Journal of
Animal Breeding and Genetics

J. Anim. Breed. Genet. ISSN 0931-2668

ORIGINAL ARTICLE

A comparison of methods to estimate genomic relationships using pedigree and markers in livestock populations

N.S. Forneris[1], J.P. Steibel[2], A. Legarra[3], Z.G. Vitezica[3,4], R.O. Bates[2], C.W. Ernst[2], A.L. Basso[5] & R.J.C. Cantet[1,6]

1 Departamento de Producción Animal, Facultad de Agronomía, Universidad de Buenos Aires, Ciudad Autónoma de Buenos Aires, Argentina
2 Department of Animal Science, Michigan State University, East Lansing, MI, USA
3 INRA, GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), Castanet-Tolosan, France
4 INP, ENSAT, GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), Université de Toulouse, Castanet-Tolosan, France
5 Departamento de Biología Aplicada y Alimentos, Facultad de Agronomía, Universidad de Buenos Aires, Ciudad Autónoma de Buenos Aires, Argentina
6 INPA-CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas), Buenos Aires, Argentina

Keywords
Accuracy; genomic prediction; genome sharing; identity by descent; SNP.

Correspondence
N.S. Forneris, Departamento de Producción Animal, Facultad de Agronomía, Universidad de Buenos Aires, 1417, Ciudad Autónoma de Buenos Aires, Argentina.
Tel: 54-11-4524-8000 extension 8184;
Fax: 54-11-4524-8735;
E-mail: forneris@agro.uba.ar

Received: 11 September 2015;
accepted: 30 March 2016

Summary

Accurate prediction of breeding values depends on capturing the variability in genome sharing of relatives with the same pedigree relationship. Here, we compare two approaches to set up genomic relationship matrices for precision of genomic relationships (GR) and accuracy of estimated breeding values (GEBV). Real and simulated data (pigs, 60k SNP) were analysed, and GR were estimated using two approaches: (i) identity by state, corrected with either the observed ($\boldsymbol{G}_{\mathbf{VR\text{-}O}}$) or the base population ($\boldsymbol{G}_{\mathbf{VR\text{-}B}}$) allele frequencies and (ii) identity by descent using linkage analysis ($\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$). Estimators were evaluated for precision and empirical bias with respect to true pedigree IBD GR. All three estimators had very low bias. $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$ displayed the lowest sampling error and the highest correlation with true genome-shared values. $\boldsymbol{G}_{\mathbf{VR\text{-}B}}$ approximated $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$'s correlation and had lower error than $\boldsymbol{G}_{\mathbf{VR\text{-}O}}$. Accuracy of GEBV for selection candidates was significantly higher when $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$ was used and identical between $\boldsymbol{G}_{\mathbf{VR\text{-}O}}$ and $\boldsymbol{G}_{\mathbf{VR\text{-}B}}$. In real data, $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$'s sampling standard deviation was the closest to the theoretical value for each pedigree relationship. Use of pedigree to calculate GR improved the precision of estimates and the accuracy of GEBV.

Introduction

In traditional pedigree-based evaluation, the numerator or additive relationship matrix (Henderson 1976), which is equal to twice the matrix of pairwise kinship (or coancestry) coefficients, has been widely used to estimate genetic covariances and breeding value of individuals. Additive relationships carry information on genetic resemblance from common inheritance and are based on probabilities that gene pairs are identical by descent or IBD (Wright 1922). In animal breeding, and throughout this paper, it is assumed that there is an accepted founder population relative to which IBD is to be measured, consisting of the founder members of a defined pedigree, with the implication that more remote coancestry of current gametes is ignored (Thompsom 2013). Therefore, individuals whose genes are copies from an ancestral one in the base population are likely to share on average the same causal loci, so that phenotypic data from related individuals are informative for the prediction of the breeding value of either animal. Inbreeding and kinship coefficients, and more generally probabilities of any IBD state, are expectations of random variables that indicate IBD at a given point in the genome (Thompsom 2013). In the absence of inbreeding,

  doi:10.1111/jbg.12217

additive relationships (Wright 1922) represent the expected proportion of genome-shared IBD.

Finite size of the genome and recombination introduce randomness and variation on the amount of genome-shared IBD for any particular type of relatives (Risch and Lange, 1979; Guo, 1996; Hill & Weir 2011), which makes actual relationships to differ from their expected value. The availability of dense panels of SNP markers in livestock species allows estimating these actual relationships using marker data. The genomic relationship matrix (***G***) calculated with markers has a paramount role in the prediction of breeding values from animal models, when using best linear unbiased predictors. Elements of ***G*** are estimates of the actual proportion of the genome that two individuals share (*realized* relationships), whereas the pedigree-based relationship matrix is the expectation of this proportion (*expected* relationships) (Goddard *et al.* 2011). The use of realized relationships is responsible for the gain in accuracy while predicting breeding values in genomic selection schemes. This gain in accuracy can be shown to be due to the reduction in the variance of Mendelian residuals of the genomic breeding values (Cantet & Vitezica 2014). The efficiency of the BLUP (accuracy) depends on how well marker-derived genomic relationships capture the patterns of realized genetic relationships at causal loci (VanRaden 2007, 2008; De los Campos *et al.* 2013).

VanRaden (2007, 2008) proposed a calculus of genomic relationships by adding cross-products of marker data deviated from mean gene frequencies and divided by the total heterozigosity at the markers. These relationships reflect the actual proportion of marker alleles shared by identity by state (IBS), as a deviation from the expected proportion of alleles shared in the population (Vela-Avitúa *et al.* 2015). As a result, likeness among alleles at all markers constitutes the information on which genetic resemblance among animals is carried to ***G***. An alternative way of using marker information to estimate realized relationships is to trace IBD inheritance of haplotypes within the known pedigree (Thompson 2013). The efficiency of either method depends on how well they can capture the signals from the true IBD process in the genome continuum, which in turn is affected by linkage disequilibrium, incomplete pedigree information and inbreeding. VanRaden's estimates of genomic relationships require accurate estimates of the true allele frequencies of the unselected base population, which can be difficult to obtain. Simple frequency estimates obtained as means of only the subset of known genotypes either from the current or from the base population (founders), or even base frequency estimates using the algorithm of Gengler *et al.* (2007), can lead to biased relationship coefficients. If base allele frequencies are unknown, incorporating pedigree information into these calculations could be a strategy when dealing with large families with a small number of genotyped animals.

The purpose of this research was to compare two approaches to estimate the true pairwise-realized relationships between genotyped animals, in terms of the precision of the relationships, by analysing real and simulated data. We define the true realized relationship as the proportion of total genome that two individuals share IBD relative to the specified founders of a pedigree. The first one is the IBS-derived approach that is widely employed in genomic BLUP (GBLUP) methods (VanRaden 2008) and uses only markers to infer genome sharing across individuals. The second approach (IBD) infers relationships tracing transmission of markers throughout the pedigree (linkage analysis) even if there are many ungenotyped family members, while accounting for population linkage disequilibrium or background sharing beyond the pedigree. We further illustrate the consequences of using either approach on accuracy of genomic estimated breeding values (GEBV).

Materials and methods

Two approaches to estimate genomic relationships were evaluated using both simulated and real pig data. To ascertain the precision of these estimates, the true relationships – or realized proportion of genome shared by relatives of a given degree – need to be known. These are available only for simulated data, yet unknown with real data (it is impossible to know without error which of the alleles from the founder allele set an individual has inherited at every genome location). Still, for real data, we can compare the mean and variance of the true relationships, which can be calculated using theoretical formulae (Hill & Weir 2011) that depend only on map length and on the pedigree relationship between the individuals, with the corresponding estimated mean and variance. Thus, we used an existing real pig data set from an F_2 cross, in which pedigree relationships were precisely defined and had many pairs of individuals within each type of pedigree relationship. The simulated data are a more conventional population.

Simulated data set

Data were simulated using QMSim (Sargolzaei & Schenkel 2009), by considering a simplified scenario

for the breeding programme of a pig nucleus. The simulated genome consisted of 5 autosomal chromosomes of 160 cM each. Bi-allelic markers (35 000) were distributed randomly across the genome, with equal allele frequency in the first historical generation. A mutation rate of 2×10^{-4} per locus per generation was applied, assuming a recurrent model. The historical population was simulated by considering an equal number of males and females, discrete generations, random mating, no selection and no migration. Offspring were produced by the union of gametes randomly sampled from the male and female gametic pools. Recombination was modelled at a rate of 1 cM/Mb assuming a Poisson distribution. After 2500 generations with a constant size of 500, followed by a severe bottleneck during 30 generations with a constant size of 75, a historical population at mutation-drift equilibrium that produced realistic level of linkage disequilibrium was established. Sex ratio was constant across historical generations, except for the last generation, in which 20 males and 200 females were generated by random choice of two gametes from the male and female gametic pools. These animals constituted the founders for the recent population (G_0). Among the marker loci with MAF >0.01 in G_0, 16 000 SNPs (spaced on average every 0.05 cM) were randomly chosen. A polygenic trait with heritability (h^2) of 0.25 and phenotypic variance of 1 was simulated by assigning to each founder an additive effect sampled from a normal distribution with mean 0 and variance 0.25. Then, the following selection scheme was followed for five generations. In each generation, 20 boars were mated with 200 sows to produce 2000 offspring (half of them males). Mating design was optimized to minimize inbreeding (Sonesson & Meuwissen 2000) using the 'minf' option in QMSim. For the next generation, the 20 boars with the highest estimated BV were selected based on best linear unbiased prediction (BLUP) via an animal model, whereas 200 sows were randomly selected. Pedigree was available for all 5 generations (10 220 animals). For estimation purposes, it was assumed that 140 animals (i.e. G_0 boars, the 20 selected boars from generations 1 to 4, and 40 boars randomly chosen from the selection candidates from generation 5) were genotyped. The rest of animals in the pedigree were assumed non-genotyped. The whole simulation process was replicated 50 times.

Real data set

Pedigree and genotypic data used in our analyses were collected on 411 animals from an outbred resource pig population Duroc × Pietrain elapsing three generations (F_0, F_1 and F_2) that was raised at Michigan State University Swine Teaching and Research Farm (Edwards *et al.* 2008). Animal protocols were approved by the Michigan State University All-University Committee on Animal Use and Care. The population was established from 4 F_0 Duroc sires and 15 F_0 Pietrain dams. From the F_1 progeny, 50 females and 6 males were selected as parents of the F_2 generation while avoiding full- or half-sib matings. A total of 1259 F_2 pigs were born alive in 141 litters across 11 farrowing groups. All animals were produced through the artificial insemination. From the F_2 animals, 336 pigs were selected for genotyping to represent all full-sib families (Gualdrón Duarte *et al.* 2013). A total of 411 pigs (19 F_0, 56 F_1 and 336 F_2) out of 1334 were genotyped with the Illumina PorcineSNP60 chip (Ramos *et al.* 2009). Genotyping was performed at a commercial laboratory (GeneSeek, a Neogen Company, Lincoln, NE, USA). Of 62 163 SNPs, 38 263 were employed for all analyses after quality-control procedures, which involved removing non-autosomal SNPs (15 298), SNPs with MAF <0.01, call rate <90% or Mendelian inconsistencies >2%.

Computation of Pairwise identical-by-descent (IBD) genome sharing in simulated data

Following Hill & Weir (2011)'s notation, let $\breve{R}_{ij}$ be the 'true' realized relationship or proportion of the total genome individuals i and j share IBD, with respect to the specified founders of a pedigree (i.e. starting from G_0 in the simulated data and from F_0 in the real data). We will call $\breve{R}_{ij}$ shortly hereafter the pedigree IBD genome sharing. Assume initially that, at any genome location, it can be determined which of the $2n$ alleles from the founder set an individual inherited (this is not possible with real data). Furthermore, let $S_l (l = 1, \ldots, 9)$ be an indicator variable for the event of observing the condensed identity state l (Jacquard 1974). Thus, S_l is equal to 0 or 1, depending on the observed IBD pattern among the four alleles present in two individuals. Then, the realized coancestry or kinship coefficient between a pair of individuals i and j at location t is $\breve{\theta}_{ij}(t) = S_1 + \frac{1}{2}(S_3 + S_5 + S_7) + \frac{1}{4}S_8$. This directly provides that the realized additive relationship coefficient at location t is $\breve{\theta}_{ij}(t)$ and that the actual relationship $\breve{R}_{ij}$, considering a genome of length L, is as follows (Guo 1995)

$$\breve{R}_{ij} = \frac{1}{L}\int_0^L 2\breve{\theta}_{ij}(t)\, dt \qquad (1)$$

For simulated data, we used equation (1) to compute $\breve{R}_{ij}$ for each pair of animals and also computed

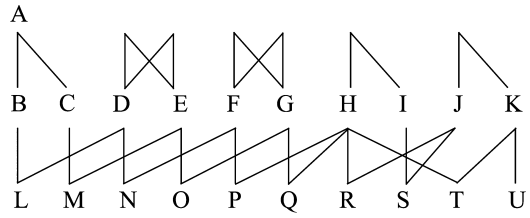
the overall mean and variance of $\breve{R}_{ij}$ across the whole set of analysed pairs (10 220) for each replicate. For the real data set, we cannot compute the value of $\breve{R}_{ij}$ as we cannot observe $\breve{\theta}_{ij}(t)$, but we can compute theoretically its mean and variance. All pairs of animals in the real data set (1334 animals) were classified into 14 different pedigree relationships (e.g. half-sibs, full-sibs, see Fig. 1). For each relationship, the mean of $\breve{R}_{ij}$, $E(\breve{R}_{ij_k})$ ($k = 1, \ldots, 14$), was obtained from the pedigree, and the variance, $\mathrm{Var}(\breve{R}_{ij_k})$, was computed using the theoretical formulae derived by Hill & Weir (2011) (formulae are for non-inbred individuals, as is the case for the real data set), which depends only on the number of chromosomes and their map length. Sex-averaged map length (cM) was taken from recombination rates reported by Tortereau *et al.* (2012). The overall mean and variance of $\breve{R}_{ij}$ across the whole real set of analysed pairs can be derived from the theory of finite mixture distributions (Frühwirth-Schnatter 2006). Let $\breve{R}_{ij}$ denotes the IBD genome sharing for a pair of animals from a mixture distribution whose probability density function is as follows:

$$f(\breve{R}_{ij}) = \sum_{k=1}^{K} \eta_k \, p(\breve{R}_{ij_k}) \tag{2}$$

In (2), $p(\breve{R}_{ij_k})$ denotes the conditional probability density function of $\breve{R}_{ij}$ given relationship class k ($k = 1, \ldots, 14$) (Fig. 1), and η_k is the mixture coefficient for class k such that $\sum_{k=1}^{K} \eta_k = 1$. Then,

$$E(\breve{R}_{ij}) = \sum_{k=1}^{K} \eta_k E(\breve{R}_{ij_k}) \tag{3}$$

$$\mathrm{Var}(\breve{R}_{ij}) = \sum_{k=1}^{K} \eta_k \left[E(\breve{R}_{ij_k}) + \mathrm{Var}(\breve{R}_{ij_k})\right] \left[E(\breve{R}_{ij})\right]^2 \tag{4}$$



Relationship	Example	Relationship	Example
Lineal relatives		*Bilineal relatives*	
Parent-offspring	AB	Full sibs	DE
Grandparent-grandoffspring	AL	Double half cousins	ST
Half sibs	HI	Triple half cousins	LM
Half uncle-nephew	JU	Double first cousins	NO
Half cousins	SU	Half sibs, mothers (fathers) half sibs	RS
Uncle-nephew	DO	Three quarter sibs	PQ
Cousins	NQ	Unrelated	BD

Figure 1 Examples of relationships on the real dataset.

Estimated IBD genome sharing between genotyped animals

Two approaches to estimate pairwise relationships based on markers, using or not pedigree information, were compared. These estimates will constitute the elements of the genomic relationship matrix for genotyped animals, $\boldsymbol{G}$, of order 140 (411) for the simulated (real) data set.

The first one is an identity-by-state (IBS)-based approach, which is widely employed in genomic BLUP (GBLUP) methods (VanRaden 2008), and uses centred genotypes to measure the number of alleles shared between individuals, sums over SNPs and divides by the total heterozigosity at the markers. Thus, the following estimated pairwise relationship between animals i and j, $G_{\mathrm{VR}_{ij}}$ is equal to the following:

$$G_{\mathrm{VR}_{ij}} = \frac{\sum_{m=1}^{M} (x_{im} - \mu_m)(x_{jm} - \mu_m)}{2 \sum_{m=1}^{M} p_m(1 - p_m)}, \tag{5}$$

where x_m is coded as -1, 0 and 1 for homozygote, heterozygote and other homozygotes, respectively, $\mu_m = 2(p_m - 0.5)$ is the population mean of the genotypic values, and p_m is the population frequency of the second allele at locus m. Relationships were calculated with PREGSF90 (Aguilar *et al.* 2011) using either the observed allele frequency of each SNP ($\boldsymbol{G}_{\mathbf{VR\text{-}O}}$) or the frequencies from all base population animals ($\boldsymbol{G}_{\mathbf{VR\text{-}B}}$). The observed allele frequencies refer to all the genotyped animals (140 in simulated data and 411 in real data). The base allele frequencies refer to the 20 G_0 boars and the 200 G_0 sows in the simulated data and to the 4 F_0 boars and 15 F_0 dams in real data. In either case, for each SNP, we counted the number of '2' alleles across individuals and divided by the total number of alleles (two times the number of individuals used for the computation). To avoid singularity issues, matrix $\boldsymbol{G}_{\mathbf{VR}}$ was calculated as $\boldsymbol{G} = w\,\boldsymbol{G}^* + (1 - w)\,\boldsymbol{A}_{22}$, where $w = 0.95$, $\boldsymbol{G}^*$ is the genomic matrix before weighting and $\boldsymbol{A}_{22}$ is the matrix of relationships across genotyped animals, that is a submatrix (of dimension 140 and 411 for simulated and real

data, respectively) of the whole pedigree-based relationship matrix $\boldsymbol{A}$ (of dimension 10 220 and 411, respectively). In real data, the 411 genotyped individuals constituted the pedigree, as non-genotyped individuals were F_2 individuals with no descendants. Matrix $\boldsymbol{G}_{\mathbf{VR\text{-}O}}$ was also scaled based on $\boldsymbol{A}_{22}$ to control bias as $\boldsymbol{G}^{*} = \boldsymbol{1}\boldsymbol{1}'\alpha + \beta\boldsymbol{G}$, where parameters α and β are estimated by equating means of diagonal elements and all elements in the two matrices ($\boldsymbol{A}_{22}$ and $\boldsymbol{G}$) (Vitezica *et al.* 2011).

The second approach infers relationships tracing transmission of markers throughout the known pedigree (linkage analysis, $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$). We used the hidden Markov model (HMM) proposed by Li *et al.* (2010). For this, the forward–backward algorithm implemented in the software PEDIBD (Li *et al.* 2010) was used. This algorithm can deal with a pedigree composed of individuals with and without genotypes, as is the case here. For any given pair of genotyped individuals, the hidden state (q_m) of the HMM is the number (0, 1 or 2) of pairs of IBD alleles at the SNP position m. The observable state, o_m, is the number of pairs of alleles that are IBS at the same position. First, the HMM is built for a pair of alleles with three possible hidden states: (i) non-IBD, (ii) IBD within the known pedigree and (iii) background IBD to fit the hidden relatedness beyond the relatedness that is observed through the available pedigree structure. Separating this background IBD from the IBD within the pedigree prevents biased inference of true IBD status, as we aim at estimating IBD from the founders of the pedigree but not further back in time. Transition probabilities between states do not only depend on the marker interval, but also on all possible inheritance paths within the pedigree linking two marker alleles. Based on this basic model, the HMM for a pair of individuals is built by assuming independence between two homologous chromosomes within an individual, which is an approximation in a pedigree with loops. Thus, for two individuals i and j, the estimated genome sharing ($G_{\mathrm{IBD-L}_{ij}}$) can be calculated as

$$G_{\mathrm{IBD-L}_{ij}} = \sum_{m=1}^{M} w_m \left[\frac{1}{2} P\left(q_m = 1 | o_1, \ldots, o_M\right) + P\left(q_m = 2 | o_1, \ldots, o_M\right) \right] \quad (6)$$

where w_m is the weight of the mth SNP and $P(q_m = 1|\, o_1,\ldots,o_M)$ ($P(q_m = 2|\, o_1,\ldots,o_M)$ is the posterior probability of sharing 1(2) pair(pairs) of alleles IBD at position m, conditional on the information of all marker loci. Each weight w_m was calculated as the mth SNP's coverage related to the physical length of the genome. Our approach differs from that of Fernando and Grossman (1989) essentially in that IBD probabilities at each SNP are estimated conditionally not only on the marker genotype of that locus but on the whole sequence of observable genotypes throughout the genome.

Matrix $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$ may be indefinite showing (small) negative eigenvalues. The reason for this is that elements of $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$ (the genomic relationships) are computed on a pairwise basis instead of globally. Thus, the 'nearPD' function in the *R* package 'Matrix' was used to compute the nearest positive definite matrix to the original $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$ (Cheng & Higham 1998; Higham 2002). These estimates were retained for the statistical analysis.

Statistical analysis

For the real data, the mean and variance of the estimated genome sharing ($G_{\mathrm{VR\text{-}O}_{ij}}$, $G_{\mathrm{VR\text{-}B}_{ij}}$ or $G_{\mathrm{IBD\text{-}L}_{ij}}$) within each class of relationship (Fig. 1) and for all the pedigrees were calculated and compared against the theoretical values. Correlations between the estimated relationship or genome sharing values and their corresponding additive relationship coefficients obtained from pedigree were also calculated.

For each replicate of simulated data, estimators were evaluated for precision by means of mean square error (MSE) and the Pearson correlation coefficient, ρ, between the estimated ($G_{\mathrm{VR\text{-}O}_{ij}}$, $G_{\mathrm{VR\text{-}B}_{ij}}$ or $G_{\mathrm{IBD\text{-}L}_{ij}}$) and the true values of genome sharing ($\breve{R}_{ij}$). The estimators were also evaluated for empirical bias, which was calculated by taking the difference $G_{ij} - \breve{R}_{ij}$ for each pair of animals and averaging them across pairs. Finally, the regression of true values of genome sharing on the estimated values was calculated as a measure of the closeness between estimators and the true relationships.

Consequences of using different *G* on accuracy of breeding values

The simulated data were used to test whether the improved estimates to set up $\boldsymbol{G}$ could result in significant gains in accuracy of genomic estimated breeding values (GEBV) for selection candidates. A single trait animal model $\boldsymbol{y} = \boldsymbol{1}_n\mu + \boldsymbol{Z}\boldsymbol{a} + \boldsymbol{e}$, with one phenotypic record per animal, except those from generation five (2000 selection candidates), was used. Hence, the left-hand side (LHS) of the mixed-model equations was equal to the following:

$$\text{LHS} = \begin{bmatrix} \mathbf{1}_n'\mathbf{1}_n\sigma_e^{-2} & \mathbf{1}_n'\mathbf{Z}\sigma_e^{-2} \\ \mathbf{Z}'\mathbf{1}_n\sigma_e^{-2} & \mathbf{Z}'\mathbf{Z}\sigma_e^{-2} + \mathbf{H}^{-1}\sigma_a^{-2} \end{bmatrix} \quad (7)$$

In (7), $\mathbf{Z}'\mathbf{Z}$ is a diagonal matrix with $d_{ii} = 1$ when animal i has a record and zero; otherwise, $\mathbf{H}^{-1}$ is the inverse of the covariance matrix of BV that combines pedigree and genomic information (Aguilar *et al.*, 2010), σ_a^2 is the additive genetic variance and σ_e^2 is the residual variance. Accuracy of GEBV for each animal was taken to be equal to the following:

$$r_i = \sqrt{1 - \frac{\text{PEV}_i}{\sigma_A^2}} \quad (8)$$

where PEV_i is the prediction error variance of animal i. To compare the different genomic relationship matrices, it was assumed that the correct covariance matrix of BV was $\breve{R}_{ij}$ with elements $\breve{R}_{ij}$ obtained using Equation (1). In the 'true' model (i.e. $\mathbf{H} = \breve{\mathbf{R}}$), PEV can be computed based on the inverse of LHS. When the covariance matrix of BV is misspecified, PEV can be calculated as in Henderson (1975):

$$\text{PEV} = \mathbf{C}^{aa} + \mathbf{C}^{aa}\mathbf{H}^{-1}\sigma_A^{-2}\left(\breve{\mathbf{R}} - \mathbf{H}\right)\mathbf{H}^{-1}\mathbf{C}^{aa} \quad (9)$$

with

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}{}^{-1} \end{bmatrix} \quad (10)$$

where $\mathbf{C}^{aa}$ is the inverse of LHS, obtained using (10) in (7), $\mathbf{G}^{-1}$ is the inverse of the genomic relationship matrix ($\mathbf{G_{VR\text{-}O}}$, $\mathbf{G_{VR\text{-}B}}$ or $\mathbf{G_{IBD\text{-}L}}$) and $\mathbf{A}_{22}{}^{-1}$ is the inverse of the pedigree-based numerator relationship matrix for genotyped animals (Aguilar *et al.*, 2010). Accuracies were computed under two heritability scenarios: $h^2 = 0.25$ and $h^2 = 0.15$.

Scheffé's multiple comparison procedure was used to test the significance of differences in accuracies between the covariance matrix estimators. Accuracies (acc_{ijk}) of selection candidates ($k = 1, \ldots, 2000$) were analysed using the mixed model (Proc Mixed SAS version 9.3.1, SAS Institute, Cary, NC, USA) $acc_{ijk} = \tau_i + r_j + \varepsilon_{ijk}$, where the relationship matrix estimator was treated as fixed (τ_i, $i = 1, \ldots, 4$ for $\mathbf{A}_{22}$, $\mathbf{G_{VR\text{-}O}}$, $\mathbf{G_{VR\text{-}B}}$ and $\mathbf{G_{IBD\text{-}L}}$, respectively), and the replicate (r_j, $j = 1, \ldots, 50$) was treated as a random effect. A banded main diagonal covariance matrix was used for errors ε_{ijk}, in which all observations having the same level of the fixed effect (τ_i) have the same variance parameter or component.

Estimating accuracy using (9), we assume that IBD relationships are a perfect description of genetic covariances across individuals (i.e. they correspond to the 'true' model), which in turn implies the hypothesis that all base alleles are different. This is wrong in the presence of large QTLs, but seems a reasonable assumption for most cases, as most genomic information comes from close relatives (i.e. Habier *et al.* 2013).

Results

Real data

In the real data, the estimated genome sharing was computed for a total of 84 254 pairs of genotyped animals. The mean and standard deviation of the absolute difference between the observed and the base allele frequency were 0.083 and 0.074, respectively. The observed pattern for the three estimators of genome sharing ($\mathbf{G_{VR\text{-}O}}$, $\mathbf{G_{VR\text{-}B}}$ and $\mathbf{G_{IBD\text{-}L}}$) within each pedigree relationship was similar: the estimated mean decreased as relationships become more distant (Table 1). However, the mean of $\mathbf{G_{IBD\text{-}L}}$ was closer to its theoretical value on nine of fourteen pedigree relationships; $\mathbf{G_{VR\text{-}O}}$ was the closest to the theoretical value for the grandparent–grand offspring and half-cousin relationships. The latter relationship involves the former one, as half-cousins have one grandparent in common. Besides, the mean of the estimated relationship between half-cousins followed the same pattern as the overall mean, and it was computed with

Table 1 Sample size (*N*), expectation of actual relationships (***R***) and sample mean of estimated genome sharing (***G***) using three different estimators for a real pig data set for specific types of relatives

Relationship	*N*	Expected	Mean		
		R	$\mathbf{G_{IBD\text{-}L}}$	$\mathbf{G_{VR\text{-}O}}$	$\mathbf{G_{VR\text{-}B}}$
Parent–offspring	784	0.5000	0.5000	0.4299	0.4824
Full-sibs	639	0.5000	0.5046	0.4286	0.4886
Three-quarter sibs (horizontal)	816	0.3750	0.3730	0.3126	0.3588
Half-sibs, mothers' (fathers) half-sibs	2848	0.3125	0.3231	0.2522	0.2997
Grandparent–grand offspring	1344	0.2500	0.2067	0.2299	0.2709
Half-sibs	7061	0.2500	0.2537	0.2185	0.2811
Uncle–nephew	1716	0.2500	0.2282	0.2279	0.2468
Double first cousins	544	0.2500	0.2343	0.2193	0.3150
Triple half-cousins	2912	0.1875	0.1754	0.1533	0.2197
Double half-cousins	5408	0.1250	0.1313	0.1076	0.1229
Half-uncle–nephew	6800	0.1250	0.1344	0.1216	0.1266
First cousins	6960	0.1250	0.1169	0.1097	0.1780
Half-cousins	22 944	0.0625	0.0735	0.0585	0.1019
Unrelated	23 478	0.0000	0.0000	0.0444	0.0599

the highest number of pairs. The estimators $\boldsymbol{G}_{\mathbf{VR\text{-}B}}$ for uncle–nephew, half-uncle–nephew and double half-cousins were closest to the true means. Note that uncle–nephew can be regarded as a two-way half-uncle–nephew relationship, whereas double half-cousins can be viewed as descendants of four half-uncle–nephew pairs.

Table 2 reports the theoretical standard deviations (SD) of actual relationships and the sampling SD of the estimated genome sharing for each type of relatives. The IBD-based values of estimated SD were always smaller than their IBS-based counterparts, whether the observed or base allele frequencies were used: on average, $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$, $\boldsymbol{G}_{\mathbf{VR\text{-}O}}$ and $\boldsymbol{G}_{\mathbf{VR\text{-}B}}$ were 7.50, 60.37 and 174.07% higher than the theoretical SD, for each pedigree relationship, respectively. Thus, the overlapping in the amount of IBD sharing from quite different pedigree relationships was higher for the IBS-based estimates.

For the real data set with pig records, the overall mean and standard deviation (SD) of the estimated genome sharing were compared against their theoretical values (Table 3) calculated using Equations (3) and (4) and based on pedigree and porcine genetic maps. The mean of genomic relationships was equal to the theoretical value when $\boldsymbol{G}_{\mathbf{VR\text{-}O}}$ was used, as this estimator was scaled based on $\boldsymbol{A}$ so that the means of diagonals and off-diagonals are the same as in the pedigree relationship matrix (Vitezica *et al.* 2011). The overall mean of $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$ was very close to the theoretical value. The estimator that differed most from the overall theoretical mean was $\boldsymbol{G}_{\mathbf{VR\text{-}B}}$. With respect to the overall SD of the estimated genome sharing, the value for $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$ was closer to the theoretical value than $\boldsymbol{G}_{\mathbf{VR\text{-}O}}$ or $\boldsymbol{G}_{\mathbf{VR\text{-}B}}$.

Table 2 SD of actual relationships (***R***) and estimated genome sharing (***G***) using three different estimators for a real pig data set for specific types of relatives

Relationship	$\boldsymbol{R}$	$\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$	$\boldsymbol{G}_{\mathbf{VR\text{-}O}}$	$\boldsymbol{G}_{\mathbf{VR\text{-}B}}$
Parent–offspring	0.0000	0.0000	0.0573	0.1188
Full-sibs	0.0527	0.0578	0.0826	0.1317
Three-quarter sibs (horizontal)	0.0476	0.0478	0.0711	0.1180
Half-sibs, mothers' (fathers) half-sibs	0.0447	0.0438	0.0641	0.1086
Grandparent–grand offspring	0.0456	0.0465	0.0993	0.1454
Double first cousins	0.0419	0.0472	0.0581	0.1017
Half-sibs	0.0373	0.0344	0.0609	0.0895
Uncle-nephew	0.0348	0.0361	0.0512	0.1204
Triple half-cousins	0.0386	0.0420	0.0560	0.1038
Double half-cousins	0.0350	0.0385	0.0504	0.0862
Half-uncle–nephew	0.0335	0.0375	0.0465	0.1110
First cousins	0.0297	0.0321	0.0535	0.0793
Half-cousins	0.0248	0.0279	0.0495	0.0709
Unrelated	0.0000	0.0000	0.0651	0.0854

Table 3 Overall mean and standard deviation (SD) of actual relationships (***R***) and estimated genomic relationships (***G***) across all pairs of genotyped individuals in a real pig data set

	R	$\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$	$\boldsymbol{G}_{\mathbf{VR\text{-}O}}$	$\boldsymbol{G}_{\mathbf{VR\text{-}B}}$
Mean	0.1062	0.1087	0.1062	0.1416
SD	0.1100	0.1090	0.0985	0.1273

The Pearson correlation coefficients between the estimated values of genome sharing and their corresponding pedigree-based additive relationship coefficient were 0.959, 0.797 and 0.702 for $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$, $\boldsymbol{G}_{\mathbf{VR\text{-}O}}$ and $\boldsymbol{G}_{\mathbf{VR\text{-}B}}$ respectively.

Simulation

In the simulated data, the estimated genome sharing was computed for a total of 9730 pairs of genotyped animals. For the simulated data set, the mean and standard deviation of the absolute difference between the base and the observed allele frequency were 0.072 and 0.068, respectively. Table 4 summarizes the precision and bias averaged over replicates that were achieved by the three different estimators ($\boldsymbol{G}_{\mathrm{VR\text{-}O}_{ij}}$, $\boldsymbol{G}_{\mathrm{VR\text{-}B}_{ij}}$ and $\boldsymbol{G}_{\mathrm{IBD\text{-}L}_{ij}}$) of the pairwise pedigree IBD genome sharing between simulated genotypes ($\breve{R}_{ij}$). All three estimators had very low empirical bias, being $\boldsymbol{G}_{\mathbf{VR\text{-}O}}$ the least unbiased. $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$ displayed lower sampling MSE and higher correlation with true values of genome sharing than $\boldsymbol{G}_{\mathbf{VR\text{-}O}}$. When allele frequencies in the base population were used, the correlation between $\boldsymbol{G}_{\mathbf{VR\text{-}B}}$ and the true value approximated the corresponding correlation for $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$, while having lower MSE than $\boldsymbol{G}_{\mathbf{VR\text{-}O}}$. The $\boldsymbol{G}_{\mathbf{VR\text{-}B}}$ estimator, although not always feasible to calculate (as the frequencies from the base population are not always available), assured a better scenario. The last column in Table 4 displays the regression of the true genomic relationships on the estimated genomic relationships. The regression coefficient was close to 1 for $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$, being significantly lower for both $\boldsymbol{G}_{\mathbf{VR}}$ estimators.

To analyse the consequences of using different $\boldsymbol{G}$ matrices in the accuracy of prediction of BV, the accuracy of GEBV for selection candidates was computed under two heritability scenarios: $h^2 = 0.25$ and $h^2 = 0.15$ (Table 5). As expected, the use of any of the genomic matrices resulted in greater accuracy of GEBV for selection candidates when compared to the pedigree-only-based relationship matrix. Accuracy of

Table 4 Performance of estimators of pairwise genomic relationships with the simulated data

	MSE(×100)	Pearson correlation	Bias	b_1*
$\boldsymbol{G}_{\mathbf{VR\text{-}O}}$	0.9352 ± 0.2847	0.678 ± 0.048	−0.0086 ± 0.0095	0.7483
$\boldsymbol{G}_{\mathbf{VR\text{-}B}}$	0.5703 ± 0.2059	0.876 ± 0.022	0.0180 ± 0.0185	0.7285
$\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$	0.1886 ± 0.0535	0.946 ± 0.008	0.0122 ± 0.0091	0.9723

*b_1 is the regression coefficient of the true genomic relationship on the estimated genomic relationships.

Table 5 Mean (SE) accuracy of GEBV for selection candidates under different relationship matrices over replicates

h^2		$\boldsymbol{A}$	$\boldsymbol{G}_{\mathbf{VR\text{-}O}}$	$\boldsymbol{G}_{\mathbf{VR\text{-}B}}$	$\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$
0.25	Genotyped*	0.498a (0.002)	0.538b (0.002)	0.538b (0.002)	0.559c (0.002)
	All*	0.497a (0.001)	0.518b (0.001)	0.518b (0.001)	0.521c (0.001)
0.15	Genotyped*	0.460a (0.003)	0.501b (0.003)	0.501b (0.003)	0.528c (0.003)
	All*	0.458a (0.002)	0.481b (0.002)	0.481b (0.002)	0.486c (0.002)

*Different letters in the same row indicate a statistically significant difference between the covariance matrices ($p < 0.0001$)

$\boldsymbol{A}$: pedigree-based relationship matrix; $\boldsymbol{G}_{\mathbf{VR}}$: IBS-based genomic relationship matrix constructed with either the observed allele frequencies ($\boldsymbol{G}_{\mathbf{VR\text{-}O}}$) or the frequencies of all base population animals ($\boldsymbol{G}_{\mathbf{VR\text{-}B}}$); $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$: IBD-based genomic relationship matrix.

GEBV for selection candidates was statistically higher when matrix $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$ was used. In fact, differences were larger for genotyped animals. The differences among the IBS-based estimators were not statistically significant. The accuracies dropped in the same magnitude when $h^2 = 0.15$ for the three estimators.

Discussion

De los Campos *et al.* (2013) found that 'the effectiveness of GBLUP depends critically on the extent to which marker-derived genomic relationships reflect the patterns of realized genetic relationships at causal loci'. The current research attempted to compare two approaches to estimate true realized relationships to be used in the set-up of genomic relationship matrices. One was the widely used VanRaden (2008) approach, which estimates relationships using only markers ($\boldsymbol{G}_{\mathbf{VR}}$). The second was an approach that uses genomic data to estimate realized relationships based on IBD sharing of marker alleles relative to the known pedigree ($\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$).

The real data set allowed comparing the empirical variation in genome sharing of relatives with the same pedigree relationship, from either IBD- or IBS-based estimators. The SD of the estimated genome sharing for $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$ was notably closer to the theoretical value than $\boldsymbol{G}_{\mathbf{VR\text{-}O}}$ or $\boldsymbol{G}_{\mathbf{VR\text{-}B}}$. In contrast, it was extremely difficult to distinguish different pedigree relationships from the actual fraction of the genome shared estimated by $\boldsymbol{G}_{\mathbf{VR}}$. Although $\boldsymbol{G}_{\mathbf{VR}}$ is an estimate of the realized proportion of genome-shared IBD, it does not take either the parent–offspring transmission or the segmental nature of inheritance of DNA into account (Thompson 2013). Indeed, permutation of the genotypes for each SNP will result in the same IBS-based $\boldsymbol{G}$ matrix. The mean of $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$ was extremely close to its theoretical value for most pedigree relationships. $\boldsymbol{G}_{\mathbf{VR\text{-}O}}$ was unbiased for the overall mean, yet it did not behave as well as $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$ when comparisons were made on a relationship basis. The most biased estimator was $\boldsymbol{G}_{\mathbf{VR\text{-}B}}$ (Table 3), which tended to overestimate pedigree IBD genome sharing. This can be explained in part by the fact that base allele frequencies were computed from a small number of animals that belonged to two different breeds (4 Duroc sires and 15 Pietrain dams) so that estimates of true base allele frequencies suffered from a lack of precision. In fact, $\boldsymbol{G}_{\mathbf{VR\text{-}B}}$ was the most biased for the half-cousins and unrelated relationships, which account for 27.2 and 27.9% of the pairwise estimated relationships, respectively, and are expected to have the lowest (or zero) theoretical mean pedigree IBD genome sharing (Table 1).

Results from our simulation allowed us to compare the precision and bias achieved by the different estimators of the true pedigree IBD genome sharing between genotyped animals. $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$ displayed higher precision than $\boldsymbol{G}_{\mathbf{VR\text{-}O}}$. This can be because $\boldsymbol{G}_{\mathbf{VR\text{-}O}}$ could not capture the unobserved history of relatedness within a small livestock population as the one simulated when dealing with a small number of genotyped animals. A better scenario was assured when allele frequencies in the base population were used, allowing the precision of $\boldsymbol{G}_{\mathbf{VR\text{-}B}}$ to approximate that of $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$. This result also agrees with the fact that

$\boldsymbol{G}_{\mathbf{VR\text{-}B}}$ was nearly unbiased in our simulation, in contrast to the results from real data, where base allele frequencies were not well represented by frequencies of F_0 genotyped animals. A solution, as in VanRaden (2008), could be to estimate base allele frequencies with a linear model that solves for gene content of non-genotyped ancestors and descendants using pedigree (Gengler *et al.* 2007).

Vela-Avitúa *et al.* (2015), in a simulated aquaculture breeding scheme, showed that differences in accuracies of GEBVs among $\boldsymbol{G}$ estimators depend on marker density: IBS-based GEBVs were slightly more accurate than their IBD-based counterparts using dense markers, but also considerably more sensitive to a reduction in density. Yet, these authors found that accuracy of IBD-based GEBV was stable across marker densities and, in fact, greater at low densities (≤100 SNP/M) than that achieved using the IBS-based $\boldsymbol{G}$ matrix. In our simulation using dense markers, accuracy of GEBV for selection candidates was statistically higher when matrix $\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$ was used. This slight superiority in accuracy could be explained by the fact that our IBD-based approach differs from that used in the above-mentioned article in that it models LD information. This is achieved by adding a background IBD state to fit the hidden relatedness beyond the relatedness that is observed through the available pedigree structure. Yet, this comes at the expense of using HMM methods that are computationally intensive (~4 hours per chromosome on a computer having a Quad-core 2.7 GHz AMD Opteron 8384 processor with 128 GB of memory).

Characterizing actual relationships in animal, human and agricultural populations is a key aspect in genetic analysis. QTL detection models in association analysis generally correct for structure and relatedness between individuals using a relationship matrix (either genomic or pedigree-based) or even using the methods of estimating genome-wide pairwise IBD within families (Kennedy *et al.* 1992; Kang *et al.* 2010; Legarra *et al.* 2015). Legarra *et al.* (2015) obtained similar results when comparing methods to detect QTL in four livestock species using markers, whether a genomic or a pedigree-based numerator relationship matrix was used. Yet, no further investigation on the subject has been carried out so far. A more precise genomic relationship matrix such as the one proposed in our research ($\boldsymbol{G}_{\mathbf{IBD\text{-}L}}$) may potentially imply higher power to detect QTL in livestock populations, where pedigree is (up to some extent) known.

With respect to the differences in accuracy of GEBVs among the IBS-based estimators, these were not statistically significant. Strandén & Christensen (2011) showed that changes in the numerator of $\boldsymbol{G}_{\mathbf{VR}}$ (as can the allele frequencies used to centre genotypes) do not change relative differences between the estimated GEBVs, because they are just shifted by a constant. However, modifying the denominator that scales $\boldsymbol{G}_{\mathbf{VR}}$ is like dividing or multiplying $\boldsymbol{G}$ by a constant and will, in principle, change results, although in our case this did not affect the results greatly.

Conclusion

Incorporating pedigree data to trace IBD inheritance in the calculation of genomic relationships improved the precision of estimates of actual relationships or proportion of genome shared between individuals in livestock populations. Moreover, the IBD-based method presented here better captures the extent of the variation in the actual proportion of genome shared by relatives that have the same kind or degree of pedigree relationship. When dealing with small numbers of genotyped animals, marker-only-based methods could be good estimators of $\boldsymbol{G}$ as well, provided that accurate inferences of allele frequencies in the base population were available. Using pedigree and markers, the gain in accuracy in elements of $\boldsymbol{G}$ was translated into higher accuracies in genomic breeding value predictions for selection candidates.

Acknowledgements

This project was supported by Agriculture and Food Research Initiative Competitive Grant no. 2010-65205-20342 from the USDA National Institute of Food and Agriculture and by funding from the National Pork Board Grant no. 11-042. Partial funding was also provided by the US Pig Genome Coordination Program. Computer resources were provided by the Michigan State University High Performance Computing Center (HPCC) and by Toulouse Midi-Pyrénées bioinformatic platform. NSF and RJCC were funded by grants of CONICET (PIP 2013-00833) and ANPCyT (PICT 2013-1661) from Argentina. Part of this work was made possible by a visit of NSF to INRA, Toulouse, France, financed by the Saint-Exupéry Scholarship Program 2013–2014 (MinCyT Argentina–French Embassy).

References

Aguilar I., Misztal I., Johnson D.L., Legarra A., Tsuruta S., Lawlor T.J. (2010) A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.*, **93**, 743–752.

Aguilar I., Misztal I., Legarra A., Tsuruta S. (2011) Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. *J. Anim. Breed. Genet.*, **128**, 422–428.

Cantet R.J.C., Vitezica Z.G. (2014) Properties of Mendelian residuals when regressing breeding values using a genomic covariance matrix. In: Proceedings of the 10th World Congress on Genetics Applied to Livestock Production. Vancouver (Canada), 17–22 August 2014. Paper number 687.

Cheng S.H., Higham N.J. (1998) A modified Cholesky algorithm based on a symmetric indefinite factorization. *SIAM J. Matrix Anal. Appl.*, **19**, 1097–1110.

De los Campos G., Vazquez A.I., Fernando R., Klimentidis Y.C., Sorensen D. (2013) Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet.*, **9**, e1003608.

Edwards D.B., Ernst C.W., Tempelman R.J., Rosa G.J.M., Raney N.E., Hoge M.D., Bates R.O. (2008) Quantitative trait loci mapping in an F_2 Duroc×Pietrain resource population: I. Growth traits. *J. Anim. Sci.*, **86**, 241–253.

Fernando R.L., Grossman M. (1989) Marker assisted selection using best linear unbiased prediction. *Genet. Sel. Evol.*, **21**, 467–477.

Frühwirth-Schnatter S. (2006) Finite Mixture Modeling. In: Bickel, P. et al. (ed), Finite Mixture and Markov Switching Models. Springer Series in Statistics, Springer, New York, pp. 1–23.

Gengler N., Mayeres P., Szydlowski M. (2007) A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. *Animal*, **1**, 21–28.

Goddard M.E., Hayes B.J., Meuwissen T.H.E. (2011) Using the genomic relationship matrix to predict the accuracy of genomic selection. *J. Anim. Breed. Genet.*, **128**, 409–421.

Gualdrón Duarte J.L., Bates R.O., Ernst C.W., Raney N.E., Cantet R.J.C., Steibel J.P. (2013) Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. *BMC Genet.*, **14**, 38.

Guo S.W. (1995) Proportion of genome shared identical by descent by relatives: concept, computation, and applications. *Am. J. Hum. Genet.*, **56**, 1468–1476.

Guo S.W. (1996) Variation in genetic identity among relatives. *Hum. Hered.*, **46**, 61–70.

Habier D., Fernando R.L., Garrick D.J. (2013) Genomic BLUP decoded: a look into the black box of genomic prediction. *Genetics*, **194**, 597–607.

Henderson C.R. (1975) Comparison of alternative sire evaluation methods. *J. Anim. Sci.*, **41**, 760–770.

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.

Higham N.J. (2002) Computing the nearest correlation matrix–a problem from finance. *IMA J. Numer. Anal.*, **22**, 329–343.

Hill W.G., Weir B.S. (2011) Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genet. Res. (Camb.)*, **93**, 47.

Jacquard A. (1974) The genetic structure of populations. Springer Verlag, Berlin.

Kang H.M., Sul J.H., Service S.K., Zaitlen N.A., Kong S.Y., Freimer N.B., Sabatti C., Eskin E. (2010) Variance component model to account for sample structure in genome- wide association studies. *Nat. Genet.*, **42**, 348–354.

Kennedy B.W., Quinton M., van Arendonk J.A. (1992) Estimation of effects of single genes on quantitative traits. *J. Anim. Sci.*, **70**, 2000–2012.

Legarra A., Croiseau P., Sanchez M.P., Teyssèdre S., Sallé G., Allais S., Fritz S., Moreno C.R., Ricard A., Elsen J.M. (2015) A comparison of methods for whole-genome QTL mapping using dense markers in four livestock species. *Genet. Sel. Evol.*, **47**, 6.

Li X., Yin X., Li J. (2010) Efficient identification of identical-by-descent status in pedigrees with many untyped individuals. *Bioinformatics*, **26**, 191–198.

Risch N., Lange K. (1979) An alternative model of recombination and interference. *Ann. Hum. Genet.*, **43**, 61–70.

Ramos A.M., Crooijmans R.P.M.A., Affara N.A., Amaral A.J., Archibald A.L., Beever J.E., Bendixen C., Churcher C., Clark R., Dehais P., Hansen M.S., Hedegaard J., Hu Z.-L., Kerstens H.H., Law A.S., Megens H.-J., Milan D., Nonneman D.J., Rohrer G.A., Rothschild M.F., Smith T.P.L., Schnabel R.D., Van Tassell C.P., Taylor J.F., Wiedmann R.T., Schook L.B., Groenen M.A.M. (2009) Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PLoS ONE*, **4**, e6524.

Sargolzaei M., Schenkel F.S. (2009) QMSim: a large-scale genome simulator for livestock. *Bioinformatics*, **25**, 680–681.

Sonesson A.K., Meuwissen T.H.E. (2000) Mating schemes for optimum contribution selection with constrained rates of inbreeding. *Genet. Sel. Evol.*, **32**, 231–248.

Strandén I., Christensen O.F. (2011) Allele coding in genomic evaluation. *Genet. Sel. Evol.*, **43**, 25.

Thompson E.A. (2013) Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*, **194**, 301–326.

Tortereau F., Servin B., Frantz L., Megens H.J., Milan D., Rohrer G., Wiedmann R., Beever J., Archibald A.L., Schook L.B., Groenen M. (2012) A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC Genom.*, **13**, 586.

VanRaden P.M. (2007) Genomic measures of relationship and inbreeding. *Interbull Bull*, **37**, 33–36.

VanRaden P.M. (2008) Efficient methods to compute genomic predictions. *J. Dairy Sci.*, **91**, 4414–4423.
Vela-Avitúa S., Meuwissen T.H.E., Luan T., Ødegård J. (2015) Accuracy of genomic selection for a sib-evaluated trait using identity-by-state and identity-by-descent relationships. *Genet. Sel. Evol.*, **47**, 9.
Vitezica Z.G., Aguilar I., Misztal I., Legarra A. (2011) Bias in genomic predictions for populations under selection. *Genet. Res. (Camb.)*, **93**, 357–366.
Wright S. (1922) Coefficients of inbreeding and relationship. *Am. Nat.*, **56**, 330–338.