Ancestral Relationships Using Metafounders: Finite Ancestral Populations and Across Population Relationships

Andres Legarra,*[,1] Ole F. Christensen,[†] Zulma G. Vitezica,[‡] Ignacio Aguilar,[§] and Ignacy Misztal**
*Institut National de la Recherche Agronomique and [‡]Université de Toulouse, INP, ENSAT, GenPhySE, Génétique, Physiologie et Systèmes d'Elevage, F-31326 Castanet-Tolosan, France, [†]Aarhus University, Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, DK-8830 Tjele, Denmark, [§]Instituto Nacional de Investigación Agropecuaria, Canelones 90200, Uruguay, and **Animal and Dairy Science, University of Georgia, Athens, Georgia 30602

ABSTRACT Recent use of genomic (marker-based) relationships shows that relationships exist within and across base population (breeds or lines). However, current treatment of pedigree relationships is unable to consider relationships within or across base populations, although such relationships must exist due to finite size of the ancestral population and connections between populations. This complicates the conciliation of both approaches and, in particular, combining pedigree with genomic relationships. We present a coherent theoretical framework to consider base population in pedigree relationships. We suggest a conceptual framework that considers each ancestral population as a finite-sized pool of gametes. This generates across-individual relationships and contrasts with the classical view which each population is considered as an infinite, unrelated pool. Several ancestral populations may be connected and therefore related. Each ancestral population can be represented as a "metafounder," a pseudo-individual included as founder of the pedigree and similar to an "unknown parent group." Metafounders have self- and across relationships according to a set of parameters, which measure ancestral relationships, *i.e.*, homozygozities within populations and relationships across populations. These parameters can be estimated from existing pedigree and marker genotypes using maximum likelihood or a method based on summary statistics, for arbitrarily complex pedigrees. Equivalences of genetic variance and variance components between the classical and this new parameterization are shown. Segregation variance on crosses of populations is modeled. Efficient algorithms for computation of relationship matrices, their inverses, and inbreeding coefficients are presented. Use of metafounders leads to compatibility of genomic and pedigree relationship matrices and to simple computing algorithms. Examples and code are given.

KEYWORDS relationships; pedigree; genetic drift; base populations; marker genotypes; shared data resource; GenPred

POWELL *et al.* (2010) pointed out the conceptual conflict between identity-by-descent (IBD) relationships based on pedigree and identity-by-state (IBS) relationships based on marker genotypes. These are also known as pedigree and genomic (VanRaden 2008) relationships, respectively, and we use this terminology hereinafter. Whereas reference for pedigree relationships is formed by founders of the pedigree, reference for the genomic relationships is most often the current genotyped population (*e.g.*, Powell *et al.* 2010; Vitezica *et al.* 2011). Powell *et al.* (2010) showed that one can (at least conceptually) refer genomic relationship coefficients to the pedigree scale and vice versa. In the context of applied genetic evaluation of livestock, similar notions were introduced by VanRaden (2008) and Vitezica *et al.* (2011), explicitly modifying genomic relationships to refer to pedigree coefficients. However, an implicit assumption in these proposals is that the genotyped population has no pedigree structure, *e.g.*, no sib groups and only one generation (Christensen 2012), and the proposals are also difficult to extend to several base populations (Harris and Johnson 2010; Misztal *et al.* 2013; Makgahlela *et al.* 2014).

In addition, pedigree relationships have several problems. Pedigrees, which are incomplete by definition, end up in one or several base populations (lines or breeds). For instance, the pedigree of the Romane sheep synthetic breed traces back to

doi: 10.1534/genetics.115.177014
Manuscript received February 12, 2015; accepted for publication April 3, 2015; published Early Online April 14, 2015.
Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.177014/-/DC1.
[1]Corresponding author: INRA, UMR 1388 GenPhySE, CS 52627, 31326 Castanet-Tolosan, France. E-mail: andres.legarra@toulouse.inra.fr

two base populations, the Romanov and Berrichon du Cher breeds, and the pedigree of global Holstein cattle population can often be traced back to "European" and "North American" base populations. In another more complex case, pedigrees are incomplete for some categories of animals. For instance, in dairy sheep, the father of all males is known, but only 5–80% of females have a known father. To further complicate things, in the presence of selection, assuming that all unknown parents belong to the same base population and have the same genetic level is unfair since younger (or sometimes foreign) animals are selected and therefore "better" than the base population. If not properly accounted for, this structure in several base populations results in biases (*e.g.*, Ugarte *et al.* 1996; Misztal *et al.* 2013). Therefore, unknown fathers are assigned to different base populations, *e.g.*, depending on year of birth, country of origin, sex, or path of selection. Current practice of genetic evaluations assumes that individuals in the different base populations (typically known as genetic groups or unknown parent groups) have different *a priori* average values, and these values are estimated as fixed effects within the model (Thompson 1979; Quaas 1988). However, the quantitative genetics theory of unknown parent groups has not been much further developed. For instance, Kennedy (1991) pointed out that genetic groups are incorrectly assumed to be unrelated to each other and that reduction of variance due to drift and selection should be accounted for. With molecular markers, there are more and more examples of observed relationships across base populations that were *a priori* unrelated (Kijaas *et al.* 2009 ; Gibbs *et al.* 2009; Ter Braak *et al.* 2010 ; VanRaden *et al.* 2011).

The hypothesis of unrelatedness of founders in a base population implies that the base population is drawn from a very large ancestral population. Not only is this false but it also contradicts marker-based information (animals seemingly unrelated share alleles at markers). Although unrelatedness can simply be seen as an arbitrary starting point, we suggest that relaxing this hypothesis gives more flexibility to the models.

On the other hand, genomic measures of relationships are not dependent on knowledge of pedigree. Further, they are more accurate because they consider realized, not expected, relationships (VanRaden 2008; Hayes, *et al.* 2009; Hill and Weir 2011). Genomic relationships can be projected along the pedigree for animals with no genotypes (Legarra *et al.* 2009; Christensen and Lund 2010). The so-called single-step GBLUP (SSGBLUP) thus mixes pedigree and genomic relationships, and is becoming the *de facto* standard in genomic evaluations for livestock (*e.g.*, Legarra *et al.* 2014b). However, SSGBLUP requires genomic and pedigree relationship to refer to the same base. This base is however, hard to define. Genomic relationships of the current population change as more individuals are being included and are poorly defined if populations are structured (*i.e.*, in lines, breeds or origins) (*e.g.*, Harris and Johnson 2010). Defining a base is also difficult for pedigree relationships as pedigrees are incomplete and possibly end up in several base populations. An alternative is truncation of the pedigree, to have a more homogeneous base population (*e.g.*, Lourenco *et al.* 2014), but this is not always a feasible option. Furthermore, defining pedigree founders as unrelated is contradictory with results obtained if these individuals are genotyped. Christensen (2012) suggested taking for genomic relationships an arbitrary reference and an ideal population with 0.5 allele frequency at the markers, and referring pedigree relationships to this base population. By doing so, he showed that founders of the base population should become related, and this extra relatedness can be understood as an excess of identical-by-descent homozygosity. The approach can be understood as a marginalization with respect to uncertainty in allelic frequencies, and a stable definition of the genetic base across time and different populations is obtained. Extension of this method to several founder populations is, alas, not straightforward.

In this work, we present a theory to consider relationships within and across base, or founder, populations. This theory provides the tools, on the one hand, to generalize the "unknown parent groups" used in genetic evaluations and, on the other hand, to generalize Christensen's results, which conciliate pedigree and genomic relationships. The concepts developed in this work aim to be rather general and are based on pedigree considerations, but their use is of large interest in two cases: first, when combining genomic and pedigree relationships across individuals (as the SSGBLUP mentioned above) and, second, when considering several base populations simultaneously.

The outline of this article is as follows. First, we show that base populations with related individuals can be understood as issued from finite size ancestral populations. This, although not strictly necessary for practical purposes, gives a conceptual model and a genetic interpretation (Jacquard 1974). Second, such an ancestral population can be represented as a single pseudo-individual (a metafounder) with a particular self-relationship (a measure of homozygosity) and represents a pool of gametes. Several base populations can be represented as several, possibly related, metafounders. Metafounders are convenient because they simplify the representation and the algorithms for computing relationships and inbreeding. Finally, we show how parameters (ancestral homozygosities and relationships across populations) of ancestral populations can be estimated from the combined use of marker and pedigree data. Our work is an extension and unification of existing works by Jacquard (1969, 1974), VanRaden (1992), Aguilar and Misztal (2008), VanRaden *et al.* (2011), Colleau and Sargolzaei (2011), and Christensen (2012).

Theory

Relationships in a finite population

Relationships across base individuals: Let "ancestral" be the population from which founders of the pedigree are drawn and "base" population be the set of these pedigree founders

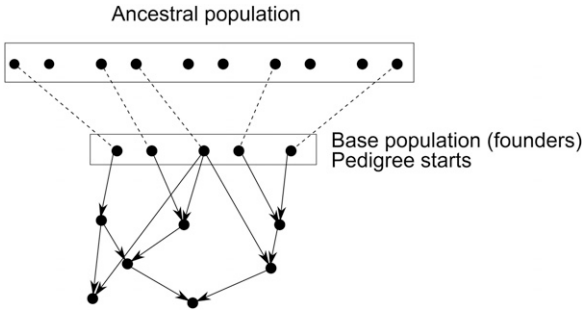

Figure 1 Ancestral and base population and pedigree.

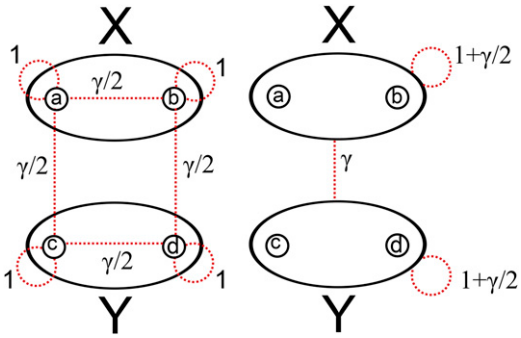

Figure 2 Relationships (red dashed line), in a related base population, of two individuals X and Y seen as gametes (left) or individuals (right). Individual X (Y) contains gametes a and b (c and d). Relationships across and within gametes are respectively $\boldsymbol{\gamma}/2$ and 1; relationships across and within individuals are $\boldsymbol{\gamma}$ and $1+\boldsymbol{\gamma}/2$.

(Figure 1). Typically, individuals in the base population are assumed to be drawn from a large, unrelated, ancestral population mating at random, so that the base population individuals will not be related. Jacquard (1969, 1974) considered relationships in a finite-size population, showing that inbreeding and relationships increase steadily. We redevelop his treatment in a simplified form. Pedigree founders in the base population are drawn at random, with replacement, from an ancestral finite monoecious population with effective size N_e, $2N_e$ gametes, "true" average breeding value μ, and genetic variance σ^2_u. In this ancestral population gametes are assumed to be independent (in a sense, the ancestral population becomes the new base). Imagine two gametes sampled with replacement from the finite ancestral population to form the base population. The second gamete will be identical to the first gamete $1/(2N_e)$ of the times. Therefore, the relationship coefficient (probability of identity by descent) between all pairs of gametes is $\gamma/2 = 1/2N_e$, and this relationship $\gamma/2$ can be understood as the correlation between gametes of Wright (1922). Jacquard (1974) used $\alpha = \gamma/2$ and called it the "inbreeding coefficient of a population."

Across-individual relationships in the base population are depicted in Figure 2. Consider diploid individuals *X* and *Y*. They are constituted by four gametes, a,b,c,d. These gametes have been drawn from a pool of gametes where the probability of being identical (by descent) is $\gamma/2$ across gametes and 1 with itself (Figure 2, left). Therefore, the coancestry coefficient between *X* and *Y* is the four-way average of probabilities of being identical for each possible pair of gametes, which sums to $\gamma/2$. Additive relationship between *X* and *Y* is twice the coancestry and therefore γ (Figure 2, right). Now consider individual *X*. The self-coancestry considers four ways of sampling alleles a and b (with replacement), and because $P(a \equiv b) = \gamma/2$, self-coancestry is equal to $1/2 + \gamma/4$, and therefore self-relationship is equal to $1 + \gamma/2$.

The base population has associated breeding values $\mathbf{u}_0$. From the developments above, the variance-covariance matrix of breeding values is $\mathrm{Var}(\mathbf{u}_0) = [\mathbf{I}(1-\gamma/2) + \mathbf{J}\gamma]\sigma^2_u$, where $\mathbf{I}$ is the identity matrix and $\mathbf{J}$ is a matrix of ones. This covariance structure was suggested by Christensen (2012) to correctly compare genomic relationships and pedigree relationships. Due to random sampling of a limited number of founders, the mean of the base population composed of n individuals ($\overline{u}_0 = \mathbf{1}'\mathbf{u}_0/n$) will drift around the mean of the ancestral population with variance $\mathrm{Var}(\overline{u}_0) = \gamma + (1-\gamma/2)/n$.

Pedigree relationships from related base populations: VanRaden (1992) (unaware of the work of Jacquard 1974) assigned nonzero relatedness to animals in the base populations to correctly estimate inbreeding when pedigree information is missing. The value assigned to this relatedness, which is equivalent to γ, was set to the average relatedness of contemporary individuals with known relationships. Lutaaya *et al.* (1999) showed that the classical algorithm for calculating inbreeding is very sensitive to even a small loss of pedigree while VanRaden (1992) algorithm is much better although not perfect. This idea was also applied by Aguilar and Misztal (2008), and Colleau and Sargolzaei (2011) used a closely related idea in a similar setting.

Obtaining pedigree relationships from related base populations is conceptually straightforward, can be done following the tabular rules (Emik and Terrill 1949), and leads to a matrix of additive relationships

$$\mathbf{A}^{\gamma} = \mathbf{A}\left(1 - \frac{\gamma}{2}\right) + \gamma\mathbf{J},$$

where $\mathbf{A}$ is the matrix with regular relationships and $\mathbf{J}$ a matrix of 1's. In Jacquard (1974, p. 169), this formula is presented using coancestries instead of relationships. However, algorithms for computation of inbreeding (*e.g.*, Quaas 1976; Meuwissen and Luo 1992), Henderson's (1976) sparse inverse of the pedigree relationships, and other algorithms (Colleau 2002) need to be modified to account for nonzero relatedness of founders (*e.g.*, Aguilar and Misztal 2008; Christensen 2012). These changes are rather complex and do not generalize well to the case of several base populations that are presented later. For this reason, and for its conceptual appeal, we have conceived the use of metafounders.

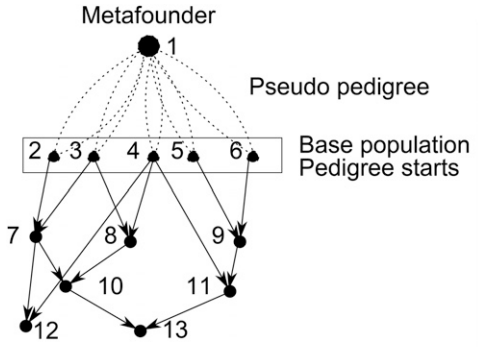

Figure 3 Base population with a metafounder and corresponding pedigree.

Metafounder

We now introduce a different, but equivalent, representation of related base populations that allows a greater flexibility. This representation uses so-called metafounders.

Definition: The notion of metafounder comes as an extension of VanRaden (1992) method for estimation of across-breed relationships. Imagine a pseudo-individual who can be considered as, simultaneously, father and mother of all base animals (Figure 3). We call this pseudo-individual a metafounder. The metafounder in Figure 3 represents the ancestral population in Figure 1.

In Figure 3, the metafounder (individual 1) represents a finite-size pool of gametes, from which the gametes constituting individuals 2–6 (the base population) are drawn. Picking two gametes at random with replacement, these gametes have an across-gamete relationship of $\gamma/2$. Therefore, the metafounder can be considered as having a self-relationship of $a_{11} = \gamma$ and an individual inbreeding coefficient of $F_i = a_{11} - 1 = \gamma - 1$, which will usually be negative. Inbreeding means departure from Hardy–Weinberg equilibrium, and negative inbreeding represents excess of heterozygotes. Therefore, negative inbreeding means that in most cases two gametes are different, *i.e.*, the size of the pool is large, which is a tenable genetic hypothesis. For instance, considering $\gamma = 0$ (and therefore $F = -1$) means that the two gametes are always different (by descent) and unrelated, *i.e.*, the size of the pool is infinite, heterozygosity (by descent) is complete, and all individuals in the base population are unrelated. Considering $\gamma = 2$ (and $F = 1$) means that two gametes drawn at random are always identical, *i.e.*, the pool consists of one gamete, there is complete homozygosity, and all individuals in the base population are identical and completely inbred.

Algorithms for relationships and inbreeding with a single metafounder: With this representation using metafounders, regular rules for computation of relationships and inbreeding change only slightly. Consider the Emik and Terrill (1949) rules for computation of additive relationship coefficients. They start by assigning self-relationships of 1 to all animals in the base population and later two rules are used,

$$a_{ij} = 0.5(a_{dj} + a_{sj})$$

$$a_{ii} = 1 + 0.5(a_{sd}),$$

where d and s are the dam and sire of i, which must be younger than j. To include the metafounder, the only change is to set its self-relationship (a_{11} in the example) to γ. The Emik and Terrill rules do not otherwise need to be changed. For instance, for individual 2 in Figure 3, $a_{22} = 1 + 0.5a_{11} = 1 + \gamma/2$, and for individuals 1 and 2, $a_{12} = 0.5(a_{11} + a_{11}) = \gamma$. For individuals 2 and 3, $a_{23} = 0.5(a_{12} + a_{12}) = \gamma$. Therefore, assigning a metafounder with self-relationship γ is strictly equivalent to considering across-founder population relationships γ and founder self-relationships $1 + \gamma/2$. The recursive algorithms of Karigl (1981) and Aguilar and Misztal (2008) are versions of Emik and Terrill (1949) and therefore need no modification beyond setting a_{11} to γ. Using these rules, $\mathbf{A}^{\gamma}$ is easily created.

Consider Henderson's (1976) inverse of the relationship matrix $\mathbf{A}$. This consists in a product on the form $\mathbf{A}^{-1} = \mathbf{L}^{-1}\mathbf{D}^{-1}\mathbf{L}^{-1'}$, where $\mathbf{D}$ is usually a diagonal matrix containing variances of the Mendelian sampling terms (deviation of an individual's breeding value from its parents' average) and $\mathbf{L}^{-1}$ contains ones in the diagonal and 0.5 coefficients linking parents to offspring. Elements of $\mathbf{D}$ are a function of inbreeding of the parents (see Thompson 1977 for the proof and Elzo (2008) for a detailed explanation). This reasoning applies equally well to the use of one metafounder. Thus, using pedigrees with a metafounder, all the information about covariance of gametes transmitted from base animals to their descendants is contained in the inbreeding of the base animals, and the algorithm of Henderson (1976) works without changes, provided (and this is important) that inbreeding for all individuals is computed previously. This is opposite to Christensen (2012), who had to devise modifications of the algorithm.

Inbreeding coefficients can be computed by Emik and Terrill (1949) or, equivalently, using recursion (Karigl 1981; Aguilar and Misztal 2008). However, efficient algorithms for computation of inbreeding use Henderson's (1976) decomposition of the numerator relationship matrix. These algorithms (*e.g.*, Quaas 1976; Meuwissen and Luo 1992) proceed by computing the variance of the Mendelian sampling term, D_{ii}. Meuwissen and Luo (1992) presented one rule,

$$D_{ii} = 0.5 - 0.25(F_s + F_d),$$

where, in the case of unknown ancestor, $s = 0$ (or $d = 0$), their programming set $F_0 = -1$. The same rule for computation of D_{ii} applies to the pedigree with one metafounder in Figure 3, by setting $F_1 = \gamma - 1$. In fact, the Meuwissen and Luo (1992) algorithm can be understood as having one

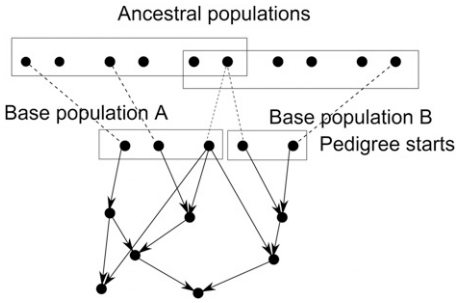

Figure 4 Several related base populations.

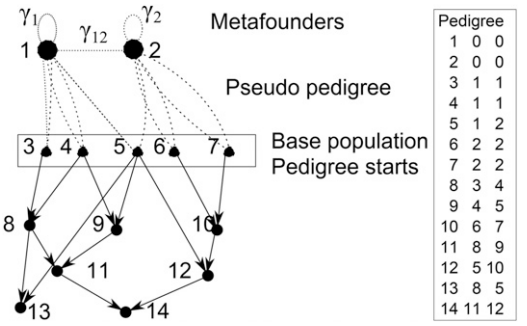

Figure 5 Population with two related metafounders 1 and 2, self-relationship coefficients $\boldsymbol{\gamma}^1$, $\boldsymbol{\gamma}^2$, and relationship coefficient $\boldsymbol{\gamma}^{1,2}$ and associated pedigree.

metafounder with $\gamma = 0$. Finally, the algorithm of Colleau (2002) for fast multiplication of matrix **A** with vector **x**, **Ax**, or extraction of elements of **A** also works.

Multiple base populations

Across-population relationships: An important case is the analysis of several populations at the same time, possibly with crosses. The conceptual model can easily be extended to several base populations, possibly with overlap as represented in Figure 4. In this case, we need to define within- and across-population relationships

$$\boldsymbol{\Gamma} = \begin{pmatrix} \gamma^{\mathrm{A}} & \gamma^{\mathrm{A,B}} \\ \text{symm} & \gamma^{\mathrm{B}} \\ & \cdots \end{pmatrix}.$$

This was suggested by VanRaden (1992) and used by VanRaden *et al.* (2011). The interpretation of the across-base population coefficients like $\gamma^{\mathrm{A,B}}$ is that the ancestor populations overlap, as seen in Figure 4. If population A is composed of n_{A} gametes, population B of n_{B} gametes, and they overlap to an extent of n_{AB} gametes (for instance, in Figure 4 these are 6, 6, and 2, respectively), then $\gamma^{\mathrm{A}} = 1/n_{\mathrm{A}}$, $\gamma^{\mathrm{B}} = 1/n_{\mathrm{B}}$, and $\gamma^{\mathrm{A,B}} = n_{\mathrm{AB}}/n_{\mathrm{A}}n_{\mathrm{B}}$. The last result can be explained as follows: $\gamma^{\mathrm{A,B}}$ is the probability that the gamete from A comes from the overlap $(n_{\mathrm{AB}}/n_{\mathrm{A}})$, times the probability that the gamete from B comes from the overlap $(n_{\mathrm{AB}}/n_{\mathrm{B}})$, times the probability that both gametes are actually the same, given that they come from the overlap $(1/n_{\mathrm{AB}})$. We allow values of γ^{A}, γ^{B}, and $\gamma^{\mathrm{A,B}}$ in a continuous range, even though the formulas only support values corresponding to integer values of n_{A}, n_{B} and n_{AB}. We also allow $\gamma^{\mathrm{A,B}}$ to potentially be negative, in order to consider the situation where populations have diverged due to selection in opposite directions. However, there is the restriction that the matrix $\boldsymbol{\Gamma}$ should be positive definite.

Metafounders: The consideration of each ancestral population as a metafounder is straightforward. Metafounders would be related by relationships

$$\boldsymbol{\Gamma} = \begin{pmatrix} \gamma^{\mathrm{A}} & \gamma^{\mathrm{A,B}} \\ & \gamma^{\mathrm{B}} \\ & \cdots \end{pmatrix}$$

(Figure 5). Actual numbers for the relationships within and across metafounders in $\boldsymbol{\Gamma}$ either can come from knowledge of the history of the populations (*i.e.*, they diverged so many generations ago) or can be inferred from genomic relationships; this is detailed later.

Algorithms for relationships and inbreeding with several metafounders: A pedigree with several metafounders defines a relationship matrix $\mathbf{A}^{\boldsymbol{\Gamma}}$. Algorithms for creation of this matrix are extensions of previous ones. To form $\mathbf{A}^{\boldsymbol{\Gamma}}$ using the tabular rules (Emik and Terrill 1949), the first step is to set $\boldsymbol{\Gamma}$ as relationships of the metafounders and then apply the regular rules. Rules for the inverse $\mathbf{A}^{\boldsymbol{\Gamma}-1}$ consist in, first, inverting $\boldsymbol{\Gamma}$ to create a small submatrix of $\boldsymbol{A}^{\boldsymbol{\Gamma}-1}$ and then using Henderson's rules (1976) with the elements D_{ii} for all individuals modified according to self-relationships of metafounders, as in the previous section. Using generalized inverses for inversion of $\boldsymbol{\Gamma}$ results in an algorithm that, for $\boldsymbol{\Gamma} = 0$, gives the same $\boldsymbol{A}^{\boldsymbol{\Gamma}-1}$ as with unknown parent groups as in Thompson (1979) or Quaas (1988). The reason for this is that the generalized inverse of $\boldsymbol{\Gamma} = 0$ is 0, and otherwise the rules for inversion and the values of D_{ii} are identical. This shows that metafounders are a generalization of unknown parent groups.

Computing D_{ii} involves computation of inbreeding coefficients, which can be done by recursion or modifying Meuwissen and Luo (1992). The Meuwissen and Luo (1992) algorithm goes up the ancestors of a given animal i and adds contributions $L_{ij}D_{jj}$ to the inbreeding coefficient of i; then animal j is deleted from the list of ancestors, and L_{ij} is set to zero. However, this does not work in the particular case of a crossbred individual issued directly from two related metafounders, *i.e.*, an F1 crossbred individual with unknown parents. This is a case that does sometimes exist, *e.g.*, in sheep and cattle. In this case, the contribution from the metafounders to A_{ii} is a sum over all metafounders $\sum_{k=1,\mathrm{nmf}}(\boldsymbol{L}_{i,k}\mathbf{K}_{:,k})^2$, where $\mathbf{K}_{:,k}$ is the kth column of $\mathbf{K}$, the lower triangular Cholesky decomposition of $\boldsymbol{\Gamma} = \mathbf{K}\mathbf{K}'$, and nmf is the number of metafounders. Therefore, in the case of several metafounders, their contributions need to be processed for simultaneously. The core modification for the Meuwissen and Luo code is

$K = \text{lower\_choleski}(\text{gamma})$

for (i in metafounders){ $F(i) = 1 - \text{gamma}(i)$}

for (i in (metafounders + 1, all animals)){

for (j in ancestors(i)){

...

if (j not in metafounders) then

Add $L_{ij}D_{jj}$ to A_{ii}

$L_{ij} = 0$

endif

}

Add $\sum\left(\left(\boldsymbol{L}_{i,1:\text{nmf}}\mathbf{K}\right)^2\right)$ to A_{ii}

$L_{i,1:\text{nmf}} = 0$

}

Finally, the algorithm of Colleau (2002) to efficiently compute products $\mathbf{A}^{\boldsymbol{\Gamma}}\mathbf{x}$ as $\mathbf{LDL}'\mathbf{x}$ multiplies the result of $\boldsymbol{L}'x$ by $\boldsymbol{D}$, which has an upper diagonal block equal to $\boldsymbol{\Gamma}$ but that is diagonal otherwise. A complete code is furnished in Supporting Information, File S2.

Genetic variance considering related base populations

Single base population: The additive genetic variance is the variance of the breeding values of the set of individuals constituting a population. This definition does not involve a notion of (un)relatedness in itself. However, in the base population, these individuals are typically assumed unrelated, which simplifies the reasoning. A question is how to relate the genetic variance of a population modeled as "related" to the genetic variance of a population modeled as "unrelated." The breeding value is defined as relative to the average of the population. For this reason, any statistical model relating phenotypes to breeding values is forced to include an overall mean or an environmental effect confounded with it. A typical model for the phenotype can be written as

$$\mathbf{y} = 1\mu + \mathbf{u} + \mathbf{e}.$$

We follow the argument of Strandén and Christensen (2011), but for the sake of discussion, consider the mean as a random variable with variance σ^2_μ. The covariance of $\mathbf{y}$ is, for the classical model with unrelated base animals, $\text{Var}(\mathbf{y}) = \mathbf{J}\sigma^2_\mu + \mathbf{A}\sigma^2_{u-\text{unrelated}} + \mathbf{R}$, where $\text{Var}(\mathbf{e}) = \mathbf{R}$. As for the new model with related base animals $\text{Var}(\mathbf{y}) = \mathbf{J}\sigma^2_{\mu^*} + \mathbf{J}\gamma\sigma^2_{u-\text{related}} + \mathbf{A}(1-\gamma/2)\sigma^2_{u-\text{related}} + \mathbf{R}$. Two equivalent models (with equivalent likelihoods under multivariate normality) should have the same covariance for $\mathbf{y}$ and therefore

$$\sigma^2_{u-\text{related}} = \frac{\sigma^2_{u-\text{unrelated}}}{1-\gamma/2}$$

and $\sigma^2_\mu = \sigma^2_{\mu^*} + \gamma\sigma^2_{u-\text{related}}$. In other words, the general across-individual covariance γ is absorbed by the overall mean (and it will be the case even if the mean is considered as a "fixed" effect; Strandén and Christensen 2011). An intuitive explanation is that, when sampling a finite number of animals from a population, animals will tend to be related and therefore the mean will drift from zero; but this drift of the mean will be accounted for by the general mean of the model. The expression above agrees with the numerical results in Christensen (2012).

This result looks puzzling because it suggests that an "inbred" population has higher genetic variance than a non-inbred one, but this is not actually the case. The parameter $\sigma^2_{u-\text{related}}$ has to be interpreted as a parameter of the statistical linear model used for the analysis, and it cannot be interpreted as a genetic variance within the population (whereas $\sigma^2_{u-\text{unrelated}}$ can be). In fact, the $\sigma^2_{u-\text{related}}$ would be genetic variance in their hypothetical unrelated ancestral monoecious parents, and it would be reduced to $\sigma^2_{u-\text{unrelated}}$ assuming a rate of inbreeding $\gamma/2$ from parents to offspring, as relatedness γ decreases the genetic variance within a population. Thus, the genetic variance *within* the population is always $\sigma^2_{u-\text{unrelated}}$, and the variance component associated to the linear model is $\sigma^2_{u-\text{related}}$. Along the same line, genetic gain in the "related" base population is not proportional to $\sigma^2_{u-\text{related}}$ (because when selecting individuals, they will be related) but to $\sigma^2_{u-\text{unrelated}}$.

Multiple base populations: The reasoning extends to the case with several populations but no crosses. For simplicity, we consider only two purebred populations. For breeds $b = \text{A}, \text{B}$ the model for phenotypes is

$$\mathbf{y}^{\mathbf{b}} = \mathbf{1}\mu^b + \mathbf{u}^{\mathbf{b}},$$

where the variance–covariance matrix of the combined vector of breeding values is

$$\text{var}\begin{pmatrix}\mathbf{u}^{\text{A}}\\ \mathbf{u}^{\text{B}}\end{pmatrix} = \sigma^2\begin{pmatrix}\mathbf{A}_{\text{A,A}}\left(1-\frac{\gamma^{\text{A}}}{2}\right) & 0\\ 0 & \mathbf{A}_{\text{B,B}}\left(1-\frac{\gamma^{\text{B}}}{2}\right)\end{pmatrix} + \sigma^2\begin{pmatrix}\gamma^{\text{A}}\mathbf{J}_{\text{AA}} & \gamma^{\text{A,B}}\mathbf{J}_{\text{AB}}\\ \gamma^{\text{A,B}}\mathbf{J}_{\text{BA}} & \gamma^{\text{B}}\mathbf{J}_{\text{BB}}\end{pmatrix},$$

with $\mathbf{A}_{b,b}$ being the relationship matrix of breed $b = A, B$ and $\mathbf{J}_{AA}$, $\mathbf{J}_{AB}$, $\mathbf{J}_{BA}, \mathbf{J}_{BB}$ being matrices consisting of 1's. Therefore, the vector of breeding values can be expressed as

$$\begin{pmatrix}\mathbf{u}^{\mathrm{A}}\\ \mathbf{u}^{\mathrm{B}}\end{pmatrix}=\begin{pmatrix}\mathbf{u}_u^{\mathrm{A}}\left(1-\frac{\gamma^{\mathrm{A}}}{2}\right)^{0.5}\\ \mathbf{u}_u^{\mathrm{B}}\left(1-\frac{\gamma^{\mathrm{B}}}{2}\right)^{0.5}\end{pmatrix}+\begin{pmatrix}\beta_{\mathrm{A}}1_{n_1}\\ \beta_{\mathrm{B}}1_{n_2}\end{pmatrix},$$

where subindex u on breeding values denotes that they are in the model with unrelated base populations, and

$$\begin{pmatrix}\beta_{\mathrm{A}}\\ \beta_{\mathrm{B}}\end{pmatrix}\sim N\left(\begin{pmatrix}0\\ 0\end{pmatrix},\sigma^2\begin{pmatrix}\gamma^{\mathrm{A}} & \gamma^{\mathrm{A,B}}\\ \gamma^{\mathrm{A,B}} & \gamma^{\mathrm{B}}\end{pmatrix}\right)$$

and assumed independence of breeding values $\mathbf{u}_u^{\mathrm{A}}$ and $\mathbf{u}_u^{\mathrm{B}}$. By an argument similar to above (*i.e.*, Stranden and Christensen 2011), the parameters β_{A} and β_{B} are absorbed into the two general mean parameters μ^{A} and μ^{B}, respectively. Therefore, the two models are equivalent in the sense that genetic variance parameters are just scaled by $(1-\gamma^b/2)$ and breeding values are just scaled and shifted. This model implies that phenotypes are separate by population and a mean (or distinct levels of fixed effects, *e.g.*, herds) has to be fit by population. The argument above is not difficult to generalize to any number of populations, as far as crosses that do not exist.

Multiple base populations with crossing: For crossbred populations the equivalence above does not hold because γ^{AB} enters into the covariances across individuals. A different approximate equivalence of variances can be constructed as follows. Assume a set of n base population individuals (n is assumed large) drawn from each of m populations. Let the genetic values of the across-breed base populations be $\mathbf{u}_0=\begin{pmatrix}\mathbf{u}_{\mathrm{A}0}\\ \mathbf{u}_{\mathrm{B}0}\end{pmatrix}$. The variance–covariance matrix is

$$\mathrm{Var}(\mathbf{u}_0)=\begin{pmatrix}1+\frac{\gamma^{\mathrm{A}}}{2} & \gamma^{\mathrm{A}} & \cdots & \gamma^{\mathrm{AB}} & \gamma^{\mathrm{AB}} & \cdots\\ \gamma^{\mathrm{A}} & 1+\frac{\gamma^{\mathrm{A}}}{2} & \cdots & \gamma^{\mathrm{AB}} & \gamma^{\mathrm{AB}} & \cdots\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ \gamma^{\mathrm{AB}} & \gamma^{\mathrm{AB}} & \cdots & 1+\frac{\gamma^{\mathrm{B}}}{2} & \gamma^{\mathrm{B}} & \cdots\\ \gamma^{\mathrm{AB}} & \gamma^{\mathrm{AB}} & \cdots & \gamma^{\mathrm{B}} & 1+\frac{\gamma^{\mathrm{B}}}{2} & \cdots\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\end{pmatrix}\sigma^2_{\mathrm{related}}.$$

The sample variance of $\mathbf{u}_0$, across all populations, is

$$S_u^2=\frac{\mathbf{u}_0{}'\mathbf{u}_0}{nm}-\overline{\mathbf{u}_0}^2,$$

which, for $\mathrm{Var}(\mathbf{u}_0)=\mathbf{K}\sigma_u^2$ (σ_u^2 is a parameter), has expectation (Searle 1982, p. 355)

$$E\left(S_u^2\right)=\left(\overline{\mathrm{diag}(\boldsymbol{K})}-\overline{\boldsymbol{K}}\right)\sigma_u^2.$$

In the classical parameterization (unrelated founders) $\mathbf{K}=\mathbf{I}$ and thus

$$E\left(S_u^2\right)=\left(\overline{\mathrm{diag}(\mathbf{K})}-\overline{\mathbf{K}}\right)\sigma_u^2=\left(1-\frac{1}{nm}\right)\sigma^2_{\mathrm{unrelated}},$$

which is equal to $\sigma^2_{\mathrm{unrelated}}$ if the population is reasonably large (a popular assumption) and therefore $\sigma_u^2=\sigma^2_{\mathrm{unrelated}}$ if founders are unrelated. This means that when the founders are unrelated, the genetic variance is, on expectation, equal to the variance component of the covariance structure.

Consider now the structure above for $\mathrm{Var}(\mathbf{u}_0)$. The two terms are equal to

$$\overline{\mathrm{diag}(\boldsymbol{K})}=1+\frac{\overline{\mathrm{diag}(\boldsymbol{\Gamma})}}{2}$$

$$\begin{aligned}\overline{\boldsymbol{K}}&=\frac{\left(n^2m^2\overline{\boldsymbol{\Gamma}}+nm-nm\left(\overline{\mathrm{diag}(\boldsymbol{\Gamma})}/2\right)\right)}{(nm)^2}\\&=\overline{\boldsymbol{\Gamma}}+\frac{1-\overline{\mathrm{diag}(\boldsymbol{\Gamma})}/2}{nm}\end{aligned}$$

$$\begin{aligned}E\left(S_u^2\right)&=\left(\overline{\mathrm{diag}(\boldsymbol{K})}-\overline{\boldsymbol{K}}\right)\sigma_u^2\\&=\left(1+\frac{\overline{\mathrm{diag}(\boldsymbol{\Gamma})}}{2}-\overline{\boldsymbol{\Gamma}}-\frac{1-\overline{\mathrm{diag}(\boldsymbol{\Gamma})}/2}{nm}\right)\sigma^2_{\mathrm{related}}\end{aligned}$$

in which we neglect the last term. This means that the genetic variance is, on expectation, equal to the variance component $\sigma^2_{\mathrm{related}}$ times a constant $(1+\overline{\mathrm{diag}(\boldsymbol{\Gamma})}/2-\overline{\boldsymbol{\Gamma}})$, which is <1. Equating these two expressions for $E(S_u^2)$ gives

$$\sigma^2_{\mathrm{related}}\approx\frac{\sigma^2_{\mathrm{unrelated}}}{\left(1+\overline{\mathrm{diag}(\boldsymbol{\Gamma})}/2-\overline{\boldsymbol{\Gamma}}\right)}.$$

This expression gives the previous result $\sigma^2_{\mathrm{related}}=\sigma^2_{\mathrm{unrelated}}/(1-\gamma/2)$ for a single population. Compared to the result in the previous subsection about multiple populations, this approximate equivalence is quite different. The result in the previous subsection is an equivalence between one genetic variance in a model with related base individuals and breed-specific genetic variances in a model with unrelated base individuals, whereas the result here is an approximate equivalence between two genetic variances, one being in a related base population and another being in an unrelated base population. This last expression $\sigma^2_{\mathrm{related}}\approx\sigma^2_{\mathrm{unrelated}}/\left[\left(1+\overline{\mathrm{diag}(\boldsymbol{\Gamma})}/2-\overline{\boldsymbol{\Gamma}}\right)\right]$ is more general because it can consider correctly crosses across individuals. The difference comes also because in the previous expression there were separate means for each population, something that is not required here.

Segregation variance: When crossing pure breeds, there is an increase of genetic variance due to the increase of heterozygosity of the QTL; for instance, if alternative alleles are fixed at each line. The additional variance in the F2 cross compared to the variance in the F1 cross is termed segregation variance (Lande 1981; Lo *et al.* 1993). This is typically

ignored in a classical framework, although methods exist (Lo *et al.* 1993; Garcia-Cortes and Toro 2006). This increase in the genetic variance can be considered using related metafounders, as we show here. Two individuals in an F1 population (assuming—in a pedigree sense—unrelated and non-inbred parents, and factorizing out $\sigma^2_{\text{related}}$) have

$$\text{Var}(\boldsymbol{u}_{\text{AB}}) = \begin{pmatrix} 1+\frac{\gamma^{\text{AB}}}{2} & \frac{\gamma^{\text{A}}}{4}+\frac{\gamma^{\text{B}}}{4}+\frac{\gamma^{\text{AB}}}{2} \\ \frac{\gamma^{\text{A}}}{4}+\frac{\gamma^{\text{B}}}{4}+\frac{\gamma^{\text{AB}}}{2} & 1+\frac{\gamma^{\text{AB}}}{2} \end{pmatrix}$$

whereas two individuals in an F2 population (parents in F1 above) have

$$\begin{aligned}\text{Var}(\mathbf{u}_{\text{AB}\times\text{AB}}) &= \begin{pmatrix} 1+\frac{\gamma^{m}}{4}+\frac{\gamma^{\text{AB}}}{4} & \frac{\gamma^{m}}{2}+\frac{\gamma^{\text{AB}}}{2} \\ \frac{\gamma^{m}}{2}+\frac{\gamma^{\text{AB}}}{2} & 1+\frac{\gamma^{m}}{4}+\frac{\gamma^{\text{AB}}}{4} \end{pmatrix} \\ &= \begin{pmatrix} 1+\frac{\gamma^{\text{F2}}}{2} & \gamma^{\text{F2}} \\ \gamma^{\text{F2}} & 1+\frac{\gamma^{\text{F2}}}{2} \end{pmatrix},\end{aligned}$$

with $\gamma^m = (\gamma^{\text{A}} + \gamma^{\text{B}})/2$ and $\gamma^{\text{F2}} = (\gamma^{\text{A}} + \gamma^{\text{B}})/4 + \gamma^{\text{AB}}/2$. The γ^{F2} is transmitted forward and does not change in the F3, F4, etc. The genetic variance of such a population is thus $1 - \gamma^{\text{F2}}/2 = 1 - [(\gamma^{\text{A}} + \gamma^{\text{B}})/4 - \gamma^{\text{AB}}/2]/2$. The variance–covariance matrix of two individuals in the F2 can be expressed as

$$\begin{aligned}\text{Var}(\mathbf{u}_{\text{AB}\times\text{AB}}) &= \text{Var}(\mathbf{u}_{\text{AB}}) \\ &\quad + \begin{pmatrix} (\gamma^{m} - \gamma^{\text{AB}})/4 & 0 \\ 0 & (\gamma^{m} - \gamma^{\text{AB}})/4 \end{pmatrix}\end{aligned}$$

showing that the segregation variance is $(\gamma^m - \gamma^{\text{AB}})/4$. Because $\boldsymbol{\Gamma}$ is positive definite, then this term must be ≥ 0. Slatkin and Lande (1994) showed that segregation variance is a function of within-loci squared differences of means at the two breeds, plus cross-products of differences across loci weighted by linkage. If γ^{AB} is estimated using markers as above, then it is implicitly assumed that genotypes at loci for the trait of interest have the same distribution across breeds and within the genome as marker genotypes. Reports of segregation variances in the livestock genetics literature are scarce (*e.g.*, Cardoso and Tempelman 2004; Munilla-Leguizamon and Cantet 2011), partly because of poor data sets, partly because of computational difficulties, and partly because the bulk of crossbred animals is in poultry and swine, where crosses do not go beyond F1 populations. So it is uncertain whether accounting for segregation variance is of any practical relevance.

Estimation of metafounders ancestral relationships from genomic data

Because the within- and across-founder relationships cannot be inferred from pedigree, we suggest estimating these relationships using molecular markers, referring them to a genetic base defined according to genomic relationships (Christensen 2012). The objective of this section is to obtain estimators of $\boldsymbol{\Gamma}$ based on two kinds of statistical inference: a method of maximum likelihood and a method of moments (roughly, make first- and second-order statistics of genomic and pedigree relationships comparable).

Maximum likelihood: Genomic information sheds light on relationships across breeds (Gibbs *et al.* 2009; Kijaas *et al.* 2009; VanRaden *et al.* 2011; Legarra *et al.* 2014a). Genomic relationships (VanRaden 2008; Hayes *et al.* 2009) are estimators of relatedness based on the observation of thousands of molecular markers, and typically matrix $\mathbf{G} = \mathbf{Z}\mathbf{Z}'/s$ is used, where $\mathbf{Z}$ contains centered genotypes and s is a measure of global heterozygosity, for instance, $s = 2\sum p_i q_i$, the total heterozygosity at the markers. This information can in principle be used to infer the $\boldsymbol{\Gamma}$ coefficients as follows. Marker genotypes follow Mendelian transmission, and therefore the covariance of genotypes of two individuals is determined by their relationship. Christensen (2012) used this to estimate γ in a single population. First, he integrated the likelihood over the unknown allelic frequencies, which results in using allelic frequencies of 0.5 as a reference ($\mathbf{Z}$ coded as $\{-1, 0, 1\}$). Assuming multivariate normality for $\mathbf{Z}$, the markers' genotype, the likelihood of observed genotypes conditional to γ and s is

$$\begin{aligned}\log p(\mathbf{Z}|\gamma, s) = \text{const} &- \frac{pn_2}{2}\log(s) - \frac{p}{2}\log\left(\left|\mathbf{A}^{\gamma}_{22}\right|\right) \\ &- \frac{p}{2s}\text{tr}\left(\mathbf{A}^{\gamma-1}_{22}\mathbf{Z}\mathbf{Z}'\right),\end{aligned}$$

where n_2 is the number of genotyped individuals and $\mathbf{A}^{\gamma}_{22}$ is the submatrix of $\mathbf{A}^{\gamma}$ corresponding to the genotyped individuals. The parameter s is a measure of heterozygosity in the genotyped population, and it is *not* equal to observed $2\sum p_i q_i$. The extension of this likelihood to multiple populations with different γ's in $\boldsymbol{\Gamma}$ is straightforward

$$\begin{aligned}\log p(\boldsymbol{Z}|\boldsymbol{\Gamma}, s) = \text{const} &- \frac{pn_2}{2}\log(s) - \frac{p}{2}\log\left(\left|\boldsymbol{A}^{\boldsymbol{\Gamma}}_{22}\right|\right) \\ &- \frac{p}{2s}tr\left(\boldsymbol{A}^{\boldsymbol{\Gamma}-1}_{22}\boldsymbol{Z}\boldsymbol{Z}'\right),\end{aligned}$$

where $\mathbf{A}^{\boldsymbol{\Gamma}}$ is the relationship matrix constructed with a given $\boldsymbol{\Gamma}$ matrix and $\mathbf{A}^{\boldsymbol{\Gamma}}_{22}$ is the submatrix corresponding to the genotyped individuals. This likelihood can be factorized by markers as

$$\begin{aligned}\log p(\mathbf{Z}|\boldsymbol{\Gamma}, s) = \text{const} &- \frac{pn_2}{2}\log(s) - \frac{p}{2}\log\left(\left|\mathbf{A}^{\boldsymbol{\Gamma}}_{22}\right|\right) \\ &- \frac{p}{2s}\sum_{i=1}^{nsnp}\mathbf{z}'_i\mathbf{A}^{\boldsymbol{\Gamma}-1}_{22}\mathbf{z}_i.\end{aligned}$$

The procedure can be completed by adding a prior distribution to $\boldsymbol{\Gamma}$ and using a Bayesian estimator instead of maximum likelihood. The prior distribution for $\boldsymbol{\Gamma}$ can be assigned

based on spatial or temporal distances; for instance, Latxa sheep founders in 1990 and 1992 should be closer than 1990 and 2000. However, in none of these forms of the likelihood can $\boldsymbol{\Gamma}$ be factorized out, and the maximization of the likelihood needs to be done by a search method such as Simplex or Monte Carlo methods. For this reason, we present a method based on summary statistics.

Method of moments based on summary statistics: This method matches summary statistics of across-individual and within-individual relationships in both $\mathbf{A}_{22}^{\boldsymbol{\Gamma}}$ (the matrix of extended pedigree-based relationships) and $\mathbf{G}$ (VanRaden *et al.* 2011; Vitezica *et al.* 2011; Christensen *et al.* 2012). This forces the equivalence between expected changes of the mean and variance under genetic drift (Vitezica *et al.* 2011; Christensen *et al.* 2012) for the populations described by either the pedigree or the genomic relationship matrices. For a set of n random variables $\mathbf{u}$ with variance–covariance matrix $\mathbf{K}$, the sample average $\overline{\mathbf{u}} = \mathbf{1}'\mathbf{u}/n$ has a variance $\mathrm{Var}(\overline{\mathbf{u}}) = \overline{\mathbf{K}}$, whereas the sample variance $S_u^2 = \mathbf{u}'\mathbf{u}/n - \overline{\mathbf{u}}^2$ has expectation $E(S_u^2) = \mathrm{tr}(\mathbf{K})/n - \overline{\mathbf{K}} = \overline{\mathrm{diag}(\mathbf{K})} - \overline{\mathbf{K}}$ (Searle, 1982, p. 355). The idea in the method is to force these two statistics of $\mathbf{K}$ ($\mathrm{Var}(\overline{\mathbf{u}})$ and $E(S_u^2)$) to be equivalent across both parameterizations ($\mathbf{K} = \mathbf{A}^{\boldsymbol{\Gamma}}$ and $\mathbf{K} = \mathbf{G}$). We consider three situations.

Single population: Two single unknowns need to be estimated: γ and s. Since $\boldsymbol{\gamma} = \mathbf{A}(1-\gamma/2) + \gamma\mathbf{J}$, the average of all elements is $\overline{\mathbf{A}_{22}^{\gamma}} = \overline{\mathbf{A}_{22}}(1-\gamma/2) + \gamma$, and the average of the diagonal is $\overline{\mathrm{diag}(\mathbf{A}_{22}^{\gamma})} = \overline{\mathrm{diag}(\mathbf{A}_{22})}(1+\gamma/2) + \gamma$, where $\mathbf{A}_{22}$ is the regular pedigree-relationship matrix for genotyped individuals. Therefore, a system of two equations needs to be set up,

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

$$\overline{\mathrm{diag}(\boldsymbol{A}_{22})}\left(1-\frac{\gamma}{2}\right) + \gamma = \overline{\mathrm{diag}(\mathbf{Z}\mathbf{Z}')}/s$$

with solutions

$$s = \frac{\overline{\mathrm{diag}(\mathbf{Z}\mathbf{Z}')}\left(1-\overline{\mathbf{A}_{22}}/2\right) - \overline{\mathbf{Z}\mathbf{Z}'}\left(1-\overline{\mathrm{diag}(\mathbf{A}_{22})}/2\right)}{\overline{\mathrm{diag}(\mathbf{A}_{22})} - \overline{\mathbf{A}_{22}}}$$

$$\gamma = \frac{\overline{\mathbf{Z}\mathbf{Z}'}/s - \overline{\mathbf{A}_{22}}}{1-\overline{\mathbf{A}_{22}}/2}.$$

These solutions have an interpretation in terms of measures of inbreeding in the population. In a population large enough and mating at random, inbreeding of the individuals is equal to half the relationships of their parents, $\overline{\mathrm{diag}(\mathbf{A}_{22})} = 1 + \overline{\mathbf{A}_{22}}/2 = 1 + \overline{F_{\mathrm{A}}}$ ($\overline{F_{\mathrm{A}}}$ is average pedigree inbreeding), and $\overline{\mathrm{diag}(\mathbf{Z}\mathbf{Z}')}/s = 1 + \overline{\mathbf{Z}\mathbf{Z}'}/2s = 1 + \overline{F_{\mathrm{G}}}$ ($\overline{F_{\mathrm{G}}}$ is average genomic inbreeding). Therefore, in this case $\gamma/2 = (\overline{F_{\mathrm{G}}} - \overline{F_{\mathrm{A}}})/(1-\overline{F_{\mathrm{A}}})$. This is basically the reverse of the expression derived by Vitezica *et al.* (2011), who adjusted $\mathbf{G}$ to match $\mathbf{A}$ and called $\gamma = \alpha$. The expression shows that γ is a correction for underestimation of inbreeding of $\mathbf{A}$ with respect to $\mathbf{G}$, following Wright's F coefficients theory. An advantage of the method is that it needs only statistics of the $\mathbf{A}_{22}$ and $\mathbf{G}$ matrices, which might be more available than full matrices.

Multiple pure populations: Assume that a sample from each pure breed is genotyped. Consider the purebred parts of $\mathbf{G}$ and $\mathbf{A}_{22}^{\boldsymbol{\Gamma}}$, for simplicity 2 breeds A and B:

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}^{\mathrm{A,A}} & \mathbf{G}^{\mathrm{A,B}} \\ \mathbf{G}^{\mathrm{B,A}} & \mathbf{G}^{\mathrm{B,B}} \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_{\mathrm{A}}\mathbf{Z}_{\mathrm{A}}' & \mathbf{Z}_{\mathrm{A}}\mathbf{Z}_{\mathrm{B}}' \\ \mathbf{Z}_{\mathrm{B}}\mathbf{Z}_{\mathrm{A}}' & \mathbf{Z}_{\mathrm{B}}\mathbf{Z}_{\mathrm{B}}' \end{pmatrix} \Big/ s$$

$$\boldsymbol{A}_{22}^{\boldsymbol{\Gamma}} = \begin{pmatrix} \mathbf{A}_{22\mathrm{A,A}}^{\boldsymbol{\Gamma}} & \mathbf{A}_{22\mathrm{A,B}}^{\boldsymbol{\Gamma}} \\ \mathbf{A}_{22\mathrm{B,A}}^{\boldsymbol{\Gamma}} & \mathbf{A}_{22\mathrm{B,B}}^{\boldsymbol{\Gamma}} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{22\mathrm{A,A}}\left(1-\frac{\gamma^{\mathrm{A}}}{2}\right) + \mathbf{J}\gamma^{\mathrm{A}} & \mathbf{J}\gamma^{\mathrm{A,B}} \\ \mathbf{J}\gamma^{\mathrm{A,B}} & \mathbf{A}_{22\mathrm{B,B}}\left(1-\frac{\gamma^{\mathrm{B}}}{2}\right) + \mathbf{J}\gamma^{\mathrm{B}} \end{pmatrix}.$$

To meet the conditions of unbiasedness we need to force the equality of average diagonal and averages of $\mathbf{G}$ and $\mathbf{A}^{\boldsymbol{\Gamma}}$ and set up the four equations

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

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

$$\overline{\mathbf{A}_{22,\mathrm{A,B}}} + \gamma^{\mathrm{A,B}} = \overline{\mathbf{Z}_{\mathbf{A}}\mathbf{Z}_{\mathbf{B}}'}/s$$

$$\overline{\mathrm{diag}(\mathbf{A}_{22,\mathrm{A,A}})}\left(1-\frac{\gamma^{\mathrm{A}}}{2}\right) + \gamma^{\mathrm{A}} + \overline{\mathrm{diag}(\mathbf{A}_{22,\mathrm{B,B}})}\left(1-\frac{\gamma^{\mathrm{B}}}{2}\right) + \gamma^{\mathrm{B}} = \overline{\mathrm{diag}(\mathbf{Z}\mathbf{Z}')}/s.$$

The solution is a generalization of the solutions for single populations. The scaling estimates for single populations are $s_{\mathrm{A}} = d_{\mathrm{A}}/m_{\mathrm{A}}$ and $s_{\mathrm{B}} = d_{\mathrm{B}}/m_{\mathrm{B}}$ with

$$d_{\mathrm{A}} = \overline{\mathrm{diag}(\mathbf{Z}_{\mathbf{A}}\mathbf{Z}_{\mathbf{A}}')}\left(1-\overline{\mathbf{A}_{22,\mathrm{A,A}}}/2\right) - \overline{\mathbf{Z}_{\mathbf{A}}\mathbf{Z}_{\mathbf{A}}'}\left[1-\frac{\overline{\mathrm{diag}(\mathbf{A}_{22,\mathrm{A,A}})}}{2}\right]$$

and $m_{\mathrm{A}} = \overline{\mathrm{diag}(\mathbf{A}_{22,\mathrm{A,A}})} - \overline{\mathbf{A}_{22,\mathrm{A,A}}}$, and d_{B} and m_{B} defined similarly. The solutions for two populations are

$$s = \frac{d_{\mathrm{A}}n_{\mathrm{A}} + d_{\mathrm{B}}n_{\mathrm{B}}}{m_{\mathrm{A}}n_{\mathrm{A}} + m_{\mathrm{B}}n_{\mathrm{B}}}$$

$$\gamma^{\mathrm{A}} = \frac{\overline{\mathbf{G}}_{\mathrm{A,A}} - \overline{\mathbf{A}}_{22\mathrm{A,A}}}{1-\overline{\mathbf{A}_{22,\mathrm{A,A}}}/2};\ \gamma^{\mathrm{B}} = \frac{\overline{\mathbf{G}}_{B,B} - \overline{\mathbf{A}}_{22\mathrm{B,B}}}{1-\overline{\mathbf{A}_{22,\mathrm{B,B}}}/2};\ \gamma^{\mathrm{A,B}} = \overline{\mathbf{G}}_{\mathrm{A,B}}$$

so that within-breed and across-breed average relationships agree. Assuming $\overline{\mathbf{A}_{22,A,A}}$ $\overline{\mathbf{A}_{22,B,B}}$ are close to zero, $\gamma^{A} = (\overline{\mathbf{G}}_{A,A} - \overline{\mathbf{A}}_{A,A})$, $\gamma^{B} = (\overline{\mathbf{G}}_{B,B} - \overline{\mathbf{A}}_{B,B})$, $\gamma^{A,B} = \overline{\mathbf{G}}_{A,B}$, which consist in setting $p = 0.5$ to construct the **G** matrices (VanRaden 2008) and then simply quantify average relationships across breeds. This is the simple method used by VanRaden *et al.* (2011), although they did not define scaling s as we have done. This reasoning can be extended to as many breeds as needed. Again, this method can be used from published statistics without access to raw data.

Populations with crosses: In some cases, pure populations may not be genotyped. For instance, Angus bulls may be mated to Limousine females and only the crossbreds and Angus genotyped. Another example is unknown parent groups (Quaas 1988), base populations that account for missing parentages. However, at some point descendants of these populations may be genotyped, and this information is usable. We propose an algorithm very similar to that of Harris and Johnson (2010). Let **Q** be a matrix containing in the i,j cell the expected fraction of metafounder j in the individual i (Quaas 1988). This matrix can be efficiently obtained using Colleau (2002), recursion, or tracing down the pedigree. The following identity, which is an extension of $\mathbf{A}^{\gamma} = \mathbf{A}(1-\gamma/2) + \gamma\mathbf{J}$, approximately holds

$$\mathbf{A}^{\boldsymbol{\Gamma}} \approx \mathbf{A}(\mathbf{I} - 0.5\ \mathrm{diag}(\mathbf{Q}\boldsymbol{\Gamma}\mathbf{Q}^{'})) + \mathbf{Q}\boldsymbol{\Gamma}\mathbf{Q}^{'}.$$

And therefore, $\mathbf{A}_{22}^{\boldsymbol{\Gamma}} \approx \mathbf{A}_{22}[\mathbf{I} - 0.5\ \mathrm{diag}(\mathbf{Q}_2\boldsymbol{\Gamma}\mathbf{Q}_2^{'})] + \mathbf{Q}_2\boldsymbol{\Gamma}\mathbf{Q}_2^{'}$. A linear model can be fit as $\mathbf{G} = \mathbf{A}_{22}^{\boldsymbol{\Gamma}} + \mathbf{E}$, where **E** is an error term and $\mathbf{Q}_2$ is the section of **Q** containing proportions of metafounders in genotyped individuals. We neglect the term $-0.5\ \mathrm{diag}(\mathbf{Q}_2\boldsymbol{\Gamma}\mathbf{Q}_2^{'})$, which is small with respect to the rest of elements and obtain a further approximation $\mathbf{G} = \mathbf{A}_{22} + \mathbf{Q}_2\boldsymbol{\Gamma}\mathbf{Q}_2^{'} + \mathbf{E}$, in which $\boldsymbol{\Gamma}$ is explicit. This expression can be linearized using the vec operator (Henderson and Searle 1979), and the least-squares estimator can be transformed back to a matrix form. This least-squares estimator of $\boldsymbol{\Gamma}$ is

$$\widehat{\boldsymbol{\Gamma}} = \left(\mathbf{Q}_2^{'}\mathbf{Q}_2\right)^{-1}\mathbf{Q}_2^{'}(\mathbf{G} - \mathbf{A}_{22})\mathbf{Q}_2\left(\mathbf{Q}_2^{'}\mathbf{Q}_2\right)^{-1}$$

using $\mathbf{G} = \mathbf{Z}\mathbf{Z}'/s$ and assuming that the value of s is known. If only pure population animals are genotyped, this is identical to the approximation of the estimator above for "pure populations." This solution for $\boldsymbol{\Gamma}$ is identical to the estimator proposed by Harris and Johnson (2010, Equations 13 and 14). As for s, one can use

$$s = \frac{\overline{\mathrm{diag}(\mathbf{Z}\mathbf{Z}^{'})} - \overline{\mathbf{Z}\mathbf{Z}^{'}}}{\overline{\mathrm{diag}(\mathbf{A}_{22}^{\boldsymbol{\Gamma}})} - \overline{\mathbf{A}_{22}^{\boldsymbol{\Gamma}}}},$$

where the approximation is used for $\mathbf{A}_{22}^{\boldsymbol{\Gamma}}$ [in this case including $-0.5\ \mathrm{diag}(\mathbf{Q}\boldsymbol{\Gamma}\mathbf{Q}^{'})$] such that $\overline{\mathrm{diag}(\mathbf{A}_{22}^{\boldsymbol{\Gamma}})}$ and $\overline{\mathbf{A}_{22}\boldsymbol{\Gamma}}$ are linear functions of $\boldsymbol{\Gamma}$. This system of two equations with two unknown is iterated until convergence. If there is little information for some metafounders (as is the case in ruminants), Bayesian estimation using a prior structure for $\boldsymbol{\Gamma}$ can be considered.

Combining pedigree relationships with metafounders and genomic relationships when not all individuals are genotyped

The SSGBLUP method for genomic evaluation (Aguilar *et al.* 2010; Christensen and Lund 2010; Legarra *et al.* 2014b) completes genomic information with pedigree-based information and in fact proceeds by correcting pedigree relationships in view of genomic relationships. Pedigree relationships are modified as (Legarra *et al.* 2009; Christensen and Lund 2010)

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

where **H** is a matrix with relationships after including pedigree and genomic relationships, **G** is a matrix including genomic relationships for genotyped individuals ($\mathbf{u}_2$), which is projected upon relationships of ungenotyped animals ($\mathbf{u}_1$), **A** is the pedigree-based relationship matrix, and $\mathbf{A}_{22}$ is a relationship matrix across genotyped individuals. This joint matrix **H** can be understood as a linear imputation of genotypes over all nongenotyped individuals (Christensen and Lund 2010), considering also the uncertainty in the imputation. This covariance matrix is increasingly used in genomic predictions of genetic merit (Aguilar *et al.* 2010; Christensen *et al.* 2012) and also in QTL detection (Dikmen *et al.* 2013).

The algebraic development of matrix **H** assumes that base allelic frequencies are known or, equivalently, that mean and variance of the population do not change with time. This is notoriously false with small populations, deep pedigrees, or in presence of selection. Different adjustments had been suggested to modify genomic relationships so that their genetic base is the same as that of pedigree relationships (Vitezica *et al.* 2011; Christensen *et al.* 2012). This implicitly estimates the shift in breeding values (or allelic frequencies) from the pedigree base to the genotyped population (Vitezica *et al.* 2011). However, these adjustments do not consider the pedigree structure of the populations, and their generalizations to crosses of lines or breeds are neither completely satisfactory nor well understood (but see Harris and Johnson 2010; Makgahlela *et al.* 2014).

Christensen (2012) argued that, contrary to pedigree relationships, genomic relationships are independent of pedigree completeness and they should define the genetic base. He thus considered matching pedigree relationships to genomic relationships instead of the opposite. He showed that after marginalizing the allelic frequencies from the joint likelihood, the result was a related base population and suggested estimating γ and s using maximum likelihood. All our developments rely on this base and therefore, the extended pedigrees with metafounders do automatically conciliate marker and pedigree-based relationships, using estimates of $\boldsymbol{\Gamma}$ and s from markers. In particular, the inverse of the joint pedigree and markers relationship matrix is

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

This matrix can be fit into the mixed model equations of the SSGBLUP.

We have seen that the variance component assuming "related" founders is not the same as the genetic variance assuming "unrelated" founders; the latter is the one classically estimated and used. The most straightforward solution is to reestimate the variance using metafounders. Alternatively, to use current estimates of genetic variance in the implementation, the variance of the breeding values needs to be scaled to $\sigma^2_{\text{unrelated}}$. On expectation, the following equivalence holds:

$$\sigma^2_{\text{related}} = \sigma^2_{\text{unrelated}}/k, \quad \text{with } k = \left(1 + \frac{\overline{\text{diag}(\boldsymbol{\Gamma})}}{2} - \overline{\boldsymbol{\Gamma}}\right).$$

Thus

$$\text{Var}(\mathbf{u}) = \mathbf{H}^{\boldsymbol{\Gamma}}\sigma^2_{\text{related}} = \mathbf{H}^{\boldsymbol{\Gamma}}\frac{\sigma^2_{\text{unrelated}}}{k}$$

and

$$\text{Var}(\mathbf{u})^{-1} = k\mathbf{H}^{\boldsymbol{\Gamma}-1}\sigma^{-2}_{\text{unrelated}} = k\mathbf{A}^{\boldsymbol{\Gamma}-1}\sigma^{-2}_{\text{unrelated}} + \begin{pmatrix} 0 & 0 \\ 0 & k\mathbf{G}^{-1} - k\mathbf{A}_{22}^{\boldsymbol{\Gamma}-1} \end{pmatrix}\sigma^{-2}_{\text{unrelated}},$$

such that the inverse of the combined relationship matrix ($\mathbf{H}^{\boldsymbol{\Gamma}-1}$) can be multiplied by a single scalar, $k = (1 + \overline{\text{diag}(\boldsymbol{\Gamma})}/2 - \overline{\boldsymbol{\Gamma}})$.

Examples

Example 1: How pedigree relationships are modified

Consider the pedigree in Figure 4 and the relationships between the subset of individuals 8 (pure breed A), 10 (pure breed B) , and 14 (crossbred, 56% breed A and 44% breed B, grandson of 8 and of 10). Regular relationships ($\mathbf{A}_{\text{subset}}$) are

$$\mathbf{A}_{\text{subset}} = \begin{pmatrix} 1 & 0 & 0.31 \\ 0 & 1 & 0.25 \\ 0.31 & 0.25 & 1.06 \end{pmatrix}.$$

Consider now $\boldsymbol{\Gamma} = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.2 \end{pmatrix}$. Then

$$\mathbf{A}_{\text{subset}} = \begin{pmatrix} 1.05 & 0 & 0.35 \\ 0 & 1.03 & 0.31 \\ 0.35 & 0.31 & 1.08 \end{pmatrix}.$$

All within-breed relationships have increased, because each base population is now assumed self-related. However, animals 8 and 10 are unrelated. Considering across-base population relationships in $\boldsymbol{\Gamma} = \begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.2 \end{pmatrix}$ gives

$$\mathbf{A}_{\text{subset}} = \begin{pmatrix} 1.05 & 0.05 & 0.37 \\ 0.05 & 1.10 & 0.34 \\ 0.37 & 0.34 & 1.09 \end{pmatrix},$$

where the relationship between 8 and 10 appears, which in turn slightly increases the inbreeding coefficient of 14. To standardize to the genetic variance estimated assuming unrelated base individuals, $\mathbf{A}_{\text{subset}}$ must be divided by $(1 + \overline{\text{diag}(\boldsymbol{\Gamma})}/2 - \overline{\boldsymbol{\Gamma}}) = 0.975$.

Example 2: Interpretation of γ in a single population

Legarra *et al.* (2014a) used dairy sheep data (Manech Tête Rousse) for genomic prediction including 38,287 markers and 1295 rams. The relevant statistics are (observed) $2\sum p_i q_i = 14771$, $\overline{\text{diag}(\mathbf{ZZ}')} = 22798$, $\overline{\mathbf{ZZ}'} = 8654$, $\overline{\text{diag}(\mathbf{A}_{22})} = 1.011$, $\overline{\mathbf{A}_{22}} = 0.04$. Using the single population method above yields $\gamma = 0.434$, $s = 18602$. What do these numbers mean? They imply that heterozygosity of markers at the base population should have been $s = 18602$ (instead of observed 14771), to appropriately match the fact that the heterozygosity at the markers reduced from the base to the observed population, according to inbreeding observed in the pedigree. Based on this estimate, average genomic inbreeding is $1 - \overline{\text{diag}(\mathbf{ZZ}')}/s = 0.22$, which can be achieved with an effective size of the founder population $N_e = 1/0.43$ and therefore $\gamma = 0.43$. Although this effective size is very small, it refers to a reference with allelic frequencies equal to 0.5. This has to be taken as a reference point for the linear model and has no clear biological meaning.

Example 3: Numerical example of two breeds and crossbred individuals

VanRaden *et al.* (2011) estimated relationship coefficients across Jersey, Holstein, and Brown Swiss using 43,385 markers. Based on their published statistics and using the method based on summary statistics outlined above, we obtained an estimate of $\boldsymbol{\Gamma} = \begin{pmatrix} 0.55 & 0.48 \\ 0.48 & 0.77 \end{pmatrix}$ for Holstein and Jersey. Assuming the pedigree in Figure 5, we constructed $\mathbf{A}^{-1}$ using $\boldsymbol{\Gamma} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$, which is equivalent to use of regular unknown parent group rules (Quaas 1988) and we also constructed $\mathbf{A}^{\boldsymbol{\Gamma}-1}$ as described before with $\boldsymbol{\Gamma} = \begin{pmatrix} 0.55 & 0.48 \\ 0.48 & 0.77 \end{pmatrix}$; we scaled $\mathbf{A}^{\boldsymbol{\Gamma}-1}$ to refer to the same regular genetic variance multiplying it by the constant $k = 1 + \overline{\text{diag}(\boldsymbol{\Gamma})}/2 - \overline{\boldsymbol{\Gamma}}$. Results are shown in Figure 6.

It can be observed that the sparsity pattern does not change, except for the nonnull values across metafounders. Also, the numbers do not change greatly but diagonal elements are higher because there is shrinkage associated to the metafounders, which is not the case for regular unknown parent groups.

Consider now the variance in the hypothetical crossed Holstein–Jersey individuals. The segregation variance is, by the

```
 2.25  0.25 -1    -1    -0.50  0     0     0     0     0     0     0     0     0
 0.25  2.25  0     0    -0.50 -1    -1     0     0     0     0     0     0     0
-1     0     1.50  0.50  0     0     0    -1     0     0     0     0     0     0
-1     0     0.50  2     0.50  0     0     -1    -1     0     0     0     0     0
-0.50 -0.50  0     0.50  2.50  0     0     0.50 -1     0.50  0    -1    -1     0
 0    -1     0     0     0     1.50  0.50  0     0    -1     0     0     0     0
 0    -1     0     0     0     0.50  1.50  0     0    -1     0     0     0     0
 0     0    -1    -1     0.50  0     0     3     0.50  0    -1     0    -1     0
 0     0     0    -1    -1     0     0     0.50  2.50  0    -1     0     0     0
 0     0     0     0     0.50 -1    -1     0     0     2.50  0    -1     0     0
 0     0     0     0     0     0     0    -1    -1     0     2.53  0.53  0    -1.07
 0     0     0     0    -1     0     0     0     0    -1     0.53  2.53  0    -1.07
 0     0     0     0    -1     0     0    -1     0     0     0     0     2     0
 0     0     0     0     0     0     0     0     0     0    -1.07 -1.07  0     2.13
```

```
 5.41 -1.61 -1.05 -1.05 -0.57  0     0     0     0     0     0     0     0     0
-1.61  4.92  0     0    -0.57 -1.24 -1.24  0     0     0     0     0     0     0
-1.05  0     1.57  0.52  0     0     0    -1.05  0     0     0     0     0     0
-1.05  0     0.52  2.08  0.51  0     0    -1.05 -1.02  0     0     0     0     0
-0.57 -0.57  0     0.51  2.71  0     0     0.51 -1.02  0.55  0    -1.11 -1.02  0
 0    -1.24  0     0     0     1.85  0.62  0     0    -1.24  0     0     0     0
 0    -1.24  0     0     0     0.62  1.85  0     0    -1.24  0     0     0     0
 0     0    -1.05 -1.05  0.51  0     0     3.13  0.52  0    -1.04  0    -1.02  0
 0     0     0    -1.02 -1.02  0     0     0.52  2.57  0    -1.04  0     0     0
 0     0     0     0     0.55 -1.24 -1.24  0     0     3.02  0    -1.11  0     0
 0     0     0     0     0     0     0    -1.04 -1.04  0     2.64  0.57  0    -1.14
 0     0     0     0    -1.11  0     0     0     0    -1.11  0.57  2.78  0    -1.14
 0     0     0     0    -1.02  0     0    -1.02  0     0     0     0     2.05  0
 0     0     0     0     0     0     0     0     0     0    -1.14 -1.14  0     2.28
```

Figure 6 Inverse of the numerator relationship matrix with (up) unknown parent groups (equivalently, with $\mathbf{\Gamma}=0$) or (down) with metafounders and Holstein–Jersey $\mathbf{\Gamma}$ coefficients () scaled to the same genetic variance. Pedigree as in Figure 5.

formula above, increased by $[(0.55+0.77)/2-0.48]/4=0.045$ compared to the variance in the F1.

Discussion

Conceptual developments

This work presents new conceptual developments for pedigree relationships, including ancestral relationships at the founders due to finite size of ancestral population and across-base population relationships due to overlapping. Such development is of conceptual interest *per se* (Kennedy 1991; VanRaden 1992; ter Braak 2010), but it is obliged for genomic evaluations integrating genotyped and nongenotyped individuals. In practice regular genetic evaluations including several base populations and their crosses assume that ancestral populations are of infinite size and unrelated. This leads to unsolved questions. For instance, assume three pure breeds A, B, C and all of their F1 crosses, all in the same environment. If breed A and B are more similar to each other than to breed C, does this need to be included in the genetic analysis? Another typical case is with ruminant population with missing parentages, which are modeled as animals entering from new base populations. These base populations will become gradually more inbred (VanRaden 1992) and they will drift from the oldest base population. Also, they will be related (*i.e.*, the unknown parent group "Holstein2004" will be more related to "Holstein2002" than to "Holstein1994"). All this can be conveniently modeled, estimated, and included in the genetic evaluations using metafounders. As genomic evaluation procedures are becoming more comprehensive, examples of these kind of problems are showing up in the animal breeding literature: Harris and Johnson (2010), Misztal *et al.* (2013), Makgahlela *et al.* (2014), Winkelman *et al.* (2015).

Metafounders and unknown parent groups

Metafounders are closely related to unknown parent groups or genetic groups (Thompson 1979; Quaas 1988). Genetic groups allow estimation of different genetic bases across the same population, which is necessary if the selection process is unknown (*i.e.*, importing animals or missing pedigrees). Genetic values of individuals in a genetic group model can be written as $\mathbf{u}=\mathbf{u}^*+\mathbf{Qg}$, where $\mathbf{g}$ has average values of the genetic groups. Genetic groups are usually considered as fixed but they can be conceived as random (Sullivan and Schaeffer 1994). For random $\mathbf{g}$ and $\mathrm{Var}(\mathbf{u}^*)=\mathbf{A}$ and $r(\mathbf{g})=\mathbf{\Gamma}$, $\mathrm{Var}(\mathbf{u}^*)=\mathbf{A}+\mathbf{Q\Gamma Q}'$. This is similar to $\mathbf{A}^{\mathbf{\Gamma}}$, but does not correctly model crosses and overestimates inbreeding. As pointed out by Kennedy (1991) this traditional formulation of genetic groups did not consider inbreeding or drift. Our work can be seen as a generalization of genetic groups to include inbreeding, drift, and across-group relationships. This generalization overcomes the problems mentioned by Misztal *et al.* (2013), who realized that inclusion of unknown parent groups into single-step methods involved approximations in the setup of joint matrix $\mathbf{H}$.

Inclusion of finite size ancestral populations in genetic evaluation procedures has been largely neglected. Jacquard (1969, 1974) work on relationships in closed populations has been ignored. Independently, VanRaden (1992) made a first contribution to palliate the lack of genealogical information in cattle. He used inbreeding coefficients for unknown parent groups based on inbreeding of contemporaries; here we suggest using genomic information instead. Both ideas can possibly be merged.

A notion related to that of metafounders is *partial relationships* across pairs of individuals due to sharing alleles from some particular origin. This allows modeling the genetic value of an individual as a sum of genetic values from several breeds, and this is known as "splitting breeding values"

(Garcia-Cortes and Toro 2006). The relationship matrix with metafounders can be decomposed into such a structure, as explained in the Supporting Information, File S1.

Metafounders and pedigree and genomic relationships

The use of metafounders with $\boldsymbol{\Gamma}$ relationships allows a reconciliation of pedigree and genomic relationships and inbreeding (Powell *et al.* 2010; Vitezica *et al.* 2011). Homozygosity (or identity) can be considered as deviation from Hardy–Weinberg equilibrium (Wright 1922). These deviations cannot be easily measured because they depend on the assumed allelic frequencies, which change in time. Considering unrelated founders assumes that all founder alleles are different, which is not tenable in view of marker information. By assuming 0.5 allelic frequencies, the reference is constant and there are no ambiguities.

The fact that inbreeding automatically increases when considering metafounders may seem worrying. If the objective of quantifying inbreeding is to describe the incertitude *a priori* of inbred animals (*i.e.*, inbred animals tend to be more variable), this does not seem a concern. Use of pedigree inbreeding with metafounders to quantify inbreeding depression should not be problematic, for two reasons. The first is that adding a constant (roughly $\gamma/2$) to inbreeding will not change estimates. The second is that, due to purge, only "new" rate of inbreeding (ΔF) seems to have a measurable effect (*e.g.*, Hinrichs *et al.* 2007). Recent inbreeding could even be better estimated using metafounders (for instance, in incomplete pedigrees; VanRaden 1992).

Genomic relationships are based on markers, and commercial marker chips are often biased toward intermediate frequencies or toward specific breeds. For instance many markers conceived for *Bos taurus* are monomorphic in *Bos indicus* and their use will result in biased estimates of $\boldsymbol{\Gamma}$. For this reason, the approaches in this work should be considered with caution for such populations. Use of unbiased markers (*e.g.*, from sequence data or from random genotyping across the genome) will result in more accurate estimates of relationships across metafounders, if the populations are distant ones.

Genetic background across populations

Use of metafounders assumes a common genetic background across all base populations. This is typically accepted as true within breed, but breed itself is somewhat ill defined. Some genomic predictions across breeds assume identical genetic background (*i.e.*, Hayes *et al.* 2009; Harris and Johnson 2010). If the hypothesis of a homogeneous genetic background is not acceptable, for instance, in the case of genetic–environment interactions or scale effects, a genetic correlation model can be used (Wei and Vanderwerf 1994; Karoui *et al.* 2012).

Empirical checking

Practical performance of our model has to be ascertained with real data but we give an example of its interest. Winkelman *et al.* (2015), using a simplified single-step GBLUP, reported better performance of the Euclidean distance matrix relationship matrix (Gianola and van Kaam 2008) across breeds and their crosses, compared to **G** adjusted as in Harris and Johnson (2010). We have observed that numerically, **G** matrices based on EDM and **G** matrices based on 0.5 allelic frequencies tend to be similar (unpublished). It would seem that the appeal of the EDM relationship matrix is therefore its independence of within-breed allelic frequencies, as proposed by Christensen (2012). In this work, we have aimed at creating tools to make pedigree relationships compatible with this kind of **G** matrices.

Conclusion

We have defined the notion of metafounders, which can be understood as a limited pool of gametes from which the founders of the pedigree are drawn. Metafounders can also be understood as a generalization of unknown parent groups or genetic groups, which are essential in genetic evaluation of livestock. Use of metafounders makes it possible to analyze pedigreed populations allowing for relatedness within and across base populations, something that is desirable for genetic evaluations combining pedigree and genetic markers. Metafounders can account for extra segregation variances due to crosses of populations. Efficient algorithms exist for computation of relationship matrices and their inverses and inbreeding. Relationships across metafounders can be inferred from marker data. By doing so, compatibility of pedigree and genomic relationships is warranted by construction. This work provides new tools and concepts for genetic evaluation and management of populations.

Acknowledgments

We are grateful to the genotoul bioinformatics platform Toulouse Midi-Pyrenees for providing computing resources. We thank reviewers and editor for useful comments and suggestions. A.L. and Z.G.V. acknowledge financing from INRA SelGen projects X-Gen and SelDir. OFC was supported by Center for Genomic Selection in Animals and Plants (GenSAP) funded by the Danish Council for Strategic Research.

Literature Cited

Aguilar, I., and I. Misztal, 2008 Technical note: Recursive algorithm for inbreeding coefficients assuming nonzero inbreeding of unknown parents. J. Dairy Sci. 91: 1669–1672.

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta *et al.*, 2010 Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93: 743–752.

Cardoso, F., and R. Tempelman, 2004 Hierarchical Bayes multiple-breed inference with an application to genetic evaluation of a Nelore-Hereford population. J. Anim. Sci. 82: 1589–1601.

Christensen, O. F., 2012 Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. Genet. Sel. Evol. 44: 37.

Christensen, O. F., and M. S. Lund, 2010 Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42: 2.

Christensen, O., P. Madsen, B. Nielsen, T. Ostersen, and G. Su, 2012 Single-step methods for genomic evaluation in pigs. Animal 6: 1565–1571.

Colleau, J. J., 2002 An indirect approach to the extensive calculation of relationship coefficients. Genet. Sel. Evol. 34: 409–421.

Colleau, J., and M. Sargolzaei, 2011 MIM: an indirect method to assess inbreeding and coancestry in large incomplete pedigrees of selected dairy cattle. J. Anim. Breed. Genet. 128: 163–173.
Dikmen, S., J. B. Cole, D. J. Null, and P. J. Hansen, 2013 Genome-wide association mapping for identification of quantitative trait loci for rectal temperature during heat stress in Holstein cattle. PLoS ONE 8: e69202.
Elzo, M., 2008 *Animal Breeding Notes*. University of Florida, Gainesville, FL (http://www.animal.ufl.edu/elzo/).
Emik, L. O., and C. E. Terrill, 1949 Systematic procedures for calculating inbreeding coefficients. J. Hered. 40: 51–55.
Garcia-Cortes, L. A., and M. Toro, 2006 Multibreed analysis by splitting the breeding values. Genet. Sel. Evol. 38: 601–615.
Gianola, D., and J. B. C. H. M. van Kaam, 2008 Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178: 2289–2303.
Gibbs, R. A., J. F. Taylor, C. P. Van Tassell, W. Barendse, K. A. Eversole *et al.*, 2009 Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science 324: 528–532.
Harris, B. L., and D. L. Johnson, 2010 Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. J. Dairy Sci. 93: 1243–1252.
Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard, 2009 Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet. Sel. Evol. 41: 51.
Henderson, C. R., 1976 A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32: 69–83.
Henderson, H. V., and S. Searle, 1979 Vec and vech operators for matrices, with some uses in Jacobians and multivariate statistics Can. J. Stat. 7: 65–81.
Hill, W. G., and B. S. Weir, 2011 Variation in actual relationship as a consequence of Mendelian sampling and linkage. Gen. Res. 93: 47–64.
Hinrichs, D., T. H. Meuwissen, J. Ødegard, M. Holt, O. Vangen *et al.*, 2007 Analysis of inbreeding depression in the first litter size of mice in a long-term selection experiment with respect to the age of the inbreeding. Heredity 99: 81–88.
Jacquard, A., 1969 Evolution of genetic structures of small populations. Biodemography and social biology 16: 143–157.
Jacquard, A., 1974 *The Genetic Structure of Populations*. Springer Verlag, Berlin/Heidelberg/New York.
Karigl, G., 1981 A recursive algorithm for the calculation of identity coefficients. Ann. Hum. Genet. 45: 299–305.
Karoui, S., M. J. Carabaño, C. Díaz, and A. Legarra, 2012 Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. Genet. Sel. Evol. 44: 39.
Kennedy, B., 1991 CR Henderson: The unfinished legacy. J. Dairy Sci. 74: 4067–4081.
Kijaas, J. W., D. Townley, B. P. Dalrymple, M. P. Heaton, J. F. Maddox *et al.*, 2009 A genome-wide survey of SNP variation reveals the genetic structure of sheep breeds. PLoS ONE 4: e4668.
Lande, R., 1981 The minimum number of genes contributing to quantitative variation between and within populations. Genetics 99: 541–553.
Legarra, A., I. Aguilar, and I. Misztal, 2009 A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92: 4656–4663.
Legarra, A., G. Baloche, F. Barillet, J. Astruc, C. Soulas *et al.*, 2014a Within-and across-breed genomic predictions and genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech, and Basco-Béarnaise. J. Dairy Sci. 97: 3200–3212.
Legarra, A., O. F. Christensen, I. Aguilar, and I. Misztal, 2014b Single step, a general approach for genomic selection. Livest. Sci. 166: 54–65.
Lo, L. L., R. L. Fernando and M. Grossman, 1993 Covariance between relatives in multibreed populations - additive-model. Theor. Appl. Genet. 87: 423–430.
Lourenco, D. A. L., I. Misztal, S. Tsuruta, I. Aguilar, T. J. Lawlor *et al.*, 2014 Are evaluations on young genotyped animals benefiting from the past generations? J. Dairy Sci. 97: 3930–3942.
Lutaaya, E., I. Mitsztal, J. K. Bertrand, and J. W. Mabry, 1999 Inbreeding in populations with incomplete pedigrees. J. Anim. Breed. Genet. 116: 475–480.
Makgahlela, M., I. Strandén, U. Nielsen, M. Sillanpää, and E. Mäntysaari, 2014 Using the unified relationship matrix adjusted by breed-wise allele frequencies in genomic evaluation of a multibreed population. J. Dairy Sci. 97: 1117–1127.
Meuwissen, T., and Z. Luo, 1992 Computing inbreeding coefficients in large populations. Genet. Sel. Evol. 24: 305–313.
Misztal, I., Z. G. Vitezica, A. Legarra, I. Aguilar, and A. A. Swan, 2013 Unknown-parent groups in single-step genomic evaluation. J. Anim. Breed. Genet. 130: 252–258.
Munilla-Leguizamón, S, and R. J. Cantet, 2010 Equivalence of multibreed animal models and hierarchical Bayes analysis for maternally influenced traits. Genet. Sel. Evol. 42: 20.
Powell, J. E., P. M. Visscher, and M. E. Goddard, 2010 Reconciling the analysis of IBD and IBS in complex trait studies. Nat. Rev. Genet. 11: 800–805.
Quaas, R. L., 1976 Computing the diagonal elements and inverse of a large numerator relationship matrix. Biometrics 32: 949–953.
Quaas, R. L., 1988 Additive genetic model with groups and relationships. J. Dairy Sci. 71: 1338–1345.
Searle, S. R., 1982 *Matrix Algebra Useful for Statistics*. Wiley, Hoboken, NJ.
Slatkin, M., and R. Lande, 1994 Segregation variance after hybridization of isolated populations. Genet. Res. 64: 51–56.
Strandén, I., and O. F. Christensen, 2011 Allele coding in genomic evaluation. Genet. Sel. Evol. 43: 25.
Sullivan, P., and L. Schaeffer, 1994 Fixed *vs.* random genetic groups, pp. 483–486 in *Proceedings, 6th World Congr. Genet. Appl. to Lives. Prod.*, Guelph, Canada, edited by C. Smith, J. S. Gavora, B. Benkel, J. Chesnais, W. Fairfull, J. P. Gibson, B. W. Kennedy, and E. B. Burnside, University of Guelph, Guelph, Ontario, Canada.
ter Braak, C. J., M. P. Boer, L. R. Totir, C. R. Winkler, O. S. Smith *et al.*, 2010 Identity-by-descent matrix decomposition using latent ancestral allele models. Genetics 185: 1045–1057.
Thompson, R., 1977 The estimation of heritability with unbalanced data. II. Data available on more than two generations. Biometrics 33: 497–504.
Thompson, R., 1979 Sire evaluation. Biometrics 35: 339–353.
Ugarte, E., E. Urarte, F. Arrese, J. Arranz, L. Silio *et al.*, 1996 Genetic parameters and trends for milk production of blond-faced Latxa sheep using Bayesian analysis. J. Dairy Sci. 79: 2268–2277.
VanRaden, P., 1992 Accounting for inbreeding and crossbreeding in genetic evaluation of large populations. J. Dairy Sci. 75: 3136–3144.
VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423.
VanRaden, P., K. Olson, G. Wiggans, J. Cole, and M. Tooker, 2011 Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss. J. Dairy Sci. 94: 5673–5682.
Vitezica, Z., I. Aguilar, I. Misztal, and A. Legarra, 2011 Bias in genomic predictions for populations under selection. Genet. Res. 93: 357–366.
Wei, M., and J. Van der Werf, 1994 Maximizing genetic response in crossbreds using both purebred and crossbred information. Anim. Prod. 59: 401–413.
Winkelman, A. M., D. L. Johnson, and B. L. Harris, 2015 Application of genomic evaluation to dairy cattle in New Zealand. J. Dairy Sci. 98: 1–17.
Wright, S., 1922 Coefficients of inbreeding and relationship. Am. Nat. 56: 330–338.

Communicating editor: G. A. Churchill

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.177014/-/DC1

Ancestral Relationships Using Metafounders: Finite Ancestral Populations and Across Population Relationships

Andres Legarra, Ole F. Christensen, Zulma G. Vitezica, Ignacio Aguilar, and Ignacy Misztal

DOI: 10.1534/genetics.115.177014

FILE S1

Decomposition of relationships by population of origin and crosses (splitting breeding values).

García-Cortés and Toro (2006) suggested splitting crossbred relationships as sums of several relationship matrices, one by breed and segregation term. In our proposal, the overall relationship matrix can be decomposed in a similar manner, summing covariances across partial relationships matrices, each one by breed of origin or crosses of breeds. It can be shown that

$$\boldsymbol{A}^{\boldsymbol{\Gamma}} = \sum_{b=1}^{B} \left\{ \boldsymbol{A}^{(b)} \left(1 - \frac{\gamma_b}{2}\right) + \boldsymbol{C}^{(b)} \gamma_b \right\} + \sum_{b=1}^{B} \sum_{b'>b} \boldsymbol{C}_{ii'}^{(b,b')} \gamma_{bb'}$$

where $\boldsymbol{A}^{(b)}$ are partial relationship matrices for origin b and matrices $\boldsymbol{C}^{(b)}$ and $\boldsymbol{C}^{(b,b')}$ describe the covariance of breed fractions across individuals, and they are calculated using rules that are derived below.

Here, a recursive method for constructing $\boldsymbol{A}^{\boldsymbol{\Gamma}}$ is presented in complete generality. An additive relationship matrix should satisfy the following recursions

$$\boldsymbol{A}_{ii}^{\boldsymbol{\Gamma}} = 1 + \boldsymbol{A}_{f(i)m(i)}^{\boldsymbol{\Gamma}}/2$$

$$\boldsymbol{A}_{ii'}^{\boldsymbol{\Gamma}} = (\boldsymbol{A}_{f(i)i'}^{\boldsymbol{\Gamma}} + \boldsymbol{A}_{m(i)i'}^{\boldsymbol{\Gamma}})/2$$

where $f(i)$ and $m(i)$ denote the two parents of individual i, and individual i' is not a direct descendant of individual i. Matrix $\boldsymbol{A}^{\boldsymbol{\Gamma}}$ is defined by base individuals in breed $b = 1, \ldots, p$ being related with relationship coefficient γ_b and inbred with inbreeding coefficient $\gamma_b / 2$, and by base individuals in different breeds $b \neq b'$ being related with relationship coefficient $\gamma_{bb'}$. This may for the base population be expressed as

$$A_{ii'}^{\Gamma} = \sum_{b} f_i^b A_{ii'}^b (1 - \gamma_b / 2) + \sum_{b} \sum_{b'} f_i^b f_{i'}^{b'} \gamma_{bb'},$$

where f_i^b is the breed b proportion of individual i, and $A_{ii'}^b$ is the usual additive relationship when $i \neq i'$ and self-relationships when $i = i'$ between individuals in breed b (and zero when individuals are not breed b), and $\gamma_{bb} = \gamma_b$. The recursions required for an additive relationship matrix are satisfied for this expression as long as all animals are purebred. The recursions are also satisfied when individuals are crossbred with purebred parents with $A_{ii'}^b$ then denoting partial relationships (Garcia-Cortes and Toro, 2006), but they are not satisfied in general.

Here, we split $\mathbf{A}^{\Gamma}$ into several components and derive how the recursions look like for the components. The formula above suggests that

$$A^{\Gamma}_{ii'} = \sum_b A^b_{ii'}(1-\gamma_b/2) + \sum_b C^b_{ii'}\gamma_b + \sum_b \sum_{b':b'>b} C^{b,b'}\gamma_{bb'},$$

where $\mathbf{A}^b$ the breed b specific partial relationship matrix (Garcia-Cortes and Toro, 2006) such that

$$A^b_{ii} = f^b_i + A^b_{f(i)m(i)}/2,$$

$$A^b_{ii'} = (A^b_{f(i)i'} + A^b_{m(i)i'})/2,$$

when individual i' is not a direct descendant of individual i. Inserting the suggested form of $\mathbf{A}^{\Gamma}$ into the recursive formulas of $\mathbf{A}^{\Gamma}$ we obtain (having used $1 = \sum_b f^b_i$) that diagonal elements should satisfy

$$A^b_{ii}(1-\gamma_b/2) + C^b_{ii}\gamma_b = f^b_i + (A^b_{f(i)m(i)}(1-\gamma_b/2) + C^b_{f(i)m(i)}\gamma_b)/2,$$

$$C^{b,b'}_{ii}\gamma_{bb'} = C^{b,b'}_{f(i)m(i)}\gamma_{bb'}/2,$$

from which we obtain (having used $f^b_i = f^b_i(1-\gamma_b/2) + f^b_i\gamma_b/2$) that $A^b_{ii} = f^b_i + A^b_{f(i)m(i)}/2$ is satisfied and

$$C^b_{ii} = (f^b_i + C^b_{f(i)m(i)})/2,$$

$$C^{b,b'}_{ii} = C^{b,b'}_{f(i)m(i)},$$

and off-diagonal elements should satisfy

$$A^b_{ii'} = (A^b_{f(i)i'} + A^b_{m(i)i'})/2,$$

$$C^b_{ii'} = (C^b_{f(i)i'} + C^b_{m(i)i'})/2,$$

$$C^{b,b'}_{ii'} = (C^{b,b'}_{f(i)i'} + C^{b,b'}_{m(i)i'})/2.$$

The description is completed by specifying that the recursions start by $A^b_{ii} = 1$, $A^b_{ii'} = 0$, $C^b_{ii} = C^b_{ii'} = 1$ when i and $i' \neq i$ are base animals of breed b, and $C^{b,b'}_{ii'} = 1$ when i and i' are base animals in different breeds b and b'.

We note that elements in matrices $\mathbf{A}^b$ and $\mathbf{C}^b$ are only non-zero for individuals which contain a breed b proportion. The matrix $\mathbf{C}^{b,b'}$ is complicated, but the rule is that off-diagonal elements are nonzero for pairs of individuals where one of them contain a breed b proportion and the other a b' proportion, and diagonal elements are non-zero for individuals where one of its parent contain a breed b and the other a breed b' proportion (i.e. certain crossbred animals).

FILE S2

Code for algorithms related to metafounders

File S2 is available for download as a compressed file (metafounders_code.tar.gz) at

www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.177014/-/DC1