Short communication: Genomic selection in a crossbred cattle population using data from the Dairy Genetics East Africa Project

A. Brown,[*] J. Ojango,[†] J. Gibson[+], M. Coffey,[*] M. Okeyo,[†] R Mrode[*†1]

* Animal & Veterinary Sciences

Scotland's Rural College

Easter Bush, Midlothian EH25 9RG, Scotland, United Kingdom

† International Livestock Research Institute (ILRI)

Box 30709, Nairobi, Kenya

[+]University of New England

Armidale, NSW 2351, Australia

[1] Corresponding author: raphael.mrode@sruc.ac.uk

ABSTRACT

Due to the absence of accurate pedigree information, it has not been possible to implement genetic evaluations for crossbred cattle in African small-holder systems. Genomic selection techniques that do not rely on pedigree information could, therefore, be a useful alternative. The objective of this study was to examine the feasibility of using genomic selection techniques in a crossbred cattle population using data from Kenya provided by the Dairy Genetics East Africa Project. Genomic estimated breeding values for milk yield were estimated using 2 prediction methods,

GBLUP and BayesC, and accuracies were calculated as the correlation between yield deviations and genomic breeding values included in the estimation process, mimicking the situation for young bulls. The accuracy of evaluation ranged from 0.28 to 0.41, depending on the validation population and prediction method used. No significant differences were found in accuracy between the 2 prediction methods. The results suggest that there is potential for implementing genomic selection for young bulls in crossbred small-holder cattle populations, and targeted genotyping and phenotyping should be pursued to facilitate this.

SHORT COMMUNICATION

Genomic selection is now widely used in the dairy industry, with genomic estimated breeding values (**GEBV**) now being commercially produced for several breeds worldwide, as part of routine genetic evaluations. However, the majority of these evaluation schemes are carried out in developed countries, where most animals evaluated are purebred, and have large volumes of phenotype, genotype, and pedigree data. In developing countries, such as those in Eastern Africa, a large proportion of dairy production is carried out by small holders, who in many cases keep fewer than 10 cattle. These cattle are mostly crosses between indigenous African breeds and exotic dairy breeds, and have little phenotypic or pedigree data available. It has, therefore, not been possible to implement conventional genetic evaluation methods in these populations. As a result, bulls cannot currently be effectively ranked for genetic progress, preventing effective genetic improvement. If the level of phenotypic recording can be increased, and sufficient funding is available to cover the costs of genotyping, genomic selection may be a suitable tool for estimation of breeding values in these crossbred cattle.

Several studies have highlighted the potential for crossbred genomic evaluations using a training population made up of crossbred animals (Ibánez-Escriche et al., 2009; Toosi et al., 2010; Mucha et al., 2015; Vanraden and Cooper, 2015).

Earlier studies, such as those by Ibáñez-Escriche et al. (2009) and Toosi et al. (2010), used simulated data to investigate the potential for using a crossbred reference population to estimate breeding values of purebred animals for the performance of their crossbred offspring. Results suggested that there is potential for using crossbred reference populations to predict GEBV in purebreds, with no necessity to use complex models to assign breed-specific allele frequencies. More recently, VanRaden et al. (2015) used empirical data to show that genomic-predicted transmitting abilities can be computed for crossbred animals by applying purebred marker effects that have been weighted by the crossbred animal's genomic breed composition. In a study involving UK dairy goats, Mucha et al. (2015) computed milk yield GEBV for crossbred goats using a crossbred training population. The results suggested that there was no additional benefit to using SNP-BLUP to estimate breeding values, compared with pedigree-based BLUP, but higher accuracies were achieved when the single step method was implemented.

The above studies used a range of statistical methods for prediction of GEBV, with Ibánez-Escriche et al. (2009) and Toosi et al. (2010) using Bayesian methods of prediction, whereas Mucha et al. (2015) implemented SNP-BLUP and single step approaches. Simulation studies have suggested that Bayesian methods have a slight advantage over GBLUP methods for genomic prediction (Hayes et al., 2009); however, the methods have not been compared using real-world data in the analysis of dairy traits.

This study aims to investigate the feasibility of using genomic selection in a small population of African crossbred cattle, using 2 statistical methods, GBLUP and BayesC. The method of assessing achieved accuracy mimics the situation of young bulls.

The data set consisted of genotype data for 1,013 cows aged 4 to 8 yr, from the Kenyan component of the Dairy Genetics East Africa Project (Ojango et al., 2014, Gibson et al. 2014) Animals consisted of varying crosses between indigenous African breeds (N'dama–*Bos taurus*, and Nellore–*Bos indicus*) and 5 exotic dairy breeds (Ayrshire, Friesian, Holstein, Guernsey, Jersey). All individuals were genotyped using the Illumina BovineHD BeadChip (Illumina, San Diego, CA). Genotype data were edited by loci; SNP with a minor allele frequency of <0.05, a call rate of <0.95, or with no chromosomal position, were removed, along with those that were detected as not being in Hardy-Weinberg equilibrium and SNP on the X chromosome. After applying these filters, 665,408 autosomal SNP were available for analysis.

The phenotypes used were milk yield deviations (**YD**). These were computed from a fixed test-day model using test-day records for the first 3 lactations with management group, year-month of test, parity, and dairy group by breed interaction fitted as fixed effects. In addition, fixed lactation curves of Legendre polynomials of order 4, nested within dairy group by breed interaction, were fitted to account for crossbreeding effects in the model (J. Ojango, unpublished data). Random effects of animal and permanent environment were also included in the model. The YD were averaged by cow and the corresponding weight for YD for each cow used in the genomic analysis

was computed as the inverse of the standard error. The heritability of milk yield based on this model was 0.30.

A genomic relationship matrix was computed for all animals using VanRaden's first definition of **G** (VanRaden, 2008), where $\mathbf{G} = \frac{\mathbf{ZZ'}}{2\sum p_i(1-p_i)}$, where **Z** is a design matrix of centered genotypes, and p_i is the allele frequency estimated across breeds for the major allele at SNP *i*. Principal components analysis (**PCA**) was then carried out on the **G** matrix using the R function "princomp" (R Core Team, 2013), to investigate the genomic relationships between individuals.

Figure 1 illustrates the results of the PCA. Although there is no distinct separation between clusters of animals, the first principal component divides the animals into 5 groups based on the proportion of their genetics that is contributed by exotic dairy breeds, their so-called percentage exotic breeds. Due to this clustering, we chose to investigate how well GEBV for animals with the highest and lowest percentage exotic breeds could be estimated using the remainder of the population. Two groups with higher percentage exotic breeds were chosen for validation: (1) animals with percentage exotic breeds above 87.5%, and (2) animals with 60 to 87.5% exotic breeds. However, the number of animals with a low percentage of exotic breeds was too low to create a third validation population based purely on this category. The data were therefore re-organized into 6 categories, with each category defined by the combination of exotic breeds that contributed most of the exotic genes to the cross. These categories were (a) Ayrshires; (b) Friesians; (c) Ayrshires and Friesians; (d) Guernseys and Friesians; (e) Ayrshires, Friesians, and Guernsey; and (f) mixed exotic. For animals in category f, the exotic genes came from more than 3 exotic

breeds (average percentage exotic breeds was approximately 46%), with indigenous breeds contributing ≥40% of genetics in most cows. To represent animals with mainly indigenous genetics, a third validation group was created using animals from category f. Figure 2 shows the same PCA with animals labeled according to the 6 categories described above. Summary statistics for the 3 validation groups are shown in Table 1.

Figure 1 Principal components 1 and 2 based on the analysis of the genomic relationship matrix of 1,013 crossbred cows. Animals are labelled according to the percentage of their genetics contributed by exotic dairy breeds.

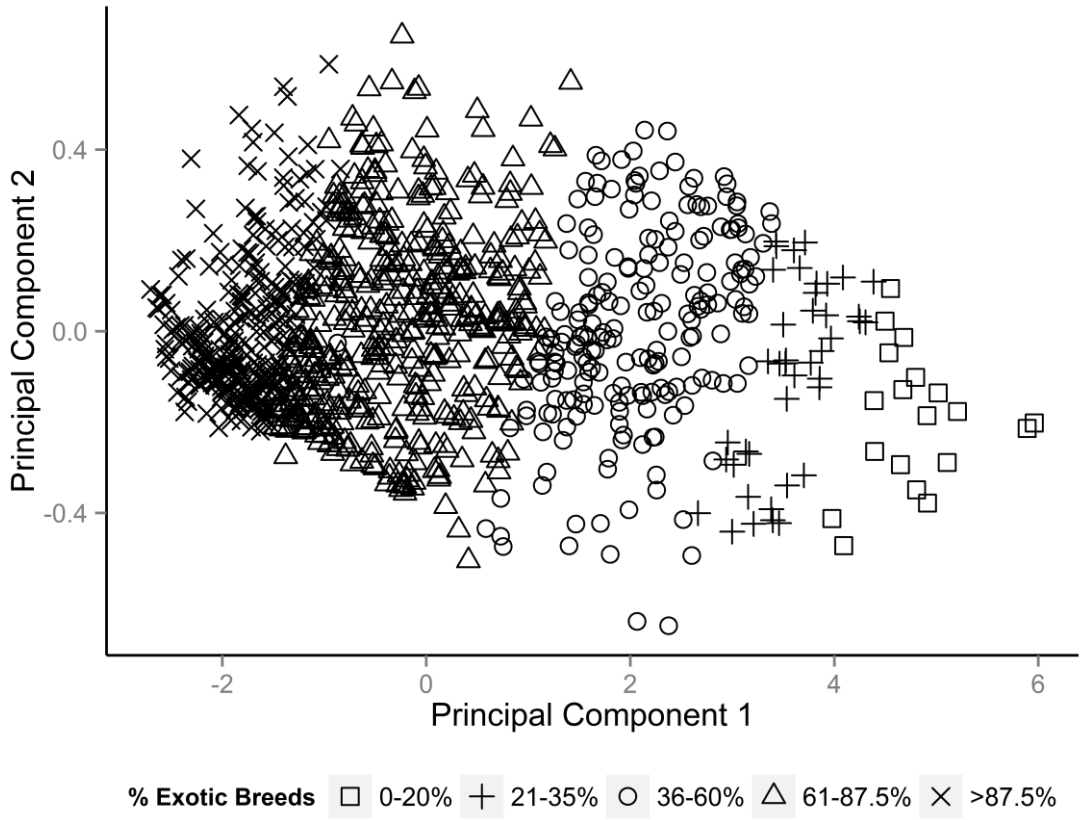

Figure 2 Principal components 1 and 2 based on the analysis of the genomic relationship matrix of 1,013 crossbred cows. Animals are split into 6 categories, with each category defined by the number of exotic breeds that contributed most of the exotic genes to the cross. a) Ayrshires, b) Friesians, c) Ayrshires and Friesians, d) Guernseys and Friesians e) Ayrshires, Friesians and Guernsey and f) Mixed exotic.

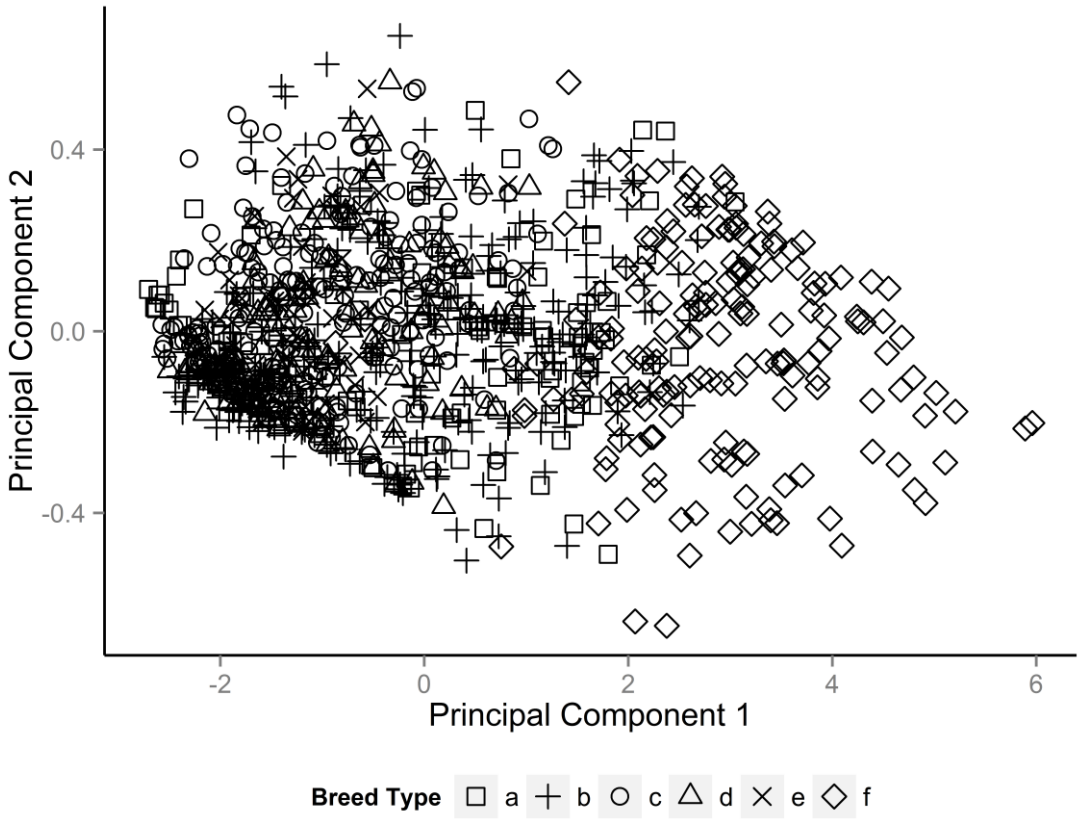

Table 1 Summary statistics for each of the three groups chosen for GEBV estimation and validation.

Validation group	Description	N	Mean yield deviation (s.d)	Range
1	>87.5% exotic	297	0.39 (1.60)	-2.34 – 7.75
2	61-87.5% exotic	448	0.00 (1.34)	-3.41 – 7.32
3	33-50% exotic	178	-0.61 (1.08)	-2.87 – 3.45

Two statistical models were used to compare their performance, GBLUP and BayesC. The model for the GBLUP analysis was $\mathbf{y} = \boldsymbol{\mu} + \mathbf{Zg} + \mathbf{e}$, where $\mathbf{y}$ is the vector of the weighted YD, $\boldsymbol{\mu}$ is the overall mean, $\mathbf{Z}$ is an incidence matrix relating individuals to records, $\mathbf{g}$ is a vector of random animal effects with an assumed distribution of $N(0,\sigma^2_g\mathbf{G})$, where σ^2_g is the additive genomic variance and $\mathbf{G}$ is the genomic relationship matrix calculated as detailed above, and $\mathbf{e}$ is a vector of residual

effects with an assumed distribution of $N(0,\sigma^2_e\mathbf{I})$, where σ^2_e is the residual variance and **I** is an identity matrix. The software package Mix99 (Lidauer and Strandén, 1999) was used for GBLUP analysis.

The BayesC method used the same basic model as detailed above, but in this case vector **g** is defined as $\sum_{i=1}^{N}(z_i a_i I_i)$, where z_i is the genotype at SNP *i*, a_i is the effect of SNP*i*, and I_i is an indicator variable which is set to 1 if the *i*th SNP has an effect on the trait of interest, or 0 if it has no effect. The distribution of a_i is assumed $N(0, \sigma^2_a)$, where σ^2_a is the SNP variance. The SNP effects were assumed to be normally distributed, and variable *I* was assumed to be binomially distributed with probability π. Previous analyses using a BayesCπ model suggested a value of π of 0.23, and so in this study the value of π was set to 0.3. A custom written Fortran program following the method by Mrode (2014) was used for BayesC analysis.

The accuracy of prediction was calculated for both methods as the correlation between the YD and the GEBV within each of the 3 validation populations described above. In each case, the reference population comprised all animals in the data set that were not chosen for the validation. Reference and validation population sizes for each analysis are shown in Table 2.

Table 2 Accuracies of GEBV based on GBLUP and BayesC models, for each of three validation groups; 1) animals with percentage exotic breeds above 87.5%, 2) animals with 60 - 87.5% exotic breeds, and 3) animals with predominantly indigenous genetics.

Validation Group	$N_{validation}$	$N_{reference}$	Accuracy (s.e)	
			GBLUP	BayesC
1	297	716	0.41 (0.04)	0.39 (0.05)
2	448	565	0.35 (0.04)	0.35 (0.04)
3	178	835	0.32 (0.06)	0.28 (0.06)

Accuracies of prediction ranged from 0.28 to 0.41 dependent upon the validation group and the statistical method used (Table 2). The highest accuracies were observed for animals with percentage exotic breeds of above 87.5% (group 1), and lower accuracies for the mixed exotic group (group 3). In general, the validation accuracies reported for milk yield are much lower than observed in developed countries (Hayes et al., 2009). Differences in size and type of data could be considered as major factors in this difference. However, this analysis provides the first estimates of genetic merit for this population and is, therefore, valuable for identifying extreme animals and selecting teams of young bulls that can be used for breeding. Small sample size in group 3 may be a factor contributing to a lower accuracy, as a small number of badly performing individuals can have a large effect on the overall accuracy. However, differences between accuracies achieved in the 3 validation groups were tested for significance using Fisher's r to z transformation, with no significant difference in accuracy between the 3 groups observed for either method of prediction ($P = 0.19$ to $P = 0.70$). Fisher's r to z transformation was also used to test the comparative performance of GBLUP and BayesC; no significant differences were found in performance between the 2 methods ($P = 0.68$ to $P = 1$).

It was particularly interesting that the BayesC method did not perform significantly better than the GBLUP model, as previous studies have suggested that Bayesian models should predict genomic breeding values with a higher accuracy than GBLUP (Hayes et al., 2009). Bayesian methods of prediction require more computational time and greater computational power to run than GBLUP-based methods. Due to this difference in running time, GBLUP methods are often preferred in commercial situations; Bayesian methods must, therefore, produce substantially higher accuracies

of prediction than GBLUP for the increased computational time to be worthwhile. As such, we suggest that the GBLUP model is more suitable for commercial evaluations of polygenic traits, such as milk yield, in crossbred populations. However, considering that Bayesian methods of prediction are expected to perform better for traits controlled by a small number of genes of large effect (Hayes et al., 2009), we suggest that Bayesian models should still be considered when implementing evaluations for less polygenic traits.

The accuracies obtained in this study are similar to those reported by Mucha et al. (2015), who estimated GEBV for milk yield in a UK population of dairy goats. In the study by Mucha et al. (2015), the SNP-BLUP model did not outperform pedigree-based BLUP, and to see any benefit of implementing genomic selection, the authors had to incorporate further data using the single step method. We are unable to implement pedigree-based evaluation methods in this population of cattle; as such, we are comparing our predictions to a baseline accuracy of zero. The results presented above are, therefore, extremely positive, and provide an opportunity for undertaking selection and consequently increasing the rate of genetic progress within this population. This study used high-density genotypes to capture as much genetic variation as possible within this crossbred population; however, it is unlikely that genomic selection will be implemented commercially using this chip due to the costs associated with high density genotyping. Work is currently on-going to develop a lower density chip that is suitable for use in the wider African small holder cattle population. As indicated earlier, the prediction of genomic merit in this study provides an opportunity for the selection of teams of young bulls for breeding, and will also help to identify extreme animals. It therefore provides the incentive for

more targeted recording schemes that will allow the collection of more phenotypic data, with the aim of improving the accuracy achieved by increasing the size of the reference population. Innovative ways of giving timely and targeted feedbacks to farmers, based on such data, would help to support data collection and should be pursued.

ACKNOWLEDGEMENTS

The authors thank the Bill and Melinda Gates Foundation for funding the Dairy Genetics East Africa (DGEA) project. A. Brown also acknowledges the Biotechnology and Biological Sciences Research Council (Swindon, UK) and the Knowledge Transfer Network (London, UK) for funding.

3.1 Conclusion

This study demonstrates that there is potential for applying genomic evaluation techniques in crossbred cattle populations, but, as in the previous chapter, more data is needed to validate the work. As in chapter 2, this chapter has focussed on using high-density genotype data for genomic evaluations. Whole-genome sequence data is now being generated for multiple breeds of cattle, and so the final experimental chapter will focus on whether there is an advantage to using SNPs extracted from sequence data for across-breed genomic evaluations.

3.2 References

Gibson, J., Ojango, J., and Mwai, O. (2014) Dairy Genetics East Africa (DGEA) Phase 2 - The final narrative report to the Bill and Melinda Gates Foundation. (Unpublished)

Hayes, B.J., P.J. Bowman, a J. Chamberlain, and M.E. Goddard. 2009. Invited

review: Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* 92:433–443. doi:10.3168/jds.2008-1646.

Ibánez-Escriche, N., R.L. Fernando, A. Toosi, and J.C.M. Dekkers. 2009. Genomic selection of purebreds for crossbred performance. *Genet. Sel. Evol.* 41:12. doi:10.1186/1297-9686-41-12.

Lidauer, M., and I. Strandén. 1999. Fast and flexible program for genetic evaluation in dairy cattle. *Interbull Bull.* 20:19–24.

Mrode, R.. 2014. Linear Models for the Prediction of Animal Breeding Values. 3rd ed. CABI. 199-200 pp.

Mucha, S., R. Mrode, I. MacLaren-Lee, M. Coffey, and J. Conington. 2015. Estimation of genomic breeding values for milk yield in UK dairy goats. *J. Dairy Sci.* 98:8201–8208. doi:10.3168/jds.2015-9682.

Ojango, J.M.K., A. Marete, D. Mujibi, J. Rao, J. Pool, J.E.O. Rege, C. Gondro, W.M.S.P. Weerasinghe, J.P. Gibson, and A.M. Okeyo. 2014. A novel use of high density SNP assays to optimize choice of different crossbred dairy cattle genotypes in small-holder systems in East Africa. *Proc. 10th World Congr. Genet. Appl. to Livest. Prod.* 2–4.

R Core Team. 2013. R Core Team. *R A Lang. Environ. Stat. Comput. R Found. Stat. Comput. Vienna, Austria.* ISBN 3–900051–07–0, URL http://www.R–project.org/.

Toosi, a., R.L. Fernando, and J.C.M. Dekkers. 2010. Genomic selection in admixed and crossbred populations. *J. Anim. Sci.* 88:32–46. doi:10.2527/jas.2009-1975.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91:4414–4423. doi:10.3168/jds.2007-0980.

Vanraden, P.M., and T.A. Cooper. 2015. Genomic Evaluations and Breed

Composition for Crossbred U . S . Dairy Cattle. 9–13.