


RESEARCH ARTICLE

Open Access



# Estimating genomic relationships of metafounders across and within breeds using maximum likelihood, pseudo-expectation–maximization maximum likelihood and increase of relationships

Andres Legarra<sup>1\*</sup> , Matias Bermann<sup>2</sup>, Quanshun Mei<sup>3</sup> and Ole F. Christensen<sup>4</sup>

## Abstract

**Background** The theory of “metafounders” proposes a unified framework for relationships across base populations within breeds (e.g. unknown parent groups), and base populations across breeds (crosses) together with a sensible compatibility with genomic relationships. Considering metafounders might be advantageous in pedigree best linear unbiased prediction (BLUP) or single-step genomic BLUP. Existing methods to estimate relationships across metafounders  $\mathbf{\Gamma}$  are not well adapted to highly unbalanced data, genotyped individuals far from base populations, or many unknown parent groups (within breed per year of birth).

**Methods** We derive likelihood methods to estimate  $\mathbf{\Gamma}$ . For a single metafounder, summary statistics of pedigree and genomic relationships allow deriving a cubic equation with the real root being the maximum likelihood (ML) estimate of  $\mathbf{\Gamma}$ . This equation is tested with Lacaune sheep data. For several metafounders, we split the first derivative of the complete likelihood in a term related to  $\mathbf{\Gamma}$ , and a second term related to Mendelian sampling variances. Approximating the first derivative by its first term results in a pseudo-EM algorithm that iteratively updates the estimate of  $\mathbf{\Gamma}$  by the corresponding block of the  $\mathbf{H}$ -matrix. The method extends to complex situations with groups defined by year of birth, modelling the increase of  $\mathbf{\Gamma}$  using estimates of the rate of increase of inbreeding ( $\Delta F$ ), resulting in an expanded  $\mathbf{\Gamma}$  and in a pseudo-EM+ $\Delta F$  algorithm. We compare these methods with the generalized least squares (GLS) method using simulated data: complex crosses of two breeds in equal or unsymmetrical proportions; and in two breeds, with 10 groups per year of birth within breed. We simulate genotyping in all generations or in the last ones.

**Results** For a single metafounder, the ML estimates of the Lacaune data corresponded to the maximum. For simulated data, when genotypes were spread across all generations, both GLS and pseudo-EM(+ $\Delta F$ ) methods were accurate. With genotypes only available in the most recent generations, the GLS method was biased, whereas the pseudo-EM(+ $\Delta F$ ) approach yielded more accurate and unbiased estimates.

**Conclusions** We derived ML, pseudo-EM and pseudo-EM+ $\Delta F$  methods to estimate  $\mathbf{\Gamma}$  in many realistic settings. Estimates are accurate in real and simulated data and have a low computational cost.

\*Correspondence:

Andres Legarra

andres.legarra@uscpcb.com

Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

The theory of “metafounders” (abbreviated MF in the following) [1, 2] proposes a unified framework for relationships across base populations within breeds (that are usually modelled using unknown parent groups for different pathways of selection and periods), and base populations across breeds e.g. in crossbred animals (that are also sometimes modelled with unknown parent groups) together with a sensible compatibility with genomic relationships. Relationships across base populations are defined using an “absolute” reference point which is an ideal population with allele frequencies at biallelic markers of 0.5 [3, 4]. These relationships are contained in a matrix called  $\Gamma$ . In essence, matrix  $\Gamma$  contains average (unobserved) relationships across (unobserved) pools of founder gametes, and these are the so-called metafounders.

It is of interest to use metafounders in predictions that include pedigree, either in pedigree best linear unbiased prediction (BLUP) or (more commonly) in single-step genomic BLUP. The reasons are to obtain both more accurate, less biased, and more robust solutions of metafounders themselves, in particular in the presence of a genetic trend [5, 6], while at the same time ensuring compatibility of pedigree relationships with genomic relationships. This requires an estimate of matrix  $\Gamma$ , which is typically based on genotyped individuals that rarely belong to the base populations of interest.

Legarra et al. [1] suggested a series of methods, which were improved, first, by the discovery that  $\Gamma$  is actually a function of base allele frequencies [3] and, second, by modelling the increase of relationships within breed [6–9].

Still, there is no consensus and computational efficient methods are lacking. This is true in particular for pedigrees composed of several breeds, possibly with crossings, and with MF defined within and across breeds. For instance, in their study, Kudinov et al. [6] considered in genetic evaluations of four dairy cattle breeds (Holstein, Nordic Red Dairy Cattle, Finncattle and “Other”), each of them, in turn, including 16 to 61 MF. Their method requires that the genotypes are well distributed in time to fit a covariance function across all unknown parent groups. Wicki et al. [10] considered two sub-populations of Lacaune dairy sheep, each with 22 MF; the method uses pedigree inbreeding to model the steady increase in  $\Gamma$ , but requires estimates of  $\Gamma$  at the earliest generation. All current methods have drawbacks: either they require that MF are within short genetic (time) distances of genotyped individuals [3], or that genotypes are distributed in time to obtain  $\Gamma$  from a covariance function [6], or methods are adapted to particular cases [10]. Moreover, some methods can provide estimates that are outside of

the admissible parametric space (matrix  $\Gamma$  must be positive semidefinite, and diagonal elements must be within the range from 0 to 2).

The companion paper [11] shows that a definition of  $\Gamma$  in a quantitative genetics context is such that  $\Gamma_{i,j} = \frac{2}{k}(2\mathbf{p}_i - \mathbf{1})(2\mathbf{p}_j - \mathbf{1})'$ , with  $\mathbf{p}_i$  and  $\mathbf{p}_j$  being the row vectors of allele frequencies of  $k$  markers in base populations  $i$  and  $j$ , i.e.  $\Gamma$  is a “genomic relationship” that is based on “genotypes” of their population, i.e. allele frequencies. Using this result and new developments, here we present: (1) a Maximum Likelihood (ML) estimation for a single MF, (2) an estimation of  $\Gamma$  for several MF by pseudo-Expectation–Maximization (pseudo-EM) (actually, EM of part of the derivative of the complete log-likelihood), which involves repeated set-ups of part of matrix  $\mathbf{H}$ -inverse, and (3) in the case of MF structured by year of birth, we couple the pseudo-EM with a heuristic method for within-breed estimation of  $\Gamma$ . We describe the theory and examine the results obtained with a simulated dataset.

## Theory Likelihood

The likelihood of given markers is as follows [1, 4]. Let's define,  $\mathbf{A}_{\Gamma 22}$ , the pedigree relationship matrix of genotyped individuals set up with the MF relationship matrix  $\Gamma$ . Thus, we estimate an unobserved quantity  $\Gamma$  using a statistical model that involves  $\Gamma$  as a parameter. We assume Gaussian distributions for convenience. The joint density of the observed genotypes in matrix  $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_k]$  with  $z$  coded as  $\{-1, 0, 1\}$ , assuming multivariate normality for markers, is, for  $k$  markers and given  $\Gamma$  and pedigree:

$$f(\mathbf{M}|\Gamma) \propto \prod_{j=1}^k \det(0.5\mathbf{A}_{\Gamma 22})^{-1/2} \exp\left(-\mathbf{z}'_j(0.5\mathbf{A}_{\Gamma 22})^{-1}\mathbf{z}_j/2\right),$$

where a proportionality constant is ignored. Because the product of exponential terms is the exponential of the sum, and that  $\sum_j \mathbf{z}'_j(0.5\mathbf{A}_{\Gamma 22})^{-1}\mathbf{z}_j/2 = \text{Tr}\left((\mathbf{A}_{\Gamma 22})^{-1}\mathbf{Z}\mathbf{Z}'\right)$ , the likelihood function is:

$$L(\Gamma) \propto \det(0.5\mathbf{A}_{\Gamma 22})^{-\frac{k}{2}} \exp\left(-\text{Tr}\left((\mathbf{A}_{\Gamma 22})^{-1}\mathbf{Z}\mathbf{Z}'\right)\right).$$

Taking the logarithm of the likelihood function and introducing the notation  $\mathbf{G} = \mathbf{Z}\mathbf{Z}'/(k/2)$ , we obtain the log-likelihood function:

$$\log_L(\Gamma) = \text{constant} - \left(\frac{k}{2}\right) \log(\det(\mathbf{A}_{\Gamma 22})) - \left(\frac{k}{2}\right) \text{Tr}\left(\mathbf{A}_{\Gamma 22}^{-1}\mathbf{G}\right).$$

Note that the likelihood uses  $\mathbf{G}$ , i.e. a direct crossproduct of marker readings, and the inverse of  $\mathbf{G}$  is not needed, so  $\mathbf{G}$  does not need to be strictly positive definite. Some options to compute  $l(\mathbf{\Gamma})$  are presented in [Appendix](#). The constant in the log-likelihood function is invariant to  $\mathbf{\Gamma}$  and is ignored in the following.

Maximization of this likelihood function has proven difficult for the general case. The main difficulty compared to variance component estimation is, first, that it is not possible to factorize  $\mathbf{A}_{\Gamma 22}$  into a Kronecker product of a parameter-free relationship matrix and  $\mathbf{\Gamma}$ , and second, that the elements of  $\mathbf{\Gamma}$  are propagated through the Mendelian sampling variance of the animals. Furthermore, derivative-free and Markov chain Monte Carlo methods proved to be difficult to use with simulated data (not shown). Below, we show an exact solution for a single MF, an EM maximization using part of the derivative of the complete log-likelihood function, and a heuristic extension to within-breed across-time metafounders.

### Maximum likelihood for a single metafounder

The theory for a single MF was presented in [12] and we reintroduce it here for completion and for later discussion. For the case of a single MF, there is an explicit solution that maximizes the log-likelihood function  $log_L(\mathbf{\Gamma}) = -\left(\frac{k}{2}\right)log(det(\mathbf{A}_{\Gamma 22})) - \left(\frac{k}{2}\right)Tr(\mathbf{A}_{\Gamma 22}^{-1}\mathbf{G})$ . Let us call  $\gamma$  the (single) scalar value of  $\mathbf{\Gamma}$  for the single MF case. In that case,  $\mathbf{A}_{\gamma 22} = \mathbf{A}_{22}(1 - \frac{\gamma}{2}) + \mathbf{1}\mathbf{1}'\gamma$ , where  $\mathbf{A}_{22}$  is the matrix of pedigree relationships across genotyped individuals. As detailed in [Appendix](#), we used this to obtain, in short-hand notation,  $a = \mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{1}$ ,  $b = Tr(\mathbf{A}_{22}^{-1}\mathbf{G})$  and  $c = Tr(\mathbf{A}_{22}^{-1}\mathbf{1}\mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{G}) = \mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{1}$ , which are later used in the cubic equation ( $n$  being the number of individuals with a genotype):

$$e_3\gamma^3 + e_2\gamma^2 + e_1\gamma + e_0 = 0,$$

where  $e_3 = -n(-1/2 + a)^2/2$ ,  $e_2 = n((-3/2 + a) + a - b(-1/2 + a) + c)(-1/2 + a)$ ,  $e_1 = (n - 1)(-3/2 + 2a) - (-1 + 2a)(-3/2 + a + b)$  and  $e_0 = n - 2a - b + 2c$ .

The real roots of this equation are the ML estimate of  $\gamma$ , in our simulated and real examples we have only found one real root. In addition, if the ML estimate of  $\gamma$  is outside the parametric space, the estimate is at the boundary. Some methods to compute  $b$  and  $c$  are presented in [Appendix](#) and their estimation is actually easy when matrices  $\mathbf{G}$  and  $\mathbf{A}_{22}^{-1}$  can be explicitly computed.

### Maximum likelihood with multiple metafounders

Here we just sketch what could be done in principle. Using the same log-likelihood, it is conceptually possible to split the likelihood among breeds and pairs of breeds using partial relationship matrices [1, 13]:

$$\mathbf{A}_{\Gamma} = \sum_b \mathbf{A}^b \left(1 - \frac{\gamma_b}{2}\right) + \sum_{b,b',b'>b} \mathbf{A}^{b,b'} \left(\frac{\gamma'_b + \gamma_b}{8} - \frac{\gamma_{b,b'}}{4}\right) + \sum_b \mathbf{C}^b \gamma_b + \sum_{b,b',b'>b} \mathbf{C}^{b,b'} \gamma_{b,b'},$$

with  $\mathbf{A}^b$  being the breed  $b$  specific partial relationship matrix,  $\mathbf{A}^{b,b'}$  being the matrix of partial relationships due to segregation across breeds  $b$  and  $b'$  [13], matrix  $\mathbf{C}^b$  having entries  $\mathbf{C}_{i,i'}^b = f_i^b f_{i'}^b$ , and matrix  $\mathbf{C}^{b,b'}$  having entries  $\mathbf{C}_{i,i'}^{b,b'} = f_i^b f_{i'}^{b'} + f_i^{b'} f_{i'}^b$  for  $f_i^b$  being the fraction of breed “ $b$ ” origin of individual  $i$ . From this expression, matrix derivatives with respect to MF parameters can be obtained. However, this has proven to be difficult because the expressions quickly get too complicated.

### Derivatives of the “complete” likelihood

Following expectation–maximization ideas, we consider the derivatives of a “complete” likelihood in which all animals (including MF) are genotyped. The particular block of genomic relationships for MF will be named  $\mathbf{G}_{MF}$ , and indeed by definition  $\mathbf{G}_{MF} = \mathbf{\Gamma}$  as described in the companion paper [11]. From [11], we know the definition of elements of  $\mathbf{\Gamma}$ ,  $\mathbf{\Gamma}_{b,b'} = \frac{2}{k}(2\mathbf{p}_b - \mathbf{1})(2\mathbf{p}_{b'} - \mathbf{1})'$  for the  $\mathbf{p}_b$  and  $\mathbf{p}_{b'}$  row vectors of allele frequencies, although these allele frequencies are typically unknown (if they are known, estimation is immediate). In other words, it is meaningful to assign genomic relationships to MF. Consider now the form of the “complete” log-likelihood where  $\mathbf{A}_{\Gamma}$  and  $\mathbf{G}$  include all animals and the MF:

$$log_L(\mathbf{\Gamma}) = -\frac{k}{2}log(det(\mathbf{A}_{\Gamma})) - \frac{k}{2}Tr(\mathbf{A}_{\Gamma}^{-1}\mathbf{G}) = -\frac{k}{2}\left[log(det(\mathbf{A}_{\Gamma})) + Tr(\mathbf{A}_{\Gamma}^{-1}\mathbf{G})\right].$$

To maximize this “complete” log-likelihood function, we need to derive formulas of the derivatives of  $log(det(\mathbf{A}_{\Gamma}))$  and  $Tr(\mathbf{A}_{\Gamma}^{-1}\mathbf{G})$  with respect to elements in  $\mathbf{\Gamma}$ . For convenience, we use  $\gamma$ 's to represent each of the several parameters in  $\mathbf{\Gamma}$  in the following.

Consider matrix  $\mathbf{A}_{\Gamma} = \mathbf{T}\mathbf{D}_{\Gamma}\mathbf{T}'$ . This matrix includes relationship among individuals, among MF, and among both individuals and MF. Its inverse is

$(\mathbf{A}_\Gamma)^{-1} = (\mathbf{T}^{-1})' \mathbf{D}_\Gamma^{-1} \mathbf{T}^{-1}$  with  $\mathbf{T}^{-1} = \mathbf{I} - \mathbf{S}$  linking individuals to ancestors and MF to themselves; for instance an individual with one parent known and the other a MF has a -0.5 in the (individual, MF) element of  $\mathbf{S}$  [14] (where matrix  $\mathbf{S}$  was called  $\mathbf{P}$ ). The dependence of  $(\mathbf{T}^{-1})' \mathbf{D}_\Gamma^{-1} \mathbf{T}^{-1}$  on  $\Gamma$  is only through matrix  $\mathbf{D}_\Gamma$ .

Matrix  $\mathbf{D}_\Gamma$  is a block diagonal matrix, consisting of the  $\Gamma$  matrix for MF and the usual diagonal matrix with Mendelian sampling terms for non-metafounders;  $\mathbf{T}$  and  $\mathbf{T}^{-1}$  are lower triangular matrices.

After some algebra that is shown in Appendix, we can show that:

$$\frac{\partial \log_L(\Gamma)}{\partial \gamma} = -\frac{k}{2} \left[ \text{Tr} \left[ \Gamma^{-1} \frac{\partial \Gamma}{\partial \gamma} (\mathbf{I} - \Gamma^{-1} \mathbf{G}_{MF}) \right] \right] - \frac{k}{2} \left[ \sum_i \left[ \frac{\partial D_{\Gamma,ii}}{\partial \gamma} \left( \frac{1}{D_{\Gamma,ii}} - \frac{B_{i,i}}{D_{\Gamma,ii}^2} \right) \right] \right].$$

The structure of  $\mathbf{B}$  and  $\mathbf{D}_\Gamma$  is:

$$\mathbf{D}_\Gamma = \begin{cases} \Gamma, & \text{for metafounders} \\ \text{diagonal matrix,} & \text{for non-metafounders} \end{cases},$$

$$\mathbf{B} = \mathbf{W}\mathbf{W}' = \begin{cases} \mathbf{G}_{MF}, & \text{for metafounders} \\ \mathbf{W}_{NMF} \mathbf{W}_{NMF}', & \text{for non-metafounders} \end{cases},$$

where  $\mathbf{W} = \mathbf{T}^{-1} \mathbf{Z} / \sqrt{k/2}$ . Note that the block of  $\mathbf{T}^{-1}$  corresponding to MF is simply an identity matrix, and thus the block of  $\mathbf{B}$  corresponding to MF is simply  $\mathbf{G}_{MF}$ .

As for matrix  $\mathbf{G}_{MF}$ , it is the genomic relationships across MF: in order to derive the algorithm, we use the complete likelihood, in which it is assumed that all individuals, including MF, have been genotyped.

The first term in the partial derivative involves  $\mathbf{G}_{MF}$  and is simple to manipulate. The last term in the partial derivative involves individual terms of Mendelian sampling variances  $D_{\Gamma,ii}$  and it is difficult to compute and derive algebraically in a recognizable form. Instead, we approximate  $\frac{\partial \log_L(\Gamma)}{\partial \gamma}$  by the first term, i.e. as:

$$\frac{\partial \log_L(\Gamma)}{\partial \gamma} \approx -\frac{k}{2} \left[ \text{Tr} \left[ \Gamma^{-1} \frac{\partial \Gamma}{\partial \gamma} (\mathbf{I} - \Gamma^{-1} \mathbf{G}_{MF}) \right] \right].$$

### Pseudo-EM algorithm

#### E-step

The EM algorithm uses the expectation of the log-likelihood over the distribution of unknown data, conditional to the actual value of the parameters ( $\Gamma$ ). However, we know from the single step theory that (using the notation  $o$  for observed and  $n$  for not observed, in fact  $\mathbf{G}_{o,o}$  is  $\mathbf{G}$  actually observed) [15]:

$$E_G(\mathbf{G}) = E \left( \left[ \begin{array}{c} \mathbf{G}_{n,n} \mathbf{G}_{n,o} \\ \mathbf{G}_{o,n} \mathbf{G}_{o,o} \end{array} \right] \middle| \mathbf{G}_{o,o} \right) = \begin{pmatrix} \mathbf{H}_{n,n} \mathbf{H}_{n,o} \\ \mathbf{H}_{o,n} \mathbf{H}_{o,o} \end{pmatrix} = \mathbf{H},$$

and in fact,  $\mathbf{H}_{o,o} = \mathbf{G}_{o,o}$ . Using this together with  $E(\text{tr}(\mathbf{A})) = \text{tr}(E(\mathbf{A}))$  we get:

$$\begin{aligned} E_G(\log_L(\Gamma)) &= E_G \left( -\frac{k}{2} \left[ \log(\det(\mathbf{A}_\Gamma)) + \text{Tr}(\mathbf{A}_\Gamma^{-1} \mathbf{G}) \right] \right) \\ &= -\frac{k}{2} \left[ \log(\det(\mathbf{A}_\Gamma)) + E_G(\text{Tr}(\mathbf{A}_\Gamma^{-1} \mathbf{G})) \right] \\ &= -\frac{k}{2} \left[ \log(\det(\mathbf{A}_\Gamma)) + \text{Tr}(\mathbf{A}_\Gamma^{-1} E_G(\mathbf{G})) \right] \\ &= -\frac{k}{2} \left[ \log(\det(\mathbf{A}_\Gamma)) + \text{Tr}(\mathbf{A}_\Gamma^{-1} \mathbf{H}) \right] \end{aligned}$$

This means that, in the following derivations, we can use  $\mathbf{H}$  (computed at the actual value of  $\Gamma$ ) in the place of  $\mathbf{G}$ .

#### M-step

The approximate first derivative of the complete log-likelihood shown before is:

$$\frac{\partial \log_L(\Gamma)}{\partial \gamma} \approx -\frac{k}{2} \left[ \text{Tr} \left[ \Gamma^{-1} \frac{\partial \Gamma}{\partial \gamma} (\mathbf{I} - \Gamma^{-1} \mathbf{G}_{MF}) \right] \right],$$

and because the previous E-step uses the conditional expectation of  $\mathbf{G}$ , i.e.  $\mathbf{H}$ , this becomes:

$$\frac{\partial \log_L(\Gamma)}{\partial \gamma} \approx -\frac{k}{2} \left[ \text{Tr} \left[ \Gamma^{-1} \frac{\partial \Gamma}{\partial \gamma} (\mathbf{I} - \Gamma^{-1} \mathbf{H}_{MF}) \right] \right],$$

setting to 0, factorizing  $-\frac{k}{2}$  and introducing  $\gamma_{ij}$  to indicate which of the elements of  $\Gamma$  we work with, gives:

$$\text{Tr} \left[ \Gamma^{-1} \frac{\partial \Gamma}{\partial \gamma_{ij}} \right] = \text{Tr} \left[ \Gamma^{-1} \frac{\partial \Gamma}{\partial \gamma_{ij}} \Gamma^{-1} \mathbf{H}_{MF} \right].$$

After some algebra that is shown in Appendix this yields:

$$\Gamma = \mathbf{H}_{MF},$$

where  $\mathbf{H}_{MF}$  is the block of  $\mathbf{H}$  corresponding to the MF with themselves.

Remember that  $\mathbf{G}_{MF}$  is the (unobserved) genomic relationship matrix across MF, which at iteration  $t$  in the E-step is ‘‘augmented’’ by its conditional expectation,  $\mathbf{H}_{MF}$ , the corresponding submatrix of  $\mathbf{H}$ . The previous expression simply says that the algorithm proceeds by updating  $\mathbf{H}_{(t)}$  from the previous estimate  $\hat{\Gamma}_{(t-1)}$  and then setting  $\hat{\Gamma}_{(t)} \leftarrow \mathbf{H}_{MF(t)}$ , the last being the block of  $\mathbf{H}_{(t)}$  corresponding to MF.

**Algorithm**

- (1) Either  $\mathbf{Z}$  where  $\mathbf{Z}$  contains  $\{-1, 0, 1\}$  readings, or  $\mathbf{G} = \frac{2}{k}\mathbf{Z}\mathbf{Z}'$  may be used. Note that  $\mathbf{G}$  does not change across iterations and does not need to be full rank depending on the algorithm.
- (2) Use a starting value  $\hat{\mathbf{\Gamma}}_{(t=0)}$  different from  $\mathbf{0}$ , for instance  $\hat{\mathbf{\Gamma}}_{(t=0)} = 0.1\mathbf{I}$ .
- (3) Do the following steps until convergence, at iteration  $t$ :
  - (a) compute  $\mathbf{H}_{MF(t)}$  from  $\hat{\mathbf{\Gamma}}_{(t-1)}$ ;
  - (b) update:  $\hat{\mathbf{\Gamma}}_{(t)} \leftarrow \mathbf{H}_{MF(t)}$  using one of the options below;
  - (c) optionally, compute (exact) log-likelihood  $\log_L(\mathbf{\Gamma}) = -\left(\frac{k}{2}\right)\log(\det(\mathbf{A}_{\Gamma 22})) - \left(\frac{k}{2}\right)Tr(\mathbf{A}_{\Gamma 22}^{-1}\mathbf{G})$  (just for checking);
  - (d) at convergence,  $\hat{\mathbf{\Gamma}}_{(t)}$  is the estimate of  $\hat{\mathbf{\Gamma}}$ .

Convergence can be checked by comparing, for instance, the elements of the Cholesky decomposition  $\mathbf{U}_{(t)}$ ,  $\mathbf{U}_{(t-1)}$ , respectively of  $\mathbf{\Gamma}_{(t)}$  and of  $\mathbf{\Gamma}_{(t-1)}$  and using  $\frac{\sum_i \sum_j ((U_{(t)}[i,j] - U_{(t-1)}[i,j])^2)}{\sum_i \sum_j (U_{(t-1)}[i,j])^2}$ . To update  $\mathbf{H}_{MF(t)}$  there are several options:

**Option 1:**

- (1) Compute  $\mathbf{A}_{\Gamma(t)}^{-1}$ ,  $\mathbf{A}_{22\Gamma(t)}^{-1}$  from pedigree and  $\hat{\mathbf{\Gamma}}_{(t)}$ .
- (2) Compute  $\mathbf{H}_{(t)}^{-1} = \mathbf{A}_{\Gamma(t)}^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{\Gamma 22(t)}^{-1} \end{pmatrix}$  for all individuals including MF. Note that  $\mathbf{G}$  does not change across iterations, but  $\mathbf{A}_{\Gamma(t)}^{-1}$  and  $\mathbf{A}_{\Gamma 22(t)}^{-1}$  do, and so does  $\mathbf{H}_{(t)}^{-1}$ . In addition, this requires  $\mathbf{G}$  to be full rank. Note that  $\mathbf{H}_{(t)}^{-1}$  includes a square block for MF.
- (3) Extract the (MF, MF) block of  $\mathbf{H}_{MF(t)}$  from  $\mathbf{H}_{(t)}$ , which can be done:
  - (a) by repeated “solving” of  $\mathbf{H}_{(t)}^{-1}\mathbf{x} = \mathbf{v}$ , where  $\mathbf{v}$  is a vector with 1 in the MF position and 0 elsewhere;
  - (b) or by sparse inversion of  $\mathbf{H}_{(t)}^{-1}$ .

**Option 2:**

- (1) Use the expression  $\mathbf{\Gamma}_{t+1} = \mathbf{\Gamma}_t + \mathbf{A}_{\Gamma mf, 2(t)}\mathbf{A}_{\Gamma 22(t)}^{-1}(\mathbf{G} - \mathbf{A}_{22\Gamma(t)})\mathbf{A}_{\Gamma 22(t)}^{-1}\mathbf{A}_{\Gamma 2, mf(t)}$  [16], i.e. using the subblocks (genotyped individuals, MF) of  $\mathbf{A}_{\Gamma}$ .

- (2) Equivalently, use  $\mathbf{\Gamma}_{t+1} = \mathbf{\Gamma}_t + \left(\mathbf{A}_{\Gamma(t)}^{mf, mf}\right)^{-1}\mathbf{A}_{\Gamma(t)}^{mf, 2}(\mathbf{G} - \mathbf{A}_{22\Gamma(t)})\mathbf{A}_{\Gamma(t)}^{2, mf}\left(\mathbf{A}_{\Gamma(t)}^{mf, mf}\right)^{-1}$ , because  $\left(\mathbf{A}_{\Gamma(t)}^{mf, mf}\right)^{-1}\mathbf{A}_{\Gamma(t)}^{mf, 2} = \mathbf{A}_{\Gamma mf, 2(t)}\mathbf{A}_{\Gamma 22(t)}^{-1}$ .
- (3) Note that  $\mathbf{A}_{mf, 2\Gamma(t)} = \mathbf{Q}_2\mathbf{\Gamma}_t$  for  $\mathbf{Q}_2$ , which is a matrix with MF proportions in genotyped individuals.

Noting that  $\mathbf{A}_{\Gamma mf, 2(t)}\mathbf{A}_{\Gamma 22(t)}^{-1}\mathbf{Z}$  is a matrix which contains two times the estimates of allele frequencies, minus one [3, 17], we can put results above as a function of estimated allelic frequencies  $\hat{\mathbf{P}}$ , where  $\mathbf{P}$  has as many rows as markers ( $k$ ) and as many columns as populations ( $npop$ ), as follows:

$$\begin{aligned} \mathbf{\Gamma}_{t+1} &= \mathbf{\Gamma}_t + \mathbf{A}_{mf, 2\Gamma(t)}\mathbf{A}_{22\Gamma(t)}^{-1}(\mathbf{G} - \mathbf{A}_{22\Gamma(t)})\mathbf{A}_{22\Gamma(t)}^{-1}\mathbf{A}_{2, mf\Gamma(t)} \\ &= \left(\mathbf{\Gamma}_t - \mathbf{A}_{mf, 2\Gamma(t)}\mathbf{A}_{22\Gamma(t)}^{-1}\mathbf{A}_{2, mf\Gamma(t)}\right) \\ &\quad + \mathbf{A}_{mf, 2\Gamma(t)}\mathbf{A}_{22\Gamma(t)}^{-1}\mathbf{G}\mathbf{A}_{22\Gamma(t)}^{-1}\mathbf{A}_{2, mf\Gamma(t)} \\ &= \left(\mathbf{\Gamma}_t - \mathbf{A}_{mf, 2\Gamma(t)}\mathbf{A}_{22\Gamma(t)}^{-1}\mathbf{A}_{2, mf\Gamma(t)}\right) \\ &\quad + \left(\frac{2}{k}\right)\left(\mathbf{A}_{mf, 2\Gamma(t)}\mathbf{A}_{22\Gamma(t)}^{-1}\mathbf{Z}\right)\left(\mathbf{Z}'\mathbf{A}_{22\Gamma(t)}^{-1}\mathbf{A}_{2, mf\Gamma(t)}\right) \\ &= \left(\mathbf{\Gamma}_t - \mathbf{A}_{mf, 2\Gamma(t)}\mathbf{A}_{22\Gamma(t)}^{-1}\mathbf{A}_{2, mf\Gamma(t)}\right) \\ &\quad + \left(\frac{2}{k}\right)\left(2\hat{\mathbf{P}}_{\Gamma(t)} - \mathbf{1}_{npop}\mathbf{1}'_k\right)\left(2\hat{\mathbf{P}}_{\Gamma(t)} - \mathbf{1}_{npop}\mathbf{1}'_k\right)' \end{aligned}$$

This yields one extra option:

**Option 3:**

- (1) Use  $\mathbf{\Gamma}_{t+1} = \left(\mathbf{\Gamma}_t - \mathbf{A}_{mf, 2\Gamma(t)}\mathbf{A}_{22\Gamma(t)}^{-1}\mathbf{A}_{2, mf\Gamma(t)}\right) + \left(\frac{2}{k}\right)\left(2\hat{\mathbf{P}}_{\Gamma(t)} - \mathbf{1}_{npop}\mathbf{1}'_k\right)\left(2\hat{\mathbf{P}}_{\Gamma(t)} - \mathbf{1}_{npop}\mathbf{1}'_k\right)'$  where  $\hat{\mathbf{P}}_{\Gamma(t)}$  contains the current estimate of all allele frequencies across populations using the current value of  $\mathbf{A}_{\Gamma}$  (in this case  $\mathbf{G}$  is not needed).
- (2) Equivalently, use  $\mathbf{\Gamma}_{t+1} = \left(\mathbf{A}_{\Gamma(t)}^{mf, mf}\right)^{-1} + \left(\frac{2}{k}\right)\left(2\hat{\mathbf{P}}_{\Gamma(t)} - \mathbf{1}_{npop}\mathbf{1}'_k\right)\left(2\hat{\mathbf{P}}_{\Gamma(t)} - \mathbf{1}_{npop}\mathbf{1}'_k\right)'$ , where  $\mathbf{A}_{\Gamma(t)}^{mf, mf}$  is the block of  $\mathbf{A}_{\Gamma(t)}^{-1}$  e.g. that is set up using Henderson’s rules corresponding to MF, which then has to be inverted. This is because  $\left(\mathbf{A}_{\Gamma(t)}^{mf, mf}\right)^{-1} = \left(\mathbf{\Gamma}_t - \mathbf{A}_{mf, 2\Gamma(t)}\mathbf{A}_{22\Gamma(t)}^{-1}\mathbf{A}_{2, mf\Gamma(t)}\right)$ .

**Methods to estimate  $\Gamma$  across and within breeds using the increase in relationships**

The previous methods do not apply well to populations, such as ruminant species with unknown parent groups defined by year of birth within breed and sometimes within sexes or selection paths. It is often the case that

$$Var(\epsilon_t) = \Delta \bar{A}_t,$$

Note that, because inbreeding is half the relationship between the parents (and assuming mating at random),  $\Delta \bar{A}_t = \bar{A}_t - \bar{A}_{t-1} \approx 2\bar{F}_{t+1} - 2\bar{F}_t = 2\Delta F_{t+1}$ .

Thus, we can describe  $\Gamma$  in the same manner:

$$\Gamma = \begin{bmatrix} \Gamma_0 & \Gamma_0 & \Gamma_0 & \Gamma_0 & \dots \\ \Gamma_0 & \Gamma_0 + \Delta\Gamma_1 & \Gamma_0 + \Delta\Gamma_1 & \Gamma_0 + \Delta\Gamma_1 & \dots \\ \Gamma_0 & \Gamma_0 + \Delta\Gamma_1 & \Gamma_0 + \Delta\Gamma_1 + \Delta\Gamma_2 & \Gamma_0 + \Delta\Gamma_1 + \Delta\Gamma_2 & \dots \\ \Gamma_0 & \Gamma_0 + \Delta\Gamma_1 & \Gamma_0 + \Delta\Gamma_1 + \Delta\Gamma_2 & \Gamma_0 + \Delta\Gamma_1 + \Delta\Gamma_2 + \Delta\Gamma_3 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix}.$$

there are too many of these groups or they are too far away (in time, or to be more specific, in number of meiosis) from genotypes to be estimated accurately.

Here, we show how the previous method for estimating MF parameters for separate populations by pseudo-EM combines with previous work [8, 10] on how to estimate MF structured by year of birth, with the same objectives (although different methods) than Kudinov et al. [6, 9]. First, we model the change in relationship across individuals with time. Then, we plug-in the methods from the previous sections.

For a closed population, the change of mean in time can be expressed as  $\mu_t = \mu_{t-1} + \epsilon_t$ , when  $Var(\epsilon)$  is described by coancestry, this leads to the expression [18]:

$$Var \begin{pmatrix} \mu_0 \\ \mu_1 \\ \mu_2 \\ \mu_3 \\ \dots \end{pmatrix} = \begin{bmatrix} \bar{A}_0 & \bar{A}_0 & \bar{A}_0 & \bar{A}_0 & \dots \\ \bar{A}_0 & \bar{A}_1 & \bar{A}_1 & \bar{A}_1 & \dots \\ \bar{A}_0 & \bar{A}_1 & \bar{A}_2 & \bar{A}_2 & \dots \\ \bar{A}_0 & \bar{A}_1 & \bar{A}_2 & \bar{A}_3 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix} = \begin{bmatrix} \bar{A}_0 & \bar{A}_0 & \bar{A}_0 & \bar{A}_0 & \dots \\ \bar{A}_0 & \bar{A}_0 + \Delta\bar{A}_1 & \bar{A}_0 + \Delta\bar{A}_1 & \bar{A}_0 + \Delta\bar{A}_1 & \dots \\ \bar{A}_0 & \bar{A}_0 + \Delta\bar{A}_1 & \bar{A}_0 + \Delta\bar{A}_1 + \Delta\bar{A}_2 & \bar{A}_0 + \Delta\bar{A}_1 + \Delta\bar{A}_2 & \dots \\ \bar{A}_0 & \bar{A}_0 + \Delta\bar{A}_1 & \bar{A}_0 + \Delta\bar{A}_1 + \Delta\bar{A}_2 & \bar{A}_0 + \Delta\bar{A}_1 + \Delta\bar{A}_2 + \Delta\bar{A}_3 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix}$$

and so on. This is simply the covariance structure of the process (similar but not identical to an autoregressive process):

$$\mu_t = \mu_{t-1} + \epsilon_t,$$

$$Var(\mu_0) = \bar{A}_0,$$

We will obtain those elements from pedigree-based inbreeding. We use the equivalence between inbreeding “with MF”  $\Delta F_\gamma$  and inbreeding with “unrelated founders”  $\Delta F$ , such that  $\Delta F_\gamma = \Delta F(1 + \frac{\gamma}{2})$ . Then, we consider the fact that two times average inbreeding is equal to average coancestry:  $\Delta\Gamma_t = 2\Delta F_{\gamma,t+1}$  and we assume that  $\Delta F_\gamma$  is the same across all periods, and our MF are separated by the same time distances (this can easily be modified), leading to:

$$\Gamma = \begin{bmatrix} \Gamma_0 & \Gamma_0 & \Gamma_0 & \Gamma_0 & \dots \\ \Gamma_0 & \Gamma_0 + 2\Delta F_{(\gamma)} & \Gamma_0 + 2\Delta F_{(\gamma)} & \Gamma_0 + 2\Delta F_{(\gamma)} & \dots \\ \Gamma_0 & \Gamma_0 + 2\Delta F_{(\gamma)} & \Gamma_0 + 4\Delta F_{(\gamma)} & \Gamma_0 + 4\Delta F_{(\gamma)} & \dots \\ \Gamma_0 & \Gamma_0 + 2\Delta F_{(\gamma)} & \Gamma_0 + 4\Delta F_{(\gamma)} & \Gamma_0 + 6\Delta F_{(\gamma)} & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix}.$$

This covariance structure can be described in matrix terms as  $\Gamma = \mathbf{1}\mathbf{1}'\gamma_0 + \mathbf{K}\mathbf{K}'\Delta F(1 + \frac{\gamma_0}{2})$ , where  $\gamma_0$  (or  $\Gamma_0$ ) is the self-relationship of the very first metafounder and

$$\mathbf{K} = \begin{bmatrix} 0 & 0 & 0 & 0 & \dots \\ 1 & 0 & 0 & 0 & \dots \\ 1 & 1 & 0 & 0 & \dots \\ 1 & 1 & 1 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix}.$$

Therefore, we have a para-

metric structure for  $\mathbf{\Gamma}$  in which we need only the elements  $\Delta\Gamma_t$ . Note that other definitions of  $\mathbf{K}$  are possible, e.g. with different or fractional time steps.

To consider two populations, we used the following structure:

$$\mathbf{\Gamma} = \begin{bmatrix} \Gamma_{1,1} & \Gamma_{1,1} & \Gamma_{1,1} & \dots & \Gamma_{1,2} & \Gamma_{1,2} & \Gamma_{1,2} & \dots \\ \Gamma_{1,1} & \Gamma_{1,1} + 2\Delta F^{(1)}(1 - \Gamma_{1,1}) & \Gamma_{1,1} + 2\Delta F^{(1)}(1 - \Gamma_{1,1}) & \dots & \Gamma_{1,2} & \Gamma_{1,2} & \Gamma_{1,2} & \dots \\ \Gamma_{1,1} & \Gamma_{1,1} + 2\Delta F^{(1)}(1 - \Gamma_{1,1}) & \Gamma_{1,1} + 4\Delta F^{(1)}(1 - \Gamma_{1,1}) & \dots & \Gamma_{1,2} & \Gamma_{1,2} & \Gamma_{1,2} & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \Gamma_{2,1} & \Gamma_{2,1} & \Gamma_{2,1} & \dots & \Gamma_{2,2} & \Gamma_{2,2} & \Gamma_{2,2} & \dots \\ \Gamma_{2,1} & \Gamma_{2,1} & \Gamma_{2,1} & \dots & \Gamma_{2,2} & \Gamma_{2,2} + 2\Delta F^{(2)}(1 - \Gamma_{2,2}) & \Gamma_{2,2} + 2\Delta F^{(2)}(1 - \Gamma_{2,2}) & \dots \\ \Gamma_{2,1} & \Gamma_{2,1} & \Gamma_{2,1} & \dots & \Gamma_{2,2} & \Gamma_{2,2} + 2\Delta F^{(2)}(1 - \Gamma_{2,2}) & \Gamma_{2,2} + 4\Delta F^{(2)}(1 - \Gamma_{2,2}) & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{bmatrix}.$$

The extension to more populations (e.g. breeds, countries, pathways of selection or combinations thereof) is

immediate i.e.  $\mathbf{\Gamma} = \mathbf{X}\mathbf{\Gamma}_0\mathbf{X}' + \mathbf{K} \begin{bmatrix} \mathbf{I}\Delta F_{(\gamma)}^{(1)} & & \\ & \mathbf{I}\Delta F_{(\gamma)}^{(2)} & \\ & & \dots \end{bmatrix} \mathbf{K}'$ ,

with  $\mathbf{X}$  and  $\mathbf{K}$  defined appropriately. If there are  $n$  populations, the model needs  $n$  values of  $\Delta F$  and  $n(n + 1)/2$

- (1) Start with values of  $\mathbf{\Gamma}$  of the oldest MF only (in the two breeds example above,  $\Gamma_{1,1}, \Gamma_{1,2}, \Gamma_{2,2}$ ). Expand them to full  $\mathbf{\Gamma}$  using the function of  $\Delta F$ .
- (2) Obtain matrix  $\mathbf{H}_{MF(t)}$  as described in one of the three options before, as a function of  $\mathbf{A}_{\mathbf{\Gamma}(t)}$  and observed  $\mathbf{G}$ . Pick up the elements corresponding to the oldest MF, i.e.,  $\Gamma_{1,1}, \Gamma_{1,2}, \Gamma_{2,2}$ . These are the new values. From them, expand to the whole matrix  $\mathbf{\Gamma}$  as above.

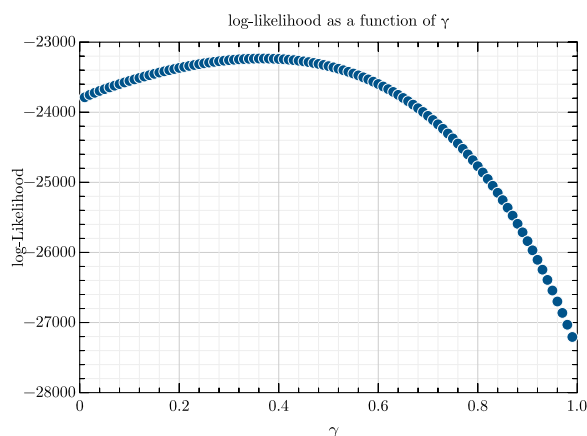
**Numerical example**

This is an example of the pseudo-EM algorithm. We took the two metafounders and 12 animals in [1] with a simulated matrix  $\mathbf{G} = \begin{pmatrix} 1.16 & 0.16 & 0.96 & 0.36 \\ 0.16 & 1.18 & 0.69 & 0.96 \\ 0.96 & 0.69 & 1.80 & 0.87 \\ 0.36 & 0.96 & 0.87 & 0.96 \end{pmatrix}$  and starting value  $\mathbf{\Gamma} = 0.1\mathbf{I}$ . Pedigree is sorted so that MF precede real individuals. The first iteration yields:

$$\mathbf{H}_{MF(0)} = \begin{pmatrix} 0.103 & 0.015 & 0.093 & 0.119 & 0.074 & 0.015 & 0.08 & 0.096 & 0.122 & 0.048 & 0.089 & 0.061 & 0.085 & 0.075 \\ 0.015 & 0.105 & 0.027 & 0.07 & 0.073 & 0.105 & 0.123 & 0.06 & 0.115 & 0.114 & 0.111 & 0.093 & 0.066 & 0.102 \\ 0.093 & 0.027 & 1.019 & 0.058 & 0.001 & 0.027 & 0.242 & 0.515 & 0.018 & 0.135 & 0.219 & 0.068 & 0.258 & 0.143 \\ 0.119 & 0.07 & 0.058 & 1.178 & 0.245 & 0.07 & 0.564 & 0.557 & 0.886 & 0.317 & 0.599 & 0.281 & 0.401 & 0.44 \\ 0.074 & 0.073 & 0.001 & 0.245 & 1.166 & 0.073 & 0.156 & 0.05 & 0.956 & 0.114 & 0.356 & 0.64 & 0.608 & 0.498 \\ 0.015 & 0.105 & 0.027 & 0.07 & 0.073 & 1.054 & 0.123 & 0.06 & 0.115 & 0.589 & 0.111 & 0.331 & 0.066 & 0.221 \\ 0.08 & 0.123 & 0.242 & 0.564 & 0.156 & 0.123 & 1.188 & 0.565 & 0.689 & 0.655 & 0.956 & 0.405 & 0.36 & 0.68 \\ 0.096 & 0.06 & 0.515 & 0.557 & 0.05 & 0.06 & 0.565 & 0.956 & 0.347 & 0.313 & 0.539 & 0.181 & 0.503 & 0.36 \\ 0.122 & 0.115 & 0.018 & 0.886 & 0.956 & 0.115 & 0.689 & 0.347 & 1.81 & 0.402 & 0.867 & 0.679 & 0.651 & 0.773 \\ 0.048 & 0.114 & 0.135 & 0.317 & 0.114 & 0.589 & 0.655 & 0.313 & 0.402 & 1.096 & 0.533 & 0.605 & 0.213 & 0.569 \\ 0.089 & 0.111 & 0.219 & 0.599 & 0.356 & 0.111 & 0.956 & 0.539 & 0.867 & 0.533 & 0.966 & 0.444 & 0.447 & 0.705 \\ 0.061 & 0.093 & 0.068 & 0.281 & 0.64 & 0.331 & 0.405 & 0.181 & 0.679 & 0.605 & 0.444 & 1.109 & 0.411 & 0.777 \\ 0.085 & 0.066 & 0.258 & 0.401 & 0.608 & 0.066 & 0.36 & 0.503 & 0.651 & 0.213 & 0.447 & 0.411 & 1.042 & 0.429 \\ 0.075 & 0.102 & 0.143 & 0.44 & 0.498 & 0.221 & 0.68 & 0.36 & 0.773 & 0.569 & 0.705 & 0.777 & 0.429 & 1.195 \end{pmatrix},$$

values in  $\mathbf{\Gamma}_0$ . It is even possible to consider crossbred cases, e.g. using  $\Delta\Gamma_{1,2}$  elements. Finally, we fit this structure into the pseudo-EM described before. The algorithm (which we call “pseudo-EM with  $\Delta F$ ”) is similar to the pseudo-EM above with the following modifications:

from which the upper left  $2 \times 2$  block, which corresponds to the two MF, is the new estimate  $\hat{\mathbf{\Gamma}}_1 = \begin{pmatrix} 0.103 & 0.015 \\ 0.015 & 0.105 \end{pmatrix}$ . After 14 iterations,  $\hat{\mathbf{\Gamma}} = \begin{pmatrix} 0.408 & 0.367 \\ 0.367 & 0.412 \end{pmatrix}$ .



**Fig. 1** Log-likelihood of genotypes as a function of  $\gamma$

### Tests

#### Maximum likelihood with one metafounder

We used 29,138 genotyped animals of Lacaune dairy sheep [10]. For the sake of experimentation, we considered a single MF. Matrices  $\mathbf{G}$  and  $\mathbf{A}_{22}$  were constructed and written to disk. Then, a Julia program: (1) got ML estimates using the cubic equation described above and (2) did a one-dimensional grid search of the likelihood from functions of  $\mathbf{G}$  and  $\mathbf{A}_{22}$  as detailed in Appendix. The outcome of the program was estimates of  $\gamma$  and an exploration of the log-likelihood  $l(\gamma)$  curve.

Both the cubic equation and the one-dimensional grid search agreed on an ML estimate  $\hat{\gamma} = 0.37$ , however this value does not need to be taken seriously because the pedigree is *not* complete and therefore more than one MF should be used. The cubic equation had only one real solution. The shape of the likelihood is shown in Fig. 1. The likelihood was reasonably peaked, and using a quadratic approximation of the information matrix gives an asymptotic standard error of 0.0755.

### Simulation

We did three simulations. First, two “mixture” simulations (similar to, e.g. a synthetic breed or the introduction of US Holstein into European Friesian) with complete pedigree up to each parental breed. In the “mixture” simulations, we considered a symmetrical case (both breeds have the same genetic drift) and an asymmetrical case (breeds have different genetic drifts) (see details below).

Second, a more complex scenario of “two breeds with groups per year of birth”. This is similar to the current use of dairy cattle where most animals are purebreds but some are crossbreds, e.g. Jersey and Holstein, Nordic dairy cattle breeds, and it is similar to some

sheep breeds where crossbreeding is frequent. This scenario includes two mildly related breeds.

In the “mixture, symmetrical” case, using the program macs [19], we simulated two “cattle” populations of  $N_e = 300$  which split 50 generations ago, resulting in 353,850 polymorphisms with a value of  $F_{ST} = 0.082$ . We selected a subset of 40K loci that were polymorphic in both breeds (with a minor allele frequency (MAF)  $> 0.01$  in both breeds) to declare them “single nucleotide polymorphisms” (SNPs) with a value of  $F_{ST} = 0.079$  (for the 40K SNPs). With these SNPs, we computed  $\mathbf{\Gamma} = \begin{pmatrix} 0.75 & 0.64 \\ 0.64 & 0.75 \end{pmatrix}$ . The SNPs were aggregated into 30 chromosomes of 1 Morgan each. Then, we gene dropped the 40K SNPs in a complex pedigree of 10 generations with 84,200 individuals (200 sires and 4000 dams founders, followed by 10 more generations, with a progeny size of 2), completed at the top with the two MF. Individuals in the first generation were assigned to a breed origin at random with equal probability. Then, matings proceed at random, e.g. the second generation had random proportions (25/50/25) of purebreds and F1 individuals, the third generation has purebreds, F1 and F2 individuals and probably some backcrosses. For instance, looking at animals with “first breed” proportions of 0, 0.25, 0.50, 0.75, 1, each proportion had respectively an animal count of 445, 2092, 3021, 1974 and 468. Eventually, the population becomes a mixture of the two breeds; in the last generation, the proportions of breed 1 oscillate between 0.46 and 0.56. This makes estimation of  $\mathbf{\Gamma}$  more challenging as time advances. There were only two MF corresponding to the two breeds.

Then, the individuals in this complex pedigree were “genotyped” in two different ways that constitute two scenarios. The first scenario (“all generations”) consists in genotyping every 10th animal, i.e. all generations are represented with 8400 genotyped animals. The second scenario (“last generations”) considers the last 2000 of these 8400 animals, i.e. it considers animals of generations 9 to 11.

In the “mixture, unsymmetrical case”, we simulated that, after the split of populations, the  $N_e$  of population 1 was 450, whereas the  $N_e$  of population 2 was 45, leading to  $\mathbf{\Gamma} = \begin{pmatrix} 0.58 & 0.25 \\ 0.25 & 0.65 \end{pmatrix}$ , with all other settings being identical. Again, there were only two MF.

In the “two breeds with groups per year of birth” scenario, the coalescent simulation was as above, using program macs, with the initial  $\mathbf{\Gamma} = \begin{pmatrix} 0.68 & 0.57 \\ 0.57 & 0.68 \end{pmatrix}$ . Then, we plugged the true complex pedigree of dairy sheep Latxa Cara Negra (LCN) and Manech Tête Noire (MTN) [20] with 220 K individuals spanning 30 years





**Fig. 2** Simulated gamma for the “two breeds with groups per year of birth” scenario. The lower left block is one breed and the upper left block is another breed. Metafounders are defined every 3 years within breed

and ~10 generations (e.g. periods of 3 years), which were mainly purebreds with a few sporadic crosses (474 F1 animals, of which 58 rams with at least 10 daughters 1/4 MTN and 3/4 LCN), with missing parentships in all

generations (25% missing sires and 9% missing dams after the initial generation), as it happens in real ruminant populations. There were 20 MF, 10 per breed distributed every three years (i.e. 1 to 10 for breed 1 and 11 to 20 for breed 2). We “gene dropped” markers, in 25 chromosomes of 1 Morgan each, through the pedigree. For animals in the two earliest MF (1 and 11, respectively for each breed) alleles at markers were drawn at random from corresponding allele frequencies. However, for animals in subsequent generations with missing parents, each missing parent was sampled from contemporary animals. After the simulation, we computed allele frequencies for each of the 10 generations within each of the two breeds, and we obtained true  $\Gamma$  of size  $20 \times 20$  from the cross-product  $\Gamma_{b,b'} = \frac{2}{k}(2\mathbf{p}_b - \mathbf{1})(2\mathbf{p}_{b'} - \mathbf{1})'$  with  $\mathbf{p}_b$  and  $\mathbf{p}_{b'}$  being row vectors, as shown in Fig. 2. Again, the scenario “all generations” considered 10% genotyped animals across all generations, for a total of 22,433 animals, and we also considered a “last generations” scenario of 2000 animals corresponding (roughly) to the last three generations.

We applied the generalized least squares (GLS) algorithm [3] and the pseudo-EM algorithm (the stopping criterion was  $10^{-6}$ ) to all these scenarios. The GLS algorithm was applied in its raw form, e.g. there was no correction for estimates of allele frequencies or of  $\Gamma$  that were outside the boundaries, and also, we

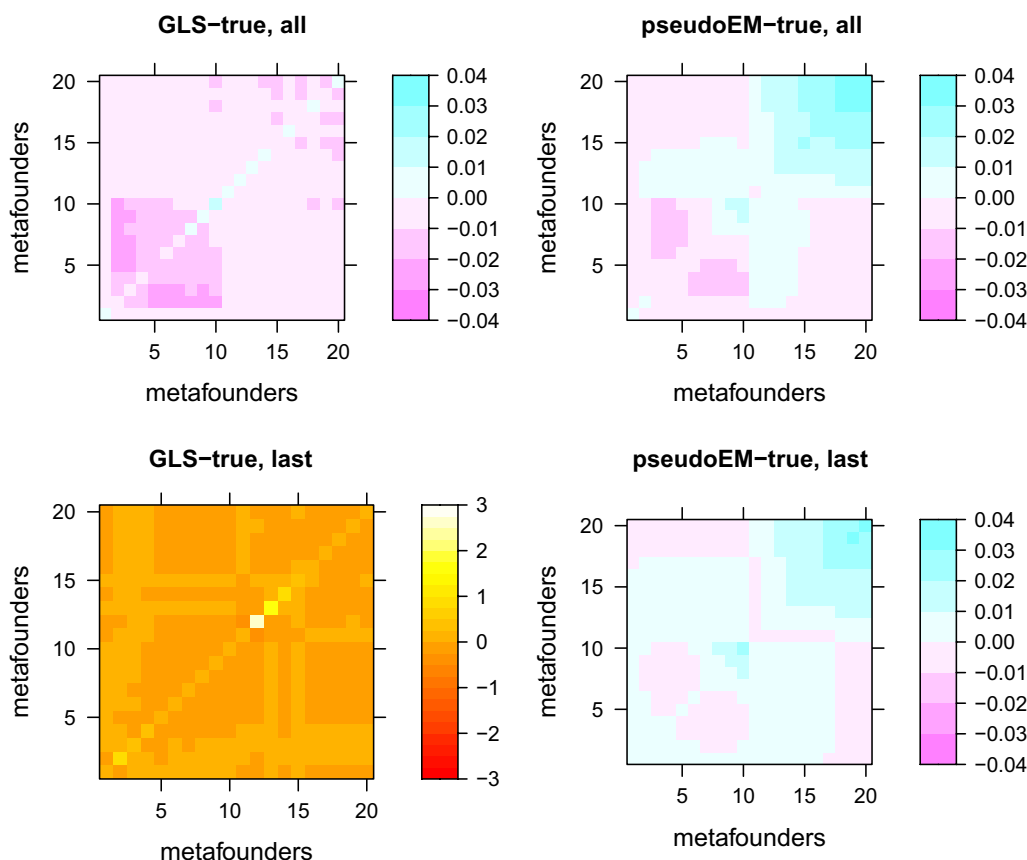
**Table 1** True (simulated) and estimates of gamma using GLS or pseudo-EM and using animals from all generations (“all”) or from the last two generations (“last”)

	Symmetrical					Asymmetrical				
	True	All		Last		True	All		Last	
		GLS	Pseudo-EM	GLS	Pseudo-EM		GLS	Pseudo-EM	GLS	Pseudo-EM
$\Gamma_{1,1}$	0.75	0.75	0.74	1.07	0.76	0.58	0.58	0.58	0.96	0.58
$\Gamma_{1,2}$	0.64	0.63	0.64	0.31	0.64	0.25	0.24	0.25	-0.13	0.25
$\Gamma_{2,2}$	0.75	0.75	0.74	1.09	0.72	0.65	0.65	0.65	1.06	0.64

**Table 2** Statistics of true and estimated values of gamma for the “two breeds with groups per year of birth” scenario

	All		Last	
	GLS	Pseudo-EM+ $\Delta F$	GLS	Pseudo-EM+ $\Delta F$
Correlation, diagonal	0.967	0.783	-0.448	0.897
Correlation, off-diagonal	0.998	0.990	0.914	0.996
Median (estimator-true), diagonal	0.000	0.007	0.045	0.009
Median (estimator-true), off-diagonal	-0.007	0.000	-0.005	0.002
Maxdiff (estimator-true), diagonal	0.016	0.036	2.542	0.031
Maxdiff (estimator-true), off-diagonal	-0.022	0.033	-0.278	0.029

Maxdiff is the maximum absolute difference, with sign



**Fig. 3** Difference between estimated and true gamma when genotyped animals are distributed in “all” generations or in the three “last” generations. Simulation “two breeds with groups per year of birth”

did not attempt a  $GLS+\Delta F$  method. In the “two breeds with groups per year of birth” case, we used the model pseudo-EM+ $\Delta F$ , using literature values of pedigree-based  $\Delta F$  of 0.0021 and 0.0016 per year for MTN and LCN [21, 22], and a time interval of 3 years between each MF. Note that although we do not use genotypes from either of these breeds, we do use their real pedigrees and hence the literature estimates are adequate.

**Results from simulation**

Table 1 shows the results with the “mixture” case. When genotyped animals are present in all generations, both GLS and pseudo-EM estimate  $\Gamma$  correctly. However, when information is available only for the last generations, GLS tends to overestimate the diagonal of  $\Gamma$ , because errors in the estimate of allele frequencies, when squared, cumulate in the diagonal. Regarding the values of the off-diagonal elements of  $\Gamma$ , they tend to be underestimated because errors across two MF tend to cancel out, i.e.  $\hat{p}_b\hat{p}_{b'} < p_b p_{b'}$ .

Results of “two breeds with groups per year of birth” scenario are in Table 2 and Figs. 2 and 3. Figure 2 shows

that the true, simulated relationships in  $\Gamma$  are structured within- and across-populations, and there is a slow increase in  $\Gamma$  due to increased coancestry within breed. Values go from 0.69 to 0.74 (MTN) or 0.70 (LCN) within breed, and are  $\sim 0.57$  across breeds. The simulated  $\Gamma$  increases with time as expected due to increased coancestry within the breed.

Table 2 and Fig. 3 (note that in Fig. 3 the scales differ for each panel) show the performance of the estimates. In the “all generations genotyped” scenario, both GLS and pseudo-EM+ $\Delta F$  are very accurate. GLS underestimates off-diagonal relationships in the first breed, because (again) estimated allele frequencies are not perfect, whereas pseudo-EM+ $\Delta F$  overestimates them in the second breed, probably because the literature estimates of  $\Delta F$  that we used do not consider correctly missing pedigrees, which is shown by the bias being larger for the block that would correspond to LCN (upper right) that has more missing pedigree. In any case, both GLS and pseudo-EM+ $\Delta F$  estimates of  $\Gamma$  should perform adequately for genetic evaluations because the differences with the true  $\Gamma$  are very small.

When only “last” generations have individuals with genotypes, GLS does not provide a good estimate because, although it does capture the block structure of the two populations, it has values that are too high in the diagonal, exceeding the biological limits of 2. On the contrary, pseudo-EM+ $\Delta F$  obtains values that are quite close to true values, even if there are always small biases towards the ends. It is also worth mentioning that pseudo-EM errors are of the same (small) magnitude in the “all” and “last” scenarios, which is not true for GLS, that has a performance that is strongly affected by distance of the estimated MF to genotyped individuals.

Convergence to the required value was fast, about  $\sim 7$  iterations for all simulations, except for the “asymmetrical, last generations genotyped” scenario, which took 59 iterations.

In all three simulations, we genotyped every 10th individual within each genotyped generation, which ensures a homogeneous genotyping. This is not true in real life where elite animals and lines are over-represented in the genotypes, which might lead to biases in the estimation of  $\Gamma$ .

## Discussion

The lack of a general method to estimate  $\Gamma$  is a real problem for use of MF in genetic evaluations. Indeed, several studies used ad hoc techniques to estimate  $\Gamma$ , e.g. [6, 8] and the lack of a general method is a frequent complaint. In our experience, the simple GLS method yields some parts of  $\Gamma$  that are well estimated (i.e. for large breeds) whereas other parts are not (e.g. for small breeds).

Our method of ML for one population is rather simple and not more expensive computationally than the GLS method. Although a bit more complex to implement, the ML method provides estimates that are guaranteed to be in the parametric space and more robust than the GLS method estimates. For the cases where there are no genetic groups or MF, Garcia-Baccino et al. [3] proved that using a single MF was better than the current methods of “tuning” the  $\mathbf{G}$  matrix [23], and that the ML method or the GLS method provided good estimates. Another option is, of course, to use base allele frequencies to build  $\mathbf{G}$ .

Our methods of pseudo-EM and pseudo-EM+ $\Delta F$  do a good job, in particular when information is “asymmetrical” or genotypes exist only in the last generations. Our method uses already available  $\Delta F$  based on pedigree. It may be argued that  $\Delta F$  changes with time, but past  $\Delta F$  does not change, and the method can easily accommodate different  $\Delta F$  along time by changing  $\mathbf{K}$ ,  $\Delta F$ , or both. Estimation of  $\Delta F$  is challenging in itself [7, 24] and underestimation of  $\Delta F$  will result in values of  $\Gamma$  too close

to each other. In our sheep-based example, we have not applied any particular technique for correction of missing pedigree and the results are reasonable, which seems to imply that the method is robust to a small number of missing pedigree records.

An alternative method by Kudinov et al. [6, 9] models  $\Gamma$  using covariance functions in a rather general manner, i.e. it would be of the form  $\Gamma = \Phi \Gamma_0 \Phi'$  where  $\Phi$  is a matrix that is similar to our  $\mathbf{K}$  matrices and  $\Gamma_0$  can be estimated from data. This method is expected to properly describe the increase in coancestry in closed populations, provided enough information (genotypes) across time is given, which may not be true in all cases, for instance, for beef or sheep. It is possibly of interest to dig into the similarities of the two methods and combine them.

Another alternative method in the literature is GLS to obtain allele frequencies [3, 17, 25], followed by the equation  $\hat{\Gamma}_{b,b'} = 8Cov(\hat{p}_b, \hat{p}_{b'})$  [3]. We have improved this method in two manners. First, we use a better definition of  $\Gamma_{b,b'} = \left(\frac{2}{k}\right)(2\mathbf{p}_b - \mathbf{1})(2\mathbf{p}_{b'} - \mathbf{1})'$ , where  $\mathbf{p}_b$  and  $\mathbf{p}_{b'}$  are row vectors of allele frequencies, which does not rely on random coding of alleles. Second, the equation  $\hat{\Gamma}_{b,b'} = 8Cov(\hat{p}_b, \hat{p}_{b'})$  uses estimated  $\hat{p}$  in place of true  $p$  and this leads to cumulation of errors within MF (upward bias in the diagonal) or negative covariance of errors across MF (downward bias off-diagonal) globally leading to biased estimates of  $\Gamma$  for extreme cases (i.e. our simulations with “last” individual genotyped). On the one hand, the GLS method can be refined and made into a sort of GLS+ $\Delta F$  method [10] although we have not attempted to do this. On the other hand, the pseudo-EM strategy is not more expensive computationally and has better theoretical properties.

Compared to GLS, pseudo-EM should be a more robust method because it is an (approximation of) EM algorithm, i.e., it should yield estimates within the parametric space, as far as the approximation is a good one. Pseudo-EM+ $\Delta F$  yielded estimates within the parameter space in the case “two breeds with groups per year of birth”, whereas GLS did not. Computing times of GLS and pseudo-EM (with or without  $\Delta F$ ) are similar if efficiently programmed because both require some form of either  $\mathbf{A}_{\Gamma}^{-1}$  or  $\mathbf{A}^{-1}$  and manipulation of  $\mathbf{G}$  or  $\mathbf{Z}$  depending on the actual form of the algorithms. A difference is that pseudo-EM is an iterative algorithm that requires some iterations, which in our case were mostly a small number ( $\leq 7$ ) but not always (59 for one case).

It is also of interest to present pseudo-EM compared to previous algorithms and true ML. Consider the update

$$\Gamma_{t+1} = \left(\mathbf{A}_{\Gamma(t)}^{mf,mf}\right)^{-1} + \left(\frac{2}{k}\right) \left(2\hat{\mathbf{P}}_{\Gamma(t)} - \mathbf{1}_{npop} \mathbf{1}'_k\right)$$

$(2\widehat{\mathbf{P}}_{\Gamma(t)} - \mathbf{1}_{npop}\mathbf{1}'_k)'$ . The second part of the update,  $(\frac{2}{k})(2\widehat{\mathbf{P}}_{\Gamma(t)} - \mathbf{1}_{npop}\mathbf{1}'_k)(2\widehat{\mathbf{P}}_{\Gamma(t)} - \mathbf{1}_{npop}\mathbf{1}'_k)'$ , corresponds to the GLS estimator of [3], the differences being that the latter: (1) used  $\mathbf{A}$  (not  $\mathbf{A}_{\Gamma}$ ), (2) was not iterated and (3) used the covariance across allele frequencies instead of the cross-product. In addition, the first part of the update  $(\mathbf{A}_{\Gamma(t)}^{mf,mf})^{-1}$  considers the prediction error variance in the prediction of allele frequencies in  $\mathbf{P}$ , i.e. the more the genotyped animals are far from MF, the less the EM algorithm relies on estimates of allele frequencies. The prediction error variance will be different for MF that have less information in genotyped animals.

This expression also allows us to see that, the left part of the update  $(\mathbf{\Gamma}_t - \mathbf{A}_{mf,2\Gamma(t)}\mathbf{A}_{22\Gamma(t)}^{-1}\mathbf{A}_{2,mf\Gamma(t)})$  is the prediction error covariance matrix of genotypes for MF given the observed  $\mathbf{G}$  [15], and this prediction error covariance which was included in the approximate ML estimator suggested (but not actually used) by Garcia-Baccino et al. [3].

Matrix  $\mathbf{A}_{2,mf\Gamma(t)}$  in  $\widehat{\mathbf{\Gamma}}_{t+1} = \widehat{\mathbf{\Gamma}}_t + \mathbf{A}_{mf,2\Gamma(t)}\mathbf{A}_{22\Gamma(t)}^{-1}(\mathbf{G} - \mathbf{A}_{22\Gamma(t)})\mathbf{A}_{22\Gamma(t)}^{-1}\mathbf{A}_{2,mf\Gamma(t)}$  is actually a matrix of breed proportions  $\mathbf{A}_{2,mf\Gamma(t)} = \mathbf{Q}_2\widehat{\mathbf{\Gamma}}_t$ , which gives  $\widehat{\mathbf{\Gamma}}_{t+1} = \widehat{\mathbf{\Gamma}}_t + \widehat{\mathbf{\Gamma}}_t\mathbf{Q}'_2\mathbf{A}_{22\Gamma(t)}^{-1}(\mathbf{G} - \mathbf{A}_{22\Gamma(t)})\mathbf{A}_{22\Gamma(t)}^{-1}\mathbf{Q}_2\widehat{\mathbf{\Gamma}}_t$ . At convergence  $\widehat{\mathbf{\Gamma}} = \widehat{\mathbf{\Gamma}}_{t+1} = \widehat{\mathbf{\Gamma}}_t$  which implies that  $\widehat{\mathbf{\Gamma}}$  is the solution to the (non-linear) equation  $\mathbf{0} = \mathbf{Q}'_2\mathbf{A}_{22\Gamma(t)}^{-1}(\mathbf{G} - \mathbf{A}_{22\Gamma(t)})\mathbf{A}_{22\Gamma(t)}^{-1}\mathbf{Q}_2$ .

As for the comparison with true ML, consider this last non-linear equation  $\mathbf{0} = \mathbf{Q}'_2\mathbf{A}_{22\Gamma(t)}^{-1}(\mathbf{G} - \mathbf{A}_{22\Gamma(t)})\mathbf{A}_{22\Gamma(t)}^{-1}\mathbf{Q}_2$  in the case of single MF:  $\mathbf{0} = \mathbf{1}'\mathbf{A}_{22\Gamma(t)}^{-1}(\mathbf{G} - \mathbf{A}_{22\Gamma(t)})\mathbf{A}_{22\Gamma(t)}^{-1}\mathbf{1}$ . After applying the identities to  $\mathbf{A}_{\gamma 22} = \mathbf{A}_{22}(1 - \frac{\gamma}{2})$  and some algebra, the preceding non-linear equation yields the linear equation on  $\gamma$ :

$$1 - \frac{\gamma}{2} + \gamma\mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{1} = \frac{\mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{1}}{\mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{1}}.$$

The solution of this equation is an estimate  $\widehat{\gamma}$ . However, we note that compared to ML, the term  $Tr(\mathbf{A}_{22}^{-1}\mathbf{G})$  does not appear here. Therefore, the solution to this equation is not the ML estimate. We also note that for  $n$  unrelated individuals  $\mathbf{A}_{22} = \mathbf{I}_n$ , and we obtain  $\gamma \approx \frac{1}{n^2}\mathbf{1}'\mathbf{G}\mathbf{1} = mean(\mathbf{G})$  as expected.

Last, it has to be recalled that the Mendelian sampling variances  $D_{i,i}$  contain likelihood information about  $\mathbf{\Gamma}$ , which is used in true ML but the maximization of which is unclear in pseudo-EM. For instance, F1 and F2

individuals  $\mathbf{A}\times\mathbf{B}$  or  $(\mathbf{A}\times\mathbf{B})\times(\mathbf{A}\times\mathbf{B})$ , in spite of having the same breed proportions will have different Mendelian sampling variances of their respective gametes [13].

## Conclusions

The theory of MF allows a general method that accommodates pedigree and genomic relationships correctly. However, its use demands estimation of relationships across base populations ( $\mathbf{\Gamma}$ ) which is complex, in particular in the complex pedigrees used in livestock genetics. Using Gaussian likelihoods, we derived ML, pseudo-EM and pseudo-EM+ $\Delta F$  methods to estimate  $\mathbf{\Gamma}$  in many realistic settings. These methods require either set up and comparison of genomic and pedigree relationships, or use of allele frequency estimates based on observed markers and pedigree, sometimes completed with additional information (evolution of inbreeding) from pedigrees. Computational cost is therefore low. Estimates are accurate in real and simulated data. These methods will help testing and using MF for genetic evaluations in livestock species.

## Appendix

### Computational tricks to obtain the likelihood and parts of the maximization algorithm

#### Likelihood

The likelihood is composed of two parts,  $log(det(\mathbf{A}_{\Gamma 22}))$  and  $Tr((\mathbf{A}_{\Gamma 22})^{-1}\mathbf{G})$ . The value of  $log(det(\mathbf{A}_{\Gamma 22}))$  can be computed as follows:

(option 1) get log-determinant of  $\mathbf{A}_{\Gamma 22}$  using:

- (1) Compute  $\mathbf{A}_{\Gamma 22}$
- (2)  $\mathbf{L} = cholesky(\mathbf{A}_{\Gamma 22})$  i.e., the lower triangular factor of its Cholesky decomposition.
- (3)  $ln|\mathbf{A}_{\Gamma 22}| = 2 \sum log L_{ii}$  where  $L_{ii}$  are the diagonals of the  $\mathbf{L}$  Cholesky factor.

(option 2)  $ln|\mathbf{A}_{\Gamma 22}|$  can be obtained from sparse matrices  $\mathbf{A}_{\Gamma}^{-1}$  and  $\mathbf{A}_{\Gamma}^{11}$  (the (sparse) block of the (sparse)  $\mathbf{A}_{\Gamma}^{-1}$  corresponding to non-genotyped individuals and meta-founders) using partitioned matrix theory as:

$$log|\mathbf{A}_{\Gamma 22}| = log|\mathbf{A}_{\Gamma}^{11}| - log|\mathbf{A}_{\Gamma}^{-1}|,$$

where  $log|\mathbf{A}_{\Gamma}^{11}|$  is computed e.g. using sparse inversion. The  $log|\mathbf{A}_{\Gamma}^{-1}|$  is computed as follows.  $\mathbf{A}_{\Gamma}^{-1} = \mathbf{T}^{-1}\mathbf{D}^{-1}\mathbf{T}'^{-1}$  and therefore  $|\mathbf{A}_{\Gamma}^{-1}| = |\mathbf{T}^{-1}||\mathbf{D}^{-1}||\mathbf{T}'^{-1}| = |\mathbf{D}^{-1}|$  because  $\mathbf{T}^{-1}$  is a triangular matrix with a diagonal of 1 s.

Matrix  $|\mathbf{D}^{-1}|$  contains two parts, the part linked to MF and the part linked to “animals”:

$$|\mathbf{D}^{-1}| = |\mathbf{\Gamma}^{-1}| \times \prod_i \frac{1}{d_{ii}},$$

where  $1/d_{ii}$  are the contributions summed in Henderson’s algorithm for the  $\mathbf{A}$ -inverse, but here depends on  $\mathbf{\Gamma}$ . Thus:

$$\log|\mathbf{A}_{\Gamma}^{-1}| = \log|\mathbf{\Gamma}^{-1}| + \sum \log\left(\frac{1}{d_{ii}}\right),$$

where the summation extends to all animals in the pedigree.

The value of  $Tr(\mathbf{A}_{\Gamma 22}^{-1}\mathbf{G})$  can be obtained in at least two different ways.

$$b = Tr(\mathbf{A}_{\Gamma 22}^{-1}\mathbf{G}) = \sum_i \sum_j \left( (\mathbf{A}_{\Gamma 22}^{-1})_{ij} \mathbf{G}_{ij} \right),$$

in other words, as the sum of elements of the direct product  $\mathbf{A}_{\Gamma 22}^{-1} \odot \mathbf{G}$ , which is very easy if they have already been computed.

Alternatively, for instance if  $\mathbf{A}_{\Gamma 22}^{-1}$  is not directly computed but products  $\mathbf{A}_{\Gamma 22}^{-1}\mathbf{z}$  can be computed [26]:

$$b = Tr(\mathbf{A}_{\Gamma 22}^{-1}\mathbf{G}) = Tr\left(\mathbf{A}_{\Gamma 22}^{-1}\mathbf{Z}\mathbf{Z}'\frac{2}{k}\right) = \frac{2}{k} \sum \mathbf{z}'_j \mathbf{A}_{\Gamma 22}^{-1} \mathbf{z}_j,$$

where the summation includes  $j = 1, \dots, k$  loci.

**Parts of the single-metafounder exact ML equation**

The value of  $a = \mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{1}$  can be computed as the sum of the elements of  $\mathbf{A}_{22}^{-1}$  (if explicitly computed) or using an indirect method to yield products  $\mathbf{A}_{\Gamma 22}^{-1}\mathbf{1}$ .

$$\begin{aligned} \log_L(\gamma) = & -\left(\frac{k}{2}\right)\log(\det(\mathbf{A}_{22})) - \left(\frac{k}{2}\right)(n-1)\log\left(1 - \frac{\gamma}{2}\right) - \left(\frac{k}{2}\right)\log\left(1 - \frac{\gamma}{2} + \gamma a\right) \\ & - \left(\frac{k}{2}\right)\frac{b}{\left(1 - \frac{\gamma}{2}\right)} + \left(\frac{k}{2}\right)c\frac{\gamma}{\left(1 - \frac{\gamma}{2}\right)\left(1 - \frac{\gamma}{2} + \gamma a\right)}. \end{aligned}$$

For the value of  $b$ , see above.

The value of  $c = \mathbf{1}'\mathbf{A}_{\Gamma 22}^{-1}\mathbf{G}\mathbf{A}_{\Gamma 22}^{-1}\mathbf{1}$  can be obtained explicitly if matrices are available or indirectly, noting that  $c = \mathbf{1}'\mathbf{A}_{\Gamma 22}^{-1}\mathbf{G}\mathbf{A}_{\Gamma 22}^{-1}\mathbf{1} = \mathbf{1}'\mathbf{A}_{\Gamma 22}^{-1}\left(\frac{2}{k}\mathbf{Z}\mathbf{Z}'\right)\mathbf{A}_{\Gamma 22}^{-1}\mathbf{1} = \frac{2}{k}\mathbf{t}'\mathbf{t} =$  where  $\mathbf{t}$  is a vector of size  $k$  number of markers,  $\mathbf{t}' = \mathbf{1}'\mathbf{A}_{\Gamma 22}^{-1}\mathbf{Z}$ . Elements of  $\mathbf{t}$  can be computed by pre-computing the vector  $\mathbf{q} = \mathbf{1}'\mathbf{A}_{\Gamma 22}^{-1}$  and then looping through markers as  $t_j = \mathbf{q}'\mathbf{z}_j$  or directly, in a loop similar as above, as:

$$c = \frac{2}{k} \sum_j (\mathbf{q}'\mathbf{z}_j)^2.$$

**Exact ML for one metafounder**

Consider  $\log_L(\mathbf{\Gamma}) = -(k/2)\log(\det(\mathbf{A}_{\Gamma 22})) - (k/2)Tr\left((\mathbf{A}_{\Gamma 22})^{-1}\mathbf{G}\right)$ . Using formulas from [27], we obtain:

$$(\mathbf{A}_{\gamma 22})^{-1} = \mathbf{A}_{22}^{-1}(1 - \gamma/2) - \frac{\mathbf{A}_{22}^{-1}\mathbf{1}\mathbf{1}'\gamma}{\left(\left(1 - \frac{\gamma}{2}\right)\left(1 - \frac{\gamma}{2} + \gamma\mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{1}\right)\right)},$$

$$\begin{aligned} \det(\mathbf{A}_{\gamma 22}) &= \det(\mathbf{A}_{22}^{-1})(1 - \gamma/2)^n \left(1 + \mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{1}/(1 - \gamma/2)\right) \\ &= \det(\mathbf{A}_{22}^{-1})(1 - \gamma/2)^{n-1} \left(1 + \mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{1}\right). \end{aligned}$$

Inserting those into the log-likelihood function, and using short-hand notation  $a = \mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{1}$ ,  $b = Tr(\mathbf{A}_{22}^{-1}\mathbf{G})$  and  $c = Tr(\mathbf{A}_{22}^{-1}\mathbf{1}\mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{G}) = \mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{1}$ , i.e. the sum of all elements of  $\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}$ , we obtain:

The ML estimate is obtained by differentiating  $l(\gamma)$  and setting equal to zero, where we note that the  $k/2$  in front of each term in the expression can be ignored,

$$\begin{aligned} 0 = & \frac{(n-1)/2}{(1 - \gamma/2)} - \frac{-1/2 + a}{1 - \gamma/2 + \gamma a} - \frac{b/2}{(1 - \gamma/2)^2} \\ & + c\frac{\left(1 - \frac{\gamma}{2}\right)\left(1 - \frac{\gamma}{2} + \gamma a\right) - \gamma\left(\left(1 - \frac{\gamma}{2}\right)\left(-\frac{1}{2} + a\right) - \frac{\left(1 - \frac{\gamma}{2} + \gamma a\right)}{2}\right)}{(1 - \gamma/2)^2(1 - \gamma/2 + \gamma a)^2}. \end{aligned}$$

Multiplying by  $((1 - \gamma/2)^2)(1 - \gamma/2 + \gamma a)^2$  on both sides of the equation followed by some algebraic manipulations, we obtain a cubic equation:

$$e_3\gamma^3 + e_2\gamma^2 + e_1\gamma + e_0 = 0,$$

where  $e_3 = -n(-1/2 + a)^2/2$ ,  $e_2 = n((-3/2 + a) + a - b(-1/2 + a) + c)(-1/2 + a)$ ,  $e_1 = (n - 1)(-3/2 + 2a) - (-1 + 2a)(-3/2 + a + b)$  and  $e_0 = n - 2a - b + 2c$ .

Solving a cubic equation, first requires the computation of the discriminant  $\Delta = 18e_3e_2e_1e_0 - 4e_2^3e_0 + e_2^2e_1^2 - 4e_3e_1^3 + 27e_3^2e_0^2$ . If  $\Delta > 0$ , then there are three distinct real roots, and if  $\Delta < 0$ , then there is only one real root, which is the ML estimate of  $\gamma$  (if in the parametric space), and two complex roots.

**Derivatives of the likelihood**

**The first derivative of  $Tr(\mathbf{A}_\Gamma^{-1}\mathbf{G})$**

Based on the definition of trace, and using rules for differentiation of a trace [28]:

$$\begin{aligned} Tr(\mathbf{A}_\Gamma^{-1}\mathbf{G}) &= Tr(\mathbf{T}^{-1'}\mathbf{D}_\Gamma^{-1}\mathbf{T}^{-1}\mathbf{G}) \\ &= Tr(\mathbf{D}_\Gamma^{-1}\mathbf{T}^{-1}\mathbf{G}\mathbf{T}^{-1'}) \\ &= Tr\left(\mathbf{D}_\Gamma^{-1}\mathbf{T}^{-1}\frac{\mathbf{Z}\mathbf{Z}'}{\frac{k}{2}}\mathbf{T}^{-1'}\right) \\ &= Tr(\mathbf{D}_\Gamma^{-1}\mathbf{W}\mathbf{W}') \\ &= Tr(\mathbf{D}_\Gamma^{-1}\mathbf{B}), \end{aligned}$$

where  $\mathbf{W} = \mathbf{T}^{-1}\mathbf{Z}/\sqrt{0.5k}$ ,  $\mathbf{B} = \mathbf{W}\mathbf{W}'$ . Note that the block of  $\mathbf{T}^{-1}$  corresponding to MF is simply an identity matrix, and thus the block of  $\mathbf{B}$  corresponding to MF is simply  $\mathbf{G}_{MF}$ . Thus,

$$\frac{\partial Tr(\mathbf{A}_\Gamma^{-1}\mathbf{G})}{\partial \gamma} = \frac{\partial Tr(\mathbf{D}_\Gamma^{-1}\mathbf{B})}{\partial \gamma} = Tr\left(\frac{\partial \mathbf{D}_\Gamma^{-1}}{\partial \gamma}\mathbf{B}\right).$$

**The first derivative of  $\log(\det(\mathbf{A}_\Gamma))$**

Based on the definition of Jacobi formula (see for instance chap. 15, 8.5 in [28]):

$$\begin{aligned} \frac{\partial \det(\mathbf{A}_\Gamma)}{\partial \gamma} &= \det(\mathbf{A}_\Gamma)Tr\left(\mathbf{A}_\Gamma^{-1}\frac{\partial \mathbf{A}_\Gamma}{\partial \gamma}\right) \\ &= \det(\mathbf{A}_\Gamma)Tr\left(\mathbf{T}^{-1'}\mathbf{D}_\Gamma^{-1}\mathbf{T}^{-1}\frac{\partial \mathbf{A}_\Gamma}{\partial \gamma}\right) \\ &= \det(\mathbf{A}_\Gamma)Tr\left(\mathbf{T}^{-1'}\mathbf{D}_\Gamma^{-1}\mathbf{T}^{-1}\mathbf{T}\frac{\partial \mathbf{D}_\Gamma}{\partial \gamma}\mathbf{T}'\right), \\ &= \det(\mathbf{A}_\Gamma)Tr\left(\mathbf{T}'\mathbf{T}^{-1'}\mathbf{D}_\Gamma^{-1}\mathbf{T}^{-1}\mathbf{T}\frac{\partial \mathbf{D}_\Gamma}{\partial \gamma}\right) \\ &= \det(\mathbf{A}_\Gamma)Tr\left(\mathbf{D}_\Gamma^{-1}\frac{\partial \mathbf{D}_\Gamma}{\partial \gamma}\right) \end{aligned}$$

where we note that  $\partial D_\Gamma/\partial \gamma$  depends on which parameter the function is differentiated with respect to, but it does not depend on the actual values of any of the parameters. Therefore,

$$\frac{\partial \log(\det(\mathbf{A}_\Gamma))}{\partial \gamma} = \frac{1}{\det(\mathbf{A}_\Gamma)} \frac{\partial \det(\mathbf{A}_\Gamma)}{\partial \gamma} = Tr\left(\mathbf{D}_\Gamma^{-1}\frac{\partial \mathbf{D}_\Gamma}{\partial \gamma}\right).$$

**The first derivative of  $\log L(\Gamma)$**

Based on the preceding equations, the first derivative of log-likelihood function would be:

$$\begin{aligned} \frac{\partial \log_L(\Gamma)}{\partial \gamma} &= -\frac{k}{2} \left[ \frac{\partial \log(\det(\mathbf{A}_\Gamma))}{\partial \gamma} + \frac{\partial Tr(\mathbf{A}_\Gamma^{-1}\mathbf{G})}{\partial \gamma} \right] \\ &= -\frac{k}{2} \left[ Tr\left(\mathbf{D}_\Gamma^{-1}\frac{\partial \mathbf{D}_\Gamma}{\partial \gamma}\right) + Tr\left(\frac{\partial \mathbf{D}_\Gamma^{-1}}{\partial \gamma}\mathbf{B}\right) \right] \\ &= -\frac{k}{2} \left[ Tr\left(\frac{\partial \mathbf{D}_\Gamma^{-1}}{\partial \gamma}(\mathbf{B} - \mathbf{D}_\Gamma)\right) \right] \end{aligned}$$

The structure of  $\mathbf{B}$  and  $\mathbf{D}_\Gamma$  is:

$$\mathbf{D}_\Gamma = \begin{cases} \mathbf{\Gamma}, & \text{for metafounders} \\ \text{diagonal matrix,} & \text{for non-metafounders} \end{cases}$$

$$\mathbf{B} = \begin{cases} \mathbf{G}_{MF}, & \text{for metafounders} \\ \mathbf{W}_{NMF}\mathbf{W}_{NMF}', & \text{for non-metafounders} \end{cases}$$

Thus,  $\frac{\partial \log_L(\Gamma)}{\partial \gamma}$  can be expressed as follows:

$$\begin{aligned} \frac{\partial \log_L(\Gamma)}{\partial \gamma} &= -\frac{k}{2} \left[ Tr\left[\frac{\partial \mathbf{\Gamma}^{-1}}{\partial \gamma}(\mathbf{G}_{MF} - \mathbf{\Gamma})\right] \right. \\ &\quad \left. - \frac{k}{2} \left[ \sum_i \left[ \frac{\partial D_{\Gamma_{ii}}^{-1}}{\partial \gamma} (B_{i,i} - D_{\Gamma_{ii}}) \right] \right] \right], \end{aligned}$$

where sum over  $i$  is for non-metafounders and  $\frac{\partial \mathbf{\Gamma}^{-1}}{\partial \gamma} = -\mathbf{\Gamma}^{-1}\frac{\partial \mathbf{\Gamma}}{\partial \gamma}\mathbf{\Gamma}^{-1}$  and  $\frac{\partial D_{\Gamma_{ii}}^{-1}}{\partial \gamma} = -\frac{\partial D_{\Gamma_{ii}}}{\partial \gamma}/(D_{\Gamma_{ii}})^2$ . Inserting this, we obtain:

$$\frac{\partial \log_L(\Gamma)}{\partial \gamma} = -\frac{k}{2} \left[ \text{Tr} \left[ \Gamma^{-1} \frac{\partial \Gamma}{\partial \gamma} (\mathbf{I} - \Gamma^{-1} \mathbf{G}_{MF}) \right] \right] - \frac{k}{2} \left[ \sum_i \left[ \frac{\partial D_{\Gamma_{i,i}}}{\partial \gamma} \left( \frac{1}{D_{\Gamma_{i,i}}} - \frac{B_{i,i}}{D_{\Gamma_{i,i}}^2} \right) \right] \right].$$

The first term involves  $\mathbf{G}_{MF}$  and is simple to manipulate. The last term implies individual terms of Mendelian sampling variances  $D_{\Gamma_{i,i}}$  and it is difficult to compute and derive algebraically in a recognizable form. Instead, we approximate  $\frac{\partial \log_L(\Gamma)}{\partial \gamma}$  by its first term as:

$$\frac{\partial \log_L(\Gamma)}{\partial \gamma} \approx -\frac{k}{2} \left[ \text{Tr} \left[ \Gamma^{-1} \frac{\partial \Gamma}{\partial \gamma} (\mathbf{I} - \Gamma^{-1} \mathbf{G}_{MF}) \right] \right].$$

### Derivation of the pseudo-EM algorithm

Consider the approximation above:

$$\frac{\partial \log_L(\Gamma)}{\partial \gamma} \approx -\frac{k}{2} \left[ \text{Tr} \left[ \Gamma^{-1} \frac{\partial \Gamma}{\partial \gamma} (\mathbf{I} - \Gamma^{-1} \mathbf{G}_{MF}) \right] \right],$$

setting to 0 and factorizing  $-\frac{k}{2}$  and introducing  $\gamma_{ij}$  to indicate which of the points of  $\Gamma$  we work with gives:

$$\text{Tr} \left[ \Gamma^{-1} \frac{\partial \Gamma}{\partial \gamma_{ij}} \right] = \left[ \text{Tr} \left[ \Gamma^{-1} \frac{\partial \Gamma}{\partial \gamma_{ij}} \Gamma^{-1} \mathbf{G}_{MF} \right] \right].$$

Now we focus on  $\text{Tr} \left[ \Gamma^{-1} \frac{\partial \Gamma}{\partial \gamma_{ij}} \right]$ . In fact,  $\frac{\partial \Gamma}{\partial \gamma_{ij}}$  is a matrix  $E(i, j)$  that contains 1 in the  $(i, j)$  position and 0 otherwise. The direct product  $\Gamma^{-1} \odot E(i, j)$  produces a matrix with 0 s and picking up elements  $(i, j)$  of the matrix  $\Gamma^{-1}$ . Then we use the fact that:

$$\text{Tr}(\mathbf{A}\mathbf{B}) = \sum \sum a_{ij}b_{ij} = \sum_{\text{all elements}} (\mathbf{A} \odot \mathbf{B}),$$

to observe that in fact  $\text{Tr} \left[ \Gamma^{-1} \frac{\partial \Gamma}{\partial \gamma_{ij}} \right] = \text{Tr} \left[ \Gamma^{-1} E(i, j) \right]$  as follows:

for  $i = j$  (diagonal elements) then we obtain the diagonal element  $\gamma^{ij}$  (the element  $(i, j)$  of  $\Gamma^{-1}$ , not of  $\Gamma$ );

for  $i \neq j$  (off-diagonal elements) then we obtain  $2\gamma^{ij}$  (because there are two 1 s then in  $E(i, j)$ ).

Now we focus on the expression  $\text{Tr} \left[ \Gamma^{-1} \frac{\partial \Gamma}{\partial \gamma_{ij}} \Gamma^{-1} \mathbf{G}_{MF} \right] = \text{Tr} \left[ \frac{\partial \Gamma}{\partial \gamma_{ij}} \Gamma^{-1} \mathbf{G}_{MF} \Gamma^{-1} \right]$ . By the same reasoning, when we consider  $\gamma_{ij}$  we obtain:

for  $i = j$  (diagonal elements) we get the element  $(i, j)$  of  $\Gamma^{-1} \mathbf{G}_{MF} \Gamma^{-1}$ , i.e.  $\Gamma^{-1} \mathbf{G}_{MF} \Gamma^{-1} [i, j]$ ;

for  $i \neq j$  (diagonal elements) then we get twice the element  $(i, j)$  of  $\Gamma^{-1} \mathbf{G}_{MF} \Gamma^{-1}$ , i.e.  $2\Gamma^{-1} \mathbf{G}_{MF} \Gamma^{-1} [i, j]$ .

Thus, for each  $(i, j)$  combination we have one (for  $i = j$  in the position  $(i, i)$ ) or two (for  $i \neq j$ , in the positions  $(i, j)$  and  $(j, i)$ ) scalar equations of the form:

$$\gamma^{ij} = \Gamma^{-1} \mathbf{G}_{MF} \Gamma^{-1} [i, j],$$

(note that for  $i \neq j$  the factor of 2 cancels out). Putting all these equations together we obtain:

$$\Gamma^{-1} = \Gamma^{-1} \mathbf{G}_{MF} \Gamma^{-1},$$

which after pre- and post-multiplication by  $\Gamma$  results in:

$$\Gamma = \mathbf{G}_{MF}.$$

### Acknowledgements

The reviewers and the editor are thanked for many details and insightful comments.

### Author contributions

All four authors derived different parts of the theory. AL and MB prepared the examples. AL ran the test with real Lacaune data. The first version was written by AL with input from all four authors. All authors have read and approved the final manuscript.

### Funding

Not applicable.

### Availability of data and materials

Not applicable.

### Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup>CDCB, 4201 Northview Drive, Bowie, MD 20716, USA. <sup>2</sup>Animal and Dairy Science, University of Georgia, 425 River Rd, Athens, GA 30602, USA. <sup>3</sup>Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA. <sup>4</sup>Center for Quantitative Genetics and Genomics, Aarhus University, C. F. Møllers Allé 3, bld. 1130, 8000 Aarhus C, Denmark.

Received: 8 September 2023 Accepted: 18 March 2024

Published online: 02 May 2024

### References

- Legarra A, Christensen OF, Vitezica ZG, Aguilar I, Misztal I. Ancestral relationships using metafounders: finite ancestral populations and across population relationships. *Genetics*. 2015;200:455–68.
- Christensen OF, Legarra A, Lund MS, Su G. Genetic evaluation for three-way crossbreeding. *Genet Sel Evol*. 2015;47:98.
- Garcia-Baccino CA, Legarra A, Christensen OF, Misztal I, Pocrnic I, Vitezica ZG, et al. Metafounders are related to Fst fixation indices and reduce bias in single-step genomic evaluations. *Genet Sel Evol*. 2017;49:34.
- Christensen OF. Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. *Genet Sel Evol*. 2012;44:37.
- Macedo FL, Astruc JM, Meuwissen THE, Legarra A. Removing data and using metafounders alleviates biases for all traits in Lacaune dairy sheep predictions. *J Dairy Sci*. 2022;105:2439–52.

6. Kudinov AA, Koivula M, Aamand GP, Strandén I, Mäntysaari EA. Single-step genomic BLUP with many metafounders. *Front Genet.* 2022;13:1012205.
7. VanRaden PM. Accounting for inbreeding and crossbreeding in genetic evaluation of large populations. *J Dairy Sci.* 1992;75:3136–44.
8. Bermann M, Aguilar I, Lourenco D, Misztal I, Legarra A. Reliabilities of estimated breeding values in models with metafounders. *Genet Sel Evol.* 2023;55:6.
9. Kudinov AA, Mäntysaari EA, Aamand GP, Uimari P, Strandén I. Meta-founder approach for single-step genomic evaluations of Red Dairy cattle. *J Dairy Sci.* 2020;103:6299–310.
10. Wicki M, Raoul J, Legarra A. Effect of subdivision of the Lacaune dairy sheep breed on the accuracy of genomic prediction. *J Dairy Sci.* 2023;106:5570–81.
11. Legarra A, Bermann M, Mei Q, Christensen OF. Short communication: redefining and interpreting the genomic relationships of metafounders. *Genet Sel Evol.* 2024. <https://doi.org/10.1186/s12711-024-00891-w>.
12. Christensen OF, Legarra A. Maximum likelihood estimation of meta-founder parameters for single and multiple breeds. In: *Proceedings of 12th world congress on genetics applied to livestock production: 03–08 July 2022; Rotterdam; 2022.*
13. Garcia-Cortes LA, Toro M. Multibreed analysis by splitting the breeding values. *Genet Sel Evol.* 2006;38:601–15.
14. Quaas RL. Additive genetic model with groups and relationships. *J Dairy Sci.* 1988;71:91–8.
15. Christensen OF, Lund MS. Genomic prediction when some animals are not genotyped. *Genet Sel Evol.* 2010;42:2.
16. Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic information. *J Dairy Sci.* 2009;92:4656–63.
17. Gengler N, Mayeres P, Szydlowski M. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. *Animal.* 2007;1:21–8.
18. Sorensen DA, Kennedy BW. The use of the relationship matrix to account for genetic drift variance in the analysis of genetic experiments. *Theor Appl Genet.* 1983;66:217–20.
19. Chen GK, Marjoram P, Wall JD. Fast and flexible simulation of DNA sequence data. *Genome Res.* 2009;19:136–42.
20. Garcia-Baccino CA, Pineda-Quiroga C, Astruc JM, Ugarte E, Legarra A. High genetic correlation for milk yield across Manech and Latxa dairy sheep from France and Spain. *JDS Commun.* 2022;3:260–4.
21. Granado-Tajada I, Rodríguez-Ramilo S, Legarra A, Ugarte E. Inbreeding, effective population size, and coancestry in the Latxa dairy sheep breed. *J Dairy Sci.* 2020;103:5215–26.
22. Rodríguez-Ramilo S, Elsen JM, Legarra A. Inbreeding and effective population size in French dairy sheep: comparison between genomic and pedigree estimates. *J Dairy Sci.* 2019;102:4227–37.
23. Vitezica Z, Aguilar I, Misztal I, Legarra A. Bias in genomic predictions for populations under selection. *Genet Res (Camb).* 2011;93:357–66.
24. Colleau JJ, Sargolzaei M. MIM: an indirect method to assess inbreeding and coancestry in large incomplete pedigrees of selected dairy cattle. *J Anim Breed Genet.* 2011;128:163–73.
25. McPeck MS, Wu X, Ober C. Best linear unbiased allele-frequency estimation in complex pedigrees. *Biometrics.* 2004;60:359–67.
26. Masuda Y, Misztal I, Legarra A, Tsuruta S, Lourenco DAL, Fragomeni BO, et al. Technical note: avoiding the direct inversion of the numerator relationship matrix for genotyped animals in single-step genomic best linear unbiased prediction solved with the preconditioned conjugate gradient. *J Anim Sci.* 2017;95:49–52.
27. Henderson HV, Searle SR. On deriving the inverse of a sum of matrices. *SIAM Rev.* 1981;23:53–60.
28. Harville DA. *Matrix algebra from a statistician's perspective.* New York: Springer-Verlag; 1997.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.