Approximated prediction of genomic selection accuracy when reference and candidate populations are related

Jean-Michel Elsen

Elsen *Genet Sel Evol (2016) 48:18*
DOI 10.1186/s12711-016-0183-3

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

Approximated prediction of genomic selection accuracy when reference and candidate populations are related

Jean-Michel Elsen[1,2*]

Abstract

Background: Genomic selection is still to be evaluated and optimized in many species. Mathematical modeling of selection schemes prior to their implementation is a classical and useful tool for that purpose. These models include formalization of a number of entities including the precision of the estimated breeding value. To model genomic selection schemes, equations that predict this reliability as a function of factors such as the size of the reference population, its diversity, its genetic distance from the group of selection candidates genotyped, number of markers and strength of linkage disequilibrium are needed. The present paper aims at exploring new approximations of this reliability.

Results: Two alternative approximations are proposed for the estimation of the reliability of genomic estimated breeding values (GEBV) in the case of non-independence between candidate and reference populations. Both were derived from the Taylor series heuristic approach suggested by Goddard in 2009. A numerical exploration of their properties showed that the series were not equivalent in terms of convergence to the exact reliability, that the approximations may overestimate the precision of GEBV and that they converged towards their theoretical expectations. Formulae derived for these approximations were simple to handle in the case of independent markers. A few parameters that describe the markers' genotypic variability (allele frequencies, linkage disequilibrium) can be estimated from genomic data corresponding to the population of interest or after making assumptions about their distribution. When markers are not in linkage equilibrium, replacing the real number of markers and QTL by the "effective number of independent loci", as proposed earlier is a practical solution. In this paper, we considered an alternative, *i.e.* an "equivalent number of independent loci" which would give a GEBV reliability for unrelated individuals by considering a sub-set of independent markers that is identical to the reliability obtained by considering the full set of markers.

Conclusions: This paper is a further step towards the development of deterministic models that describe breeding plans based on the use of genomic information. Such deterministic models carry low computational burden, which allows design optimization through intensive numerical exploration.

Background

The effectiveness of genomic selection comes from the possibility of predicting breeding values on un-phenotyped and young animals [1]. Genomic selection promised and proved to be extremely efficient and beneficial for dairy cattle (e.g. [2–7]), but debate continues for other species and production sectors (e.g. [8–12]). A key criterion to decide whether or not selection schemes (also referred to here as breeding plans) should include genomic information is the reliability of the genomic predictor. It was clearly shown that this reliability depends on the structure of the reference population and on the characteristics of the marker set used. The size of this reference population, its diversity, the genetic distance between the reference and the group of selection candidates genotyped, the number of markers, and the degree

*Correspondence: elsen@toulouse.inra.fr; jean-michel.elsen@toulouse.inra.fr
[2] Animal Genetics and Breeding Unit, University of New England, Armidale, Australia
Full list of author information is available at the end of the article

or strength of the linkage disequilibrium are the main factors that influence this reliability [13–23].

An extensive literature exists on the mathematical modeling of selection schemes prior to their implementation, in order, for instance, to optimize their design, or to evaluate the usefulness of new technologies such as embryo transfer, sperm selection, DNA markers and others (e.g. [24] for a review). These models account for factors such as selection intensities and maintenance or loss of genetic variability. Among these parameters, the precision of breeding value estimates is central. To model genomic selection schemes, equations that predict this reliability as a function of the factors cited above are needed (e.g. [6, 25, 26]).

The quantitative influence of these factors (size of the reference population, its diversity, etc.) was assessed by simulation studies [18–21, 27, 28]. An equation that predicts the reliability of genomic evaluation in the very simple situation of independent quantitative trait loci (QTL), that are perfectly marked by single nucleotide polymorphisms (SNPs) and populations (reference and candidates) of unrelated individuals was derived [13]. This approach was extended to the case when only a part of the genetic variability is imperfectly marked by SNPs [16, 19], and the situation of non-independence between reference and candidate populations was explored [17]. It was demonstrated that genomic information captures historical linkage disequilibrium, short-term linkage between QTL and markers and additive relationships between reference and candidate individuals, the equation of the reliability accounting for these three phenomena being derived in a very simple case of one QTL marked by a single SNP [22].

A Taylor expansion of a matrix inverse involved in the reliability formula was suggested [18], which led to the algebraic development of an approximation. This approximation seems to work well in the simple situation but lacks generality. In this paper, an alternative approximation is proposed, opening a way to include non-independence between reference and candidate populations, and between markers.

After a formalization of the genomic selection context, the principles that underlie these approximations are presented and their properties are compared by using a simple example. Then, the new approximation is derived when reference and candidate animals are related. This is illustrated by some numerical examples. Finally, the extension to the linkage disequilibrium situation is described.

Methods

General framework

Although the prediction equations derived below were based on a number of simplifying assumptions, it is important to first draw a complete description of the biological framework, as a basis to subsequently simplify the discussion.

The SNP effects are estimated in a reference population, P_r, comprising n_r individuals. The genomic estimated breeding values (GEBV) are calculated for a population of candidates for selection and used in breeding, P_c, comprising n_c individuals.

Let $\mathcal{P} = (\mathcal{P}_r, \mathcal{P}_c)$ the population structure (including pedigree relationships between individuals and marker allele frequencies, but not including genotypes and phenotypes).

Individuals are characterized by their genotypes at n_M markers (observed) and at n_Q QTL (unknown). Alleles will be noted A_m and B_m for the marker m and A_q and B_q for the QTL q. Let $a_{tim} \in \{0, 1, 2\}$ and $a_{tiq} \in \{0, 1, 2\}$ be the numbers of B_m (and respectively, B_q) alleles that an individual i from population P_t (P_r or P_c) carries at marker m (respectively, QTL q). Let p_{tm} and p_{tq} be the frequencies of alleles B_m and B_q in P_t.

Genotypic values will be assigned to the different markers and QTL genotypes. Following [18], genotypes will be coded as $x_{tim} = a_{tim} - 2p_{tm}$ and $w_{tiq} = a_{tiq} - 2p_{tq}$. Different codifications can be proposed [15]. In particular, as described for instance in [29], genotypic values may be standardized, *i.e.* $x_{tim} = (a_{tim} - 2p_{tm})/\sigma_{tm}$ and $w_{tiq} = (a_{tiq} - 2p_{tq})/\sigma_{tq}$, with variances $\sigma^2_{tm} = 2p_{tm}(1 - p_{tm})$ and $\sigma^2_{tq} = 2p_{tq}(1 - p_{tq})$. Most of the following developments are given with the first codification here, and the results with the second codification are described in a specific section.

These genotypic values are assembled in matrices $\mathbf{X}$ (dim $(\mathbf{X}) = (n_r + n_c) \times n_M$) and $\mathbf{W}$ (dim $(\mathbf{W}) = (n_r + n_c) \times n_Q$). Sub-matrices corresponding to sub-populations will be noted in the following way: $\mathbf{X}' = (\mathbf{X}'_\mathbf{r}, \mathbf{X}'_\mathbf{c})$ and $\mathbf{W}' = (\mathbf{W}'_\mathbf{r}, \mathbf{W}'_\mathbf{c})$.

The genetic model assumes additivity of QTL effects. The additive genetic value of an individual is described as $g_{ti} = \sum_{q=1}^{n_Q} w_{tiq}\alpha_q$ and, in general, $\mathbf{g} = \mathbf{W}\alpha$. The phenotypic values when observed are $\mathbf{y} = \mathbf{g} + \boldsymbol{\varepsilon}$.

A statistical model describes the performances in the reference population as random variables for which the expectations are linear combinations of SNP effects: $y_{ri} = \sum_{m=1}^{n_S} X_{rim}\beta_m + e_{ri}$ and, in general, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$.

In these models, the SNP (or QTL) effects may be considered as fixed, or random. Since the number of SNPs is much bigger than the number of individuals ($n_M \gg n_r$) the second solution is generally chosen in the statistical model (but not always see [1, 13]).

In the random model, a distribution $\mathcal{L}(\boldsymbol{\theta}_\beta, \mathbf{V}_\beta)$ (respectively $\mathcal{L}(\boldsymbol{\theta}_\alpha, \mathbf{V}_\alpha)$) of the SNP (respectively QTL) effects is assumed, with $\boldsymbol{\theta}_\boldsymbol{\beta}$ (respectively, $\boldsymbol{\theta}_\alpha$) being the vector of expectations and $\mathbf{V}_\beta$ (respectively, $\mathbf{V}_\alpha$) being the matrix of variances. For a full description of the variability, the $\mathbf{V}_\beta$ and $\mathbf{V}_\alpha$ matrices are each subdivided into four blocks

corresponding to the reference and candidate populations and their covariances. Covariances between the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ vectors have also to be considered. Most generally, the SNP (QTL) effects are supposed *i.i.d.* giving $\mathbf{V}_\beta = \mathbf{I}\sigma^2_\beta$ ($\mathbf{V}_\alpha = \mathbf{I}\sigma^2_\alpha$). The interpretation of these QTL effects is nicely debated in Gianola et al. [30]. In the frequentist view, we simply have to imagine that QTL effects are randomly sampled from a distribution with a σ^2_α variance. In the Bayesian context, the prior variability of the SNP effects was most generally described as heteroskedastic or even coming from mixtures of SNPs with or without an effect on the trait.

The expectations $\boldsymbol{\theta}_\beta(\boldsymbol{\theta}_\alpha)$ are generally assumed equal to zero, but when information about population history is available (in particular, when we know it is a mixed population), non-zero values should be considered.

The vector $\mathbf{q} = \mathbf{X}\boldsymbol{\beta}$ is a quantity similar but not equal to the genetic value $\mathbf{g}$. Its element q_{ti} is the molecular score of individual i in population t. This vector may be segmented in two parts: $\mathbf{q}' = (\mathbf{q}'_\mathbf{r}, \mathbf{q}'_\mathbf{c})$.

Since the variances may be defined within a population, we have $v(\mathbf{q_r}|\mathbf{X}) = \mathbf{X_r}\mathbf{V}_{\beta\mathbf{r}}\mathbf{X}'_\mathbf{r}$, and $v(\mathbf{q_c}|\mathbf{X}) = \mathbf{X_c}\mathbf{V}_{\beta\mathrm{c}}\mathbf{X}'_\mathbf{c}$. The residual variance is $v(\mathbf{e}) = \mathbf{I}\sigma^2_e$. Assuming that the distribution of marker effects is centered ($\boldsymbol{\theta}_\boldsymbol{\beta} = \mathbf{0}$) and *i.i.d.* ($\mathbf{V}_{\beta r} = \mathbf{I}\sigma^2_{\beta r}$ and $\mathbf{V}_{\beta c} = \mathbf{I}\sigma^2_{\beta c}$), and extending Gianola et al. [30], we have $v(q_{ri}) = \mathrm{E_X}[v(q_{\mathrm{ri}}|\mathbf{X})] = \sigma^2_{\beta r}\sum_m 2p_{mr}(1-p_{mr}) = \sigma^2_{\beta r}\sum_m \sigma^2_{mr} = \sigma^2_{\beta r}\tau_r$ in the reference population, and $v(q_{ci}) = \sigma^2_{\beta\mathrm{c}}\tau_c$ in the candidate population. Assuming that the distribution of the marker effects and genotypes are the same in $\mathrm{P_r}$ and $\mathrm{P_c}$, *i.e.* $p_{rm} = p_{cm} = p_m, p_{rq} = p_{cq} = p_q$, thus $\tau_r = \tau_C = \tau$ and $\sigma^2_{\beta r} = \sigma^2_{\beta c} = \sigma^2_\beta$, we define $\sigma^2_q = \tau\sigma^2_\beta$. Thus, $v(\mathbf{q}|\mathbf{X}) = \frac{1}{\tau}\mathbf{X}\mathbf{X}'\sigma^2_q$. These equations hold even if the markers are in linkage disequilibrium (LD) as shown in Eq. A2 from Gianola et al. [30].

We note σ^2 as the total phenotypic variance, *i.e.* $\sigma^2 = \sigma^2_q + \sigma^2_e$, and ν^2 as the proportion of this variance explained by the molecular score $\left(\nu^2 = \frac{\sigma^2_q}{\sigma^2}\right)$. The ratio $\frac{\sigma^2_q}{\sigma^2_e}$ will be noted γ.

The SNP effects $\boldsymbol{\beta}$ may be estimated in different ways. The genomic best linear unbiased prediction (BLUP) will only be considered here, with $\hat{\boldsymbol{\beta}} = \mathrm{cov}(\boldsymbol{\beta}, \mathbf{y})\mathrm{var}(\mathbf{y})^{-1}\mathbf{y}$. Classically, this equation becomes $\hat{\boldsymbol{\beta}} = \sigma^2_\beta\mathbf{X}'_\mathbf{r}\left[\sigma^2_\beta(\mathbf{X_r}\mathbf{X}'_\mathbf{r} + \mathbf{I}\lambda_\beta)\right]^{-1}\mathbf{y} = (\mathbf{X}'_\mathbf{r}\mathbf{X_r} + \mathbf{I}\lambda_\beta)^{-1}\mathbf{X}'_\mathbf{r}\mathbf{y}$ with $\lambda_\beta = \sigma^2_e/\sigma^2_\beta$. The linear combination $\hat{\boldsymbol{q}}_\boldsymbol{c} = \mathbf{X_c}\hat{\boldsymbol{\beta}}$ is the GBLUP vector for candidates in $\mathrm{P_c}$. It must be emphasized that these estimations and predictions are conditional on the genotypic structures defined by $\mathbf{X}$ ($\mathbf{X_r}$ and $\mathbf{X_c}$).

Given $\mathbf{X}$, the reliability of the GBLUP is $r^2(g_{ci}, \hat{q}_{ci}|\mathbf{X}) = \frac{cov^2(g_{ci},\hat{q}_{ci}|\mathbf{X})}{v(g_{ci}|\mathbf{X})v(\hat{q}_{ci}|\mathbf{X})}$.

In [16], the reliability is described (Eq. 6 in [16]) as $r(g_{ci}, \hat{q}_{ci}) = r(g_{ci}, q_{ci}) \times r(q_{ci}, \hat{q}_{ci})$, by ignoring the conditioning on $\mathbf{X}$. In Goddard et al. [18], the reliability is described as $r^2_{g_{ci},\hat{q}_{ci}} = \frac{v(\hat{q}_{ci})}{v(g_{ci})} = \frac{v(q_{ci})}{v(g_{ci})}\frac{v(\hat{q}_{ci})}{v(q_{ci})}$. In this formulation, $\frac{v(q_{ci})}{v(g_{ci})}$ is the proportion of the genetic variance explained by the markers and $\frac{v(\hat{q}_{ci})}{v(q_{ci})}$ is the accuracy of estimated marker effects. This is similar to the $\mathrm{qr}_{\hat{\mathrm{Q}}}$ reported by Dekkers et al. [25]. All these reliability formulae are approximations since $cov^2(g_{ci}, \hat{q}_{ci}) = cov^2\left(\sum w_{ciq}\alpha_q, \sum x_{cis}\hat{\beta}_s\right) \neq v(\hat{q}_{ci}) = v\left(\sum x_{cis}\hat{\beta}_s\right)$, in general.

Situation analyzed in this paper

In the following, ignoring the difficulty that was mentioned above, we will assume $r^2(q_{\mathrm{ci}}, \hat{q}_{ci}|\mathbf{X}) = \frac{cov^2(q_{ci},\hat{q}_{ci}|\mathbf{X})}{v(q_{ci}|\mathbf{X})v(\hat{q}_{ci}|\mathbf{X})} = \frac{v(\hat{q}_{ci}|\mathbf{X})}{v(q_{ci}|\mathbf{X})}$. We are interested in a single candidate in $\boldsymbol{P_c}$ with a $\mathbf{x_c}$ vector of marker genotypes.

Formulae were simplified in two ways. (1) the i index of the candidate was omitted in the following developments: the genetic value of the candidate is noted q_c, estimated by $\hat{q}_c = cov(q_c, \mathbf{y})v(\mathbf{y})^{-1}\mathbf{y}$, and its precision is $r^2(q_\mathrm{c}, \hat{q}_c|\mathbf{X}) = \frac{v(\hat{q}_c|\mathbf{X})}{v(q_c|\mathbf{X})}$, with $v(q_\mathrm{c}|\mathbf{X}) = \sigma^2_\beta\mathbf{x_c}\mathbf{x}'_\mathbf{c}$ and $v(\hat{q}_c|\mathbf{X}) = \sigma^2_\beta\mathbf{x_c}\mathbf{X}'_\mathbf{r}(\mathbf{X_r}\mathbf{X}'_\mathbf{r} + \mathbf{I}\lambda_\beta)^{-1}\mathbf{X_r}\mathbf{x}'_\mathbf{c}$ (where $\mathbf{x_c}$ is a row vector); (2) the r indices of reference individuals were most often omitted, which resulted in y_i for their phenotypes and q_i for their molecular scores.

In fact, our objective was to estimate the expectation of this precision across the variation domain of $\mathbf{X_r}$ and $\mathbf{x_c}$ given the pedigree structure $(\mathcal{P}_\boldsymbol{r}, \mathcal{P}_\boldsymbol{c})$: $\mathrm{E_X}\left[r^2(q_\mathrm{c}, \hat{q}_c|\mathbf{X})|\mathcal{P}\right]$. It will be noted $E\left[r^2_{q_\mathrm{c},\hat{q}_\mathrm{c}}\right]$.

The following approximation was made: $\mathrm{E}\left[r^2_{q_\mathrm{c},\hat{q}_\mathrm{c}}\right] = \frac{\mathrm{E_X}[v(\hat{q}_c|\mathbf{X})]}{\mathrm{E_X}[v(q_c|\mathbf{X})]} = \frac{\mathrm{E}[v(\hat{q}_c)]}{\mathrm{E}[v(q_c)]}$.

Let $\mathbf{A}$ be the pedigree relationship matrix between individuals in $\mathcal{P}$. Its blocks are $\mathbf{A} = \begin{pmatrix} \mathrm{a_{cc}} & \mathbf{A_{cr}} \\ \mathbf{A_{rc}} & \mathbf{A_{rr}} \end{pmatrix}$. Let $\mathbf{G}^* = \mathbf{X}\mathbf{X}' = \begin{pmatrix} \mathbf{x_c}\mathbf{x}' & \mathbf{x_c}\mathbf{X}'_\mathbf{r} \\ \mathbf{X_r}\mathbf{x}'_\mathbf{c} & \mathbf{X_r}\mathbf{X}'_\mathbf{r} \end{pmatrix}$, which results in $\mathbf{V} = \frac{1}{\tau}\mathbf{G}^*\sigma^2_q + \mathbf{I}\sigma^2_e$. It must be noted that the σ^2_e term in the diagonal of the $\mathbf{V}$ submatrix corresponding to the candidate population is artificial since candidates are not phenotyped.

We have $\mathrm{E}[\mathbf{G}^*] = \mathbf{A}\tau$. The limits of this equality will be discussed below. As indicated above, the denominator of the expected reliability $\mathrm{E_X}[v(q_\mathrm{c}|\mathbf{X})]$, is $\tau\sigma^2_\beta = \sigma^2_q$. Approximating $\mathrm{E}\left[v(\hat{q}_c)\right]$ by $\mathrm{E}[cov(q_c, \mathbf{y})]E[v(\mathbf{y})]^{-1}E[cov(\mathbf{y}, q_c)]$ is useless because it makes an oversimplification of the relationships between the

reference and the candidate population: it considers separately the marginal distributions of $\mathbf{x_c}\mathbf{X_r'}$ and $(\mathbf{X_r}\mathbf{X_r'} + \mathbf{I}\lambda_\beta)^{-1}$, while these random matrices are correlated. Estimating directly $\mathrm{E}[cov(q_c,\ \mathbf{y})v(\mathbf{y})^{-1}cov(\mathbf{y}, q_c)]$ seems impossible in the general case. The approach of Goddard et al. [18] avoids this difficulty, *i.e.* the variance $v(\hat{q}_c|\mathbf{X}) = \sigma^2_\beta \mathbf{x_c}\mathbf{x_c'} + \sigma^2_e - \frac{1}{\{\mathbf{V}^{-1}\}_{cc}}$, and $\mathbf{V}^{-1}$ is approximated by a second degree Taylor expansion ($\mathbf{V}^{-1} \sim \mathbf{\Lambda}(\mathbf{X})$), giving $v(\hat{q}_c|\mathbf{X}) \sim \sigma^2_\beta \mathbf{x_c}\mathbf{x_c'} + \sigma^2_e - \frac{1}{\Lambda_{cc}(\mathbf{x_c},\mathbf{X_r})}$.

Alternative approximations of the reliability

Extension of Goddard's formula

In their "heuristic approximation for $\mathbf{V}^{*-1}$", Goddard et al. [18] considered the situation where unrelated individuals are included in the reference and candidate populations, that is $\mathrm{E}[\mathbf{G}^*] = \mathbf{I}\tau$ and $\mathbf{G}^* = \mathbf{I}\tau + \mathbf{E}$, with $\mathbf{E}$, a "noise" matrix centered on the null matrix $\mathbf{0}$. A direct extension of their development would be the following. The matrix $\mathbf{V} = \frac{1}{\tau}\mathbf{G}^*\sigma^2_q + \mathbf{I}\sigma^2_e$ can be written as:

$$\mathbf{V} = \sigma^2_e(\mathbf{I} + \mathbf{A}\gamma)[\mathbf{I} + \mathbf{D}\gamma],$$

$$\text{with } \mathbf{D} = (\mathbf{I} + \mathbf{A}\gamma)^{-1}\left(\tfrac{1}{\tau}\mathbf{G}^* - \mathbf{A}\right) = \mathbf{T}\left(\tfrac{1}{\tau}\mathbf{G}^* - \mathbf{A}\right),$$

and $\gamma = \frac{\sigma^2_q}{\sigma^2_e}$. Thus, $\mathbf{V}^{-1} = \frac{1}{\sigma^2_e}[\mathbf{I} + \mathbf{D}\gamma]^{-1}\mathbf{T}$.

The inverse matrix $[\mathbf{I} + \mathbf{D}\gamma]^{-1}$ will be approximated using a Taylor series. It must be emphasized that the Taylor series $\mathbf{I} - \mathbf{D}\gamma + (\mathbf{D}\gamma)^2 - (\mathbf{D}\gamma)^3 + \cdots$ converges towards $[\mathbf{I} + \mathbf{D}\gamma]^{-1}$ only if the highest Eigen value of $\mathbf{D}\gamma$ is smaller than 1, *i.e.* if $(\mathbf{D}\gamma)^{\mathrm{t}} \to \mathbf{0}$ when $\mathrm{t} \to \infty$.

The second order approximation of $\mathbf{V}^{-1}$ is equal to $\frac{1}{\sigma^2_e}(\mathbf{I} - \mathbf{D}\gamma + \mathbf{D}^2\gamma^2)\mathbf{T}$. As $\mathrm{E}[\mathbf{D}] = \mathbf{0}$ and $\mathrm{E}\left[\mathbf{D}^2\right] = \mathbf{T}\left(\frac{1}{\tau^2}\mathrm{E}\left[\mathbf{G}^*\mathbf{T}\mathbf{G}^*\right] - \mathbf{A}\mathbf{T}\mathbf{A}\right)$, its expectation $\mathrm{E}[\mathbf{\Lambda}] = \frac{1}{\sigma^2_e}(\mathbf{I} - \mathrm{E}[\mathbf{D}]\gamma + \mathrm{E}\left[\mathbf{D}^2\right]\gamma^2)\mathbf{T}$

i.e. $\mathrm{E}[\mathbf{\Lambda}] = \frac{1}{\sigma^2_e}\left(\mathbf{I} - \gamma^2\mathbf{T}\mathbf{A}\mathbf{T}\mathbf{A} + \frac{\gamma^2}{\tau^2}\mathrm{E}\left[\mathbf{T}\mathbf{G}^*\mathbf{T}\mathbf{G}^*\right]\right)\mathbf{T}$.

Finally, the reliability of the candidate GBLUP is approximated by:

$$\tilde{\mathrm{E}}\left[r^2_{q_c,\hat{q}_c}\right] \sim \frac{1}{v^2} - \frac{1}{\gamma\mathbf{T}_{cc} - \gamma^3\{\mathbf{T}\mathbf{A}\mathbf{T}\mathbf{A}\mathbf{T}\}_{cc} + \frac{\gamma^3}{\tau^2}\left\{\mathbf{T}\mathrm{E}\left[\mathbf{G}^*\mathbf{T}\mathbf{G}^*\right]\mathbf{T}\right\}_{cc}}. \quad (1)$$

A difficulty with this approximation comes from the $\mathbf{T}$ term. As an example, consider a reference population composed of n_r half-sibs of the candidate, $\mathbf{T} = \xi\mathbf{I} + \psi\mathbf{J}$ with $\xi = \frac{4}{4+3\gamma}$. As $\mathbf{T^t} = \xi^{\mathbf{t}}\mathbf{I} + \left[n_r^{\mathbf{t}}\xi^{\mathbf{t}} + \cdots\right]\mathbf{J}$, the $\mathbf{J}$ coefficient will tend to ∞ as soon as $n_r\xi = \frac{4n_r}{4+3\gamma} > 1$, a very realistic situation. Thus, the convergence of the Taylor series will be a balance between the increase of $\mathbf{T^t}$ and decrease of $[\mathbf{D}\gamma]^{\mathbf{t}}$.

Another approximation of the reliability

Principle

Using the classical matrix inversion lemma, the variance $v(\hat{q}_c|\mathbf{x_c},\mathbf{X_r}) = \sigma^2_\beta\mathbf{x_c}\mathbf{X_r'}\left(\mathbf{X_r}\mathbf{X_r'} + \mathbf{I}\lambda_\beta\right)^{-1}\mathbf{X_r}\mathbf{x_c'}$ may also be defined as $\boldsymbol{v}(\hat{q}_c|\mathbf{x_c},\mathbf{X_r}) = \sigma^2_\beta\mathbf{x_c}\mathbf{x_c'} - \sigma^2_e\mathbf{x_c}\left(\mathbf{X_r'}\mathbf{X_r} + \mathbf{I}\lambda_\beta\right)^{-1}\mathbf{x_c'}$.

$\mathbf{X_r'}\mathbf{X_r}$ is a very large matrix ($n_M \times n_M$) that describes the LD between markers: its elements tend to be smaller when they are more distant from the diagonal.

Elements of $\mathrm{E}[\mathbf{X_r'}\mathbf{X_r}]$ are the following: $\mathrm{E}\left[\mathbf{X_r'}\mathbf{X_r}\right]_{\mathrm{ml}} = \mathrm{E}\left[\sum_{\mathrm{i}}(a_{im} - 2p_m)(a_{il} - 2p_l)\right] = 2n_r\Delta_{ml}$, with Δ_{ml} the LD between loci m and l.

$\mathrm{E}[\mathbf{X_r'}\mathbf{X_r}]_{\mathrm{mm}} = \mathrm{E}[\sum_{\mathrm{i}}(a_{im} - 2p_m)^2] = n_r\sigma^2_{\mathrm{m}}$. Let $\mathbf{C} = \mathbf{I}\lambda_\beta + n_r\mathrm{diag}\left[\sigma^2_1,\ldots,\sigma^2_{n_M}\right]$, the $\mathbf{X_r'}\mathbf{X_r} + \mathbf{I}\lambda_\beta$ matrix may be written as:

$$\mathbf{X_r'}\mathbf{X_r} + \mathbf{I}\lambda_\beta = \left[\left(\mathbf{X_r'}\mathbf{X_r} - n_r\mathrm{diag}\left[\sigma^2_1,\ldots,\sigma^2_{n_M}\right]\right)\mathbf{C}^{-1} + \mathbf{I}\right]\mathbf{C},$$

which results in:

$$\mathbf{X_r'}\mathbf{X_r} + \mathbf{I}\lambda_\beta = [\mathbf{I} + \mathbf{B}]\mathbf{C}.$$

The convergence of the Taylor series $\mathbf{I} - \mathbf{B} + \mathbf{B}^2 - \mathbf{B}^3 + \cdots$ to $(\mathbf{I} + \mathbf{B})^{-1}$ depends on the structure of the $\mathbf{B}$ matrix, which varies depending on the sample. However, we can examine the case of its expectation $\mathrm{E}[\mathbf{B}]$.

$\mathrm{E}[\mathbf{B}]_{mm} = 0$ and $\mathrm{E}[\mathbf{B}]_{ml} = \frac{2n_r\Delta_{ml}}{\lambda_\beta + n_r\sigma^2_1}$. The ratio λ_β is proportional to the number of markers ($\lambda_\beta = n_M\frac{\bar{\sigma}^2_{\mathrm{m}}\sigma^2_e}{\sigma^2_{\mathrm{q}}}$) and dominates the denominator when $n_M \gg n_r$. The (m, l) term in $\mathrm{E}[\mathbf{B}]^2$, *i.e.* $\mathrm{E}[\mathbf{B}]^2_{\boldsymbol{ml}} = \sum_k \frac{4n_r^2\Delta_{mk}\Delta_{kl}}{(\lambda_\beta + n_r\sigma^2_{\mathrm{k}})(\lambda_\beta + n_r\sigma^2_{\mathrm{l}})}$, is of order $\frac{1}{n_M}$. Thus, we expect the Taylor series to converge to $(\mathbf{I} + \mathrm{E}[\mathbf{B}])^{-1}$.

First order approximation

At the first order, $v(\hat{q}_c|\mathbf{x_c},\mathbf{X_r}) \sim \sigma^2_\beta\mathbf{x_c}\mathbf{x_c'} - \sigma^2_e\mathbf{x_c}\mathbf{C}^{-1}(\mathbf{I} - \mathbf{B})\mathbf{x_c'}$ and the expectation of the reliability of the candidate GBLUP is approximated by $\tilde{\tilde{\mathrm{E}}}\left[r^2_{q_c,\hat{q}_c}\right] = 1 - \frac{\sigma^2_e\mathrm{E}\left[\mathbf{x_c}\mathbf{C}^{-1}(\mathbf{I}-\mathbf{B})\mathbf{x_c'}\right]}{\sigma^2_\beta\mathrm{E}\left[\mathbf{x_c}\mathbf{x_c'}\right]}$.

$$\mathbf{x_c}\mathbf{C}^{-1}(\mathbf{I} - \mathbf{B})\mathbf{x_c'} = \sum_m \frac{x^2_{cm}}{\lambda_\beta + n_r\sigma^2_m} + n_r\sum_m \frac{\sigma^2_m x^2_{cm}}{\left(\lambda_\beta + n_r\sigma^2_m\right)^2} - \mathbf{x_c}\mathbf{C}^{-1}\left(\mathbf{X_r'}\mathbf{X_r}\right)\mathbf{C}^{-1}\mathbf{x_c'}.$$

Using $\mathbf{x_c}\mathbf{C}^{-1}\mathbf{X_r'} = \left(\sum_{\mathrm{m}}\frac{\mathrm{x_{cm}X_{r1m}}}{\lambda_\beta + \mathrm{n_r}\sigma^2_{\mathrm{m}}} \cdots \sum_{\mathrm{m}}\frac{\mathrm{x_{cm}X_{rn_rm}}}{\lambda_\beta + \mathrm{n_r}\sigma^2_{\mathrm{m}}}\right)$, the last term is: $\sum_i\left(\sum_m \frac{x_{cm}X_{rim}}{\lambda_\beta + n_r\sigma^2_m}\right)^2$. Finally, the expectation is:

$$\tilde{\tilde{E}}\left[r^2_{q_c,\hat{q}_c}\right] = 1 - \frac{\lambda_\beta}{\tau}\left\{\sum_m\left[\frac{\sigma^2_m}{\lambda_\beta + n_r\sigma^2_m} + \frac{n_r\sigma^4_m}{\left(\lambda_\beta + n_r\sigma^2_m\right)^2}\right] - \sum_i\sum_m\left[\frac{\mathrm{E}\left[x^2_{cm}X^2_{rim}\right]}{\left(\lambda_\beta + n_r\sigma^2_m\right)^2} + \sum_{l\neq m}\frac{\mathrm{E}[x_{cm}X_{rim}x_{cl}X_{ril}]}{\left(\lambda_\beta + n_r\sigma^2_m\right)\left(\lambda_\beta + n_r\sigma^2_l\right)}\right]\right\}. \quad (2)$$

Results

Application in the case of independent markers

This situation either assumes low density marker information, or corresponds to the idea of an effective number of loci that was developed by Goddard [16, 31]. In the first case, the proportion of the genetic variance explained by the markers $\frac{v(q_{ci})}{v(g_{ci})}$ is small and this quantity should be considered when estimating the genomic precision.

First approximation

Using the notation $\mathbf{X}' = (\mathbf{x}'_\mathbf{c}, \mathbf{X}'_\mathbf{r})$, the (i,j) element of $\mathbf{G}^*\mathbf{T}\mathbf{G}^*$ is: $\left\{\mathbf{X}\mathbf{X}'\mathbf{T}\mathbf{X}\mathbf{X}'\right\}_{\mathrm{ij}} = \sum_l\sum_k t_{kl}\left(\sum_m X_{im}X_{km}\right)\left(\sum_m X_{jm}X_{lm}\right)$. Thus, elements of $\mathrm{E}[\mathbf{X}\mathbf{X}'\mathbf{T}\mathbf{X}\mathbf{X}']$ will involve expectations of fourth level moments of X_{im} within m joint distributions: $\mathrm{E}[\mathrm{E}\left[X^2_{im}X_{jm}X_{km}\right], \mathrm{E}\left[X^2_{im}X^2_{jm}\right], \mathrm{E}\left[X^3_{im}X_{jm}\right]X_{im}X_{jm}X_{km}X_{lm}]$, and $\mathrm{E}[X^4_{im}]$. Defining and $\tau_2 = \sum_m[2p_m(1-p_m)]^2$ a_{ij} the coancestry coefficient between individuals i and j, we found that [See Additional file 1]: $\left\{\mathrm{E}\left[\mathbf{X}\mathbf{X}'\mathbf{T}\mathbf{X}\mathbf{X}'\right]\right\}_{\mathrm{ij}} = \sum_l\sum_k t_{kl}\left(\frac{1}{2}\tau\alpha^{1111}_{ijkl} - \frac{1}{4}\tau_2\gamma^{1111}_{ijkl} + 4a_{ik}a_{jl}\left[\tau^2 - \tau_2\right]\right)$, where parameters $\alpha^{d_id_j\cdots d_K}_{ij\cdots K}$ and $\gamma^{d_id_j\cdots d_K}_{ij\cdots K}$ are functions of the probabilities of the identity states between gametes of $ij\cdots K$ individuals at marker m (Table 1). In the summations above, when individuals are repeated (e.g. $i = j$), the corresponding exponents are summed (e.g. $\alpha^{1111}_{iiil} = \alpha^{31}_{il}$). The resulting X_{im} moments are in Table 2.

Second approximation

The expectations $\mathrm{E}[x^2_{cm}x^2_{rim}]$ and $\mathrm{E}\left[x^2_{cm}x_{rim}x_{cl}x_{ril}\right]$ are also obtained from the coefficients in Table 1, i.e.: $\mathrm{E}\left[x^2_{cm}x^2_{rim}\right] = \frac{1}{2}\sigma^2_m\alpha^{22}_{ci} - \frac{1}{4}\sigma^4_m\gamma^{22}_{ci}$ and, when markers are independent, $\mathrm{E}[x_{cm}X_{rim}x_{cl}X_{ril}] = \mathrm{E}[x_{cm}X_{rim}].\mathrm{E}[x_{cl}X_{ril}] = 4a^2_{ci}\sigma^2_{\mathrm{m}}\sigma^2_{\mathrm{l}}$. Let $\rho_m = \frac{n_r\sigma^2_m}{\lambda_\beta + n_r\sigma^2_m}$. After some algebra, it appears that:

$$\tilde{\tilde{E}}\left[r^2_{q_c,\hat{q}_c}\right] = 1 - \frac{\lambda_\beta}{n_r\tau}\left\{\left(\sum_m\rho_m\right) + \left(\sum_m\rho^2_m\right)\left(1 + 4\bar{a}^2_{ci} + \frac{1}{4}\bar{\gamma}^{22}_{ci}\right) - \left(\sum_m\rho_m\right)^2\left(4\bar{a}^2_{ci}\right) - \left(\sum_m\frac{\rho^2_m}{\sigma^2_m}\right)\left(\frac{1}{2}\bar{\alpha}^{22}_{ci}\right)\right\} \quad (3)$$

where $\bar{a}^2_{ci}$, $\bar{\alpha}^{22}_{ci}$ and $\bar{\gamma}^{22}_{ci}$ are the means of the corresponding coefficients, considering all possible i reference individuals.

Parameter estimation

The parameters $\tau = \sum_m\sigma^2_m$ and $\tau_2 = \sum_m\sigma^4_m$ that appear in the first approximation, and the parameters $\sum_m\rho_m$, $\sum_m\rho^2_m$ and $\sum_m\frac{\rho^2_m}{\sigma^2_m}$ that appear in the second approximation, are

Table 1 Coefficients describing the genotypes' distributions moments when using the relation $\mathbf{E}\left[X^{d_i}_{im}X^{d_j}_{jm}\cdots X^{d_K}_{Km}\right] = p_m(1-p_m)\alpha^{d_id_j\cdots d_K}_{ij\cdots K} - \left[p_m(1-p_m)\right]^2\gamma^{d_id_j\cdots d_K}_{ij\cdots K}$ from Additional file 1

$\mathrm{E}[X_{im}]$	$\alpha^1_i = 0$ $\gamma^1_i = 0$
$\mathrm{E}[X^2_{im}]$	$\alpha^2_i = 2$ $\gamma^2_i = 0$
$\mathrm{E}[X^4_{im}]$	$\alpha^4_i = 2$ $\gamma^4_i = 0$
$\mathrm{E}[X_{im}X_{jm}]$	$\alpha^{11}_{ij} = 4\delta_1 + 2[\delta_2 + \delta_3 + \delta_4 + \delta_5 + \delta_9 + \delta_{12}] + \delta_{10} + \delta_{11} + \delta_{13} + \delta_{14} = 4a_{ij}$ $\gamma^{11}_{ij} = 0$
$\mathrm{E}[X_{im}X^3_{jm}]$	$\alpha^{13}_{ij} = 16\delta_1 + 2(\delta_2 + \delta_3) + 8(\delta_4 + \delta_5) + 2(\delta_9 + \delta_{12}) + \delta_{10} + \delta_{11} + \delta_{13} + \delta_{14}$ $\gamma^{13}_{ij} = 24\delta_1 + 12(\delta_4 + \delta_5)$
$\mathrm{E}[X^2_{im}X^2_{jm}]$	$\alpha^{22}_{ij} = 16\delta_1 + 4(\delta_2 + \delta_3 + \delta_4 + \delta_5) + 2(\delta_9 + \delta_{12}) + \delta_{10} + \delta_{11} + \delta_{13} + \delta_{14}$ $\gamma^{22}_{ij} = 48\delta_1 + 8(\delta_2 + \delta_3 + \delta_4 + \delta_5) - 4\delta_{15} - 16\delta_6 - 8(\delta_7 + \delta_8)$

δ_s are the 15 classical identity states probabilities between two individuals [33–35]

Coefficients of expectations $\mathrm{E}\left[X_iX_jX^2_k\right]$ and $\mathrm{E}[X_iX_jX_kX_l]$ involve IBD status between three or four different individuals and are explained in Additional file 1

Table 2 Moments of genotypes' distributions depending on genotype codification

Expectations	Genotype codification	
	$x_{tim} = a_{tim} - 2p_{tm}$	$x_{tim} = (a_{tim} - 2p_{tm})/\sigma_{tm}$
$E[X_{im}]$	0	0
$E[X_{im}^2]$	σ_m^2	1
$E[X_{lm}^4]$	σ_m^2	$1/\sigma_m^2$
$E[X_{im}X_{jm}]$	$2a_{ij}\sigma_m^2$	$2a_{ij}$
$E[X_{im}X_{jm}^3]$	$\frac{1}{2}\alpha_{ij}^{13}\sigma_m^2 - \frac{1}{4}\gamma_{ij}^{13}\sigma_m^4$	$\frac{1}{2}\alpha_{ij}^{13}/\sigma_m^2 - \frac{1}{4}\gamma_{ij}^{13}$
$E[X_{im}^2X_{jm}^2]$	$\frac{1}{2}\alpha_{ij}^{22}\sigma_m^2 - \frac{1}{4}\gamma_{ij}^{22}\sigma_m^4$	$\frac{1}{2}\alpha_{ij}^{22}/\sigma_m^2 - \frac{1}{4}\gamma_{ij}^{22}$

unknown. Their expectations can be derived by making assumptions about the distribution of the marker allele frequencies. They were derived assuming either a uniform distribution of allele frequencies or the U-shaped distribution of allelic frequencies proposed by Goddard [16]: $f(p) = k/2p(1-p)$ with the constant k estimated as $1/\log 2N_e$, N_e being the effective size of the reference population. The expectations of the parameters are in Table 3. The corresponding algebra is detailed in Additional file 2.

The parameters τ and τ_2 are linked to the number M_e of independent segments. This quantity M_e was defined by Goddard [16] as the number of independent chromosomal segments which would give the same variance of genomic covariances c_{ij} between individuals i and j as that observed, *i.e.* when LD exists. Conditional on the genotypic observation, the genomic covariance between two individuals is $\mathrm{cov}(q_i,q_j|\mathbf{X}) = \sigma_\beta^2 \sum_q X_{iq}X_{jq} = c_{ij}$. Thus, $v_X(c_{ij}) = \sigma_\beta^4 v[\sum_q X_{iq}X_{jq}]$, or $v_X(c_{ij}) = \sigma_\beta^4(\sum_q v(X_{iq}X_{jq}) + \sum_q\sum_{q'\neq q}\mathrm{cov}(X_{iq}X_{jq},X_{iq'}X_{jq'}))$. When the markers are in linkage equilibrium, the covariance term is null, and $v_X(c_{ij}) = \sigma_\beta^4\left[\frac{1}{2}\tau\alpha_{ij}^{22} - \frac{1}{4}\tau_2\gamma_{ij}^{22} - \frac{1}{4}a_{ij}^2\tau_2\right]$. If individuals are unrelated, $\alpha_{ij}^{22} = 0$, $\gamma_{ic}^{22} = -4$ and $a_{ij} = 0$. Thus, $v_X(c_{ij}) = \sigma_\beta^4\tau_2$. As $\sigma_q^2 = \sigma_\beta^2\tau$, $v_X(c_{ij}) = \sigma_q^4\frac{\tau_2}{\tau^2}$. From the appendix in the paper of Goddard [16], this variance is $v_X(c_{ij}) = \sigma_q^4/M_e$. Thus:

$$M_e = \tau^2/\tau_2. \tag{4}$$

It must be emphasized that M_e, which depends on the variability of allele frequencies, is not the number of markers n_M.

Table 3 Expectation of elements involved in precision formulae when a uniform ($f(p) = 1$) or a U shaped distribution of allelic frequencies is assumed ($f(p) = k/2p(1-p)$)

Element	Expectation	
	Uniform	U shaped
$E[\sigma_m^2]$	1/3	k
$E[\sigma_m^4]$	2/15	$k/3$
$E[\rho_m]$	$1 - 2\frac{h}{\omega}\theta$	$\frac{k}{\omega}\theta$
$E[\rho_m^2]$	$\left(\frac{4\theta}{\omega} + \frac{2}{h}\right)\left(\frac{1+h}{1+4h}\right)^2 - \frac{4\theta h}{\omega}$	$\frac{k}{\omega^2}\left[\theta\left(\omega - \frac{2h}{\omega}\right) - 1\right]$
$E[\rho_m^2/\sigma_m^2]$	$\frac{1}{\omega^2}\left[\theta\left(\omega - \frac{2h}{\omega}\right) - 1\right]$	$\frac{k}{2\omega^3}\left\{2\theta + \frac{\omega}{h}\right\}$

A large effective size N_e of the population was assumed to make $1/N_e$ negligible
$\theta = \log\left(\left|\frac{1+\omega}{1-\omega}\right|\right), \omega = \sqrt{1+4h}\, h = \lambda_\beta/2n_r, \lambda_\beta = \sigma_e^2/\sigma_\beta^2$

The case of unrelated individuals

The first approximation gives results similar to Goddard et al. [18] when individuals are not related. In this case, $\mathbf{A} = \mathbf{I}$ then $\mathbf{T} = \frac{1}{1+\gamma}\mathbf{I} = \frac{\sigma_e^2}{\sigma_q^2+\sigma_e^2}\mathbf{I} = \frac{\sigma_e^2}{\sigma^2}\mathbf{I}$. The ratio $\frac{\gamma}{1+\gamma} = \frac{\sigma_q^2}{\sigma^2} = \nu^2$ is the proportion of the phenotypic variance explained by the molecular score.

$$\begin{aligned}\mathrm{E}[\mathbf{\Lambda}_{cc}] &= \frac{1}{\sigma_e^2}\left\{\mathbf{T}_{cc} - \gamma^2\{\mathbf{TATAT}\}_{cc} + \frac{\gamma^2}{\tau^2}\left\{\mathbf{TE}\left[\mathbf{G}^*\mathbf{TG}^*\right]\mathbf{T}\right\}_{cc}\right\}\\ &= \frac{1}{\sigma_e^2}\left\{\frac{1}{1+\gamma} - \frac{\gamma^2}{(1+\gamma)^3} + \frac{\gamma^2}{\tau^2(1+\gamma)^2}\mathrm{E}\left[\mathbf{G}^*\mathbf{TG}^*\right]_{cc}\right\}\end{aligned}$$

$$\left\{\mathrm{E}\left[\mathbf{G}^*\mathbf{TG}^*\right]\right\}_{cc} = \sum_l\sum_k t_{kl}\left(\frac{1}{2}\tau\alpha_{cckl}^{1111} - \frac{1}{4}\tau_2\gamma_{cckl}^{1111} + 4a_{ck}a_{cl}\left[\tau^2 - \tau_2\right]\right)$$

$\mathbf{T}$ being diagonal, this equation simplifies to $\left\{\mathrm{E}\left[\mathbf{G}^*\mathbf{TG}^*\right]\right\}_{cc} = \sum_k t_{kk}\left(\frac{1}{2}\tau\alpha_{ck}^{22} - \frac{1}{4}\tau_2\gamma_{ck}^{22} + 4a_{ck}^2\left[\tau^2 - \tau_2\right]\right)$, with $t_{kk} = \frac{1}{1+\gamma}$, $\alpha_{ck}^{22} = 0$ and $\gamma_{ck}^{22} = -4$ if $c \neq k$, $\alpha_{cc}^{22} = \alpha_c^4 = 2$ and $\gamma_c^4 = 0$, $a_{cc} = a_{kk} = \frac{1}{2}$ and $a_{ck} = 0$. Hence $\left\{\mathrm{E}\left[\mathbf{G}^*\mathbf{TG}^*\right]\right\}_{cc} = \frac{1}{1+\gamma}\left\{\tau + \tau^2 - \tau_2 + n_r\tau_2\right\}$, and $\mathrm{E}[\mathbf{\Lambda}_{cc}] = \frac{1}{\sigma_e^2}\left\{\frac{1}{1+\gamma} + \frac{\gamma^2}{(1+\gamma)^3}\left(\frac{1}{\tau} + (n_r-1)\frac{\tau_2}{\tau^2}\right)\right\}$.

$$\begin{aligned}\mathrm{E}\left[v\left(\hat{q}_c\right)\right] &= \mathrm{E}\left[\mathbf{V}_{cc}^*\right] - \frac{1}{\mathrm{E}[\mathbf{\Lambda}_{cc}]}\\ &= \sigma^2 - \sigma^2\frac{1}{1+\nu^4\left(\frac{1}{\tau} + \frac{n_r-1}{M_e}\right)}\\ &= \sigma^2\frac{\nu^4\left(\frac{1}{\tau} + \frac{n_r-1}{M_e}\right)}{1+\nu^4\left(\frac{1}{\tau} + \frac{n_r-1}{M_e}\right)}.\end{aligned}$$

If we neglect $\frac{1}{\tau} - \frac{1}{M_e}$ and use $\nu^2 = \frac{\sigma_q^2}{\sigma^2}$, we get $\mathrm{E}\left[v\left(\hat{q}_c\right)\right] = \sigma_q^2\frac{\nu^2\frac{n_r}{M_e}}{1+\nu^4\frac{n_r}{M_e}}$, which is similar but not identical to the equation in Goddard et al. [18] ($\sigma_q^2\frac{\nu^2\frac{n_r}{M_e}}{1+\nu^2\frac{n_r}{M_e}}$). Finally, the precision is estimated as:

$$\tilde{\mathrm{E}}\left[r_{q_c,\hat{q}_c}^2\right] = \frac{\nu^2\frac{n_r}{M_e}}{1+\nu^4\frac{n_r}{M_e}}. \tag{5}$$

In this situation of unrelatedness between the candidate and the reference population, the second approximation simplifies to $\tilde{\tilde{E}}\left[r_{q_c,\hat{q}_c}^2\right] = 1 - \lambda_\beta\frac{E[\sum_m \rho_m]}{n_r\tau}$.

From Table 3, we have $E\left[\sum_m \rho_m\right] = n_M \frac{k}{\omega}\theta$ with $\theta = \log\left(\left|\frac{1+\omega}{1-\omega}\right|\right), \omega = \sqrt{1+4h}, \quad h = \lambda_\beta / 2n_r$ and $k = 1/\log 2N_e$. As $\lambda_\beta = \tau/\gamma$, we found:

$$\tilde{\tilde{E}}\left[r^2_{q_c,\hat{q}_c}\right] = 1 - \frac{n_M \mathrm{k}\theta}{\gamma n_r \omega}. \quad (6)$$

Non-independence between reference and candidate population, a simple example

We consider the situation of a candidate that is the son of one of the n_r individuals in $\mathrm{P_r}$ (say the first in the list) while still assuming that reference individuals are unrelated. In this situation, the pedigree relationship matrix is $\begin{pmatrix} 1 & 0.5 & \\ 0.5 & 1 & \mathbf{0} \\ & \mathbf{0} & \mathbf{I}_{\mathrm{n_r}-1} \end{pmatrix}$, which results in a $\mathbf{T}$ matrix $\begin{pmatrix} a & b & \\ b & a & \mathbf{0} \\ & \mathbf{0} & \frac{1}{1+\gamma}\mathbf{I}_{\mathrm{n_r}-1} \end{pmatrix}$ with $\gamma = \frac{\sigma^2_q}{\sigma^2_e}, a = \frac{1+\gamma}{(1+\gamma)^2 - 1/4\gamma^2}$ and $b = -\frac{\gamma/2}{(1+\gamma)^2 - 1/4\gamma}$. Applications of formulae (2) and (3) are described in Additional file 3. The expected approximate precision with the first approach is:

$$\tilde{\mathrm{E}}\left[r^2_{q_c,\hat{q}_c}\right] \sim \frac{1}{\nu^2} - \frac{1}{\gamma \mathrm{a} + \gamma^3 \frac{\tau - \tau_2}{\tau^2}\mathrm{c1} + \gamma^3 \frac{\tau_2}{\tau^2}\left[\mathrm{c2} + \frac{\mathrm{n_r}-1}{1+\gamma}\mathrm{c3}\right]}, \quad (7)$$

where $c1 = (\mathrm{a}+\mathrm{b})^3 + \left(\mathrm{a}^2 + \mathrm{b}^2\right)\frac{1}{2}\mathrm{a}$, $c2 = \frac{1}{4}\mathrm{a}\left(\mathrm{b}^2 - \mathrm{a}^2\right)$ and $c3 = \mathrm{a}^2 + \mathrm{b}^2 + \frac{1}{2}\mathrm{ab}$. And with the second approach:

$$\tilde{\tilde{E}}\left[r^2_{q_c,\hat{q}_c}\right] = 1 - \frac{\mathrm{n_M} k\theta}{\gamma \mathrm{n_r}\omega} - \frac{\mathrm{n_M} k}{4\gamma n_r^2 \omega^2}\left(5\theta\omega - \frac{(10h+2)\theta}{\omega} - 5 - \mathrm{n_M k}\theta^2 - \frac{1}{h}\right). \quad (8)$$

Alternative genotypes codification

In all the previous developments, genotypes were coded $x_{tim} = a_{tim} - 2p_{tm}$ and $w_{tiq} = a_{tiq} - 2p_{tq}$. Alternatively, we could define $x_{tim} = (a_{tim} - 2p_{tm})/\sigma_{tm}$ and $w_{tiq} = (a_{tiq} - 2p_{tq})/\sigma_{tq}$. The relation between genetic and marker variances becomes $\sigma^2_q = n_M \sigma^2_\beta$ and the relation between pedigree and genomic matrices becomes $\mathrm{E}[\mathbf{G}^*] = \mathbf{A} n_M$. Thus, formulae (1) and (2) are still valid when replacing τ by n_M. The $\mathrm{E}\left[X^{d_i}_{im} X^{d_j}_{jm} \cdots X^{d_K}_{Km}\right]$ elements derived in Additional file 1, need to be divided by $\sigma^{d_i+d_j+\cdots+d_K}_m$. Table 2 gives the expectations with this alternative codification of genotypes. The quantity $\{\mathrm{E}[\mathbf{XX}'\mathbf{TXX}']\}_{\mathrm{ij}}$ has to be changed, using $\zeta = \frac{1}{n_M}\sum_m \frac{1}{\sigma^2_m}$. We have: $\sum_{\mathrm{m}} \mathrm{E}\left[X_{im}X_{km}X_{jm}X_{lm}\right] = \frac{n_M}{2}\zeta\alpha^{1111}_{\mathrm{ijkl}} - \frac{n_M}{4}\gamma^{1111}_{\mathrm{ijkl}}$, $\sum_{\mathrm{m}} \mathrm{E}[X_{im}X_{km}] = 2n_M a_{ik}$, and $\sum_{\mathrm{m}}(\mathrm{E}[X_{im}X_{km}]\mathrm{E}[X_{jm}X_{lm}]) = 4n_M a_{ik}a_{jl}$. Thus:

$$\{\mathrm{E}[\mathbf{XX}'\mathbf{TXX}']\}_{\mathrm{ij}} = \sum_l \sum_k t_{kl}\left(\frac{n_M}{2}\zeta\alpha^{1111}_{ijkl} - \frac{n_M}{4}\gamma^{1111}_{ijkl} + 4n_M(n_M-1)a_{ik}a_{jl}\right).$$

When applied to the case of unrelated individuals and no LD, *i.e.* when $t_{kk} = \frac{1}{1+\gamma}$, $\alpha^{22}_{ck} = 0$ and $\gamma^{22}_{ck} = -4$ if $c \neq k$, $\alpha^{22}_{cc} = \alpha^4_c = 2$ and $\gamma^4_c = 0$, $\mathrm{a_{cc}} = \mathrm{a_{kk}} = \frac{1}{2}$ and $\mathrm{a_{ck}} = 0$, we have:

$$\mathrm{E}[\mathbf{\Lambda}_{\mathrm{cc}}] = \frac{1}{\sigma^2_e}\left\{\frac{1}{1+\gamma} - \frac{\gamma^2}{(1+\gamma)^3} + \frac{\gamma^2}{n^2_M(1+\gamma)^2} \cdot \sum_k \frac{1}{1+\gamma}\left(\frac{n_M}{2}\zeta\alpha^{1111}_{cckk} - \frac{n_M}{4}\gamma^{1111}_{cckk} + 4n_M(n_M-1)a_{ck}a_{ck}\right)\right\},$$

which gives:

$$\mathrm{E}[\mathbf{\Lambda}_{\mathrm{cc}}] = \frac{1}{\sigma^2_e(1+\gamma)}\left\{1 - \frac{\gamma^2}{(1+\gamma)^2}\left(1 - \frac{\zeta + n_R + n_M - 1}{n_M}\right)\right\}$$

i.e. $\mathrm{E}[\mathbf{\Lambda}_{\mathrm{cc}}] = \frac{1}{\sigma^2}\left\{1 + \nu^4\left(\frac{\zeta + n_R - 1}{n_M}\right)\right\}$ and $\tilde{\mathrm{E}}\left[r^2_{q_c,\hat{q}_c}\right] = \frac{\nu^2 \frac{\zeta + n_R - 1}{n_M}}{1 + \nu^4 \frac{\zeta + n_R - 1}{n_M}}$.

Based on Additional file 2, the expectation of ζ parameter is $\frac{k}{4}\left[2\log(N_e - 1) + 2\frac{N_e(N_e-2)}{N_e - 1}\right]$ for a U-shaped distribution of alleles frequencies and $\log(N_e - 1)$ for a uniform distribution.

The case of markers in linkage disequilibrium

So far, following Goddard [16], we considered the situation of n_M independent segments that each carries a single QTL in LD with a single marker. More typically, the genomic information consists of a large number of non-independent markers. This non-independence comes from long-term effects due to bottlenecks, mutations, migrations, etc. and short-term effects due to family structure.

Effective and equivalent numbers of independent loci

We based our developments on the very fruitful concept of the effective number of loci that Goddard defined as "the number of independent loci that gives the same variance of realized relationships as that obtained in the more realistic situation" (Goddard [16] appendix). Since our objective was to predict the reliability of GEBV, we now suggest the alternative definition of an "equivalent number of independent loci" which would give the reliability of GEBV for unrelated individuals

when considering a sub-set of independent markers that would be identical to the reliability obtained when considering the full set of markers. From the derivation of the reliability given previously, defining $\mathbf{x}_\mathbf{c}^\mathbf{i}$ and $\mathbf{X}_\mathbf{r}^\mathbf{i}$ as the genotype matrices of the independent loci, we need $\mathrm{E}_{\mathbf{x_c},\mathbf{X_r}}\left[v\left(\hat{q}_c|\mathbf{x_c},\mathbf{X_r}\right)\right] = \mathrm{E}_{\mathbf{x}_\mathbf{c}^\mathbf{i},\mathbf{X}_\mathbf{r}^\mathbf{i}}\left[v\left(\hat{q}_c|\mathbf{x}_\mathbf{c}^\mathbf{i},\mathbf{X}_\mathbf{r}^\mathbf{i}\right)\right]$. With a few simplifying assumptions (identical distribution of genotypes in the reference and candidate populations and equal genotypic variance at all loci) a simple formula can be derived [see Additional file 4]:

$$n_{M_i} = n_M \frac{1+\gamma}{\gamma}\left(1 - tr\left[\left(\mathrm{E}\left[\mathbf{X}_\mathbf{r}'\mathbf{X}_\mathbf{r}\right] + \lambda_\beta \mathbf{I}\right)^{-1}\right]\frac{\sum_m \sigma_m^2/n_M}{\gamma}\right), \tag{9}$$

where tr[M] is the trace of matrix M.

Once marker allele frequencies and between-marker LD are estimated in a population of interest, the equivalent number of independent loci which can be estimated from formula (9) and this parameter can be used in models that predict the genetic gain expected from a genomic selection scheme applied to this population.

In the more general situation, prior to the observation of the $\mathbf{X}_\mathbf{r}$ matrix, a simple approximation for n_{M_i} is obtained assuming equal variances $\sigma_m^2 = s^2$, and using the relation between expected LD and effective population size N_e as derived by Sved [32]: $\mathrm{E}[2\Delta_{ml}] = \sigma_m\sigma_l \Big/ \sqrt{1+4N_e d_{lm}}$ with d_{lm} the distance between ordered loci l and m, such that $d_{lm} = |l-m|L/n_M$ with L the genome length in Morgan. With those hypotheses, let $\mathrm{U} = tr[(\gamma n_R \mathbf{R} + n_M \mathbf{I})^{-1}]$ with $\mathbf{R}_{\mathrm{ml}} = \sqrt{n_M/\left(n_M + 4N_e|l-m|L\right)}$.

In this simplified situation, the equivalent number of loci is [See Additional file 4]:

$$n_{M_i} = n_M \frac{n_R\gamma(1-U)}{n_R\gamma - n_M(1-U)}. \tag{10}$$

Towards an exact treatment of linkage disequilibrium

For a complete treatment of the LD situation, it is necessary to estimate the expectations of the product of four genetic values. For instance, with the second approximation [formula (2)], we need to compute $\mathrm{E}[x_{cm}X_{rim}x_{cl}X_{ril}]$. Let $X_{im} = g_{imf} + g_{imd}$, where g_{imf} and g_{imd} are the "values" of the alleles transmitted to individual i by its father and its dam, with g_{imf} and $g_{imd} = (0\ or\ 1) - p_m$. They will be called allelic values in the following. Equivalent terms are defined for x_{cl},x_{cm} and X_{il}. The random variable M_{cls} is the allele of individual c received from s at locus $l\left(f \text{ or } d\right)$. M_{cmt}, M_{ilu} and M_{imv} are defined similarly.

$$\mathrm{E}[x_{cl}x_{cm}X_{il}X_{im}] = \sum_{s\in\{f,d\}}\sum_{t\in\{f,d\}}\sum_{u\in\{f,d\}}\sum_{v\in\{f,d\}}\mathrm{E}[g_{cls}g_{cmt}g_{ilu}g_{imv}]\cdot \tag{11}$$

For the candidate c as for the reference individual i, the pair of genetic values may originate from the same parent (and coded on the same chromosome) or not, giving four types of $\left(g_{cls}, g_{cmt}, g_{ilu}, g_{imv}\right)$ vectors. In type 1 ($s = t\ and\ u = v$), both alleles (belonging to loci m and l) of each pair of loci (one for c and one for i) are on the same chromosome (may be from the two fathers, the two dams, c's father and i's dam or i's father and c's dam). In type 2 ($s = t\ and\ u \neq v$), both alleles (belonging to loci m and l) of the candidate are on the same chromosome, while alleles of the reference individual i are not on the same chromosome.Type 3 ($s \neq t\ and\ u = v$) is the reverse from type 2. In type 4 ($s \neq t\ and\ u \neq v$), alleles of loci m and l of both individuals and i are on different chromosomes.

For each of these situations, the identity by descent (IBD) status between alleles at locus m on chromosomes ct and iv, and at locus l on chromosomes cs and iu are considered. There are four, as follows:

$$\mathcal{S}_{ml} = \{\, M_{cmt} \equiv M_{imv}\ and\ M_{cls} \equiv M_{ilu}\} \text{ with a probability } \varphi_{ml}^{stuv},$$

$$\mathcal{S}_{m\not l} = \{\, M_{cmt} \equiv M_{imv}\ and\ M_{cls} \not\equiv M_{ilu}\} \text{ with a probability } \varphi_{m\not l}^{stuv},$$

$$\mathcal{S}_{\not m l} = \{\, M_{cmt} \not\equiv M_{imv}\ and\ M_{cls} \equiv M_{ilu}\} \text{ with a probability } \varphi_{\not m l}^{stuv},$$

$$\mathcal{S}_{\not m\not l} = \{\, M_{cmt} \not\equiv M_{imv}\ and\ M_{cls} \not\equiv M_{ilu}\} \text{ with a probability } \varphi_{\not m\not l}^{stuv}.$$

Thus, 16 terms involved in $E[x_{cl}x_{cm}X_{il}X_{im}]$ are given by:

$$E[g_{cls}g_{cmt}g_{ilu}g_{imv}] = \sum_{k\in\{ml,m\cancel{l},\cancel{m}l,\cancel{ml}\}} \varphi_k^{stuv} E[g_{cls}g_{cmt}g_{ilu}g_{imv}|\mathcal{S}_k]. \quad (12)$$

As described in Additional file 5, only seven $E[g_{cls}g_{cmt}g_{ilu}g_{imv}|\mathcal{S}_k]$ are non-null (Table 4). Principles on which the probabilities φ_k^{stuv} are estimated and basic examples are described in Additional file 5.

As an illustration, we consider again the situation of a candidate (c), that is the son of one of the n_r individuals in P_r and assume that c's dam is unrelated to the sire. In formula (2), the summation over the reference individuals i comprises a single term for the sire of the candidate and $n_r - 1$ terms for the individual that are unrelated to the c members of this reference population.

Based on Additional files 1 and 5, expectations involved in the precision formulae (2) are:

$E[x_{cm}^2 X_{rim}^2] = p_m(1-p_m)$, and

$$E[x_{cl}x_{cm}X_{il}X_{im}] = (1-p_m)(1-p_l)p_m p_l + \Delta_{lm}(1-2p_l)(1-2p_m) + 2\Delta_{lm}^2\left[r_{ml}\left(\frac{p_m^3+(1-p_m)^3}{p_m(1-p_m)} + \frac{p_l^3+(1-p_l)^3}{p_l(1-p_l)}\right) + (1-r_{ml})(1-2p_m)(1-2p_l)\right],$$

when i is the sire of c; and $E[x_{cm}^2 X_{rim}^2] = 4[p_m(1-p_m)]^2$ and $E[x_{cl}x_{cm}X_{il}X_{im}] = 4\Delta_{lm}^2(1-2p_m)(1-2p_l)$, when i and c are unrelated.

Numerical evaluation

Simulation of allele frequencies

In the following numerical evaluation of the formulae derived above, allele frequencies were simulated following an inverse transform sampling (e.g. [32]): n_M allele frequency cumulative distribution function values u_m were simulated in a uniform $\mathcal{U}(0,1)$, and corresponding allele frequencies p_m, *i.e.* such as $u_m = \int_{1/2n_r}^{p_m} f(p)dp$, computed by $p_m = \frac{(2n_r-1)^{(2u_m-1)}}{1+(2n_r-1)^{(2u_m-1)}}$.

Table 4 Expectations of products of four allelic values received by two individuals at two loci depending on the IBD status and parental origins of the alleles

$\mathcal{S}$	$\mathcal{T}$	$E[g_{cls}g_{cmt}g_{ilu}g_{imv}\|\mathcal{S}\ \&\ \mathcal{T}]$
$\mathcal{S}_{ml}$	$s=t$ and $u=v$	$(1-p_m)(1-p_l)p_m p_l + \Delta_{lm}(1-2p_l)(1-2p_m)$
	$s=t$ and $u\neq v$	$(1-p_m)(1-p_l)p_m p_l + \Delta_{lm}(1-2p_l)(1-2p_m)$
	$s\neq t$ and $u=v$	$(1-p_m)(1-p_l)p_m p_l + \Delta_{lm}(1-2p_l)(1-2p_m)$
	$s\neq t$ and $u\neq v$	$(1-p_m)(1-p_l)p_m p_l$
$\mathcal{S}_{m\cancel{l}}$	$s=t$ and $u=v$	$\Delta_{lm}^2 \times [p_m^3+(1-p_m)^3]/[p_m(1-p_m)]$
$\mathcal{S}_{\cancel{m}l}$	$s=t$ and $u=v$	$\Delta_{lm}^2 \times [p_l^3+(1-p_l)^3]/[p_l(1-p_l)]$
$\mathcal{S}_{\cancel{ml}}$	$s=t$ and $u=v$	$\Delta_{lm}^2 \times (1-2p_m)(1-2p_l)$

Only non-null terms are given

p_m and p_l are the frequencies of the most frequent alleles at loci m and l. Δ_{lm} is the linkage disequilibrium measure between m and l

$g_{cls} = (0 \text{ or } 1) - p_l$ is the allelic value the candidate received from its parent s at locus l etc

$\mathcal{S}_{ml}$ means c and i genes are IBD at m and l, $\mathcal{S}_{m\cancel{l}}$ only at m etc

Basic situation: no LD and unrelated individuals

Convergence of Taylor series and quality of expectation of the reliability approximations were tested for different population sizes ($n_r = 500, 1000, 1500$ and 2500), numbers of markers ($n_M = 50, 100, 250, 1000, 1500, 2000$ and 2500) and proportions of the phenotypic variance explained by the molecular score ($\nu^2 = 0.1, 0.4$ and 0.7). Given the set of allele frequencies $p_m (m = 1 \ldots n_M)$, genotypes $\mathbf{X}$ of $n_r + 1$ individuals were generated and the $\mathbf{G}$ matrix was built. The reliability of the candidate individual GEBV, $r^2 = \frac{v(\hat{q}_c|\mathbf{X})}{v(q_c|\mathbf{X})}$ was computed as described in the section «Situation analyzed» , as well as approximations considering 1–10 elements in the Taylor series $\mathbf{I} - \mathbf{D}\gamma + \mathbf{D}^2\gamma^2 - \mathbf{D}^3\gamma^3 \cdots$ The convergence of the series as predicted by the value (lower or higher than 1) of the matrix's largest eigenvalue was checked numerically, by estimating the mean of this largest eigenvalue from five simulations in each case studied ($n_r = 200$ to 1000; $n_M = 100$ to 2000 and $\nu^2 = 0.1, 0.4, 0.7$).

This limited number of replications was chosen after observation of a very limited variance of this eigenvalue. Finally, the asymptotic values of the suggested approximations [formulae (5) and (6)] were computed using the number of independent segments as described by [4]. The process was repeated 50 times and the means of those exact or approximated reliabilities computed.

Figure 1a and b illustrates the convergence of the Taylor series when 2000 markers are used, and Tables 5 and 6 give the results for both approximations when $\nu^2 = 0.4$. The Taylor series converged when the proportion ν^2 of the phenotypic variance explained by the molecular score was low, with oscillations and divergence observed when $\nu^2 = 0.4$ or 0.7 with the first approximation and $\nu^2 = 0.7$ with the second approximation. These observations were in accordance with the deviation to one of the largest eigenvalue of the matrix involved in the series (Fig. 2a, b). When the series converged, the approximations rapidly reached a plateau, at the 3rd (respectively, 2nd) order for the first (respectively, second) approximation.

Table 6 shows that the second Taylor series converges always when $\nu^2 = 0.4$. The proposed approximation was generally biased upwards. This over-estimation of the precision was generally limited but increased as the number of markers and the reference population size decreased. The maximum over-estimation observed was

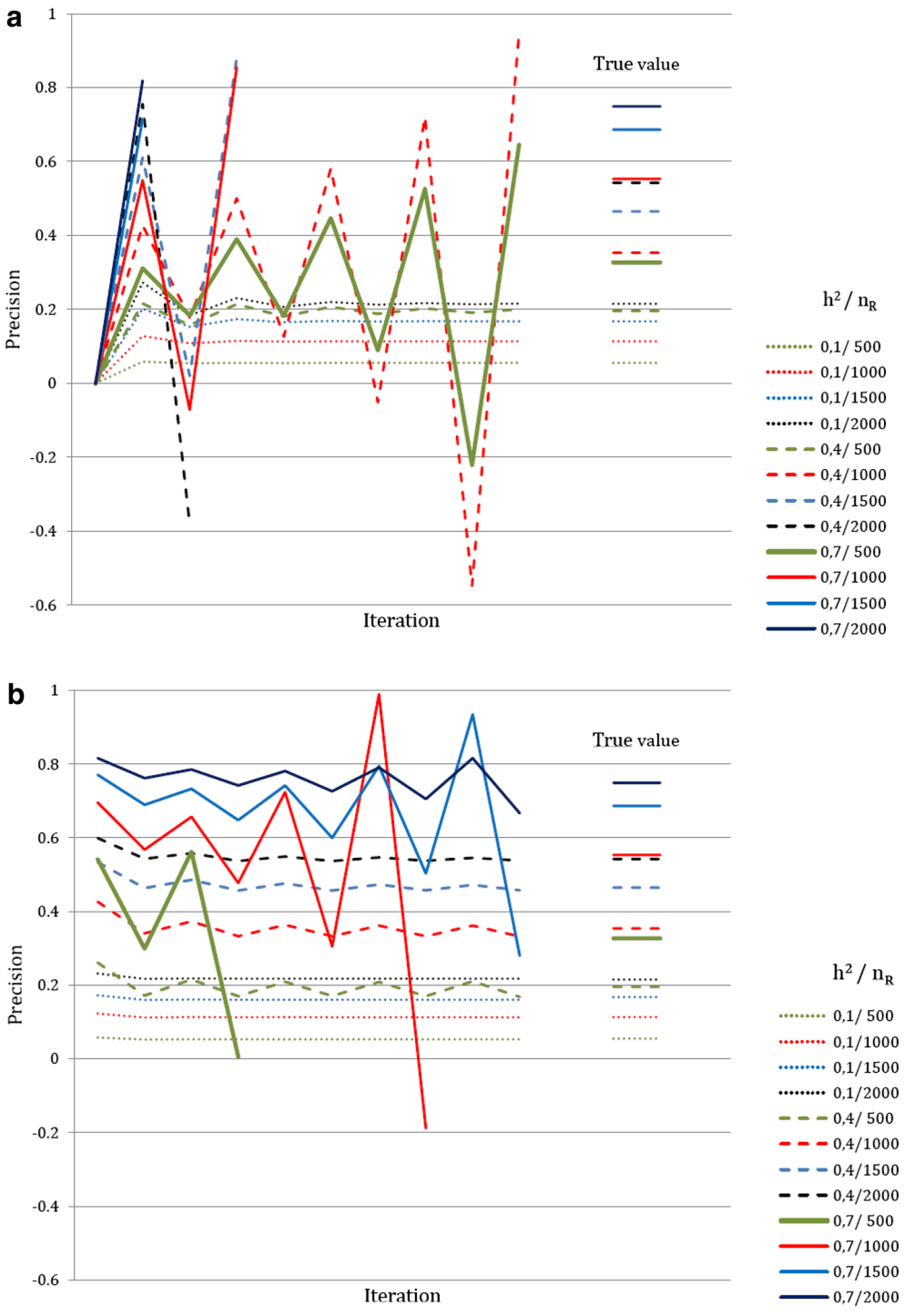

Fig. 1 Convergence of the Taylor series as a function of heritability and reference population size ($n_M = 2000$). **a** First approximation. **b** Second approximation

37.5 % (0.22 instead of 0.16 with a standard error less than 0.02). Based on the results in Table 5, it appears that when the first Taylor series converges, the proposed approximation is also slightly over-estimated. The expectation of the approximations, as given in formulae (5) and (6) are very close to the observations.

Table 5 Performances of the first approximation $\left(\tilde{r}^2_{q_c,\hat{q}_c}\right)$ for an unrelated reference population as a function of the number of markers (n_M) and reference population size (n_R), assuming $v^2 = 0.4$

n_M	n_R	True value $\left(r^2_{q_c,\hat{q}_c}\right)$	Approximation $\left(\hat{r}^2_{q_c,\hat{q}_c}\right)$	Convergence criteria	$E\left[\hat{r}^2_{q_c,\hat{q}_c}\right]$
50	500	0.92	2.14	2.74	2.05
50	1000	0.96	2.30	2.62	2.23
50	1500	0.96	2.31	2.57	2.28
50	2000	0.97	2.43	2.60	2.37
100	500	0.85	1.71	2.58	1.69
100	1000	0.90	2.01	2.49	2.05
100	1500	0.95	2.21	2.58	2.18
100	2000	0.96	2.31	2.60	2.26
250	500	0.67	1.09	2.53	1.09
250	1000	0.81	1.56	2.53	1.55
250	1500	0.85	1.72	2.54	1.72
250	2000	0.89	1.96	2.53	1.97
1000	500	0.32	0.39	0.61	0.38
1000	1000	0.52	0.72	2.45	0.73
1000	1500	0.64	0.99	2.51	0.99
1000	2000	0.71	1.18	2.51	1.18
1500	500	0.25	0.28	0.27	0.28
1500	1000	0.42	0.54	1.88	0.54
1500	1500	0.54	0.76	2.49	0.76
1500	2000	0.61	0.91	2.50	0.91
2000	500	0.20	0.22	0.20	0.22
2000	1000	0.35	0.43	0.95	0.43
2000	1500	0.46	0.61	2.28	0.61
2000	2000	0.54	0.76	2.48	0.76
2500	500	0.16	0.17	0.16	0.17
2500	1000	0.30	0.35	0.44	0.35
2500	1500	0.40	0.50	1.61	0.50
2500	2000	0.49	0.65	2.40	0.66

The convergence criterion is the value of the Taylor series at order 10

$E\left[\hat{r}^2_{q_c,\hat{q}_c}\right]$ is the expectation of the first approximation across the distribution of allele frequencies as given in Goddard [16]

No LD and non-independence between reference and candidate population

The quality of the approximation was tested as above, by considering the case of a candidate having one of its parents in the reference population and all other individuals being unrelated. Tables 7 and 8, which summarize the results of the simulation, show that the second approximation is still the most efficient (systematic convergence of the Taylor series and consistency between first order approximation and its expectation). Again, an overestimation of about 20 % is observed with this approximation.

Example of the use of the second approach

As an illustration of formula (3) different situations that differ in the relationships between the candidate and reference populations were compared. Coefficients of formula (3) were estimated using the elements in Table 3. An effective reference population size of 200, the genotyping of 10,000 markers and a heritability of 0.4 were assumed. Scenarios included no individuals related to the candidate in the reference population, its sire, both parents, 1–10 half-sibs (or uncles), and a combination of parental and half-sib information.

The results are in Fig. 3. The linearity of the precision increases with the number of half-sibs, which is consistent with the approximation, but unsatisfactory, as discussed below.

Equivalent number of independent loci

This number was computed using formula (8), for various effective population sizes ($N_e = 100$ to 1000), heritabilities ($h^2 = 0.1$ to 0.5), total numbers of loci

Table 6 Performances of the second approximation $\left(\tilde{r}^2_{q_c,\hat{q}_c}\right)$ for an unrelated reference population as a function of the number of markers (n_M) and reference population size (n_R), assuming $v^2 = 0.4$

n_M	n_R	True value $\left(r^2_{q_c,\hat{q}_c}\right)$	Approximation $\left(\tilde{r}^2_{q_c,\hat{q}_c}\right)$	10th order approximation	$E\left[\tilde{r}^2_{q_c,\hat{q}_c}\right]$
50	500	0.92	0.91	0.91	0.91
50	1000	0.96	0.95	0.94	0.95
50	1500	0.96	0.97	0.97	0.96
50	2000	0.97	0.97	0.97	0.97
100	500	0.85	0.83	0.82	0.83
100	1000	0.90	0.91	0.90	0.91
100	1500	0.95	0.94	0.94	0.94
100	2000	0.96	0.95	0.95	0.95
250	500	0.67	0.71	0.67	0.71
250	1000	0.81	0.83	0.81	0.82
250	1500	0.85	0.88	0.87	0.88
250	2000	0.89	0.90	0.90	0.90
1000	500	0.32	0.41	0.31	0.40
1000	1000	0.52	0.59	0.52	0.57
1000	1500	0.62	0.68	0.64	0.67
1000	2000	0.69	0.73	0.70	0.73
1500	500	0.24	0.32	0.23	0.31
1500	1000	0.42	0.50	0.42	0.48
1500	1500	0.52	0.60	0.53	0.59
1500	2000	0.60	0.66	0.61	0.67
2000	500	0.19	0.26	0.17	0.28
2000	1000	0.34	0.43	0.33	0.44
2000	1500	0.46	0.53	0.46	0.53
2000	2000	0.53	0.60	0.54	0.61
2500	500	0.16	0.22	0.14	0.23
2500	1000	0.30	0.38	0.28	0.38
2500	1500	0.40	0.48	0.39	0.47
2500	2000	0.47	0.55	0.48	0.56

The convergence criterion is the value of the Taylor series at order 10

$E\left[\tilde{r}^2_{q_c,\hat{q}_c}\right]$ is the expectation of the second approximation across the distribution of allele frequencies as given in Goddard [16]

($n_M = 1000$ to $10,000$) and reference population sizes ($n_R = 1000$ to 2500).

Figure 4 shows how equivalent numbers of independent loci (n_{M_i}) vary with the total number of markers (n_M) and reference population size (n_R). As n_M increases, the number n_{M_i} rapidly converges to a value which strongly depends on the size of the reference population size. This dependence on n_R of the equivalent number of independent loci does not exist in the Goddard's effective number of loci and clearly shows the difference in nature between these concepts. Three phenomena, observed when considering the extreme case of two markers (see Additional file 5), explain this behavior:

(1) The trace T of $(E[\mathbf{X}_\mathbf{r}'\mathbf{X}_\mathbf{r}] + \lambda_\beta \mathbf{I})^{-1}$ is a decreasing function of n_r: as a consequence, the larger is the population size, the smaller is T, which is proportional to the marker effects conditional variances $v(\boldsymbol{\beta}) - cov(\boldsymbol{\beta}, \mathbf{y})v(\mathbf{y})^{-1}cov(\mathbf{y}, \boldsymbol{\beta}))$ and the higher is the variance of the estimated molecular score $(v(q_c|\mathbf{y}) = \mathbf{x_c}cov(\boldsymbol{\beta}, \mathbf{y})v(\mathbf{y})^{-1}cov(\mathbf{y}, \boldsymbol{\beta})\mathbf{x_c}')$.

(2) The trace T is always higher in the situation of LD than for independent markers ($T_{LD} > T_{LE}$).

(3) The rate of decrease is higher for T_{LD} than for T_{LE}. On the whole, the reliability for a given number of observed markers corresponds to the reliability that is reached with a larger number of independent loci when the size of the reference population is larger.

Figure 5 indicates that the equivalent number of independent loci increases with heritability and effective population size. This last observation was expected since with larger effective population sizes, the LD between two loci decreases and this increases the effective number of loci. The effect of heritability is less direct.

Discussion

The objective of this paper was to explore approximations of the precision of genomic selection when the selection candidate has relatives in the reference population. Two approximations were developed and numerically compared.

These approximations were based on Taylor expansions of a matrix inverse $\mathbf{M}^{-1}$. In both cases, the initial matrix is the sum of the identity matrix and a perturbation ($\mathbf{M} = \mathbf{I} + \mathbf{E}$). Convergence of these series is not guaranteed and depends on the behavior of the perturbation ($\mathbf{I} - \mathbf{E} + \mathbf{E}^2 - \mathbf{E}^3 \rightarrow (\mathbf{I} + \mathbf{E})^{-1}$ if $\mathbf{E}^t \rightarrow 0$ when $t \rightarrow \infty$). With the first approximation, derived from the appendix in [18], this convergence failed when the number of makers was too small (less than 1500 in our example) or the heritability was greater than 0.1. This was only observed when $v^2 = 0.7$ with the second approximation. This is fully consistent with the deviation to one of the largest eigenvalue of the **E** matrix.

The expectation of the proposed approximation, when data were simulated with the model corresponding to the hypotheses underlying their algebraic development, was very close to the mean value after 50 simulations. Thus, extremely fast estimation of the precision is possible, which allows intensive optimization and comparison of selection schemes.

When individuals are unrelated and markers are in linkage equilibrium, we obtain an estimation of the GEBV accuracy which differs from that of Goddard et al. [18].

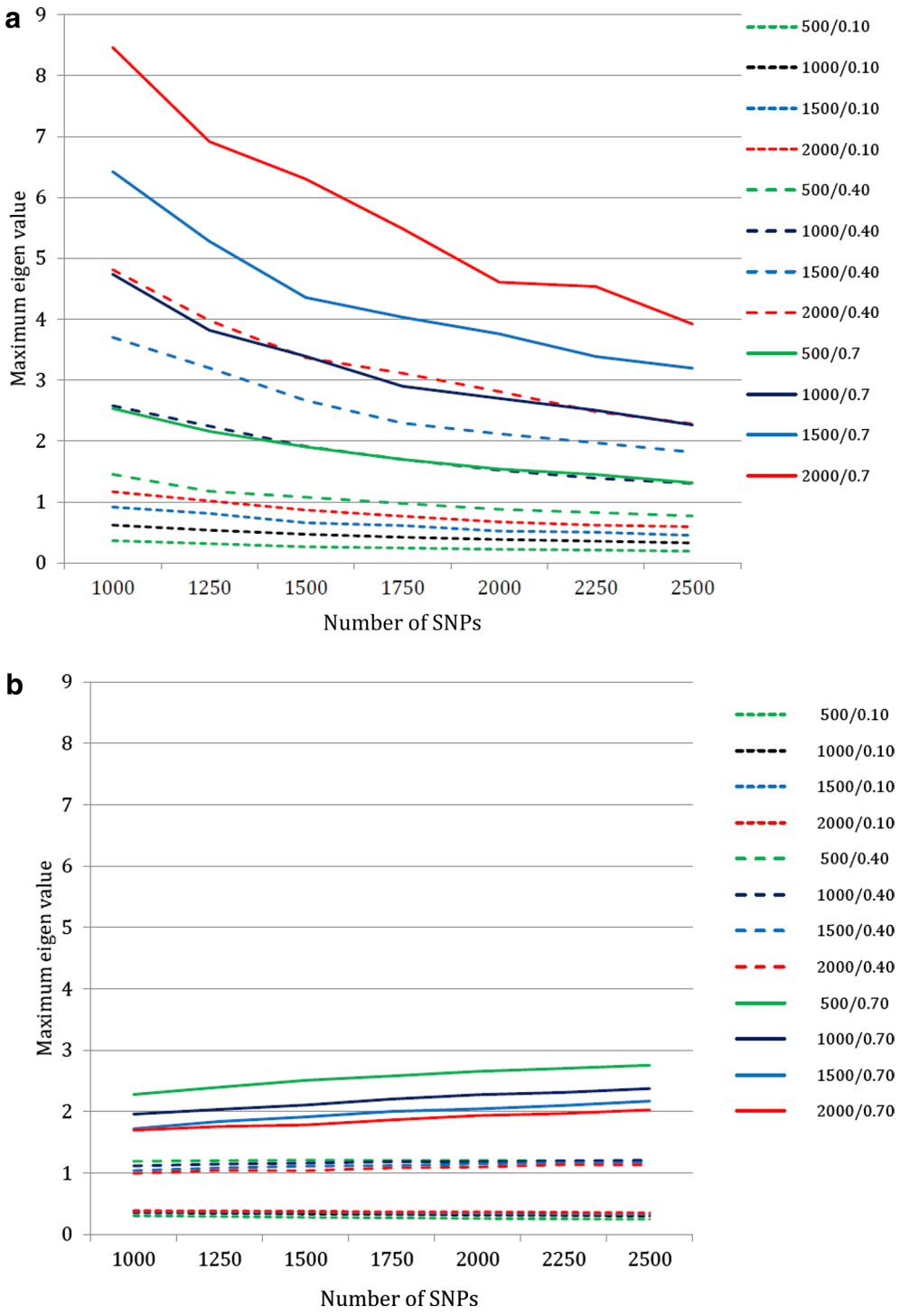

Fig. 2 Largest Eigen value of the noise matrix $\mathbf{D}\boldsymbol{\gamma}$ involved in the Taylor expansion of the phenotypic variances matrix **V** as a function of heritability, reference population size and number of markers. **a** First approximation. **b** Second approximation

This is surprising since that approach was said to be based on the Taylor approximation used here. Their formula may be obtained in a simpler way [see Additional file 6]. However, relaxing the assumption of "absence of between-individual relationship" is not straightforward using this approach.

Table 7 Performances of the first approximation $\left(\hat{r}^2_{q_c,\hat{q}_c}\right)$ when the parents of candidate belong to the reference population as a function of the number of markers (n_M) and reference population size (n_R), assuming $v^2 = 0.4$

n_M	n_R	True value $\left(r^2_{q_c,\hat{q}_c}\right)$	Approximation $\left(\hat{r}^2_{q_c,\hat{q}_c}\right)$	10th order approximation	$E\left[\hat{r}^2_{q_c,\hat{q}_c}\right]$
1000	500	0.37	0.42	0.58	0.47
1000	1000	0.56	0.73	2.47	0.82
1000	1500	0.65	0.95	2.50	1.04
1000	2000	0.72	1.17	2.52	1.26
1500	500	0.31	0.34	0.33	0.37
1500	1000	0.46	0.56	1.87	0.63
1500	1500	0.56	0.73	2.44	0.81
1500	2000	0.62	0.87	2.50	0.96
2000	500	0.27	0.29	0.27	0.32
2000	1000	0.40	0.46	0.89	0.52
2000	1500	0.50	0.62	2.24	0.69
2000	2000	0.57	0.76	2.48	0.84
2500	500	0.24	0.25	0.24	0.27
2500	1000	0.36	0.40	0.49	0.45
2500	1500	0.46	0.55	1.79	0.61
2500	2000	0.52	0.67	2.35	0.74

The convergence criterion is the value of the Taylor series at order 10

$E\left[\hat{r}^2_{q_c,\hat{q}_c}\right]$ is the expectation of the first approximation across the distribution of allele frequencies as given in Goddard [16]

Table 8 Performances of the second approximation $\left(\tilde{r}^2_{q_c,\hat{q}_c}\right)$ when the parents of the candidates belong to the reference population as a function of the number of markers (n_M) and reference population size (n_R), assuming $v^2 = 0.4$

n_M	n_R	True value $\left(r^2_{q_c,\hat{q}_c}\right)$	Approximation $\left(\tilde{r}^2_{q_c,\hat{q}_c}\right)$	10th order approximation	$E\left[\tilde{r}^2_{q_c,\hat{q}_c}\right]$
1000	500	0.37	0.46	0.35	0.46
1000	1000	0.53	0.60	0.54	0.61
1000	1500	0.64	0.70	0.65	0.69
1000	2000	0.71	0.75	0.72	0.75
1500	500	0.30	0.39	0.26	0.40
1500	1000	0.47	0.55	0.46	0.51
1500	1500	0.56	0.63	0.56	0.61
1500	2000	0.63	0.69	0.64	0.68
2000	500	0.27	0.36	0.22	0.35
2000	1000	0.40	0.49	0.38	0.48
2000	1500	0.50	0.58	0.50	0.56
2000	2000	0.57	0.64	0.57	0.62
2500	500	0.24	0.33	0.20	0.32
2500	1000	0.34	0.44	0.31	0.45
2500	1500	0.44	0.53	0.43	0.53
2500	2000	0.37	0.46	0.35	0.46

The convergence criterion is the value of the Taylor series at order 10

$E\left[\tilde{r}^2_{q_c,\hat{q}_c}\right]$ is the expectation of the second approximation across the distribution of allele frequencies as given in Goddard [16]

A strong limit of our new approximation comes from the limitation to the first order term of the Taylor series. Deriving algebra was only possible at this stage. The side effect is that no genotypic covariance terms between reference individuals appear in this approximation. As a consequence, only the direct relationships between candidate and reference individuals play a role in the estimation, but not the structure within the reference population. This is unfortunate, because accuracies of genomic prediction are obviously affected by the construction of the reference population. Our last numerical example, in which there is a linear trend with the number of half-sibs, reveals this drawback: two half-sibs of the candidates are treated as unrelated and the information that they carry is just the double of that of a single half-sib. Future developments should focus on this limitation, for instance to derive the expectation of the $\mathbf{x_c C^{-1} B^2 x_c'}$ term.

The U-shaped density function $f(p)$ of allele frequencies was defined as in [16]. A Beta distribution $\mathcal{B}(\phi_a, \phi_b)$ for the allele frequencies was assumed by Gianola et al. [30], following Wright [34]. Assuming that the frequency distribution is centered on 0.5, *i.e.* $\Phi_a = \Phi_b = \Phi$, this quantity Φ can be adjusted to fit the distribution of Goddard. Using the Chi^2 test as a fitting option, we observed that the adjusted $\hat{\phi}$ rapidly decreased as the population size increased (Fig. 6), with a slower and slower evolution as the population size grew larger (with $n_r = 200,000$ the adjusted $\hat{\phi}$ is 0.9750000). Using a Beta distribution could give more generality to the results. If the expectation of τ and τ_2 are easily derived from the moments generating function of Beta distribution ($E[\tau] = \frac{n_r a}{2a+1}$

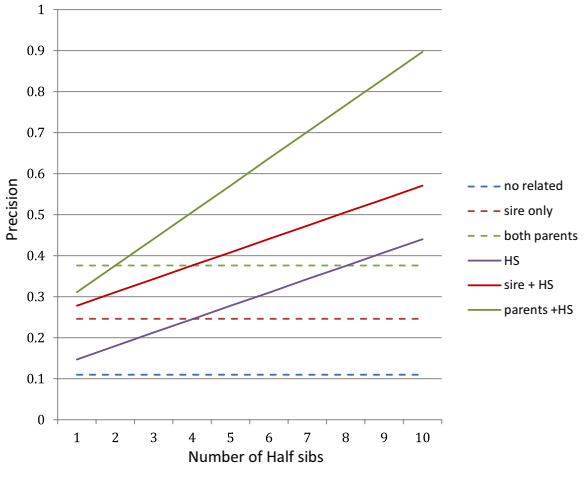

Fig. 3 Example of approximated precision [from Eq. (3)] corresponding to various relations between the candidate and reference populations. ($n_R = 1000$; $n_M = 10,000$; $v^2 = 0.4$)

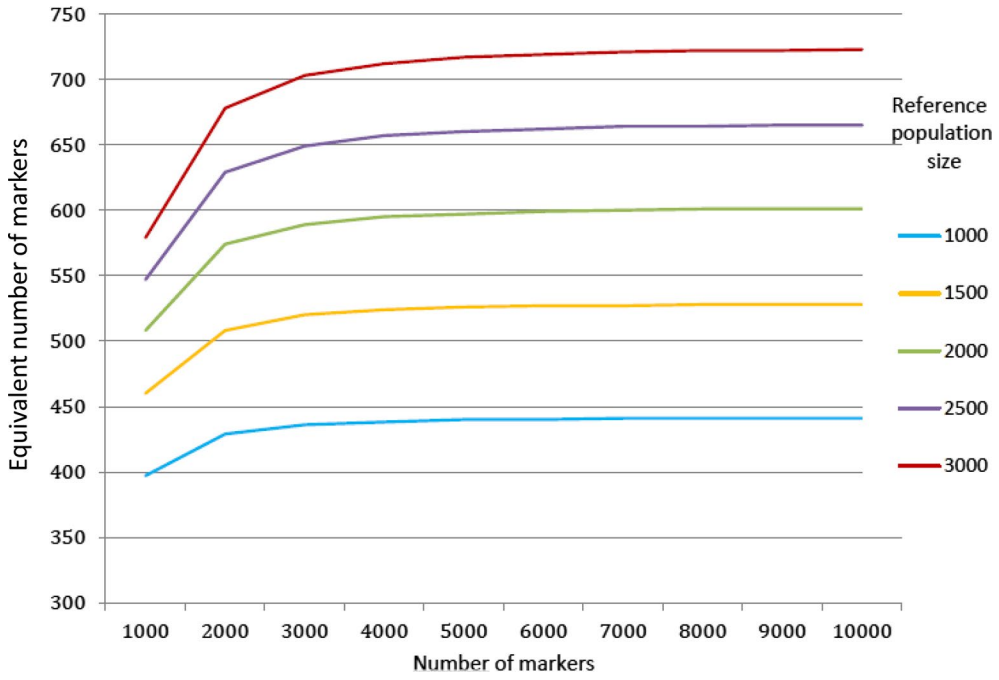

Fig. 4 Number of equivalent markers [from Eq. (8)] as a function of the total number of markers (n_M) and reference population size (n_R). ($Ne = 200$; $\nu^2 = 0.4$)

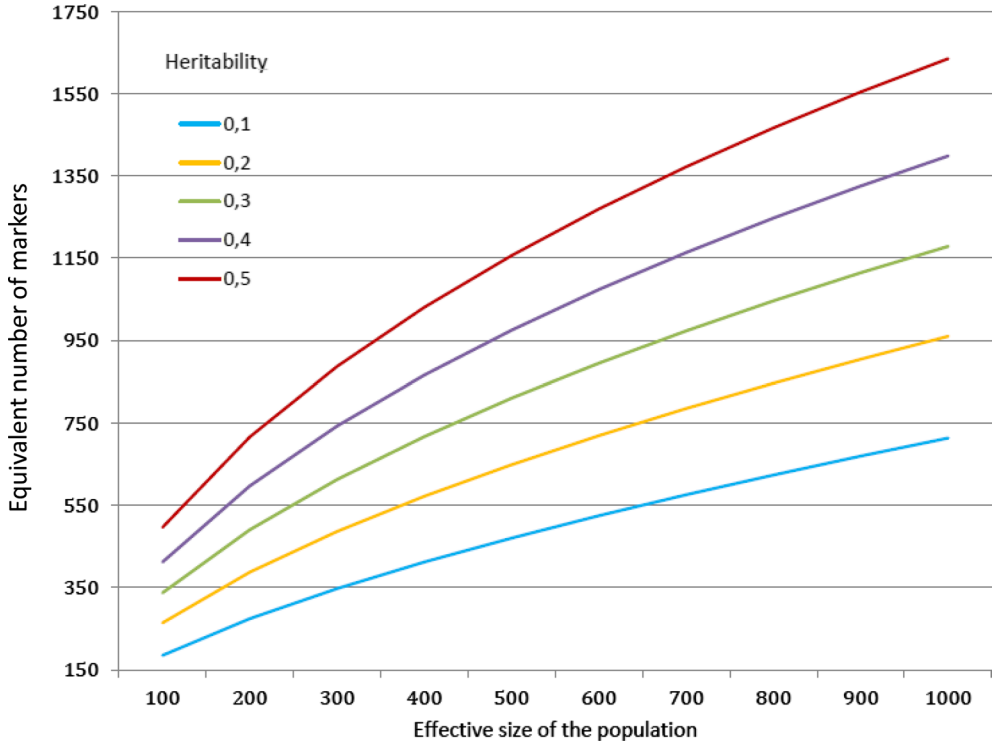

Fig. 5 Number of equivalent markers [from Eq. (8)] as a function of the effective population size (Ne) and heritability (v^2) ($n_M = 5000$; $n_R = 2000$)

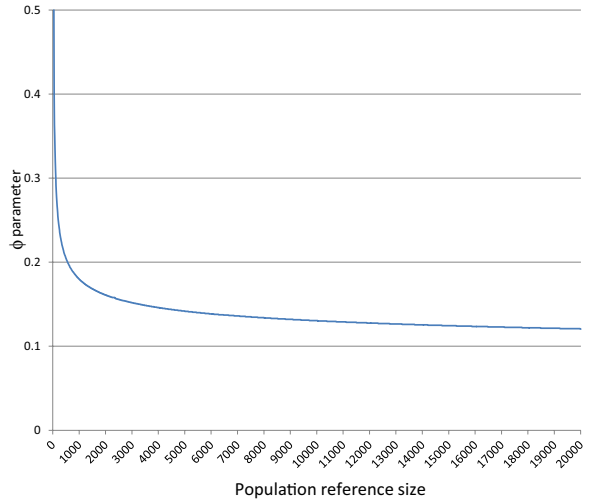

Fig. 6 Parameter of the beta distribution $\mathcal{B}(\phi,\phi)$ that best fits Goddard's distribution of allele frequencies

and $\mathrm{E}[\tau_2] = n_r \frac{4a^2+16a+18}{4a^2+8a+3}$), deriving the expectation of parameters $\sum_m \rho_m$, $\sum_m \rho_m^2$ and $\sum_m \frac{\rho_m^2}{\sigma_m^2}$ is not simple. However, these quantities are quite easily obtained by numerical integration. Thus, adjusting a Beta distribution to observed allele frequencies and numerically computing formula (3) parameters would be a feasible and more versatile implementation of our second genomic precision approximation.

Our work focused on the BLUP precision of the molecular score $r^2\left(q_{ci}, \hat{q}_{ci}|\mathbf{X}\right) = \frac{v(\hat{q}_{ci})}{v(q_{ci})}$ but left aside the proportion of the genetic variance that is captured by the markers $\left(\frac{v(q_{ci})}{v(g_{ci})}\right)$. This last term could be treated as in Goddard et al. [18]: $\frac{v(q_{ci})}{v(g_{ci})} = b = \frac{n_M}{n_M+M_e}$ with M_e the number of independent segments. As noted in the section on the general framework, the quantity $\frac{v(\hat{q}_{ci})}{v(g_{ci})} = b \times r(q_{ci}, \hat{q}_{ci}|\mathbf{X})$ is only an approximation of these GEBV reliabilities i.e. $r^2\left(g_{ci}, \hat{q}_{ci}|\mathbf{X}\right) = \frac{cov^2(g_{ci},\hat{q}_{ci}|\mathbf{X})}{v(g_{ci}|\mathbf{X})v(\hat{q}_{ci}|\mathbf{X})}$. Equality between those quantities is obtained when $\mathbf{X} = \mathbf{W}$(identity between statistical and genetical models), a condition assumed in Goddard [16] where markers and QTL are modeled as a series of uncorrelated pairs.

All the developments shown in this paper are based on the hypothesis that the reliability of GEBV based on non-independent markers for a trait controlled by n_Q QTL that are in incomplete LD with the markers can be approached by the reliability of GEBV when there are n_M independent segments that carry a single QTL in LD with a single marker. A few difficulties arose when applying this approach proposed by Goddard [16]. How many independent markers should be considered? The reasoning in Goddard [16] was based on the idea of an effective number of loci (M_e) corresponding to a given variance of realized relationships. Here, we proposed the alternative equivalent number of independent loci (M_i) which corresponds to a given reliability. We showed that this M_i number depends on the size of the reference population and on heritability, a dependence that does not occur with M_e. If we invert the argument, controlling the level of realized relationships variance with the effective number of loci (M_e) does not seem to be a good approach to control the estimated GEBV reliability.

As detailed by Hayes et al. [17], the effective number of independent chromosome segments depends on the population structure. The higher is the mean relationship level, the smaller is this effective number. However, we suggest the use of this number as estimated from a set of unrelated individuals, or of its expectation prior to any observation, assuming independence between individuals. Without formal proof, the idea was that long-term LD was considered by using an effective (or equivalent) number of independent loci while short-term non-independence was taken into account with our formalization of the matrix's expectations that is developed in Additional file 1. A complete proof of the procedure is still needed.

Regardless of the definition of M_e or M_i, there is no reason that the number of independent loci must equal the number of QTL, which is unknown, contrary to the hypothesis about pairs of marker-QTL (in practice, since the QTL effects are random variables, many segments will only have very small effects on the trait, thus simulating the more likely situation of a limited number of "real" QTL). Equating $\mathbf{X}$ and $\mathbf{W}$ as well as σ^2_β and σ^2_α has no clear justification. The variance $v\left(\hat{q}_c|\mathbf{X}\right)$ of the molecular score should not be $\sigma^2_\beta \mathbf{x_c}\mathbf{X_r}'(\mathbf{X_r}\mathbf{X_r}' + \mathbf{I}\lambda_\beta)^{-1}\mathbf{X_r}\mathbf{x_c}'$ but $\sigma^2_\alpha \mathbf{x_c}\mathbf{X_r}'(\mathbf{X_r}\mathbf{X_r}' + \mathbf{I}\lambda_\beta)^{-1}(\mathbf{W_r}\mathbf{W_r}' + \mathbf{I}\lambda_\alpha)(\mathbf{X_r}\mathbf{X_r}' + \mathbf{I}\lambda_\beta)^{-1}\mathbf{X_r}\mathbf{x_c}'$. This other formula assembles two sets of unknown parameters: the variances σ^2_α and σ^2_β, and the genotypes $\mathbf{X}$ and $\mathbf{W}$. It is often assumed that $\sigma^2_\beta = \sigma^2_g/(n_M\bar{\tau})$ (e.g. [1]), which results in an overestimation of the λ_β parameter since LD is not considered. Working on the number of independent loci (M_e or M_i) apparently solves this difficulty. The QTL variance $\sigma^2_\alpha = \sigma^2_g/\left(n_Q\bar{\tau}\right)$ could be derived based on a hypothesis about the number of QTL. The situation is more difficult for the genotype matrices since the $\mathbf{W_r}$ matrix is not observed.

If the framework considered so far (n_M markers-QTL pairs with strong LD within pairs and no LD between pairs) is partly retained, a slight improvement is possible considering the element b of the genetic variability

explained by SNPs. The idea would be to replace, in the formulae used in this paper, σ_q^2 by $b \times \sigma_g^2$. Element b can be derived by considering that the markers' ($\boldsymbol{\beta}$) and QTL' ($\boldsymbol{\alpha}$) effects are fixed in the genetic and statistical models. Leaving aside the singularity of $\mathbf{X_r'X_r}$ when the number of SNPs is large, the marker effects are now estimated by $\hat{\boldsymbol{\beta}} = \left(\mathbf{X_r'X_r}\right)^{-1}\mathbf{X_r'y}$ and the molecular score defined as $\hat{q} = \mathbf{X_r}\hat{\boldsymbol{\beta}}$, while the genetic value was $\mathbf{g} = \mathbf{W_r}\boldsymbol{\alpha}$. Given the genotype matrices, the sample genetic variability is $v_g = \boldsymbol{\alpha}'\mathbf{W_r'W_r}\boldsymbol{\alpha}$ and the sample molecular score variability $\mathbf{y'X_r}(\mathbf{X_r'X_r})^{-1}\mathbf{X_r'y}$ with an expectation $v_q = \boldsymbol{\alpha}'\mathbf{W_r'X_r}(\mathbf{X_r'X_r})^{-1}\mathbf{X_r'W_r}\boldsymbol{\alpha}$. The part of the genetic variability explained by the SNPs is the ratio $b = v_q/v_g$.

Expectations of the matrices' product elements $\left\{\mathbf{X_r'X_r}\right\}_{ml}$ are $2n_r\Delta_{ml}$ off diagonal and $2n_rp_m(1-p_m) = n_r\sigma_m^2$ in the diagonal, with similar expressions for $\mathbf{W_r'X_r}$ and $\mathbf{W_r'W_r}$ elements.

Following Goddard [16], approximating expectations of the matrices' functions by the function of their expectation, and assuming that (1) markers are independent, (2) each QTL q is in LD with only one marker $m(q)$, with a LD value $\Delta_{qm(q)}$, and (3) individuals are unrelated: $v_g = n_r \sum_q \alpha_q^2\sigma_q^2$, we get

$$v_q \sim 4n_r \sum_q \frac{\Delta_{qm(q)}^2}{\sigma_{m(q)}^2}\alpha_q^2 = n_r \sum_q r_{qm(q)}^2\alpha_q^2\sigma_q^2, \quad \text{and}$$

$$b = \frac{\sum_q r_{qm(q)}^2\alpha_q^2\sigma_q^2}{\sum_q \alpha_q^2\sigma_q^2}, \text{ corresponding to Eq. (4) in [16].}$$

Table 9 Expectation of the ratio of variances vs. the ratio of the variance expectations considering different reference population sizes and numbers of markers ($v^2 = 0.4$, 50 simulations)

n_r	n_M	$E\left[v\left(\hat{q}_c\right)/v(q_c)\right]$	$E\left[v\left(\hat{q}_c\right)\right]/E\left[v(q_c)\right]$
500	1000	0.403	0.401
1000	1000	0.726	0.725
1500	1000	1.010	1.008
2000	1000	1.212	1.212
500	1500	0.270	0.269
1000	1500	0.535	0.534
1500	1500	0.753	0.753
2000	1500	0.944	0.944
500	2000	0.213	0.213
1000	2000	0.414	0.413
1500	2000	0.597	0.597
2000	2000	0.760	0.759
500	2500	0.175	0.175
1000	2500	0.349	0.348
1500	2500	0.515	0.514
2000	2500	0.670	0.669

The ratio b is the weighted mean of LD r^2. Unfortunately, neither α_q^2 nor σ_q^2 are known. The unweighted mean $\frac{\sum_q r_{qm(q)}^2}{n_q} = \bar{r}^2$ may be a fruitful approximation. Following Sved [33], the expectation of $r_{qm(q)}^2$ is $\frac{1}{1+4N_ec}$ with c being the distance, in Morgan, between the QTL and its marker. Let L be the total length of the genome, and assume an equal distance L/n_M between each successive marker $b \sim \int_0^{L/2n_M} \frac{1}{1+4N_ec}\frac{1}{L/2n_M}dc = \frac{n_M}{2N_eL}\left[\log\left(1+2N_eL/n_M\right)\right]$.

The expectation of the reliability $E\left[r_{q_c,\hat{q}_c}^2\right]$, which is a ratio of variances $E_{\mathbf{X}}\left[v\left(\hat{q}_c|\mathbf{X}\right)/v(q_c|\mathbf{X})\right]$ was approximated by the ratio of the variance expectations $E_{\mathbf{X}}\left[v\left(\hat{q}_c|\mathbf{X}\right)\right]/E_{\mathbf{X}}[v(q_c|\mathbf{X})]$. The usual second degree approximation ($E[N/D] = E[N]/E[D] - cov[N, D]/E^2[D] + v[D]E[N]/E^3[D]$) could not be used here due to algebra complexity. However, in the case of unrelated individuals and independent markers, numerical evaluation of the difference between exact and approximated results for various reference population sizes and numbers of markers shows a very small underestimation of the reliability (Table 9).

The theory presented here was developed by considering a single selection candidate. When candidates are diversely related to the reference population, as suggested in Goddard et al. [18], the candidates should be examined one by one. Moreover, non-independence between candidates should be considered. A further step towards the modeling of genomic selection could be an approximation of the mean genetic values of selected individuals when GEBV reliabilities are heterogeneous.

A few other hypotheses were made in this paper, including additivity and *i.i.d.* of QTL effects, and the use of GBLUP. As long as the objective is to model and optimize breeding plans, only relative values are interesting and we assumed that these hypotheses were not critical.

Conclusions

The objective of this paper was to provide a further step towards the development of deterministic models that describe genomic breeding plans. Such deterministic models carry low computational burden and thus allow design optimization through intensive numerical exploration.

We proposed two alternative approximations of the estimation of GEBV reliability in the case of non-independence between candidate and reference populations. Both were derived from the Taylor series heuristic approach suggested by Goddard [16]. A numerical exploration of their properties showed that the series were not equivalent in terms of convergence

to the exact reliability, that the approximations may overestimate GEBV precision and that they perfectly converged toward their theoretical expectations.

Formulae derived for these approximations were simple to handle in the case of independent markers. A few parameters that describe the markers' genotypic variability (allele frequencies, linkage disequilibrium) can be estimated from genomic data corresponding to the population of interest or estimated after assumption about their distribution.

When markers are not in linkage equilibrium (*i.e.* there is LD), replacing the real number of markers and QTL by an effective or equivalent number of independent loci, as proposed by Goddard [16] and Hayes et al. [17], is a practical solution. Research efforts are still needed to overcome some strong limits of this approach.

Additional files

Additional file 1. Computation of $E\left[X_{im}^{d_i} X_{jm}^{d_j} \cdots X_{Km}^{d_K}\right]$ as a function of between-chromosome identity coefficients, in the case of independent markers. Using an extension of the identity coefficients theory, the document shows how to compute the elements of the expectation $E[XX'TXX']$ (*i.e.* $E\left[X_i X_j\right]$, $E\left[X_i X_j^2\right]$, $E\left[X_i X_j^3\right]$, $E\left[X_i^2 X_j^2\right]$, $E\left[X_i X_j X_k^2\right]$ and $E[X_i X_j X_k X_l]$) when markers are independent and individuals are related.

Additional file 2. Expectations of $\sum_m \sigma_m^2$, $\sum_m \sigma_m^4$, $\sum_m \rho_m$, $\sum_m \rho_m^2$, $\sum_m \rho_m^2 / \sigma_m^2$ and $\sum_m 1/\sigma_m^2$. The expectations of the listed quantities are computed assuming either a U-shaped or a uniform distribution of allele frequencies.

Additional file 3. Precision formulae when the candidate is related to reference individuals. The approximated formulae derived in the main text are applied to the case of a candidate for which the sire belongs to the reference population.

Additional file 4. Equivalent numbers of independent loci. (1) equation of the equivalent number of independent loci which gives the precision $E\left[r^2_{q_c,\hat{q}_c}\right] \sim \frac{E_{\mathbf{X}}[v(\hat{q}_c|\mathbf{X})]}{E_{\mathbf{X}}[v(q_c|\mathbf{X})]}$ obtained with the total number of non-independent markers; (2) simple approximation in a very simplified situation; (3) relations between number of independent markers and size of the reference population.

Additional file 5. The case of markers in linkage disequilibrium. Derivation of the crossed terms expectation of genomic values ($E[x_{cl}x_{cm}X_{il}X_{im}]$) in the situation of LD between loci *l* and *m*. Many examples are given for diverse situations.

Additional file 6. Another demonstration of Goddard et al. [18] accuracy. A complete demonstration is given using the notations of the present paper.

Author details
[1] GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), INRA, 31326 Castanet-Tolosan, France. [2] Animal Genetics and Breeding Unit, University of New England, Armidale, Australia.

Acknowledgements
This work was partly done when the author was on sabbatical leave in the Animal Genetic and Breeding Unit (AGBU) in Armidale, Australia. This sabbatical was supported by a grant from AGBU and from INRA (métaprogramme Selgen). Andrew Swan, Julius van der Werf, Mike Goddard, Anne Ricard and Bruno Goffinet are thanked for their many useful comments. Rob Banks is particularly thanked for his help at many levels.

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

Received: 23 July 2015 Accepted: 8 January 2016
Published online: 03 March 2016

References

1. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2011;157:1819–29.
2. Schaeffer LR. Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet. 2006;123:218–23.
3. König S, Simianer H, Willam A. Economic evaluation of genomic breeding programs. J Dairy Sci. 2009;92:382–91.
4. 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.
5. de Roos P, Schrooten WC, 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.
6. 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.
7. Meuwissen THE, Hayes BJ, Goddard ME. Accelerating improvement of livestock with genomic selection. Annu Rev Anim Biosci. 2013;1:221–37.
8. Sonesson AK, Meuwissen THE. Testing strategies for genomic selection in aquaculture breeding programs. Genet Sel Evol. 2009;41:37.
9. Ibáñez-Escriche N, Fernando RL, Toosi A, Dekkers JCM. Genomic selection of purebreds for crossbred performance. Genet Sel Evol. 2009;41:12.
10. 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;43:33.
11. Tribout T, Larzul C, Phocas F. Efficiency of genomic selection in a purebred pig male line. J Anim Sci. 2012;45:4164–76.
12. 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.
13. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One. 2008;3:e3395.
14. Legarra A, Robert-Granié C, Manfredi E, Elsen JM. Performance of genomic selection in mice. Genetics. 2008;180:611–8.
15. Van Raden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
16. Goddard ME. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 2009;136:245–57.
17. Hayes B, Visscher P, Goddard M. Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res. 2009;91:47–60.
18. 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.
19. 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.
20. 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.
21. 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.
22. Habier D, Fernando RL, Garrick DJ. Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics. 2013;194:597–607.

23. 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.
24. Weller JI. Economic aspects of animal breeding. London: Chapman & Hall; 1994.
25. Dekkers JCM. Prediction of response to marker-assisted and genomic selection using selection index theory. J Anim Breed Genet. 2007;124:331–41.
26. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams A. The impact of genetic architecture on genome-wide evaluation methods. Genetics. 2010;185:1021–31.
27. Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007;177:2389–97.
28. 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.
29. Wientjes YCJ, Veerkamp RF, Bijma P, Bovenhuis H, Schrooten C, Calus MPL. Empirical and deterministic accuracies of across-population genomic prediction. Genet Sel Evol. 2015;47:5.
30. Gianola D, De Los Campos G, Hill WG, Manfredi E, Fernando R. Additive genetic variability and the bayesian alphabet. Genetics. 2009;183:347–63.
31. 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.
32. Sigman K. Lecture notes introduction to discrete-time Markov chains. http://www.columbia.edu/~ks20/stochastic-I/stochastic-I-MCI.pdf. 2009.
33. Sved JA. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor Popul Biol. 1971;2:125–41.
34. Wright S. The distribution of gene frequencies in populations. Proc Natl Acad Sci USA. 1937;23:307–20.
35. La Gillois M. relation d'identité en génétique. Ann Inst Henri Poincaré. 1964;B2:1–94.
36. Harris DL. Genotypic covariances between inbred relatives. Genetics. 1964;50:1319–48.
37. Jacquard A. Logique du calcul des coefficients d'identité entre deux individus. Populations. 1966;21:751–76.