

Can genomic data enable genetic evaluation with phenotypes recorded on smallholder farms?

Owen Powell¹, R. Chris Gaynor¹, Janez Jenko¹, Gregor Gorjanc¹, Raphael Mrode^{2,3}, & John M. Hickey¹

¹The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Research Centre, Midlothian EH25 9RG, UK

²Livestock Genetics, International Livestock Research Institute (ILRI), P.O. Box 30197 Nairobi 00100, Kenya

³Scotland Rural College, The Roslin Institute Building, Easter Bush Research Centre, Midlothian EH25 9RG, UK

Abstract

The huge increases in dairy cattle milk yield in advanced economies over the past century is a powerful example of the role of breeding in improving livestock productivity. However, breeding practices have had poor efficacy and penetrance in smallholder farming systems in regions such as East Africa. Therefore, to meet the continually growing expectations of a more discerning global population for a more varied and nutritious diet, effective dairy cattle breeding programmes need to reach smallholder dairy producers. In advanced economies, large data sets from commercial farms with modest to large herd sizes (e.g. 20 to several thousand cows) and widespread use of AI have provided sufficient animals within each herd and sufficient genetic connectedness between herds. This has enabled the genetic and environmental components of an individual animal's phenotype to be accurately separated, thus providing accurate genetic evaluations with pedigree information. Typically, herds are neither large nor have high genetic connectedness in smallholder farming systems, such as in East Africa, which limits genetic evaluation with pedigree information. Genomic information keeps track of shared haplotypes rather than animals. This information could capture and strengthen connectedness between herds and through this may enable genetic evaluations based on phenotypes recorded on smallholder dairy farms. The objective of this study was to use simulation to quantify the power of genomic information to enable genetic evaluation under such conditions. The results show; (i) GBLUP produced higher accuracies than PBLUP at all population sizes and herd sizes, (ii) Models with herd fitted as a random effect produced equal or higher accuracies than the model with herd fitted as a fixed effect across all herd size scenarios, (iii) At low levels of genetic connectedness, with four offspring per sire and one to two animals per herd, GBLUP produced EBV accuracies greater than 0.5. Generally, a decrease in the number of sires mated showed consistently higher accuracies compared to when more sires were used. These results suggest that effective breeding programs that use data recorded on smallholder dairy farms in East Africa are possible.

Keywords: dairy cattle, genetic evaluations, genomic selection, smallholder farms, genetic connectedness.

Introduction

The huge increases in dairy cattle milk yield in advanced economies over the past century is a powerful example of the role of breeding in improving livestock productivity. For example, in the US dairy industry, production of milk per cow doubled from an average of 6,000kg to 12,000kg between 1955 and 2005. Approximately 50% of this improvement can be attributed to breeding (CDCB 2017). However, breeding practices have had poor efficacy and penetrance in smallholder farming systems in regions such as East Africa. Recent estimates from Kenyan smallholder farms

suggest that average productivity per cow is as low as 3 litres of milk per day. Despite this, the smallholder dairy economy is very important, as farms with an average of five cows or less account for greater than 70% of the milk produced (East African Dairy Development Program 2012; Abdulsamad & Gereffi 2016). Therefore, to meet the continually growing expectations of a more discerning global population for a more varied and nutritious diet, the improvements in dairy cattle breeding programmes need to reach smallholder dairy producers.

Genetic evaluations are a central component of genetic improvement programs. Statistical models for genetic evaluation that use

pedigree information to model the relatedness between individuals have been very successful in advanced economies. Large data sets from commercial farms with modest to large herd sizes (e.g. 20 to several thousand cows) and widespread use of AI (Wiggans et al. 1988) have provided sufficient animals within each herd and sufficient genetic connectedness between herds. This has enabled the genetic and environmental components of an individual animal's phenotype to be accurately separated (Henderson 1953), thus providing accurate genetic evaluations.

In contrast, smallholder dairy producers in Kenya, have small herd sizes (e.g. <five cows) and artificial insemination is not widely used (Muia et al. 2011; Mutavi et al. 2016; Ojango et al. 2016). This has been a barrier to effective genetic evaluations using pedigree best linear unbiased predictions (PBLUP).

The decreasing costs of SNP chip technologies have resulted in the recent accessibility of genotype information to farmers in East Africa. This in combination with developments in the statistical methodologies of genomic selection (Meuwissen et al. 2001) has, for the first time, meant genetic evaluations are a realistic goal in East Africa. The computation of a genomic relationship matrix from SNP genotypes in GBLUP tracks shared haplotypes amongst animals rather than animals. This information could capture and strengthen connectedness between herds and through this may enable genetic evaluations based on phenotypes recorded on smallholder dairy farms.

To aid in the separation of the genetic potential of an individual from the environmental effects, the management group or herd should also be modelled (Schaeffer 2009; Visscher & Goddard 1993; Ugarte et al. 1992; Frey et al. 1997). Since the publication of Henderson's selection bias theory (1975), there has been debate over the best way to model the herd effect. Currently, most genetic evaluations in advanced economies treat herd as a fixed effect which is appropriate in such situations because herd sizes are large. However, little research has been undertaken to quantify the impact of treating herds as fixed or random effects in smallholder farming systems. These farming systems are characterised by small herd sizes, large differences between production environments and likely differences in management practices on different farms. Modelling herd as a random effect in smallholder scenarios could maximise the number of effective records available for use in

genetic evaluations and in theory should provide estimated breeding values (EBV) with higher accuracies and reduced bias.

The objectives of this study were: (i) to quantify the power of genomic information to enable genetic evaluation based on phenotypes recorded on small-holder dairy farms; and (ii) to quantify the benefit of treating herds as fixed or random effects. We tested a range of scenarios with three different models to characterise herd groups that varied in the number of sires per generation, the number and size of herds and overall population size.

Material and methods

Simulation of trait and SNP genotypes

True breeding values were simulated using a genetic model with 10,000 QTL with additive effects that were sampled from a normal distribution following the parameters described by Jenko et al (2015). The trait simulated had a narrow sense heritability of 0.1 and herd heritability of 0.4 to reflect variance components previously estimated for milk production in East African smallholder dairy cattle populations. Phenotypes for individuals were simulated by adding a residual and herd effect to the true breeding value. Each individual had 50,000 single nucleotide genotypes simulated following the procedure described by Hickey & Gorjanc (2012).

Population history

A population resembling modern dairy cattle was simulated over seven generations. Preceding these seven generations, 100,000 years of evolution was simulated using a coalescent model following the parameters described by Jenko et al. (2015). In the first generation of the recent historical animal breeding population, the chromosomes of the individuals were sampled from the 1,000 simulated haplotypes. In later generations, individuals had their chromosomes sampled from parental chromosomes with recombination. The first four generations were used to burn-in the simulation. The first three of these generations comprised 1,000 animals of equal sex ratio with 50 of the males selected to be sires and 500 females selected to be dams of the next generation. In the final generation of the burn-in, 40,000 animals of equal sex ratio were produced.

Recent breeding and herd connectedness

Following historical breeding, six different connectedness scenarios were simulated over three generations of selection. Connectedness was varied using six different numbers of sires per generation: 50, 100, 250, 500, and 1,000 sires. The first and second generations of selection produced 40,000 animals of equal sex ratio. Selection on sires was based on true breeding values (TBV). No selection was performed on dams. The final generation of selection produced 4,000 animals.

Population and herd size

Four different population sizes were achieved by a balanced selection of individuals in the final generation based on the identifier of their sire. The full 4,000 individual dataset was sub-setted to produce multiple data sets with overall population sizes of 500, 1,000, 2,000 and 4,000 individuals, respectively. Herd counts were sampled from a Poisson distribution with λ equal to the targeted mean herd size. These herd counts were then randomly assigned to herds. Herds were then randomly assigned to individuals in the final generation. This process was repeated for each of the six herd size scenarios of 1, 2, 4, 8, 16 and 32 animals per herd.

The simulation process resulted in 120 different populations to evaluate. Each scenario was replicated 10 times.

Estimation of breeding values

Breeding values were estimated by analysing the data with univariate models using the Bayesian Generalized Linear Regression (BGLR) software (Pérez & De Los Campos 2014). Three models were fit; (i) excluding a herd effect; (ii) herd fitted as a fixed effect; and (iii) herd fitted as a random effect. All models fitted the animal as a random effect. For PBLUP models, the variance of breeding values was equal to $A\sigma_a^2$, where σ_a^2 is the genetic variance associated with the numerator relationship matrix, A. The A matrix was built using a complete pedigree back until the grandparental generation. For GBLUP models, the variance of breeding values was equal to $G\sigma_g^2$, where σ_g^2 is the genetic variance associated with the genomic relationship matrix, G. The G matrix was built only using individuals from the final

generation and calculated from the simulated 50k SNP genotypes following the VanRaden 1 approach (VanRaden 2008). An identity matrix and additive relationship matrix, combined with their respective variances were used to structure the herd and animal covariance's, respectively. To assess the accuracy of the genetic evaluations, the PBLUP and GBLUP EBVs generated by BGLR were correlated with the simulated TBVs. Separate accuracies were calculated for every combination of model and breeding programme scenario.

Results***Impact of the method on EBV accuracy***

GBLUP produced higher accuracies than PBLUP across all herd sizes (Figure 1). With a population size of 1,000 individuals, the accuracy was 0.445 for GBLUP and 0.350 for PBLUP. When the population size was doubled to 2,000 individuals the accuracy was 0.505 for GBLUP and 0.350 for PBLUP. The largest population size of 4,000 records produced the highest accuracies with an accuracy of 0.601 for GBLUP and 0.355 for PBLUP. Of note is the lack of increase in accuracy of PBLUP as the population size increased. This was due to the increased number of herd effects that were estimated. The trend of increased EBV accuracy using GBLUP compared to PBLUP was also observed for models excluding the herd effect and herd fitted as a fixed effect. Therefore, only GBLUP results will be presented from here on.

Fixed versus Random Herd Effect

Genetic evaluations were run using three models: (i) excluding a herd effect; (ii) herd fitted as a fixed effect; and (iii) herd fitted as a random effect. Including a herd effect in the model was important. The model excluding a herd effect produced an accuracy of 0.355. While accuracies of 0.433 and 0.433 were achieved with herd fitted as fixed and random effects, respectively. At herd sizes of eight or greater, there was no difference in accuracy achieved by fitting herd as a fixed or random effect (Figure 2a). At a herd size of eight, models with herd fitted as a fixed or a random effect produced accuracies of 0.417 and 0.420, respectively. At the largest herd size of 32, similar accuracies of 0.432 and 0.431 were observed.

However, at low herd sizes, the choice of the model became increasingly important (Figure 2b). With a population size of 4,000 cows and a

herd size of four, the model with herd fitted as a random effect increased EBV accuracy by 0.016. This difference more than tripled to 0.057 with a decrease in mean herd size to two animals per herd. With herd fitted as a random effect, accuracies were 0.597, while with herd fitted as a fixed effect, accuracies were 0.540. Finally, with a herd size of one, the model with herd fitted as a fixed effect produced accuracies of 0.000, as expected. While the model with herd fitted as a random effect produced accuracies of 0.549. Something of note is that the accuracies of the model excluding a herd effect, 0.554 and 0.552, outperformed the model with herd fitted as a fixed effect, 0.000 and 0.540, with a mean herd size of two or less. At a herd size of one, the model with herd fitted as a random effect produced comparable EBV accuracies as the model excluding the herd effect.

Impact of Connectedness

Results from GBLUP evaluations with a population a size of 4,000 are presented in Table 1. Accuracies of 0.580 and 0.530 were achieved at low levels of genetic connectedness with four offspring per sire spread across 2,000 and 4,000 herds, respectively. No statistically significant changes in accuracy were observed when the number of offspring per sire was increased to eight and sixteen.

However, the higher levels of genetic connectedness produced with forty offspring per sire increased accuracies by 0.029 and 0.030. When dams were spread across 2,000 herds an accuracy of 0.609 was achieved, while analysis including 4,000 herds produced an accuracy of 0.560. The highest accuracies were achieved with 80 offspring per sire. This equates to the use of 50 sires per generation. Accuracies of 0.635 and 0.591 were achieved with 2,000 and 4,000 herds, respectively.

Discussion

There have been increasing efforts to improve productivity in East African small-holder dairy economies (Mutavi et al. 2016). This has included new methods to collect data from rural farms more effectively, establish national genetic evaluation schemes, as well as, some initial studies into the suitability of genomic selection in small-holder cattle populations (Ojango et al. 2016; Brown et al. 2016). Missing or incomplete mating records and low levels of genetic connectedness between herds

have often hampered previous pedigree-based evaluations, while genomic studies have often been limited by dataset size. This simulation study aimed to demonstrate the capability of genomic based evaluations to provide reliable results when using data from smallholder farms.

The higher accuracies achieved by GBLUP over PBLUP, as well as, the reliable accuracies achieved by GBLUP at low numbers of offspring per sire and low herd sizes, indicate keeping track of haplotypes rather than animals captures and strengthens genetic connectedness between herds. While, the increased accuracies observed using the model with herd fitted as a random effect over a fixed effect indicates that the random model can increase the number of effective of records used in evaluations.

However, these simulations were based upon a random association between the genetic value of the dam and herd value. This may not be the case in reality, where preferential treatment and a positive correlation between farm inputs and animal genetic value could exist. Previous simulation studies have demonstrated that non-random associations between sire genetic value and management group value can introduce bias to sire EBVs (Ugarte et al. 1992; Visscher & Goddard 1993). The influence of non-random associations on bias in EBVs has been investigated less in dams, where the trend is expected to be less extreme. However, this warrants further investigation in a smallholder dairy systems context.

Conclusions

This simulation study aimed to demonstrate the capability of genomic based evaluations to provide reliable results when using data from smallholder farms. GBLUP was shown to produce higher accuracies than PBLUP at all population sizes and herd sizes. Models with herd fitted as a random effect produced equal or higher accuracies than the model with herd fitted as a fixed effect across all herd size scenarios. The differences in accuracy between the two modelling approaches were at their largest at low herd sizes. While at low levels of genetic connectedness, with four offspring per sire and one to two animals per herd, GBLUP produced EBV accuracies greater than 0.5.

This study has demonstrated the potential of genomic information to be an enabling technology in smallholder dairy economies by facilitating genetic evaluations with records

collected from farms with herd sizes of four cows or less. The inclusion of smallholder dairy data in genetic evaluations could provide increases in local

and national milk production with downstream impacts upon wider societal, nutritional and economic outcomes.

References

- Abdulsamad, A. & Gereffi, G., 2016. East Africa dairy value chains: Firm capabilities to expand regional trade. , (October).
- Brown, A. et al., 2016. Short communication: Genomic selection in a crossbred cattle population using data from the Dairy Genetics East Africa Project. *Journal of Dairy Science*, 0(0), pp.433–443. Available at: <http://dx.doi.org/10.3168/jds.2016-11083>.
- CDCB, 2017. Trend in Milk BV for Holstein or Red & White. Available at: <https://queries.uscdcb.com/eval/summary/trend.cfm> [Accessed January 30, 2018].
- East African Dairy Development Program, 2012. The Dairy Value Chain in Kenya. , (September), pp.2011–2012.
- Frey, M., Hofer, A. & Künzi, N., 1997. Comparison of models with a fixed or a random contemporary group effect for the genetic evaluation for litter size in pigs. *Livestock Production Science*, 48, pp.135–141. Available at: file:C:/Users/xwei/Reprints/F/FreyM comp mdl fxd_randm grp effect 4 g_eval.pdf.
- Henderson, C.R., 1975. Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics*, 31(2), pp.423–447.
- Henderson, C.R., 1953. Estimation of Variance and Covariance Components. *Biometrics*, 9(2), pp.226–252.
- Hickey, J.M. & Gorjanc, G., 2012. Simulated Data for Genomic Selection and Genome-Wide Association Studies Using a Combination of Coalescent and Gene Drop Methods. *G3: Genes/Genomes/Genetics*, 2(4), pp.425–427. Available at: <http://g3journal.org/lookup/doi/10.1534/g3.111.001297>.
- Jenko, J. et al., 2015. Potential of promotion of alleles by genome editing to improve quantitative traits in livestock breeding programs. *Genetics, selection, evolution : GSE*, 47(1), p.55. Available at: <http://www.gsejournal.org/content/47/1/55>.
- Meuwissen, T.H.E., Hayes, B.J. & Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4), pp.1819–1829.
- Muia, J. et al., 2011. Smallholder dairy production in high altitude Nyandarua milk-shed in Kenya: Status, challenges and opportunities. , 23.
- Mutavi, S.K. et al., 2016. Innovativeness and Adaptations : The Way forward for Small scale Peri- Urban Dairy Farmers in Semi-Arid Regions of South Eastern Kenya. , 3(5), pp.1–14.
- Ojango, J.M.K. et al., 2016. Dairy production systems and the adoption of genetic and breeding technologies in Tanzania, Kenya, India and Nicaragua. *Animal Genetic Resources*, 59, pp.81–95. Available at: <https://www.cambridge.org/core/terms.%5Cnhttps://www.cambridge.org/core/terms.%5Cnhttps://www.cambridge.org/core/terms>.
- Pérez, P. & De Los Campos, G., 2014. Genome-wide regression and prediction with the BGLR statistical package. *Genetics*, 198(2), pp.483–495.
- Schaeffer, L., 2009. Contemporary Groups Are Always Random. *Personal*, pp.1–6. Available at: <http://www.aps.uoguelph.ca/~lrs/LRSsite/ranfix.pdf>.
- Ugarte, E., Alenda, R. & Carabaño, M.J., 1992. Fixed or Random Contemporary Groups in Genetic Evaluations. *Journal of Dairy Science*, 75(1), pp.269–278. Available at: <http://linkinghub.elsevier.com/retrieve/pii/S0022030292777625>.
- VanRaden, P.M., 2008. Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science*, 91(11), pp.4414–4423. Available at: <http://linkinghub.elsevier.com/retrieve/pii/S0022030208709901>.
- Visscher, M.P. & Goddard, M.E., 1993. Fixed and Random Contemporary Groups. *Journal of Dairy Science - J DAIRY SCI*, 76, pp.1444–1454.
- Wiggans, G.R., Misztal, I. & Van Vleck, L., 1988. Implementation of an Animal Model for Genetic Evaluation of Dairy Cattle in the United States. *Journal of Dairy Science*, 71, pp.54–69.

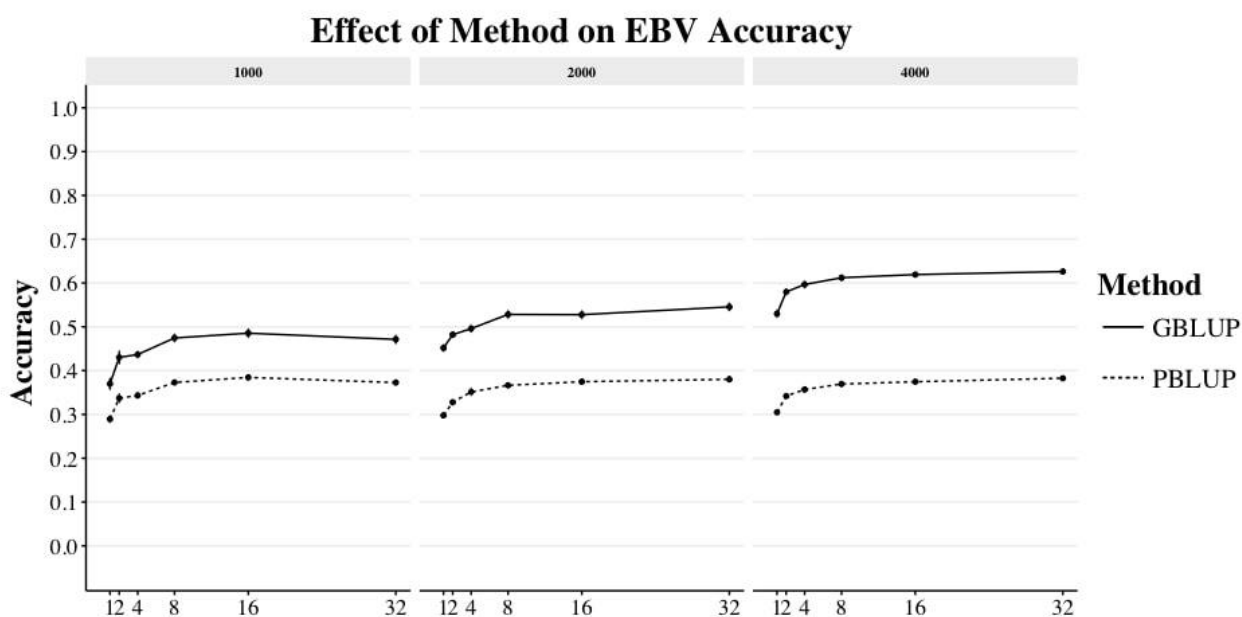


Figure 1. Comparison of the genetic evaluation method. The accuracy of estimated breeding values from PBLUP (--) & GBLUP (–) as the size of herds (1-32) and population size (1,000-4,000) is increased. The number of offspring per sire was held constant at 4. Herd fitted as a random effect.

Table 1. Impact of connectedness among herds. GBLUP accuracies presented for a population size of 4,000 individuals.

<i>n</i> Herds	HerdSize	Offspring/Sire	Accuracy	<i>se</i>
4000	1	4	0.530	0.008
		8	0.539	0.008
		16	0.530	0.009
		40	0.560	0.010
		80	0.591	0.012
2000	2	4	0.580	0.007
		8	0.575	0.007
		16	0.588	0.006
		40	0.609	0.005
		80	0.635	0.008

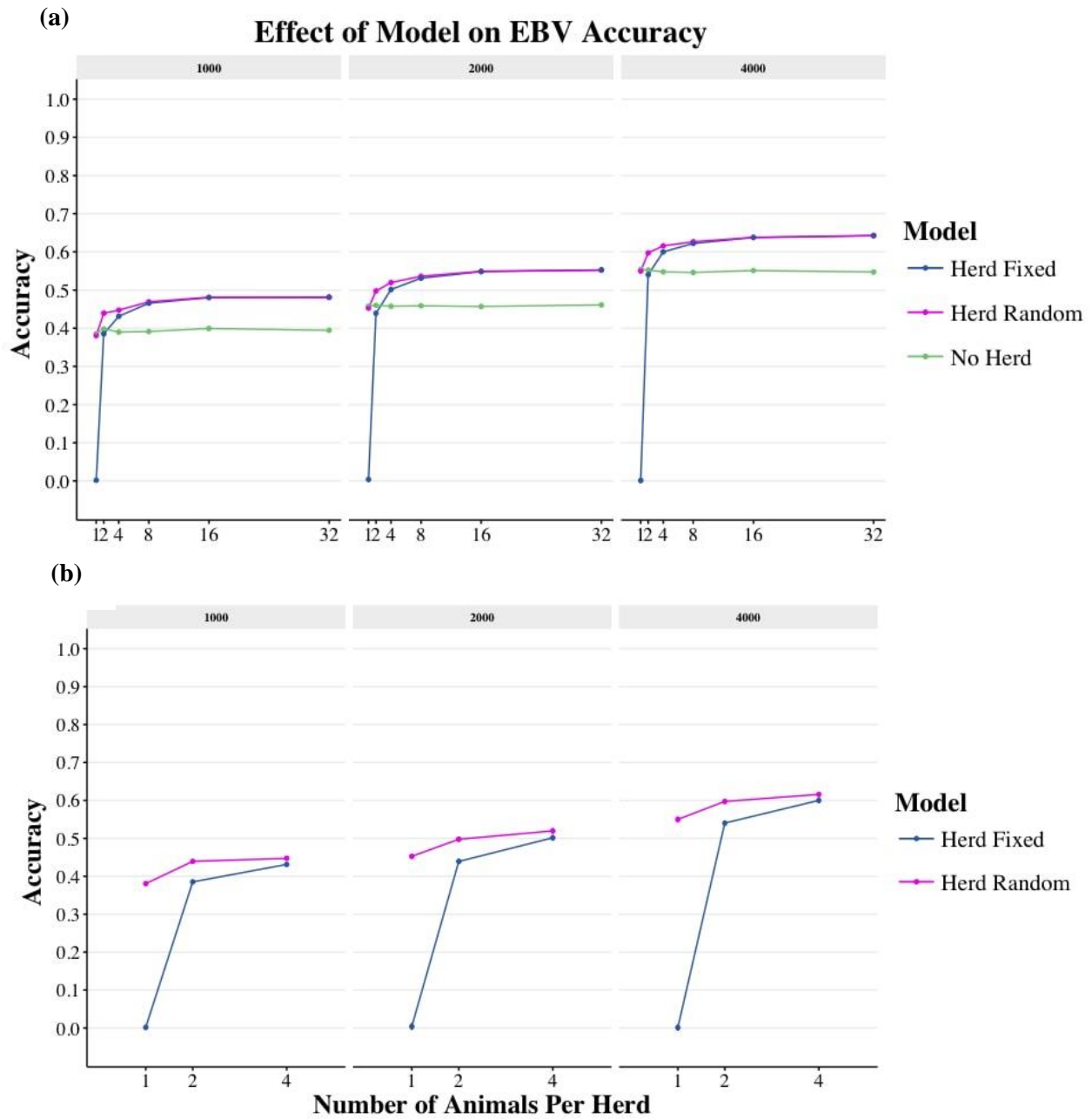


Figure 2. Comparison of modelling the herd effect with GBLUP. The accuracy of estimated breeding values with increasing population size (1,000-4,000) and with the herd; (i) excluded from the model (—), (ii) fitted as a fixed effect (—) and (iii) fitted as a random effect (—). a) Accuracies achieved across all herd sizes (1-32). b) Accuracies achieved at small herd sizes (1-4). Model (i) not included.