An analytical framework to derive the expected precision of genomic selection

Jean-Michel Elsen

Elsen *Genet Sel Evol (2017) 49:95*
https://doi.org/10.1186/s12711-017-0366-6

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

An analytical framework to derive the expected precision of genomic selection

Jean-Michel Elsen*

Abstract

Background: Formulae to predict the precision or accuracy of genomic estimated breeding values (GEBV) are important when modelling selection schemes. Simple versions of such formulae have been proposed in the past, based on a number of simplifying hypotheses, including absence of linkage disequilibrium and linkage between loci, a population made up of unrelated individuals, and that all genetic variability of the trait is explained by the genotyped loci. These formulae were based on approximations that were not always clear. The objective of this paper is to offer a unique framework to derive equations that predict the precision of GEBV from the size of the reference population and the heritability of and number of QTL controlling the quantitative trait.

Results: The exact formulation of the precision of GEBV involves the expectation of the inverse of a linear function of the genomic matrix, which cannot be calculated from simple algebra but can be approximated using a Taylor polynomial expansion. First order approximations performed better than the initial prediction equations published in the literature. Second order approximations produced almost perfect estimates of precision when compared to results obtained when simulating situations that agreed with the assumptions that were required to derive the precision equations. Using this proposed framework, we present several generalizations, including multi-trait genomic evaluation.

Conclusions: Although further improvements are needed to account for the complexity of practical situations, the equations proposed here can be used to derive the precision of GEBV when comparing breeding schemes a priori.

Background

After the seminal work of Meuwissen et al. [1], who provided statistical methods to exploit linkage disequilibrium (LD) between genotyped marker loci and quantitative trait loci (QTL) in animal and plant breeding, as previously proposed by Lande and Thompson [2], genomic selection was launched, which has since revolutionised both research in quantitative and applied genetics and practical breeding plans. The benefits of this technology are considerable in dairy cattle (e.g. [3–8]) and dairy cattle breeders very rapidly changed their schemes in order to adopt genomic selection methods. Thus, it became possible to improve the reliability of estimated breeding values (EBV) at a young age, avoid costly and lengthy progeny tests, and limit the detrimental evolution of inbreeding. However, the application of genomic selection was not so clear in other breeding sectors, for various reasons: the high relative costs of genotyping (compared to the value of reproducers), the limited size of the (reference) populations required to calibrate the effects of single nucleotide polymorphisms (SNPs), and the fact that basic schemes were already organised with short generation intervals (e.g. [9–13]).

Mathematical models to describe and evaluate breeding plans can be useful to decide whether a breeding scheme based on genomic evaluations should be implemented or not e.g. [14]. These models are often based on stochastic simulations, in which the characteristics of single individuals (their genotypes at a number of SNPs, including QTL located across the genome, and their phenotypes for traits influenced by QTL) are generated, in order to produce data files that can be used as in "real life" (e.g. [15, 16]). Alternatively, models that describe populations

*Correspondence: jean-michel.elsen@inra.fr
GenPhySE (Génétique Physiologie et Systèmes d'Elevage), Université de Toulouse, INRA, ENVT, 31326 Castanet-Tolosan, France

at a higher level (generations, cohorts, classes of reproducers defined by their role in the scheme) offer a more rapid and flexible alternative to evaluate alternate breeding programs. In such approaches, deterministic equations link population characteristics such as heritability, mean LD, replacement rates, and the number of genotyped individuals to expected genetic progress by unit of time. Some of the most important equations in these models are the formulae that predict the precision of genomic EBV (GEBV). Analyses of simulations and real data have clearly demonstrated that the precision of GEBV depends on the structure of the reference population and the characteristics of the marker set used. The size of this reference population, its diversity, the genetic distance between the reference population and the group of selection candidates genotyped, the number of markers, and the degree or strength of LD are the main factors that influence this precision [17–29].

A very simple formula to obtain the precision of GEBV was given by Daetwyler et al. [17], based on a number of simplifying hypotheses that included: absence of LD and linkage between loci, a population made up of unrelated individuals, and all genetic variation of the trait is explained by the genotyped loci. Under this approach, the regression of phenotypes on SNP genotypes was performed one locus at a time. This equation has been widely used and cited more than 100 times in the literature. Adjustments have since been proposed to deal with the distribution of marker allele frequencies [20], include dependence between marker loci through the definition of an effective number of independent loci [30], include the proportion of genetic variance explained by markers [22], and account for a smaller error variance when multiple marker loci are considered simultaneously [8]. Brard and Ricard [31] reviewed and challenged these formulae, using the results reported in 13 publications based on real data. They showed that the size of the reference population and the number of independent segments had a considerable impact on precision, and that the different formulae produced very different results. Other situations were explored by Hayes et al. [21] by considering dependence between the reference and candidate populations, by Wientjes et al. [32], who studied multi-population scenarios, and by Elsen [33], who suggested opportunities for the more systematic exploration of dependence between SNPs and between individuals.

In the present paper, using the simple situation that was initially studied by Daetwyler et al. [17], we propose a framework to derive equations that predict the precision of GEBV based on the size of the reference population, and the heritability of and number of QTL controlling the quantitative trait. We are interested in the expectation of the precision of GEBV, before implementing possible genotyping and selection schemes, as a tool for optimizing resources. With this prior approach, the variability summarized when computing the expectation of GEBV precision comes from marker locus polymorphisms as well as from QTL and environmental random effects. After demonstrating the performance of the solutions obtained, we explore extensions to more complex situations. Ten equations are successively proposed: (1) a general formulation of the expectation of the precision of GEBV; (2) the Daetwyler et al. [17] equation that assumes that the error variance is not modified after correction for SNP effects; (3) the Daetwyler et al. [17] equation that accounts for the corresponding reduction error variance; (4) an approximation of Eq. (1) based on a Taylor series expansion; (5) and (9) applications of Eq. (4) to the first order, assuming that all SNPs contribute equally to the genetic variance (Eq. 5) or that their effects share the same prior variance (Eq. 9); (6) an extension of Eq. (5) to the second order; (7) and (8) applications of Eq. (6) when assuming that distribution of allele frequencies is uniform (Eq. 7) or U-shaped (Eq. 8); (10) an extension to the multivariate situation.

Methods

Proposed framework

Notations and hypotheses

A list of abbreviations is in Table 1. Genomic predictions are based on a set of M biallelic SNPs, with alleles A_k and B_k at locus k, the frequency of allele B_k being f_k. All SNPs are assumed to be in linkage equilibrium. The reference population, which is considered to be a random subset of a larger population, is made up of N unrelated individuals, which are genotyped and phenotyped. We are interested in the precision of the GEBV of a selection candidate that is not related to individuals in the reference population, but belongs to a selection population that is another subset of the larger population. The GEBV is derived as a SNP best linear unbiased prediction (BLUP) based on SNP genotypes.

The random elements of the prediction model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ are as follows:

$\mathbf{y}$ is a vector of phenotypes recorded in the reference population, assumed to be centred at zero. $\boldsymbol{\beta}$ is a vector of SNP effects and is randomly distributed with a mean of 0 and covariance matrix $v(\boldsymbol{\beta}) = \begin{pmatrix} \sigma^2_{\beta_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2_{\beta_M} \end{pmatrix} = \mathbf{B}$.

Note that based on this matrix, the β_k effects are supposed to be uncorrelated.

$\mathbf{X}$ is the genotype matrix defined by $X_{ik} = n_{ik} - 2f_k$, where $n_{ik} \in \{0, 1, 2\}$ is the number of B_k alleles carried by individual i at locus k. We assume that allele frequencies f_k are known. The expectation of X_{ik} is null, and its variance is $\sigma^2_k = 2f_k(1 - f_k)$. Under linkage equilibrium

Table 1 List of abbreviations used in alphabetical order

Abbreviation	Full meaning
$\mathbf{a}$	Vector of economic weights in γ
α	Lower bound of the distribution of minor allele frequencies
A_k and B_k	Alleles at SNP k
β_k	Effect of SNP k
$\hat{\beta}_k$	Prediction of the effect of SNP k
$\boldsymbol{\beta}_j$	Vector of SNP effects for trait j
$\boldsymbol{\beta}$	Vector of SNP effects
$\mathbf{B}$	Covariance matrix $v(\boldsymbol{\beta})$
$\tilde{\boldsymbol{\beta}}$	Estimates of fixed effects
γ	Selection objective
$\hat{\gamma}$	BLUP of γ
$\mathbf{D}$	Expectation of $\mathbf{X}'\mathbf{X}+\boldsymbol{\Lambda}$
δ_k	k^{th} diagonal term of $\mathbf{D}$
$\mathbf{E}$	Deviation of $\mathbf{X}'\mathbf{X}+\boldsymbol{\Lambda}$ from $\mathbf{D}$
$\mathbf{e}$	Vector of residuals
f_{min}	Minimum minor allele frequency
f_k	Frequency of allele B_k
$\mathbf{F}$	Matrix of SNP genotypes
g	Genetic value of the candidate
$\hat{g}$	GEBV of the candidate
$\mathbf{g}$	Vector of genetic values
$\hat{\mathbf{g}}$	BLUP of $\mathbf{g}$
h_0^2	Heritability
$\boldsymbol{\Lambda}$	Diagonal matrix with elements λ_k
λ	Ratio $\sigma_e^2/\sigma_{\boldsymbol{g}_k}^2$
λ_k	Ratio $\sigma_e^2/\sigma_{\boldsymbol{\beta}_k}^2$
M_e	Effective number of loci
M	Number of SNPs
N_e	Effective population size
N	Size of the reference population
$\mathbf{P}$	Working matrix $(\mathbf{D}^{-1}\mathbf{E}\mathbf{D}^{-1}\mathbf{E})$
$\mathbf{R}$	Covariance matrix $v(\mathbf{e})$
$\widetilde{r^2}$	Estimate of the precision of GEBV based on [17]
$\hat{r}^2$	Approximation of r^2 proposed here
$r^2(\mathbf{X},\mathbf{w})$	Expected precision of GEBV, given $\boldsymbol{X}$ and $\boldsymbol{w}$
r^2	Marginal expected precision of GEBV
$\sigma_{\boldsymbol{\beta}_k}^2$	Variance of the effects β_k
σ_k^2	Variance of the number of B_k alleles
σ_g^2	Genetic variance
σ_e^2	Environmental variance
σ_γ^2	Phenotypic variance
$\mathbf{w}$	Vector of SNP genotypes
$\mathbf{X}$	Genotype matrix
$\mathbf{y}$	Vector of the phenotypes of the reference population

between SNPs, the expectation of matrix $\mathbf{X}'\mathbf{X}$ is $N\begin{pmatrix}\sigma_1^2 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sigma_M^2\end{pmatrix}=N\mathbf{F}$.

All genetic variability is assumed to be explained by the SNPs.

$\mathbf{e}$ is a vector of residuals with a mean of 0 and covariance matrix $v(\mathbf{e})=\begin{pmatrix}\sigma_e^2 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sigma_e^2\end{pmatrix}=\sigma_e^2\mathbf{I}_N=\mathbf{R}$.

$g=\mathbf{w}\boldsymbol{\beta}$ is the true genomic breeding value of the candidate to be predicted, with $\mathbf{w}$ the vector of SNP genotypes, defined as the rows in $\mathbf{X}$. The variance of $\mathbf{w}$ is $v(\mathbf{w})=\mathbf{F}$. Assuming all genetic variability is explained by the SNPs, we have $v(g)=E\left[\mathbf{w}\mathbf{B}\mathbf{w}'\right]=E\left[\mathbf{X}\mathbf{B}\mathbf{X}'\right]_{ii}\forall i$.

$\hat{g}=\mathbf{w}\hat{\boldsymbol{\beta}}$ is the GEBV of the candidate, where $\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}+\mathbf{B}^{-1})^{-1}\mathbf{X}'\mathbf{R}^{-1}\mathbf{y}$ is the BLUP of the SNP effects.

For a given set of genotypes $\mathbf{X}$, variance $v(\hat{\boldsymbol{\beta}}|\mathbf{X})=\mathbf{B}-\left(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}+\mathbf{B}^{-1}\right)^{-1}$. Defining matrix $\boldsymbol{\Lambda}=\begin{pmatrix}\lambda_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \lambda_M\end{pmatrix}$, with $\lambda_k=\sigma_e^2/\sigma_{\boldsymbol{\beta}_k}^2$, this variance is also $v(\hat{\boldsymbol{\beta}}|\mathbf{X})=\mathbf{B}-\sigma_e^2\left(\mathbf{X}'\mathbf{X}+\boldsymbol{\Lambda}\right)^{-1}$.

Expected precision of GEBV

Four sources of variation underlie the correlation between genomic breeding values (g) and their prediction ($\hat{g}$): the SNP genotypes ($\mathbf{X}$ and $\mathbf{w}$), their effects ($\boldsymbol{\beta}$), and the environmental effects ($\mathbf{e}$). Quite often, we are interested in the precision of GEBV, given the population genotypes ($\mathbf{X}$ and $\mathbf{w}$), and the randomness arising from the variability of $\boldsymbol{\beta}$ and $\mathbf{e}$, i.e. $r^2(\mathbf{X},\mathbf{w})=\frac{v(\hat{g}|\mathbf{X},\mathbf{w})}{v(g|\mathbf{X},\mathbf{w})}$, which is a function of matrices $\mathbf{X}$ and $\mathbf{w}$. A priori, before genotyping, for instance when different SNP chip densities or reference population sizes are compared, the criterion of interest is $r^2=\frac{v(\hat{g})}{v(g)}$. This is the situation explored in this paper.

The denominator in the previous equation for the precision of the GEBV is the genetic variance in the selection population: $v(g)=E_w\left[\mathbf{w}\mathbf{B}\mathbf{w}'\right]=tr[v[\mathbf{w}]\mathbf{B}]=tr[\mathbf{F}\mathbf{B}]=\sum_k\sigma_k^2\sigma_{\boldsymbol{\beta}_k}^2$. The variances of SNP effects are not known but must be estimated (e.g. [1]). In our a priori estimation of the precision of the GEBV, simplifying assumptions are needed. Following VanRaden [19], all variances of SNP effects are assumed to be equal to σ_β^2 and, thus, $\sigma_g^2=\sigma_\beta^2\sum_k\sigma_k^2$. Alternatively, following Wientjes et al. [32], all SNPs contribute equally to $\sigma_e^2=\sigma_g^2$, i.e. $\sigma_k^2\sigma_{\boldsymbol{\beta}_k}^2=\sigma_g^2/M$. This is the situation considered in the present paper. The ratio σ_e^2/σ_g^2 will be denoted by λ.

The numerator of the equation for precision is $v(\hat{g}) = E_{X,w}\left[\mathbf{w}v\left(\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{w}\right)\mathbf{w}'\right]$ since: (1) $v(\hat{g}) = E_{X,w}\left[v(\hat{g}|\mathbf{X},\mathbf{w})\right] + v_{X,w}\left[E(\hat{g}|\mathbf{X},\mathbf{w})\right]$ and (2) $E(\hat{g}|\mathbf{X},\mathbf{w}) = \mathbf{w}E\left(\hat{\boldsymbol{\beta}}|\mathbf{X}\right) = 0$.

Since $\hat{\boldsymbol{\beta}}$ and $\mathbf{X}$ are independent from $\mathbf{w}$ and $E[\mathbf{w}] = 0$ $:v(\hat{g}) = E_w\left[\mathbf{w}E_X\left[v\left(\hat{\boldsymbol{\beta}}|\mathbf{X}\right)\right]\mathbf{w}'\right] = tr\left[v[\mathbf{w}]E_X\left[v\left(\hat{\boldsymbol{\beta}}|\mathbf{X}\right)\right]\right]$.

Finally $r^2 = \frac{tr[\mathbf{FB}]-\sigma_e^2 tr\left[\mathbf{F}E_X\left[\left(\mathbf{X}'\mathbf{X}+\boldsymbol{\Lambda}\right)^{-1}\right]\right]}{\sigma_g^2}$,

i.e. $r^2 = 1 - \lambda\, tr\left[\mathbf{F}\, E_X\left[\left(\mathbf{X}'\mathbf{X}+\boldsymbol{\Lambda}\right)^{-1}\right]\right]$. (1)

Approximation of the precision of GEBV proposed by Daetwyler et al. [17]

In their derivation of the precision of GEBV, Daetwyler et al. [17] considered marker effects as both random and fixed effects. With our notations, they used $\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{R}^{-1}\mathbf{y} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$ as a fixed effect estimator of $\boldsymbol{\beta}$. In this context, $var\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = var\left(\tilde{\boldsymbol{\beta}}\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\sigma_e^2$. However, when considering $\boldsymbol{\beta}$ as a random effect, they used $cov\left(\boldsymbol{\beta}, \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = 0$, giving $cov(\tilde{g}, g) = v(g)$, and found $r^2 = \frac{v(g)}{v(\tilde{g})}$, i.e. the inverse of the classical $r^2 = \frac{v(\hat{g})}{v(g)}$. Assuming that SNPs are in linkage equilibrium, have uncorrelated effects, and independence between SNP effects and genotypes ($cov(w_j, \beta_j) = 0$), the variance $v(\tilde{g}) = v\left(\mathbf{w}\left(\boldsymbol{\beta} + \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right)$ is $\sum_j var(w_j)v(\beta_j) + \sum_j var(w_j)v\left(\tilde{\beta}_j - \beta_j\right)$. When the reference population size is sufficiently large, then $\left(\mathbf{X}'\mathbf{X}\right)^{-1} \approx E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]$, giving $v(\tilde{g}) = \sigma_g^2 + \sum_j var(w_j)\sigma_e^2 / Nvar(x_{ij})$.

Initially, Daetwyler et al. [17] assumed inconsistently that both $\sigma_p^2 = 1$ ("assuming the phenotypic variance is 1") giving $\sigma_g^2 = h_0^1$ and $\sigma_e^2 = 1$ ("for the present, we shall conservatively take $\sigma_e^2 = 1$"). Since the candidate and reference individuals belong to the same population, $var(w_j) = var(x_{ij})$ and $v(\tilde{g}) = h_0^2 + M/N$, which gives:

$$\widetilde{r^2}_{(1)} = \frac{Nh_0^2}{Nh_0^2 + M}. \quad (2)$$

A correction was proposed to relax the approximation $\sigma_e^2 = 1$, which resulted in an upward correction of $\widetilde{r^2}$. The idea was to replace $\sigma_e^2 = 1$ by $\sigma_e^2 = 1 - h_0^2 + h_0^2(1 - r^2)$, giving a quadratic equation in r^2 and

$$\widetilde{r^2}_{(2)} = \frac{M + Nh_0^2 \pm \sqrt{\left(M + Nh_0^2\right)^2 - 4NMh_0^4}}{2Mh_0^2}. \quad (3)$$

An alternative derivation of Eq. (2) was proposed by Wientjes et al. [32]. The main idea was that, assuming all SNPs are independent, their effects can be estimated in single random effect models, with $\mathbf{y} = \mathbf{X}_k\beta_k + \mathbf{e}_k$ for locus k, giving $\hat{\beta}_k = \left(\mathbf{X}'_k\mathbf{X}_k + \frac{\sigma_{e_k}^2}{\sigma_{\beta_k}^2}\right)^{-1}\mathbf{X}'_k\mathbf{y}$. They assumed that (1) the reference population was large ($\mathbf{X}'_k\mathbf{X}_k \approx N\sigma_k^2$), (2) the SNPs contributed equally to the genetic variance ($\sigma_k^2\sigma_{\beta_k}^2 = \sigma_g^2/M$), and (3) the individual contribution of each SNP was very small ($\sigma_{e_k}^2 = \sigma_Y^2 - \sigma_g^2/M \cong \sigma_Y^2$). Applying these assumptions, the BLUP of β_k is $\hat{\beta}_k = \frac{\mathbf{X}'_k\mathbf{y}}{\sigma_k^2\left(N + M\sigma_Y^2/\sigma_g^2\right)}$, with variance $v\left(\hat{\beta}_k\right) = \frac{N\sigma_{\beta_k}^2}{N + M/h_0^2}$, and the precision of GEBV is $r^2 = \frac{v(\hat{g})}{v(g)} = \frac{\mathbf{w}v\left(\hat{\boldsymbol{\beta}}\right)\mathbf{w}'}{\mathbf{w}v(\boldsymbol{\beta})\mathbf{w}'} = \frac{\frac{N}{N+M/h_0^2}\sum_k w_k^2\sigma_{\beta_k}^2}{\sum_k w_k^2\sigma_{\beta_k}^2} = \frac{Nh_0^2}{Nh_0^2+M}$.

Another approach to calculate the precision of GEBV

In the following, we do not assume that $\sigma_Y^2 = \sigma_e^2$ or $\sigma_Y^2 = \sigma_{e_k}^2$, and a unique multi QTL random model ($\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$) is used to describe relationships between phenotype and genotype. As in Wientjes et al. [32], we assume that $\sigma_k^2\sigma_{\beta_k}^2 = \sigma_g^2/M$. In Eq. (1), the expectation of the inverse of matrix $\mathbf{X}'\mathbf{X} + \boldsymbol{\Lambda}$ appears. This matrix can be broken down into diagonal ($\mathbf{D} = E\left[\mathbf{X}'\mathbf{X}\right] + \boldsymbol{\Lambda}$) and non-diagonal elements ($\mathbf{E} = \mathbf{X}'\mathbf{X} - E\left[\mathbf{X}'\mathbf{X}\right]$). As in Goddard et al. [22] and Elsen [33], a Taylor series expansion for matrix $\mathbf{E}$ is used to find approximations:

$$\begin{aligned}\left(\mathbf{X}'\mathbf{X} + \boldsymbol{\Lambda}\right)^{-1} &= (\mathbf{E} + \mathbf{D})^{-1} = \mathbf{D}^{-1}\left(\mathbf{I} + \mathbf{E}\mathbf{D}^{-1}\right)^{-1} \\ &= \mathbf{D}^{-1}\left(\mathbf{I} - \mathbf{E}\mathbf{D}^{-1} + \mathbf{E}\mathbf{D}^{-1}\mathbf{E}\mathbf{D}^{-1}\cdots\right).\end{aligned}$$

Because $\mathbf{D}^{-1}$ is not random and $E_X[\mathbf{E}] = 0$, the second order approximation is $E_X\left[\left(\mathbf{X}'\mathbf{X} + \boldsymbol{\Lambda}\right)^{-1}\right] = \mathbf{D}^{-1} + E_X\left[\mathbf{D}^{-1}\mathbf{E}\mathbf{D}^{-1}\mathbf{E}\right]\mathbf{D}^{-1}$ and the precision of GEBV can be approximated by:

$$\hat{r}^2_{(2)} \cong 1 - \lambda\left(tr\left[\boldsymbol{F}\boldsymbol{D}^{-1}\right] + tr\left[\boldsymbol{F}E_X\left[\boldsymbol{D}^{-1}\boldsymbol{E}\boldsymbol{D}^{-1}\boldsymbol{E}\right]\boldsymbol{D}^{-1}\right]\right). \quad (4)$$

First order approximation

Matrix $\mathbf{D}^{-1}$ is diagonal with terms $\frac{1}{\delta_k} = \frac{1}{E[\mathbf{X}'\mathbf{X}]_{kk}+\lambda_k}$. In the first order, $\hat{r}^2_{(1)} \cong 1 - \frac{\sigma^2_e}{\sigma^2_g}\left(\sum_k \frac{\sigma^2_k}{\delta_k}\right)$. Assuming as above that $\delta_k = N\sigma^2_k + \frac{\sigma^2_e}{\sigma^2_{\beta_k}} = \sigma^2_k(N + M\lambda)$, we find:

$$\begin{aligned} \hat{r}^2_{(1)} &\cong 1 - \lambda\frac{M}{N + M\lambda}, \\ \hat{r}^2_{(1)} &= \frac{N}{N + M\lambda} = \frac{Nh^2_0}{Nh^2_0 + M\left(1 - h^2_0\right)}. \end{aligned} \tag{5}$$

This equation differs from formula (2) of Daetwyler et al. [17] by a factor $(1 - h^2_0)$.

Second order approximation

When the size of the reference population is limited, elements of matrix $\mathbf{X}'\mathbf{X}$ differ from their expectations: non-zero non-diagonal terms are present even if the SNPs are in linkage equilibrium and diagonal elements diverge from the genetic variances. The second order approximation of Eq. (4) partly captures these deviations. In Eq. (4), matrices $\mathbf{F}$ and $\mathbf{D}^{-1}$ are both diagonal. Thus, we only need the diagonal of matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{E}\mathbf{D}^{-1}\mathbf{E}$ when computing the traces. Additional file 1 shows that terms $E_X[P_{kk}]$ simplify to $\frac{N\sigma^2_k}{\delta_k}\left\{\frac{1-2\sigma^2_k}{\delta_k} + \sum_t \frac{\sigma^2_t}{\delta_t}\right\}$. A second order approximation of the precision of GEBV is thus $\hat{r}^2_{(2)} \cong 1 - \lambda\left(\sum_k \frac{\sigma^2_k}{\delta_k}\left(1 + \frac{N\sigma^2_k}{\delta_k}\left\{\frac{1-2\sigma^2_k}{\delta_k} + \sum_t \frac{\sigma^2_t}{\delta_t}\right\}\right)\right)$, which, using $\delta_k = \sigma^2_k(N + M\lambda)$, simplifies to:

$$\hat{r}^2_{(2)} \cong \frac{N}{N + M\lambda} - \lambda\frac{NM}{(N + M\lambda)^3}\left(M - 2 + \frac{1}{M}\sum_k \frac{1}{\sigma^2_k}\right). \tag{6}$$

This expression includes genetic variances σ^2_k, which in practice can be estimated from the genomic data available. A priori, when these variances are not available, we can approximate the last term by using $E\left[\frac{1}{\sigma^2_k}\right] = E\left[\frac{1}{2f_k(1-f_k)}\right]$. A general situation is a uniform distribution of the frequencies between f_{min}, the minimum minor allele frequency of genotyped SNPs (MAF), and $1 - f_{min}$ (i.e. a probability density function $f(f_k) = \frac{1}{1-2f_{min}}$), which results in $E\frac{1}{\sigma^2_k} = \int_{f_{min}}^{1-f_{min}} \frac{1}{1-2f_{min}}\frac{1}{2}\left[\frac{1}{f_k} + \frac{1}{1-f_k}\right]df_k = \frac{1}{2(1-2f_{min})}\left[\log\left(\frac{1-f_{min}}{f_{min}}\right) - \log\left(\frac{f_{min}}{1-f_{min}}\right)\right] = \frac{\log\left(\frac{1-f_{min}}{f_{min}}\right)}{1-2f_{min}}$. A practical approximation of the expected precision of GEBV is thus:

$$\hat{r}^2_{(2)} \cong \frac{N}{N + M\lambda} - \lambda\frac{NM}{(N + M\lambda)^3}\left(M - 2 + \frac{\log\left(\frac{1-f_{min}}{f_{min}}\right)}{1 - 2f_{min}}\right). \tag{7}$$

Numerical comparison of estimates of the precision of GEBV

Equations (1), (2), (3), (5) and (7) were evaluated by simulating data corresponding to the hypotheses that underlie their development: unrelated reference and candidate individuals, SNPs in linkage equilibrium, GEBV from a SNP-based BLUP model, and all causal SNPs are included in the SNP panel. Heritability ranged from 0.1 to 0.7, genotypes were available for 500 or 1000 SNPs and the size of the reference population ranged from 1000 to 10,000. The minimum MAF (f_{min}) was 0.025, 0.05, 0.075 or 0.10. One hundred replicates were simulated for each scenario, with allele frequencies generated for each. Computations were made in FORTRAN with the help of the Nag library [34].

The results obtained when $M = 500$, $f_{min} = 0.05$ and $h^2_0 = 0.3$ are in Fig. 1. Table 2 shows the effects of heritability and f_{min} on the results. They showed consistently that (i) the prediction proposed by Daetwyler et al. [17] is closer to the "true" precision given by Eq. (1) when the size of the reference population is limited, (ii) the first order approximation proposed in the present paper (Eq. (5)) improves when the size of this population increases, i.e. when the assumptions that underlie these equations are more realistic, and (iii) the second order approximations of Eqs. (3) and (7) were nearly perfect in all cases, with a slight advantage for those of Daetwyler et al. [17] in most cases.

Extensions

The framework that we propose is flexible to accommodate alternative situations without major problems, as illustrated in the following.

Distribution of allelic frequencies

In our a priori approach leading to Eq. (7), a uniform distribution of frequencies $f(f_k) = \frac{1}{1-2f_{min}}$ was assumed, corresponding to $E\left[\frac{1}{\sigma^2_k}\right] = \frac{\log\left(\frac{1-f_{min}}{f_{min}}\right)}{1-2f_{min}}$. Following Hayes et al. [21], a U-shaped distribution could be assumed, with $f(f_k) = C/2f_k(1-f_k)$. The C constant must be estimated from the constraint $\int_\alpha^{1-\alpha} f(f_k)df_k = 1$, where α and $1 - \alpha$ are the bounds of the f_k domain. Hayes et al. [21] argued that $\alpha = 1/2N_e$, with N_e being the effective size of the reference population, leading to

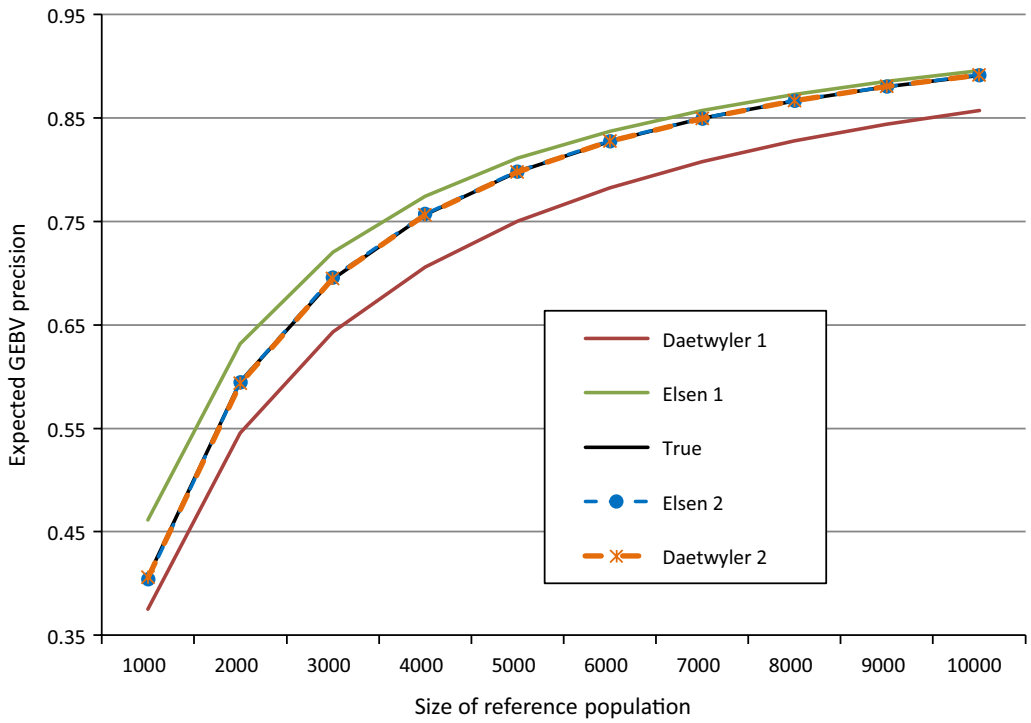

Fig. 1 Expected precision of GEBV as predicted by Eqs. (1), (2), (3), (5) and (7), when 500 SNPs are genotyped, minimum minor allele frequency is 0.05, and heritability is 0.3

Table 2 Expected precision of GEBV as predicted by Eq. 1 (true precision), Daetwyler's formula 2 and 3, Eqs. 5 and 7 when 500 SNPs are genotyped in a reference population of size 5000, depending on minimum minor allele frequency (MAF) and heritability

MAF	h_0^2	True Eq. (1)	Daetwyler formula (2)	Daetwyler formula (3)	Elsen Eq. (5)	Elsen Eq. (7)
0.05	0.1	0.514	0.500	0.513	0.526	0.513
0.05	0.3	0.798	0.750	0.798	0.811	0.798
0.05	0.5	0.901	0.833	0.901	0.909	0.902
0.05	0.7	0.955	0.875	0.955	0.959	0.955
0.025	0.1	0.513	0.500	0.513	0.526	0.513
0.025	0.3	0.797	0.750	0.798	0.811	0.798
0.025	0.5	0.901	0.833	0.901	0.909	0.902
0.025	0.7	0.955	0.875	0.955	0.959	0.955

$C = 1/\log(2N_e)$. Alternatively, we could set $\alpha = f_{min}$ and $E\left[\frac{1}{\sigma_k^2}\right] = 1 + \frac{1-2f_{min}}{2f_{\min}(1-f_{\min})\log\left(\frac{1-f_{min}}{f_{min}}\right)}$. Figure 2 shows the values of these expectations for different minimum MAF. When the minimum MAF is higher than 5%, this correction factor $E\left[\frac{1}{\sigma_k^2}\right]$ is almost 2 and, in most cases, the expected precision of GEBV is close to:

$$\hat{r}^2_{(2)} \cong \frac{N}{N+M\lambda} - \lambda\left(\frac{M}{N}\right)^2\left(\frac{N}{N+M\lambda}\right)^3. \quad (8)$$

Goddard [20] also derived the precision of GEBV in this case of a U-shaped distribution of allele frequencies. In his formulation (formula (8) of his paper), the expectation of the precision of GEBV depends on the ratio $\lambda = \frac{1-h^2}{h^2}\frac{M_e}{\log(2N_e)}$, where M_e is the effective number of SNPs genotyped. Replacing $\log(2N_e)$ by $-\log(f_{min})$, as above, and assuming all SNPs are unlinked, resulting in $M_e = M$, we compared the two approaches by using simulated data, as explained in the previous section. Results in Table 3 suggest that, in most cases, Goddard's formula

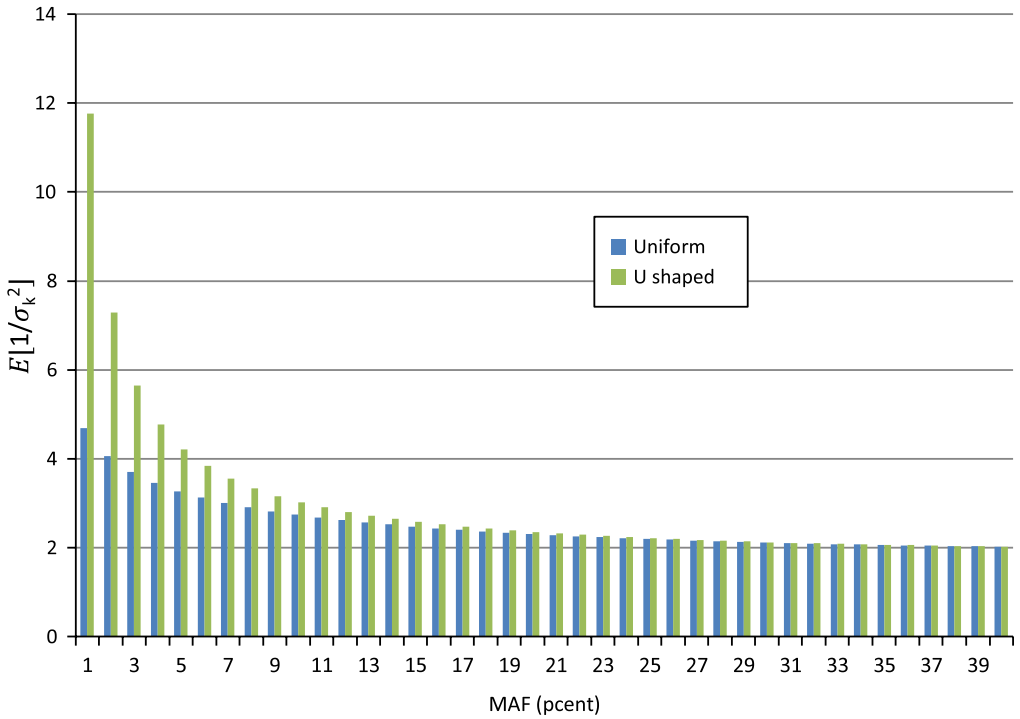

Fig. 2 Expectation of the inverse of variance of allele frequencies as a function of minimum allele frequencies (MAF), assuming a uniform or U-shaped distribution of allele frequencies

Table 3 Expected precision of GEBV when the distribution of allele frequencies is U-shaped, as predicted by Eqs. 1 and 8, by Daetwyler formula (3) and according to Goddard [20], when 500 SNPs are genotyped in a reference population of size 5000, depending on minimum minor allele frequency (MAF) and heritability

MAF	h_0^2	True Eq. (1)	Elsen Eq. (8)	Daetwyler formula (3)	Goddard [20]
0.05	0.1	0.513	0.513	0.513	0.491
0.05	0.3	0.798	0.798	0.798	0.759
0.05	0.5	0.901	0.902	0.901	0.867
0.05	0.7	0.955	0.955	0.955	0.930
0.025	0.1	0.513	0.513	0.513	0.537
0.025	0.3	0.798	0.798	0.798	0.791
0.025	0.5	0.901	0.901	0.901	0.886
0.025	0.7	0.955	0.955	0.955	0.941

underestimates the precision, while the two other formulations are close to the expected value obtained from simulation.

Distribution of SNP effects

In previous developments, it was assumed that each SNP had its own distribution of effects with a variance of $\sigma^2_{\beta_k}$. This was the condition assumed by Meuwissen et al. [1] when defining BayesA and BayesB Markov chain Monte Carlo approaches to genomic evaluation. This was justified in practice because the authors did not work at a single locus level but considered haplotypes of markers around each tested position, while theoretical justifications were given in the Bayesian LASSO context by Park and Cassela [35]. Alternatively, Meuwissen et al. [1] considered a unique variance σ^2_β under the GBLUP approach, which is also the case for the fitted SNPs in model BayesC π [36]. The assumption of an equal contribution of each SNP to the genetic variance is no longer valid and variance σ^2_β is linked to genetic variance by $\sigma^2_g = \left(\sum_k \sigma^2_k\right)\sigma^2_\beta$.

In this case, approximations of Eq. (1) for the expected precision of GEBV can be obtained using the same approach as before, using a matrix Taylor series expansion. As shown in Additional file 2, the first order approximation is given by:

$$\hat{r}^2_{(9)} = 1 - \frac{\lambda M}{N}\left(1 + \frac{1}{1-2f_{min}}\frac{\hat{\lambda}_\beta}{N\sqrt{1+\frac{2\hat{\lambda}_\beta}{N}}}\log\left(\frac{2f_{min}-1+\sqrt{1+\frac{2\hat{\lambda}_\beta}{N}}}{1-2f_{min}+\sqrt{1+\frac{2\hat{\lambda}_\beta}{N}}}\right)\right), \quad (9)$$

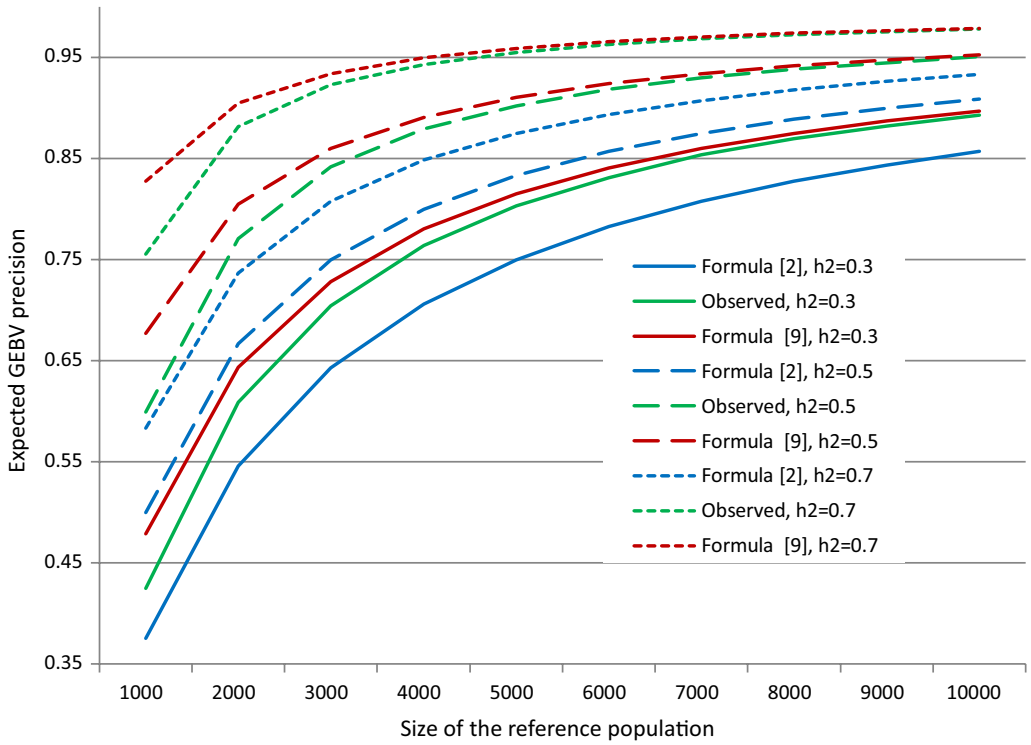

Fig. 3 Expected precision of GEBV as predicted by Eqs. (1), (2) and (9) when 500 SNPs are genotyped and minimum minor allele frequency is 0.05. Case of a single prior variance of SNP effects and three levels of heritability ($h^2 = 0.3$, 0.5 and 0.7)

where $\hat{\lambda}_\beta = \frac{M}{1-2f_{min}} \frac{1-6f_{min}^2+2f_{min}^3}{3} \lambda$. An illustration of the quality of this approximation is in Fig. 3. The quality of the first order approximation of Eq. (1), which always overestimates precision, increased with population size and heritability and appeared to be satisfactory when $N \geq 5000$.

Multivariate prediction

A simple generalisation of the expected precision of GEBV can be obtained when retaining the previous assumption. A total of n_c traits are recorded in the reference population and this information is used to predict the global genetic value of the candidate: $\gamma = \sum_j a_j g_j = \mathbf{ag}$, where $\mathbf{a}$ is a vector of n_c economic weights and $\mathbf{g}$ the column vector of n_c genetic values, i.e. $g_j = \mathbf{w}\beta_j$. Vector $\boldsymbol{\beta}_j$ is a vector of the M SNP effects on trait j ($\boldsymbol{\beta'}_j = (\beta_{j1}, \ldots, \beta_{jM})$). We assume that the vector of genotypes ($\mathbf{w}$) is the same for all traits. All previous assumptions are retained: all SNPs have an effect on all traits, all SNPs have an equal contribution to genetic variance for each trait, and individuals are unrelated. SNP effects are distributed with specific prior variances of $\sigma^2_{\beta_{jk}}$, with zero correlations between SNPs. It is also assumed that the effects of SNP k on traits j and j' are correlated, with a covariance of $\sigma^2_{\beta_{jj'k}}$.

The objective is to predict precision $r^2 = \frac{cov(\gamma,\hat{\gamma})^2}{v(\gamma)v(\hat{\gamma})}$, where $\hat{\gamma} = \mathbf{a}\hat{\mathbf{g}}$, with $\hat{\mathbf{g}}$ being the vector of GEBV. Thus, we need the variance $v(\hat{\mathbf{g}})$ of these GEBV, a $n_c \times n_c$ matrix. As detailed in Additional file 3, this variance is estimated at the first order using:

$$\widehat{v(\hat{\mathbf{g}})} = Nv(\mathbf{g})(Nv(\mathbf{g}) + Mv(\mathbf{e}))^{-1}v(\mathbf{g}), \quad (10)$$

which is an obvious generalisation of the equivalent equation in the single-trait situation, which led to Eq. (5).

Discussion

Using classical statistical theory, the expected precision of GEBV based on marker-based BLUP was derived simply. Numerical approximations, based on a matrix Taylor series expansion, were produced for simple situations. From simulations that were consistent with the assumptions corresponding to these situations, these approximations performed similarly to and often better than the formulae for precision of GEBV that were previously published. However, the framework developed here is simpler and enables direct generalisations.

The first order approximation proposed here (Eq. 5) differs from formula (2) of Daetwyler et al. [17] by a $(1 - h_0^2)$ term. Those approximations differ by the way

the error term variance is defined when a single SNP effect is estimated. In Eq. (2), it was assumed that this error term variance is the total phenotypic variance, because when estimating a unique SNP effect, all other SNP effects participate to the error term. Too much error is assumed with this approximation and the precision is under evaluated. Equation (5) behaves as if all other SNP effects were perfectly estimated, limiting the error term to the only non-genetic part. This gives an overestimation of the GEBV precision. The second order approximations try to correct for these under- or overestimations: Eq. (3) replaces the $1 - h_0^2$ term of Eq. (5) by $1 - r^2 h_0^2$, which corrects for the non-perfect estimation of other SNPs effects. Equation (7) accounts for the lack of orthogonality between the SNPs.

Asymptotic behaviours of first order approximations are not the same: when all the observed variability has an (additive) genetic origin, i.e. when $h_0^2 = 1$, formula (2) simplifies to $\frac{N}{N+M}$, while our Eq. (5) predicts a perfect precision of GEBV. This discrepancy disappears when correcting Eq. (2) for non-perfect estimation of other SNPS (Eq. 3). With this correction Eq. 3 predicts a perfect precision of GEBV when $h_0^2 = 1$. In spite of being algebraically very different, the second order approximations underlying the Eqs. (3) and (7) worked very similarly and produced results that were very close to those observed from simulations.

The hypotheses that underlie the equations derived here are strong and efforts should be made to overcome these constraints. First, it was assumed that all genetic variability is explained by the SNPs included in the evaluation. Although this is increasingly true as the size of SNP chips grows towards a full knowledge of genomes by resequencing and imputation, other polymorphisms, including copy number variations (CNV), may play a role and the genotype information obtained is still far from sufficient to fully explain genetic variability. It has been suggested [8, 22] that the proportion b of the genetic variance explained by the markers should be taken into account through a reduction in the heritability (from h^2 to bh^2) in the equations used and, using path coefficient theory, through a regression of precision of GEBV by b (from r^2 to br^2). This is easily implemented in the equations provided in this paper.

A second central hypothesis was independence between SNPs. With the current sizes of SNP chips, which will be even larger in the future, close SNPs are in LD and cannot be considered to be independent. This dependence means that non diagonal terms of $E\left[\mathbf{X}'\mathbf{X}\right]$ are non-null, with $E\left[\mathbf{X}'\mathbf{X}\right]_{kl} = 2N\Delta_{kl}$, where Δ_{kl} is the LD between SNPs k and l. Equations can be derived for this situation, based on principles similar to the theory given here, but they are cumbersome, e.g. [33]. The concept of effective independent chromosomal segments has been discussed [17] and formalised [20] as an alternative to the true number of markers. The idea is that the precision of the genomic prediction model "depends on the variation in the realised relationship between pairs of animals" [20]. Then, the effective number of loci is defined as the "number of independent loci that gives the same variance of realised relationships as obtained in the more realistic situation" [20]. Solutions have also been proposed to estimate the effective number of loci, from population genetics considerations [20, 21, 37] or from real data (e.g. [32]).

A third assumption was the absence of relationships between individuals in the reference population and between candidates in the reference population. A formalisation of situations where individuals are related was proposed [33] but only for the first order approximations of precision. Although relationships between reference individuals and candidates were accounted for by using this first order approximation, this was not the case for the structure of the reference population itself. Therefore, further efforts are needed, which is particularly important since it is clear that (1) genomic predictions of breeding values arise only partly from historical LD and increase in precision when individuals in the reference population and candidates are more closely related [26, 38–40], and (2) the structure of the reference population is a key factor in the precision on GEBV, e.g. [41].

The predicted variances of SNP effects calculated by Eq. (10) in the multivariate situation were obtained under strong assumptions. First, it was assumed that GEBV are computed using a multivariate approach that considers correlations between the effects of SNPs on different traits. However, in practice, GEBV are often computed using single-trait algorithms. In our formulation, this is equivalent to omitting the off-diagonal terms in matrices $\mathbf{B}_{kk}$ and $\mathbf{E}$ when estimating the SNP effects $\widehat{\boldsymbol{\beta}}$. In this case, the variance of those effects, and $v\left(\hat{\mathbf{g}}\right)$, do not simplify to the equations derived in the case studied. A second important assumption was that all SNPs contributed equally to genetic covariances, as a direct extension of the single trait situation studied. The alternative assumption of unique (regardless of the SNPs) variances and covariances (e.g. $\sigma^2_{\beta_{jk}} = \sigma^2_{\beta_j} \forall k$) is also possible, as described in the previous section. Both these assumptions are, however, questionable, in particular because genetic correlations lower than 1 suggest that only a limited proportion of the SNPs (and underlying QTL) affect all traits. Extra prior information about the underlying genetic architecture of these correlations would be useful in this regard.

A few other assumptions are used in the current paper, including the additivity and *i.i.d.* of QTL effects, and the use of GBLUP. As long as the objective is to model and optimise breeding plans, then only relative values will be

of interest and we assume that these assumptions are not critical for those comparisons.

Conclusions

The objective of this paper was to provide a clear framework to derive predictive equations to estimate the precision of GEBV. Such equations can generate results in a second and thus enable the optimisation of a breeding program design through intensive numerical exploration. Not the entire complexity of practical breeding programs was included in the simple formulae derived here and in previously published papers. The purpose was to support the a priori comparison of breeding schemes, rather than to evaluate actual breeding schemes. The exact formulation of precision involves the expectation of the inverse of a linear function of the genomic relationship matrix, which cannot be calculated from simple algebra but can be approximated by a Taylor series development, as was already suggested by Goddard [20]. Second order approximations produced nearly perfect estimates of this precision, when compared to the results obtained by simulating data that are in agreement with the assumptions required to obtain the equations to estimate the precision of GEBV. We proposed several generalisations for the estimates of precision for this initial case, including multi-trait evaluation. Other situations can also be derived within the framework presented here.

Additional files

Additional file 1. Diagonal elements of $E_X\left[D^{-1}ED^{-1}E\right]$. This file provides details about the algebraic derivation of the expectation given in the title.

Additional file 2. Algebra when the SNP effect variance σ^2_β is unique. This file shows the derivation in the alternative hypothesis for the SNP effect variance.

Additional file 3. The multivariate case. This file provides details about the algebraic derivation in the multivariate case.

Acknowledgements
Many thanks to Yvonne CJ Wientjes for the discussions we had on how precision of GEBV should be calculated. I am very grateful to the referee who gave excellent suggestions for the discussion part.

Competing interests
The author declares that he has no competing interests.

Ethics approval and consent to participate
Not applicable.

Publisher's Note

Received: 3 July 2017 Accepted: 1 December 2017
Published online: 27 December 2017

References

1. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.
2. Lande R, Thompson R. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics. 1990;124:743–56.
3. Schaeffer LR. Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet. 2006;123:218–23.
4. König S, Simianer H, Willam A. Economic evaluation of genomic breeding programs. J Dairy Sci. 2009;92:382–91.
5. McHugh N, Meuwissen THE, Cromie AR, Sonesson AK. Use of female information in dairy cattle genomic breeding programs. J Dairy Sci. 2011;94:4109–18.
6. de Roos AP, Schrooten C, Veerkamp RF, van Arendonk JAM. Effects of genomic selection on genetic improvement, inbreeding, and merit of young versus proven bulls. J Dairy Sci. 2011;94:1559–67.
7. Pryce JE, Goddard ME, Raadsma HW, Hayes BJ. Deterministic models of breeding scheme designs that incorporate genomic selection. J Dairy Sci. 2010;93:5455–66.
8. Meuwissen THE, Hayes BJ, Goddard ME. Accelerating improvement of livestock with genomic selection. Annu Rev Anim Biosci. 2013;1:221–37.
9. Sonesson AK, Meuwissen THE. Testing strategies for genomic selection in aquaculture breeding programs. Genet Sel Evol. 2009;41:37.
10. Ibáñēz-Escriche N, Fernando RL, Toosi A, Dekkers JCM. Genomic selection of purebreds for crossbred performance. Genet Sel Evol. 2009;41:12.
11. Wolc A, Arango J, Settar P, Fulton JE, O'Sullivan NP, Preisinger R, et al. Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. Genet Sel Evol. 2011;43:23.
12. Tribout T, Larzul C, Phocas F. Efficiency of genomic selection in a purebred pig male line. J Anim Sci. 2012;45:4164–76.
13. Shumbusho F, Raoul J, Astruc JM, Palhiere I, Elsen JM. Potential benefits of genomic selection on genetic gain of small ruminant breeding programs. J Anim Sci. 2013;91:3644–57.
14. Weller JI. Economic aspects of animal breeding. London: Chapman & Hall; 1994.
15. Tribout T, Larzul C, Phocas F. Efficiency of genomic selection in a purebred pig male line. J Anim Sci. 2012;90:4164–76.
16. Bastiaansen JW, Coster A, Calus MP, van Arendonk JA, Bovenhuis H. Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. Genet Sel Evol. 2012;44:3.
17. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One. 2008;3:e3395.
18. Legarra A, Robert-Granié C, Manfredi E, Elsen JM. Performance of genomic selection in mice. Genetics. 2008;180:611–8.
19. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
20. Goddard ME. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 2009;136:245–57.
21. Hayes BJ, Visscher PM, Goddard ME. Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res (Camb). 2009;91:47–60.
22. Goddard ME, Hayes BJ, Meuwissen THE. Using the genomic relationship matrix to predict the accuracy of genomic selection. J Anim Breed Genet. 2011;128:409–21.
23. Buch LH, Kargo M, Berg P, Lassen J, Sørensen AC. The value of cows in reference populations for genomic selection of new functional traits. Animal. 2011;6:880–6.
24. Clark SA, Hickey JM, Daetwyler HD, van der Werf JHJ. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet Sel Evol. 2012;44:4.
25. Wientjes YCJ, Veerkamp RF, Calus MPL. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics. 2013;193:621–31.
26. Habier D, Fernando RL, Garrick DJ. Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics. 2013;194:597–607.
27. Erbe M, Gredler B, Seefried FR, Bapst B, Simianer H. A function accounting for training set size and marker density to model the average accuracy of genomic prediction. PLoS One. 2013;8:e81046.

28. Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007;177:2389–97.
29. Pszczola M, Strabel T, Mulder HA, Calus MPL. Reliability of direct genomic values for animals with different relationships within and to the reference population. J Dairy Sci. 2012;95:389–400.
30. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams A. The impact of genetic architecture on genome-wide evaluation methods. Genetics. 2010;185:1021–31.
31. Brard S, Ricard A. Is the use of formulae a reliable way to predict the accuracy of genomic selection? J Anim Breed Genet. 2015;132:207–17.
32. Wientjes YCJ, Bijma P, Veerkamp RF, Calus MPL. An equation to predict the accuracy of genomic values by combining data from multiple traits, breeds, lines or environments. Genetics. 2016;202:799–823.
33. Elsen JM. Approximated prediction of genomic selection accuracy when reference and candidate populations are related. Genet Sel Evol. 2016;48:18.
34. Numerical Algorithms Group. https://www.nag.co.uk/.
35. Park T, Cassela G. The Bayesian LASSO. J Am Stat Assoc. 2008;103:681–6.
36. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics. 2011;12:186.
37. Visscher PM, Medland SE, Ferreira MAR, Morley KI, Zhu G, Cornes BK, et al. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet. 2006;2:e41.
38. Schopp P, Müller D, Technow F, Melchinger AE. Accuracy of genomic prediction in synthetic populations depending on the number of parents, relatedness, and ancestral linkage disequilibrium. Genetics. 2017;205:441–54.
39. Sun X, Fernando R, Dekkers J. Contributions of linkage disequilibrium and co-segregation information to the accuracy of genomic prediction. Genet Sel Evol. 2016;48:77.
40. Scutari M, Mackay I, Balding D. Using genetic distance to infer the accuracy of genomic prediction. PLoS Genet. 2016;12:e1006288.
41. Rincent R, Laloë D, Nicolas S, Altmann T, Brunel D, Revilla P, et al. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). Genetics. 2012;192:715–28.