Redefining and interpreting genomic relationships of metafounders

Andres Legarra, Matias Bermann, Quanshun Mei, Ole F. Christensen

Genetics Selection Evolution

SHORT COMMUNICATION

Open Access

Redefining and interpreting genomic relationships of metafounders

Andres Legarra[1*], Matias Bermann[2], Quanshun Mei[3] and Ole F. Christensen[4]

Abstract

Metafounders are a useful concept to characterize relationships within and across populations, and to help genetic evaluations because they help modelling the means and variances of unknown base population animals. Current definitions of metafounder relationships are sensitive to the choice of reference alleles and have not been compared to their counterparts in population genetics—namely, heterozygosities, F_{ST} coefficients, and genetic distances. We redefine the relationships across populations with an arbitrary base of a maximum heterozygosity population in Hardy–Weinberg equilibrium. Then, the relationship between or within populations is a cross-product of the form $\mathbf{\Gamma}_{(b,b')} = \left(\frac{2}{n}\right)(2\mathbf{p}_b - \mathbf{1})(2\mathbf{p}_{b'} - \mathbf{1})'$ with $\mathbf{p}$ being vectors of allele frequencies at n markers in populations b and b'. This is simply the genomic relationship of two pseudo-individuals whose genotypes are equal to twice the allele frequencies. We also show that this coding is invariant to the choice of reference alleles. In addition, standard population genetics metrics (inbreeding coefficients of various forms; F_{ST} differentiation coefficients; segregation variance; and Nei's genetic distance) can be obtained from elements of matrix $\mathbf{\Gamma}$.

Background

Because selection proceeds within breeds, animal breeders have not often dealt with relationship across populations, contrary to evolutionary geneticists, e.g. [1]. Thus, pedigree-based modelling of relationships across animals for genetic evaluation assumed that base populations (Unknown Parent Groups or Genetic Groups) were unrelated and of infinite size. However, populations differ in heterozygosity and are more or less close to each other [2]. In theory, this can be modelled using phylogenetic trees, which can be converted into covariances of gene content at loci [3]. However, these trees are notoriously difficult to estimate in practice. VanRaden [4] proposed methods to model relationships across populations, both within and across breeds, in particular to correctly estimate inbreeding when pedigree information is missing, but his ideas were not broadly applied. With the introduction of genomic evaluation and selection, it was noticed that the assumption of unrelated populations was untenable, and differences across pedigree bases of the different breeds (or groups within breeds) had to be explicitly modelled when pedigree and genomic data were combined. Defining a relationship implies defining a genetic base, which is difficult in practice due to the lack of a clear "starting point". This motivated the theory of "metafounders" (abbreviated MF in the following) [5–7]. The theory is actually composed of two parts, which are somewhat mixed up in the literature. The first part consists in defining pseudo-individuals (MF) which represent populations. The relationships across these MF, encapsulated in a matrix $\mathbf{\Gamma}$, model covariances between the means of these populations [6], populations' homozygosities, and their similarity. These relationships $\mathbf{\Gamma}$ can be extended via the tabular method [7], in a manner that is

*Correspondence:
Andres Legarra
andres.legarra@uscdcb.com
[1] CDCB, 4201 Northview Drive, Bowie, MD 20716, USA
[2] Animal and Dairy Science, University of Georgia, 425 River Rd, Athens, GA 30602, USA
[3] Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA
[4] Center for Quantitative Genetics and Genomics, Aarhus University, C. F. Møllers Allé 3, Bld. 1130, 8000 Aarhus C, Denmark

a generalization of the regular theory for pedigree relationships, to model covariances across individuals within and across breeds [6, 7], including segregation variances e.g. in F2 animals. The modelling of the covariance across breeds using $\mathbf{\Gamma}$ implies that the allele substitution effects are defined across breeds [6, 8]. The second part of the theory is the definition of a genetic base from which to define the population means and their covariances. It turns out that a convenient reference is an "absolute" reference point, which is an ideal population with allele frequencies of 0.5 at biallelic markers and therefore with the maximum possible heterozygosity in Hardy–Weinberg equilibrium (HWE) [9]. This is also convenient for compatibility with genomic relationships based on the same 0.5 reference point [6]. The use of 0.5 as a reference leads to a mathematical definition of $\mathbf{\Gamma}$ as (co)variances of allele frequencies across and within populations [9]. However, this definition is (empirically) sensitive to the choice of reference alleles. In addition, the meaning of $\mathbf{\Gamma}$ is not yet fully understood in terms of commonly used population genetics metrics, such as inbreeding coefficients, heterozygosity, and genomic relationships across breeds or populations [2]. For instance, a potential user of the theory of MF may be at odds on how to actually compute (or estimate) $\mathbf{\Gamma}$ from known allele frequencies. Moreover, the user may want to compare inbreeding coefficients or heterozygosities to other population genetics metrics. This is increasingly important with the growing use of genomic measurements for managing genetic diversity [10].

The aim of this short note is to clarify the following two points: (1) give equivalent definitions of $\mathbf{\Gamma}$ that are invariant to the (maybe non-random) choice of reference alleles; and (2) explain how to interpret $\mathbf{\Gamma}$ in terms of inbreeding and heterozygosity. These results are used in the companion paper [11] that is dedicated to methods for estimation of $\mathbf{\Gamma}$ in complex populations.

Theory

Definition of Γ invariant to the choice of reference alleles

The definition of $\mathbf{\Gamma}$ in [5] can be understood as "the relationship across individuals in the base pedigree population(s), relative to a conceptual base population with all allele frequencies $p = 0.5$". Note that, here, the population for which $p = 0.5$ is merely conceptual.

Garcia-Baccino et al. [9] later found out that $\gamma_{b,b'} = 8cov(p_b, p_{b'})$ for populations b and b'. This comes from the fact that the mean and the homozygosity of each population refer to a conceptual base population where the expectation of allele frequencies is $\overline{p} = 0.5$. In other words, some p_i will be lower than 0.5 and some will be higher, but they average 0.5. This is reasonable to assume, conceptually, by randomly labeling an allele as the reference. However, empirical treatment of observed genomic data often delivers $\overline{p} \neq 0.5$, even when addressing multiple populations, as populations are real (observed). For this reason, two researchers using different choice of reference alleles for the same dataset may get different numbers from $\mathbf{\Gamma}$ if they apply blindly $\gamma_{b,b'} = 8cov(p_b, p_{b'})$. The same happens if one uses sequences simulated by coalescence, which call "1" the mutant and "0" the wild allele.

Consider the matrix $\mathbf{M}$ which contains genotypes coded {0,1,2}. The values of genomic relationships obtained as cross-product $\mathbf{G} = \frac{1}{s}\mathbf{ZZ}' = \frac{1}{s}(\mathbf{M} - \mathbf{2p}')(\mathbf{M} - \mathbf{2p}')'$ [12] with s a scale factor (typically $s = 2\sum p_i q_i$ or $s = n/2$ for n markers) are invariant to changes in the reference alleles used to define $\mathbf{M}$ and $\mathbf{p}$. Although rarely explicitly stated, this invariance is well known. We show proof in the Appendix.

In the same spirit, next we need an alternative definition of $\mathbf{\Gamma}$ which is invariant to the choice of the reference allele. In [7], $\mathbf{\Gamma}$ and metafounders are defined from alleles in the base-population being sampled from pools of alleles, and counting how many are identical or not. Similarly, for a given labelling of alleles, we need to define unambiguously $\mathbf{\Gamma}$, without imposing the condition $\overline{p} = 0.5$. To arrive to a meaningful definition, we notice that $\gamma_{b,b}$ (the self-relationship of MF b) is simply the average (genomic) relationship across animals that form the corresponding base population b, and the relationship $\gamma_{b,b'}$ of populations b and b' is the average relationship across all possible pairs of individuals, one from b and the other one from b'. This definition was already presented in [13–15] and (unaware of these works) was rediscovered and accommodated to genomic relationships [7].

It follows (as described in the Appendix) that the self-relationship of a population b with itself is $\gamma_{b,b} = \frac{1}{s}\sum_{k=1}^{n}(2p_{b(k)} - 1)^2 = \frac{1}{s}(2\mathbf{p}_b - \mathbf{1})(2\mathbf{p}_b - \mathbf{1})'$ with $s = \frac{n}{2}$, n being the number of markers, and the relationship across populations b and b' is $\gamma_{b,b'} = \frac{1}{s}(2\mathbf{p}_b - \mathbf{1})(2\mathbf{p}_{b'} - \mathbf{1})'$. This is purely a quantitative genetics definition, i.e. $\mathbf{\Gamma}$ is a feature of the population(s).

Equivalently, we can see $\mathbf{\Gamma}$ as genomic relationships of the base populations means, seen as individuals, which requires the "genotypes" of each population. If $\mathbf{p}_b$ is a vector of allele frequencies of the base population b, we can see $2\mathbf{p}_b$ as the "genotype" of the base population. The centered "genotype" of the base population, with respect to the fictitious population with all $p = 0.5$, is simply $\mathbf{z}_b = 2\mathbf{p}_b - \mathbf{1}$ where 1 is twice 0.5, i.e. the reference allele frequency. Thus, the genomic relationship matrix across populations is simply $\mathbf{\Gamma} = \frac{1}{s}\mathbf{ZZ}'$ where $\mathbf{Z}$ contains twice the allele frequencies of the populations, minus 1: $z_{b,k} = 2p_{b,k} - 1$. We note that this is strictly the same definition as in VanRaden [7], if we consider

that allele frequencies are "genotypes" of populations—this idea is e.g. in Tier [16]. For statistical inference, $\mathbf{\Gamma}$ is a parameter of a distribution from which "genotypes" (twice the allele frequencies minus 1) of base populations are sampled.

We also want to stress that if $E(p_b) = E(p_{b'}) = 0.5$, then $\gamma_{b,b'} = \frac{1}{s}(2\mathbf{p}_b - \mathbf{1})(2\mathbf{p}_{b'} - \mathbf{1})' = 8Cov(p_{b(i)}, p_{b'(i)})$ as in [9]. However, the new formulation is more general, and correctly considers the cases where $\overline{p}_b \neq 0.5$, for instance across several breeds or when one of the "wild" or "mutant" alleles is the reference allele.

Interpretation of Γ as heterozygosities or inbreeding coefficients of populations

In this section, we try to relate the values in $\mathbf{\Gamma}$ to diversity and homozygosity of the population. Consider average heterozygosity of a population, $\overline{\mathcal{H}} = \overline{2p_i q_i}$. The conceptual population with $p = 0.5$ has $\overline{\mathcal{H}}_{max} = 0.5$, whereas the observed population b has $\overline{\mathcal{H}}_b = \overline{(2p_{b(i)}q_{b(i)})}$. We can obtain, after some algebra:

$$\frac{\gamma_{b,b}}{2} = \frac{1}{2}\frac{2}{n}\sum_{i=1,n}(2p_{b(i)} - 1)^2 = \frac{0.5 - \overline{(2p_{b(i)}q_{b(i)})}}{0.5} = \frac{(\overline{\mathcal{H}}_{max} - \overline{\mathcal{H}}_b)}{\overline{\mathcal{H}}_{max}}.$$

From this, it follows that $\overline{\mathcal{H}}_b = \overline{\mathcal{H}}_{max}(1 - \frac{\gamma_{b,b}}{2})$, and $\frac{\gamma_{b,b}}{2}$ can be seen as an inbreeding coefficient. In other words, $\frac{\gamma_{b,b}}{2}$ measures the relative change in heterozygosity from average $\overline{\mathcal{H}}_{max} = 0.5$ to $\overline{\mathcal{H}}_b = \frac{1}{2} - \frac{\gamma_{b,b}}{4} = \overline{(2p_{b(i)}q_{b(i)})}$. Indeed, Jacquard [17] called $\frac{\gamma_{b,b}}{2}$ the inbreeding coefficient of a population.

Meuwissen et al. [10] reviewed different measurements of inbreeding for genomic management. Among these, we can find a first inbreeding coefficient based on homozygosity:

$$F_{hom} = 1 - \frac{\mathcal{H}_t}{\mathcal{H}_0} = 1 - \frac{1}{n}\sum\frac{2p_{b(t,i)}q_{b(t,i)}}{2p_{b(0,i)}q_{b(0,i)}},$$

and when we impose $p_{b(0,i)} = q_{b(0,i)} = 0.5$, this expression yields:

$$F_{hom} = 1 - 2\overline{(2p_{b(i)}q_{b(i)})} = \frac{\gamma_{b,b}}{2}.$$

Thus, $\frac{\gamma_{b,b}}{2}$ has the same interpretation as above, i.e. in terms of change in heterozygosity.

The second inbreeding coefficient in [10] is based on drift:

$$F_{drift} = \frac{1}{n}\sum_{i=1,n}\frac{(p_{b(i)} - p_{b(0,i)})^2}{p_{b(0,i)}q_{b(0,i)}},$$

and again, when we impose $p_{b(0,i)} = q_{b(0,i)} = 0.5$, this yields:

$$F_{drift} = \frac{1}{n}\sum_{i=1,n}\frac{(2p_{b(i)} - 1)^2}{0.25} = \frac{\gamma_{b,b}}{2},$$

identically to the previous one. However, note that here we are imposing $p_{b(0,i)} = q_{b(0,i)} = 0.5$, which means that, in fact, the value $\frac{\gamma_{b,b}}{2}$ is not truly due to genealogical drift from a real, existing population (rather, it describes change from a merely conceptual one), thus describing different values of $\frac{\gamma_{b,b}}{2}$ as due to drift would be a misnomer.

The third inbreeding coefficient is defined as follows. If $\gamma_{b,b}$ is a relationship coefficient, then:

$$F_b = \gamma_{b,b} - 1,$$

can be seen as an inbreeding coefficient—a measure of homozygosity of the population b, not of any individual. Substituting $\gamma_{b,b}$ by $\gamma_{b,b} = 2\frac{0.5 - \overline{(2p_{b(i)}q_{b(i)})}}{0.5}$ (obtained before) gives:

$$F_b = 1 - 4\overline{(2p_{b(i)}q_{b(i)})}.$$

If average heterozygosity $\overline{(2p_{b(i)}q_{b(i)})}$ is 0, then $F_b = 1$, meaning that there is complete inbreeding and lack of heterozygosity. If average heterozygosity (under HWE conditions) is maximal: $\overline{(2p_{b(i)}q_{b(i)})} = 0.5$, then inbreeding $F_b = -1$, meaning complete heterozygosity (under HWE conditions). Again, $\gamma_{b,b} - 1$ describes a feature of the population—the homozygosity compared to a population in HWE with maximum heterozygosity.

Interpretation of Γ in terms of segregation variance, genetic distances and Fst

A commonly used measure of genetic distance across populations is Nei's minimum genetic distance, $D_{b,b'}$, which is also the numerator of the F_{ST} differentiation index, and is simply [1]:

$$D_{b,b'} = \frac{1}{n}\sum(p_{b(i)} - p_{b'(i)})^2.$$

After some algebra, we get (as described in the Appendix):

$$D_{b,b'} = \frac{\gamma_b}{8} + \frac{\gamma_{b'}}{8} - \frac{\gamma_{bb'}}{4},$$

which also corresponds to the segregation variance, i.e. the difference in genetic variance from F1 to F2 crosses of b and b' [7]. Thus, we can use γ coefficients to describe genetic distances.

The F_{ST} coefficient, applying the Hudson et al. [18] definition as $F_{ST} = \frac{(H_{between} - H_{within})}{H_{between}}$ is shown in the Appendix to be:

$$F_{ST} = \frac{\frac{\gamma_b}{8} + \frac{\gamma_{b'}}{8} - \frac{\gamma_{bb'}}{4}}{\frac{1}{2} - \frac{\gamma_{bb'}}{4}},$$

which again shows that γ relates to already known descriptors of differentiation. Note that this formula takes into account the covariance of allele frequencies in both populations but also the heterozygosity in each population. For instance, assume two breeds fixed for opposite alleles as follows:

Breed b	Breed b'
A	a
c	C
D	d
e	E

,

and so on. We have $\Gamma_{b,b} = \Gamma_{b',b'} = 2$ and $\Gamma_{b,b'} = -2$. These yield $F_{ST} = 1$ as expected.

Other reference base populations

The theory of MF uses 0.5 as the frequency of the reference allele because it is convenient for many purposes. However, one could define relationships from a particular "reference" base population—for instance, in single breed evaluations, it could be the oldest base population in the breed; but it could be a wild ancestor, or an outgroup population. Then, equations should include frequencies in the outgroup (p_o) as:

$$\gamma_{b,b} = \frac{1}{2\sum p_{o(i)}q_{0(i)}} \sum \left(2p_{b(i)} - 2p_{o(i)}\right)^2,$$

$$\gamma_{b,b'} = \frac{1}{2\sum p_{o(i)}q_{0(i)}} \sum \left(2p_{b(i)} - 2p_{o(i)}\right)\left(2p_{b'(i)} - 2p_{o(i)}\right).$$

For MF that describe missing parents across years within breed (typically modelled as unknown parent groups), choosing as reference base population the very first MF in chronological order may be convenient. This would yield a self-relationship of the reference base population of $\gamma_{o,o} = 0$ and would naturally lead to use the genetic variance of the base population as the parameter of models using $\mathbf{\Gamma}$ [7]. The problems are (a) $\mathbf{\Gamma}$ would be no longer full rank and (b) $p_{o(i)}$ is often unknown.

Discussion

Description of the genetic features of a population in itself is a subject that has not been frequently addressed by animal breeders, because the assumption of unrelated base populations is a simple and efficient one [19], even if the theory could be improved [20–22]. However, the advent of genomic selection led to reconsider modelling means and variances of these populations, in particular because of an acute need for the so-called single step genomic best linear unbiased prediction (ssGBLUP) [23, 24]. At the same time, the concepts of inbreeding, heterozygosity, and drift have been thoroughly revisited with the advent of genomic evaluation [10, 12, 25].

The concept of MF tries to merge the genetic description of populations and the relationships across them [17, 26] with a relationship formulation that can be used for pedigree and genomic selection, giving an explicit modelling to differences in means, segregation variance, or covariances across crossbreds with variable composition. It does this in a manner that is, by construction, compatible (at least in principle) with individual single nucleotide polymorphism (SNP)-based measurements of relationships.

This short note presents an alternative derivation of MF relationships in terms of cross-products of gene content (of the populations), which had not been fully described so far [6, 7, 9]. This derivation is fully compatible to previous derivations and allows to derive estimators more easily for relationships across MF (see in the companion paper [11]). Moreover, we also derive other subproducts that frame our theory with population genetics metrics such as F_{ST} or heterozygosity. These relationships have been derived assuming the conceptual base population with $p = 0.5$. In addition, the now more coherent theory could be used e.g. to establish priorities for management of diversity across breeds including crosses [27]. Note that whereas values of $\mathbf{\Gamma}$ itself assume the conceptual base population with $p = 0.5$, using them for management of diversity would lead to increase heterozygosities at markers, which may not be desirable [10], whereas on the other hand it gives a unified framework which may be attractive. To solve the issue, Colleau et al. [28] suggested "... [converting] the results into more conventional scales..." through scale and shift factors, but that does not resolve the problem of increasing homozygosities versus conserving existing allele frequencies.

On the other side, this theory is somehow compromised because the markers used are not random—they have been tailored, for commercial chips, to be polymorphic in major commercial breeds. For this reason, the relationships obtained in this way, in particular for minor breeds, should not be taken at face value.

Overall, we believe that this note contributes towards a more general and encompassing theory of diversity and relationships, which would be useful both for management diversity and for prediction.

Conclusions

Metafounders are a concept that describes genetic variation and co-variation within and across finite populations. We presented alternative, new definitions of the concept of MF in terms of cross-product of allele frequencies of populations. The new definitions are more general and can be related to existing concepts of genetic distances, heterozygosity or inbreeding, and they can be naturally integrated into genomic and pedigree-based predictions. We expect that these new definitions will help develop conceptual and practical tools for population management and selection.

Appendix

Matrix G is invariant to changes in reference alleles

This can be shown as follows. Consider the genotypes of two individuals, row vectors $\mathbf{z}_i$ and $\mathbf{z}_j$, which contain values of $-1, 0, 1$. The genomic relationship of individuals i, j $G_{(i,j)} = \frac{1}{s}\sum_k z_{i,k}z_{j,k}$ where s a scaling factor (for instance $s = 2\sum p_i q_i$; or $s = 2\sum 0.5^2 = \frac{n}{2}$ with n the number of markers, e.g. assumed to have a frequency of 0.5) and $z_{i,k} = m_{i,k} - 2p_k$ where $m_{i,k} = \{0, 1, 2\}$ copies of the reference allele and p_k an assumed frequency for the reference allele at locus k. Change of the reference allele results in switching to $m^{new}_{i,k} = \{2, 1, 0\}$ *i.e.* $m^{new}_{i,k} = 2 - m_{i,k}$ and $p^{new}_k = 1 - p_k$. As a result $z^{new}_{i,k} = m^{new}_{i,k} - 2p^{new}_k = -z_{i,k}$ and the negative sign cancels at the crossproduct: $z^{new}_{i,k} z^{new}_{j,k} = \left(-z_{i,k}\right)\left(-z_{j,k}\right) = z_{i,k}z_{j,k}$. A similar argument holds for the value of s, i.e. even if the reference allele is swapped, the values of $p_i q_i$ do not change. In particular, the proof does not assume *any* value for allele frequencies. Thus, the value of $G_{i,j}$ is invariant to the choice of the reference allele.

Definition of Γ invariant to changes in reference alleles

Within populations

To define unambiguously $\mathbf{\Gamma}$ as a function of observed allele frequencies in each base population, without imposing the condition $\bar{p} = 0.5$, we notice that $\gamma_{(b)}$ is simply the average genomic relationship across animals in the corresponding base population b. Then, we derive the expected value of the average $\mathbf{G}$ taking into account the allele frequencies in HWE.

First, we consider a single population, b. The crossproducts $z_i z_j$ with scalars z_i (z_j) the genotype at one locus for individual i (j) coded as {− 1,0,1} are either 1 (for same homozygotes) or −1 (for opposite homozygotes), and these values occur with frequencies that can be obtained from the following Punnet square (here we omit the subindex b for clarity) with crossproducts $z_i z_j$ with gamete frequencies of individual i (rows) and j (columns):

	p^2	$2pq$	q^2
p^2	1	0	−1
$2pq$	0	0	0
q^2	−1	0	1

The expected value of $z_i z_j$ for the founders of a population is therefore:

$$E_{\text{founders}}(z_i z_j) = p^2\left(p^2 - q^2\right) - q^2\left(p^2 - q^2\right) = \left(p^2 - q^2\right)\left(p^2 - q^2\right) = (p-q)^2 = (2p-1)^2.$$

Then, we sum all n loci and we divide by the scale $s = \frac{n}{2}$ (which is equivalent to assuming a conceptual base population with maximum heterozygosity), which gives:

$$\gamma_{b,b} = \frac{2}{n}\sum\nolimits_i \left(2p_{b(i)} - 1\right)^2 = \frac{2}{n}(2\mathbf{p}_b - \mathbf{1})(2\mathbf{p}_b - \mathbf{1})'. \quad (1)$$

for $\mathbf{p}_b$ the row vector with frequencies in population b.

Note that if $\overline{p_b} = \frac{1}{n}\sum p_{b(i)} = 0.5$, this is equivalent (as expected) to $\gamma = 8var(p_i)$ in Garcia-Baccino et al. [9], where random labelling of alleles is assumed, and thus $\overline{p_b} = 0.5$ holds.

Anyway, Eq. (1) is invariant to choosing $p = freq(A)$ or to choosing $p = freq(a)$ (in other words, to the choice of reference allele "A" or "a") since all that it counts is the absolute deviation of $p_{b(i)}$ from 0.5.

Across populations

Now we compute the average genomic relationship across *two* populations in HWE, b and b', with respective frequencies p_b and $p_{b'}$ as follows:

	$p^2_{b'}$	$2p_{b'}q_{b'}$	$q^2_{b'}$
p^2_b	1	0	−1
$2p_b q_b$	0	0	0
q^2_b	−1	0	1

This gives that across all founders in both populations:

$$\begin{aligned} E_{\text{founders}}(z_i z_j) &= p^2_b\left(p^2_{b'} - q^2_{b'}\right) - q^2_b\left(p^2_{b'} - q^2_{b'}\right) \\ &= \left(p^2_b - q^2_b\right)\left(p^2_{b'} - q^2_{b'}\right) = (p_b - q_b)(p_{b'} - q_{b'}). \\ &= (2p_b - 1)\left(2p_{b'} - 1\right) \end{aligned}$$

As before, this is invariant to the reference alleles. For instance, assume that the reference allele is switched, so that the new allele frequency is $p^* = 1 - p$. This would give

$$\left(2p_b^* - 1\right)\left(2p_{b'}^* - 1\right) = (2(1-p_b)-1)(2(1-p_{b'})-1)$$
$$= (-2p_b+1)(-2p_{b'}+1) = (2p_b-1)(2p_{b'}-1).$$

Now, summing across all loci and using the scaling $s = \frac{n}{2}$ as before gives:

$$\gamma_{b,b'} = \frac{2}{n}\sum\left(2p_{b(i)}-1\right)\left(2p_{b'(i)}-1\right) = \frac{2}{n}(2\mathbf{p}_b - \mathbf{1})(2\mathbf{p}_{b'} - \mathbf{1})'. \quad (2)$$

Again, if $E(p_b) = E(p_{b'}) = 0.5$, then $\gamma_{b,b'} = 8cov\left(p_{b(i)}, p_{b'(i)}\right)$ as in [9]. However, this formulation is much more general, and correctly considers the cases where $\overline{p}_b \neq 0.5$, for instance across several breeds or when the "wild" or "mutant" allele is the reference allele.

Nei's genetic distance

Nei's minimum genetic distance, $D_{b,b'}$, which is also the numerator of the F_{ST} differentiation index, is simply:

$$D_{b,b'} = \frac{1}{n}\sum\left(p_{b(i)} - p_{b'(i)}\right)^2.$$

This can be obtained in terms of Γ as follows. First, expand the equation above:

$$D_{b,b'} = \frac{1}{n}\sum\left(p_{b(i)}\right)^2 + \frac{1}{n}\sum\left(p_{b'(i)}\right)^2 - \frac{1}{n}2\sum\left(p_{b(i)} - p_{b'(i)}\right).$$

Then express each term as a function of Γ coefficients:

$$\frac{1}{n}\sum\left(p_{b(i)}\right)^2 = \frac{\gamma_b}{8} + \frac{1}{n}\sum(p_{b(i)}) - \frac{1}{4},$$

$$\frac{1}{n}\sum\left(p_{b'(i)}\right)^2 = \frac{\gamma_{b'}}{8} + \frac{1}{n}\sum(p_{b'(i)}) - \frac{1}{4},$$

$$\frac{1}{n}\sum\left(p_{b(i)}p_{b'(i)}\right) = \frac{\gamma_{bb'}}{8} + \frac{1}{2}\frac{1}{n}\sum(p_{b(i)}) + \frac{1}{2}\frac{1}{n}\sum(p_{b'(i)}) - \frac{1}{4}.$$

Substituting above we obtain:

$$D_{b,b'} = \frac{\gamma_b}{8} + \frac{\gamma_{b'}}{8} - \frac{\gamma_{bb'}}{4},$$

which corresponds as well to the segregation variance in an F2 from b and b' [7].

Derivation of the F_{ST}

The F_{ST} in Hudson et al. [18] as described by Bhatia et al. [29] is:

$$F_{ST} = \frac{H_{between} - H_{within}}{H_{between}},$$

where $H_{between} - H_{within} = D_{b,b'} = \frac{\gamma_b}{8} + \frac{\gamma_{b'}}{8} - \frac{\gamma_{bb'}}{4}$ as above. Then, using the identity for $\frac{1}{n}\sum\left(p_{b(i)}p_{b'(i)}\right)$ above we get:

$$H_{between} = \frac{1}{n}\sum\left(p_{b(i)}\left(1-p_{b'(i)}\right) + \left(1-p_{b(i)}\right)p_{b'(i)}\right)$$
$$= \frac{1}{n}\sum\left(p_{b(i)}\right) + \frac{1}{n}\sum\left(p_{b'(i)}\right)$$
$$+ -2\frac{1}{n}\sum\left(p_{b(i)}p_{b'(i)}\right).$$
$$= \frac{1}{2} - \frac{\gamma_{b,b'}}{4}$$

Combining both terms gives:

$$F_{ST} = \frac{\frac{\gamma_b}{8} + \frac{\gamma_{b'}}{8} - \frac{\gamma_{bb'}}{4}}{\frac{1}{2} - \frac{\gamma_{b,b'}}{4}} = \frac{\gamma_b + \gamma_{b'} - 2\gamma_{bb'}}{4 - 2\gamma_{bb'}}.$$

Acknowledgements
Authors thank the reviewers for their meticulous advice.

Author contributions
AL put together the first version of the theory that was corrected and completed by the other authors. All authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

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

Received: 8 September 2023 Accepted: 18 March 2024
Published online: 02 May 2024

References

1. Nei M. Molecular evolutionary genetics. NewYork: Columbia University Press; 1987.
2. VanRaden PM, Olson KM, Wiggans GR, Cole JB, Tooker ME. Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss. J Dairy Sci. 2011;94:5673–82.
3. Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, Blott S, et al. Detecting selection in population trees: the Lewontin and Krakauer test extended. Genetics. 2010;186:241–62.
4. VanRaden PM. Accounting for inbreeding and crossbreeding in genetic evaluation of large populations. J Dairy Sci. 1992;75:3136–44.
5. Christensen OF. Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. Genet Sel Evol. 2012;44:37.
6. Christensen OF, Legarra A, Lund MS, Su G. Genetic evaluation for three-way crossbreeding. Genet Sel Evol. 2015;47:98.
7. Legarra A, Christensen OF, Vitezica ZG, Aguilar I, Misztal I. Ancestral relationships using metafounders: finite ancestral populations and across population relationships. Genetics. 2015;200:455–68.
8. Stuber CW, Cockerham CC. Gene effects and variances in hybrid populations. Genetics. 1966;54:1279–86.

9. Garcia-Baccino CA, Legarra A, Christensen OF, Misztal I, Pocrnic I, Vitezica ZG, et al. Metafounders are related to Fst fixation indices and reduce bias in single-step genomic evaluations. Genet Sel Evol. 2017;49:34.
10. Meuwissen THE, Sonesson AK, Gebregiwergis G, Woolliams JA. Management of genetic diversity in the era of genomics. Front Genet. 2020;11:880.
11. Legarra A, Bermann M, Mei Q, Christensen OF. Estimating genomic relationships of metafounders across and within breeds using maximum likelihood, pseudo- expectation-maximization maximum likelihood and increase of relationships. Genet Sel Evol. 2024. https://doi.org/10.1186/s12711-024-00892-9.
12. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
13. Wright S. The genetical structure of populations. Ann Eugen. 1949;15:323–54.
14. Jacquard A. Inbreeding: one word, several meanings. Theor Popul Biol. 1975;7:338–63.
15. Cockerham CC. Group inbreeding and coancestry. Genetics. 1967;56:89–104.
16. Tier B, Meyer K, Swan A. On implied genetic effects, relationships and alternate allele coding. In: Proceedings of the 11th World Congress on Genetics Applied to Livestock Production: 11–16 February 2018; Auckland. 2018.
17. Jacquard A. The genetic structure of populations. Berlin: Springer-Verlag; 1974.
18. Hudson RR, Slatkin M, Maddison WP. Estimation of levels of gene flow from DNA sequence data. Genetics. 1992;132:583–9.
19. Quaas RL. Additive genetic model with groups and relationships. J Dairy Sci. 1988;71:1338–45.
20. Lo LL, Fernando RL, Grossman M. Covariance between relatives in multibreed populations—additive-model. Theor Appl Genet. 1993;87:423–30.
21. Kennedy BW, Henderson CR. CR Henderson: the unfinished legacy. J Dairy Sci. 1991;74:4067–81.
22. Garcia-Cortes LA, Toro MA. Multibreed analysis by splitting the breeding values. Genet Sel Evol. 2006;38:601–15.
23. Strandén I, Aamand GP, Mäntysaari EA. Single-step genomic BLUP with genetic groups and automatic adjustment for allele coding. Genet Sel Evol. 2022;54:38.
24. Misztal I, Vitezica Z-G, Legarra A, Aguilar I, Swan AA. Unknown-parent groups in single-step genomic evaluation. J Anim Breed Genet. 2013;130:252–8.
25. Toro MÁ, García-Cortés LA, Legarra A. A note on the rationale for estimating genealogical coancestry from molecular markers. Genet Sel Evol. 2011;43:27.
26. Wright S. Isolation by distance. Genetics. 1943;28:114–38.
27. Caballero A, Toro MA. Analysis of genetic diversity for the management of conserved subdivided populations. Conserv Genet. 2002;3:289–99.
28. Colleau J-J, Palhière I, Rodríguez-Ramilo ST, Legarra A. A fast indirect method to compute functions of genomic relationships concerning genotyped and ungenotyped individuals, for diversity management. Genet Sel Evol. 2017;49:87.
29. Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting FST: the impact of rare variants. Genome Res. 2013;23:1514–21.

Publisher's Note