Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Estimation of Genomic Breeding Values for Milk Yield in UK Dairy Goats

S. Mucha, R. Mrode, M. Coffey and J. Conington.
Animal & Veterinary Sciences, Scotland's Rural College, Easter Bush, Midlothian EH25 9RG, Scotland, United Kingdom

ABSTRACT: Genomic selection in UK dairy goats is in its initial stage. One of the main challenges is the small size and structure of the reference population (mixture of males and females). The objective of this study is to estimate genomic breeding value for milk yield in dairy goats. The research was based on data provided by two farms in the UK comprising 590,409 milk yield records on 14,453 goats. The pedigree contained 30,139 individuals. In total 1960 animals were genotyped with Illumina 50K caprine chip. BLUP-SNP and single-step approach were performed on the data and the two methods were compared. The highest accuracy was obtained with the single-step method. The results indicate that this method provides the best accuracy for populations with a small number of genotyped individuals, where the number of males is relatively low, and females are predominant in the reference population.
Keywords: dairy goats; genomic selection; milk yield

Introduction

Genomic selection has become routine in many species such as dairy and beef cattle. Thanks to exchange of genotypes between countries, reference populations for those species are large consisting of thousands of bulls with high reliability breeding values. This allows predicting genomic breeding values for young animals, which have no phenotypic records, with high accuracy. In the case of dairy goats, the breeding industry is not so well developed worldwide. Routine breeding value estimation is carried out in such countries as Canada, France, US, and Norway (Bélichon et al. 2000; Montaldo and Manfredi 2002). Currently genomic selection in dairy goats has been introduced only in France (Carillier et al. 2013) using 2810 genotyped Saanen and Alpine goats. In the UK, the number of genotyped goats is also relatively small which poses certain restrictions with respect to the estimation of genomic breeding values. The accuracy of the methods that use only phenotypes of the genotyped animals, and ignore the records of the non-genotyped part of the population (GBLUP, BLUP-SNP) is limited when the reference population is small. Therefore, an alternative approach is considered, which integrates all of the available phenotypic, pedigree, and genomic information in a single step procedure (Legarra et al. 2009; Misztal et al. 2009; Christiansen and Lund 2010). This approach has been regarded as computationally demanding in the case of large datasets and complex models. However in goats, the amount of data used in genetic evaluations is considerably lower than that of dairy cattle. Moreover the method is easy to implement as it can use raw phenotypic records without the need to calculate deregressed proofs. It also facilitates the evaluation of all animals (with and without genotypes) simultaneously. The objective of this study was to evaluate BLUP-SNP and single-step approaches for the estimation of genomic breeding values in dairy goats. Additionally, the level of linkage disequilibrium in the reference population was investigated.

Materials and Methods

Phenotypic data. The lactation data were from two separate farm units in the UK owned by a single farming business. A more detailed overview of the lactation data and genetic parameters is described by Mucha et al., 2014. The dataset comprised 590,409 records on 14,453 dairy goats kidding between 1987 and 2013. The population was created in 1985 by crossing three breeds: Alpine, Saanen, and Toggenburg. There was no particular crossing strategy. In each generation the best performing animals were selected for breeding and as a result, a synthetic breed was created. The breed composition of the animals was not recorded, and thus could not be included in the analysis. To mitigate this problem SNP information was used to assess breed composition of the animals. A total of 1961 goats from the same population were genotyped with Illumina caprine 50K chip (Illumina Inc., San Diego, CA; Tosser-Klopp et al. 2012). Clustering based on principal component analysis, performed with SNP & Variation Suite v7.7.8 (Golden Helix Inc.), did not reveal any major distinct groups. This suggests that the analysed population is mostly homogenous and therefore breed was not included as a factor in the analysis. The pedigree file contained 30,139 individuals, of which 2,799 were considered as founders. There were 296 sires and 12,468 dams in the pedigree. The dataset contained test day records of milk yield, along with information about lactation number (1 to 6), farm (2 farms), age at kidding (12 to 90 months), year (1987 to 2013) and season of kidding [summer (June to August), autumn (September to November), winter (December to February), and spring (March to May)]. Fat and protein content were not included in these first analyses, as they have only recently started to formally record these traits.

Genotypes. Two thousand animals were selected for the reference population. The selection of candidate animals was done based on two criteria: average daily lifetime yield (ADLY), and the genetic relationship between the animals. The process was optimized in such a way as to select animals from the upper (group 1) and lower (group 2) tail of the distribution of ADLY. Animals had been selected so that the relationship within the two groups was minimized, and relationship between the two groups was maximized. This was undertaken using software package Coro-

na (Brian Kinghorn, pers.comm). Mean relationship of animals in the reference population was 0.03.

Animals were genotyped with the Illumina Caprine 50K BeadChip (Illumina Inc., San Diego, CA). After filtering out SNP that were not in Hardy-Weinberg equilibrium, had minor allele frequency below 0.05, were monomorphic, had call rate below 0.95 or the GC content below 0.6, the dataset contained 47,306 markers. Additionally, animals with a call rate below 0.9 were removed from further analyses. This resulted in 1902 animals in the data set that were born between 2003-2012.

Linkage disequilibrium. Linkage disequilibrium (LD) was measured as r^2, which is the squared correlation of the alleles at two loci (Hill and Robertson 1968):

$$r^2 = \frac{[f(AB) - f(A)f(B)]^2}{f(A)f(a)f(B)f(b)}$$

Where f(AB), f(A), f(a), f(B), f(b) are observed frequencies of haplotype AB and of alleles A, a, B and b, respectively. LD was calculated for all syntenic marker pairs (markers from the same chromosome). SNP markers that could not be mapped to any chromosome were excluded from these analyses.

Estimation of genomic breeding values. Two methods were used to estimate genomic breeding values (GEBV). The first method was BLUP-SNP where de-regressed sire and female proofs were used as phenotypes. The software package MIX99 (Lidauer et al. 2011) was used for the de-regression, using a full animal pedigree with effective daughter contributions used as weighting factors. SNP effects were estimated with the following statistical model:

$$y_i = \mu + v_i + \sum_{j=1}^{m} z_{ij} u_j + e_i$$

where: y_i is the de-regressed proof, μ is the overall mean, v_i is the residual polygenic effect of i-th goat (10% of additive genetic variance), z_{ij} is the genotype value coded as 0, 1, or 2 for homozygote, heterozygote, and the opposing homozygote, u_j is the random regression coefficient for j-th SNP, and e_i is the residual effect.

The second approach to calculate GEBV was based on the single-step method - HBLUP (Legarra et al. 2009; Misztal et al. 2009; Christiansen and Lund 2010). The software package BLUPf90 (Misztal et al. 2002) was used to fit the following random regression model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{Wp} + \mathbf{e}$$

where **y** is the vector of test-day observations; **b** the vector of fixed effects consisting of herd test day, year-season, age at kidding, and fixed lactation curves modelled by fitting Legendre polynomials (Kirkpatrick et al. 1990) of fourth order; **a** is a 1x3 vector of random regression coefficients (Legendre polynomials of second order) for the animal effect; **p** is the 1x3 vector of random regression coefficients (Legendre polynomials of second order) for the permanent environment effect; **e** is the vector of random residual effect. The matrix **X** is the incidence matrix for fixed effects; **Z** and **W** are matrices of Legendre polynomials of days in milk of second order for random animal and permanent environment effect, respectively. Random effects were assumed to be normally distributed with zero means and the following covariance structure:

$$Var\begin{bmatrix}\mathbf{a}\\ \mathbf{p}\\ e\end{bmatrix} = \begin{bmatrix}\mathbf{H}\otimes\mathbf{U} & 0 & 0\\ & \mathbf{I}\otimes\mathbf{P} & 0\\ symm & & \mathbf{I}\sigma_e^2\end{bmatrix}$$

Where **U** and **P** are 3 x 3 (co)variance matrices of the random regression coefficients for the animal and permanent environment effects respectively, **I** are identity matrices, and **H** is the relationship matrix calculated using Van Raden's (2008) genomic relationship matrix **G** and pedigree relationship matrix **A** as:

$$\mathbf{G} = 0.95\frac{\mathbf{SS}'}{2\sum_{i=1}^{n} p_i(1-p_i)} + 0.05\mathbf{A}$$

Where **S** is a centered incidence matrix of SNP genotypes, n is the number of SNP markers, and p_i is allele frequency of marker i. The inverse of H is:

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

Where $\mathbf{A}_{22}^{-1}$ is the inverse of pedigree relationship matrix for the genotyped animals.

Accuracy of genomic breeding values. The reference population was divided into a training and a validation set consisting of 1474 (1410 females and 64 males) and 305 animals (302 females and 3 males), respectively. Reference animals were born between 2003 and 2010, and had a minimum reliability of EBVs of 0.76. Validation animals were born in 2011 and had a minimum reliability of EBVs of 0.69. Females in the reference and validation set had 1 to 6, and 1 to 2 lactations respectively. The accuracies of genomic predictions were calculated as correlations between de-regressed proofs (DRP) and GEBVs from HBLUP, or by direct genomic values (DGV) using BLUP-SNP of validation animals. Additionally, the accuracy of pedigree-based predictions (PBLUP) were calculated as correlations of DRP and EBV of the validation animals. The EBVs from PBLUP were obtained from the same model as for HBLUP, but the H matrix was replaced with the pedigree-based A matrix. The gain of using SNP information was calculated as the difference between accuracy of HBLUP and PBLUP. Additionally, to verify the gain in accuracy of the non-

genotyped candidates in HBLUP, the genotypes of 100 animals from the validation data set were removed (HBLUP-cut). The accuracy was then compared with the accuracy calculated for the same 100 animals with genotypes (HBLUPall).

Results and Discussion

Average r^2 among syntenic markers as a function of marker distance is presented in Figure 1. The largest decline of LD was for distances below 100 kb. In the studied population the mean r^2 at 50 kb (distance between two SNP) was 0.18. LD declined from 0.14 at 100 kb to 0.09 at 1000 kb, and 0.07 at 2000 kb. The extent of LD found in the current population is similar to that reported by Carillier et al. (2013) in the French pure bred dairy goats (r^2 of 0.17 at 50 kb) but higher than the value obtained for the cross-bred population (r^2 of 0.14 at 50 kb). This might indicate that this population is mostly homogeneous with respect to breed composition, and can be treated as a synthetic pure-breed. The average LD in dairy goats appears to be lower than that reported in dairy cattle (0.20-0.23 at 40 kb, Khatkar et al. 2008; de Ross et al. 2008; Habier et al. 2010) or pigs (0.47-0.49 at 30 kb, Uimari and Tapio 2011).

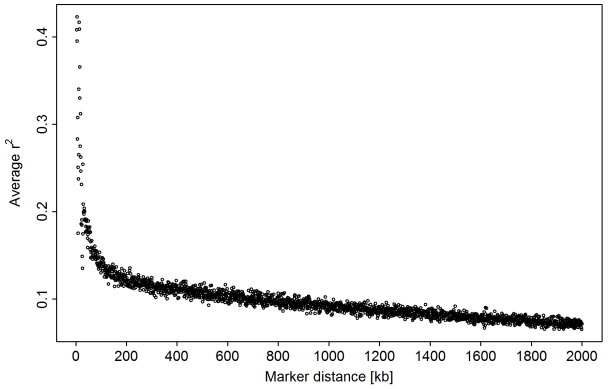

Figure 1. Linkage disequilibrium

Table 1. Correlations (R), regression coefficients (b_1) between de-regressed proofs (DRP) and parent average (PA), EBVs from PBLUP, DGVs from BLUP-SNP, and GEBVs from HBLUP for animals in the validation population.

Method	b_1	R
PA	1.08	0.45
PBLUP	1.27	0.58
BLUP-SNP	0.29	0.36
HBLUP	0.99	0.61
$HBLUP_{all}$	1.13	0.56
$HBLUP_{cut}$	1.30	0.54

$HBLUP_{cut}$ = same as HBLUP, but genotypes of 100 animals from the validation removed

$HBLUP_{all}$ = same as HBLUP but regression and correlation was for the 100 validation animals without genotypes in $HBLUP_{cut}$

The accuracy of the genomic predictions was 0.36 and 0.61 for BLUP-SNP and HBLUP, respectively (Table 1). Low accuracy of BLUP-SNP could be due to having mostly females in the validation set and therefore less precise information. BLUP-SNP not only had a considerably lower accuracy, but also resulted in a low regression coefficient of 0.29. This suggests that DGV obtained from this method overpredicts the DRP. GEBVs from HBLUP appear to be less biased, as the regression coefficient was 0.99. The gain of using SNP information expressed as the difference between accuracy of PBLUP and HBLUP was 5.2%. The accuracy of young animals increased by 3.7% when comparing their evaluation with and without all genotypes

Conclusion

Single-step approach resulted in higher accuracy of genomic breeding values in comparison with BLUP-SNP. This method can be recommended for breeding programmes with reference populations containing a small number of sires supplemented with females.

Acknowledgements

This paper is part of a 3 year project co-funded by the UK's innovation agency, the Technology Strategy Board in collaboration with Illumina. The authors gratefully acknowledge co-operation with Angus Wielkopolski and Mark De Hamel from Yorkshire Dairy Goats.

Literature Cited

Bélichon, S., Manfredi, E., and Piacère, A. (2000). Genet. Sel. Evol. 30:529-534.

Carillier, C., Larroque, H., Palhière, I. et al. (2013). J. Dairy Sci. 96:7294–7305.

Christensen, O. F., and Lund, M. S. (2010). Genet. Sel. Evol. 42:2.

de Roos, A. P. W., Hayes, B. J., Spelman, R. J. et al. (2008). Genetics 179:1503–1512.

Habier, D., Tetens, J., Seefried, F-R. et al. (2010). Genet. Sel. Evol. 42:5.

Hill, W. G., and Robertson, A. (1968). Theor. Appl. Genet. 38:226–231.

Khatkar, M. S., Nicholas, F. W., Collins, A. R. et al. (2008). BMC Genomics 9:187.

Kirkpatrick, M., Lofsvold, D., and Bulmer, M. (1990). Genetics 124:979-993

Legarra, A., Aguilar, I., and Misztal, I. (2009). J. Dairy Sci. 92:4656-4663.

Lidauer, M., Matilainen, K., Mantysaari, E. et al. (2011). MiX99, MTT, Jokioinen, Finland.

Misztal, I., Tsuruta, S., Strabel, T. et al. (2002). Proc. 7th WCGALP, 743–745.

Misztal, I., Legarra, A., and Aguilar, I. (2009). J. Dairy Sci. 92:4648–4655.

Montaldo, H. H., and Manfredi, E. (2002). Proc. 7th WC GALP, 1-8.

Mucha, S., Mrode, R., Coffey, M. et al. (2014) J. Dairy Sci. 97:2455–2461.

Uimari, P., and Tapio, M. (2011). J. Anim. Sci. 89:609-614.

VanRaden, P. M. (2008). J. Dairy Sci. 91:4414–4423.