ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/314504044

Metafounders are related to Fst fixation indices and reduce bias in single-step genomic evaluations

Article *in* Genetics Selection Evolution · March 2017

DOI: 10.1186/s12711-017-0309-2

CITATIONS
8

READS
134

7 authors, including:

Carolina Andrea Garcia-Baccino
Universidad de Buenos Aires
3 PUBLICATIONS **14** CITATIONS

SEE PROFILE

Andres Legarra
French National Institute for Agricultural Research
207 PUBLICATIONS **4,269** CITATIONS

SEE PROFILE

Ignacy Misztal
University of Georgia
389 PUBLICATIONS **9,443** CITATIONS

SEE PROFILE

Ivan Pocrnic
University of Georgia
20 PUBLICATIONS **47** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Prediction of Breeding value View project

Project Signatures of selection View project

All content following this page was uploaded by Andres Legarra on 13 March 2017.

The user has requested enhancement of the downloaded file.

Garcia-Baccino *et al. Genet Sel Evol (2017) 49:34*
DOI 10.1186/s12711-017-0309-2

Genetics Selection Evolution

RESEARCH ARTICLE Open Access

Metafounders are related to F_{st} fixation indices and reduce bias in single-step genomic evaluations

Carolina A. Garcia-Baccino[1,2], Andres Legarra[3*], Ole F. Christensen[4], Ignacy Misztal[5], Ivan Pocrnic[5], Zulma G. Vitezica[3] and Rodolfo J. C. Cantet[1,2]

Abstract

Background: Metafounders are pseudo-individuals that encapsulate genetic heterozygosity and relationships within and across base pedigree populations, i.e. ancestral populations. This work addresses the estimation and usefulness of metafounder relationships in single-step genomic best linear unbiased prediction (ssGBLUP).

Results: We show that ancestral relationship parameters are proportional to standardized covariances of base allelic frequencies across populations, such as F_{st} fixation indexes. These covariances of base allelic frequencies can be estimated from marker genotypes of related recent individuals and pedigree. Simple methods for their estimation include naïve computation of allele frequencies from marker genotypes or a method of moments that equates average pedigree-based and marker-based relationships. Complex methods include generalized least squares (best linear unbiased estimator (BLUE)) or maximum likelihood based on pedigree relationships. To our knowledge, methods to infer F_{st} coefficients from marker data have not been developed for related individuals. We derived a genomic relationship matrix, compatible with pedigree relationships, that is constructed as a cross-product of {−1,0,1} codes and that is equivalent (apart from scale factors) to an identity-by-state relationship matrix at genome-wide markers. Using a simulation with a single population under selection in which only males and youngest animals are genotyped, we observed that generalized least squares or maximum likelihood gave accurate and unbiased estimates of the ancestral relationship parameter (true value: 0.40) whereas the naïve method and the method of moments were biased (average estimates of 0.43 and 0.35). We also observed that genomic evaluation by ssGBLUP using metafounders was less biased in terms of estimates of genetic trend (bias of 0.01 instead of 0.12), resulted in less overdispersed (0.94 instead of 0.99) and as accurate (0.74) estimates of breeding values than ssGBLUP without metafounders and provided consistent estimates of heritability.

Conclusions: Estimation of metafounder relationships can be achieved using BLUP-like methods with pedigree and markers. Inclusion of metafounder relationships reduces bias of genomic predictions with no loss in accuracy.

Background

Metafounders are pseudo-individuals that describe relationships within and across pedigree base populations. The concept of metafounders provides a coherent framework for the theory of genomic evaluation [1]. Genomic evaluation in agricultural species often implies partially genotyped populations, i.e. some individuals are genotyped, using high-density genetic markers across the genome, others are not, and phenotypes may be recorded in either of the two subsets. An integrated solution called single-step genomic best linear unbiased prediction (ssGBLUP) has been proposed [2–4]. This solution uses the following integrated relationship matrix:

$$\mathbf{H} = \begin{pmatrix} \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{pmatrix},$$

*Correspondence: andres.legarra@inra.fr
[3] GenPhySE, INRA, INPT, ENVT, Université de Toulouse, 31326 Castanet-Tolosan, France
Full list of author information is available at the end of the article

BioMed Central

© The Author(s) 2017. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

with inverse:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{pmatrix},$$

where $\mathbf{G}$ is the genomic relationship matrix, $\mathbf{A}$ is the pedigree-based relationship matrix, and matrices $\mathbf{A}_{11}, \mathbf{A}_{12}, \mathbf{A}_{21}, \mathbf{A}_{22}$ are submatrices of $\mathbf{A}$ with labels 1 and 2 denoting non-genotyped and genotyped individuals, respectively.

Because genotyped animals are often not a random sample from the analyzed populations (they tend to be younger or selected), it was quickly acknowledged that a proper analysis requires specifying different means for genotyped and non-genotyped individuals for the trait under consideration. These different means can be considered as parameters of the model, which are either fixed [4] or random [5, 6] effects. In the latter case, the random variables induce covariances between individuals, a situation that is informally referred to as "compatibility" of genomic and pedigree relationships. In fact, compatibility implies equality of the average breeding value of the base population and of the genetic variance [7] across the different measures of relationships. Numerically, the problem appears as follows. The formulae for matrix $\mathbf{H}$ and its inverse contain $(\mathbf{G} - \mathbf{A}_{22})$ and $(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1})$ (assuming $\mathbf{G}$ is full rank), respectively. This suggests that if $\mathbf{G}$ and $\mathbf{A}_{22}$ differ too much, biases may appear.

Genomic relationships are usually computed in one of two manners: using "cross-products" [8] or "corrected identity-by-state (IBS)" [9]. Both depend critically on assumed allele frequencies at markers in the pedigree base population [10]. Base allele frequencies are often unavailable. However, for most purposes, allele frequencies are not of interest *per se* and can be treated as nuisance parameters that can be marginalized. Christensen [11] achieved an algebraic integration of allele frequencies, leading to a very simple covariance structure with allele frequencies in genomic relationships fixed at 0.5 (e.g., using genotypes coded as {−1,0,1} in the cross-product method) and a parameter called γ that describes relationships between pedigree founders i.e. $\mathbf{A}_{base}^{(\gamma)} = \mathbf{I}\left(1 - \frac{\gamma}{2}\right) + \mathbf{1}\mathbf{1}'\gamma$ in the base population. A second parameter in Christensen's marginalisation is s, which is a measure of marker heterozygosity in the base population. Therefore, instead of inferring (thousands of) base allele frequencies, inference can be based on two simple parameters γ and s. Both can be estimated by maximizing the likelihood of observed genotypes. In addition, this approach considers the fact that pedigree depth is arbitrary and mostly based on historical availability of records.

Legarra et al. [1] showed the equivalence of the Christensen approach to the metafounder concept: pseudo-individuals that encapsulate three ideas: (a) separate means for each base population [4, 12, 13], (b) randomness of these means [5] and (c) propagation of the randomness of these means to the progeny [11], while accommodating several populations with complex crosses, e.g. [14]. Legarra et al. [1] also generalized one relationship between founders (scalar γ) to several relationships between founders in the pedigree, i.e. ancestral relationships (matrix $\mathbf{\Gamma}$), and suggested simple methods to estimate them. Legarra et al. [1] showed that construction of $\mathbf{A}^{\mathbf{\Gamma}}$ from $\mathbf{\Gamma}$ and a pedigree reduces to the use of the tabular rules [15] for construction of relationships, and its inversion is achieved by inversion of $\mathbf{\Gamma}$ followed by Henderson's rules [16]. We provide an example of matrices $\mathbf{A}^{\mathbf{\Gamma}}$ and $\mathbf{\Gamma}$ in "Appendix". However, the performance of their model has not been tested so far, either for estimation of ancestral relationships or for genomic evaluation.

This work has two objectives. The first is to show that the structure of the metafounder approach yields an alternative parameterization and method for estimation of ancestral relationships. By doing so, we found that ancestral relationships are generalizations of Wright's F_{st} fixation index [17]. The second goal is to test, by simulation, (1) methods to estimate ancestral relationship parameters, (2) the quality of genomic predictions using metafounders, and (3) the quality of variance component estimation. For the second goal, the simulated population is undergoing selection and with a complete partially genotyped pedigree.

Methods

Relationship between metafounders and allele frequencies in the pedigree base population

Single population

Let $\mathbf{M}$ be a matrix of genotypes coded as gene content, i.e. {0,1,2} and the genomic relationship matrix $\mathbf{G} = (\mathbf{M} - \mathbf{J})(\mathbf{M} - \mathbf{J})'/s$, with $\mathbf{J}$ a matrix of 1s, with reference alleles taken at random, so that the expected allele frequency p is 0.5 for a random locus [11]. In other words, the matrix $\mathbf{Z} = (\mathbf{M} - \mathbf{J})$ contains values of {−1,0,1} for each genotype. In a single population, let γ be the relationship coefficient between pedigree founders or, equivalently, the self-relationship of the metafounder [1, 11]. Parameter s (defined above) is a measure of marker heterozygosity in the population. Ancestral relationships in γ explain, for instance, genomic relationships in $\mathbf{G} = (\mathbf{M} - \mathbf{J})(\mathbf{M} - \mathbf{J})'/s$ that are not captured by available pedigree; e.g. across nominally unrelated individuals. It will be shown later that this relationship γ is relative to a population with maximum heterozygosity and is analogous to an F_{st} fixation index [18].

Christensen [11] estimated the two parameters, γ and s, using maximum likelihood, whereas [1] suggested

methods of moments. Closer inspection of Appendix A in [11] leads to the following developments that were not described in Christensen [11] (see "Appendix" of the present paper for more details).

Parameter γ is such that $\gamma = \frac{4Var(p_i)}{2Var(p_i)+E(2p_iq_i)}$, with $p_i = 1 - q_i$ the allele frequency at a random locus i. Parameter $s = n(2Var(p_i) + E(2p_iq_i))$, with n being the number of markers. However, $E(2p_iq_i) = 2E(p_i)E(q_i) - 2Var(p_i) = 0.5 - 2Var(p_i)$, such that if reference alleles are chosen at random across loci, then $E(p_i) = E(q_i) = 0.5$. From this it follows that:

$$s = \frac{n}{2} = \frac{number\ of\ markers}{2},$$

and the genomic relationship matrix is $\mathbf{G} = 2(\mathbf{M} - \mathbf{J})(\mathbf{M} - \mathbf{J})'/n$. Interestingly, this matrix is similar to a matrix of IBS relationships, that can be written as:

$$\mathbf{G}_{IBS} = (\mathbf{M} - \mathbf{J})(\mathbf{M} - \mathbf{J})'/n + \mathbf{1}\mathbf{1}',$$

so that $\mathbf{G}_{IBS} = \frac{1}{2}\mathbf{G} + \mathbf{1}\mathbf{1}'$ (see proof in "Appendix").

Substituting $E(2p_iq_i) = 0.5 - 2Var(p_i)$ into the expression $\gamma = \frac{4Var(p_i)}{2Var(p_i)+E(2p_iq_i)}$ gives: $\gamma = 8Var(p_i) = 8\sigma_p^2$, such that γ for a single population is eight times the variance of allele frequencies in the base population (this variance was described by Cockerham [19]). We stress that $Var(p_i) = \sigma_p^2$ to imply that σ_p^2 (and γ) is a parameter, the variance of allele frequencies across markers [10, 11, 20, 21]. However, s can be considered as equivalent to heterozygosity when all markers have an allele frequency of 0.5, that is, the maximum possible heterozygosity.

Multiple populations

In an analogous manner, the relationship between two metafounders b and b' is: $\gamma_{b,b'} = 8Cov(p_{b,i}, p_{b',i}) = 8\sigma_{p_b,p_{b'}}$, i.e., the covariance across loci between allele frequencies of two populations b and b'. This is almost tautological: the relationship (between two populations in this case) is the covariance between the gene content at a locus. Christensen et al. [6] implicitly show this in Appendix A of their paper. Cockerham [19] and Robertson [22] interpreted 4 $\sigma_{p_b,p_{b'}}$ as the coancestry between two populations and Fariello et al. [23] used $\sigma_{p_b,p_{b'}}$ to describe the divergence of populations. Several measures of genetic distance between populations have been developed (e.g. [24]), and most of them contain a term that is related, implicitly or explicitly, to $\sigma_{p_b,p_{b'}}$. In particular, the average square of the Euclidean distance can be written as $D^2 = E((p_b - p_{b'})^2) = -2\sigma_{p_b,p_{b'}}$. Thus, $\gamma_{b,b'} = -4D^2$.

Estimation

Estimation in a single population

Estimation of s is trivial, it is simply half the number of markers. Parameter γ is proportional to the variance of allele frequencies in the base population. If base population individuals were genotyped, computing allele frequencies and estimating γ would be trivial. In the next section, we propose methods when this is not the case, i.e. genotyped individuals are related and perhaps several generations away from the base population.

Assuming no pedigree structure i.e. naïve The simplest model assumes that genotyped individuals are unrelated and constitute the base population. For locus i, let $\boldsymbol{m}_i$ be a vector of gene contents in the form {0,1,2}, defined as before. The mean of this vector is $\mu_i = 2p_i$. For each locus, μ_i is estimated as the observed mean of $\boldsymbol{m}_i$, then $Var(\hat{\boldsymbol{\mu}})$ is computed as the empirical variance across loci of $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)$, and because $p_i = \mu_i/2$, then $\hat{\sigma}_p^2 = Var(\hat{\boldsymbol{\mu}})/4$ and $\gamma = 8\hat{\sigma}_p^2 = 2Var(\hat{\boldsymbol{\mu}})$.

Considering pedigree structure At locus i, gene content can be seen as a quantitative trait mean of $\boldsymbol{m}_i$ in the base population equal to $2p_i$, where p_i is the allele frequency in the base population and the genetic variance is $2p_iq_i$ [25–27]. Cockerham [19] showed that the covariance of gene content of marker i between individuals j and k is a function of their relationship (A_{jk}): $Cov(m_{i,j}, m_{i,k}) = A_{jk}2p_iq_i$. A linear model can therefore be written as:

$$\mathbf{m}_i = \mathbf{1}\mu_i + \mathbf{W}\mathbf{u}_i + \mathbf{e},$$

where $\mathbf{W}$ is an incidence matrix relating individuals in the pedigree to observed genotypes, and $\mathbf{u}_i$ is the deviation of each individual from the mean μ_i for all individuals [25–27]. Assuming multivariate normality: $\boldsymbol{\mu} \sim N(\mathbf{0}, \mathbf{I}\sigma_\mu^2)$ and $\mathbf{u}_i \sim N(\mathbf{0}, \mathbf{A}(2p_iq_i)) = N(\mathbf{0}, \mathbf{A}\sigma_{m_i}^2)$.

Equivalently, for the set of genotyped individuals (labelled as "2"), $\mathbf{u}_{2,i} \sim N(\mathbf{0}, \mathbf{A}_{22}(2p_iq_i))$, where $\mathbf{A}_{22} = \mathbf{WAW}'$ is an additive relationship matrix that includes only the genotyped individuals. From this formulation, there are two possible strategies to estimate σ_μ^2.

Generalized least squares (GLS) This ignores the prior distribution of $\boldsymbol{\mu}$ and estimates each μ_i as a "fixed effect", using best linear unbiased estimator (BLUE) (or, equivalently, GLS) estimators of μ_i separately for each locus. One option is to use the $\mathbf{A}^{-1}$ spanning all the pedigree and mixed model equations [25–27]. Equivalently, the corresponding GLS expression is:

$$\hat{\mu}_i = \left(\mathbf{1}'\mathbf{A}_{22}^{-1}\sigma_{m_i}^{-2}\mathbf{1}\right)^{-1}\mathbf{1}'\mathbf{A}_{22}^{-1}\boldsymbol{m}_i\sigma_{m_i}^{-2} = \left(\mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{1}\right)^{-1}\mathbf{1}'\mathbf{A}_{22}^{-1}\boldsymbol{m}_i,$$

where $\left(\mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{1}\right)$ is the sum of elements of $\mathbf{A}_{22}^{-1}$, $\sigma_{m_i}^2 = 2p_iq_i$ and $\mathbf{1}'\mathbf{A}_{22}^{-1}\boldsymbol{m}_i$ is a weighted sum of genotypes.

Then, σ^2_μ is estimated as $Var(\hat{\boldsymbol{\mu}})$, because $p_i = \mu_i/2$, $\hat{\sigma}^2_p = \sigma^2_\mu/4$, and it follows that $\hat{\gamma} = 2\hat{\sigma}^2_\mu$.

Maximum likelihood If allele frequencies in the base population have a distribution, μ_i can be considered as drawn from a normal distribution, $\boldsymbol{\mu} \sim N\left(\mathbf{0}, \mathbf{I}\sigma^2_\mu\right)$. Thus σ^2_μ is a variance component that can be estimated by maximum likelihood (ML). The equations for given values of σ^2_μ and $\sigma^2_{m_i} = 2p_iq_i$ are $\left(\mathbf{1}'\mathbf{A}^{-1}_{22}\sigma^{-2}_{m_i}\mathbf{1} + \sigma^{-2}_\mu\right)\hat{\mu}_i = \mathbf{1}'\mathbf{A}^{-1}_{22}\sigma^{-2}_{m_i}\mathbf{m}_i$. An expectation–maximization scheme [28] to obtain ML is as follows. Pick starting values for σ^2_μ and $\sigma^2_{m_i}$. Iterate until convergence on:

1. For each marker i,

   (a) estimate $\hat{\mu}_i = \left(\mathbf{1}'\mathbf{A}^{-1}_{22}\sigma^{-2}_{m_i}\mathbf{1} + \sigma^{-2}_\mu\right)^{-1}\mathbf{1}'\mathbf{A}^{-1}_{22}\sigma^{-2}_{m_i}\mathbf{m}_i$,

   (b) store $PEV_i\left(\hat{\mu}_i\right) = \left(\sigma^{-2}_\mu + \mathbf{1}'\mathbf{A}^{-1}_{22}\sigma^{-2}_{m_i}\mathbf{1}\right)^{-1}$,

   (c) update $\sigma^2_{m_i}$ as $\hat{\sigma}^2_{m_i} = 2\hat{p}_i\hat{q}_i$ with $\hat{p}_i = \hat{\mu}_i/2$;

2. Update σ^2_μ as $\hat{\sigma}^2_\mu = \frac{1}{n}\left(\hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}} + \sum PEV_i\left(\hat{\mu}_i\right)\right)$, where the second part of the expression corresponds to the trace $tr(\mathbf{IC})$, $\mathbf{I}$, the identity matrix, is the relationship structure across levels of $\boldsymbol{\mu}$ and $\mathbf{C}$ is the prediction error covariance matrix of $\hat{\boldsymbol{\mu}}$. As only the diagonal elements of $\mathbf{C}$ are needed in $tr(\mathbf{IC})$, its elements $PEV_i\left(\hat{\mu}_i\right)$ can be obtained separately from each single locus analysis.

At convergence, the estimate is $\hat{\gamma} = 2\hat{\sigma}^2_\mu$. This gives the same estimate as the method based on a Wishart likelihood function [11] with $s = n/2$ (results not shown).

Estimation in multiple populations

If t base populations are considered, the variance component σ^2_μ generalizes to $\boldsymbol{\Sigma}_0$, a $t \times t$ matrix of variances and covariances between means $\mu^{[b]}_i$ for marker i in population b. Across populations,

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} \sigma^2_{\mu^{[1]}\mu^{[1]}} & \sigma_{\mu^{[1]}\mu^{[2]}} & \cdots \\ \cdots & \sigma^2_{\mu^{[2]}\mu^{[2]}} & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix} \text{and } \hat{\boldsymbol{\Gamma}} = 2\hat{\boldsymbol{\Sigma}}_0.$$

Assuming no pedigree structure

Naïve If relationships across individuals are ignored:

$$\mathbf{m}_i = \mathbf{Q}\boldsymbol{\mu}_i + \mathbf{e}_i,$$

where $\mathbf{Q}$ is a matrix, the rows of which sum to 1, and that assigns individuals to fractions of populations, and $\boldsymbol{\mu}_i$ is a vector with t elements for the average of each population. For each locus, $\boldsymbol{\mu}_i$ can be estimated using least squares and the covariance matrix of $\boldsymbol{\mu}_i$ across loci gives an estimate of $\boldsymbol{\Sigma}_0$, e.g. for two populations $\hat{\boldsymbol{\Sigma}}_0 = Cov\left(\boldsymbol{\mu}^{[1]}, \boldsymbol{\mu}^{[2]}\right)$, a two-by-two matrix.

Considering pedigree structure

If there are no crosses between individuals from different populations in the pedigree, the estimation of allele frequencies in each base population can be split in separate analyses by population b: $\mathbf{m}^b_i = \mathbf{1}\mu^{[b]}_i + \mathbf{W}^b\mathbf{u}^b_i + \mathbf{e}$, with $\mathbf{u}^b_i \sim N\left(\mathbf{0}, \mathbf{A}^b(2p_i(1-p_i))\right)$ and $\mathbf{A}^b$ the matrix of pedigree-based relationships among individuals in population b, and the analysis proceeds as in a single population. Then, $\hat{\boldsymbol{\Sigma}}_0$ is estimated as the observed matrix of covariances for $\hat{\mu}^b_i$ across loci.

When there are crosses, there are two alternatives.

Generalized least squares (GLS)

The first alternative [27] is to use a genetic groups model [12, 13], as $\mathbf{m}_i = \mathbf{Q}\boldsymbol{\mu}_i + \mathbf{W}\mathbf{u}_i + \mathbf{e}$, where $\mathbf{Q}_{k,b}$ contains the fraction of ancestry b in individual k. This ignores the fact that the variance of gene content, $(2p_iq_i)$, differs between breeds and crosses. The second, and more exact alternative, is to use the representation where the breeding values are split into within- and across-breed components [29]:

$$\mathbf{m}_i = \mathbf{Q}\boldsymbol{\mu}_i + \sum_b \mathbf{W}^b\mathbf{u}^b_i + \sum_{b,b',b>b'} \mathbf{W}^{b,b'}\mathbf{u}^{b,b'}_i + \mathbf{e},$$

with partial relationship matrices for vectors $\mathbf{u}^b_i$ and $\mathbf{u}^{b,b'}_i$. The BLUE's of $\boldsymbol{\mu}_i$ can be obtained and then $\hat{\boldsymbol{\Sigma}}_0$ estimated as above.

Maximum likelihood (ML)

Analogously to the single population case, an expectation–maximization updated estimate can be obtained using multiple-trait formulations [28], where *PEC* is the prediction error variance–covariance, e.g. with two populations:

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} \boldsymbol{\mu}^{[1]'}\boldsymbol{\mu}^{[1]} & \boldsymbol{\mu}^{[1]'}\boldsymbol{\mu}^{[2]} \\ \boldsymbol{\mu}^{[2]'}\boldsymbol{\mu}^{[1]} & \boldsymbol{\mu}^{[2]'}\boldsymbol{\mu}^{[2]} \end{pmatrix}.$$

Our implementation of this approach is as follows:

1. For each marker i:

   (a) estimate $\hat{\boldsymbol{\mu}}_i = \left(\boldsymbol{\Sigma}^{-1}_0 + \mathbf{Q}'\mathbf{A}^{-1}_{22}\sigma^{-2}_{m_i}\mathbf{Q}\right)^{-1}\mathbf{Q}'\mathbf{A}^{-1}_{22}\sigma^{-2}_{m_i}\mathbf{m}_i$,

   (b) store $PEC_i\left(\hat{\boldsymbol{\mu}}_i\right) = \left(\boldsymbol{\Sigma}^{-1}_0 + \mathbf{Q}'\mathbf{A}^{-1}_{22}\sigma^{-2}_{m_i}\mathbf{Q}\right)^{-1}$,

   (c) update $\sigma^2_{m_i}$ as $\hat{\sigma}^2_{m_i} = 2\hat{p}^*_i\left(1-\hat{p}^*_i\right)$ with $\hat{p}^*_i = \frac{1}{Nb}\sum_{b=1,Nb}\frac{\hat{\mu}^b_i}{2}$;

2. Update $\mathbf{\Sigma}_0$ using cross-products within and across populations as, e.g., with two populations:

$$\hat{\mathbf{\Sigma}}_0 = \frac{1}{n}\left(\begin{pmatrix} \hat{\boldsymbol{\mu}}^{[1]'}\hat{\boldsymbol{\mu}}^{[1]} & \hat{\boldsymbol{\mu}}^{[1]'}\hat{\boldsymbol{\mu}}^{[2]} \\ \hat{\boldsymbol{\mu}}^{[2]'}\hat{\boldsymbol{\mu}}^{[1]} & \hat{\boldsymbol{\mu}}^{[2]'}\hat{\boldsymbol{\mu}}^{[2]} \end{pmatrix} + \sum_{i=1,n} PEC_i\right).$$

Step 1 includes an approximation in (1c) because we assume that $\sigma^2_{m_i} = 2p_iq_i$ is the same for all base populations, as in the GLS above, which could be further improved by using partial relationship matrices. This point will be addressed in future research.

Simulation

To assess the quality of genomic predictions using one metafounder, we simulated data using QMSim [30]. The simulation closely followed that in [5] to mimic a dairy cattle selection scheme scenario. A historical population undergoing mutation and drift was generated, followed by a recent population undergoing selection.

First, 100 generations of the historical population were generated with an effective population size of 100 during the first 95 generations, followed by a gradual expansion during the last five generations to an effective population size of 3000. Thirty chromosomes of 100 cM and 40,000 segregating biallelic markers distributed at random along the chromosomes in the first generation of the historical population were simulated. The 40,000 markers were resampled from a larger set of 90,000 markers in order to obtain allelic frequencies from a beta(2,2) distribution, similar to dairy cattle marker data, so that parameter γ had a true value around 0.40. There were 1500 QTL affecting the phenotype; QTL allele effects were sampled from a Gamma distribution with a shape parameter of 0.4. Mutation rate at the markers (recurrent mutation process) and QTL was assumed to be 2.5×10^{-5} per locus per generation [31]. We used a higher mutation rate than typical (10^{-8}, [32, 33]) to overcome the fact that QMSim is not a coalescent simulator. Phenotypes for a trait recorded only on females with a heritability of 0.30 were simulated.

Then, 10 overlapping generations of selection followed. In each generation, 200 males were mated with 2600 females to produce 2600 offspring by a positive assortative mating design based on EBV. Within the simulation, individuals were selected according to estimated breeding value (EBV) based on pedigree BLUP. In each generation, 40% of males and 20% of females were replaced by selected younger individuals. No restrictions were set to avoid or minimize inbreeding, so highly inbred individuals were found, as a result of strong selection and matings among highly-related individuals. A total of 100 individuals had an inbreeding coefficient higher than 0.20 (mainly found in the last generation), with some individuals having inbreeding coefficients higher than 0.40. True breeding values (TBV) and pedigree information were available for all 10 generations (28,800 individuals in the pedigree), phenotypes were available for all females except in the last generation (14,300 records). The 840 sires of females with phenotypic records were genotyped, as well as 2600 individuals in generation 9 (with records) and 2600 in generation 10 (without records). A total of 20 independent replicates were made. A two-step analysis was carried out using the simulated data. First, we compared several methods to estimate γ. Then, we tested the quality of genomic predictions using four methods (see section on genomic prediction methods), one of which included one metafounder.

Methods to estimate γ

Parameter γ was estimated using four estimation methods. First, the naïve method that does not consider the pedigree structure. Pedigree information was included in three methods: GLS, ML, and the method of moments (MM) in [1]. For a single population, the last method involves estimation of γ based on summary statistics of $\mathbf{A}_{22}$ (regular pedigree-relationship matrix for genotyped individuals) and $\mathbf{G}$ (the genomic relationship matrix).

Genomic prediction methods

The EBV of the selection candidates in generation 10 (genotyped and without phenotype records) were estimated using four methods. The first was the pedigree-based BLUP (PBLUP) based on phenotype and pedigree information. The second method was ssGBLUP, in which genomic information is also taken into account. We used the correction of [34] to equate genomic and pedigree average inbreeding and relationships, the default method used in most practical applications [34, 35]. However, the implementation that we used does not include inbreeding in the setup of $\mathbf{A}^{-1}$ [36], although it does consider inbreeding in $\mathbf{A}_{22}^{-1}$ (see below for use of these matrices). The third method was ssGBLUP that includes inbreeding in the setup of $\mathbf{A}^{-1}$ and of $\mathbf{A}_{22}^{-1}$ (ssGBLUP_F). The fourth method was ssGBLUP with the metafounder (ssGBLUP_M), using γ estimated by GLS since it turned out to be an accurate method to estimate $\mathbf{\Gamma}$ (see the Results section). The four methods used the following inverse relationship matrices: PBLUP: $\mathbf{A}^{-1}$; ssGBLUP: $\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{G}_a^{-1} - \mathbf{A}_{22}^{-1} \end{pmatrix}$ where $\mathbf{G}_a$ is as in [34] and $\mathbf{A}^{-1}$ is constructed ignoring inbreeding [36]; ssGBLUP_F: same as ssGBLUP, with $\mathbf{A}^{-1}$ correctly constructed; ssGBLUP_M: $\mathbf{H}^{(\gamma)-1} = \mathbf{A}^{(\gamma)-1} + \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{(\gamma)-1} \end{pmatrix}$ where

$\mathbf{G} = (\mathbf{M} - \mathbf{J})(\mathbf{M} - \mathbf{J})'/s$ with $s = n/2$ (see the "Methods" section) and $\mathbf{A}^{(\gamma)}$ is as in [1]. More details are given in "Appendix". For computation, we used blupf90 [37]. In the case of ssGBLUP_M, we constructed $\mathbf{H}^{(\gamma)-1}$ with own software and then used the option user_file in blupf90 (http://nce.ads.uga.edu/wiki/doku.php).

Quality of genomic prediction

Prediction quality was evaluated for all 2600 selection candidates in generation 10. The accuracy of the methods was measured as the Pearson correlation between TBV and EBV. Bias was calculated as the difference between the average TBV and average EBV with respect to the base population (i.e. to the solution of the metafounder for ssGBLUP_M or to 0 for the other methods). Thus, bias is related to estimated genetic progress in the selection candidates. The inflation (often also called bias) of the prediction method was quantified by the coefficient of regression of TBV on EBV. These two statistics correspond to the coefficients b_0 and b_1 in the Interbull validation method [38], which uses the regression $TBV = b_0 + b_1 EBV + e$. The mean square error (MSE) of prediction of EBV was calculated as the mean of the squared difference between TBV and EBV. An ideal method should have maximum accuracy, minimum MSE, zero bias, and a regression coefficient of 1. These are not only elegant statistical properties but also have relevance in livestock selection [39–41]. Changes in ranking of the selection candidates were also assessed by calculating the Spearman's rank correlation coefficients between EBV across methods.

In addition, the quality of variance component estimation was assessed by comparing estimated and simulated heritabilities. For this purpose, variance components were estimated by REML with remlf90 [37] based on the four methods (PBLUP, ssGBLUP, ssGBLUP_F, ssGBLUP_M).

Results

Estimation of γ

Figure 1 shows boxplots of the differences between the estimates of γ based on the four methods (MM, Naïve, ML and GLS) and the true values obtained by simulation, for each of the 20 replicates. The simulations were tailored to produce $\gamma = 0.40$. Methods ML and GLS estimated γ very accurately. Method MM clearly underestimated γ, whereas the Naïve method overestimated it. Based on these results, we used γ estimated by GLS for ssGBLUP_M for prediction. The effect of employing different values of γ in genomic prediction was assessed to quantify its impact on predictions. Using estimates of γ based on MM only slightly changed results. For example, the accuracies and slopes of ssGBLUP_M were not affected up to the 4th digit (not shown).

Quality of genomic prediction

Correlations between TBV and EBV of candidates in generation 10 for each prediction methods are in Table 1 and Fig. 2a. Compared with PBLUP, ssGBLUP_F and ssGBLUP_M increased accuracy by approximately 23 absolute points. This shows an important improvement by including marker information in the prediction and the possibility of generating a small extra gain when also including the metafounder. Method ssGBLUP resulted in a small loss of accuracy compared to ssGBLUP_F and ssGBLUP_M.

Table 1 and Fig. 2b display the regression coefficient of TBV on EBV, which measures the degree of inflation for each prediction method and should be close to 1. PBLUP and ssGBLUP_F produced values closest to 1. Including genomic data in the prediction based on ssGBLUP resulted in regression coefficients lower than 1, but including the metafounder in ssGBLUP_M gave values closer to 1. Methods ssGBLUP_M and ssGBLUP_F displayed a lower standard deviation compared to the other two methods. Again, method ssGBLUP showed the highest variability.

Biases of EBV obtained with each prediction method are in Table 1 and Fig. 2c. Both PBLUP and ssGBLUP_M were unbiased, whereas ssGBLUP and ssGBLUP_F were biased. The bias was higher for ssGBLUP than for ssGBLUP_F, which was largely due to a single outlier; the median bias was roughly the same for ssGBLUP and ssGBLUP_F. The bias with ssGBLUP_F was equivalent to roughly 0.5 generations of genetic improvement or 0.4 standard genetic deviations. Finally, ssGBLUP_M had the lowest MSE (closer to zero), followed by ssGBLUP_F (Table 1).

Ranking of EBV

The methods were also compared based on rank correlations of EBV with TBV and between methods. A rank correlation of 1 implies that the same candidates would be selected. Results are in Table 2. Rank correlations with TBV were similar to the Pearson correlations in Table 1. Selection decisions differed only slightly when using ssGBLUP, ssGBLUP_F or ssGBLUP_M. Note, however, that this table reports rank correlations among young selection candidates in the last generation and does not address comparisons across generations (e.g. old vs. young animals), which is sensitive to the biases that are reflected in Table 1 [41]. For instance, all young animals would be overestimated by 0.11 with ssGBLUP_F, which results in these young animals looking better than proven

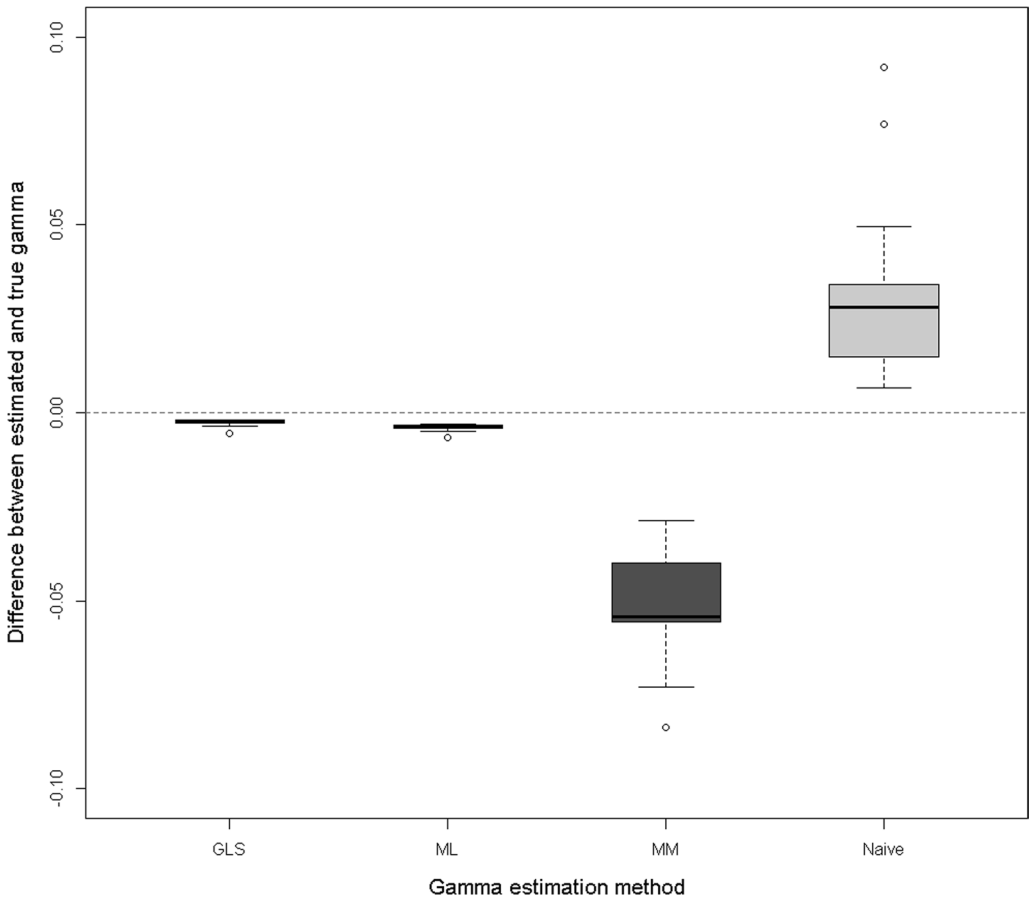

Fig. 1 Differences between estimated and true Gamma, across 20 simulation replicates. Gamma was estimated by generalized least squares (GLS), maximum likelihood (ML), method of moments (MM) and the Naive method

sires, which had an accuracy of essentially 1 and no bias. Depending on the selection scheme, this may lead to less than optimal selection decisions.

Estimation of variance components

Figure 3 shows estimates of heritability obtained with three of the four methods (PBLUP, ssGBLUP_F and ssGBLUP_M). The estimates obtained using ssGBLUP did not converge for six of the 20 simulation replicates. Convergence was achieved in those cases by weighting the submatrix $\mathbf{A}_{22}^{-1}$ in $\mathbf{H}^{-1}$ by $\omega = 0.7$ instead of 1 [42] but poor quality estimates were obtained and are, therefore, not reported.

Estimates were generally lower than the simulated true heritability (0.30). The lowest estimates were obtained with ssGBLUP_F. Including the metafounder improved estimates compared to ssGBLUP_F and reduced variability of estimates compared to PBLUP.

Table 1 Accuracy (correlation between TBV and EBV), inflation (regression coefficient of TBV on EBV), bias [average (EBV–TBV)] and mean square error (MSE) for each prediction methods

Prediction method	Accuracy	Inflation	Bias	MSE
PBLUP	0.51 (0.05)	0.98 (0.06)	−0.0003 (0.03)	0.206 (0.01)
ssGBLUP	0.72 (0.03)	0.89 (0.19)	0.2169 (0.04)	0.159 (0.03)
ssGBLUP_F	0.74 (0.02)	0.99 (0.04)	0.1167 (0.04)	0.141 (0.01)
ssGBLUP_M	0.74 (0.02)	0.94 (0.04)	0.0094 (0.03)	0.125 (0.01)

Averages across 20 replicates with standard deviations in parenthesis

Discussion

In this work, we have addressed the complex issue of conciliation of marker and pedigree information in genetic evaluation. Powell et al. [43] argued that both IBS (at the markers) and identity-by-descent (IBD) are compatible notions because they are both measures of identity at causal genes. However, incompatibility appears when mixing both types of relationships [5, 34, 44, 45]. Legarra [7] suggested that, in order to compare genetic variance across IBD, IBS or other measures of relationships,

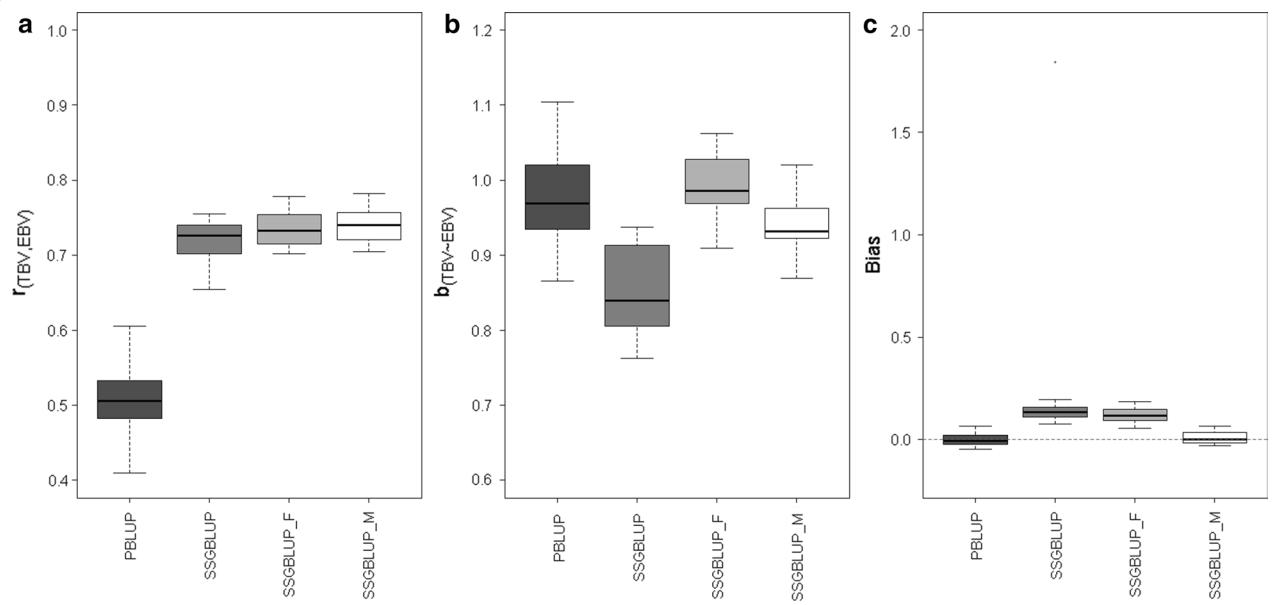

Fig. 2 **a** Correlation of TBV with EVB for each prediction method (accuracy). **b** Regression slope of TBV on EBV (overdispersion). **c** Bias [average (EBV–TBV)]

Table 2 Spearman correlations among TBV and the four EBV for each prediction methods

	EBV PBLUP	EBV ssGBLUP	EBV ssGBLUP_F	EBV ssGBLUP_M
TBV	0.49 (0.06)	0.71 (0.02)	0.72 (0.03)	0.73 (0.02)
EBV PBLUP		0.56 (0.05)	0.62 (0.04)	0.64 (0.04)
EBV ssGBLUP			0.99 (0.01)	0.98 (0.01)
EBV ssGBLUP_F				0.99 (0.002)

Averages across 20 replicates with standard deviations in parenthesis

a common reference must be chosen. Similar (but not identical) to [43], in this work we used a fixed reference ($\mathbf{G}$ constructed as a cross-product of $\{-1,0,1\}$ genotypic codes) and tailored $\mathbf{A}$ (IBD, pedigree) to fit $\mathbf{G}$ (IBS, markers). Compared to previous approaches, using a fixed reference has the advantage that genomic relationships are immutable (i.e. adding more genotyped individuals to the database does not change the existing relationships) and they do not depend on pedigree depth, which by construction is always limited and, in animal breeding, often heterogeneous. Our approach is in fact very similar to using IBS as measure of identity. We used a genomic relationship matrix $\mathbf{G} = 2(\mathbf{M} - \mathbf{J})(\mathbf{M} - \mathbf{J})'/n$, whereas the matrix of IBS is $\mathbf{G}_{IBS} = \mathbf{G}/2 + \mathbf{11}'$ (see proof in "Appendix"). In GBLUP with associated variance component estimation, when all animals are genotyped, using a model with $\mathbf{G}_{IBS}$ or the $\mathbf{G}$ matrix proposed here yields identical EBV, as the term ½ in $\mathbf{G}/2$ gets absorbed into the variance component and the constant $\mathbf{11}'$ gets absorbed into the fixed part of the linear mixed model [7, 46]. However, matrix $\mathbf{G}$ rather than $\mathbf{G}_{IBS}$ must be used in ssGBLUP_M because $\mathbf{G}_{IBS}$ is not compatible with pedigree relationships. In [4], the (fixed effect) intercept term μ_g models, identical to [5], the difference between genetic values of individuals in the base and genotyped individuals. These intercept terms play therefore a similar role as metafounders.

Easy estimation of ancestral relationships

Derivations in the Theory section show that estimation of ancestral relationships based on γ (one base population) and $\mathbf{\Gamma}$ (several base populations) can be framed within the classic linear model approach of quantitative genetics [19], which has recently been used for gene content [14, 25–27]. This approach is easy to understand and compute. Also, $\mathbf{\Gamma}$ can be understood, just like heritability, as an unobserved base population parameter that does not change with additional data (although its estimate may change). Therefore, an accurate estimate of $\mathbf{\Gamma}$ can be used repeatedly without the need for re-estimation, as is customary in livestock genetic evaluation. This contrasts with "centering" of marker covariates, which changes with every new genotype. If all base allele frequencies were known exactly, then there should be no need to use metafounders, as relationship matrices can be appropriately constructed [14].

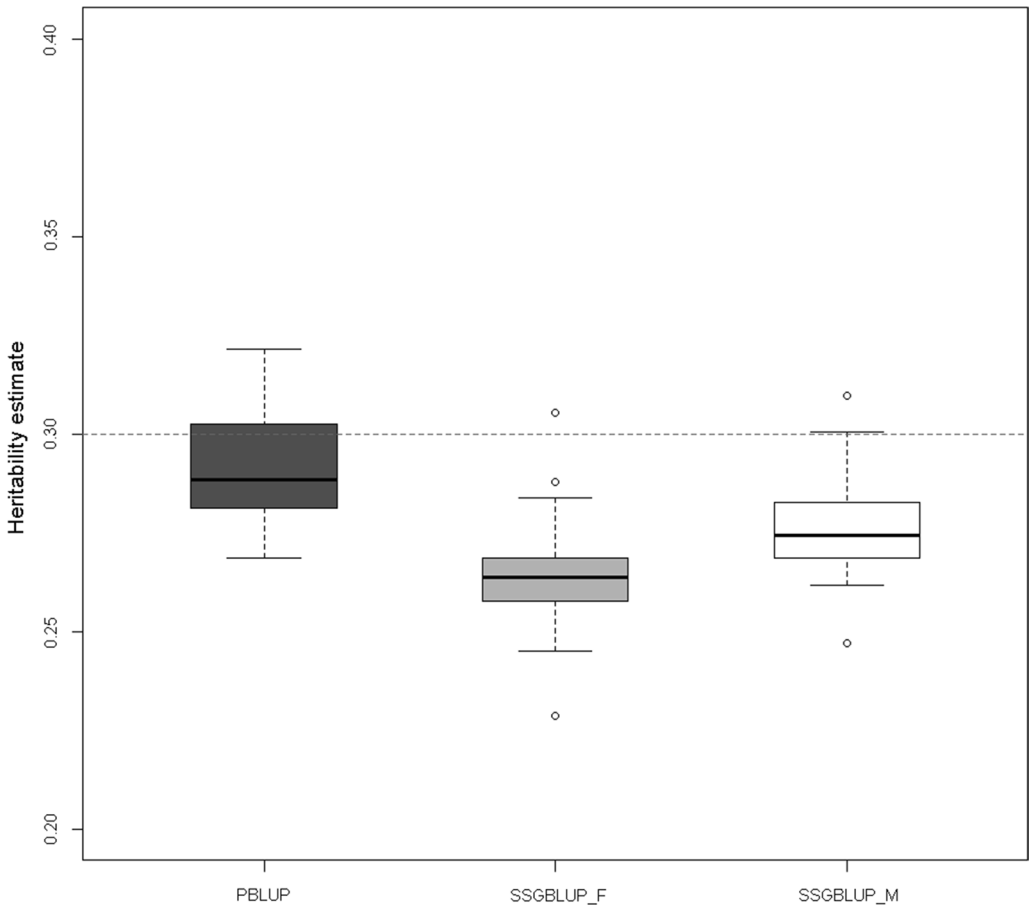

Fig. 3 Estimated heritability for PBLUP, SSGBLUP_F and SSGBLUP_M considering the 20 replicates. The *dotted line* shows the simulated heritability of 0.30

In the work presented here, the simplest methods (Naïve and method of moments) yielded biased (upwards and downwards, respectively) estimates of γ; the naïve method because it ignores that allele frequencies tend to drift to their extreme values as generations progress, and the method of moments because it implicitly assumes that genotyped individuals are a random sample from a particular generation.

Equivalence of ancestral relationships with second moments of allele frequencies also shows a strong relation with population genetics theory, which will be detailed in the next paragraph.

Relationship between metafounders parameter γ and F_{st} fixation index

The fixation index F_{st} [17] is a measure of diversity of a set of populations with respect to a reference population, usually the pool of all populations. In this approach, each population is assumed to be a random sample from all possible populations that could be sampled according to the evolutionary process described by F_{st}. Conceptually, F_{st} is a parameter to be estimated [18, 19], and it is not a statistic computed from the data. The usual definition of F_{st} for a particular biallelic locus is:

$$F_{st} = \frac{\sigma_p^2}{\bar{p}(1-\bar{p})},$$

where σ_p^2 is the variance of allelic frequencies across populations and $\bar{p}$ is the allele frequency of the conceptual combined population. If we consider that the variance of allele frequencies applies *across* loci and not *across* populations, it follows that $\bar{p} = 0.5$ because reference alleles are taken at random. In this case:

$$F_{st} = \frac{\sigma_p^2}{\bar{p}(1-\bar{p})} = \frac{\sigma_p^2}{0.5^2} = 4\sigma_p^2 = \frac{\gamma}{2}.$$

Our interpretation of this link between F_{st} and γ is as follows. Jacquard [47] called $\frac{\gamma}{2}$ the "inbreeding coefficient

of a population". Cockerham [16] modelled $\frac{\gamma}{2} = \theta_l = F_{st}$ as an intraclass correlation, "the coancestry of the line with itself", in other words, the probability that two gametes taken at random from the population are identical. Thus, it makes perfect sense to consider that the additive relationship (which is twice the coancestry value) of a group with itself is $\gamma = 2\theta_l = 8\sigma_p^2$. This is the interpretation of the $\frac{\gamma}{2}$ coefficient in Legarra et al. [1]. Note that the assumption that $\bar{p} = 0.5$ is automatically fulfilled if reference alleles are chosen at random across loci (i.e., they are neither the most frequent nor the least observed allele).

Alternatively, [1] showed that for a population with self-relationships equal to γ, the average heterozygosity is $1 - \frac{\gamma}{2}$, i.e. the variance is reduced by an amount equal to $\frac{\gamma}{2}$ from the conceptual population with heterozygosity 1. Thus $\frac{\gamma}{2}$ can be interpreted as F_{st} if the F_{st} is taken as a measure of homozygosity.

Consequences of using metafounders in genomic evaluation

Genomic estimates of breeding values are invariant to allele coding [46] when all individuals are genotyped. However, this is not the case when pedigree and marker informations are combined, as in ssGBLUP. In this work, we have shown that, even in the presence of complete pedigree and a single base population, use of metafounders in ssGBLUP_M leads to slightly more inflated and less biased EBV, lower MSE, and nearly unbiased estimates of heritability compared to ssGBLUP_F. Bias, defined as E(EBV–TBV), is typically overlooked in genomic prediction, but in an example of biased evaluation, Henderson [48] recognized that "sires of later generations appeared to be under-evaluated relative to older sires". Overdispersion, also called bias in recent literature (e.g. [38]), may also have a dramatic impact in practice [39–41] and the trade-off between bias and variance needs further study. For instance, Vitezica et al. [5] found that ssGBLUP_F was unbiased but had some overdispersion, which likely depends on the data structure, including which individuals are genotyped.

In addition, use of metafounders allows a clear definition of genomic relationships because relationships do not depend on pedigree depth or completeness or on changes in allele frequencies as new data is added. In addition, a high-dimensional parameter (i.e. base allele frequencies) is substituted by a low-dimensional one (matrix $\mathbf{\Gamma}$).

The poor performance of ssGBLUP compared to ssGBLUP_F is likely due to the presence of highly inbred individuals because ssGBLUP ignored inbreeding in the setup of $\mathbf{A}^{-1}$. This relates to the interpretation of parameter ω, as used in early studies of ssGBLUP [42]. An application of ssGBLUP for type traits in Holstein [42] experienced convergence problems, which were eliminated when $\mathbf{A}_{22}^{-1}$ was multiplied by $\omega = 0.7$ and which increased accuracy of predictions. However, the nature of parameter ω was not known [49]. In those studies, the inverse of the numerator relationship matrix $\mathbf{A}^{-1}$ was constructed using Henderson's rules while ignoring inbreeding [36], while the submatrix $\mathbf{A}_{22}^{-1}$ included inbreeding. As a result, the elements in the latter matrix were too large. In addition, genotyped animals were on average unrelated in $\mathbf{G}$ but not in $\mathbf{A}_{22}$, which can be corrected by scaling $\mathbf{G}$, as in [5]. However, this resulted in the elements in $\mathbf{A}_{22}^{-1}$ to be too large for younger animals relative to $\mathbf{G}$. Both these problems are partially circumvented by putting a weight $\omega < 1$ on $\mathbf{A}_{22}^{-1}$. When $\mathbf{A}^{-1}$ was constructed while considering inbreeding, the optimal value of ω in an analysis of Holstein dairy cattle increased from 0.7 to 0.9 (Masuda, personal communication, 2016). However, the metafounder approach provides a more principled solution to this problem. Also, following these experiences, $\mathbf{A}^{-1}$ should always be constructed while considering inbreeding to avoid infrequent but pathological problems.

Conclusions

Metafounders have relationships that are closely related to F_{st} fixation indices and proportional to covariances of allele frequencies in base populations. Use of metafounders can be simplified by new methods to estimate the covariance of base allele frequencies. We verified by simulation of a selected population that, in a single population, both GLS and ML are unbiased and computationally efficient. In the same simulation, use of metafounders in ssGBLUP led to more accurate and less biased evaluations, and also to more accurate estimates of genetic parameters. We propose a genomic relationship matrix that refers to a population with ideal base allele frequencies equal to 0.5. This matrix is similar to an IBS relationship matrix (up to scale factors), does not change with new data, and is compatible with pedigree data if metafounders are used. In the simulated data, pedigrees were perfectly known. Future work with real datasets in more complex settings—purebreds and their crosses [50, 51], and selected populations with unknown parent groups [13] will investigate the feasibility and accuracy, in practice, of using metafounders in ssGBLUP.

Authors' contributions

AL and OFC derived the theory with help from ZGV and CAGB. All authors agreed on scenarios to be tested. CAGB programmed and run all the simulations, with substantial input from IP and IM. The initial version of the manuscript was written by CAGB and AL and then completed by all authors. All authors read and approved the final manuscript.

Author details
[1] Departamento de Producción Animal, Facultad de Agronomía, Universidad de Buenos Aires, C1417DSE Buenos Aires, Argentina. [2] Instituto de Investigaciones en Producción Animal - Consejo Nacional de Investigaciones Científicas y Técnicas, Buenos Aires, Argentina. [3] GenPhySE, INRA, INPT, ENVT, Université de Toulouse, 31326 Castanet-Tolosan, France. [4] Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, 8830 Tjele, Denmark. [5] Animal and Dairy Science, University of Georgia, Athens, GA 30602, USA.

Acknowledgements
We thank S. Boitard and B. Servin for discussions concerning F_{st} and all members of AdMixSel project. We acknowledge suggestions and corrections from two anonymous referees.

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

Availability of data and materials
Software and files are available at https://github.com/alegarra/metafounders.

Funding
CAGB, SML and RJCC were partially funded by grants FONCyT PICT 2013-1661, UBACyT 861/2011 and PIP CONICET 833/2013. This work was partially financed by the AdMixSel project of the INRA SELGEN metaprogram (CAGB, AL and ZGV) as well as INP Toulouse (CAGB, AL). The project was partly supported by the Toulouse Midi-Pyrenees Bioinformatics platform. Reviewers and editors are acknowledged for useful comments and corrections.

Appendix

The appendix contains examples, details and algebraic developments that were not detailed in the main text.

Example of relationship matrix with one metafounder

Consider the pedigree:

1	0	0
2	1	1
3	1	1
4	2	3
5	2	4

where 1 is a metafounder with $\gamma = a_{1,1} = 0.2$. Using the tabular method [15], $a_{2,2} = 1 + a_{1,1}/2 = 1.1$ and $a_{1,2} = 0.5\left(a_{1,sire(2)} + a_{1,dam(2)}\right) = 0.5\left(a_{1,1} + a_{1,1}\right) = 0.2$. Proceeding with the tabular method, $\mathbf{A}^{(\gamma)}$ is:

0.2000	0.2000	0.2000	0.2000	0.2000
0.2000	1.1000	0.2000	0.6500	0.8750
0.2000	0.2000	1.1000	0.6500	0.4250
0.2000	0.6500	0.6500	1.1000	0.8750
0.2000	0.8750	0.4250	0.8750	1.3250

with inverse $\mathbf{A}^{(\gamma)-1}$, that can be obtained by inverting γ and using Henderson's rules [1, 16]:

7.2222	−1.1111	−1.1111	0.0000	0.0000
−1.1111	2.2222	0.5556	−0.5556	−1.1111
−1.1111	0.5556	1.6667	−1.1111	0.0000
0.0000	−0.5556	−1.1111	2.7778	−1.1111
0.0000	−1.1111	0.0000	−1.1111	2.2222

These compare with regular $\mathbf{A}$ that can be obtained by setting $\gamma = 0$. In this case, individual 1 is an unknown parent group and its "relationships" have been set to 0 for presentation:

0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	1.0000	0.0000	0.5000	0.7500
0.0000	0.0000	1.0000	0.5000	0.2500
0.0000	0.5000	0.5000	1.0000	0.7500
0.0000	0.7500	0.2500	0.7500	1.2500

and the inverse relationship matrix including the unknown parent group [13] is $\mathbf{A}^{-1}$:

2.0000	−1.0000	−1.0000	0.0000	0.0000
−1.0000	2.0000	0.5000	−0.5000	−1.0000
−1.0000	0.5000	1.5000	−1.0000	0.0000
0.0000	−0.5000	−1.0000	2.5000	−1.0000
0.0000	−1.0000	0.0000	−1.0000	2.0000

Analytical derivation of γ and s

For a particular population, the genetic variance–covariance structure is a function of two parameters η_1 and η_2 : $\gamma = \frac{4\eta_1}{2\eta_1+\eta_2}$ and $s = n(2\eta_1 + \eta_2)$ (n being the number of markers) which depend on the allelic frequencies Appendix A in [11]. With p_j being the allelic frequencies across the $j = 1 \ldots n$ loci, these parameters do not depend on j and are equal to:

$$\eta_1 = Var\left(p_j\right),$$

$$\eta_2 = E\left(2p_jq_j\right),$$

with $q = 1 - p$.

Now, we use the following developments.

$$E(pq) = E(p(1-p)) = E(p) - E\left(p^2\right). \tag{1}$$

Since we have $Var(p) = E\left(p^2\right) - E(p)^2$, we obtain $E\left(p^2\right) = Var(p) + E(p)^2$. We also have $E(q) = 1 - E(p)$. Substituting $E\left(p^2\right)$ in Eq. (1) gives:

$$\begin{aligned} E(pq) &= E(p) - Var(p) - E(p)^2 \\ &= E(p)(1 - E(p)) - Var(p) = E(p)E(q) - Var(p). \end{aligned}$$

If markers are biallelic and labeled at random $E(p) = E(q) = 0.5$. So the equation above gives $E(pq) = 0.25 - Var(p)$. From this we obtain:

$$2\eta_1 + \eta_2 = 2Var\left(p_j\right) + 0.5 - 2Var\left(p_j\right) = 0.5,$$

and therefore

$$s = n(2\eta_1 + \eta_2) = \frac{n}{2}, \tag{2}$$

or, in other words, s is half the number of markers. Furthermore,

$$\gamma = \frac{4\eta_1}{2\eta_1 + \eta_2} = \frac{4\eta_1}{0.5} = 8Var(p_j) = 8\sigma_p^2, \quad (3)$$

so that γ for a single population is eight times the variance of allelic frequencies at the base population.

Equivalences of genomic relationship matrices

The matrix **G** described in [11] and in this paper can be written as:

$$\mathbf{G} = \frac{2}{n}(\mathbf{M} - \mathbf{J})(\mathbf{M} - \mathbf{J})',$$

where **M** contains genotypes coded as {0,1,2} and **J** is a matrix of 1s. The purpose of this paragraph is to show the linear relationship of this matrix with a matrix describing IBS coefficients, in fact $\mathbf{G}_{IBS} = \frac{1}{2}\mathbf{G} + \mathbf{11}'$. The terms in $\mathbf{G}_{IBS}$ are usually described in terms of identities or countings (i.e. [9, 10, 52]):

$$G_{IBS_{ij}} = \frac{1}{n}\sum_{m=1}^{n} 2\frac{\sum_{k=1}^{2}\sum_{l=1}^{2} I_{kl}}{4},$$

where I_{kl} measures the identity (with value 1 or 0) of allele k in individual i with allele l in individual j, and single-locus identity measures are averaged across k loci. There is an algebraic expression for this "counting". Toro et al. [10] in their Eq. (1) show that, for biallelic markers, for a locus k (omitted in the notation for clarity):

$$f_{M_{ij}} = \frac{m_i}{2}\frac{m_j}{2} + \left(1 - \frac{m_i}{2}\right)\left(1 - \frac{m_j}{2}\right), \quad (4)$$

for coancestry (half relationship) $f_{M_{ij}}$ of individuals i and j, where $m/2$ is the "gene frequency" of the individual (half the gene content (m), i.e. {0,1/2,1} for the three genotypes).

In order to prove $\mathbf{G}_{IBS} = \frac{1}{2}\mathbf{G} + \mathbf{11}'$, first we translate the equation in [10] to the more familiar scale of relationships $g_{IBS_{ij}} = 2f_{M_{ij}}$ and gene contents m. Thus:

$$g_{IBS_{ij}} = 2f_{M_{ij}} = 2\left(\frac{m_i}{2}\frac{m_j}{2} + \left(\frac{2}{2} - \frac{m_i}{2}\right)\left(\frac{2}{2} - \frac{m_j}{2}\right)\right)$$
$$g_{IBS_{ij}} = m_i m_j - m_i - m_j + 2.$$

This expression can be easily verified in a table with the nine possible genotypes:

	AA	Aa	aa
AA	2	1	0
Aa	1	1	1
aa	0	1	2

Also,

$$g_{IBS_{ij}} = m_i m_j - m_i - m_j + 2 = (m_i - 1)(m_j - 1) + 1,$$

which extends to all individuals and averaged across loci can be written as:

$$\mathbf{G}_{IBS} = \frac{1}{n}(\mathbf{M} - \mathbf{J})(\mathbf{M} - \mathbf{J})' + \mathbf{11}'.$$

Thus, matrix $\mathbf{G}_{IBS} = \frac{1}{n}(\mathbf{M} - \mathbf{J})(\mathbf{M} - \mathbf{J})' + \mathbf{11}'$ and because $\mathbf{G} = \frac{2}{n}(\mathbf{M} - \mathbf{J})(\mathbf{M} - \mathbf{J})'$ it follows that $\mathbf{G}_{IBS} = \frac{1}{2}\mathbf{G} + \mathbf{11}'$. The equivalence can also be verified by noting that, for all nine genotypes, the cross-product $(m_i - 1)(m_j - 1)$ in the following table is identical to $g_{IBS_{ij}} - 1$ in the previous table.

	AA	Aa	aa
AA	1	0	−1
Aa	0	0	0
Aa	−1	0	1

Computation of the different H matrices

For ssGBLUP and ssGBLUP_F, matrix $\mathbf{H}^{-1}$ is constructed as follows [2, 3]:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_a^{*-1} - \mathbf{A}_{22}^{-1} \end{pmatrix},$$

with $\mathbf{G}_a^* = 0.95\mathbf{G}_a + 0.05\mathbf{A}_{22} = 0.95(a + b\mathbf{G}) + 0.05\mathbf{A}_{22}$ and $\mathbf{G} = \frac{(\mathbf{M}-\mathbf{P})(\mathbf{M}-\mathbf{P})}{2\sum p_i q_i}$ as in [8], **M** contains genotypes coded as {0,1,2} and **P** contains twice allelic frequencies p_i. These are computed from the observed genotypes so that $2p_i$ is equal to the mean of the i-th column of **M**. Constants a and b are such that the full-matrix and diagonal averages of $\mathbf{G}_a$ and $\mathbf{A}_{22}$ are the same [34] in order to make the two matrices compatible. The use of the weights 0.95 and 0.05 is in order to make $\mathbf{G}_a$ invertible. Matrix $\mathbf{A}^{-1}$ should be constructed using contributions with values described in the table below (i.e. [53]):

No parent known	1
One parent known	$\left(0.75 - \frac{F_{known}}{4}\right)^{-1}$
Two parents known	$\left(0.5 - \frac{F_{sire}}{4} - \frac{F_{dam}}{4}\right)^{-1}$

Or, in a more compact way $\left(0.5 - \frac{F_{sire}}{4} - \frac{F_{dam}}{4}\right)^{-1}$ with $F_{unknown} = -1$.

ssGBLUP uses the defaults in blupf90 suite of programs (random_type *add_animal*). ssGBLUP uses the simple method to create $\mathbf{A}^{-1}$, a method which pretends that, in all cases, inbreeding in expressions above is $F = 0$.

ssGBLUP_F uses $\mathbf{H}^{-1}$ as above but constructs $\mathbf{A}^{-1}$ correctly (blupf90 random_type *add_an_upginb*), using the rules above.

ssGBLUP_M uses the blupf90 random_type *user_file* to consider the following relationship matrix, which was constructed externally:

$$\mathbf{H}^{(\Gamma)-1} = k\left(\mathbf{A}^{(\Gamma)-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{*-1} - \mathbf{A}_{22}^{(\Gamma)-1} \end{pmatrix},\right)$$

with $\mathbf{G}^* = 0.95\mathbf{G} + 0.05\mathbf{A}_{22}^{(\Gamma)}$ (basically to make $\boldsymbol{G}$ invertible), $\mathbf{G} = \frac{1}{s}(\mathbf{M} - \mathbf{J})(\mathbf{M} - \mathbf{J})'$ and $s = n/2$, $\mathbf{M}$ contains genotypes coded as {0,1,2}, n is the number of markers, $\mathbf{A}^{(\Gamma)-1}$ and $\mathbf{A}_{22}^{(\Gamma)-1}$ are constructed with own programs as in [1] using the estimated value of $\boldsymbol{\Gamma}$. Inbreeding is fully considered in both matrices $\mathbf{A}^{(\Gamma)-1}$ and $\mathbf{A}_{22}^{(\Gamma)-1}$. The constant $k = 1 - \frac{\gamma}{2}$ puts the genetic variance associated to metafounders (i.e. to "related" founders) on the same scale as regular "unrelated" founders in $\mathbf{A}$ or $\mathbf{H}$ [1].

Received: 25 October 2016 Accepted: 3 March 2017
Published online: 10 March 2017

References

1. Legarra A, Christensen OF, Vitezica ZG, Aguilar I, Misztal I. Ancestral relationships using metafounders: finite ancestral populations and across population relationships. Genetics. 2015;200:455–68.
2. Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic information. J Dairy Sci. 2009;92:4656–63.
3. Christensen OF, Lund MS. Genomic prediction when some animals are not genotyped. Genet Sel Evol. 2010;42:2.
4. Fernando RL, Dekkers JC, Garrick DJ. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genet Sel Evol. 2014;46:50.
5. Vitezica Z, Aguilar I, Misztal I, Legarra A. Bias in genomic predictions for populations under selection. Genet Res (Camb). 2011;93:357–66.
6. Christensen OF, Legarra A, Lund MS, Su G. Genetic evaluation for three-way crossbreeding. Genet Sel Evol. 2015;47:98.
7. Legarra A. Comparing estimates of genetic variance across different relationship models. Theor Popul Biol. 2016;107:26–30.
8. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
9. Ritland K. Estimators for pairwise relatedness and individual inbreeding coefficients. Genet Res (Camb). 1996;67:175–85.
10. Toro MÁ, García-Cortés LA, Legarra A. A note on the rationale for estimating genealogical coancestry from molecular markers. Genet Sel Evol. 2011;43:27.
11. Christensen OF. Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. Genet Sel Evol. 2012;44:37.
12. Thompson R. Sire evaluation. Biometrics. 1979;35:339–53.
13. Quaas RL. Additive genetic model with groups and relationships. J Dairy Sci. 1988;71:1338–45.
14. Makgahlela ML, Strandén I, Nielsen US, Sillanpää MJ, Mäntysaari EA. Using the unified relationship matrix adjusted by breed-wise allele frequencies in genomic evaluation of a multibreed population. J Dairy Sci. 2014;97:1117–27.
15. Emik LO, Terrill CE. Systematic procedures for calculating inbreeding coefficients. J Hered. 1949;40:51–5.
16. Henderson CR. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics. 1976;32:69–83.
17. Wright S. Isolation by distance. Genetics. 1943;28:114–38.
18. Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nat Rev Genet. 2009;10:639–50.
19. Cockerham CC. Variance of gene frequencies. Evolution. 1969;23:72–84.
20. Wright S. Evolution in Mendelian populations. Genetics. 1931;16:97–159.
21. Crow J, Kimura M. An introduction to population genetics theory. New York: Harper and Row; 1970.
22. Robertson A. Gene frequency distributions as a test of selective neutrality. Genetics. 1975;81:775–85.
23. Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. Genetics. 2013;193:929–41.
24. Laval G, SanCristobal M, Chevalet C. Measuring genetic distances between breeds: use of some distances in various short term evolution models. Genet Sel Evol. 2002;34:481–508.
25. McPeek MS, Wu X, Ober C. Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics. 2004;60:359–67.
26. Gengler N, Mayeres P, Szydlowski M. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. Animal. 2007;1:21–8.
27. Forneris NS, Legarra A, Vitezica ZG, Tsuruta S, Aguilar I, Misztal I, et al. Quality control of genotypes using heritability estimates of gene content at the marker. Genetics. 2015;199:675–81.
28. Mäntysaari E, Van Vleck LD. Restricted maximum likelihood estimates of variance components from multitrait sire models with large number of fixed effects. J Anim Breed Genet. 1989;106:409–22.
29. Garcia-Cortes LA, Toro M. Multibreed analysis by splitting the breeding values. Genet Sel Evol. 2006;38:601–15.
30. Sargolzaei M, Schenkel FS. QMSim: a large-scale genome simulator for livestock. Bioinformatics. 2009;25:680–1.
31. Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE. Genomic selection using different marker types and densities. J Anim Sci. 2008;86:2447–54.
32. Hickey JM, Gorjanc G. Simulated data for genomic selection and genome-wide association studies using a combination of coalescent and gene drop methods. G3 (Bethesda). 2012;2:425–7.
33. MacLeod IM, Hayes BJ, Goddard ME. The effects of demography and long-term selection on the accuracy of genomic prediction with sequence data. Genetics. 2014;198:1671–84.
34. Christensen O, Madsen P, Nielsen B, Ostersen T, Su G. Single-step methods for genomic evaluation in pigs. Animal. 2012;6:1565–71.
35. Masuda Y, Misztal I, Tsuruta S, Legarra A, Aguilar I, Lourenco DAL, et al. Implementation of genomic recursions in single-step genomic best linear unbiased predictor for US Holsteins with a large number of genotyped animals. J Dairy Sci. 2016;99:1968–74.
36. Mehrabani-Yeganeh H, Gibson JP, Schaeffer LR. Including coefficients of inbreeding in BLUP evaluation and its effect on response to selection. J Anim Breed Genet. 2000;117:145–51.
37. Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee DH. BLUPF90 and related programs (BGF90). In: Proceedings of the 7th World Congress on Genetics Applied to Livestock Production, 19–23 Aug 2002, Montpellier. 2002. CD-ROM communication no. 28-07.
38. Mantysaari E, Liu Z, VanRaden P. Interbull validation test for genomic evaluations. Interbull Bull. 2010;41:17–22.
39. Sargolzaei M, Chesnais J, Schenkel FS. Assessing the bias in top GPA bulls. 2012. cgil.uoguelph.ca/dcbgc/Agenda1209/DCBGC1209_Bias_Mehdi.pdf. Accessed 21 July 2016.
40. Spelman RJ, Arias J, Keehan MD, Obolonkin V, Winkelman AM, Johnson DL, et al. Application of genomic selection in the New Zealand dairy cattle industry. In: Proceedings of the 9th World Congress on Genetics Applied to Livestock Production, 1–6 Aug 2010, Leipzig. 2010.
41. Winkelman AM, Johnson DL, Harris BL. Application of genomic evaluation to dairy cattle in New Zealand. J Dairy Sci. 2015;98:659–75.
42. Tsuruta S, Misztal I, Aguilar I, Lawlor T. Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. J Dairy Sci. 2011;94:4198–204.
43. Powell JE, Visscher PM, Goddard ME. Reconciling the analysis of IBD and IBS in complex trait studies. Nat Rev Genet. 2010;11:800–5.
44. Harris BL, Johnson DL. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. J Dairy Sci. 2010;93:1243–52.
45. Meuwissen THE, Luan T, Woolliams JA. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. J Anim Breed Genet. 2011;128:429–39.

46. Strandén I, Christensen OF. Allele coding in genomic evaluation. Genet Sel Evol. 2011;43:25.
47. Jacquard A. The genetic structure of populations. Berlin: Springer; 1974.
48. Henderson C. Sire evaluation and genetic trends. J Anim Sci. 1973: symposium 10-41. doi:10.2527/1973.1973Symposium10x.
49. Misztal I, Vitezica ZG, Legarra A, Aguilar I, Swan AA. Unknown-parent groups in single-step genomic evaluation. J Anim Breed Genet. 2013;130:252–8.
50. Christensen OF, Madsen P, Nielsen B, Su G. Genomic evaluation of both purebred and crossbred performances. Genet Sel Evol. 2014;46:23.
51. Lourenco DAL, Tsuruta S, Fragomeni BO, Chen CY, Herring WO, Misztal I. Crossbreed evaluations in single-step genomic best linear unbiased predictor using adjusted realized relationship matrices. J Anim Sci. 2016;94:909–19.
52. Nejati-Javaremi A, Smith C, Gibson JP. Effect of total allelic relationship on accuracy of evaluation and response to selection. J Anim Sci. 1997;75:1738–45.
53. Meuwissen THE, Luo Z. Computing inbreeding coefficients in large populations. Genet Sel Evol. 1992;24:305–13.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

View publication stats