

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

3

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

6

7 ^{*}Departamento de Producción Animal, Facultad de Agronomía, Universidad de Buenos
8 Aires, C1417DSE Buenos Aires, Argentina

9 ^{§§}Instituto de Investigaciones en Producción Animal - Consejo Nacional de Investigaciones
10 Científicas y Técnicas, Buenos Aires, Argentina

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

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

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

16 ^{**}Center for Quantitative Genetics and Genomics, Department of Molecular Biology and
17 Genetics, Aarhus University, DK-8830 Tjele, Denmark

18

19 Running title: metafounders in single step GBLUP

20

21

22

ABSTRACT

23

24 BACKGROUND:

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

29 RESULTS:

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

39 A compatible genomic relationship matrix constructed as a crossproduct of $\{-1,0,1\}$ codes,
40 and equivalent (up to scale factors) to an identity by state relationship matrix at the
41 markers, is derived. Using a simulation with a single population under selection, in which
42 only males and youngest animals were genotyped, we observed that generalized least

43 squares or maximum likelihood gave accurate and unbiased estimates of the ancestral
44 relationship parameter (true value: 0.40) whereas the other two (naïve and method of
45 moments) were biased (estimates of 0.43 and 0.35). We also observed that genomic
46 evaluation by Single Step GBLUP using metafounders was less biased in terms of accurate
47 genetic trend (0.01 instead of 0.12 bias), slightly overdispersed (0.94 instead of 0.99) and
48 as accurate (0.74) than the regular Single Step GBLUP. Single Step GBLUP using
49 metafounders also provided consistent estimates of heritability.

50 CONCLUSIONS:

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

54

55 **Keywords:** BLUP, Fst, relationships, genomic selection

56

57

BACGROUND

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

$$64 \quad \mathbf{H} = \begin{pmatrix} \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{pmatrix},$$

65 with inverse

$$66 \quad \mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{pmatrix}$$

67 where \mathbf{G} is the genomic relationship matrix, \mathbf{A} is the pedigree-based relationship matrix, and
68 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
69 genotyped individuals, respectively.

70 Because genotyped animals are not a random sample from the analyzed populations (they
71 are younger or selected), it was quickly acknowledged that a proper analysis requires
72 specifying different means for genotyped and non-genotyped individuals for the trait
73 under consideration. These different means can be considered as parameters of the
74 model, which are either fixed [4] or random [5,6]. In the latter case, the random variables
75 induce covariances across individuals, a situation that is referred to as “compatibility” of
76 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 \mathbf{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. $\mathbf{A}^{(\gamma)} = \mathbf{I} \left(1 - \frac{\gamma}{2}\right) + \mathbf{1}\mathbf{1}'\gamma$ 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

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

108

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

117

118

119

METHODS

120

Relationship between metafounders and allelic frequencies at the base

121 **Single population.** Let \mathbf{M} be a matrix of genotypes coded as gene content, i.e. $\{0,1,2\}$ and
122 the genomic relationship matrix $\mathbf{G} = (\mathbf{M} - \mathbf{J})(\mathbf{M} - \mathbf{J})'/s$ with \mathbf{J} a matrix of 1's, with
123 reference alleles taken at random so that for a random locus the expected allelic
124 frequency p is 0.5. [10]. In other words, the matrix $\mathbf{Z} = (\mathbf{M} - \mathbf{J})$ contains values of $\{-1,0,1\}$
125 for each genotype. In a single population, let γ be a relationship coefficient across
126 pedigree founders or, equivalently the self-relationship of the metafounder [1,10].
127 Parameter γ is the relationship coefficient among the founders of a population, so that
128 $\mathbf{G} = (\mathbf{M} - \mathbf{J})(\mathbf{M} - \mathbf{J})'/s$ is most likely given the observed pedigree. This relationship γ is
129 relative to a population with maximum heterozygosity and it is analogous to an F_{st}
130 fixation index. The parameter s is a measure of maximum heterozygosity in the
131 population.

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

136 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
137 a random locus i . The parameter $s = n(2Var(p_i) + E(2p_iq_i))$ with n being the number
138 of markers. However, $E(2p_iq_i) = 2E(p_i)E(q_i) - 2Var(p_i) = 0.5 - 2Var(p_i)$, where it
139 was used that if alleles are labelled at random across loci then $E(p_i) = E(q_i) = 0.5$. From
140 this it follows that $s = \frac{n}{2}$ and the genomic relationship matrix is $\mathbf{G} = 2(\mathbf{M} - \mathbf{J})(\mathbf{M} - \mathbf{J})'/n$.
141 Interestingly, this matrix is similar to a matrix of IBS relationships, that can be written

142 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 the
143 supplementary Material).

144

145 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

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

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

151

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

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

154 i.e., the covariance across loci between allelic frequencies of two populations b and b' .
155 This is almost tautological: the relationship is the covariance across gene contents at a
156 locus, here applied for populations. Christensen et al. (2015) show this in Appendix A,
157 somehow implicitly. Cockerham [13] and Robertson [17] interpret $4\sigma_{p_b,p_{b'}}$ as the
158 coancestry across two populations and Fariello et al. [18] use $\sigma_{p_b,p_{b'}}$ to describe the
159 divergence of populations. There are several measures of genetic distance between
160 populations (e.g. [19]), and most of them contain a term related, implicitly or explicitly, to

161 $\sigma_{p_b, p_{b'}}$. In particular, the average square of the Euclidean distance can be written as $D^2 =$
162 $E((p_b - p_{b'})^2) = -2\sigma_{p_b, p_{b'}}$. Thus, $\gamma_{b, b'} = -4D^2$.

163

164 **Estimation**

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

170

171 *1-Assuming no pedigree structure. NAIVE:* The simplest model assumes that genotyped
172 individuals are unrelated and constitute the base population. For locus i , let \mathbf{m}_i be a
173 vector of gene contents in the form $\{0,1,2\}$, defined as before. The mean of this vector is
174 $\mu_i = 2p_i$. For each locus, estimate μ_i as the observed mean of \mathbf{m}_i , then compute $Var(\hat{\boldsymbol{\mu}})$
175 as the empirical variance across loci of $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \dots, \hat{\mu}_n)$, and because $p_i = \mu_i/2$ then $\hat{\sigma}_p^2 =$
176 $Var(\hat{\boldsymbol{\mu}})/4$ and $\gamma = 8\hat{\sigma}_p^2 = 2Var(\hat{\boldsymbol{\mu}})$.

177

178 *2-Considering pedigree structure.* At locus i , gene content can be seen as a quantitative
179 trait where the mean of \mathbf{m}_i in the base population is $2p_i$, where p_i is the allelic frequency
180 at the base population, and the genetic variance is $2p_iq_i$ [20]. Cockerham (1969) showed
181 that the covariance of gene content of marker i across individuals j and k is a function of
182 relationship $Cov(m_{i,j}, m_{i,k}) = A_{jk}2p_iq_i$. A linear model can therefore be written as:

183
$$\mathbf{m}_i = \mathbf{1}\mu_i + \mathbf{W}\mathbf{u}_i + \mathbf{e}$$

184 where \mathbf{W} is an incidence matrix relating individuals in pedigree to genotypes, and with \mathbf{u}_i
 185 being the deviation of each individual from the mean μ_i for all individuals (Gengler et al.,
 186 2007; Forneris et al., 2015). Assuming multivariate normality:

187
$$\boldsymbol{\mu} \sim N(\mathbf{0}, \mathbf{I}\sigma_\mu^2)$$

188
$$\mathbf{u}_i \sim N(\mathbf{0}, \mathbf{A}(2p_iq_i)) = N(\mathbf{0}, \mathbf{A}\sigma_{m_i}^2)$$

189 Equivalently, for the set of genotyped individuals (labelled as “2”),
 190 $\mathbf{u}_{2,i} \sim N(\mathbf{0}, \mathbf{A}_{22}(2p_iq_i))$ where $\mathbf{A}_{22} = \mathbf{W}\mathbf{A}\mathbf{W}'$ is an additive relationship matrix spanning
 191 only the genotyped individuals. From this formulation, there are two possible strategies to
 192 estimate σ_μ^2 .

193

194 Generalized Least Squares (GLS). This ignores the prior distribution of $\boldsymbol{\mu}$ and estimates
 195 each μ_i as a “fixed effect” using for each locus separate BLUP (or, equivalently, GLS)
 196 estimators of μ_i . One option is to use the complete \mathbf{A}^{-1} and mixed model equations
 197 [20,21]. Equivalently, the corresponding GLS expression is

198
$$\hat{\mu}_i = (\mathbf{1}'\mathbf{A}_{22}^{-1}\sigma_{m_i}^{-2}\mathbf{1})^{-1}\mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{m}_i\sigma_{m_i}^{-2} = (\mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{1})^{-1}\mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{m}_i$$

199 where $(\mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{1})$ is the sum of elements of \mathbf{A}_{22}^{-1} , $\sigma_{m_i}^2 = 2p_iq_i$ and $\mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{m}_i$ is simply a
 200 weighted sum of genotypes. Then, estimate σ_μ^2 as $Var(\hat{\mu})$ and because $p_i = \mu_i/2$, $\hat{\sigma}_p^2 =$
 201 $\sigma_\mu^2/4$, and it follows that $\hat{\gamma} = 2\hat{\sigma}_p^2$.

202

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

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

208 1. For each marker i ,

209 a. estimate $\hat{\mu}_i = (\mathbf{1}'\mathbf{A}_{22}^{-1}\sigma_{m_i}^{-2}\mathbf{1} + \sigma_\mu^{-2})^{-1}\mathbf{1}'\mathbf{A}_{22}^{-1}\sigma_{m_i}^{-2}\mathbf{m}_i$

210 b. store $PEV_i(\hat{\mu}_i) = (\sigma_\mu^{-2} + \mathbf{1}'\mathbf{A}_{22}^{-1}\sigma_{m_i}^{-2}\mathbf{1})^{-1}$

211 c. update $\sigma_{m_i}^2$ as $\hat{\sigma}_{m_i}^2 = 2\hat{p}_i\hat{q}_i$ with $\hat{p}_i = \hat{\mu}_i/2$

212 2. Update σ_μ^2 as $\hat{\sigma}_\mu^2 = \frac{1}{n}(\hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}} + \sum PEV_i(\hat{\mu}_i))$, where the second part of the
213 expression corresponds to the trace $Tr(\mathbf{IC})$, \mathbf{I} , the identity matrix, is the
214 relationship across $\boldsymbol{\mu}$ and \mathbf{C} is the prediction error covariance matrix of $\hat{\boldsymbol{\mu}}$. As only
215 the diagonal elements of \mathbf{C} are needed, the elements $PEV_i(\hat{\mu}_i)$ can be obtained
216 separately from each single locus analysis.

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

220

221

222 **Estimation in multiple populations.**

223 If t base populations are considered, the variance component σ_{μ}^2 generalizes to Σ_0 , a $t \times t$
224 matrix of variances and covariances across means μ_i^b for marker i in population b . Across

225 different populations, $\Sigma_0 = \begin{pmatrix} \sigma_{\mu^1\mu^1}^2 & \sigma_{\mu^1\mu^2} & \dots \\ \dots & \sigma_{\mu^2\mu^2}^2 & \dots \\ \dots & \dots & \dots \end{pmatrix}$ and $\hat{\Gamma} = 2\hat{\Sigma}_0$.

226

227 *1-Assuming no pedigree structure. NAIVE* If relationships across individuals are ignored:

228
$$\mathbf{m}_i = \mathbf{Q}\boldsymbol{\mu}_i + \mathbf{e}_i$$

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

232

233 *2-Considering pedigree structure.* If there are no crosses, the estimation of allelic
234 frequencies can be split in separate analysis by population b : $\mathbf{m}_i^j = \mathbf{1}\mu_i^b + \mathbf{W}^b\mathbf{u}_i^b + \mathbf{e}$
235 with $\mathbf{u}_i^b \sim N(\mathbf{0}, \mathbf{A}^b(2p_i(1-p_i)))$, and \mathbf{A}^b is the matrix of relationships concerning
236 population b . Then, $\hat{\mathbf{P}}_0$ is estimated as the observed matrix of covariances across loci for
237 estimated $\hat{\mu}_i^b$. If there are crosses, there are two alternatives.

238 GENERALIZED LEAST SQUARES (GLS). The first alternative, suggested by Forneris et al.

239 (2015) is to use a genetic groups model [11,23], as $\mathbf{m}_i = \mathbf{Q}\boldsymbol{\mu}_i + \mathbf{W}\mathbf{u}_i + \mathbf{e}$ where $\mathbf{Q}_{k,b}$
240 contains the fraction of ancestry b in individual k . This ignores the fact that the variance
241 of gene content, $(2p_iq_i)$ is different for each breed and cross. The second, and more exact
242 alternative is to use the representation where the breeding values are split into within and
243 across breed components (Garcia-Cortes and Toro, 2006), as

$$244 \quad \mathbf{m}_i = \mathbf{Q}\boldsymbol{\mu}_i + \sum_b \mathbf{W}^b \mathbf{u}_i^b + \sum_{b,b',b>b'} \mathbf{W}^{b,b'} \mathbf{u}_i^{b,b'} + \mathbf{e}$$

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

246 MAXIMUM LIKELIHOOD (ML). Analogously to the single population case, an Expectation-
 247 Maximization updated estimate can be obtained using multiple trait formulations [22]
 248 where *PEC* is the prediction error variance-covariance, e.g. with two populations:

$$249 \quad \boldsymbol{\Sigma}_0 = \begin{pmatrix} \boldsymbol{\mu}^{1'} \boldsymbol{\mu}^{1'} & \boldsymbol{\mu}^{1'} \boldsymbol{\mu}^{2'} \\ \boldsymbol{\mu}^{2'} \boldsymbol{\mu}^{1'} & \boldsymbol{\mu}^{2'} \boldsymbol{\mu}^{2'} \end{pmatrix}.$$

250 Our current implementation is as follows:

251 1. For each marker i ,

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

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

254 c. update $\sigma_{m_i}^2$ as $\hat{\sigma}_{m_i}^2 = 2\hat{p}_i^*(1 - \hat{p}_i^*)$ with $\hat{p}_i^* = \frac{1}{nb} \sum_{b=1,nb} \frac{\hat{\mu}_i^b}{2}$

255 2. Update $\boldsymbol{\Sigma}_0$ using crossproducts within and across populations as e.g. with two
 256 populations,

$$257 \quad \hat{\boldsymbol{\Sigma}}_0 = \frac{1}{n} \left(\begin{pmatrix} \hat{\boldsymbol{\mu}}^{1'} \hat{\boldsymbol{\mu}}^1 & \hat{\boldsymbol{\mu}}^{1'} \hat{\boldsymbol{\mu}}^2 \\ \hat{\boldsymbol{\mu}}^{2'} \hat{\boldsymbol{\mu}}^1 & \hat{\boldsymbol{\mu}}^{2'} \hat{\boldsymbol{\mu}}^2 \end{pmatrix} + \sum_{i=1,n} PEC_i \right).$$

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

260

261

SIMULATION

262 To assess the quality of genomic predictions using one metafounder, we simulated
263 data using QMSim [24]. The simulation closely followed Vitezica *et al.* (2011) to mimic a
264 dairy cattle selection scheme scenario. A historical population undergoing mutation and
265 drift was generated, followed by a recent population undergoing selection.

266 First, 100 generations of the historical population were generated with an effective
267 population size of 100 during the first 95 generations, followed by a gradual expansion
268 during the last 5 generations to an effective population size of 3000. In total 30
269 chromosomes of 100 cM and 40,000 segregating biallelic markers distributed at random
270 along the chromosomes in the first generation of the historical population were simulated.
271 The 40,000 markers were resampled from a larger set of 90,000 markers in order to obtain
272 allelic frequencies from a beta(2,2) distribution, similar to dairy cattle marker data, so that
273 true γ had a value around 0.40. Potentially, 1500 QTL affected the phenotype; QTL allele
274 effects were sampled from a Gamma distribution with a shape parameter of 0.4. The
275 mutation rate of the markers (recurrent mutation process) and QTL was assumed to be 2.5
276 $\times 10^{-5}$ per locus per generation (Solberg *et al.*, 2008). A female trait with a heritability of
277 0.30 was simulated.

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

284 extreme selection and matings among highly related individuals. There were 100
285 individuals (mainly found in the last generation) with an inbreeding coefficient higher than
286 0.20, with extreme cases (few individuals) with inbreeding coefficients higher than 0.40.
287 True breeding values (TBV) and pedigree information were available for all 10 generations
288 (28,800 individuals in pedigree), phenotypes were available for all females except the last
289 generation (14,300 records). All males (840 sires of females with phenotypic records) were
290 genotyped as well as 2600 individuals in generation 9 (with records) and 2600 in
291 generation 10 (with no records). All in all, 20 independent replicates were made. A two-
292 step analysis was carried out using the simulated data. First, we compared several
293 methods to estimate γ . Then, we tested the quality of genomic predictions using four
294 methods, one of them including one metafounder.

295

296 **Methods to estimate Gamma**

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

304

305

306 **Genomic prediction methods**

307 Genetic merit of the selection candidates in generation 10 (genotyped and with no
308 phenotype records) was estimated using four methods. The first one was the pedigree
309 based BLUP (PBLUP) based on phenotype and pedigree information. The second method
310 was Single-Step GBLUP (SSGBLUP) in which genomic information is also taken into account;
311 this method used the correction of [25] and is the default method used in most practical
312 applications [25,26]. However, the implementation that we used does not include
313 inbreeding in the setup of \mathbf{A}^{-1} [27], although it does consider it in \mathbf{A}_{22}^{-1} (see below for use
314 of these matrices). The third method was Single-Step GBLUP including inbreeding in the
315 setup of \mathbf{A}^{-1} and of \mathbf{A}_{22}^{-1} (SSGBLUP_F). Finally, the fourth method was SSGBLUP including
316 the metafounder (SSGBLUP_M), using γ estimated by GLS as it turned out to be an
317 accurate method to estimate gamma (see the Results section). The three methods used
318 the following inverse relationship matrices: PBLUP: \mathbf{A}^{-1} ; SSGBLUP: $\mathbf{H}^{-1} = \mathbf{A}^{-1} +$
319 $\begin{pmatrix} 0 & 0 \\ 0 & \mathbf{G}_a^{-1} - \mathbf{A}_{22}^{-1} \end{pmatrix}$ where \mathbf{G}_a is as in [25] ; SSGBLUP_M: $\mathbf{H}^{(\gamma)-1} = \mathbf{A}^{(\gamma)-1} +$
320 $\begin{pmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{(\gamma)-1} \end{pmatrix}$ where $\mathbf{G} = (\mathbf{M} - \mathbf{J})(\mathbf{M} - \mathbf{J})'/s$ with $s = n/2$ (see the Methods
321 section) and $\mathbf{A}^{(\gamma)}$ is as in [1]. More details are given in Supplementary material. For
322 computation we used blupf90 [28]. In the case of SSGBLUP_M we constructed all
323 relationship matrices with own software, and then used the option user_file in blupf90.

324

325 **Quality of genomic prediction**

326 Prediction quality was checked for all 2600 selection candidates. The accuracy of the
327 methods was measured as the Pearson correlation between TBV and EBV. Bias was
328 calculated as the difference between the average TBV and average EBV with respect to the
329 base population. Thus, bias is related to estimated genetic progress in the selection
330 candidates. The inflation (often called bias) of the prediction method was quantified by
331 the coefficient of regression of TBV on EBV. These two statistics corresponds to the
332 coefficients b_0 and b_1 in the Interbull validation method [29] which uses the regression
333 $TBV = b_0 + b_1EBV + e$. The mean square error (MSE) was calculated as the mean of the
334 squared difference between TBV and EBV. An ideal method should have maximum
335 accuracy, minimum MSE, zero bias and a regression coefficient of 1. These are not only
336 nice statistical properties but also have relevance in livestock selection [30–32]. Ranking
337 changes of the selection candidates were also assessed by calculating the Spearman's rank
338 correlation coefficients between EBVs across methods.

339

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

343

344

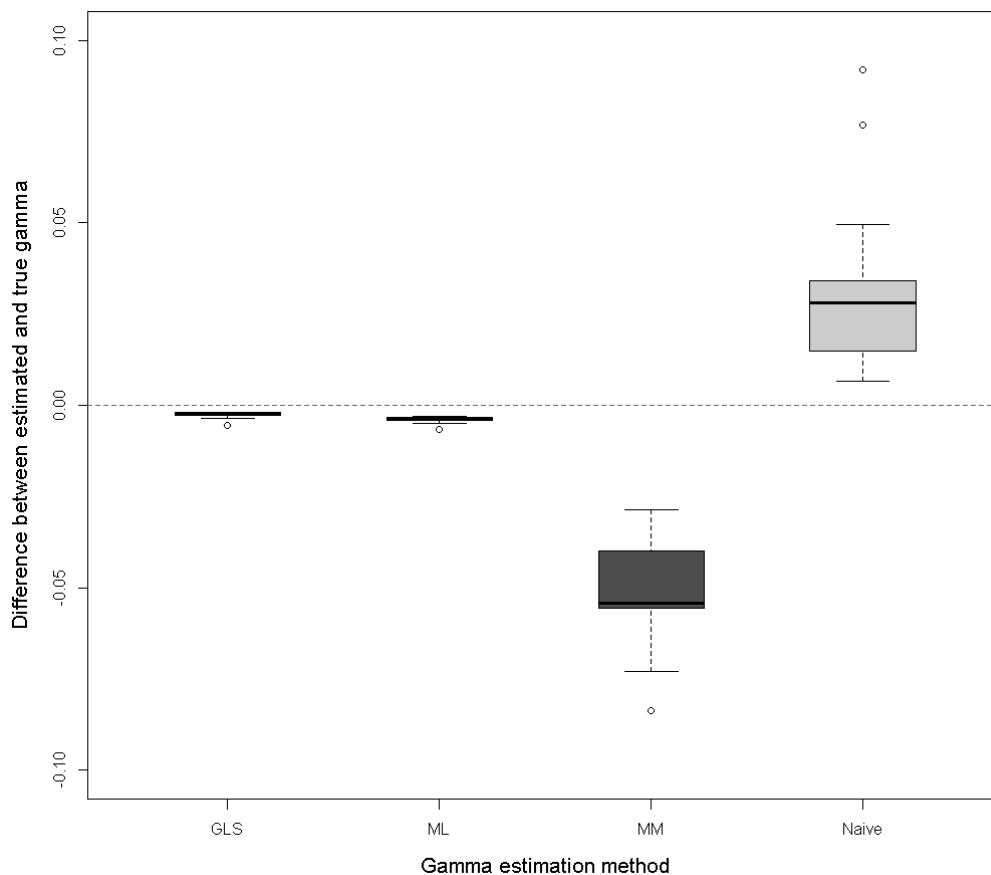
RESULTS

345 **Estimation of gamma**

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

355

356



357

358 **Figure 1** Differences between estimated and true Gamma, across 20 simulation replicates.

359 Gamma was estimated by Generalized Least Squares (GLS), Maximum Likelihood (ML),

360 Method of Moments (MM) and the Naive method.

361

362

363 **Quality of genomic prediction**

364 Correlations between TBV and EBV for each of the prediction methods are shown

365 in Table 1 and Figure 2a. Compared with PBLUP, SSGBLUP_F and SSGBLUP_M increased

366 accuracy by approximately 23 absolute points, respectively. This shows an important

367 improvement by including marker information in the prediction and the possibility of

368 generating a small extra gain when also including the metafounder. SSGBLUP resulted in a
369 small loss of accuracy as compared to SSGBLUP_F and SSGBLUP_M.

370

371

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)

372

373

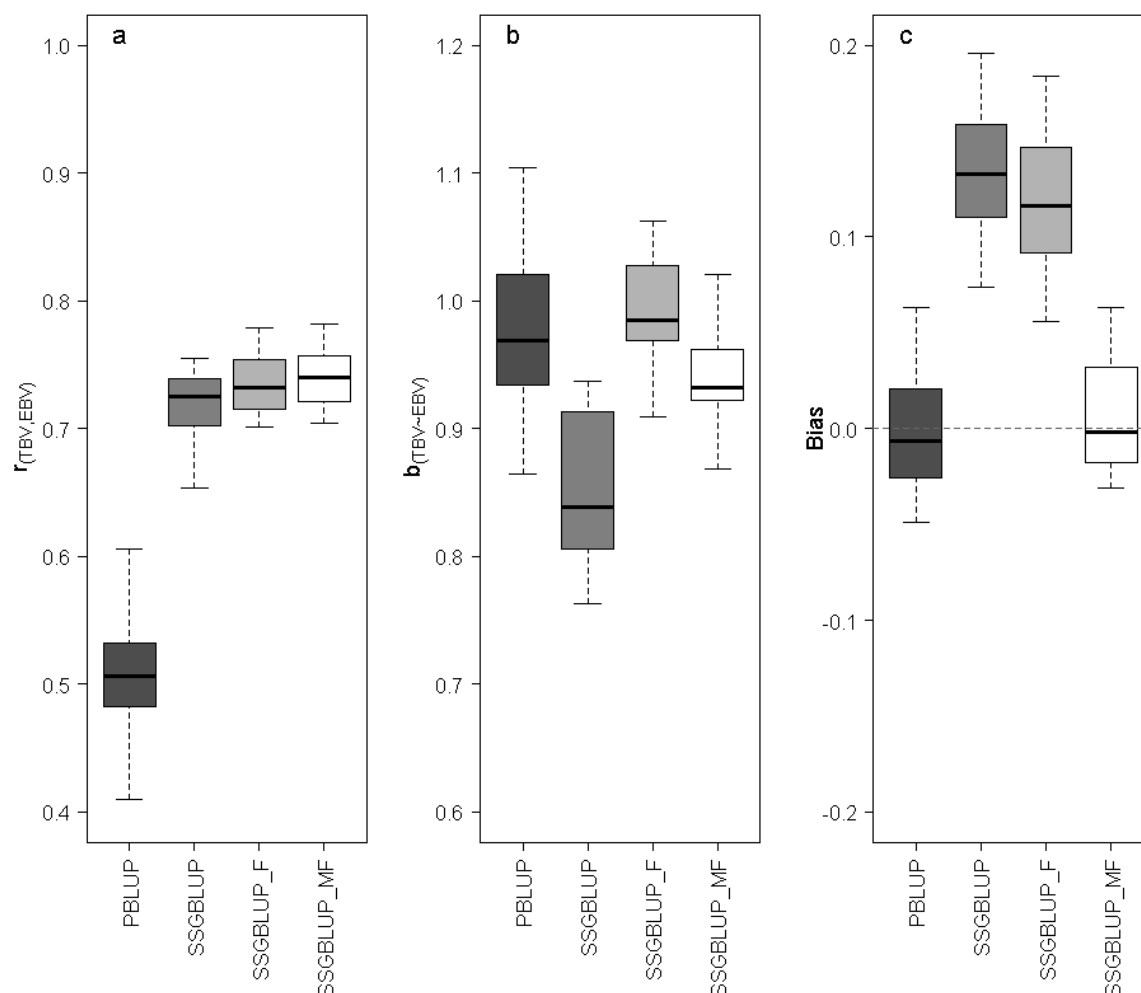
374 Bias values for each prediction method are shown in Table 1 and in Figure 2c. Both PBLUP
375 and SSGBLUP_M were unbiased, whereas SSGBLUP and SSGBLUP_F were biased. Bias in
376 SSGBLUP_F is equivalent to roughly 0.5 generations of genetic improvement or to 0.4
377 standard genetic deviations.

378

379 Table 1 and Figure 2b display the regression coefficient of TBV on EBV. This value measures
380 the inflation degree of each prediction method and should be close to 1. PBLUP and
381 SSGBLUP_F produced the values closest to one. Including genomic data in the prediction
382 using SSGBLUP resulted in regression coefficients lower than one, but including the
383 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

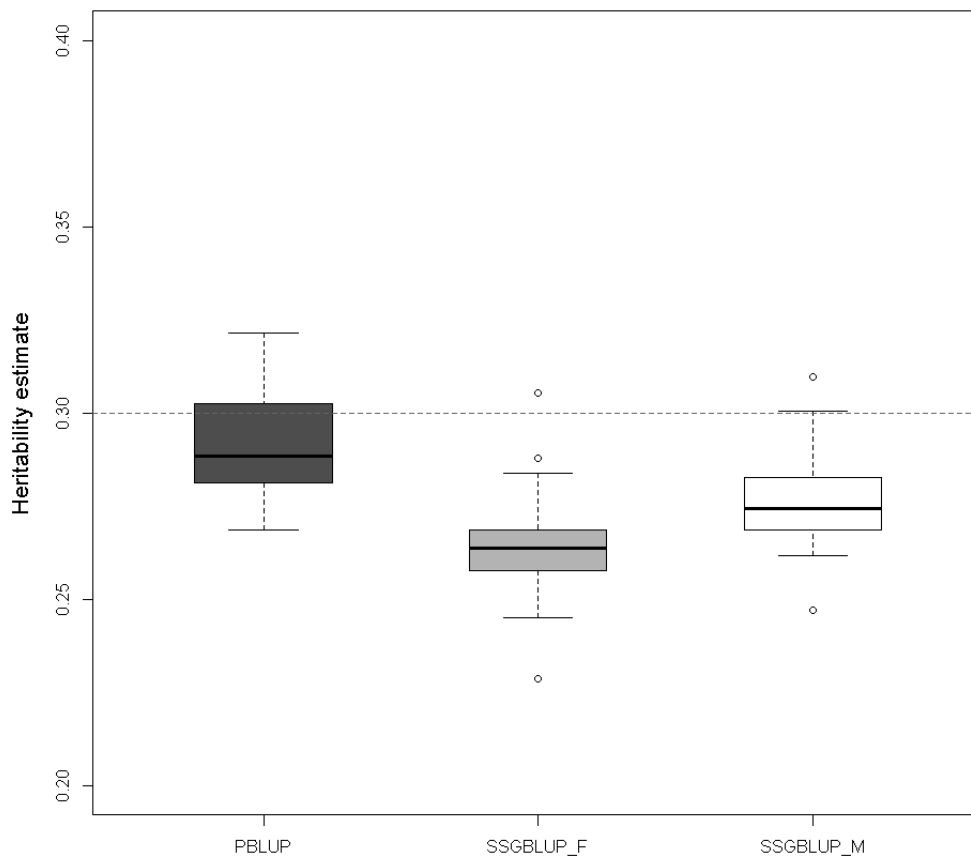
392 **Variance components estimation**

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 A_{22}^{-1} in
397 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)

416

417

418

DISCUSSION

419

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

433 matrix of IBS, or molecular, relationships is $\mathbf{G}_{IBS} = \mathbf{G}/2 + \mathbf{1}\mathbf{1}'$ (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 $\frac{1}{2}$ gets absorbed into the
436 variance component and the constant $\mathbf{1}\mathbf{1}'$ 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 \mathbf{G} and not
438 \mathbf{G}_{IBS} , because \mathbf{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 \mathbf{I} (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, \mathbf{I} 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 \mathbf{I} 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

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

456

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

460

461 **Relationship between metafounders γ and F_{st} fixation index**

462 The fixation index F_{st} [38] is a measure of diversity of a set of populations with respect to
463 a reference population, usually the pool of all populations. In this view, each population is
464 a random sample from all possible populations that could be sampled according to the
465 evolutionary process described by F_{st} . Conceptually, F_{st} is a parameter to be estimated
466 [13,39], and it is not a statistic computed from the data. A usual definition of F_{st} for a
467 particular biallelic locus is

$$468 \quad F_{st} = \frac{\sigma_p^2}{\bar{p}(1 - \bar{p})}$$

469 where σ_p^2 is the variance of allelic frequencies across populations and \bar{p} is the allelic
470 frequency of the conceptual combined population. If we consider that the variance of
471 allelic frequencies applies *across* loci and not *across* populations, it follows naturally that
472 $\bar{p} = 0.5$. In this case,

$$473 \quad F_{st} = \frac{\sigma_p^2}{\bar{p}(1 - \bar{p})} = \frac{\sigma_p^2}{0.5^2} = 4\sigma_p^2 = \frac{\gamma}{2}.$$

474 Our interpretation is as follows. Jacquard (1974) called $\frac{\gamma}{2}$ the “inbreeding coefficient of a
475 population”. Cockerham (1969) modelled $\frac{\gamma}{2} = \theta_l = F_{st}$ as an intraclass correlation, “the
476 coancestry of the line with itself”, in other words, the probability that two gametes taken
477 at random from the line are identical. Thus, it makes perfect sense to consider that the
478 additive relationship (which is twice the coancestry value) of a group with itself is $\gamma =$
479 $2\theta_l = 8\sigma_p^2$. This is the interpretation of the $\frac{\gamma}{2}$ coefficient in Legarra et al. [1]. Note that the
480 assumption $\bar{p} = 0.5$ is automatically fulfilled if reference alleles are labelled randomly
481 across loci (i.e., they are neither the most frequent nor the least observed).

482

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

487

488 **Consequences of using metafounders in genomic evaluation**

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

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

501

502

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

508

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

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

527

528

CONCLUSION

529 Metafounders are similar to F_{st} fixation indices and proportional to covariances of allelic
530 frequencies in base populations. Use of metafounders is simplified by new methods (GLS
531 and maximum likelihood) to estimate the covariance of base allele frequencies. We
532 verified by simulation of a selected population that, in a single population, both GLS and
533 ML are unbiased and computationally efficient. In the same simulation, use of
534 metafounders in Single Step GBLUP leads to more accurate and less biased evaluations,
535 and also to more accurate estimates of genetic parameters.

536

537 We propose a genomic relationship matrix that refers to a population with ideal
538 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

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

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

557 with $q = 1 - p$.

558 Now use is made of the following developments.

$$559 E(pq) = E(p(1 - p)) = E(p) - E(p^2). \quad (A1)$$

560 Since we have that $Var(p) = E(p^2) - E(p)^2$ we obtain $E(p^2) = Var(p) + E(p)^2$. We
 561 also have $E(q) = 1 - E(p)$. Substituting $E(p)^2$ in (A1) gives
 562 $E(pq) = E(p) - Var(p) - E(p)^2 = E(p)(1 - E(p)) - Var(p) = E(p)E(q) - Var(p)$.
 563 If markers are biallelic and labelled at random $E(p) = E(q) = 0.5$. So the equation above
 564 gives $E(pq) = 0.25 - Var(p)$. From this we obtain

$$565 \quad 2\eta_1 + \eta_2 = 2Var(p_j) + 0.5 - 2Var(p_j) = 0.5,$$

566 and therefore

$$567 \quad s = n(2\eta_1 + \eta_2) = \frac{n}{2}, \quad (1)$$

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

$$569 \quad \gamma = \frac{4\eta_1}{2\eta_1 + \eta_2} = \frac{4\eta_1}{0.5} = 8Var(p_j) = 8\sigma_p^2, \quad (2)$$

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

572 **Equivalences of genomic relationship matrices.**

573 The matrix \mathbf{G} described in Christensen (2012) and in this paper can be written as $\mathbf{G} =$
 574 $\frac{2}{n}(\mathbf{M} - \mathbf{J})(\mathbf{M} - \mathbf{J})'$, where \mathbf{M} contains genotypes coded as {0,1,2} and \mathbf{J} is a matrix of 1's.

575 The purpose of this paragraph is to show the linear relationship of this matrix with a
 576 matrix describing identity by state coefficients (IBS), in fact $\mathbf{G}_{IBS} = \frac{1}{2}\mathbf{G} + \mathbf{1}\mathbf{1}'$. The terms in
 577 \mathbf{G}_{IBS} are usually described in terms of identities or countings (i.e. Ritland, 1996; Toro et
 578 al., 2011; Nejati-Javaremi et al., 1997):

$$579 \quad G_{IBS_{ij}} = \frac{1}{n} \sum_{m=1}^n 2 \frac{\sum_{k=1}^2 \sum_{l=1}^2 I_{kl}}{4}$$

580 where I_{kl} measures the identity (with value 1 or 0) of allele k in individual i with allele l in
 581 individual j , and single-locus identity measures are averaged across n loci.

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

$$584 \quad f_{M_{ij}} = \frac{m_i m_j}{2} + \left(1 - \frac{m_i}{2}\right) \left(1 - \frac{m_j}{2}\right) \quad (3)$$

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

588 In order to prove $\mathbf{G}_{IBS} = \frac{1}{2} \mathbf{G} + \mathbf{11}'$, first we translate the Toro et al. (2011) equation to
 589 the more familiar scale of relationships $g_{IBS_{ij}} = 2f_{M_{ij}}$ and gene contents m . Thus

$$590 \quad g_{IBS_{ij}} = 2f_{M_{ij}} = 2 \left(\frac{m_i m_j}{2} + \left(\frac{2}{2} - \frac{m_i}{2} \right) \left(\frac{2}{2} - \frac{m_j}{2} \right) \right)$$

$$591 \quad g_{IBS_{ij}} = m_i m_j - m_i - m_j + 2$$

592 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

593

594 Also,

$$595 \quad g_{IBS_{ij}} = m_i m_j - m_i - m_j + 2 = (m_i - 1)(m_j - 1) + 1$$

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

597
$$\mathbf{G}_{IBS} = \frac{1}{n}(\mathbf{M} - \mathbf{J})(\mathbf{M} - \mathbf{J})' + \mathbf{1}\mathbf{1}'$$

598 Thus, matrix $\mathbf{G}_{IBS} = \frac{1}{n}(\mathbf{M} - \mathbf{J})(\mathbf{M} - \mathbf{J})' + \mathbf{1}\mathbf{1}'$ and because $\mathbf{G} = \frac{2}{n}(\mathbf{M} - \mathbf{J})(\mathbf{M} - \mathbf{J})'$ it

599 follows that $\mathbf{G}_{IBS} = \frac{1}{2}\mathbf{G} + \mathbf{1}\mathbf{1}'$. The equivalence can also be verified by noting that, for all

600 nine genotypes, the cross-product $(m_i - 1)(m_j - 1)$ in the following table is identical to

601 $g_{IBSij} - 1$ in the previous table.

	AA	Aa	aa
AA	1	0	-1
Aa	0	0	0
aa	-1	0	1

602

603

604 **Computation of the different H matrices**

605 For SSGBLUP and SSGBLUP_F, matrix \mathbf{H}^{-1} is constructed as follows:

606
$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_a^* - \mathbf{A}_{22} \end{pmatrix}$$

607 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

608 VanRaden (2008), \mathbf{M} contains genotypes coded as {0,1,2} and \mathbf{P} contains twice allelic

609 frequencies p_i . These are computed from the observed genotypes so that $2p_i$ is equal to

610 the the mean of the i -th column of \mathbf{M} . Constants a and b are such that the full-matrix and

611 diagonal averages of \mathbf{G}_a and \mathbf{A}_{22} are the same (Christensen et al., 2012) in order to make

612 the two matrices compatible. The use of the weights 0.95 and 0.05 is in order to make \mathbf{G}_a

613 invertible. Matrix \mathbf{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 \mathbf{A}^{-1} , method which pretends that in all cases
 618 inbreeding in expressions above is $F = 0$.

619 SSGBLUP_F uses \mathbf{H}^{-1} as above but constructs \mathbf{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
$$\mathbf{H}^{(\Gamma)-1} = \mathbf{A}^{(\Gamma)-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^* - \mathbf{A}_{22}^{(\Gamma)-1} \end{pmatrix}$$

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

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

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

627 value of Γ . Inbreeding is fully considered in both matrices.

628

629

630

631 DECLARATIONS

632 Availability of data and materials: Software and files are available 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 F_{ST} . *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