Developing flexible models for genetic evaluations in smallholder crossbred dairy farms

R. Costilla, J. Zeng, M. Al Kalaldeh, M. Swaminathan, J.P. Gibson, V. Ducrocq, B.J. Hayes

Developing flexible models for genetic evaluations in smallholder crossbred dairy farms

R. Costilla,[1,2] ◉ J. Zeng,[3] M. Al Kalaldeh,[4] ◉ M. Swaminathan,[5] ◉ J. P. Gibson,[4] ◉ V. Ducrocq,[6] ◉ and B. J. Hayes[2]

[1]Agresearch Limited, Ruakura Research Centre, Hamilton 3214, New Zealand.
[2]Centre for Animal Science, Queensland Alliance for Agriculture and Food Innovation, University of Queensland, St Lucia, QLD 4067, Australia.
[3]Institute for Molecular Biosciences, University of Queensland, St Lucia, QLD 4067, Australia.
[4]Centre for Genetic Analysis and Applications, School of Environmental and Rural Science, University of New England, Armidale, NSW 2350, Australia.
[5]BAIF Development Research Foundation, Pune 412 202, Maharashtra, India.
[6]Universite Paris-Saclay, INRAE, AgroParisTech, UMR GABI, 78350 Jouy-en-Josas, France.

ABSTRACT

The productivity of smallholder dairy farms is very low in developing countries. Important genetic gains could be realized using genomic selection but genetic evaluations need to be tailored for lack of pedigree information and very small farm sizes. To accommodate this situation, we propose a flexible Bayesian model for the genetic evaluation of milk yield which allows us to simultaneously account for non-genetic random effects for farms and varying SNP variance (BayesR model). First, we use simulations based on real genotype data from Indian crossbred dairy cattle to demonstrate that the proposed model can separate the true genetic and non-genetic parameters even for small farm sizes (2 cows on average) although with high standard errors in scenarios with low heritability. The accuracy of genomic genetic evaluation increases until farm size is approximately 5. We then apply the model to real data from 4,655 crossbred cows with 106,109 monthly test day milk records and 689,750 autosomal SNPs. We estimate a heritability of 0.16 (0.04) for milk yield and using cross-validation, a genomic estimated breeding value (GEBV) accuracy of 0.45 and bias (regression of phenotype on GEBV) of 1.04 (0.26). Estimated genetic parameters are very similar using BayesR, BayesC and genomic BLUP approaches. Candidate genes nearby the top variants, *IMMP2L* and *ARHGEF2,* have been previously associated with milk protein composition, mastitis resistance and milk cholesterol content. The estimated heritability and GEBV accuracy for milk yield are much lower than those from intensive or pasture-based systems in many countries. Further increases in the number of phenotyped and genotyped animals in farms with at least 2 cows (preferably 3–5 to allow for dropout of cows) are needed to improve the estimation of genetic effects in these smallholder dairy farms.
Keywords: Milk yield, developing countries, BayesR, QTL, GWAS

INTRODUCTION

The productivity of smallholder dairy farms is very low in developing countries due to several reasons including genetics, environmental conditions, management, and government regulations. For instance, in India despite being the world's largest milk producer, the milk yield per cow is only one-eighth of the levels achieved in the United States and Canada (OECD et al., 2019) with predominantly low-input/low-yield dairy production systems (Hemme & Deeken, 2007; Morgan, 2009).

In principle, substantial genetic gains could be realized using new breeding technologies such as genomic selection (GS), especially in smallholder systems (Burrow et al., 2021). Recent experiences for crossbred dairy cattle in Africa (Brown et al., 2016; Mrode et al., 2021; Ojango et al., 2019) and India (Al Kalaldeh et al., 2021), show that genetic progress is possible with genomic selection even in these challenging conditions.

Appropriate modeling for genetic evaluations in smallholder systems is also challenging, as these systems typically have no pedigree information, weak genetic connectedness, and very small farm sizes. An important question is then, how well can we separate genetic from management effects in smallholder farms? Current efforts that use genomic BLUP (GBLUP) approaches in these contexts (Al Kalaldeh et al., 2021; Mrode et al., 2021; Ojango et al., 2019) might not be the most accurate approaches if there are some quantitative trait loci (QTL) of moderate to large effect (eg DGAT1 for fat%, Grisart et al. (2004)). For instance, the BayesR model has provided more accurate predic-

tions for several traits for dairy and tropically adapted beef cattle (Hayes et al., 2019; Kemper et al., 2015).

The question of whether such models can disentangle genetic effects from farm effects when farm sizes are very small has been investigated using simulation (Powell et al., 2021). The authors found that modeling farm effects as random had higher accuracies than modeling them as fixed at small farm sizes ($< = 4$ cows). However, the simulation approach of Powell et al. (2021) assumes known genetic parameters and can only approximate admixture levels in real crossbred populations.

Here we propose a flexible Bayesian model that allows us to simultaneously account for non-genetic random effects for farms and varying SNP variance, a modest extension of the previously described BayesR model (Erbe et al., 2012; Kemper et al., 2015; Moser et al., 2015). Additionally, BayesR directly produces a prediction equation that can be applied to any new animal as soon as it is genotyped, speeding up selection decisions from DNA sampling to breeding values.

To answer our research question, "how well can we separate genetic from management effects in smallholder farms"?, we use simulations based on real genotype data from crossbred dairy cows from India. We are particularly interested to examine whether the proposed model can separate the true genetic and non-genetic effects despite the small sample sizes in these contexts. Lastly, we apply the BayesR model to existing monthly test day milk records from Indian cows, compare it with the GBLUP and BayesC (Habier et al., 2011) approaches, and use cross-validation to calculate the accuracy and bias for the genomic estimated breeding values (GEBV).

Methods and Materials

Phenotypic and genotypic records.

The phenotypic and genotypic data was collected by the BAIF Development Research Foundation (BAIF -https://baif.org.in). BAIF is a non-government organization established in 1967 with a mission to enhance livelihoods of Indian rural families, using smallholder dairy production as one of the tools to achieve this objective. We used monthly test day (TD) milk records (liters/day) collected from Indian smallholder dairy farms. Existing records from 4,655 cows with 106,109 monthly test day milk records and 689,750 autosomal imputed SNPs (Al Kalaldeh et al., 2021) were used here for our analyses. Cows were crossbreds between local indigenous *Bos indicus* cattle and exotic dairy breeds, mainly Holstein/Friesian and Jersey (Strucken et al., 2021). The average breed composition of these

animals was estimated to be 0.48 Holstein/Friesian, 0.15 Jersey and 0.37 Indigenous (Al Kalaldeh et al., 2021). Milk yield records came from 6 Indian states: Bihar, Jharkhand, Maharashtra, Odisha, Punjab, and Uttar Pradesh between 2016 and 2020.

TD records were collected from crossbred cows raised in smallholder farms, ranging in size from 1 to 43 cows, with an average size of 1.7 animals per farm. TD records were corrected for fixed effects, including cattle development center (CDC), season, the interaction of CDC by season, parity, the lactation curves for parities with a 3rd order Legendre polynomial, and the lactation curves for CDC modeled with a 3rd order Legendre polynomial. Fixed effects for breed were obtained as the regressions on breed proportion. An estimate of the production environment that each cow experiences was obtained as the sum of the estimated fixed CDC effect and random farm effect of that cow. The environmental estimates were then ordered and classified into bottom, middle, and top thirds, creating a fixed effect of production environment with 3 levels, low, medium, and high environments. Further details about the trait collection, construction of corrected phenotypes, breed percentage and production environment can be found in Al Kalaldeh et al. (2021). Adjusted TD records are averaged by cow and used as a phenotype for the BayesR genetic model.

Simulations.

We performed simulations using real genotype data from BAIF from smallholder farms in India. For the genotypes, we randomly chose 50,000 genome-wide markers from the imputed HD genotypes for all available 4,655 crossbred cows (Al Kalaldeh et al., 2021). The phenotypes were generated using a model with 5,000 causal markers with varying SNP variance, farm sizes and heritabilities. QTL effects for the causal markers were sampled from a mixture of normal distributions:

$$0.80 \times N(0,10^{-4}) + 0.15 \times N(0,10^{-3}) + 0.04 \times N(0,10^{-2}) + 0.01 \times (0,10^{-1}).$$

The simulated phenotypes were generated using GCTA (Yang et al., 2011).

The mixture distribution for the SNP effects allows the estimation of QTL of varying effects, i.e. few markers with large effects and many markers with very small effect. Farm sizes were sampled from a truncated Poisson distribution with means equal to 1, 2, 5, 20 (average farm size). Farm effects followed a normal distribution with zero mean and a farm variance equal to three times the additive genetic variance. We set $\sigma_f^2 = 3\sigma_g^2$

based on the estimates obtained by Al Kalaldeh et al. (2021) and other studies in smallholder contexts where the farm/environmental variance is much larger than the additive genetic variance. We used three values for the heritabilities ($h^2 = 10\%, 20\%, 50\%$) for a total of twelve scenarios. For each scenario, we generated 50 simulated datasets, giving a total of 600 simulated datasets. Breed effects were not incorporated in the simulations.

Model estimation for the simulated data used MCMC chains of 5,000 iterations, 2,000 burn-in and 10 thinning. For each scenario, we present the estimated genetic parameters, additive genetic variance and heritabilities. Models also included weights for the residuals based on the real number of phenotypic records per animal (equation 2).

Validation for the simulations.

In addition to the estimation of genetic parameters, GEBVs from simulations were validated using prediction accuracies and bias from cross-validation (five-fold). Prediction accuracy was calculated as the Pearson linear correlation (r) between the simulated phenotype and GEBVs, and bias as the regression of the simulated phenotype on the GEBVs.

Real data application.

For the real data application, we fitted a BayesR model with random farm effects (model 1, equation 1), plus another one which adds production environment (Low, Medium, High) as fixed effect (model 2). Model parameters for the real data application, were estimated using 3 MCMC chains each with 25,000 iterations, 5,000 burn-in and 10 thinning. This means that $(25{,}000{-}5{,}000)/10 = 2{,}000$ iterations per MCMC chain were used for inference. MCMC convergence was assessed using the multivariate version of the Gelman-Rubin diagnostic (Gelman & Rubin, 1992). Chains were combined and visualized using the R-package coda (Plummer et al., 2006).

Functional annotation of markers with the highest posterior inclusion probability (PIP) in the BayesR model was carried using the Variant Effect Predictor tool from Ensembl, release 107 (Ensembl, 2022). We annotated the top 10 variants with the highest PIP. Candidate genes were also mapped in Ensembl using a 100 kb window from markers.

Validation for the real data application.

Validation for the real data application was carried out using cross-validation of multi-breed groups. The validation was performed across cattle development centres (CDC), which are made up of geographically close villages within Indian districts and states, with all animals within a CDC either within the reference set or validation set. This validation strategy was a ten-fold random cross validation where all 87 CDCs across India were partitioned in ten random groups (9 groups of 9 CDCs and 1 group of 6 CDCs). In turn, each group of CDCs was then taken as a validation set, the remaining as the reference set. SNP effects were calculated from the reference set and GEBV were calculated. Prediction accuracy was calculated as the Pearson linear correlation (r) of the GEBV and the phenotype adjusted for fixed effects. Bias was calculated as the regression of GEBV on adjusted phenotype. To compare with selection based on phenotypes only, we calculated the prediction accuracy of mass selection for repeated records using the equation $\sqrt{rh^2 / (1 + (r-1)t)}$, where r is the number of records per animal, and h^2 and t are the trait heritability and repeatability, respectively (Mrode, 2014). We used $r = 18$ (average number of records per animal in the Indian real data application), and $t = 0.60$, $h^2 = 0.19$ based on previous estimates for India (Al Kalaldeh et al. (2021), Table 3).

Estimation model.

Both for the simulations and real data application we used the same genetic model, a BayesR model (Erbe et al., 2012) with random farm effects,

$$y = 1_n' \mathrm{m} + Vf + Zg + e,$$

Where y is a vector of n adjusted phenotypes, 1_n is a vector of ones, m is an overall mean, f is a vector of p farm effects with distribution $f \sim N\left(0, \sigma_f^2\right)$, g is a vector of m SNP effects with distribution $g \sim N\left(0, \sigma_i^2\right)$. The variance for each SNP was assumed to be from one of four normal distributions $\sigma_i^2 = \left\{0, 10^{-4}, 10^{-3}, 10^{-2}\right\} \times \sigma_g^2$. This specification implies that the BayesR model assigns a mixture prior of normal distributions for the SNP effects so that every SNP can belong to any of these distributions. Being a generalization of BayesC (Habier et al., 2011), the first distribution is point-mass at zero to account for SNPs with no effect in the phenotype. V and Z are design matrices of farm effects (n × p) and standardized genotypes (n × m), respectively. Finally, e is a vector of random residuals with a distribution $e \sim N\left(0, E\sigma_e^2\right)$, where E is a diagonal matrix (1/w_i) and w_i is a weight for each animal. Weights were

calculated following the formula derived by Garrick et al. (2009) for cows,

$$w_i = \frac{r_i\left(1 - h^2\right)}{1 + \left(r_i - 1\right)t - r_i h^2},$$

Where r_i is the number of records for animal i, h^2 is the heritability of the trait (single records), and t the trait repeatability. Again, we set $t = 0.60$ and $h^2 = 0.19$ based on estimates for these genetic parameters in India (Al Kalaldeh et al. (2021), Table 3).

The BayesR model simultaneously provides estimates for the $\boldsymbol{m}$ SNP effects ($\boldsymbol{g}$), and the additive genetic ($\tilde{A}_g^2$), farm ($\tilde{A}_f^2$), and residual variances ($\tilde{A}_e^2$). Note that it can also accommodate fixed effects other than the overall mean m. The model was fitted using the open-source software GCTB (Zeng et al 2018) available at https://cnsgenomics.com/software/gctb. To compare the performance of the proposed Bayesian model, we also estimated GBLUP and BayesC models in Julia v1.6.2 (Bezanson et al., 2017) using the package JWAS (Cheng et al., 2018). All models were run in a Linux server (Intel Xeon 6130, 2.10GHz, and 314GB of RAM).

RESULTS

Simulations

The estimated genetic parameters for the simulations based on real genotypes are shown as Figure 1. The true values for these parameters are shown in dashed lines and the estimated values across simulations as boxplots. Estimates for the additive genetic variance $\left(\hat{\sigma}_g^2\right)$ are in Figure 1A and estimates for the heritability $\left(\hat{h}^2\right)$ in Figure 1B. Estimates for the additive genetic variances are very close to their true values (dash line) across most scenarios (Figure 1A), e.g., across all simulations the true values for the additive genetic variances are within the first and third quartiles (25th and 75th percentiles) of the distribution of estimated values. However, estimated additive genetic variances have lower accuracy for the scenarios with lower true heritability and smaller farm sizes. For instance, with an average farm size of one animal per farm and $h^2 = 10\%$, the values for $\hat{\sigma}_g^2$ are extremely noisy, sometimes including zero. The values for $\hat{\sigma}_g^2$ become more accurate with larger farm sizes and true heritabilities. For the scenario with 20 animals per farm on average, $\hat{\sigma}_g^2$ is very close to the true value, even for the lowest heritability scenario $\left(h^2 = 10\%\right)$.

Similarly, the estimated values for the heritabilities are very close to their true values (dash line) across all scenarios (Figure 1B). The accuracy of the estimates is lowest with small farm sizes and lower true heritability, with the extreme being the scenario with the lowest heritability and farm size ($h^2 = 10\%$, and one animal per farm). The values for $\hat{h}^2$ become more accurate with increasing average farm size and true heritability, with a greater impact of farm size. Importantly, an average farm size of 2 animals and a true heritability of 20% seems to be sufficient for an accurate estimation of the heritability. This is important because this scenario with very small number of animals per farm is typical in many smallholder systems.

Validation for the simulations

Prediction accuracy and bias for the simulations based on real genotypes are shown in Figure 2. The prediction accuracy of the GEBV increases with farm size and true heritability across all scenarios. However, this increase in accuracy is not linear and seems to plateau after an average farm size of five animals. For instance, prediction accuracies plateau around 0.18, 0.29 and 0.40 for true heritabilities of $h^2 = 10\%$, 20%, 50%, respectively. Given the sample size used for this simulation (4,655 cows) these relatively small values for the accuracies are not unexpected. For typical values of smallholder farms in the real data, a heritability of about 20% and an average farm size of around 2 animals, the expected prediction accuracy is close to 0.22 (Figure 2A).

The prediction bias is also smaller with increasing farm size and heritability, although the trend is less clear when $h^2 = 10\%$ (Figure 2B). That is, the higher the heritability and the number of cows in the farm, the closer the predicted values are to the simulated phenotypes. For the extreme scenario with the lowest heritability and farm size, $h^2 = 10\%$ and average farm size = 1, the GEBVs are deflated on average.

Real data application

The parameters estimated with all models, BayesR, BayesC and GBLUP, are shown in Table 1. For the baseline model, we estimate very similar additive genetic, farm and residual variances with all 3 models which result in heritabilities of 0.16 (0.03) for BayesR, 0.15 (0.02) for BayesC and 0.16 (0.02) for GBLUP.

For all models, including production environment (model 2) barely changes the estimates for the additive genetic and residual variances, and thus the heritabilities, but greatly reduces the farm variance. For

instance for the BayesR model, it reduces from 1.94 (0.07) to 1.25 (0.07). As expected, the effect of production environment ("low," "medium," "high") in the milk yield phenotype is statistically significant and increases monotonically. This effect is negative for animals in "low" environments and positive in animals in "medium" and "high" environments with the latter being the highest. Using the BayesR model, we estimate a

heritability of 0.16 (0.04) in a model that also includes production environment (Table 1).

The parameters of the BayesR model showed MCMC convergence and good mixing for all chains (Figure 3). The multivariate Gelman-Rubin diagnostic had a value of 1.02. Running times were about 11 h for the BayesR models, 13 h for the BayesC models, and 25 min for the GBLUP models. In terms of RAM memory, the BayesR
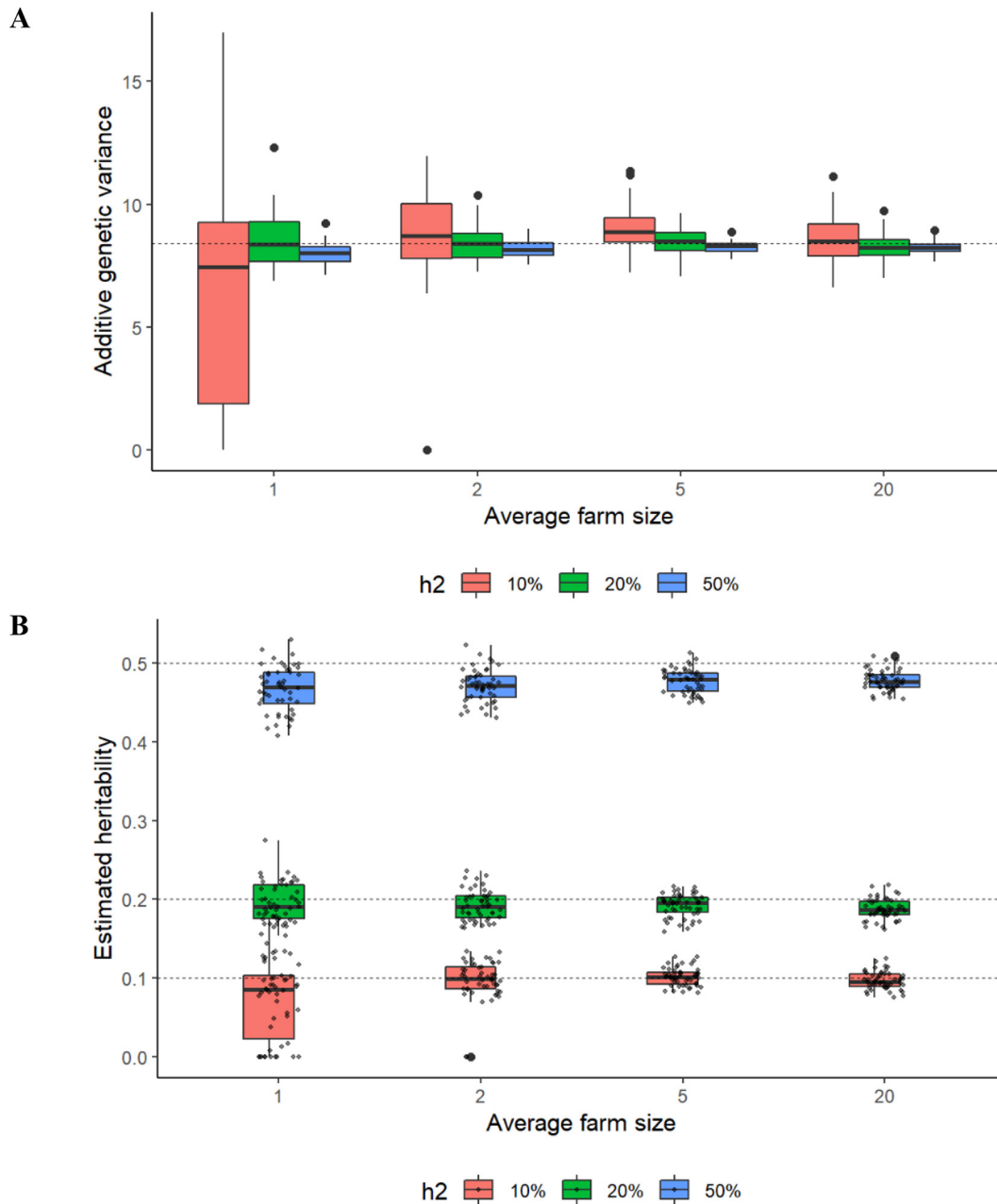

Figure 1. Estimated genetic parameters for simulated phenotypes based on real genotype data. Estimated additive genetic variance $\hat{\sigma}_g^2$ (A) and heritability $\hat{h}^2$ (B). True values are shown by dashed lines and estimated values as boxplots.
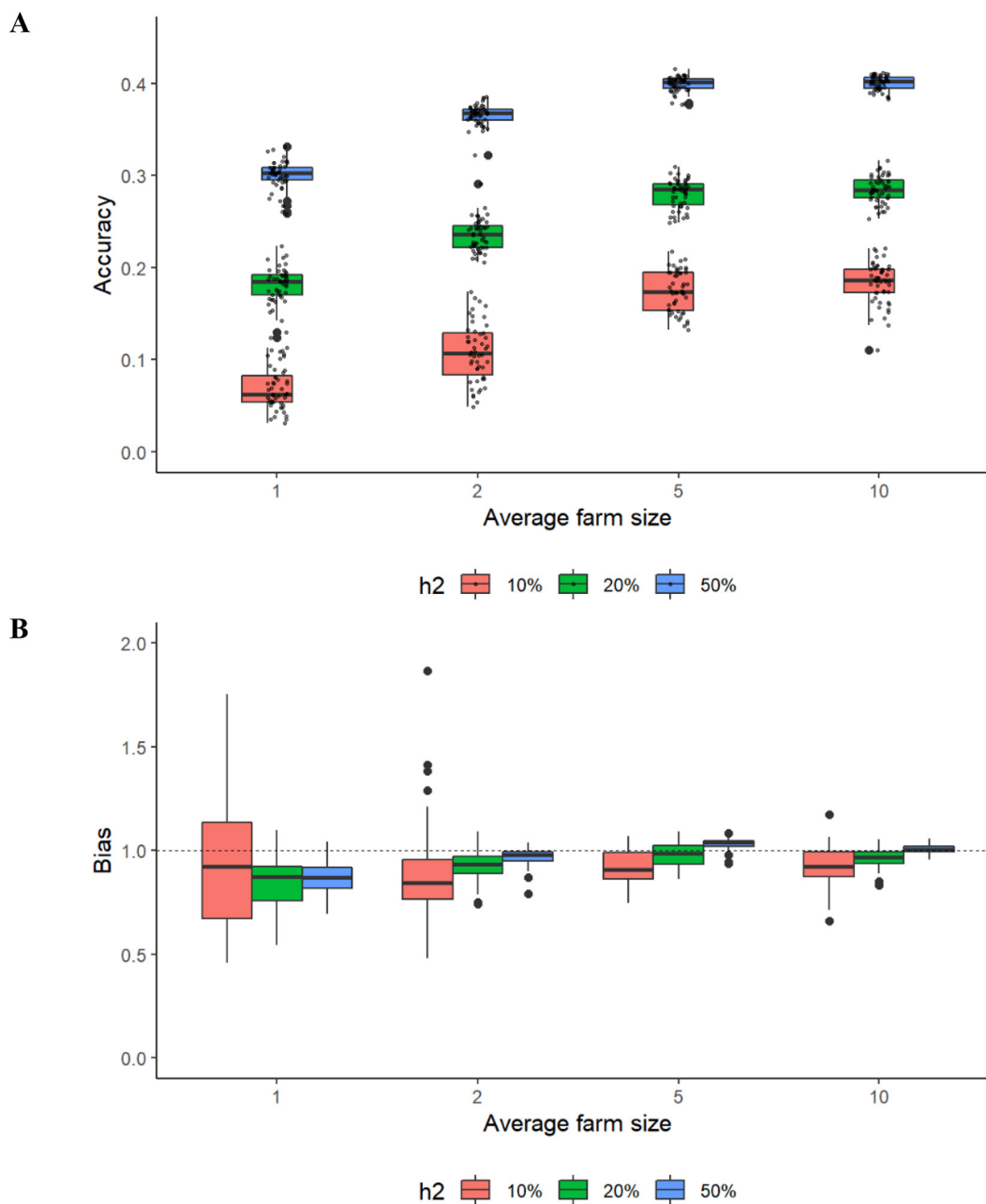
A

B

Figure 2. Prediction Accuracy (A) and bias (B) from 5-fold cross-validation for genomic prediction for simulated phenotypes based on real genotyped data. Accuracy is defined as the Pearson coefficient (r) and bias as the slope of the regression between the simulated phenotypes and GEBVs. Dashed line at 1, corresponds to unbiased GEBVs.

models in GCTB needed 13GB while the GBLUP and BayesC models used around 64GB.

We also tested breed effects in the genetic models but they were not significant after accounting for production environment and therefore not presented here. In what follows, we use the model with production environment (model 2) for QTL mapping and genomic prediction.

The variant with highest PIP, posterior inclusion probability, is located in chromosome 4 (rs109218186, minor allele frequency = 0.07) within 100 kb of the Mitochondrial Inner Membrane Protease Subunit 2 (*IMMP2L*) gene (Figure 4). This gene has been previously associated with milk protein composition (Dadousis et al., 2017) and mastitis resistance (Cai et al., 2018) in dairy cattle. The top variant in chromosome 1 (rs42222474, minor allele frequency = 0.15) is an intron

Table 1. Genetic, farm, and environment effects (SD) for the BayesR models for milk yield

Parameter	Baseline (model 1)			+ Environment (model 2)		
	BayesR	BayesC	GBLUP	BayesR	BayesC	GBLUP
Additive Genetic Variance σ_g^2	0.49 (0.08)	0.46 (0.06)	0.49 (0.07)	0.48 (0.10)	0.43 (0.06)	0.51 (0.07)
Residual Variance σ_e^2	2.66 (0.14)	2.68 (0.12)	2.64 (0.12)	2.55 (0.15)	2.60 (0.12)	2.52 (0.12)
Farm Variance σ_f^2	1.94 (0.07)	1.75 (0.09)	1.74 (0.09)	1.25 (0.07)	1.17 (0.07)	1.15 (0.07)
Heritability $h^2 = \sigma_g^2 / \left(\sigma_g^2 + \sigma_e^2 \right)$	0.16 (0.03)	0.15 (0.02)	0.16 (0.02)	0.16 (0.04)	0.14 (0.02)	0.17 (0.02)
Environment Low				−0.76 (0.05)	−0.75 (0.05)	−0.77 (0.05)
Environment Medium				0.84 (0.07)	0.84 (0.07)	0.84 (0.07)
Environment High				2.01 (0.08)	1.16 (0.08)	1.19 (0.08)

for Rho GTPase Activating Protein 26 (*ARHGEF26*) gene previously associated with milk cholesterol content (Do et al., 2018). This variant is the only intronic variant from all top 10 examined here.

Validation for the real data application

The results for the GEBV cross-validation for the real data application using random cattle development centres (CDCs), including the number of animals in each subset, prediction accuracy and bias, are presented in Table 2. The number of animals in these random groups varies from 316 to 842 with a mean of 466. On average, for the BayesR model with random farm effects, GEBV accuracy is 0.26 with a bias of 1.13 (0.19). When production environment (model 2) is incorporated, this prediction accuracy reduces to 0.18. Adjusted by heritability, we estimate a GEBV prediction accuracy of 0.45 (0.18 / $\sqrt{0.16}$) for the model with production environment as a fixed effect using this cross-validation across CDCs. By comparison, the prediction accuracy of mass selection was 0.55. In addition, prediction bias is the lowest for the model with environment 1.04 (0.26). On average across all CDCs, the confidence interval for the bias estimate includes 1, showing evidence that the GEBVs are unbiased on average.

DISCUSSION

We demonstrate by simulations that it is possible to separate genetic effects in smallholder farms using a flexible Bayesian model that simultaneously accounts for non-genetic random effects for farms and varying SNP variance. The model also provides a good way to produce interim GEBV for candidates or newly genotyped animals directly using estimated SNP effects, which provide a prediction equation. The proposed BayesR model is implemented in GCTB, a very efficient, freely available software.

The application to monthly test day milk records of 4,655 crossbred cows in India, with 106,109 records, 689,750 autosomal SNPs, and an average farm size of 1.7 animals per farm, yields an estimated heritability of 0.16 (0.04) in a model that also includes production environment. Using cross-validation, we estimated a GEBV prediction accuracy of 0.45, and a bias of 1.04 (0.26). These estimates are consistent with those obtained using a GBLUP approach, 0.42 in Al Kalaldeh et al. (2021), as well as the accuracy of mass selection (0.55). Genetic gains using genomics can therefore be substantial if GEBVs are used to shorten generation intervals. For instance, if the generation interval reduces by half when using GEBVs, and the other parameters are constant in the breeder's equation, the genetic gain will increase by 64%. When comparing our estimates with those found in other smallholder systems, we found that these accuracies are very similar to those obtained for milk yield in Kenya 0.32 to 0.41 (Brown et al., 2016) and Tanzania 0.53 to 0.59 (Mrode et al., 2021). These studies in Africa used smaller reference populations and milk yield records but also obtained smaller estimated heritabilities for this trait.

Our study has some limitations. First, we did not fit breed of origin (BOA) models (Vandenplas et al., 2016) since they only show small improvements in accuracy of estimates compared with models that ignore BOA (Eiriksson et al., 2022; Sevillano et al., 2017; VanRaden et al., 2020). In smallholder settings, crossbreeding is more complex than in intensive or pasture-based production systems. For example under Indian smallholder production, the crossbreeding program involves crossing of large numbers of indigenous cattle population, which are not categorized into specific breeds, with Indian dairy breeds (Sahiwal, Gir, Red Sindhi, Tharparkar, and Deoni) as well as with exotic dairy breeds, predominantly Holstein/Friesian and Jersey resulting in different levels of admixture in the crossbred animals (Strucken et al., 2021).
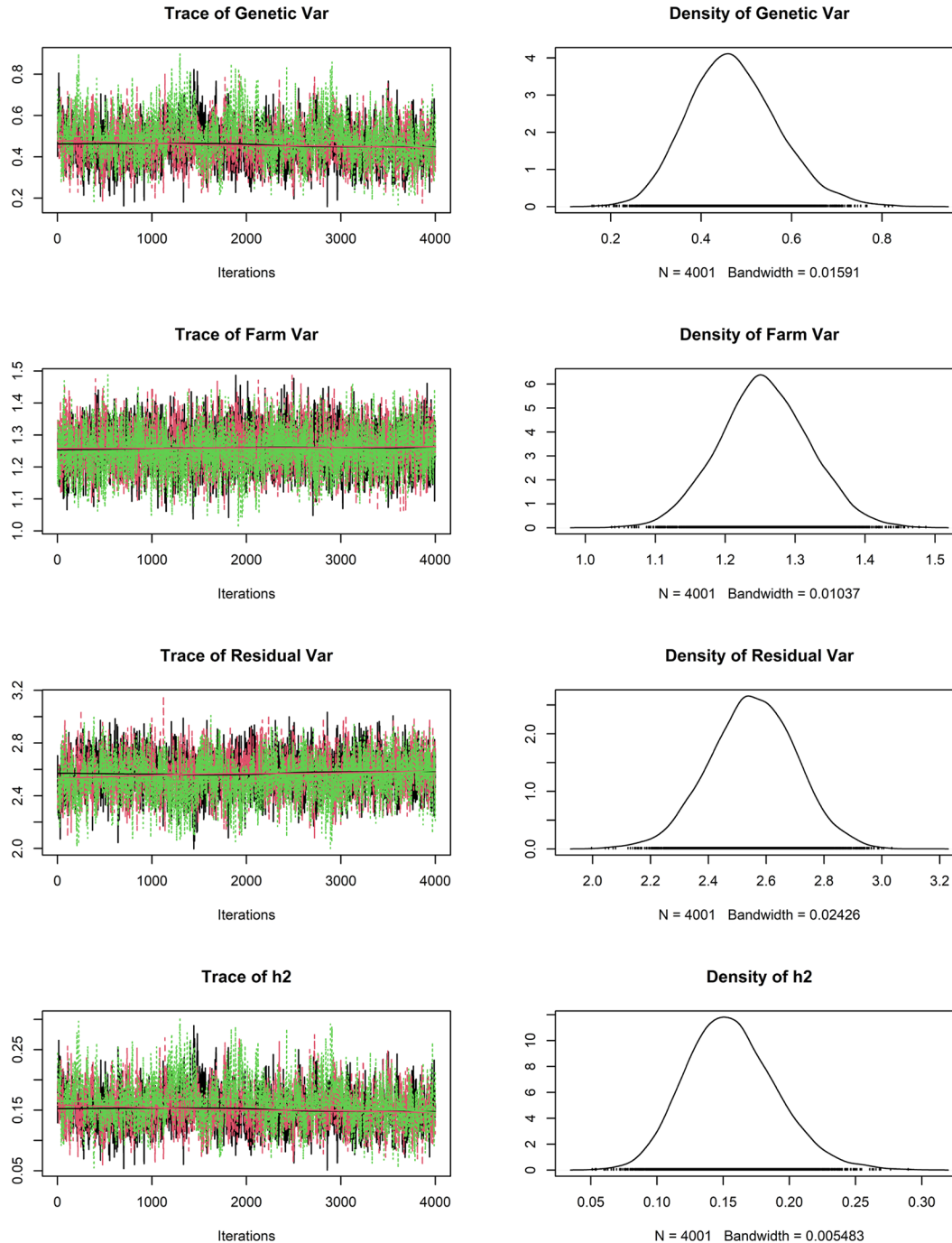
Figure 3. - MCMC convergence for all chains for the genetic parameters of the BayesR model with farm random effects and production environment (model 2).

Crossbreeding also increases the need for bigger reference populations since the inclusion of genetically divergent breeds can reduce prediction accuracy in genomic evaluations (Calus et al., 2014; Makgahlela et al., 2013) likely due to differences in causal variants and linkage disequilibrium patterns between markers and QTL across breeds. Furthermore, the results' similarity between the different approaches taken here, GBLUP, BayesC and BayesR also suggests that a much larger training population might be required for an accurate genetic evaluation in the presence of large environmental effects and crossbreeding and to exploit all the information provided by the high-density genotyping. One related limitation, and the reason why a forward-
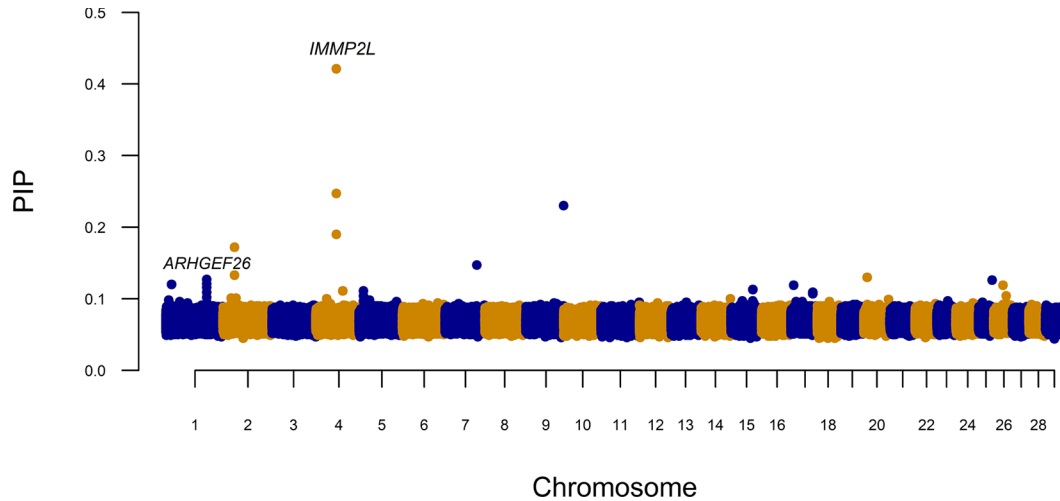
Figure 4. - Posterior inclusion probability (PIP) in the BayesR model with farm random effects and production environment (model 2). Marker in chromosome 1 is an intron for *ARHGEF26* and marker in chromosome 4 is within 100 kb of *IMMP2L*.

validation was not carried out, is that there were no new animals genotyped in 2020. In addition, there were only 402 animals genotyped in 2019 but they were from different geographical locations (farms, CDCs and States). We thus opted for the CDC cross-validation strategy presented here. As sample size grows, in terms of number of animals and records, this limitation will be alleviated.

Another potential limitation of BayesR, and other genetic models that rely on animals with both genotypic and phenotypic information, is that farmers with phenotypic records and (historical) pedigree information will have no GEBV. The problem is complex because single-step GBLUP (ssGBLUP) (Misztal et al., 2009) relies on accurate pedigree recording for several generations, which is not the case for smallholder systems and further parameter calibration to make both pedigree and genomic relationships compatible. Additionally,

SSGBLUP models have provided similar results to GBLUP in other smallholders' contexts (Mrode et al., 2021). There is therefore not perfect solution and we believe in continue the research efforts to develop genetic models that although imperfect could provide the best performance for a given context.

Typically, Bayesian models also require longer running times than the conventional GBLUP. For the Indian data set at hand (4655 animals and 689,750 SNPs), although the proposed BayesR model was 1.2x times faster than BayesC (11 vs 13 h), it was also 26.4x times slower than the conventional GBLUP (11 h vs 25 min). In terms of computer memory, the proposed BayesR model only needed about 20% of the RAM memory (13GB) required by the BayesC and GBLUP models (64GB). However, this lower RAM consumption might just reflect an efficient software implementation of BayesR in GCTB, when compared with a general-

Table 2. Cross-validation results by cattle development center (CDC) for the BayesR models

CDC group	Animals	Baseline (model 1)			+Environment (model 2)		
		Accuracy (r)	Bias	SE*	Accuracy (r)	Bias	SE*
1	399	0.14	0.51	0.18	0.07	0.36	0.24
2	427	0.18	0.58	0.15	0.10	0.46	0.21
3	412	0.23	1.08	0.23	0.21	1.64	0.37
4	465	0.30	1.31	0.19	0.19	0.98	0.23
5	566	0.22	0.84	0.16	0.12	0.62	0.21
6	533	0.37	1.74	0.19	0.34	1.79	0.21
7	316	0.33	1.44	0.24	0.22	1.28	0.32
8	842	0.37	2.13	0.19	0.30	2.10	0.23
9	324	0.29	1.08	0.20	0.16	0.77	0.26
10	371	0.17	0.60	0.18	0.08	0.45	0.29
Average	**466**	**0.26**	**1.13**	**0.19**	**0.18**	**1.04**	**0.26**

* Standard error for the bias.

purpose package like JWAS in Julia, rather than intrinsic differences between these genetic models.

Lastly, we might not be capturing environmental and management effects adequately for milk yield, overcorrecting in the model by using both farm and production environment and reducing the estimated heritability. However, we have identified 2 candidate genes, *IMMP2L* and *ARHGEF2*, and observed a prediction accuracy of 0.45 for this trait. These results are also consistent with those from a previous study using a GBLUP approach for the same data.

Taking these results together, BayesR provided similar genetic parameters and prediction accuracies but was computationally slower than GBLUP for the given Indian dairy smallholder data. Despite smaller accuracy of prediction than mass selection, both genetic models could provide higher genetic gain due to shorter generation intervals. In research settings, where running time and computational resources are not necessarily binding constraints, both models are good alternatives. However, in routine genetic evaluations time and computational resources could be important factors to consider when comparing the adoption of a genetic model. Currently, we are actively exploring alternative ways of incorporating environmental and breed effects in the genetic models for these smallholder dairy systems.

CONCLUSIONS

We demonstrate by simulations based on real genotypes that using the proposed Bayesian model it is possible to separate genetic effects in smallholder farms, even when farm sizes are small (2 cows on average). For the case study of milk yield from Indian smallholder farms, all 3 genetic models BayesR, BayesC and GBLUP provided similar genetic parameters. The heritability and GEBV accuracies for milk yield are much lower than those usually obtained in intensive or pasture-based systems. Further increases in the number of phenotype and genotyped animals in farm with at least 2 animals (preferably 3–5 to allow for dropout of cows) are needed to improve the accuracy of estimated genetic and farm effects in these smallholder dairy farms.

ACKNOWLEDGMENTS

REFERENCES

Al Kalaldeh, M., M. Swaminathan, Y. Gaundare, S. Joshi, H. Aliloo, E. M. Strucken, V. Ducrocq, and J. P. Gibson. 2021. Genomic evaluation of milk yield in a smallholder crossbred dairy production system in India. Genet. Sel. Evol. 53:73. https://doi.org/10.1186/s12711-021-00667-6.

Bezanson, J., A. Edelman, S. Karpinski, and V. B. Shah. 2017. Julia: A fresh approach to numerical computing. SIAM Rev. 59:65–98. https://doi.org/10.1137/141000671.

Brown, A., J. Ojango, J. Gibson, M. Coffey, M. Okeyo, and R. Mrode. 2016. Short communication: Genomic selection in a crossbred cattle population using data from the Dairy Genetics East Africa Project. J. Dairy Sci. 99:7308–7312. https://doi.org/10.3168/jds.2016-11083.

Burrow, H. M., R. Mrode, A. O. Mwai, M. P. Coffey, and B. J. Hayes. 2021. Challenges and Opportunities in Applying Genomic Selection to Ruminants Owned by Smallholder Farmers. Agriculture 11:1172. https://doi.org/10.3390/agriculture11111172.

Cai, Z., B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2018. Prioritizing candidate genes post-GWAS using multiple sources of data for mastitis resistance in dairy cattle. BMC Genomics 19:656. https://doi.org/10.1186/s12864-018-5050-x.

Calus, M., Huang, H., Wientjes, Y., Ten Napel, J., Bastiaansen, J., Price, M., Veerkamp, R., Vereijken, A., & Windig, J. (2014). (A) cross-breed genomic prediction.

Cheng, H., R. Fernando, and D. Garrick. 2018. JWAS: Julia implementation of whole-genome analysis software. Proceedings of the world congress on genetics applied to livestock production, Dadousis, C., Pegolo, S., Rosa, G. J. M., Bittante, G., & Cecchinato, A. (2017). Genome-wide association and pathway-based analysis using latent variables related to milk protein composition and cheesemaking traits in dairy cattle. J. Dairy Sci. 100:9085–9102. https://doi.org/10.3168/jds.2017-13219.

Do, D. N., F. S. Schenkel, F. Miglior, X. Zhao, and E. M. Ibeagha-Awemu. 2018. Genome wide association study identifies novel potential candidate genes for bovine milk cholesterol content. Sci. Rep. 8:13239. https://doi.org/10.1038/s41598-018-31427-0.

Eiriksson, J. H., K. Byskov, G. Su, J. R. Thomasen, and O. F. Christensen. 2022. Genomic predictions for crossbred dairy cows by combining solutions from purebred evaluation based on breed origin of alleles. J. Dairy Sci. 105:5178–5191. https://doi.org/10.3168/jds.2021-21644.

Ensembl. (2022). Retrieved 20/06/2022 from http://asia.ensembl.org/index.html

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 95:4114–4129. https://doi.org/10.3168/jds.2011-5019.

Garrick, D. J., J. F. Taylor, and R. L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41:1–8. https://doi.org/10.1186/1297-9686-41-55.

Gelman, A., and D. B. Rubin. 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7:457–472. https://doi.org/10.1214/ss/1177011136.

Grisart, B., F. Farnir, L. Karim, N. Cambisano, J.-J. Kim, A. Kvasz, M. Mni, P. Simon, J.-M. Frere, W. Coppieters, and M. Georges. 2004. Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. Proc. Natl. Acad. Sci. USA 101:2398–2403. https://doi.org/10.1073/pnas.0308518100.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12:1–12. https://doi.org/10.1186/1471-2105-12-186.

Hayes, B. J., N. J. Corbet, J. M. Allen, A. R. Laing, G. Fordyce, R. Lyons, M. R. McGowan, and B. M. Burns. 2019. Towards multibreed genomic evaluations for female fertility of tropical beef cattle. J. Anim. Sci. 97:55–62. https://doi.org/10.1093/jas/sky417.

Hemme, T., and E. Deeken. (2007). IFCN dairy report 2007. *IFCN Centre, Braunschweig, Germany.*

Kemper, K. E., C. M. Reich, P. J. Bowman, C. J. Vander Jagt, A. J. Chamberlain, B. A. Mason, B. J. Hayes, and M. E. Goddard. 2015. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. Genet. Sel. Evol. 47:29. https://doi.org/10.1186/s12711-014-0074-4.

Makgahlela, M., I. Strandén, U. Nielsen, M. Sillanpää, and E. Mäntysaari. 2013. The estimation of genomic relationships using breed-wise allele frequencies among animals in multibreed populations. J. Dairy Sci. 96:5364–5375. https://doi.org/10.3168/jds.2012-6523.

Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92:4648–4655. https://doi.org/10.3168/jds.2009-2064.

Morgan, N. (2009). Smallholder dairy development: Lessons learned in Asia. *RAP PUBLICATION 2009/02.* http://www.fao.org/fileadmin/user_upload/sap/docs/i0588e00.pdf

Moser, G., S. Lee, B. Hayes, M. Goddard, N. Wray, and P. Visscher. 2015. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLoS Genet. 11:e1004969. https://doi.org/10.1371/journal.pgen.1004969.

Mrode, R., J. Ojango, C. Ekine-Dzivenu, H. Aliloo, J. Gibson, and M. A. Okeyo. 2021. Genomic prediction of crossbred dairy cattle in Tanzania: A route to productivity gains in smallholder dairy systems. J. Dairy Sci. 104:11779–11789. https://doi.org/10.3168/jds.2020-20052.

Mrode, R. A. (2014). *Linear models for the prediction of animal breeding values* (3rd ed.). CABI.

OECD. Food, & Nations, A. O. o. t. U. (2019). *OECD-FAO Agricultural Outlook 2019–2028.* https://doi.org/doi:https://doi.org/10.1787/agr_outlook-2019-en

Ojango, J. M. K., R. Mrode, J. E. O. Rege, D. Mujibi, E. M. Strucken, J. Gibson, and O. Mwai. 2019. Genetic evaluation of test-day milk yields from smallholder dairy production systems in Kenya using genomic relationships. J. Dairy Sci. 102:5266–5278. https://doi.org/10.3168/jds.2018-15807.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. CODA: convergence diagnosis and output analysis for MCMC. R News 6:7–11.

Powell, O., R. Mrode, R. C. Gaynor, M. Johnsson, G. Gorjanc, and J. M. Hickey. 2021. Genomic evaluations using data recorded on smallholder dairy farms in low- to middle-income countries. JDS Commun. 2:366–370. https://doi.org/10.3168/jdsc.2021-0092.

Sevillano, C. A., J. Vandenplas, J. W. M. Bastiaansen, R. Bergsma, and M. P. L. Calus. 2017. Genomic evaluation for a three-way crossbreeding system considering breed-of-origin of alleles. Genet. Sel. Evol. 49:75. https://doi.org/10.1186/s12711-017-0350-1.

Strucken, E. M., N. Z. Gebrehiwot, M. Swaminathan, S. Joshi, M. Al Kalaldeh, and J. P. Gibson. 2021. Genetic diversity and effective population sizes of thirteen Indian cattle breeds. Genet. Sel. Evol. 53:47. https://doi.org/10.1186/s12711-021-00640-3.

Vandenplas, J., M. P. Calus, C. A. Sevillano, J. J. Windig, and J. W. Bastiaansen. 2016. Assigning breed origin to alleles in crossbred animals. Genet. Sel. Evol. 48:61. https://doi.org/10.1186/s12711-016-0240-y.

VanRaden, P. M., M. E. Tooker, T. C. S. Chud, H. D. Norman, J. H. Megonigal Jr., I. W. Haagen, and G. R. Wiggans. 2020. Genomic predictions for crossbred dairy cattle. J. Dairy Sci. 103:1620–1631. https://doi.org/10.3168/jds.2019-16634.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher. 2011. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88:76–82. https://doi.org/10.1016/j.ajhg.2010.11.011.

ORCIDS

R. Costilla ● https://orcid.org/0000-0003-0818-5065
M. Al Kalaldeh ● https://orcid.org/0000-0002-3206-6421
M. Swaminathan ● https://orcid.org/0000-0002-7544-7777
J. P. Gibson ● https://orcid.org/0000-0003-0371-2401
V. Ducrocq ● https://orcid.org/0000-0002-1503-5199