Journal of
Animal Breeding and Genetics

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

ORIGINAL ARTICLE

Use of molecular markers to improve relationship information in the genetic evaluation of beef cattle tick resistance under pedigree-based models

V.S. Junqueira[1], F.F. Cardoso[2,3], M.M. Oliveira[4], B.P. Sollero[2], F.F. Silva[1] & P.S. Lopes[1]

1 Departamento de Zootecnia, Universidade Federal de Viçosa, Viçosa Minas Gerais (MG), Brazil
2 Empresa Brasileira de Pesquisa Agropecuária, Centro de Pesquisa de Pecuária dos Campos Sul-Brasileiros, Bagé Rio Grande do Sul (RS), Brazil
3 Departamento de Zootecnia, Universidade Federal de Pelotas, Pelotas Rio Grande do Sul (RS), Brazil
4 Departamento de Zootecnia, Universidade Federal de Santa Maria, Rio Grande do Sul (RS), Brazil

Keywords
Accuracy; beef cattle; numerator relationship matrix; SNPs.

Correspondence
V.S. Junqueira, Departamento de Zootecnia, Universidade Federal de Viçosa, Viçosa, Minas Gerais (MG), Brazil.
Tel: +55 31 3899 1419;
Fax: +55 31 3899 2579;
E-mail: junqueiravinicius@hotmail.com

Received: 23 February 2016;
accepted: 1 August 2016

Summary

The selection of genetically superior individuals is conditional upon accurate breeding value predictions which, in turn, are highly depend on how precisely relationship is represented by pedigree. For that purpose, the numerator relationship matrix is essential as *a priori* information in mixed model equations. The presence of pedigree errors and/or the lack of relationship information affect the genetic gain because it reduces the correlation between the true and estimated breeding values. Thus, this study aimed to evaluate the effects of correcting the pedigree relationships using single-nucleotide polymorphism (SNP) markers on genetic evaluation accuracies for resistance of beef cattle to ticks. Tick count data from Hereford and Braford cattle breeds were used as phenotype. Genotyping was carried out using a high-density panel (BovineHD - Illumina® bead chip with 777 962 SNPs) for sires and the Illumina BovineSNP50 panel (54 609 SNPs) for their progenies. The relationship between the parents and progenies of genotyped animals was evaluated, and mismatches were based on the Mendelian conflicts counts. Variance components and genetic parameters estimates were obtained using a Bayesian approach via Gibbs sampling, and the breeding values were predicted assuming a repeatability model. A total of 460 corrections in relationship definitions were made (Table 1) corresponding to 1018 (9.5%) tick count records. Among these changes, 97.17% (447) were related to the sire's information, and 2.8% (13) were related to the dam's information. We observed 27.2% (236/868) of Mendelian conflicts for sire–progeny genotyped pairs and 14.3% (13/91) for dam–progeny genotyped pairs. We performed 2174 new definitions of half-siblings according to the correlation coefficient between the coancestry and molecular coancestry matrices. It was observed that higher-quality genetic relationships did not result in significant differences of variance components estimates; however, they resulted in more accurate breeding values predictions. Using SNPs to assess conflicts between parents and progenies increases certainty in relationships and consequently the accuracy of breeding value predictions of candidate animals for selection. Thus, higher genetic gains are expected when compared to the traditional non-corrected relationship matrix.

doi:10.1111/jbg.12239

Introduction

In pedigree-based model, parents–offspring misidentifications often exist and cannot be all solved by preliminary consistency analysis. Many of these unidentified pedigree errors could be solved when dealing with genomic relationship matrix instead of numerator relationship matrix, because the marker relationship is considered under genomic approach (VanRaden 2008). However, the use of genomic methodologies requires the availability of suitable reference population for every trait, which will often mean several thousands of animals with records and genotype, especially in scenarios where a large number of proven sires for economically relevant traits is seldom obtainable, such as for beef cattle, pigs and chickens (Brito *et al.* 2011; Van Eenennaam *et al.* 2014). In such circumstances, molecular marker information can be used to detect and to correct progenitor-progeny errors and improve pedigree-based relationships. Van Vleck (1970b) reported that errors in the pedigree and/or lack of information on kinship affect the rate of genetic gain because it reduces the correlation between the true and predicted breeding values. Many studies evaluated the losses on variance components and genetic parameters estimation (Van Vleck 1970a,b), and the consequences in genetic evaluation of animals and plants (Habier *et al.* 2010) due to incompleteness of pedigrees.

Among the traits of interest for genetic evaluation in beef cattle is the resistance to the tick *Rhipicephalus* (*Boophilus*) *microplus*, which is responsible for damage to animal health because it transmits diseases and causes anaemia, debilitation, weight loss and death. All these factors are responsible to reduce the production systems efficiency. Traditionally, the ectoparasites are controlled primarily with the topical and/or parenteral application of chemicals. However, the existence of natural variability for resistance in cattle populations can be exploited to increase the frequency of favourable alleles that control the genetic resistance. In this sense, the inheritance of resistance to ticks and the selection of animals have been the aim of several studies (Oliveira *et al.* 2013). The selection of genetically superior individuals is conditional upon accurate breeding value predictions which, in turn, are highly depend on how precisely relationship are represented by pedigree. For that purpose, the numerator relationship matrix – which is a pedigree-based matrix – is essential as *a priori* information in mixed model equations.

Nevertheless, up until now there are no studies evaluating the effect of relationship correction using SNP markers on pedigree-based genetic evaluation of beef cattle. SNP markers can be used to check Mendelian conflicts between parents and progeny genotypes with subsequent use of this information to increase the quality of relationship information among animals in a given population (Wiggans *et al.* 2010). In this context, this study aimed to evaluate whether and how the correction of the relationship information using SNP markers affects the prediction of genetic merit for tick resistance of beef cattle.

Material and methods

Phenotype and genotype data

Records of tick counts in Hereford and Braford cattle breeds used in this study were derived from joint research programme carried out by Conexão Delta G Breeding Program, Gensys Associated Consultants and Embrapa South Livestock. All institutions are located in the Rio Grande do Sul State, Brazil. Animals participating in the genetic evaluation programme were evaluated by counting adult tick females attached to one side of their bodies to evaluate individual resistance to this parasite.

The studied trait was transformed by $log_{10}(x_i + 1.0001)$, where x_i is the number of ticks. This transformation was employed to ensure residual normality required to fit the considered linear mixed model. Individuals kept in the data file were between 326 and 729 days old at the time of the count. The contemporary groups (CG) were formed by the combination of the effects of farm, sex, year of birth, management group and count date. Records of CG with less than five animals and counts above or below 3.5 standard deviations in comparison with the CG mean were discarded from the data file. After restrictions, 146 contemporary groups remained. The edited data file included records of 4363 animals raised under extensive conditions. Of these individuals, 2188 animals had three subsequent tick counts, 1934 had two counts and 241 had only one count. Therefore, the total number of records was 10 673, of which 2369 records were from Hereford animals, and 8304 were from Braford animals with a maximum of ¾ of Zebu proportion. The heterozygosity effects and recombination loss were calculated as proposed by Cardoso & Tempelman (2004) and included as linear covariate in the model. The log_{10} means and standard deviations for animals with one, two and three records were 1.3214 ± 0.4344, 1.1941 ± 0.4520 and 1.4504 ± 0.4057, respectively.

Blood, hair or semen samples were used for DNA extraction and genotyping of 3591 individuals.

Genotyping of 130 sires was performed using the high-density panel (BovineHD - Illumina® bead chip with 777 962 SNPs), while the BovineSNP50 Illumina panel (54 609 SNPs) was used for the remaining 3461 animals. The quality control criteria adopted for SNPs exclusion were the Hardy–Weinberg equilibrium chi-square test ($p = 10^{-7}$), genotype call rate (CR) (<98%), minor allele frequency (MAF) (<3%), near-perfect collinearity with another SNPs ($r > 0.98$) and SNPs in the same position. The criteria adopted to reject samples were CR <90%, heterozygosity deviation above three standard deviations, sex identification errors and identical genotypes between different individuals (more than 99.5% of similarity for all markers). A total of 41 045 markers from the BovineSNP50 chip remained after quality control. Bovine HD genotypes were filtered to retain just the 34 042 autosomal markers in common (overlapped) with the BovineSNP50, after quality control. Missing genotypes (0.89% of all genotypes) were imputed according to the sliding windows method using FImpute software (Sargolzaei *et al.* 2011).

Strategy for parentage corrections

The correctness of the pedigree relationship was evaluated for all genotyped parent–progeny pairs. Mismatches were identified based on the Mendelian conflicts counts, as proposed by Wiggans *et al.* (2009). The authors defined a conflict as the case when the progeny has a homozygous genotype, and one of the parents has a contrasting homozygous genotype. When the rate of conflicts exceeded 1% (threshold) of the total SNPs, the parentage was rejected, and an alternative parent was sought among the genotyped of the appropriate sex in the population. This approach was applied in the original pedigree file (Pedigree 1), and after corrections, a new improved pedigree was generated (Pedigree 2). The SeekParentF90 software (Aguilar 2014) was used to perform these assessments.

Approximately 60.54% of genotyped individuals in the population had unknown paternity due to the use of multiple-sire breeding management with non-genotyped sires. Thus, we adopted the approach described by Fernandez & Toro (2006) to reconstruct half-sibs families within multiple-sire groups and set up new half-sibling relationships using Pedigree 2 as starting pedigree file. This method uses the correlation coefficient between the pedigree coancestry matrix and molecular coancestry matrix as a criterion for pedigree reconstruction. A common phantom parent was attributed to each identified family of genotyped individuals who had no sire in Pedigree 2, thus generating new half-sibling genetic relationships in the Pedigree 3.

Statistical model and variance components estimation

Variance components and genetic parameter estimates were obtained using the software GIBBSF90, described by Misztal (2008), which uses a Bayesian approach via Gibbs sampling. A total of 500 000 cycles were considered with a sampling interval of 50 samples and discarding the 100 000 initial samples (burn-in). Convergence was assessed by Geweke criterion, which consists in an approximated Z-test comparing two different segments of MCMC chain (Geweke 1992). The null hypothesis assumes these two segments being from a common stationary distribution (e.g. chain convergence).

The following repeatability animal model in matrix notation was used to perform the analysis:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Wpe} + \mathbf{e},$$

where $\mathbf{y}$ is the observation vector, $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{W}$ are incidence matrices that relate $\mathbf{y}$ to the systematic effects vector $\mathbf{b} = \{b_j\} \sim \mathrm{N}(b_0, \mathbf{V}_b)$, where $\mathbf{V}_\mathrm{b}$ is a diagonal matrix of the *a priori* variance of $\mathbf{b}$, assuming $\mathbf{V}_\mathrm{b} \to \infty$. Moreover, the following distributions were assumed for additive genetic effects vector $\mathbf{u}|\sigma_u^2 \sim \mathrm{N}(0, \mathbf{A}\sigma_u^2)$, permanent environmental effects vector $\mathbf{pe}|\sigma_{pe}^2 \sim \mathrm{N}(0, \mathbf{I}\sigma_{pe}^2)$ and $\mathbf{e}|\sigma_e^2 \sim \mathrm{N}(0, \mathbf{I}\sigma_e^2)$ is the random residual vector. The scaled inverse chi-square distribution was assumed as *a priori* distribution for the additive genetic (σ_u^2), permanent environment (σ_{pe}^2) and residual (σ_e^2) variances with degrees of belief equals to $\nu_u = 5$, $\nu_{pe} = 5$ and $\nu_e = 5$, respectively.

In the systematic effects, contemporary groups were considered as classes, whereas the Zebu breed proportion, heterozygosity and recombination loss were included as linear covariates, and the age at the time of counting as a linear and quadratic covariate.

The following assumptions for the adopted model were considered:

$$\mathrm{E}(\mathbf{y}) = \mathbf{Xb},\ \mathrm{E}\begin{bmatrix}\mathbf{u}\\ \mathbf{pe}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix},$$

$$\mathrm{V}\begin{bmatrix}\mathbf{u}\\ \mathbf{pe}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{A}\sigma_u^2 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\sigma_{pe}^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2\end{bmatrix},$$

where, $\mathbf{A}$ is the numerator relationship matrix of Wright among all animals, and $\mathbf{I}$ is an identity matrix.

Under a Bayesian approach, the following joint distribution of the observed data (likelihood function) and the following *a priori* probability distributions for the parameters were considered:

$$\mathbf{y}|\mathbf{b},\mathbf{u},\mathbf{pe},\mathbf{e},\sigma_u^2,\sigma_{pe}^2,\sigma_e^2 \sim \mathrm{N}\left(\mathbf{Xb}+\mathbf{Zu}+\mathbf{Wpe},\mathbf{I}\sigma_e^2\right)$$

Models choice criteria

Each of the three pedigree files considered in this study generates a specific genetic (co)variance structure among the different levels of the additive genetic effect, that is, a different numerator relationship matrix. In this regard, each analysis was considered as a specific model and compared using the deviance information criterion (DIC) (Spiegelhalter *et al.* 2002). The DIC was used as a global fit criterion and is defined as a function of the posterior deviance (DEV):

$$\mathrm{DEV} = \frac{1}{G}\sum_{I=1}^{G} -2logp\left(y|\theta^{(I)},M\right),$$

where, $\theta^{(I)}$ is the set of parameters obtained in the Ith of a total of G Markov chain Monte Carlo (MCMC) cycles for model M, and the effective number of parameters (p_D) :

$$p_D = \mathrm{DEV} - D(\bar{\theta}),$$

where, $D(\bar{\theta}) = -2logp(y|\bar{\theta},M)$ is the deviance evaluated at the posterior mean of all parameters ($\bar{\theta}$) in model M. Thus, the DIC for the models is determined as follows:

$$\mathrm{DIC} = \mathrm{DEV} + p_D.$$

Smaller DIC estimates are indicative of better model fit and smaller p_D of lower degree of model complexity.

Prediction accuracy of breeding values

We performed best linear unbiased prediction (BLUP) analyses to estimate the prediction error variance (PEV) assuming *a priori* variance components estimated under the Bayesian approach. This strategy was used to derive accuracies of individual breeding values estimated from mixed model equations, which were then used to evaluate the accuracy gain when performing pedigree corrections, based on SNP data. For a given animal i, the model derived prediction accuracy ($\hat{\rho}_{u\hat{u}_i}$) of its genetic value $\hat{u}_i$ was estimated by :

$$\hat{\rho}_{u\hat{u}_i} = \sqrt{1-\frac{PEV_i}{\hat{\sigma}_u^2}}$$

where PEV_i is the prediction error variance of $\hat{u}_i$ and $\hat{\sigma}_u^2$ is the estimated additive genetic variance under Bayesian approach.

Cross-validation

Cross-validation is a robust technique for model evaluation, and it was used in this study to compare the predictive ability of the different pedigree models. In summary, the K-means clustering method was applied to the dissimilarity matrix obtained from genomic relationship matrix. Each dissimilarity matrix value was computed as $d_{ij} = 1 - g_{ij}/\sqrt{g_{ii}\cdot g_{jj}}$, where d_{ij} is a measure of genomic distance between individual i and individual j; g_{ij} is the additive genomic relationship between individual i and j; g_{ii} and g_{jj} are diagonal elements of genomic relationship matrix of individual i (or j). A total of 5 clusters (fivefold) were defined. K-means method intends to group the most genetic related individuals together and split less related individuals in different clusters (Saatchi *et al.* 2011). Alternatively, groups were formed by randomly sampled the individuals to the training and validation sets, such that relationships were equivalent within and between groups. Additional details about the two cross-validation strategies and about the considered population can be found in Cardoso *et al.* (2015). The posterior means of the variance components were used as *a priori* information to derive BLUP for the validation group using phenotypic data just of the training set. The prediction ability calculated from cross-validation was defined as the correlation between the adjusted tick counts ($\mathbf{y}^* = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{W}\widehat{\mathbf{pe}}$) and the predicted values ($\hat{\mathbf{y}}^* = \hat{\mathbf{u}}$) from the validation dataset, $\hat{r}_{y^*\hat{y}^*} = r(\mathbf{y}^*, \hat{\mathbf{y}}^*)$. Aiming to assess significant statistical differences between the predicted breeding value accuracy among validation sets, an analysis of variance was performed using the replicates obtained from fivefold cross-validation strategy. The Tukey's test ($\alpha = 0.05$) was adopted to evaluate significance of mean differences.

Results

Using SNP markers information, corrections in the original pedigree (Pedigree 1) were made to generate Pedigree 2. A total of 460 corrections in relationship

definitions were made (Table 1) corresponding to 1018 (9.5%) tick count records. Among these changes, 97.2% (447) were related to the sire's information, and 2.8% (13) were related to the dam's information. We observed 27.2% (236/868) of Mendelian conflicts for sire–progeny genotyped pairs and 14.3% (13/91) for dam–progeny genotyped pairs. In Pedigree 3, besides the corrections on parentage information present in Pedigree 2, phantom parents were assigned to 2174 individuals, generating 704 new half-sib families.

The distribution of the number of SNP markers of Mendelian conflicts identified in Pedigree 1 was remarkably different for matching and mismatching parent–progeny pairs (Figure 1). The mean ± SD number of conflicting genotypes for matching pairs was 17.73 ± 9.28 SNPs (range: 0–75), which corresponds to 0.04% of the total number of SNPS (41 045) with the observed maximum of 0.18% Mendelian conflicts. This value was far below the established threshold of 1% conflict tolerance. For mismatching pairs, the corresponding mean was 2704.13 ± 408.68 SNPs (range: 595–4993), which is equivalent to 6.59% of the total markers evaluated and ranges between 1.45 and 12.16%.

Table 1 Mendelian conflicts between parents and offspring according to gender

	Progeny		
Relationship	Male	Female	Total
Sire			
Confirmed	452	180	632
Reassigned	61	23	84
Rejected	96	56	152
Assigned	185	26	211
Unchecked	1080	7254	8334
Dam			
Confirmed	78	0	78
Reassigned	5	0	5
Rejected	8	0	8
Assigned	0	0	0
Unchecked	2931	8900	11 831

Confirmed: genotyped parent–progeny pair that maintained the same sire/dam assignment after evaluating Mendelian conflicts; Reassigned: Conflicting genotyped parent–progeny pair with an alternative sire/dam match found among other genotyped parents; Rejected: Conflicting genotyped parent–progeny pair and no alternative sire/dam match found; Assigned: individual with no sire/dam in the pedigree file and compatible matching sire identified; Unchecked: Pair parent/progeny was not genotyped or just progeny genotype and not compatible with any alternative genotyped parent.

Model choice

The DIC indicated that the model considering Pedigree 3 had a better fit than the models that considered Pedigree 1 or Pedigree 2 (Table 2). The improvement of the relationship matrix resulted in better overall fit (decreased DEV) and also smaller effective number of parameters, determining factors for the smaller DIC value of the model that used Pedigree 3. As pedigree information was augmented, we improved the correlation structure between breeding values of related individuals. This higher interdependency between genetic parameters may be the factor driving the observed decrease in effective number of parameters from Pedigree 1 to Pedigree 2 and from Pedigree 2 to Pedigree 3 (Table 2).

The difference between the DIC values calculated in this study ranged from 34 to 47 between the models that used pedigrees 1 and 2 and between models that used pedigrees 1 and 3, respectively. Following Spiegelhalter *et al.* (2002), DIC differences greater than seven units are indicative of substantial support (evidence) of the model with the smaller DIC.

Variance components and genetic parameters

The Geweke convergence criterion (Table 3) indicates convergence of all dispersion parameters in the three models when generating 500 000 MCMC chains, 100 000 samples for burn-in and a sampling interval of 50, totalling 8000 effective samples used for variance component estimation. The effective sample size (ESS) was used to estimate the number of independent samples with information equivalent to that contained within the dependent sampling. Even though Stock *et al.* (2007) in a simulate study showed that variance components and genetic parameters were not statically significant due to different ESS estimates. We wanted to assure that all ESS were at least 100 for all dispersion parameters in this study (Table 3).

Similar posterior mean, mode and median of variance components within each of the three pedigree files (Table 3) indicate that their marginal posterior distributions are fairly symmetric (Figure 2). As the quality of the relationship information increased from Pedigree 1 to Pedigree 2 and from Pedigree 2 to Pedigree 3, the estimated additive genetic variance tended to have greater magnitude and the permanent environmental variance had proportionally smaller values. Despite the overlapping of highest posterior density (HPD) intervals for the parameters among models (Table 3), which suggest no significant difference between them, there was a clear and expected

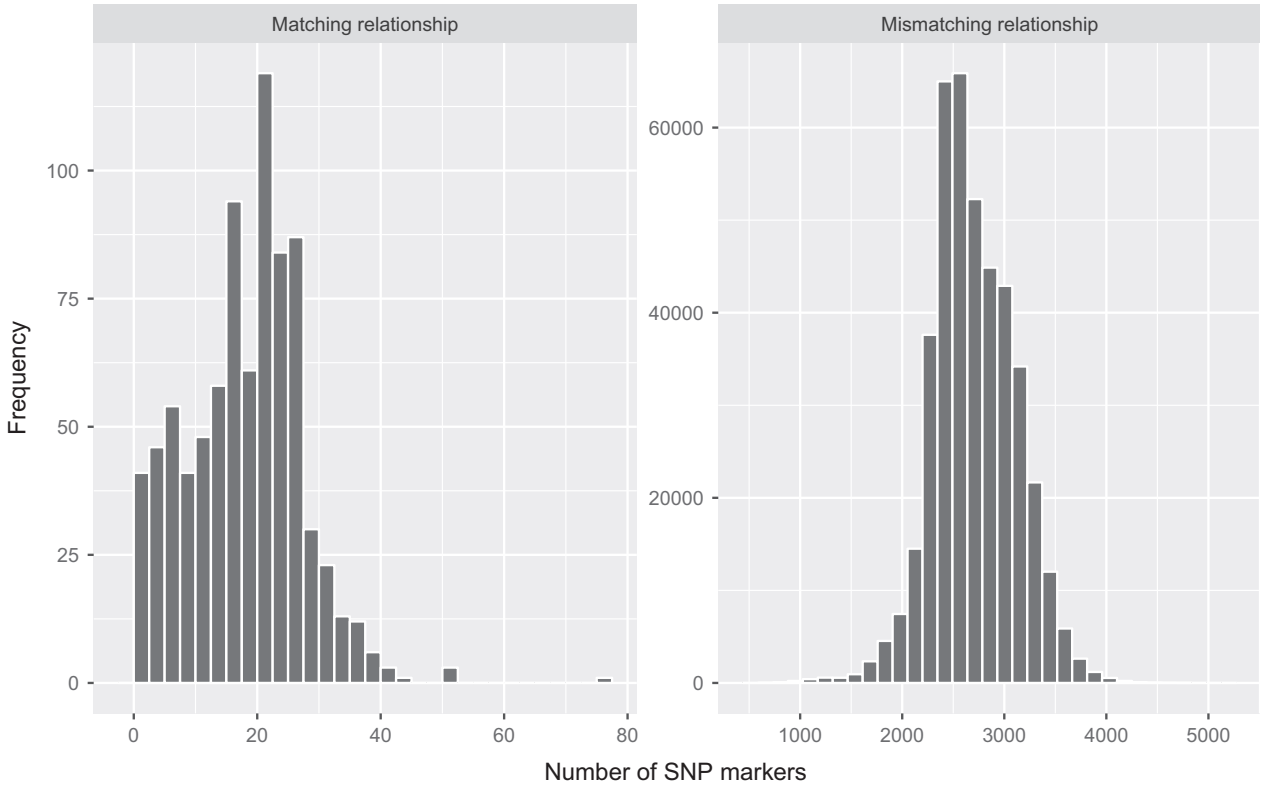

Figure 1 Distribution of the number of markers with Mendelian conflicts for parent-progeny pairs in the original pedigree (Pedigree 1) with matching or mismatching relationship based on SNPs.

Table 2 Bayesian model choice using different genetic (co)variance structures

Pedigree	DEV	P_D	DIC
1	2141.414	2224.001	4365.415
2	2136.276	2215.813	4352.089
3	2124.329	2193.268	4317.597

1: original pedigree file; 2: pedigree file with parent–progeny conflict corrections; 3: pedigree file with parent–progeny conflict corrections and definition of half-siblings based on genotypic information; mean deviance (DEV); penalty for the effective number of parameters (P_D); deviance information criterion (DIC).

trend of attributing a larger proportion of the repeatability to the additive genetic component as pedigree information improves. Thereby, heritability was greater for Pedigree 3 than Pedigree 2, which was greater than Pedigree 1. Moreover, the model using Pedigree 3 had the shortest interval of HPD for the additive genetic and permanent environmental variances. The posterior standard deviations of the estimated parameters were similar among the different models (Table 3). Finally, Figure 2 shows that there was no significant effect on the residual variance estimates when using different genetic (co)variance structures.

Predicted breeding values and accuracies

Spearman rank correlations between the predicted breeding values were assessed, considering different ratios of candidates for selection and parentage modifications (Table 4). Table 4 presents Spearman correlations between estimated breeding values considering different genetic (co)variance structures (**A** matrices based on different pedigrees) for the subset of individuals with direct changes in their parentage assignments and the alternative set of individuals that had no corrections on pedigree and remained with unchanged parentage. Considering only the animals that had changed parentage, the magnitude of the predicted breeding value correlations was higher between the models that used pedigrees 2 and 3 than between pedigrees 1 and 2 or 1 and 3. The lower the ratio of the individuals selected, the lower the correlation between the models, particularly between pedigrees 1 and 3. Despite the new relationships added in Pedigree 3, based on rank correlations presented in Table 4, the

Table 3 Variance components and genetic parameters observed using different genetic (co)variance structures

Pd	P	PM	PMD	PMO	PSD	HPD	Z	ESS
1	σ_p^2	0.100	0.100	0.100	0.0015	0.0977 – 0.1036	0.3209	6845
	σ_u^2	0.015	0.015	0.014	0.0032	0.0087 – 0.0209	0.1553	356
	σ_{pe}^2	0.014	0.014	0.014	0.0032	0.0085 – 0.0207	−0.0108	368
	σ_e^2	0.072	0.071	0.071	0.0013	0.0691 – 0.0742	−1.1853	7476
	h^2	0.146	0.146	0.144	0.0316	0.0888 – 0.2085	0.1487	401
	r	0.288	0.288	0.288	0.0121	0.2647 – 0.3125	1.5523	6679
2	σ_p^2	0.100	0.101	0.100	0.0015	0.0984 – 0.1044	−0.0883	6499
	σ_u^2	0.019	0.019	0.019	0.0033	0.0122 – 0.0248	−0.5671	348
	σ_{pe}^2	0.010	0.010	0.010	0.0031	0.0048 – 0.0166	0.7832	323
	σ_e^2	0.071	0.071	0.071	0.0013	0.0691 – 0.0741	−1.7512	7325
	h^2	0.186	0.187	0.189	0.0321	0.1230 – 0.2455	−0.5691	440
	r	0.289	0.289	0.290	0.0120	0.2661 – 0.3133	1.8360	6152
3	σ_p^2	0.101	0.101	0.101	0.0016	0.0985 – 0.1044	1.3220	5812
	σ_u^2	0.022	0.022	0.023	0.0030	0.0162 – 0.0279	0.3785	318
	σ_{pe}^2	0.007	0.007	0.007	0.0027	0.0029 – 0.0129	−0.2182	281
	σ_e^2	0.072	0.071	0.071	0.0013	0.0690 – 0.0740	0.3076	7666
	h^2	0.220	0.222	0.227	0.0286	0.1604 – 0.2716	0.3482	437
	r	0.294	0.295	0.296	0.0123	0.2708 – 0.3192	0.9596	6126

1: original pedigree file; 2: pedigree file with parent–progeny conflict corrections; 3: pedigree file with parent–progeny conflict corrections and definition of half-siblings based on genotypic information; σ_p^2: phenotypic variance; σ_u^2: additive genetic variance; σ_{pe}^2: permanent environmental variance; σ_e^2: residual variance; h^2: heritability and r: repeatability; Pedigree file (Pd); parameter (P); posterior mean (PM); posterior median (PMD); posterior mode (PMO); posterior standard deviation (PSD); posterior highest density interval (HPD); Z-Geweke (Z); effective sample size (ESS).

most significant changes in the genetic (co)variance matrix occurred from Pedigree 1 and Pedigree 2, when direct parent assignments were altered.

Scatter plots of the breeding values accuracy between different pedigree files of the 10% top individuals with parentage corrections are shown in Figure 3, which shows that quality improvement in the genetic relationship information provided higher accuracy values. This can be observed in the lower limit of the dispersions that pertain to individuals with accuracy estimates equal to 0.40 according to Pedigree 1, while they have values greater than 0.55 according to pedigree 2 and 3.

The parentage corrections provided higher average accuracy values (Table 5). The model that used Pedigree 3 exhibited a higher mean accuracy in relation to models that used pedigrees 1 or 2 for all selected proportion of individuals.

Cross-validation

The cross-validation average predictive ability using K-means and random groups for the pedigrees are shown in Figure 4. There was no interaction between the clustering strategy and pedigree correction. As expected, random selection strategy provided higher predictive ability than K-means strategy. In concordance with the behavior of individual accuracies from PEV, in general, a higher quality in the relationship information produced an average predictive ability of greater magnitude (Figure 4).

Discussion

Mendelian conflicts

The present study evaluated the gain in tick resistance breeding value prediction accuracy improvement when correcting the genetic relationships defined in beef cattle pedigree based on SNP information.

There was an increase in the quality of the genetic relationship of 460 animals, whereas most of the misidentifications were related to sire information (Table 1). Records used in the present study came from farms that use multiple-sire mating and with the availability of bulls and progeny genotypes, 211 animals that had no defined sire in the original pedigree could be associated with their actual sire. Conflicts were also observed in the identification of genotyped dams (14.3%), suggesting some problems with calving recording at the farms, but in a lesser extent compared to sire assignments (27.2%). Different results were reported by Wiggans *et al.* (2012) who assessed Mendelian conflicts between parent and progeny of Holstein, Jersey and Brown Swiss cattle using 3K SNPs and found an equal rate of bull and cow conflicts

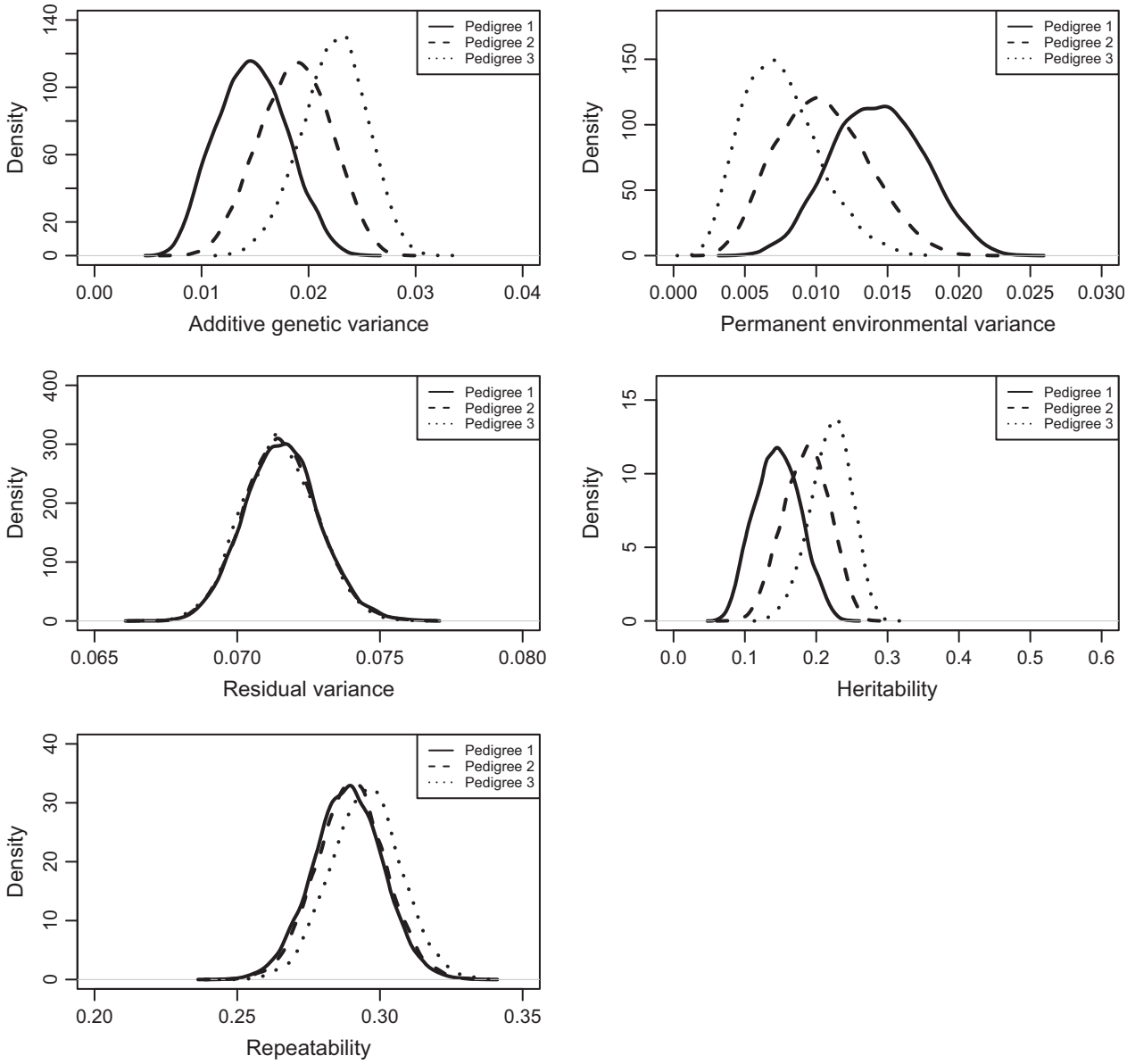

Figure 2 Posterior density of additive genetic, permanent environmental, residual variance, heritability and repeatability using different genetic (co)-variance structures of the original pedigree (Pedigree 1), a pedigree corrected for Mendelian conflicts (Pedigree 2) and a pedigree that reconstructed half-sib families (Pedigree 3).

for dairy breeds. This may be due to differences in calving management between dairy and beef operations.

Matching parent–progeny relationships had very low conflict rate between the markers' genotypes of the progeny and their parents (average of 0.18%), demonstrating high certainty to confirm that the relationship was correct. Although the identification of Mendelian conflicts was performed evaluating one parent at a time, the correction of errors in the pedigree could be performed with high accuracy using 41 045 SNPs. There was a wide gap between the maximum number of conflicts for matching pairs (75) and the corresponding minimum for mismatching pairs (595), suggesting high precision in the correction of relationships. Wiggans *et al.* (2009) using 37 811 SNP markers in a population of Holstein cattle, despite finding a smaller average number of conflicts than the present study, being of 2.3 (range: 0–89) for a match relationship and of 2411 (range: 754–3507) for mismatch relationship, also had a safe gap between the number of conflicts distribution for these two

Table 4 Spearman correlation between estimated breeding values considering different genetic relationship structures (**A** matrix based on different pedigrees) for individuals with changed and unchanged parentage

	Percentage of the best selected individuals		
Pedigree	5	20	50
Only individuals with parentage changed[a]			
1 and 2	0.626	0.706	0.728
1 and 3	0.728	0.799	0.844
2 and 3	0.770	0.806	0.881
Only individuals with parentage unchanged[b]			
1 and 2	0.966	0.972	0.977
1 and 3	0.939	0.951	0.961
2 and 3	0.978	0.985	0.988

1: original pedigree file; 2: pedigree file with parent–progeny conflict corrections; 3: pedigree file with parent–progeny conflict corrections and definition of half-siblings based on genotypic information. [a]Only those individuals that have their pedigree corrected, [b]only those individuals that did not have modifications on their parentage assignments.

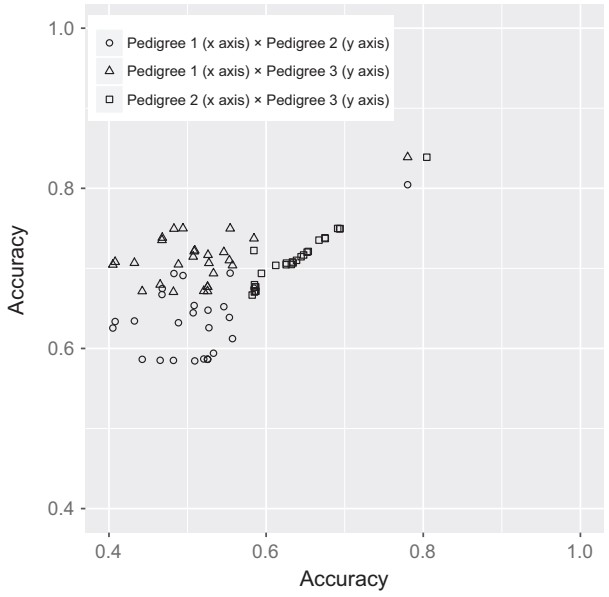

Figure 3 Scatter plots of predicted breeding value accuracies estimated from prediction error variances when using pedigrees 1 and 2 (circles), pedigrees 1 and 3 (triangles) and pedigrees 2 and 3 (squares). Individuals considered in the comparison were the 10% top individuals with parentage corrections in all combinations of pedigree files.

situations, confirming the utility of medium density SNP panels for pedigree correction.

New half-sibs' families

The identification of 2174 new half-sib relationships provided higher numeric values for additive genetic variance and heritability estimates. The positive trend in these parameters when improving relationship information indicates that usage of incomplete pedigrees may lead to downward biased estimates of the proportion of the phenotypic variance that can be attributed to additive genetic factors (Table 3). Despite the minor difference in off-diagonal correlation between genomic relationship matrix and numerator relationship matrix of 0.818 and 0.807 for pedigrees 3 and 2, respectively, accuracies presented different magnitudes. Most of this improvement in accuracies was related with the increase in nonzeros elements of pedigree-based relationship, in which Pedigree 3 had 19 814 more genetic links than Pedigree 2.

Table 5 Average accuracy from animals with records estimated from prediction error variances according to different pedigrees used to compose the numerator relationship matrix

	Percentage of the best selected individuals		
Pedigree[a]	5	20	50
1	0.455	0.461	0.465
2	0.615	0.620	0.623
3	0.700	0.704	0.707

[a]1: original pedigree file; 2: pedigree file with parent–progeny conflict corrections; 3: pedigree file with parent–progeny conflict corrections and definition of half-siblings based on genotypic information.

Model fit evaluation

The DIC of the model considering Pedigree 3 was lower than that found using Pedigree 1 or Pedigree 2 (Table 2), indicating the superior fit of the model that adopted Pedigree 3, exceeding almost five times the value of 7 units difference between models considered as relevant for this criterion (Spiegelhalter *et al.* 2002). The DIC criterion is well established and has been successfully applied in several studies as a model selection criterion. In the present study, the results for criterion reassure the importance of having reliable pedigrees to obtain accurate breeding value prediction using animal models.

Variance components and genetic parameters

The HPD, which is the interval that maximizes the posterior density, had shorter range between the 2.5 and 97.5% quantiles for the additive genetic and permanent environmental variance components under the model that used Pedigree 3 (Table 3). This suggests less uncertainty for these parameter estimates when a more accurate pedigree was used.

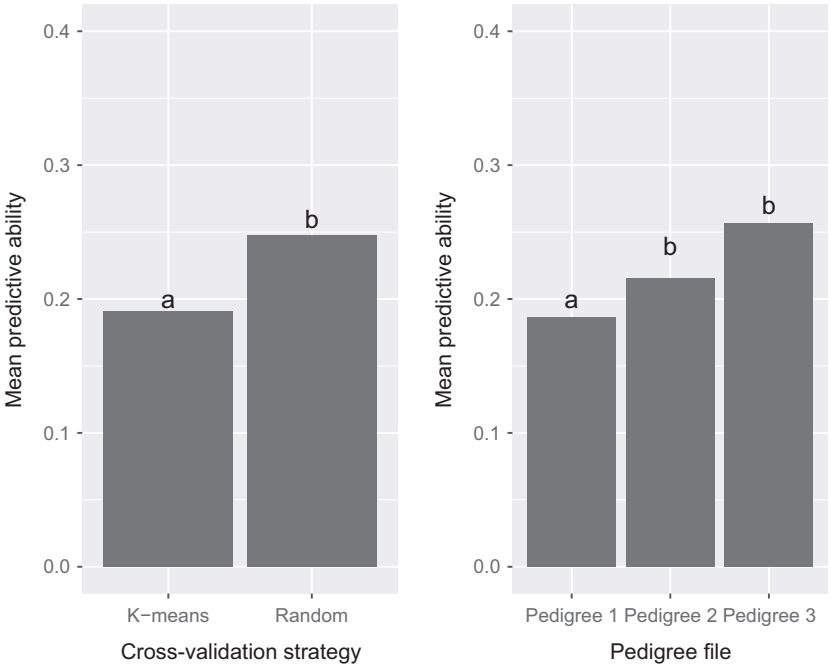

Figure 4 Average predictive ability of models based on the original pedigree (Pedigree 1), a pedigree corrected for Mendelian conflicts (Pedigree 2) and a pedigree that reconstructed half-sib families (Pedigree 3) from fivefold cross-validations using K-means and random grouping strategies. Bars sharing the same letter are not statistically different at α=0.05.

It is important to highlight that higher additive genetic variance and lower permanent environment estimates were somewhat expected as a consequence of relationship improvement in an animal model with repeated measures. Improvements in the additive genetic relationships due to pedigree corrections change the structure of **A** and, as a consequence, permanent environmental effects will also be affected as they are jointly estimated under a Bayesian framework. In this context, the variance partition was largely due to re-estimation of **u** and **pe** effects after pedigree corrections. Correcting the relationships also produced higher heritability estimates (Figure 2), suggesting the possibility of applying direct selection for cattle resistance to ticks. The main consequence of pedigree misidentifications is the reduction in the annual rate of genetic gain, mainly due to incorrect heritability estimates and incorrect prediction of breeding value, inducing inadequate selection of candidates (Sanders *et al.* 2006).

Implications of relationship corrections on genetic evaluations

The computation of traditional relationship matrix (i.e., based on identity-by-descent theory) does not account for the genetic content information shared between two individuals that are not related by pedigree. However, even in situations where molecular data is available, the decision on the adoption of SNP markers for the genetic merit estimation must consider a broad quantity of factors that directly influence the success of reliable parameters estimation. VanRaden *et al.* (2009) argue that reliability on genomic predictions strongly depends of the number of genotyped bulls in the reference population. In this sense, it is common that a large reference population size is not available for most traits of interest in beef cattle; nevertheless, misidentifications on parent–progeny pedigree data could be corrected using SNP markers aiming to get more reliable pedigree-based breeding values.

The correlations between breeding values considering different set of animals (changed and unchanged parentage after pedigree corrections based on SNP markers) and accuracy estimates are presented, respectively, in Tables 4 and 5. As expected, for individuals with changed parentage, correlations between breeding values from corrected pedigrees (2 and 3) were higher in comparison with original pedigree (1). The most important changes and consequently lower correlations were due parentage corrections (sire/dam reassignment, new assignments or rejection of original assignments) made from Pedigree 1 to Pedigree 2. The new phantom sire families introduced in Pedigree 3 had moderate effect on predictions and correlations between breeding values of animals with changed relationship in comparison with Pedigree 2 (ranged from 0.77 to 0.88). When comparing pedigrees 1 and 3,

estimated correlations were intermediate between the values obtained for pedigree 1 and 2, and 2 and 3. On the other hand, when considering only individuals with unchanged parentage, correlations were very high (>0.93) for all pedigree pair combinations, 1 and 2, 1 and 3 and 2 and 3.

In general, changes in breeding values prediction (Table 4) also reflected in changes of accuracy estimates (Table 5). In spite of true breeding values are unknown, it is reasonable to suppose that pedigree corrections provide an improvement on accuracy estimates because more realistic resemblance between relatives was achieved. This was observed by the increasing averages of PEV accuracy estimates from Pedigree 1 to Pedigree 2 and from Pedigree 2 to Pedigree 3 across different percentages of selected individuals of the studied population.

Moreover, the evaluation of accuracies dispersion (Figure 3) showed gains in magnitude of these parameters when correcting the relationship. Banos *et al.* (2001) reported that misidentifications in relationship modify the family means and also affects the estimation of inbreeding. Thus, the prediction of the genetic merit of candidates is also modified and can lead to breeding value estimates of under- or overestimated magnitudes. Similar results to our study were obtained by Munoz *et al.* (2014), who observed that pedigree corrections increase accuracy of plant breeding value estimates. These authors argued that relationship misidentifications lead to incorrect additive genetic variance estimates, which in turn decrease the accuracy of breeding values in genetic evaluations.

In general, it is expected that pedigree misidentifications would be more relevant when the phenotype under evaluation has low heritability, for example reproductive and disease resistance traits (Martinez *et al.* 2004; Cardoso *et al.* 2015). In such cases, individual performance is highly affected by environmental factors and reliable pedigrees are essential to account for phenotypic variation due to genetic factors and also to improve breeding values prediction and selection gain estimation. For traits with greater heritability, as growth and carcass traits in beef cattle (Peters *et al.* 2014; Kause *et al.* 2015), phenotypes are representative measures of genetic merit, and therefore, pedigree errors may have lower impact due to decreased importance of family information. Nonetheless, additional studies are warranted to precisely evaluate the consequences of pedigree corrections in breeding value prediction considering different heritability values and degrees of misidentifications for actual beef cattle breeding populations.

Cross-validation

We observed that the K-means cluster strategy provided lower accuracy than the random strategy. This was expected because the IBS dissimilarity matrix assumes higher values to those animals that share a large amount of identical alleles. Hence, the relationship achieved within clustering is greater than between clusters. Young candidates' breeding values are computed as a function of parent average breeding values and own performance. However, the cross-validation takes out own performance and evaluates only the utility of pedigree and relative's performance to predict the individual breeding value. Once in K-means strategy more related individuals are clustered together in the same group, family record tends to be jointly excluded from the reference population, decreasing accuracies of breeding value prediction in comparison with random grouping, where animals are clustered by chance regardless of their relatedness. Cardoso *et al.* (2015) obtained prediction accuracies for tick count varying from 0.35 to 0.39 when using K-means, and from 0.42 to 0.45 when using random clustering strategies. Taking into account these results, it is possible to infer about a trend of random clustering methods to report higher predictive ability values.

Ticks are responsible for losses in the production system by damaging the health of cattle. Thus, the identification of genotypes that are resistant to this ectoparasite should be considered a selection objective (Biegelmeyer *et al.* 2015), as well as its relationship with other traits of economic interest (Turner *et al.* 2010). The misidentification of relationships decreases the accuracy of the selection of superior genotypes, and so the importance of using SNP information to warrant correct genetic relationships becomes crucial during this process.

The corrected relationships were obtained using a medium density panel of 41 045 SNPs, because the dataset was originally generated for genomic selection and we took advantage of these data to demonstrate the impact of correcting genetic relationships on the estimation of breeding values. Even though such panel assured great precision to assess relationships, in practice, a much lower density panel would be required for parentage check purposes, for example the 100 SNP ISAG panel (Schütz & Brenig 2015). This would result in a much lower investment for genotyping of the parent–offspring pairs for pedigree correction purposes. The choice between using low-density panels that is just suitable for pedigree correction or higher densities that are also useful for genomic

prediction will depend on the relative accuracy of genomic and correct pedigree breeding value predictions, genotyping cost and economic value of the trait (s) of interest. In the case of tick resistance, this is a further research topic considering the challenge of ascertain the economic values for tick counts under different environmental challenges and breed compositions.

Conclusion

The use of SNP markers to correct pedigree information increases the reliability of parentage relationships and, consequently, improves the prediction accuracy of breeding values in genetic evaluations of beef cattle tick resistance.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

VSJ analysed the data and wrote the manuscript. FFC was involved in the statistical analyses, reviewed the manuscript and provided comments and suggestions. MMO provided scripts to deal with MolCoanc software. BPS reviewed the manuscript and was involved in parentage corrections. FFS provided comments and suggestions. PSL reviewed the manuscript, provided comments and suggestions.

Acknowledgements

The datasets used in the analyses were kindly provided by Conexão Delta G Breeding Program, Gensys Associate Consultants and Laboratory of Bioinformatics and Statistical Genomics of Embrapa South Livestock. Research partially supported by CNPq - National Council for Scientific and Technological Development grant 478992/2012-2, Embrapa - Brazilian Agricultural Research Corporation grant 01.11.07.002.07 and 02.13.14.014.00 and CAPES - Coordination for the Improvement of Higher Level Personnel. FFC, FFS and PSL are CNPq research fellows. The authors are also thankful to Dr. Jesús Fernández Martín, Departamento de Mejora Genética Animal, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid, Spain, for substantial help with MOLCOANC software analyses.

References

Aguilar I. (2014) SeekParentF90 (available at: http://nce.ads.uga.edu/wiki/doku.php?id=readme.seekparentf90; last accessed 30 October 2014).

Banos G., Wiggans G.R., Powell R.L. (2001) Impact of paternity errors in cow identification on genetic evaluations and international comparisons. *J. Dairy Sci.*, **84**, 2523–2529.

Biegelmeyer P., Nizoli L., da Silva S., dos Santos T., Dionello N., Gulias-Gomes C., Cardoso F. (2015) Bovine genetic resistance effects on biological traits of Rhipicephalus (Boophilus) microplus. *Vet. Parasit.*, **208**, 231–237.

Brito F.V., Neto J.B., Sargolzaei M., Cobuci J.A., Schenkel F.S. (2011) Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle. *BMC Genet.*, **12**, 80.

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

Cardoso F.F., Gomes C.C.G., Sollero B.P., Oliveira M.M., Roso V.M., Piccoli M.L., Higa R.H., Yokoo M.J., Caetano A.R., Aguilar I. (2015) Genomic prediction for tick resistance in Braford and Hereford cattle. *J. Anim. Sci.*, **93**, 2693–2705.

Fernandez J., Toro M.A. (2006) A new method to estimate relatedness from molecular markers. *Mol. Ecol.*, **15**, 1657–1667.

Geweke J. (1992) Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA.

Habier D., Tetens J., Seefried F., Lichtner P., Thaller G. (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet. Sel. Evol.*, **42**, 5.

Kause A., Mikkola L., Strandén I., Sirkko K. (2015) Genetic parameters for carcass weight, conformation and fat in five beef cattle breeds. *Animal*, **9**, 35–42.

Martinez G., Koch R., Cundiff L., Gregory K., Van Vleck L. (2004) Genetic parameters for six measures of length of productive life and three measures of lifetime production by 6 yr after first calving for Hereford cows. *J. Anim. Sci.*, **82**, 1912–1918.

Misztal I. (2008) Reliable computing in estimation of variance components. *J. Anim. Breed. Genet.*, **125**, 363–370.

Munoz P.R., Resende M.F.R., Huber D.A., Quesada T., Resende M.D.V., Neale D.B., Wegrzyn J.L., Kirst M., Peter G.F. (2014) Genomic relationship matrix for correcting pedigree errors in breeding populations: Impact on genetic parameters and genomic selection accuracy. *Crop Sci.*, **54**, 1115–1123.

Oliveira M.C.S., Alencar M.M., Giglioti R., Beraldo M.C.D., Aníbal F.F., Correia R.O., Boschini L., Chagas A.C.S., Bilhassi T.B., Oliveira H.N. (2013) Resistance of beef cattle of two genetic groups to ectoparasites and gastrointestinal nematodes in the state of São Paulo. *Brazil. Vet. Parasit.*, **197**, 168–175.

Peters S., Kizilkaya K., Garrick D., Fernando R., Pollak E., Enns R., De Donato M., Ajayi O., Imumorin I. (2014) Use of robust multivariate linear mixed models for estimation of genetic parameters for carcass traits in beef cattle. *J. Anim. Breed. Genet.*, **131**, 504–512.

Saatchi M., McClure M.C., McKay S.D., Rolf M.M., Kim J., Decker J.E., Taxis T.M., Chapple R.H., Ramey H.R., Northcutt S.L. (2011) Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. *Genet. Sel. Evol.*, **43**, 40.

Sanders K., Bennewitz J., Kalm E. (2006) Wrong and missing sire information affects genetic gain in the Angeln dairy cattle population. *J. Dairy Sci.*, **89**, 315–321.

Sargolzaei M., Chesnais J.P., Schenkel F.S. (2011) FImpute-An efficient imputation algorithm for dairy cattle populations. *J. Dairy Sci.*, **94**, 421.

Schütz E., Brenig B. (2015) Analytical and statistical consideration on the use of the ISAG-ICAR-SNP bovine panel for parentage control, using the Illumina BeadChip technology: example on the German Holstein population. *Genet. Sel. Evol.*, **47**, 3.

Spiegelhalter D.J., Best N.G., Carlin B.P., Van Der Linde A. (2002) Bayesian measures of model complexity and fit. *J. Roy. Stat. Soc.*, **64**, 583–639.

Stock K.F., Distl O., Hoeschele I. (2007) Bayesian estimation of genetic parameters for multivariate threshold and continuous phenotypes and molecular genetic data in simulated horse populations using Gibbs sampling. *BMC Genet.*, **8**, 1.

Turner L.B., Harrison B.E., Bunch R.J., Neto L.R.P., Li Y., Barendse W. (2010) A genome-wide association study of tick burden and milk composition in cattle. *Anim. Prod. Sci.*, **50**, 235–245.

Van Eenennaam A.L., Weigel K.A., Young A.E., Cleveland M.A., Dekkers J.C.M. (2014) Applied animal genomics: results from the field. *An. Rev. Anim. Bio.*, **2**, 105–139.

Van Vleck L.D. (1970a) Misidentification and sire evaluation. *J. Dairy Sci.*, **53**, 1697–1702.

Van Vleck L.D. (1970b) Misidentification in estimating the paternal sib correlation. *J. Dairy Sci.*, **53**, 1469–1474.

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

VanRaden P., Van Tassell C., Wiggans G., Sonstegard T., Schnabel R., Taylor J., Schenkel F. (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.*, **92**, 16–24.

Wiggans G.R., Sonstegard T.S., Vanraden P.M., Matukumalli L.K., Schnabel R.D., Taylor J.F., Schenkel F.S., Van Tassell C.P. (2009) Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. *J. Dairy Sci.*, **92**, 3431–3436.

Wiggans G.R., VanRaden P.M., Bacheller L.R., Tooker M.E., Hutchison J.L., Cooper T.A., Sonstegard T.S. (2010) Selection and management of DNA markers for use in genomic evaluation. *J. Dairy Sci.*, **93**, 2287–2292.

Wiggans G.R., Cooper T.A., VanRaden P.M., Olson K.M., Tooker M.E. (2012) Use of the Illumina Bovine3K BeadChip in dairy genomic evaluation. *J. Dairy Sci.*, **95**, 1552–1558.