Comparing estimates of genetic variance across different relationship models

Andres Legarra

To cite this version:

Andres Legarra. Comparing estimates of genetic variance across different relationship models. Theoretical Population Biology, 2016, 107, pp.26-30. 10.1016/j.tpb.2015.08.005 . hal-02637544

Theoretical Population Biology 107 (2016) 26–30

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Comparing estimates of genetic variance across different relationship models

Andres Legarra

INRA, UMR 1388 GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), F-31326 Castanet-Tolosan, France

ARTICLE INFO

Article history:
Received 29 January 2015
Available online 2 September 2015

Keywords:
Relationship
Genetic variance
Heritability
Base population
Mixed models

ABSTRACT

Use of relationships between individuals to estimate genetic variances and heritabilities via mixed models is standard practice in human, plant and livestock genetics. Different models or information for relationships may give different estimates of genetic variances. However, comparing these estimates across different relationship models is not straightforward as the implied base populations differ between relationship models. In this work, I present a method to compare estimates of variance components across different relationship models. I suggest referring genetic variances obtained using different relationship models to the same reference population, usually a set of individuals in the population. Expected genetic variance of this population is the estimated variance component from the mixed model times a statistic, D_k, which is the average self-relationship minus the average (self- and across-) relationship. For most typical models of relationships, D_k is close to 1. However, this is not true for very deep pedigrees, for identity-by-state relationships, or for non-parametric kernels, which tend to overestimate the genetic variance and the heritability. Using mice data, I show that heritabilities from identity-by-state and kernel-based relationships are overestimated. Weighting these estimates by D_k scales them to a base comparable to genomic or pedigree relationships, avoiding wrong comparisons, for instance, "missing heritabilities".

1. Introduction

Recent years have seen an enormous increase in the use of relationship matrices (Wright, 1922) in quantitative genetics (for a general review, see Speed and Balding, 2015) due to their flexibility to accommodate several purposes and also due to the computational efficiency of setting up the associated Mixed Model and the corresponding Mixed Model Equations (e.g. Henderson, 1984). In the following, I will use the term relationship for any measure of scaled covariance between individuals regardless of whether the term has a clear identity-by-descent interpretation or not.

There are several ways of modeling relationships. The first one is the use of expected identical-by-descent (IBD) relationships based on pedigree recordings (Wright, 1922; Emik and Terrill, 1949), which are efficient, reasonably accurate and are widely used in animal genetics. Finite size of the genome (i.e., there are no infinite unlinked loci) causes that true "realized" IBD relationships deviate from expected IBD relationships (Hill and Weir, 2011). Thus, more accurate measures of relationships can be obtained using identity by descent measured withmarkers (Fernando and Grossman, 1989; Almasy and Blangero, 1998; Visscher et al., 2006). Other estimators of relationships based on markers that do not use pedigree are based on identity by state (IBS) at markers, sometimes corrected to be on an IBD scale (Ritland, 1996; Toro et al., 2002; VanRaden, 2008). Finally, non-parametric theory suggests the use of kernel matrices, whose measures of similarity can be interpreted as covariances (Gianola and van Kaam, 2008; de los Campos et al., 2009; Morota and Gianola, 2014). Typical kernel matrices include a smoothing parameter that can regress low relationships towards 0, resulting in more or less local regressions.

A typical reason to use relationship matrices is to estimate so-called genetic parameters, i.e., variance components and in particular genetic variances (e.g., Henderson, 1984; Graser et al., 1987; Yang et al., 2010; Forni et al., 2011 and Rodríguez-Ramilo et al. (2014)). In particular, in human genetics, recent studies compare extensively heritabilities based on markers and based on family studies (i.e., on pedigrees), sometimes giving rise to the so-called missing heritability (e.g., Yang et al., 2010). However, there might be some confusion about how to compare estimates from different models of relationships. For instance, Forni et al. (2011) reported different estimates of heritability using different standardizations of marker-based relationships. Differences in estimates may exist between different relationship matrices (e.g. pedigree, molecular

E-mail address: andres.legarra@toulouse.inra.fr.

http://dx.doi.org/10.1016/j.tpb.2015.08.005
0040-5809/© 2015 Elsevier Inc. All rights reserved.

IBD, molecular (corrected) IBS). These differences may be due to several causes, like different data sets or noise due to uncertainty in the estimation. Estimated variance components may also refer to different genetic bases, for instance, genomic relationship matrices refer to the genotyped population whereas pedigree relationships refer to the founders of the pedigree (VanRaden, 2008; Hayes et al., 2009; Powell et al., 2010). In this work I will analyze the last point, i.e., if the researcher has several available methods for the same data set, can he/she meaningfully compare estimates of heritabilities? Speed and Balding (2015) explicitly say that "$\boldsymbol{K}$ [the relationship matrix should be] standardized to have a mean of zero and a mean diagonal value of one", in order for the estimates of variance components to be meaningful. While this is straightforward to apply numerically, the reason for this recommendation may not be obvious. Further, this condition is difficult to apply for IBD, IBS or kernel matrices, because these are all positive by definition and therefore construction. Imposing this numerical condition may result in non-invertible matrices. Further, this condition is not common practice in animal breeding and would preclude the use of Henderson's (1976, 1977) easy algorithms for inversion and inclusion of pedigree-based relationship matrices in mixed model analysis.

In this note, I will explain the suggestion of Speed and Balding and why it implies the definition of the base population. Further, I will outline how estimates of genetic variances relate to statistics of relationship matrices and how estimates based on different relationship matrices can be compared in a meaningful way taking issues related to centering and scaling into account. This may help practitioners to properly compare genetic variances and heritabilities across studies, perhaps avoiding misinterpretations of "missing heritability".

2. Theory

2.1. Genetic variance of a, possibly related, reference population

The genetic variance is the variance of the genetic values of a set of individuals who constitute the reference, or base, population. I will use the term reference population to avoid ambiguities. The term "genetic value" can be understood as the total genotypic value of an individual (e.g. Henderson, 1985, who called it "total genetic merit"; Gianola and van Kaam, 2008 and Piepho, 2009). More frequently, only the additive part of the genotypic value is considered (Wright, 1922; VanRaden, 2008 and Yang et al., 2010). For a given population, the genetic value is a real (albeit non-observable) quantity, in the sense that an experiment could potentially clone the individuals to obtain total genetic values, or mate them to a large population to obtain breeding values based on differences across offspring. The genetic values of individuals of the reference population can be stacked in a vector, $\boldsymbol{u}$. A model with relationships $\boldsymbol{K}$ across individuals implies that *a priori* these $\boldsymbol{u}$ are drawn from a certain distribution, and the common assumption is $E(\boldsymbol{u}) = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{u}) = \boldsymbol{K}\sigma_u^2$, where σ_u^2 is an associated variance component.

The genetic variance across the individuals in the reference population is simply

$$S_u^2 = \frac{1}{n}\sum (u_i - \bar{u})^2,$$

where $\bar{u}$ is the average genetic value of individuals in the reference population. In matrix notation,

$$S_u^2 = \frac{1}{n}\boldsymbol{u}'\boldsymbol{u} - \left(\frac{1}{n}\boldsymbol{1}'\boldsymbol{u}\right)^2 = \frac{1}{n}\left(\boldsymbol{u}'\boldsymbol{u}\right) - \frac{1}{n^2}\left(\boldsymbol{1}'\boldsymbol{u}'\boldsymbol{u}\boldsymbol{1}\right).$$

The key argument is that, due to the existence of relationships across individuals, the role of $\bar{u}$ cannot be neglected. For instance, if individuals are positively correlated in $\boldsymbol{K}$ (case of IBD relationships), the mean will shift in the same direction as the genetic values. In the most obvious case, if all individuals are identical (all elements in $\boldsymbol{K}$ are equal to 1) then there is no genetic variance in the reference population.

Because we do not know the genetic values of individuals, and these are drawn from a sampling distribution, the statistic S_u^2 has a certain distribution. For a set of random variables $\boldsymbol{x}$ (not necessarily normally distributed) with covariance matrix $\boldsymbol{V}$, it is known that, on expectation, $E\left(\boldsymbol{x}'\boldsymbol{x}\right) = trace(\boldsymbol{V})$ (Searle, 1982, p. 355). Taking expectations of S_u^2 on the distribution of $\boldsymbol{u}$:

$$E\left(S_u^2\right) = \frac{1}{n} trace\left(\boldsymbol{K}\right)\sigma_u^2 - \frac{1}{n^2}\left(\boldsymbol{1}'\boldsymbol{K}\boldsymbol{1}\right)\sigma_u^2$$
$$= \left(\overline{diag\left(\boldsymbol{K}\right)} - \bar{\boldsymbol{K}}\right)\sigma_u^2 = D_k\sigma_u^2 \qquad (1)$$

where $\overline{diag\left(\boldsymbol{K}\right)}$ is the average of the diagonal of $\boldsymbol{K}$ and $\bar{\boldsymbol{K}}$ is the average of $\boldsymbol{K}$, and D_k is the difference between the two. Thus, the genetic variance in the reference population (S_u^2) will be a function of the variance component associated with $\boldsymbol{K}$ in the mixed model, but the genetic variance also depends on the form of the relationship matrices. If D_k is equal to one, then the genetic variance in the reference population (S_u^2) will be equal to the variance component σ_u^2. This conclusion clarifies the statement of Speed and Balding outlined above.

Note that Eq. (1) applies to all or part of the individuals included in the analysis, so there is a need to define the reference population in (1). This choice is often not explicit in the literature. I will show some examples and consequences of Eq. (1).

2.1.1. Reference populations and relationships for which the variance component equals the genetic variance

In Hardy–Weinberg equilibrium, it can be shown that the statistic D_k is equal to 1 for "genomic" relationship matrices of the forms (VanRaden, 2008; Yang et al., 2010):

$$g_{ij} = \frac{\sum\left(x_{ik} - 2p_k\right)\left(x_{jk} - 2p_k\right)}{2\sum p_k\left(1 - p_k\right)}$$

$$g_{ij} = \sum \frac{\left(x_{ik} - 2p_k\right)\left(x_{jk} - 2p_k\right)}{2p_k\left(1 - p_k\right)}$$

where x are genotypes coded numerically across k markers for individuals i and j, and p are allelic frequencies. In this case, the genetic variance of the individuals composing the population, S_u^2, is on expectation equal to the variance component of the mixed model, σ_u^2. Sometimes, Hardy–Weinberg assumptions do not hold (e.g. there is an excess of heterozygosities). One study verified empirically that dividing $\boldsymbol{K}$ by its average diagonal yielded estimates of σ_u^2 similar to pedigree-based estimates (Forni et al., 2011). This is similar to the correction that I suggest later in this work.

Another example where variance component and expected genetic variance agree resides in pedigree relationships, because for the founders of the population it does hold that $D_k = 1$ if the population is large enough (i.e., founders are assumed unrelated and have a self-relationship of 1).

A counter example is a population composed of siblings. For instance, consider n full-sibs, all offspring of two unrelated parents. The pedigree-based relationship matrix $\boldsymbol{A}$ for the full sibs will have 1 in the diagonal ($\overline{diag\left(\boldsymbol{K}\right)} = 1$) and 0.5 off the diagonal. The mixed model would be $\boldsymbol{y} = \boldsymbol{1}\mu + \boldsymbol{u} + \boldsymbol{e}$, with $\mathrm{Var}(\boldsymbol{u}) = \boldsymbol{A}\sigma_u^2$, $\mathrm{Var}(\boldsymbol{e}) = \boldsymbol{I}\sigma_e^2$. Thus,

$$\bar{\boldsymbol{K}} = \frac{1}{n^2}\left(n + n\left(n-1\right)0.5\right) = 0.5 + \frac{1}{n}$$

and $D_k = 0.5$ for n large enough. Therefore, the genetic variance *within* the full-sibs is not σ_u^2 but $\sigma_u^2/2$, and in fact σ_u^2 corresponds to the genetic variance, not of the full-sibs, but of the unrelated population of founders from which parents were drawn, for which it does hold that $D_k = 1$. In fact, it is well described in the animal-genetics literature that, when estimating genetic variances using a pedigree-based relationship matrix, the estimate refers to the unrelated genetic population, not to the whole of the population in the analysis.

2.1.2. Inbreeding and genetic variance

We may call $\bar{F}_k$ the average inbreeding (according to relationships $\mathbf{K}$). By definition

$$\bar{F}_k = \overline{diag\ (\boldsymbol{K})} - 1.$$

Consider IBD relationships, in a population with random matings. On average, inbreeding is equal to half of parents' relationship, and therefore

$$\overline{diag\ (\boldsymbol{K})} = 1 + \bar{F}_k = 1 + \frac{\bar{\boldsymbol{K}}}{2}.$$

Thus, $D_k = 1 - \bar{F}_k$, showing the well-known result of reduction in genetic variance within populations due to relatedness across individuals.

2.1.3. Scaling effects

Consider that instead of dealing with relationships (so that the diagonal terms are similar to 1), the $\mathbf{K}$ matrix is constructed on coancestries (so that they are similar to 0.5), or they are simply real numbers without a clear interpretation. Eq. (1) provides scaling to individual-based genetic variance. For instance, assume that with pedigree relationships $\boldsymbol{A}$, the estimate of the variance is $\hat{\sigma}_u^2$. An estimate using pedigree coancestries, $\boldsymbol{C} = \boldsymbol{A}/2$, will estimate a variance component that will be doubled ($\hat{\sigma}_c^2 = 2\hat{\sigma}_u^2$). To obtain a meaningful estimate we can use (1), which gives the expected result:

$$E\left(S_u^2\right) = \left(\overline{diag\ (\mathbf{C})} - \bar{\mathbf{C}}\right)\hat{\sigma}_c^2 = 0.5 \cdot 2\hat{\sigma}_u^2 = \hat{\sigma}_u^2.$$

2.2. How to meaningfully compare estimates of genetic variances across different models of relationships

For the reference population, the genetic variance is an unobservable quantity, but "real" in some sense. Hence, I propose to refer all estimates of the variance component of the mixed model (σ_u^2) to the reference population. Assume that we have two different relationship matrices $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ (say, model 1 is based on pedigree IBD and model 2 is based on genomic relationships) and associated variance component estimates $\hat{\sigma}_{u,1}^2$, $\hat{\sigma}_{u,2}^2$. For the reference population (not necessarily the whole population), we compute D_{k1} and D_{k2}. The estimates of the genetic variance of the reference population ($S_{u,1}^2$ and $S_{u,2}^2$) will be, in turn:

$$\hat{S}_{u,1}^2 = D_{k1}\hat{\sigma}_{u,1}^2; \qquad \hat{S}_{u,2}^2 = D_{k2}\hat{\sigma}_{u,2}^2.$$

And these estimates are comparable because they refer to strictly the same thing. The choice of the reference population is not always obvious. It is better to consider a large reference population, because the expectation holds better (technically, the sampling variance of S_u^2 is reduced). On the other hand, the expectation operation can be taken again to refer the variance in the reference population to the variance in a "base" population. See the example below for comparison of pedigree versus genomic relationships.

2.2.1. Use of identity by state relationships

Identity-by-state relationships can be defined as probabilities of alikeness in state, i.e. twice the probability that two alleles, one sampled at random for each individual, are alike in state. These IBS relationships are known efficient, but biased, estimators of IBD relationships. For a single locus, (uncorrected)

$$g_{IBS} = g_{IBD} + (2 - g_{IBD})\left(p^2 + q^2\right)$$

(e.g. Ritland, 1996, Eq. (1)) where g_{IBS} (g_{IBD}) is IBS (IBD) relationship and p and q are allelic frequencies of a biallelic marker at the base population from which IBD is defined. Consider a population of unrelated individuals in the IBD sense. In terms of IBS relationships, the diagonal terms will average to $1 + p^2 + q^2$ using the expression above, as $g_{IBD} = 1$ (self relationships for all individuals are all identical to 1). The off-diagonal elements of the IBS relationships will average to $2\left(p^2 + q^2\right)$, as $g_{IBD} = 0$ for across-individual relationships. Thus, in this case, if the population is large enough,

$$\bar{\boldsymbol{K}} = 2\left(p^2 + q^2\right)$$

$$\overline{diag\ (\boldsymbol{K})} = 1 + p^2 + q^2$$

and

$$D_{k,IBS} = 1 - p^2 - q^2$$

whereas

$$D_{k,IBD} = 1.$$

Now expected genetic variances in both cases can be compared. Because $E\left(S_{u,IBS}^2\right) = D_{k,IBS}\hat{\sigma}_{u,IBS}^2$ and $E\left(S_{u,IBD}^2\right) = D_{k,IBD}\hat{\sigma}_{u,IBD}^2$, we expect $E\left(S_{u,IBS}^2\right) = E\left(S_{u,IBD}^2\right)$ and therefore, on expectation,

$$\hat{\sigma}_{u,IBS}^2 = \hat{\sigma}_{u,IBD}^2/(1 - p^2 - q^2).$$

Using the variance component $\hat{\sigma}_{u,IBS}^2$ to estimate the heritability will bias it upwards.

2.2.2. Comparing genomic and IBD estimates

When estimating genomic relationship matrix $\boldsymbol{G}$ (VanRaden, 2008; Yang et al., 2010 and Speed and Balding, 2015), the reference population is most typically equal to the genotyped population. If Hardy–Weinberg holds, then $D_G = 1$, and the estimated variance component can be interpreted as the genetic variance in the population. If this population has a pedigree, a pedigree-based relationship matrix $\boldsymbol{A}$ can be constructed, with $D_A = 1 - \bar{F}_A$ (if the reference population is sufficiently large and matings are at random). The estimated variance component refers to the founders of the pedigree, for whom $D_A = 1$. Two estimates of variance component are obtained using $\boldsymbol{G}$ or $\boldsymbol{A}$, $\hat{\sigma}_{u,G}^2$ and $\hat{\sigma}_{u,A}^2$ respectively. In order to compare them, however, we need to refer them to the same reference population. This is not the case generally. There are two options to establish a common reference population.

The first is to refer to the reference population of *genotyped individuals* and thus $\hat{S}_{u,G}^2 = \hat{\sigma}_{u,G}^2$ and $\hat{S}_{u,A}^2 = \left(1 - \bar{F}_A\right)\bar{\sigma}_{u,A}^2$, and the genetic variance of the whole genotyped population is reduced with respect to the genetic variance of the founders of the pedigree.

Another option is to refer to the reference population of *founders of the pedigree*, although computing D_G is often not possible because biological samples for the base population are often unavailable. However, D_G must behave in the same direction as D_A (it reduces with generations due to inbreeding) and therefore we may correct in the opposite sense: $\hat{S}_{u,G}^2 = \hat{\sigma}_{u,G}^2(1 - \bar{F}_A)$ which means that genetic variance should be larger in the founders of the population than in the genotyped population, because in the latter individuals are related and this reduces variance. On the other hand, for this reference population, $\hat{S}_{u,A}^2 = \hat{\sigma}_{u,A}^2$.

3. Real data example

3.1. Material and methods

I will illustrate the ideas above with a set of mouse data frequently used to test genomic prediction procedures (Valdar et al., 2006; Legarra et al., 2008). The data set includes 1884 animals

Table 1

Estimates of variance components ($\hat{\sigma}_u^2$), apparent heritabilities ($\hat{h}^2_{apparent}$), statistics of the relationship matrices (average of the diagonal $\overline{diag\,(\boldsymbol{K})}$, average relationship $\bar{\boldsymbol{K}}$, $D_k = \overline{diag\,(\boldsymbol{K})} - \bar{\boldsymbol{K}}$), genetic variances ($\hat{S}_u^2$), corrected heritabilities ($\hat{h}^2$ (corrected)), and minus twice the log-likelihood (−2logL) in a mice data set for body length, for four different modelings of the relationships (Pedigree, Genomic, Kernel and Identity by State (IBS)). The standard error of the variance components is 0.008 and of the heritabilities, 0.03.

	Pedigree	Genomic	Kernel	IBS
$\hat{\sigma}_u^2$	0.038	0.033	0.094	0.102
$\hat{h}^2_{apparent}$	0.16	0.14	0.33	0.34
$\overline{diag\,(\boldsymbol{K})}$	1	1.031	1	1.665
$\bar{\boldsymbol{K}}$	0.0045	−0.0005	0.4963	1.3090
$D_k = \overline{diag\,(\boldsymbol{K})} - \bar{\boldsymbol{K}}$	0.995	1.0005	0.5037	0.3562
$\hat{S}_u^2 = D_k\hat{\sigma}_u^2$	0.038	0.033	0.047	0.036
$\hat{h}^2$ (corrected)	0.16	0.14	0.20	0.15
−2logL	2440.67	2411.73	2410.37	2412.19

genotyped and phenotyped and roughly 10,000 markers. Pedigree includes 2272 individuals. I considered four different estimates of relationships, with the following abbreviation: (pedigree) pedigree relationships; (genomic) genomic relationships using markers, of the form

$$g_{ij} = \sum \frac{(x_{ik} - 2p_k)\left(x_{jk} - 2p_k\right)}{2p_k\,(1 - p_k)}$$

(VanRaden, 2008), using observed allelic frequencies in the population; (kernel) a Gaussian kernel matrix based on Euclidean distances (e.g. Endelman, 2011), with the form

$$g_{ij} = \exp\left(-\left(\frac{d_{ij}}{\theta}\right)^2\right)$$

where d_{ij} is a normalized distance between genotypes of individuals i and j, and θ is a smoothing parameter that was fixed at 0.5 (typical values of this parameter oscillate between 0 and 1, Endelman, 2011); (IBS) is a matrix of IBS relationships constructed using identity by state similarities (i.e., they share none, one, or two alleles), that can be expressed as follows:

$$g_{ij} = \frac{1}{n}\sum\left(x_{ik}x_{jk} - x_{ik} - x_{jk} + 2\right).$$

These IBS relationship were *not* corrected for heterozygosity at the markers, so that the bias in the estimation of genetic parameters was more apparent. The trait analyzed was body length, and the model included sex, random cage effect, and individual genetic effect as in Legarra et al. (2008). Variance components were estimated by REML using remlf90 (Misztal et al., 2002) and different matrices were constructed using preGSf90 (Aguilar et al., 2014) except for IBS relationships, which were programmed.

3.2. Results

Table 1 shows the different variance component estimates ($\hat{\sigma}_u^2$), the "apparent" heritability using this variance component ($\hat{h}^2_{apparent}$) the relevant statistics of the different relationship matrices across the genotyped and phenotyped individuals, the genetic variance taking as reference the whole genotyped population, the heritability estimated using this estimate, and minus twice the log-likelihood (lower value is better). Variance component estimates for the other effects were similar across analysis, roughly 0.050 for the "cage" effect and 0.150 for the residual. Firstly, it can be observed that all models using genomic data perform better in terms of likelihood. Secondly, and more importantly, the transformation that I propose puts estimates of genetic variances and heritabilities on a similar scale across different models.

Therefore, the much higher *apparent* heritability of models "IBS" and "Kernel" is an artifact of the form of their respective relationship matrices, and it does not imply that they are more likely or they explain better the genetic architecture (as can be seen in the log-likelihood, where they are very similar to "genomic").

Statistics of relationships D_k and "corrected" estimates in Table 1 take the 1884 animals genotyped and phenotyped as the reference. When using pedigree-based relationships (as in livestock genetics), it is customary to consider the pedigree founders as the reference population. In order to do so, estimates of the genetic variance across the pedigree founders can be obtained dividing $\hat{S}_u^2$ by the value of D_k for "pedigree".

4. Discussion

I have presented a method to obtain meaningful, and comparable, estimates of genetic variances from estimates of variance components across different structures of relationships. I stress the difference between genetic variance and variance component. The first is a biological property of the population that can, at least conceptually, be estimated with precision by an experiment. The latter is a scaling constant associated with a certain assumed structure of relationships across individuals, but, depending how these relationships are conceived, this constant may or may not be interpreted as the genetic variance of the particular population being analyzed. Here, I have presented a comprehensive theory that, firstly, defines genetic variance as associated with a set of individuals in a reference population and secondly, derives proper scaling of variance component estimates towards genetic variances. The bias that I show in this paper is of a different kind than sampling error of estimates (sampling error vanishes with large data sets), that is, even for very large data sets estimates using different relationships will differ systematically. For instance IBS relationships will estimate variance components higher than IBD relationships by a scale factor of the order $1/(1 - p^2 - q^2)$, no matter how large the data set.

The work that I present is closely related to previous attempts to reconcile genomic and pedigree relationships (Powell et al., 2010; Vitezica et al., 2011; Meuwissen et al., 2011; Christensen, 2012). These authors suggested making genomic relationship matrices comparable to pedigree-based relationship matrices by scaling and adding constants (roughly $\overline{diag\,(\boldsymbol{K})}$ and $\bar{\boldsymbol{K}}$) albeit they did not explicitly address the "comparability" of variance component estimates across different models of relationships. Implicitly, all these works, and also the present one, draw on the fixation index theory of Wright (e.g. Powell et al., 2010).

There are very few comparisons of heritability estimates *within* the same data set *across* different kind of relationships, and these have been undertaken (to my knowledge) mostly in animal genetics. Legarra et al. (2011), Forni et al. (2011) and Rodríguez-Ramilo et al. (2014) reported similar estimates across pedigree and genomic relationships. However, estimates of variance components with kernel matrices have rarely been compared with regular estimates, on the grounds that "the interpretation of this parameter is not obvious" (González-Recio et al., 2008). This is unfortunate because kernel matrices are more flexible than genomic or pedigree relationships and have the potential to smooth relationships across very distant individuals (Endelman, 2011), where we know that pedigree-based estimates of relationships, which assume infinite unlinked loci, are not reliable (Hill and Weir, 2011). In this work I show that estimates are indeed comparable, and hopefully this will help practitioners to compare estimates of genetic variances across different models.

5. Conclusion

In this work I have presented a theory for a meaningful comparison of heritability and genetic variance estimates across different models for relationships between individuals. The process involves: (a) choosing a reference population common to all models for relationships; then, for each possible relationship: (b) computing statistics D_k of relationships within the reference population, (c) estimating the genetic variance at the reference population as $\hat{S}_u^2 = D_k \hat{\sigma}_u^2$, where $\hat{\sigma}_u^2$ is the estimated variance component, (d) estimating heritabilities using $\hat{S}_u^2$.

Acknowledgments

I thank James Cussens and Nuala Sheehan for inviting me to the workshop "Statistical and computational methods for relatedness and relationship inference from genetic marker data", where I learnt much, and for encouraging me to present this manuscript. Many ideas in this manuscript are fruit of discussions with Miguel Angel Toro, Zulma Vitezica, Ole F Christensen, Jean-Michel Elsen and Anne Ricard. I acknowledge financing from INRA metaprogram SelGen in projects X-Gen and SelDir. I thank the genotoul bioinformatics platform Toulouse Midi-Pyrenees for providing computing resources. The manuscript was greatly improved by suggestions of two reviewers and the editor, and reading and corrections from Susanne Hermesch (AGBU, Australia). I want to thank Francisco Javier Mendizábal Múgica who first explained to me matrix algebra and relationships, some time ago.

References

Aguilar, I., Misztal, I., Tsuruta, S., Legarra, A., Wang, H., 2014. PREGSF90–POSTGSF90: Computational tools for the implementation of single-step genomic selection and Genome-wide association with ungenotyped individuals in BLUPF90 programs. In: 10th World Congress on Genetics Applied to Livestock Production. Asas.

Almasy, L., Blangero, J., 1998. Multipoint quantitative-trait linkage analysis in general pedigrees. Am. J. Hum. Genet. 62, 1198–1211.

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

de los Campos, G., Gianola, D., Rosa, G.J., 2009. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. J. Anim. Sci. 87, 1883–1887.

Emik, L.O., Terrill, C.E., 1949. Systematic procedures for calculating inbreeding coefficients. J. Hered. 40, 51–55.

Endelman, J.B., 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4, 250–255.

Fernando, R., Grossman, M., 1989. Marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol. 21, 467.

Forni, S., Aguilar, I., Misztal, I., 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet. Sel. Evol. 43, 1.

Gianola, D., van Kaam, J.B.C.H.M., 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178, 2289–2303.

González-Recio, O., Gianola, D., Long, N., Weigel, K.A., Rosa, G.J.M., et al., 2008. Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. Genetics 178, 2305–2313.

Graser, H.-U., Smith, S., Tier, B., 1987. A derivative-free approach for estimating variance components in animal models by restricted maximum likelihood. J. Anim. Sci. 64, 1362–1370.

Hayes, B.J., Visscher, P.M., Goddard, M.E., 2009. Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. 91, 47–60.

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, C.R., 1977. Best linear unbiased prediction of breeding values not in the model for records. J. Dairy Sci. 60, 783–787.

Henderson, C.R., 1984. Applications of Linear Models in Animal Breeding. University of Guelph, Guelph.

Henderson, C.R., 1985. Best linear unbiased prediction of nonadditive genetic merits. J. Anim. Sci. 60, 111–117.

Hill, W.G., Weir, B.S., 2011. Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet. Res. (Camb.) 1–18.

Legarra, A., Robert-Granié, C., Croiseau, P., Guillaume, F., Fritz, S., 2011. Improved Lasso for genomic selection. Genet. Res. (Camb.) 93, 77–87.

Legarra, A., Robert-Granié, C., Manfredi, E., Elsen, J.-M., 2008. Performance of genomic selection in mice. Genetics 180, 611–618.

Meuwissen, T., Luan, T., Woolliams, J., 2011. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. J. Anim. Breed. Genet. 128, 429–439.

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., et al. 2002. BLUPF90 and related programs (BGF90). In: 7th World Congress on Genetics Applied to Livestock Production. CD-ROM Communication No. 28-07.

Morota, G., Gianola, D., 2014. Kernel-based whole-genome prediction of complex traits: a review. Front. Genet. 5.

Piepho, H.P., 2009. Ridge regression and extensions for genomewide selection in maize. Crop Sci. 49, 1165–1176.

Powell, J.E., Visscher, P.M., Goddard, M.E., 2010. Reconciling the analysis of IBD and IBS in complex trait studies. Nat. Rev. Genet. 11, 800–805.

Ritland, K., 1996. Estimators for pairwise relatedness and individual inbreeding coefficients. Genet. Res. 67, 175–185.

Rodríguez-Ramilo, S.T., García-Cortés, L.A., González-Recio, Ó., 2014. Combining genomic and genealogical information in a reproducing kernel Hilbert spaces regression model for genome-enabled predictions in dairy cattle. PLoS One 9, e93424.

Searle, S.R., 1982. Matrix Algebra Useful for Statistics. John Wiley.

Speed, D., Balding, D.J., 2015. Relatedness in the post-genomic era: is it still useful? Nature Rev. Genet. 16, 33–44.

Toro, M., Barragan, C., Ovilo, C., Rodriganez, J., Rodriguez, C., et al., 2002. Estimation of coancestry in Iberian pigs using molecular markers. Conserv. Genet. 3, 309–320.

Valdar, W., Solberg, L.C., Gauguier, D., Burnett, S., Klenerman, P., et al., 2006. Genome-wide genetic association of complex traits in heterogeneous stock mice. Nat. Genet. 38, 879–887.

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91, 4414–4423.

Visscher, P.M., Medland, S.E., Ferreira, M.A.R., Morley, K.I., Zhu, G., et al., 2006. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet. 2, e41.

Vitezica, Z., Aguilar, I., Misztal, I., Legarra, A., 2011. Bias in genomic predictions for populations under selection. Genet. Res. 93, 357–366.

Wright, S., 1922. Coefficients of inbreeding and relationship. Am. Nat. 56, 330–338.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., et al., 2010. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42, 565–569.