METAFOUNDERS ARE FST FIXATION INDICES AND REDUCE BIAS IN SINGLE STEP GENOMIC EVALUATIONS

Carolina Andrea Garcia-Baccino[*,§§], Andres Legarra[§,†], Ole F Christensen[**], Ignacy Misztal[‡], Ivan Pocrnic[‡], Zulma G. Vitezica[†,§] and Rodolfo J.C. Cantet[*,§§]

[*]Departamento de Producción Animal, Facultad de Agronomía, Universidad de Buenos Aires, C1417DSE Buenos Aires, Argentina

[§§]Instituto de Investigaciones en Producción Animal - Consejo Nacional de Investigaciones Científicas y Técnicas, Buenos Aires, Argentina

[§]INRA, GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), F-31326 Castanet-Tolosan, France

[†]Université de Toulouse, INP, ENSAT, GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), F-31326 Castanet-Tolosan, France

[‡]Animal and Dairy Science, University of Georgia, 30602 Athens, GA, USA

[**]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, DK-8830 Tjele, Denmark

Running title: metafounders in single step GBLUP

ABSTRACT

BACKGROUND:

Metafounders are pseudo-individuals that condense the genetic heterozygosity and relationships within and across base pedigree populations, i.e. ancestral populations. This work addresses estimation and usefulness of metafounder relationships in Single Step GBLUP.

RESULTS:

We show that the ancestral relationship parameters are proportional to standardized covariances of base allelic frequencies across populations, like Fst fixation indexes. These covariances of base allelic frequencies can be estimated from marker genotypes of related recent individuals, and pedigree. Simple methods for estimation include naïve computation of allele frequencies from marker genotypes or a method of moments equating average pedigree-based and marker-based relationships. Complex methods include generalized least squares or maximum likelihood based on pedigree relationships. To our knowledge, methods to infer F_{st} coefficients and F_{st} differentiation have not been developed for related populations.

A compatible genomic relationship matrix constructed as a crossproduct of {-1,0,1} codes, and equivalent (up to scale factors) to an identity by state relationship matrix at the markers, is derived. Using a simulation with a single population under selection, in which only males and youngest animals were 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 other two (naïve and method of moments) were biased (estimates of 0.43 and 0.35). We also observed that genomic evaluation by Single Step GBLUP using metafounders was less biased in terms of accurate genetic trend (0.01 instead of 0.12 bias), slightly overdispersed (0.94 instead of 0.99) and as accurate (0.74) than the regular Single Step GBLUP. Single Step GBLUP using metafounders also provided consistent estimates of heritability.

CONCLUSIONS:

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

Keywords: BLUP, Fst, relationships, genomic selection

BACGROUND

The concept of metafounders gives a coherent framework for a comprehensive theory of genomic evaluation [1]. Genomic evaluation in agricultural species often implies partially genotyped populations, i.e. some individuals are genotyped, others are not, and phenotypes may be recorded in either of the two subsets. An integrated solution called Single Step has been proposed [2–4]. This solution proposes an integrated relationship matrix

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

with inverse

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

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

Because genotyped animals are not a random sample from the analyzed populations (they are 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]. In the latter case, the random variables induce covariances across individuals, a situation that is referred to as “compatibility” of genomic and pedigree relationships. In fact, compatibility implies comparability of the

77 average breeding value of the base population and of the genetic variance [7] across the
78 different measures of relationships.
79
80 Numerically, the problem shows up as follows. The formulae for matrix $\mathbf{H}$ and its inverse
81 contain $(\mathbf{G} - \mathbf{A}_{22})$ and $(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1})$ (assuming $\mathbf{G}$ is full rank), respectively. This suggests
82 that if **G** and $\mathbf{A}_{22}$ are too different, biases may appear.
83
84 Genomic relationships are usually computed in one of two manners: the "crossproducts"
85 [8] or the "corrected identity by state (IBS)" [9]. Both depend critically on assumed *base*
86 *allelic frequencies* (Toro et al., 2011). However, for most purposes allelic frequencies are
87 not of interest *per se* and can be treated as nuisance parameters to be marginalized.
88 Christensen [10] achieved an algebraic integration of allele frequencies, leading to a very
89 simple covariance structure with allele frequencies in genomic relationships fixed at 0.5
90 (e.g., using genotypes coded as {-1,0,1} in the crossproducts) and a parameter called γ
91 which describes the relationships across founders i.e. $\boldsymbol{A}^{(\gamma)} = \boldsymbol{I}\left(1 - \frac{\gamma}{2}\right) + \mathbf{1}\mathbf{1}'\gamma$ in the base
92 population. A second parameter in Christensen's marginalisation is s, which is a
93 counterpart of the heterozygosity of the markers at the base population. Therefore,
94 instead of inferring (thousands of) base allelic frequencies, inference can be based on two
95 simple parameters γ and s. Both can be estimated maximizing the likelihood of observed
96 genotypes. Also this considers the fact that pedigree depth is arbitrary and mostly based
97 on historical availability of records.
98

Legarra *et al.* [1] showed the equivalence of Christensen's ideas to metafounders: pseudo-individuals that simultaneously consider three ideas: (a) separate means for each base population [4,11], (b) randomness of these separate means [5] and (c) the propagation of the randomness of these means to the progeny [10], while accommodating several populations with complex crosses e.g. [12]. Legarra *et al.* [1] also generalized one relationship across founders (scalar γ) to several relationships across founders in the pedigree, i.e. ancestral relationships (matrix $\boldsymbol{\Gamma}$), and suggested simple methods to estimate them. However, the performance of their model, both for estimation of ancestral relationships and for genomic evaluation, has not been tested so far.

This work has two objectives. The first one is to delve into the structure of the metafounder approach to find an alternative parameterization and estimation of the ancestral relationships. By doing so we find that ancestral relationships are generalizations of Wright's F_{st} fixation index. The second goal is to test, by simulation, (i) methods to estimate ancestral relationship parameters, (ii) the quality of genomic predictions using metafounders and (iii) the quality of variance component estimation. For the second goal, the simulated population is undergoing selection and with a complete pedigree partially genotyped.

METHODS

Relationship between metafounders and allelic frequencies at the base

Single population. Let $\boldsymbol{M}$ be a matrix of genotypes coded as gene content, i.e. {0,1,2} and the genomic relationship matrix $\boldsymbol{G} = (\boldsymbol{M} - \boldsymbol{J})(\boldsymbol{M} - \boldsymbol{J})'/s$ with $\boldsymbol{J}$ a matrix of 1's, with reference alleles taken at random so that for a random locus the expected allelic frequency p is 0.5. [10]. In other words, the matrix $\boldsymbol{Z} = (\boldsymbol{M} - \boldsymbol{J})$ contains values of {-1,0,1} for each genotype. In a single population, let γ be a relationship coefficient across pedigree founders or, equivalently the self-relationship of the metafounder [1,10]. Parameter γ is the relationship coefficient among the founders of a population, so that $\boldsymbol{G} = (\boldsymbol{M} - \boldsymbol{J})(\boldsymbol{M} - \boldsymbol{J})'/s$ is most likely given the observed pedigree. This relationship γ is relative to a population with maximum heterozygosity and it is analogous to an Fst fixation index. The parameter s is a measure of maximum heterozygosity in the population.

Christensen (2012) estimated the two parameters, γ and s using maximum likelihood, whereas Legarra et al. (2015) suggested methods of moments. Closer inspection of Appendix A in Christensen (2012) leads to the following developments (see supplementary material for more details).

The parameter γ is such that $\gamma = \frac{4Var(p_i)}{2Var(p_i)+E(2p_iq_i)}$ with $p_i = 1 - q_i$ the allelic frequency at a random locus i. The parameter $s = n\big(2Var(p_i) + E(2p_iq_i)\big)$ with n being the number of markers. However, $E(2p_iq_i) = 2E(p_i)E(q_i) - 2Var(p_i) = 0.5 - 2Var(p_i)$, where it was used that if alleles are labelled at random across loci then $E(p_i) = E(q_i) = 0.5$. From this it follows that $s = \frac{n}{2}$ and the genomic relationship matrix is $\boldsymbol{G} = 2(\boldsymbol{M} - \boldsymbol{J})(\boldsymbol{M} - \boldsymbol{J})'/n$. Interestingly, this matrix is similar to a matrix of IBS relationships, that can be written

as $\boldsymbol{G}_{IBS} = (\boldsymbol{M} - \boldsymbol{J})(\boldsymbol{M} - \boldsymbol{J})'/n + \boldsymbol{11}'$, so that $\boldsymbol{G}_{IBS} = \frac{1}{2}\boldsymbol{G} + \boldsymbol{11}'$. (See proof in the supplementary Material).

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

$$\gamma = 8\,Var(p_i) = 8\sigma_p^2, \qquad (2)$$

so that γ for a single population is eight times the variance of allelic frequencies at the base population (this variance was described by Cockerham [13]). These equalities were not described in Christensen [10]. We stress that $Var(p_i) = \sigma_p^2$ to imply that σ_p^2 (and γ) is a parameter, the variance of allelic frequencies [10,14–16]. On the other hand, s can be seen as the heterozygosity in the case that all markers had an allelic frequency of 0.5.

Multiple populations. In an analogous manner, the relationship across two metafounders b and b' is

$$\gamma_{b,b'} = 8Cov(p_{b,i}, p_{b',i}) = 8\sigma_{p_b,p_{b'}} \qquad (3)$$

i.e., the covariance across loci between allelic frequencies of two populations b and b'. This is almost tautological: the relationship is the covariance across gene contents at a locus, here applied for populations. Christensen et al. (2015) show this in Appendix A, somehow implicitly. Cockerham [13] and Robertson [17] interpret $4\sigma_{p_b,p_{b'}}$ as the coancestry across two populations and Fariello et al. [18] use $\sigma_{p_b,p_{b'}}$ to describe the divergence of populations. There are several measures of genetic distance between populations (e.g. [19]), and most of them contain a term 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. If base population individuals were genotyped, computing allele frequencies and estimating γ is 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.

1-Assuming no pedigree structure. NAIVE: 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, estimate μ_i as the observed mean of $\boldsymbol{m}_i$, then compute $Var(\hat{\boldsymbol{\mu}})$ 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 $\gamma = 8\hat{\sigma}_p^2 = 2Var(\hat{\boldsymbol{\mu}})$.

2-Considering pedigree structure. At locus i, gene content can be seen as a quantitative trait where the mean of $\boldsymbol{m_i}$ in the base population is $2p_i$, where p_i is the allelic frequency at the base population, and the genetic variance is $2p_iq_i$ [20]. Cockerham (1969) showed that the covariance of gene content of marker i across individuals j and k is a function of relationship $Cov(m_{i,j}, m_{i,k}) = A_{jk}2p_iq_i$. A linear model can therefore be written as:

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

where $\boldsymbol{W}$ is an incidence matrix relating individuals in pedigree to genotypes, and with $\boldsymbol{u}_i$ being the deviation of each individual from the mean μ_i for all individuals (Gengler et al., 2007; Forneris et al., 2015). Assuming multivariate normality:

$$\boldsymbol{\mu} \sim N\left(\mathbf{0}, \boldsymbol{I}\sigma_\mu^2\right)$$

$$\boldsymbol{u}_i \sim N(\mathbf{0}, \boldsymbol{A}(2p_iq_i)) = N\left(\mathbf{0}, \boldsymbol{A}\sigma_{m_i}^2\right)$$

Equivalently, for the set of genotyped individuals (labelled as "2"), $\boldsymbol{u}_{2,i} \sim N(\mathbf{0}, \boldsymbol{A}_{22}(2p_iq_i))$ where $\boldsymbol{A}_{22} = \boldsymbol{WAW}'$ is an additive relationship matrix spanning 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 for each locus separate BLUP (or, equivalently, GLS) estimators of μ_i. One option is to use the complete $\boldsymbol{A}^{-1}$ and mixed model equations [20,21]. Equivalently, the corresponding GLS expression is

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

where $(\mathbf{1}'\boldsymbol{A}_{22}^{-1}\mathbf{1})$ is the sum of elements of $\boldsymbol{A}_{22}^{-1}$, $\sigma_{m_i}^2 = 2p_iq_i$ and $\mathbf{1}'\boldsymbol{A}_{22}^{-1}\boldsymbol{m}_i$ is simply a weighted sum of genotypes. Then, estimate σ_μ^2 as $Var(\hat{\mu})$ and because $p_i = \mu_i/2$, $\hat{\sigma}_p^2 = \sigma_\mu^2/4$, and it follows that $\hat{\gamma} = 2\hat{\sigma}_\mu^2$.

Maximum likelihood (ML). Actually (and more exactly), μ_i can be considered as drawn from a normal distribution, $\boldsymbol{\mu} \sim N\left(\mathbf{0}, \boldsymbol{I}\sigma_\mu^2\right)$. Thus σ_μ^2 is a variance component that can be

estimated by Maximum Likelihood. The equations for given values of σ_μ^2 and $\sigma_{m_i}^2 = 2p_iq_i$ are $(\mathbf{1}'\boldsymbol{A}_{22}^{-1}\sigma_{m_i}^{-2}\mathbf{1} + \sigma_\mu^{-2})\hat{\mu}_i = \mathbf{1}'\boldsymbol{A}_{22}^{-1}\sigma_{m_i}^{-2}\boldsymbol{m}_i$. An Expectation-Maximization scheme [22] is as follows. Pick starting values for $\sigma_\mu^2, \sigma_{m_i}^2$. Iterate until convergence on:

1. For each marker i,
   a. estimate $\hat{\mu}_i = \left(\mathbf{1}'\boldsymbol{A}_{22}^{-1}\sigma_{m_i}^{-2}\mathbf{1} + \sigma_\mu^{-2}\right)^{-1}\mathbf{1}'\boldsymbol{A}_{22}^{-1}\sigma_{m_i}^{-2}\boldsymbol{m}_i$
   b. store $PEV_i(\hat{\mu}_i) = \left(\sigma_\mu^{-2} + \mathbf{1}'\boldsymbol{A}_{22}^{-1}\sigma_{m_i}^{-2}\mathbf{1}\right)^{-1}$
   c. update $\sigma_{m_i}^2$ as $\hat{\sigma}_{m_i}^2 = 2\hat{p}_i\hat{q}_i$ with $\hat{p}_i = \hat{\mu}_i/2$
2. Update σ_μ^2 as $\widehat{\sigma_\mu^2} = \frac{1}{n}\left(\hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}} + \sum PEV_i(\hat{\mu}_i)\right)$, where the second part of the expression corresponds to the trace $Tr(\boldsymbol{IC})$, $\boldsymbol{I}$, the identity matrix, is the relationship across $\boldsymbol{\mu}$ and $\boldsymbol{C}$ is the prediction error covariance matrix of $\hat{\boldsymbol{\mu}}$. As only the diagonal elements of $\boldsymbol{C}$ are needed, the elements $PEV_i(\hat{\mu}_i)$ can be obtained separately from each single locus analysis.

On convergence, the estimate is $\hat{\gamma} = 2\hat{\sigma}_\mu^2$. This gives the same estimate as the method based on a Wishart likelihood function in Christensen (2012) with $s = n/2$ (results not shown).

Estimation in multiple populations.

If t base populations are considered, the variance component σ_μ^2 generalizes to $\boldsymbol{\Sigma}_0$, a $t \times t$ matrix of variances and covariances across means μ_i^b for marker i in population b. Across different populations, $\boldsymbol{\Sigma}_0 = \begin{pmatrix} \sigma^2_{\mu^1\mu^1} & \sigma_{\mu^1\mu^2} & \dots \\ \dots & \sigma^2_{\mu^2\mu^2} & \dots \\ \dots & \dots & \dots \end{pmatrix}$ and $\hat{\boldsymbol{\Gamma}} = 2\hat{\boldsymbol{\Sigma}}_0$.

1-Assuming no pedigree structure. NAIVE If relationships across individuals are ignored:

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

where $\boldsymbol{Q}$ is a matrix allocating individuals to populations and $\boldsymbol{\mu}_i$ is a vector with t elements including each population average. For each locus, $\boldsymbol{\mu}_i$ can be computed using least squares and the covariance matrix of $\boldsymbol{\mu}_i$ across loci gives an estimate of $\hat{\boldsymbol{\Sigma}}_0$.

2-Considering pedigree structure. If there are no crosses, the estimation of allelic frequencies can be split in separate analysis by population b: $\boldsymbol{m}_i^j = \mathbf{1}\mu_i^b + \boldsymbol{W}^b\boldsymbol{u}_i^b + \boldsymbol{e}$ with $\boldsymbol{u}_i^b \sim N(\mathbf{0}, \boldsymbol{A}^b(2p_i(1-p_i))$, and $\boldsymbol{A}^b$ is the matrix of relationships concerning population *b*. Then, $\hat{\boldsymbol{P}}_0$ is estimated as the observed matrix of covariances across loci for estimated $\hat{\mu}_i^b$. If there are crosses, there are two alternatives.

GENERALIZED LEAST SQUARES (GLS). The first alternative, suggested by Forneris et al. (2015) is to use a genetic groups model [11,23], as $\boldsymbol{m}_i = \boldsymbol{Q}\boldsymbol{\mu}_i + \boldsymbol{W}\boldsymbol{u}_i + \boldsymbol{e}$ where $\boldsymbol{Q}_{k,b}$ contains the fraction of ancestry b in individual k. This ignores the fact that the variance of gene content, $(2p_iq_i)$ is different for each breed and cross. The second, and more exact alternative is to use the representation where the breeding values are split into within and across breed components (Garcia-Cortes and Toro, 2006), as

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

with partial relationship matrices for vectors $\boldsymbol{u}^b$, $\boldsymbol{u}^{b,b'}$.

MAXIMUM LIKELIHOOD (ML). Analogously to the single population case, an Expectation-Maximization updated estimate can be obtained using multiple trait formulations [22] 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 current implementation is as follows:

1. For each marker i,
   a. estimate $\widehat{\boldsymbol{\mu}}_i = \left(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{Q}'\boldsymbol{A}_{22}^{-1}\sigma_{m_i}^{-2}\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'\boldsymbol{A}_{22}^{-1}\sigma_{m_i}^{-2}\boldsymbol{m}_i$
   b. store $PEC_i(\widehat{\boldsymbol{\mu}}_i) = \left(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{Q}'\boldsymbol{A}_{22}^{-1}\sigma_{m_i}^{-2}\boldsymbol{Q}\right)^{-1}$
   c. update $\sigma_{m_i}^2$ as $\hat{\sigma}_{m_i}^2 = 2\hat{p}_i^*(1-\hat{p}_i^*)$ with $\hat{p}_i^* = \frac{1}{nb}\sum_{b=1,nb}\frac{\hat{\mu}_i^b}{2}$
2. Update $\boldsymbol{\Sigma}_0$ using crossproducts within and across populations as e.g. with two populations,

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

There is an approximation in (1c) because we assume that $\sigma_{m_i}^2 = 2p_iq_i$ is equal across all base populations. This point will be addressed in future research.

SIMULATION

To assess the quality of genomic predictions using one metafounder, we simulated data using QMSim [24]. The simulation closely followed Vitezica *et al.* (2011) 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 5 generations to an effective population size of 3000. In total 30 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 true γ had a value around 0.40. Potentially, 1500 QTL affected the phenotype; QTL allele effects were sampled from a Gamma distribution with a shape parameter of 0.4. The mutation rate of the markers (recurrent mutation process) and QTL was assumed to be 2.5 $\times 10^{-5}$ per locus per generation (Solberg *et al.*, 2008). A female trait with a heritability of 0.30 was simulated.

Then, 10 overlapping generations of selection followed, where 200 males were mated with 2600 females producing 2600 offspring following a positive assortative mating design. Within the simulation, individuals were selected according to estimated breeding value (EBV) based on pedigree BLUP. In each generation 40% of the males and 20% of the females were replaced by younger and selected individuals. No restrictions were set to avoid or minimize inbreeding, so highly inbred individuals were found, as a result of

extreme selection and matings among highly related individuals. There were 100 individuals (mainly found in the last generation) with an inbreeding coefficient higher than 0.20, with extreme cases (few individuals) with inbreeding coefficients higher than 0.40. True breeding values (TBV) and pedigree information were available for all 10 generations (28,800 individuals in pedigree), phenotypes were available for all females except the last generation (14,300 records). All males (840 sires of females with phenotypic records) were genotyped as well as 2600 individuals in generation 9 (with records) and 2600 in generation 10 (with no records). All in all, 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, one of them including one metafounder.

Methods to estimate Gamma

Parameter γ was estimated using four different estimation methods. First, the NAIVE method which does not consider the pedigree structure. Then, the genealogical information was included in the estimation by three different methods: GLS, ML, and the Method of Moments (MM) presented in Legarra *et al.* (2015). For a single population, the last method involves the estimation of γ based on summary statistics of $\boldsymbol{A}_{22}$ (regular pedigree-relationship matrix for genotyped individuals) and $\boldsymbol{G}$ (the genomic relationship matrix).

Genomic prediction methods

Genetic merit of the selection candidates in generation 10 (genotyped and with no phenotype records) was estimated using four methods. The first one was the pedigree based BLUP (PBLUP) based on phenotype and pedigree information. The second method was Single-Step GBLUP (SSGBLUP) in which genomic information is also taken into account; this method used the correction of [25] and is the default method used in most practical applications [25,26]. However, the implementation that we used does not include inbreeding in the setup of $\boldsymbol{A}^{-1}$ [27], although it does consider it in $\boldsymbol{A}_{22}^{-1}$ (see below for use of these matrices). The third method was Single-Step GBLUP including inbreeding in the setup of $\boldsymbol{A}^{-1}$ and of $\boldsymbol{A}_{22}^{-1}$ (SSGBLUP_F). Finally, the fourth method was SSGBLUP including the metafounder (SSGBLUP_M), using γ estimated by GLS as it turned out to be an accurate method to estimate gamma (see the Results section). The three methods used the following inverse relationship matrices: PBLUP: $\boldsymbol{A}^{-1}$; SSGBLUP: $\boldsymbol{H}^{-1} = \boldsymbol{A}^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & \boldsymbol{G}_a^{-1} - \boldsymbol{A}_{22}^{-1} \end{pmatrix}$ where $\boldsymbol{G}_a$ is as in [25] ; SSGBLUP_M: $\boldsymbol{H}^{(\gamma)-1} = \boldsymbol{A}^{(\gamma)-1} + \begin{pmatrix} 0 & 0 \\ 0 & \boldsymbol{G}^{-1} - \boldsymbol{A}_{22}^{(\gamma)-1} \end{pmatrix}$ where $\boldsymbol{G} = (\boldsymbol{M} - \boldsymbol{J})(\boldsymbol{M} - \boldsymbol{J})'/s$ with $s = n/2$ (see the Methods section) and $\boldsymbol{A}^{(\gamma)}$ is as in [1]. More details are given in Supplementary material. For computation we used blupf90 [28]. In the case of SSGBLUP_M we constructed all relationship matrices with own software, and then used the option user_file in blupf90.

Quality of genomic prediction

Prediction quality was checked for all 2600 selection candidates. 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. Thus, bias is related to estimated genetic progress in the selection candidates. The inflation (often called bias) of the prediction method was quantified by the coefficient of regression of TBV on EBV. These two statistics corresponds to the coefficients b_0 and b_1 in the Interbull validation method [29] which uses the regression $TBV = b_0 + b_1 EBV + e$. The mean square error (MSE) 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 nice statistical properties but also have relevance in livestock selection [30–32]. Ranking changes of the selection candidates were also assessed by calculating the Spearman's rank correlation coefficients between EBVs across methods.

In addition, the quality of variance component estimation was also assessed. For this purpose variance components were estimated using the four methods (PBLUP, SSGBLUP, SSGBLUP_F, SSGBLUP_M) using REML with remlf90 [28].

RESULTS

Estimation of gamma

Figure 1 shows boxplots of the differences between the estimates of γ calculated by four different methods (MM, Naive, ML and GLS) and the true values obtained by simulation, using each of the 20 replicates. The simulations were tailored to produce $\gamma = 0.40$. ML and GLS estimated γ very accurately. The MM clearly underestimated the value of γ, whereas the Naive method overestimated it. Based on these results we used the γ estimated by GLS when using SSGBLUP_M for prediction. The effect of employing different values of γ in the genomic prediction was assessed to quantify its impact in terms of the quality of predictions. Using estimates of γ based on the Method of Moments only slightly changed the results (not shown).

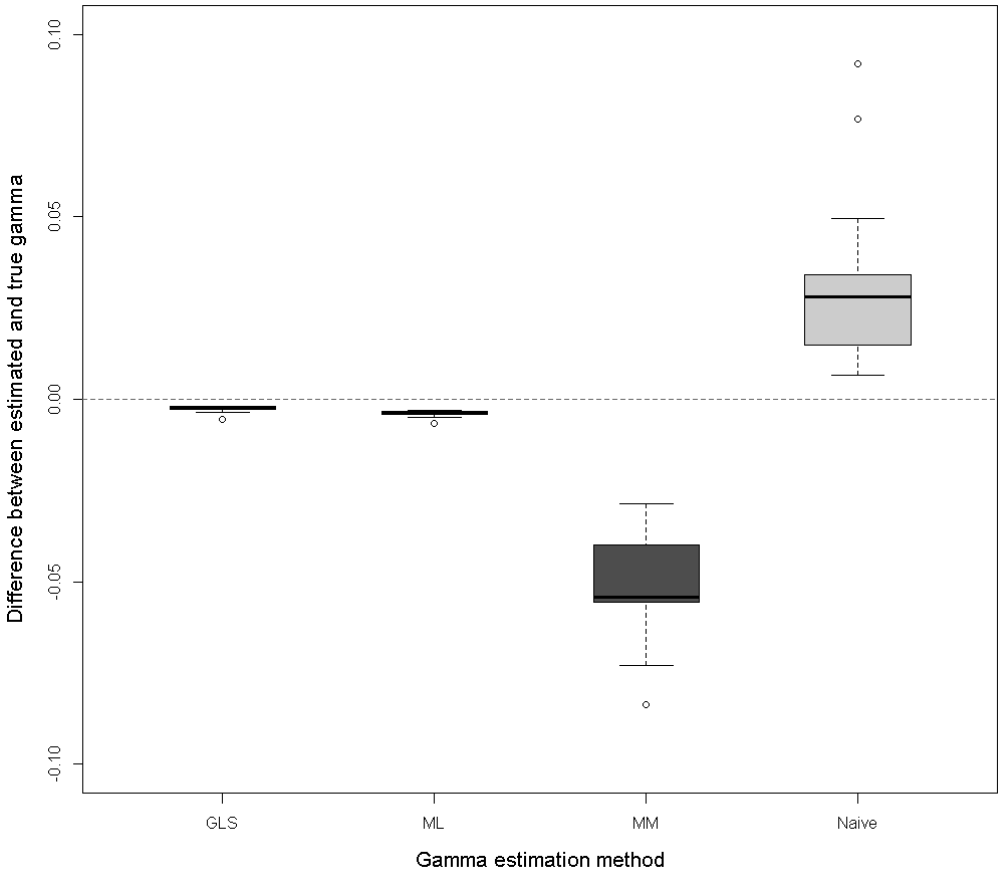


Figure 1 Differences between estimated and true Gamma, across 20 simulation replicates. Gamma was estimated by Generalized Least Squares (GLS), Maximum Likelihood (ML), Method of Moments (MM) and the Naive method.

Quality of genomic prediction

Correlations between TBV and EBV for each of the prediction methods are shown in Table 1 and Figure 2a. Compared with PBLUP, SSGBLUP_F and SSGBLUP_M increased accuracy by approximately 23 absolute points, respectively. 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. SSGBLUP resulted in a small loss of accuracy as compared to SSGBLUP_F and SSGBLUP_M.

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 of the prediction methods. Standard deviations in parenthesis.

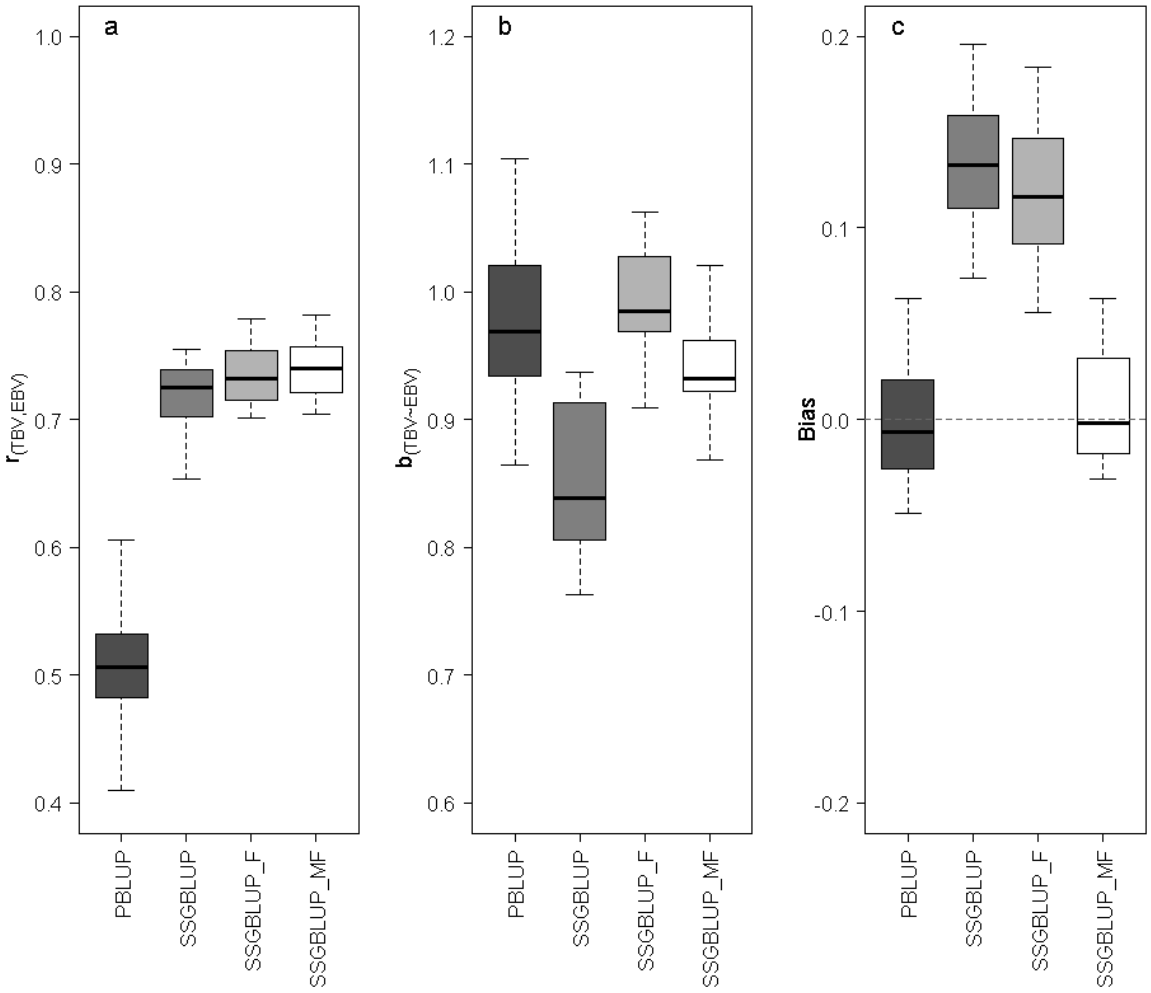
Prediction method	Accuracy	Inflation	Bias	MSE
PBLUP	0.51 (0.05)	0.98 (0.06)	-0.0003 (0.03)	0.206 (0.01)
SSGBLUP	0.72 (0.03)	0.89 (0.19)	0.2169 (0.04)	0.159 (0.03)
SSGBLUP_F	0.74 (0.02)	0.99 (0.04)	0.1167 (0.04)	0.141 (0.01)
SSGBLUP_M	0.74 (0.02)	0.94 (0.04)	0.0094 (0.03)	0.125 (0.01)

Bias values for each prediction method are shown in Table 1 and in Figure 2c. Both PBLUP and SSGBLUP_M were unbiased, whereas SSGBLUP and SSGBLUP_F were biased. Bias in SSGBLUP_F is equivalent to roughly 0.5 generations of genetic improvement or to 0.4 standard genetic deviations.

Table 1 and Figure 2b display the regression coefficient of TBV on EBV. This value measures the inflation degree of each prediction method and should be close to 1. PBLUP and SSGBLUP_F produced the values closest to one. Including genomic data in the prediction using SSGBLUP resulted in regression coefficients lower than one, but including the metafounder in SSGBLUP_M gives values closer to one. SSGBLUP_M and SSGBLUP_F

384 displayed a lower standard deviation compared to the other two methods. Again, SSGBLUP
385 showed the highest variability. SSGBLUP_M displayed the lowest MSE (closer to zero),
386 followed by SSGBLUP_F (Table 1).

387

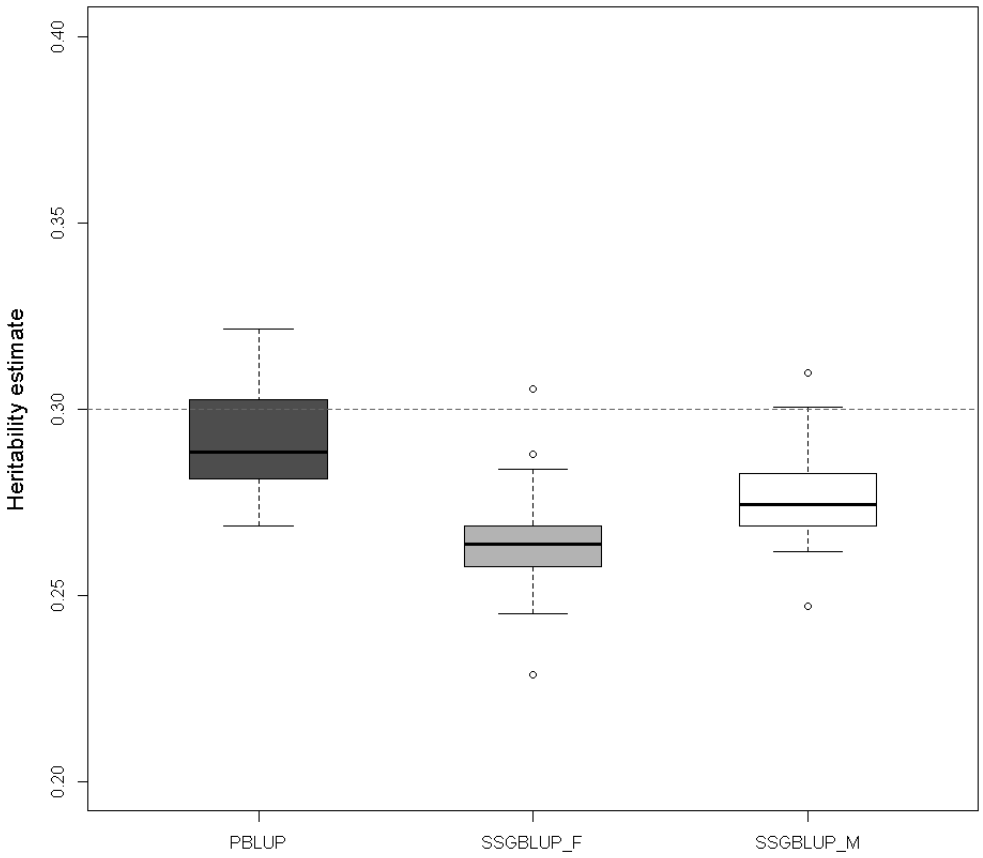


388

389 **Figure 2. a.** Correlation of TBV on EVB for each prediction method (accuracy). **b.**
390 Regression slope of TBV on EBV (overdispersion). **c.** Bias (average (EBV-TBV)).
391

392 **Variance components estimation**

393 Figure 3 shows the estimates of heritability obtained in three of the four methods assed
394 (PBLUP, SSGBLUP_F and SSGBLUP_M). The estimates obtained using SSGBLUP are not
395 displayed in Figure 3 because in 6 out of the 20 simulation replicates EM-REML did not
396 converge. Convergence was achieved in those cases by weighting the submatrix $\boldsymbol{A}_{22}^{-1}$ in
397 $\boldsymbol{H}^{-1}$ by $\omega = 0.7$ instead of 1 [33] but poor quality estimates were obtained and they are
398 not reported.
399
400 When comparing with the simulated true heritability value (0.30) the scenarios displayed
401 in general lower estimates. The lowest estimates were obtained using SSGBLUP_F.
402 Including the metafounder improved estimates compared to SSGBLUP_F and reduced
403 variability when comparing to PBLUP.
404



405

406 **Figure 3** Estimated heritability for PBLUP, SSGBLUP_F and SSGBLUP_M considering the 20
407 replicates. The dotted line shows the simulated heritability of 0.30.
408

409 **Ranking**

410 The methods were also compared based on ranking correlations of EBVs with TBV and
411 across methods. A rank correlation of 1 implies that the same candidates are selected.
412 Results are in Table 2. Rank correlations with TBV are similar to accuracies in Table 1.
413 Selection decisions are only slightly different using SSGBLUP, SSGBLUP_F or SSGBLUP_M.
414 Note however, that this table does not address the comparison across generations (e.g.
415 old vs. young animals), which is sensitive to biases reflected in Table 1 [32].

Table 2 Spearman correlation among TBV and the four EBV for each of the prediction methods. Standard deviations in parenthesis.

	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)

DISCUSSION

In this work, we have addressed the complex issue of conciliation of marker and pedigree information. Powell et al. [34] argued that both IBS (at the markers) and IBD are measures of identity at causal genes and they are compatible notions. However, the incompatibility issue appears when mixing both kind of relationships [5,25,35,36]. Legarra [7] established how to solve the issue of comparing genetic variance across IBD, IBS or other measures of relationships. In this work, we have used, similar (but not identical) to Powell et al. [34], a fixed reference ($\boldsymbol{G}$ constructed as a crossproduct of $\{-1,0,1\}$ genotypic codes) and tailored $\boldsymbol{A}$ (IBD, pedigree) to fit $\boldsymbol{G}$ (IBS, markers). Using a fixed reference has the advantage, compared to previous approaches, that genomic relationships are immutable (adding more genotypes to the database does not change the existing relationships) and they are unconditional on pedigree depth, that by construction is always limited and, in animal breeding, often heterogeneous. Our approach is in fact very similar to considering, as measures of identity, plain IBS. We use a matrix $\boldsymbol{G} = 2(\boldsymbol{M} - \boldsymbol{J})(\boldsymbol{M} - \boldsymbol{J})'/n$, whereas a

433 matrix of IBS, or molecular, relationships is $\boldsymbol{G}_{IBS} = \boldsymbol{G}/2 + \mathbf{11}'$ (see proof at the
434 supplementary material). In a GBLUP context when all animals are genotyped, using a
435 model with IBS coefficients yields identical results as the term ½ gets absorbed into the
436 variance component and the constant $\mathbf{11}'$ gets absorbed into the fixed part of the linear
437 mixed model [7,37]. However, the matrix that must be used in SSGBLUP_M is $\boldsymbol{G}$ and not
438 $\boldsymbol{G}_{IBS}$, because $\boldsymbol{G}_{IBS}$ is not compatible with pedigree relationships.

439

440 **Easy estimation of ancestral relationships**

441 The derivations in the THEORY section show that estimation of ancestral relationships in γ
442 (one base population) and $\boldsymbol{\Gamma}$ (several base populations) may be framed within the linear
443 model approach that is classical in quantitative genetics [13], and recently used for gene
444 content [12,20,21]. These methods are easy to understand and to compute. Also, $\boldsymbol{\Gamma}$ can be
445 understood, just like heritability, as an unobserved base population parameter that does
446 not change with additional data (although its estimate may change). Therefore, an
447 accurate estimate of $\boldsymbol{\Gamma}$ can be used repeatedly without the need of re-estimation, as is
448 customary in livestock genetic evaluations. This contrasts with "centering" of marker
449 covariates, which changes with every new genotype.

450

451 In the current research, the simplest methods (Naive and Method of Moments) yielded
452 biased (upwards and downwards respectively) estimates of γ; for the first method because
453 it ignores that allele frequencies drift to the extremes as generations go, and for the

second because it implicitly assumes that individuals genotyped are a random sample from a particular generation when in fact they are not.

In addition, the equivalence of ancestral relationships with second moments of allele frequencies shows a strong relation with populations genetics theory, which will be detailed in the next paragraph.

Relationship between metafounders γ and F_{st} fixation index

The fixation index F_{st} [38] is a measure of diversity of a set of populations with respect to a reference population, usually the pool of all populations. In this view, each population is 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 [13,39], and it is not a statistic computed from the data. A 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 allelic frequency of the conceptual combined population. If we consider that the variance of allelic frequencies applies *across* loci and not *across* populations, it follows naturally that $\bar{p} = 0.5$. 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 is as follows. Jacquard (1974) called $\frac{\gamma}{2}$ the "inbreeding coefficient of a population". Cockerham (1969) 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 line 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 $\bar{p} = 0.5$ is automatically fulfilled if reference alleles are labelled randomly across loci (i.e., they are neither the most frequent nor the least observed).

Alternatively, Legarra et al. (2015) showed that for a population with self-relationship of γ, the average heterozygosity was $1 - \frac{\gamma}{2} = 1 - \theta$, i.e. the variance is reduced by an amount of θ from the conceptual population with heterozygosity 1. Thus $\frac{\gamma}{2}$ can be interpreted as F_{st} if the latter is taken as a measure of homozygosity.

Consequences of using metafounders in genomic evaluation

Genomic estimates of breeding values are invariant to allele coding [37] when all individuals are genotyped. However, this is not the case when pedigree and marker information are combined as in SSGBLUP. In this work we have shown that, even in presence of complete pedigree and a single base population, use of metafounders in SSGBLUP_M leads to slightly more inflated, less biased EBVs, lower MSE and nearly unbiased estimates of heritability compared to SSGBLUP_F. Bias, defined as E(EBV-TBV)), is typically overlooked in genomic predictions, but in an example of biased evaluation "sires

of later generations appeared to be under-evaluated relative to older sires" [40]. Overdispersion, also called bias in recent literature (e.g. Mantyssari et al., 2010), may have dramatical impact as well [30–32]. The trade-off between bias and variance needs further studies. For instance, [5] found that SSGBLUP_F was unbiased but had some overdispersion; this is likely dependent on the data structure, including the genotyping.

In addition, use of metafounders allows a clear definition of genomic relationships. With this definition, relationships are not dependent on pedigree depth or completeness, and are not dependent on allelic frequencies subject to change with arrival of new data. Additionally, a high dimensional parameter (-base- allele frequencies) is substituted by a low-dimensional one (matrix $\boldsymbol{\Gamma}$).

The poor performance of SSGBLUP as compared to SSGBLUP_F (the former ignoring inbreeding in the set up of $\mathbf{A}^{-1}$) is likely due to the presence of highly inbred individuals. This relates to the interpretation of an ω parameter used in early studies of SSGBLUP. An application of SSGBLUP for type traits in Holstein [33] experienced convergence problems. The authors found that by multiplying $\mathbf{A}_{22}^{-1}$ by a $\omega = 0.7$ eliminated convergence problems and increased accuracy. However, the nature of that parameter was not known, e.g. Misztal et al. [41]. In those studies, the inverse of the numerator relationship matrix $\boldsymbol{A}^{-1}$ was constructed using Henderson's rules while ignoring inbreeding [27], while the submatrix $\boldsymbol{A}_{22}^{-1}$ included inbreeding. Subsequently, the elements in the latter were too

large. In addition, genotyped animals were on average unrelated in $\boldsymbol{G}$ but not in $\boldsymbol{A}_{22}$, which is corrected by scaling $\boldsymbol{G}$ as in Vitezica et al. (2011). But then, in $\boldsymbol{A}_{22}^{-1}$ the elements were too large for younger animals relative to $\boldsymbol{G}$. Both problems are partially circumvented but putting a weight $\omega < 1$ on $\boldsymbol{A}_{22}^{-1}$. When $\boldsymbol{A}^{-1}$ was constructed considering inbreeding, the optimal ω coefficient in an analysis of Holstein dairy cattle increased from 0.7 to 0.9 (Masuda, personal communication, 2016). However, the metafounder approach provides a clean solution to this problem. Also, following these experiences, $\boldsymbol{A}^{-1}$ should always be constructed considering inbreeding to avoid pathological problems.

CONCLUSION

Metafounders are similar to F_{st} fixation indices and proportional to covariances of allelic frequencies in base populations. Use of metafounders is simplified by new methods (GLS and maximum likelihood) 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 Single Step GBLUP leads 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 frequencies 0.5. This matrix is similar to an IBS relationship matrix (up to scale factors),

539 does not change with new data and is compatible with pedigree data if metafounders are
540 used.

541

542 In this simulated data, pedigrees are perfectly known. Future work with real data sets in
543 more complex settings - purebreds and their crosses [42,43], and selected populations
544 with unknown parent groups [11] will investigate the feasibility and accuracy in practice of
545 using metafounders on Single Step GBLUP.

546

547 APPENDIX

548 This Appendix contains several algebraic developments not detailed in the main text.

549 **Analytical derivation of γ and s**

550 For a particular population, the genetic variance-covariance structure is a function of two
551 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)
552 which depend on the allelic frequencies (Christensen 2012), Appendix A. With p_j being the
553 allelic frequencies across the $j = 1..n$ loci, these parameters do not depend on j and are
554 equal to

555 $$\eta_1 = Var(p_j)$$

556 $$\eta_2 = E\left(2p_j q_j\right)$$

557 with $q = 1 - p$.

558 Now use is made of the following developments.

559 $$E(pq) = E(p(1-p)) = E(p) - E(p^2). \qquad \text{(A1)}$$

Since we have that $Var(p) = E(p^2) - E(p)^2$ we obtain $E(p^2) = Var(p) + E(p)^2$. We also have $E(q) = 1 - E(p)$. Substituting $E(p)^2$ in (A1) gives

$$E(pq) = E(p) - Var(p) - E(p)^2 = E(p)\big(1 - E(p)\big) - Var(p) = E(p)E(q) - Var(p).$$

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

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

and therefore

$$s = n\,(2\eta_1 + \eta_2) = \frac{n}{2}, \qquad (1)$$

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

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

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 Christensen (2012) and in this paper can be written as $\boldsymbol{G} = \frac{2}{n}(\boldsymbol{M} - \boldsymbol{J})(\boldsymbol{M} - \boldsymbol{J})'$, where **M** contains genotypes coded as {0,1,2} and $\boldsymbol{J}$ is a matrix of 1's. The purpose of this paragraph is to show the linear relationship of this matrix with a matrix describing identity by state coefficients (IBS), in fact $\boldsymbol{G}_{IBS} = \frac{1}{2}\boldsymbol{G} + \boldsymbol{11}'$. The terms in $\boldsymbol{G}_{IBS}$ are usually described in terms of identities or countings (i.e. Ritland, 1996; Toro et al., 2011; Nejati-Javaremi et al., 1997):

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

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 n loci.

There is an algebraic expression for this "counting". Toro et al. (2011) expression (1), show that for biallelic markers, for a locus k (omitted in the notation for clarity):

$$f_{M_{ij}} = \frac{m_i}{2}\frac{m_j}{2} + \left(1 - \frac{m_i}{2}\right)\left(1 - \frac{m_j}{2}\right) \qquad (3)$$

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

In order to prove $\boldsymbol{G}_{IBS} = \frac{1}{2}\boldsymbol{G} + \mathbf{11}'$, first we translate the Toro et al. (2011) equation 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

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

Thus, matrix $\boldsymbol{G}_{IBS} = \frac{1}{n}(\boldsymbol{M}-\boldsymbol{J})(\boldsymbol{M}-\boldsymbol{J})' + \boldsymbol{11}'$ and because $\boldsymbol{G} = \frac{2}{n}(\boldsymbol{M}-\boldsymbol{J})(\boldsymbol{M}-\boldsymbol{J})'$ it follows that $\boldsymbol{G}_{IBS} = \frac{1}{2}\boldsymbol{G} + \boldsymbol{11}'$. 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 $\boldsymbol{H}^{-1}$ is constructed as follows:

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

with $\boldsymbol{G}_a^* = 0.95\boldsymbol{G}_a + 0.05\boldsymbol{A}_{22} = 0.95(a + b\boldsymbol{G}) + 0.05\boldsymbol{A}_{22}$, and $\boldsymbol{G} = \frac{(\boldsymbol{M}-\boldsymbol{P})(\boldsymbol{M}-\boldsymbol{P})}{2\sum p_i q_i}$ as in VanRaden (2008), **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 the mean of the i-th column of **M**. Constants a and b are such that the full-matrix and diagonal averages of $\boldsymbol{G}_a$ and $\boldsymbol{A}_{22}$ are the same (Christensen et al., 2012) in order to make the two matrices compatible. The use of the weights 0.95 and 0.05 is in order to make $\boldsymbol{G}_a$

613 invertible. Matrix $\boldsymbol{A}^{-1}$ should be constructed using contributions with values described in

614 the Table below (i.e. Meuwissen and Luo, 1992):

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

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

616 SSGBLUP uses the defaults in blupf90 suite of programs (random_type *add_animal*).

617 SSGBLUP uses the simple method to create $\boldsymbol{A}^{-1}$, method which pretends that in all cases

618 inbreeding in expressions above is $F = 0$.

619 SSGBLUP_F uses $\boldsymbol{H}^{-1}$ as above but constructs $\boldsymbol{A}^{-1}$ correctly (blupf90 random_type

620 *add_an_upginb*), using the rules above.

621 SSGBLUP_M uses the blupf90 random_type *user_file* to consider the following

622 relationship matrix:

623 $$\boldsymbol{H}^{(\Gamma)-1} = \boldsymbol{A}^{(\Gamma)-1} + \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}^{*} - \boldsymbol{A}_{22}^{(\Gamma)-1} \end{pmatrix}$$

624 with $\boldsymbol{G}^{*} = 0.95\boldsymbol{G} + 0.05\boldsymbol{A}_{22}^{(\Gamma)}$ (basically to make $\boldsymbol{G}$ invertible), $\boldsymbol{G} = \frac{1}{s}(\boldsymbol{M} - \boldsymbol{J})(\boldsymbol{M} - \boldsymbol{J})'$ and

625 $s = n/2$, **M** contains genotypes coded as {0,1,2}, n is the number of markers, $\boldsymbol{A}^{(\Gamma)-1}$ and

626 $\boldsymbol{A}_{22}^{(\Gamma)-1}$ are constructed with own programs as in Legarra et al. (2015)using the estimated

627 value of $\boldsymbol{\Gamma}$. Inbreeding is fully considered in both matrices.

628

629

630

631 DECLARATIONS

632 Availability of data and materials: Software and files are avaible at

633 https://github.com/alegarra/metafounders .

634

635 Competing interests: The authors declare that they have no competing interests

636

637 Funding: CAGB, SML and RJCC were partially funded by grants FONCyT PICT 2013-1661,

638 UBACyT 861/2011 and PIP CONICET 833/2013. This work was partially financed by the

639 AdMixSel project of the INRA SELGEN metaprogram (CAGB, AL and ZGV) as well as INP

640 Toulouse (CAGB, AL). The project was partly supported by the Toulouse Midi-Pyrenees

641 Bioinformatics platform.

642

643 Authors contribution: AL and OFC derived the theory with help from ZGV and CAGB. All

644 authors agreed on scenarios to be tested. CAGB programmed and run all the simulations,

645 with substantial input from IP and IM. The initial version of the manuscript was written by

646 CAGB and AL and then completed by all authors.

647 Acknowledgements: we thank S Boitard and B Servin for discussions concerning Fst and all

648 members of AdMixSel project.

649

650

651 **REFERENCES**

652 1. Legarra A, Christensen OF, Vitezica ZG, Aguilar I, Misztal I. Ancestral Relationships Using
653 Metafounders: Finite Ancestral Populations and Across Population Relationships.
654 Genetics. 2015;200:455–68.
655 2. Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic
656 information. J Dairy Sci. 2009;92:4656–63.
657 3. Christensen OF, Lund MS. Genomic prediction when some animals are not genotyped.
658 Genet Sel Evol. 2010;42:2.
659 4. Fernando RL, Dekkers JC, Garrick DJ. A class of Bayesian methods to combine large
660 numbers of genotyped and non-genotyped animals for whole-genome analyses. Genet Sel
661 Evol. 2014;46:50.
662 5. Vitezica Z, Aguilar I, Misztal I, Legarra A. Bias in genomic predictions for populations
663 under selection. Genet. Res. 2011;93:357–66.
664 6. Christensen OF, Legarra A, Lund MS, Su G. Genetic evaluation for three-way
665 crossbreeding. Genet. Sel. Evol. 2015;47:98.
666 7. Legarra A. Comparing estimates of genetic variance across different relationship
667 models. Theor. Popul. Biol. 2016;107:26–30.
668 8. VanRaden PM. Efficient Methods to Compute Genomic Predictions. J Dairy Sci.
669 2008;91:4414–23.
670 9. Ritland K. Estimators for pairwise relatedness and individual inbreeding coefficients.
671 Genet. Res. 1996;67:175–85.

672 10. Christensen OF. Compatibility of pedigree-based and marker-based relationship
673 matrices for single-step genetic evaluation. Genet. Sel. Evol. 2012;44:37.
674 11. Quaas RL. Additive genetic model with groups and relationships. J. Dairy Sci.
675 1988;71:1338–45.
676 12. Makgahlela M, Strandén I, Nielsen U, Sillanpää M, Mäntysaari E. Using the unified
677 relationship matrix adjusted by breed-wise allele frequencies in genomic evaluation of a
678 multibreed population. J. Dairy Sci. 2014;97:1117–27.
679 13. Cockerham CC. Variance of gene frequencies. Evolution. 1969;23:72–84.
680 14. Wright S. Evolution in Mendelian populations. Genetics. 1931;16:97–159.
681 15. Crow J, Kimura M. An introduction to population genetics theory. Harper and Row,
682 New York; 1970.
683 16. Toro MÁ, García-Cortés LA, Legarra A. A note on the rationale for estimating
684 genealogical coancestry from molecular markers. Genet. Sel. Evol. GSE. 2011;43:27.
685 17. Robertson A. Gene Frequency Distributions as a Test of Selective Neutrality. Genetics.
686 1975;81:775–85.
687 18. Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B. Detecting signatures of
688 selection through haplotype differentiation among hierarchically structured populations.
689 Genetics. 2013;193:929–941.
690 19. Laval G, SanCristobal M, Chevalet C. Measuring genetic distances between breeds: use
691 of some distances in various short term evolution models. Genet. Sel. Evol. 2002;34:481–
692 508.

693 20. Forneris NS, Legarra A, Vitezica ZG, Tsuruta S, Aguilar I, Misztal I, et al. Quality control
694 of genotypes using heritability estimates of gene content at the marker. Genetics.
695 2015;199:675–81.
696 21. Gengler N, Mayeres P, Szydlowski M. A simple method to approximate gene content in
697 large pedigree populations: application to the myostatin gene in dual-purpose Belgian
698 Blue cattle. animal. 2007;1:21–8.
699 22. Mäntysaari E, Vleck L. Restricted maximum likelihood estimates of variance
700 components from multitrait sire models with large number of fixed effects. J. Anim. Breed.
701 Genet. 1989;106:409–22.
702 23. Thompson R. Sire evaluation. Biometrics. 1979;35:339–53.
703 24. Sargolzaei M, Schenkel FS. QMSim: a large-scale genome simulator for livestock.
704 Bioinformatics. 2009;25:680–1.
705 25. Christensen O, Madsen P, Nielsen B, Ostersen T, Su G. Single-step methods for
706 genomic evaluation in pigs. Animal. 2012;6:1565–71.
707 26. Masuda Y, Misztal I, Tsuruta S, Legarra A, Aguilar I, Lourenco DAL, et al.
708 Implementation of genomic recursions in single-step genomic best linear unbiased
709 predictor for US Holsteins with a large number of genotyped animals. J. Dairy Sci.
710 2016;99:1968–1974.
711 27. Mehrabani-Yeganeh H, Gibson JP, Schaeffer L r. Including coefficients of inbreeding in
712 BLUP evaluation and its effect on response to selection. J. Anim. Breed. Genet.
713 2000;117:145–51.

714 28. Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee DH. BLUPF90 and related
715 programs (BGF90). 7th World Congr. Genet. Appl. Livest. Prod. Montpellier, France; 2002.
716 p. CD-ROM Communication N° 28-07.
717 29. Mantysaari E, Liu Z, VanRaden P. Interbull validation test for genomic evaluations.
718 Interbull Bull. 2010;41.
719 30. Sargolzaei M, Chesnais J, Schenkel FS. Assessing the bias in top GPA bulls [Internet].
720 2012 [cited 2016 Jul 21]. Available from:
721 cgil.uoguelph.ca/dcbgc/Agenda1209/DCBGC1209_Bias_Mehdi.pdf
722 31. Spelman RJ, Arias J, Keehan MD, Obolonkin V, Winkelman AM, Johnson DL, et al.
723 Application of genomic selection in the New Zealand dairy cattle industry. Proc. 9th World
724 Congr. Genet. Appl. Livest. Prod. 1-6 August 2010 Leipz. [Internet]. 2010 [cited 2016 Jul
725 26]. Available from: http://www.icar.org/Cork_2012/Manuscripts/Published/Spelman.pdf
726 32. Winkelman AM, Johnson DL, Harris BL. Application of genomic evaluation to dairy
727 cattle in New Zealand. J. Dairy Sci. 2015;98:659–75.
728 33. Tsuruta S, Misztal I, Aguilar I, Lawlor T. Multiple-trait genomic evaluation of linear type
729 traits using genomic and phenotypic data in US Holsteins. J. Dairy Sci. 2011;94:4198–204.
730 34. Powell JE, Visscher PM, Goddard ME. Reconciling the analysis of IBD and IBS in
731 complex trait studies. Nat Rev Genet. 2010;11:800–5.
732 35. Harris BL, Johnson DL. Genomic predictions for New Zealand dairy bulls and
733 integration with national genetic evaluation. J Dairy Sci. 2010;93:1243–52.

734 36. Meuwissen T, Luan T, Woolliams J. The unified approach to the use of genomic and
735 pedigree information in genomic evaluations revisited. J. Anim. Breed. Genet.
736 2011;128:429–39.
737 37. Strandén I, Christensen OF. Allele coding in genomic evaluation. Genet Sel Evol.
738 2011;43:25.
739 38. Wright S. Isolation by Distance. Genetics. 1943;28:114–38.
740 39. Holsinger KE, Weir BS. Genetics in geographically structured populations: defining,
741 estimating and interpreting FST. Nat. Rev. Genet. 2009;10:639–650.
742 40. Henderson CR. Sire evaluations and genetic trends. J Anim Sci. 1973;Symposium.
743 41. Misztal I, Vitezica Z-G, Legarra A, Aguilar I, Swan A. Unknown-parent groups in single-
744 step genomic evaluation. J. Anim. Breed. Genet. 2013;130:252–8.
745 42. Christensen OF, Madsen P, Nielsen B, Su G. Genomic evaluation of both purebred and
746 crossbred performances. Genet. Sel. Evol. 2014;46:1–9.
747 43. Lourenco DAL, Tsuruta S, Fragomeni BO, Chen CY, Herring WO, Misztal I. Crossbreed
748 evaluations in single-step genomic best linear unbiased predictor using adjusted realized
749 relationship matrices. J. Anim. Sci. 2016;94:909.
750
751