12:791859, January 2022. ISSN 1664-462X. doi: 10.3389/fpls.2021.791859. URL https://www.frontiersin.org/articles/10.3389/fpls.2021.791859/full.

25. Irwin L. Goldman. Biodiversity in Plant Breeding. In *Encyclopedia of Biodiversity*, pages 459–469. Elsevier, Madison, 2013. ISBN 978-0-12-384720-1. doi: 10.1016/B978-0-12-384719-5.00017-4. URL https://linkinghub.elsevier.com/retrieve/pii/B9780123847195000174.

26. Gregor Gorjanc, Janez Jenko, Sarah J. Hearne, and John M. Hickey. Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. *BMC Genomics*, 17(1):30, December 2016. ISSN 1471-2164. doi: 10.1186/s12864-015-2345-z. URL http://www.biomedcentral.com/1471-2164/17/30.

27. Jorge Hidalgo, Shogo Tsuruta, Daniela Lourenco, Yutaka Masuda, Yijian Huang, Kent A Gray, and Ignacy Misztal. Changes in genetic parameters for fitness and growth traits in pigs under genomic selection. *Journal of Animal Science*, 98(2):skaa032, February 2020. ISSN 0021-8812, 1525-3163. doi: 10.1093/jas/skaa032. URL https://academic.oup.com/jas/article/doi/10.1093/jas/skaa032/5717959.

28. D. A. Sorensen and B. W. Kennedy. Estimation of Genetic Variances from Unselected and Selected Populations. *Journal of Animal Science*, 59(5):1213–1223, November 1984. ISSN 0021-8812. doi: 10.2527/jas1984.5951213x. URL https://doi.org/10.2527/jas1984.5951213x.

29. J. H. J. van der Werf and I. J. M. de Boer. Estimation of additive genetic variance when base populations are selected1. *Journal of Animal Science*, 68(10):3124–3132, October 1990. ISSN 0021-8812. doi: 10.2527/1990.68103124x. URL https://doi.org/10.2527/1990.68103124x.

30. P.M. VanRaden. Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science*, 91(11): 4414–4423, November 2008. ISSN 00220302. doi: 10.3168/jds.2007-0980. URL https://linkinghub.elsevier.com/retrieve/pii/S0022030208709901.

31. Joseph E. Powell, Peter M. Visscher, and Michael E. Goddard. Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Reviews Genetics*, 11(11):800–805, November 2010. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg2865. URL http://www.nature.com/articles/nrg2865.

Supporting Information

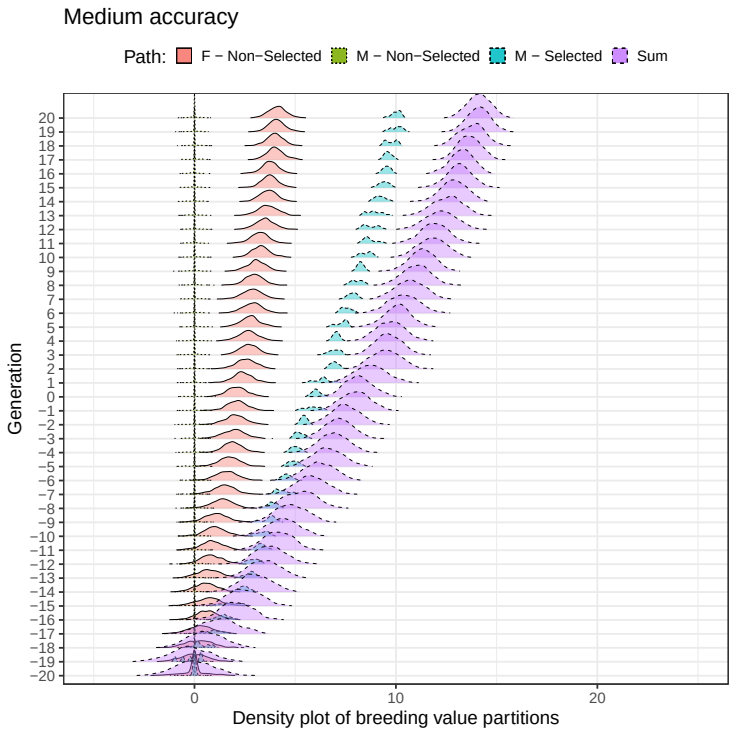

Figure S1. Distribution of breeding value partitions by sex and selection status (selected males (M(S)), non-selected males (M(N)), and females (F)) over generations for medium-accuracy scenario

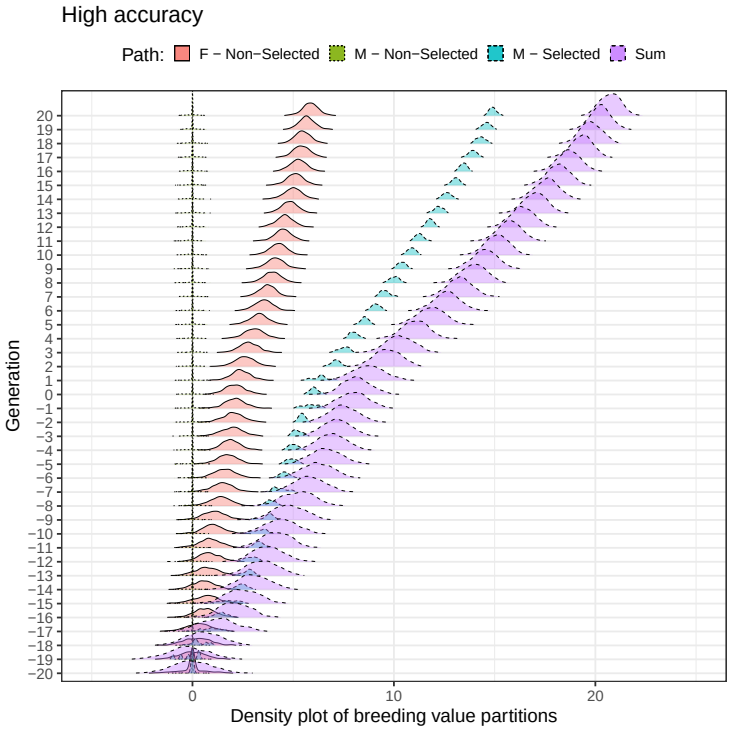

Figure S2. Distribution of breeding value partitions by sex and selection status (selected males (M(S)), non-selected males (M(N)), and females (F)) over generations for high-accuracy scenario

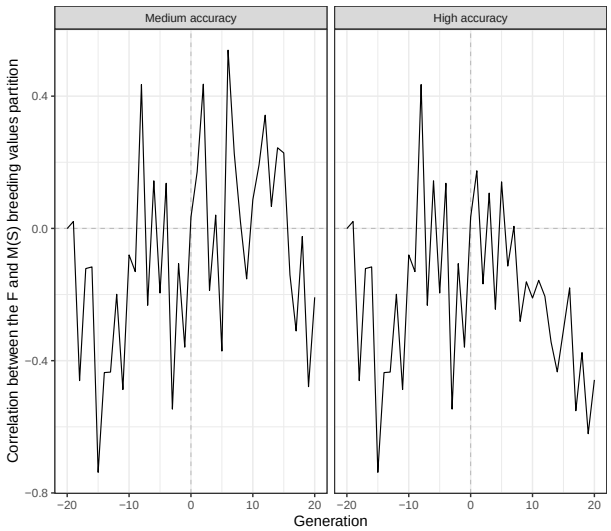

Figure S3. Correlation between Females (F) and selected males (M(S)) partitions using true breeding values for the medium- and high-accuracy scenarios and one simulation replicate

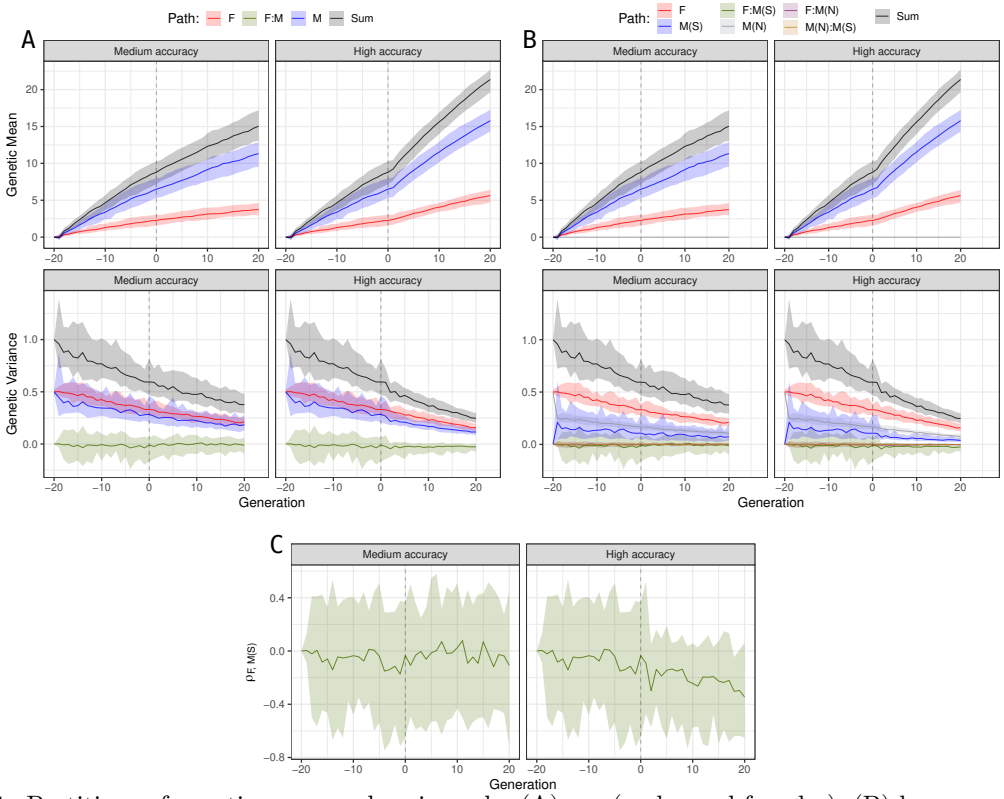

Figure S4. Partitions of genetic mean and variance by (A) sex (males and females), (B) by sex and selection status (selected males (M(S)), non-selected males (M(N)), and females (F)), and (C) the Pearson correlation between F and M(S) partitions for the medium- and high-accuracy scenarios by sex and selection status using true breeding values for 30 simulation replicate

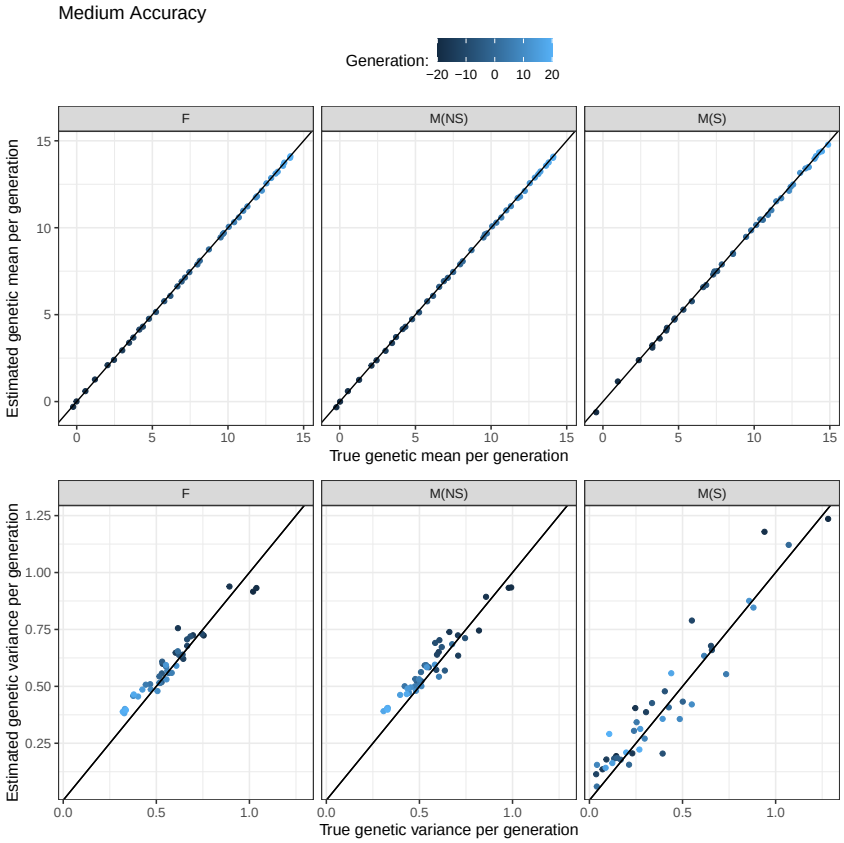

Figure S5. Estimated and true genetic means and variances over 40 generations by selected males (M(S)), non-selected males (M(N)), and females (F) in the medium-accuracy scenario. The solid line represents the equality line $y = x$ and the dots the Cartesian coordinates of estimated and true values.

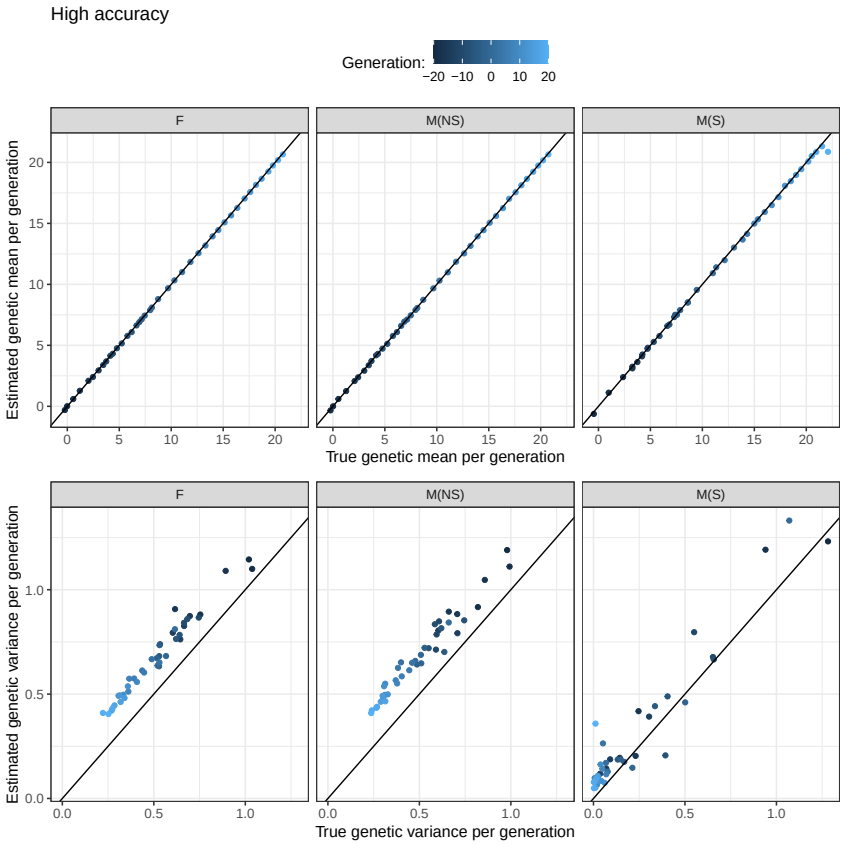

Figure S6. Estimated and true genetic means and variances over 40 generations by selected males (M(S)), non-selected males (M(N)), and females (F) in the high-accuracy scenario. The solid line represents the equality line $y = x$ and the dots the Cartesian coordinates of estimated and true values.

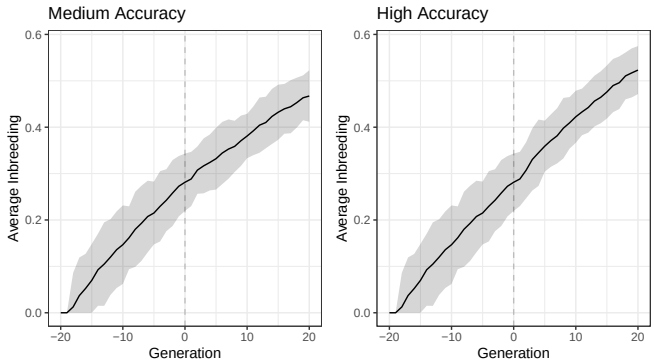

Figure S7. Point (solid line) and interval (ribbon) estimates for inbreeding over generation considering all animals in a specific generation

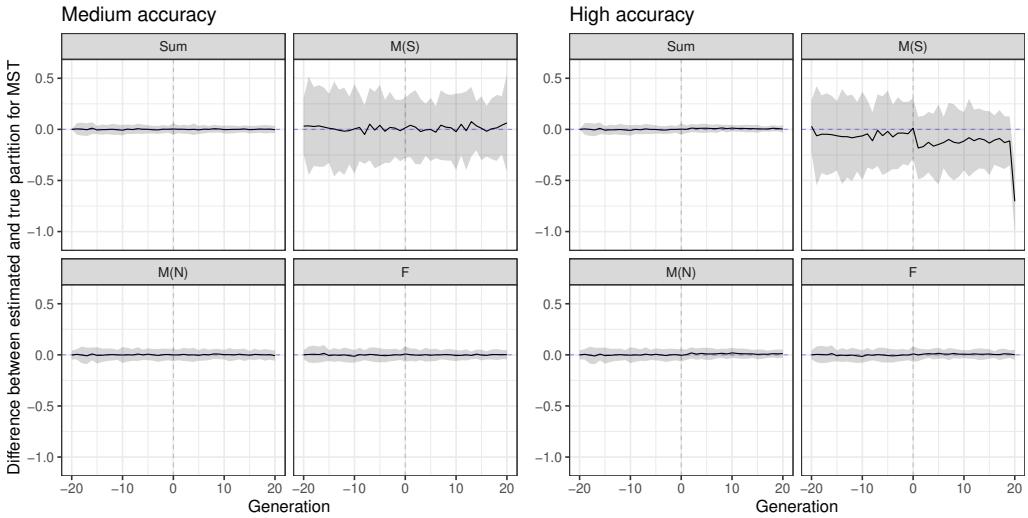

Figure S8. The difference between true and estimated Mendelian sampling terms (MST) is distributed over generations. The total (Sum) is partitioned by selected males (M(S)), non-selected males (M(N)), and females (F) paths in the medium- and high-accuracy selection scenario. We are considering 30 replicates (zero value is denoted with a dashed line and mean difference over replicates is denoted with a solid line, and 95% quantile of differences over replicates is denoted with a ribbon)

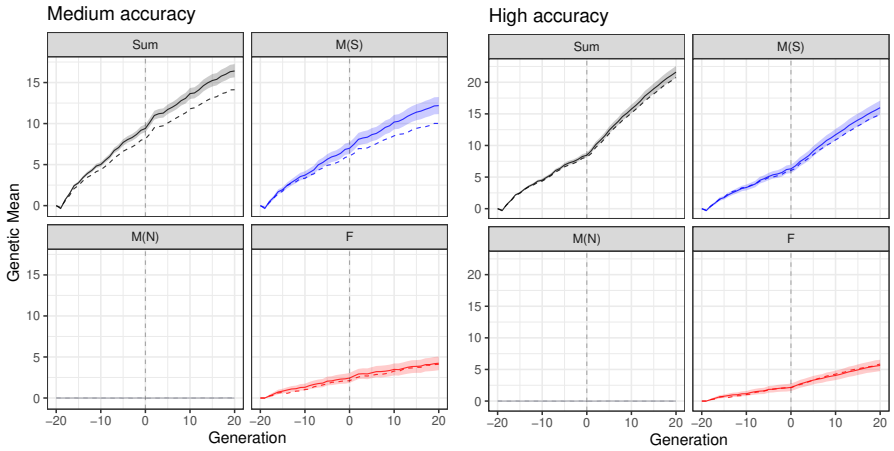

Figure S9. Partitioning of the total genetic mean (Sum) over generations by selected males (M(S)), non-selected males (M(N)), and females (F) paths in the medium-accuracy (A) and high-accuracy (B) scenario. We considered one replicate and without accounting for inbreeding in the model (true value is denoted with a dashed line and posterior mean is denoted with a solid line and 95% credible interval is denoted with a ribbon)

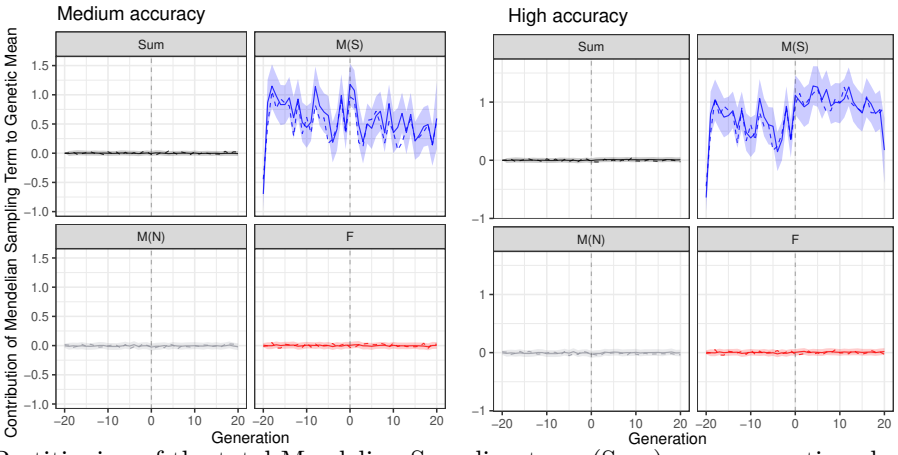

Figure S10. Partitioning of the total Mendelian Sampling term (Sum) over generations by selected males (M(S)), non-selected males (M(N)), and females (F) paths in the medium-accuracy (A) and high-accuracy (B) scenario. We considered one replicate and without accounting for inbreeding in the model (true value is denoted with a dashed line and posterior mean is denoted with a solid line and 95% credible interval is denoted with a ribbon)

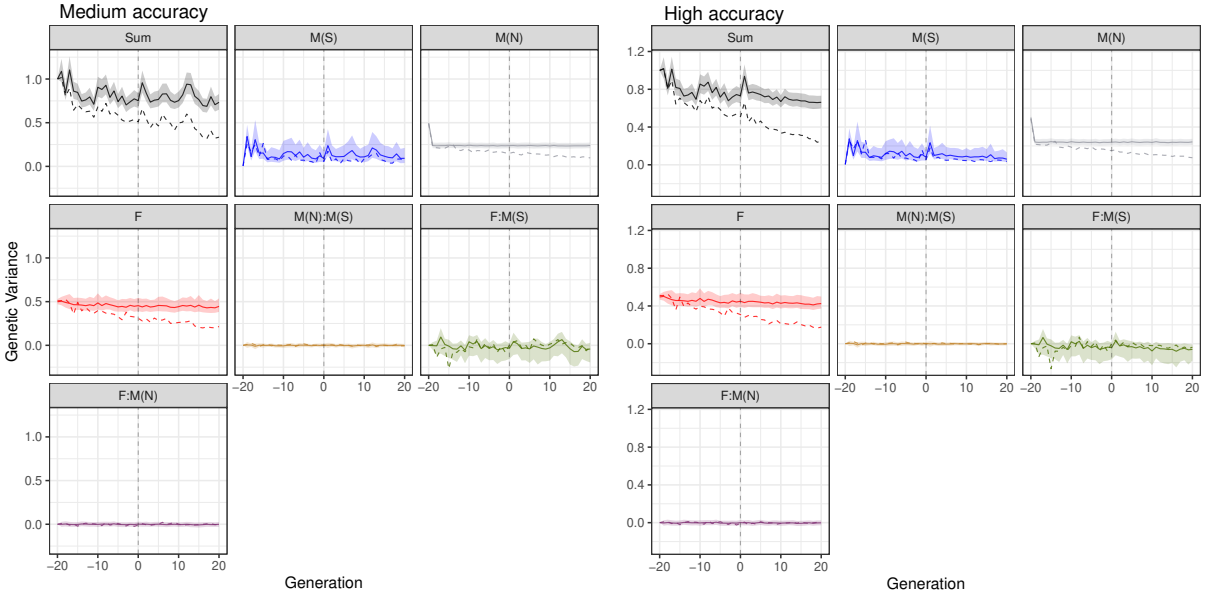

Figure S11. Partitioning of the total genetic variance (Sum) over a generation by selected males (M(S)), non-selected males (M(N)), and females (F) path in the medium-accuracy (A) and high-accuracy (B) scenario. We considered one replicate and without accounting for inbreeding in the model (true value is denoted with a dashed line and posterior mean is denoted with a solid line and 95% credible interval is denoted with a ribbon)