

Accuracy of genomic breeding values for residual feed intake in crossbred beef cattle¹

F. D. N. Mujibi,* J. D. Nkrumah,* O. N. Durunna,* P. Stothard,* J. Mah,* Z. Wang,* J. Basarab,† G. Plastow,* D. H. Crews Jr.,*‡ and S. S. Moore*²

*Department of Agriculture, Food and Nutritional Science, University of Alberta, Edmonton, Alberta, T6G 2P5, Canada; †Alberta Agriculture and Rural Development, 6000 C & E Trail, Lacombe Research Centre, Lacombe, Alberta, T4L 1W8, Canada; and ‡Department of Animal Sciences, Colorado State University, Fort Collins 80523-1171

ABSTRACT: The benefit of using genomic breeding values (GEBV) in predicting ADG, DMI, and residual feed intake for an admixed population was investigated. Phenotypic data consisting of individual daily feed intake measurements for 721 beef cattle steers tested over 5 yr was available for analysis. The animals used were an admixed population of spring-born steers, progeny of a cross between 3 sire breeds and a composite dam line. Training and validation data sets were defined by randomly splitting the data into training and testing data sets based on sire family so that there was no overlap of sires in the 2 sets. The random split was replicated to obtain 5 separate data sets. Two methods (BayesB and random regression BLUP) were used to estimate marker effects and to define marker panels and ultimately the GEBV. The accuracy of prediction (the correlation between the phenotypes and GEBV) was compared between SNP panels. Accuracy for all traits was low, ranging from 0.223 to 0.479 for marker panels

with 200 SNP, and 0.114 to 0.246 for marker panels with 37,959 SNP, depending on the genomic selection method used. This was less than accuracies observed for polygenic EBV accuracies, which ranged from 0.504 to 0.602. The results obtained from this study demonstrate that the utility of genetic markers for genomic prediction of residual feed intake in beef cattle may be suboptimal. Differences in accuracy were observed between sire breeds when the random regression BLUP method was used, which may imply that the correlations obtained by this method were confounded by the ability of the selected SNP to trace breed differences. This may also suggest that prediction equations derived from such an admixed population may be useful only in populations of similar composition. Given the sample size used in this study, there is a need for increased feed intake testing if substantially greater accuracies are to be achieved.

Key words: accuracy, beef cattle, genomic breeding value, residual feed intake, single nucleotide polymorphism

©2011 American Society of Animal Science. All rights reserved.

J. Anim. Sci. 2011. 89:3353–3361
doi:10.2527/jas.2010-3361

INTRODUCTION

A large number of genomic tools have become available because of the rapid advancement of DNA marker technology after the mapping (and sequencing) of the bovine genome. This has led to increasing demands for inclusion of DNA marker tools in traditional evaluation systems, to yield marker-assisted EBV, often with greater accuracy compared with traditional EBV (Johnston et al., 2008). Various strategies have been suggested for inclusion of marker information in genetic evaluations, but so far none of the methods is optimal (VanRaden, 2001; Dekkers, 2007; Kachman, 2008). Results from a DNA test can be used to create a molecular score (**MS**) or a molecular breeding value, which is often a weighted sum of the number of copies of the frequent alleles of several polymorphisms with the

¹The study was made possible by grants awarded to Stephen Moore from the Canadian Cattleman's Association (Calgary, Alberta, Canada), Alberta Agricultural Research Institute (Edmonton, Alberta, Canada), Alberta Beef Producers (Calgary, Alberta, Canada), Canada–Alberta Beef Industry Development Fund (Calgary, Alberta, Canada), and the Beef Cattle Research Council (Calgary, Alberta, Canada). The authors thank Jason Grant (Department of Agriculture, Food and Nutritional Science, University of Alberta, Edmonton) for Perl programming. The authors also thank John Hickey and Bruce Tier (Animal Genetics and Breeding Unit, University of New England, Armidale, New South Wales, Australia) for providing the software (AlphaBayes) for genomic analysis.

²Corresponding author: stephen.moore@ualberta.ca

Received July 27, 2010.

Accepted May 24, 2011.

weights estimated in a reference data set (Kachman, 2008). Because MS will likely account for only a small portion of the total genetic variance, it will be necessary to combine polygenic and molecular breeding value into a single selection tool (VanRaden, 2001; Dekkers, 2007; Kachman, 2008). Selection index methodologies have been shown in simulation to be useful in combining polygenic and molecular breeding values (Dekkers, 2007; Crews, 2008). Genomic selection is also seen as a viable option where selection is based solely on genomic breeding values (**GEBV**; Meuwissen et al., 2001). Recently, Bayesian estimation has emerged as the method of choice for genomic selection because it allows different variances to be fitted to each SNP (Fernando et al., 2007; Moser et al., 2009). Genomic selection has been successfully applied in the prediction of performance in dairy cattle, but such success has not been realized in beef cattle populations (MacNeil et al., 2010). In this study, Bayesian-based methods and the theory underlying genomic selection were used to select a subset of markers, and ultimately to derive GEBV whose ability to predict RFI, DMI, and ADG was then evaluated using data from an admixed population of beef cattle steers.

MATERIALS AND METHODS

The Canadian Council on Animal Care (1993) protocols and guidelines were followed when caring for the animals.

Animal Resource and Study Design

Data consisted of 721 crossbred steers sired by Angus, Charolais, or University of Alberta hybrid bulls with a composite dam line. The composition of the dam line is described in detail by Goonewardene et al. (2003). Feed intake data were collected over a 5-yr period, with 2 groups (fall-winter and winter-spring) tested every year for the first 3 yr. In yr 4, 1 group of animals was tested for 2 consecutive periods (fall-winter, and then winter-spring), first on a low-energy feedlot diet in period 1 (fall-winter), and then a high-energy feedlot diet in period 2 (winter-spring). In yr 5, 2 groups of animals were tested in 2 consecutive periods as follows: The first group was put on a high-energy feedlot diet for both periods, whereas the second group was first tested on a lower energy diet and then switched to a high-energy diet in period 2. Animals had free-choice access to feed and water. In total, 9 batches of animals were available for analysis, with a batch being a combination of year and season of testing. All batches were placed into 3 groups as follows: Fall-winter tested animals were in group 1, winter-spring test animals were in group 2, and diet-switch animals were in group 3.

Individual animal feed intake and feeding behavior data were collected using the GrowSafe automated feeding system (GrowSafe Systems Ltd., Airdrie, Alberta, Canada) at the University of Alberta Kinsella

ranch. Daily feed intake was converted into daily DMI by multiplying intake by the DM content of the diet. Daily DMI was then standardized across the different years to 10 MJ of ME/kg of DM by multiplying daily DMI with the diet ME content and then dividing by 10 (Basarab et al., 2003). Average daily gain was calculated as the slope from the regression of BW on test day. Metabolic midweight was obtained as the midweight on test raised to the power of 0.75.

Residual feed intake (**RFI**) was calculated within group using the following formula:

$$\text{RFI} = \text{DMI} - (\beta_0 + \beta_1 \text{batch} + \beta_2 \text{ADG} + \beta_3 \text{MMWT}),$$

where β_1 , β_2 , and β_3 are partial regression coefficients; β_0 is the intercept; and MMWT is metabolic midweight.

Training and validation data sets were defined by randomly splitting the data into a training set (2/3, $n = 485$) and a testing set (1/3, $n = 243$) based on sire family so that there was no overlap of sires in the 2 sets. This random split was replicated 5 times such that there were 5 training and 5 testing data sets. Random splitting by sire family reduces the ability of genetic markers to approximate the relationship between individuals in the training and testing data, thereby minimizing chances of an inflated correlation of GEBV and trait phenotype in the prediction process (Habier et al., 2007). The first replicate of the training data was used for SNP preselection, and the selected SNP were then reanalyzed in all replicates of the training data. The association between genotypes and phenotypes was tested in the training set, whereas the accuracy of prediction of the marker-derived breeding value explored in the testing set was tested as the correlation between GEBV and phenotypes.

Genetic Data

Approximately 50,000 SNP were genotyped for 745 beef steers by using the Illumina Infinium II (Illumina Inc., San Diego, CA) platform. These SNP were tested for Hardy-Weinberg equilibrium ($P > 0.05$), minor allele frequency ($>5\%$), and SNP call frequency ($>88\%$), with nonqualifying SNP being discarded. Ultimately, a total of 38,158 SNP were selected for further analysis. Genotypes were coded as 0, 1, and 2, with 0 being the SNP allele with the lesser frequency and 1 being the allele with the greater frequency, respectively, such that the 2 homozygotes were represented as 0 and 2, and 1 was the heterozygote. Missing genotypes (about 1% of all genotypes) were imputed by submitting SNP genotype calls as well as missing genotype information to fastPHASE (Scheet and Stephens, 2006) chromosome by chromosome, the SNP having been ordered according to their chromosomal position. The parameters used were as follows: 10 random starts of the expectation-maximization (EM) algorithm (T), 30 iterations of the EM algorithm (C), 15 cross-validation clusters (K), and no sampling of haplotypes from the posterior distribu-

tion of each random start of the EM algorithm (H). The most probable genotype imputed by fastPHASE was considered the true genotype. All SNP with unknown chromosomal positions were discarded. A final 37,959 SNP were included in the analysis.

The following animal model was used in the whole data set to estimate polygenic breeding values, variance components, and genetic parameters using AS-REML (Gilmour et al., 2008). The model included fixed effects of contemporary group (breed, batch, and test group combinations), with age at the start of test as a covariate, as shown below:

$$\mathbf{y}_1 = \mathbf{X}_1\beta + \mathbf{Z}_1a + \mathbf{e}, \quad [1]$$

where the design matrices \mathbf{X}_1 and \mathbf{Z}_1 relate phenotypic observations in the vector \mathbf{y}_1 to fixed (β) and polygenic (a) effects, respectively. The vector \mathbf{e} contains random residual terms specific to animals. The parameters a and \mathbf{e} were assumed to be normally distributed, with a mean of 0 and variances $\mathbf{A}\sigma_a^2$ and $\mathbf{I}_n\sigma_e^2$, respectively. The matrix \mathbf{I}_n is an identity matrix of order equal to the number of animals with RFI observations, whereas \mathbf{A} is the additive relationship matrix, σ_a^2 is the random polygenic effect variance, and σ_e^2 the residual variance, respectively. Accuracy was calculated using the formula $\text{accuracy} = \sqrt{1 - se^2/\sigma_a^2}$, with se^2 being the prediction error variance and σ_a^2 being the additive genetic variance (Gilmour et al., 2008). A bivariate model was used to compute genetic correlations between the traits by extending Eq. [1] to include a second trait.

Bayesian Estimation of Marker Effects

Estimation of marker effects was performed using 2 models:

1. Random regression BLUP (**RR-BLUP**), which assumes the same prior variance for all random SNP, as described by Meuwissen et al. (2001).
2. BayesB, in which a locus-specific variance is estimated but the loci are divided into 2 groups: one group of a relatively small number of SNP with large effects that contribute to the genetic variance with probability $(1 - \pi)$, and a second group of a large number of SNP with no effect, with probability π (Meuwissen et al., 2001). The BayesB model used was similar to that of Meuwissen et al., (2001), except that effects of SNP genotypes and not haplotype were fitted.

The BayesB model makes strong assumptions about the prior distribution of marker effects, namely, a large proportion of SNP have no effect. The BayesB and RR-BLUP models used are implemented in the AlphaBayes software (Hickey and Tier, 2009), which uses a modified version of the Gibbs sampling algorithm to solve for model effects. The SnpBlup and BayesBfast implementations in AlphaBayes were used for RR-BLUP and

BayesB analyses, respectively. Even though the real value of π was unknown for this data set, π was set at 0.95 for all analyses, such that 5% of SNP were fitted simultaneously in each cycle of the Gibbs chain.

The model of analysis used for RR-BLUP and BayesB was as follows:

$$\mathbf{y}_1 = \mathbf{X}_1\beta + \mathbf{Z}_1a^* + \mathbf{Z}_2g + \mathbf{e}, \quad [2]$$

where the design matrices \mathbf{X}_1 , \mathbf{Z}_1 , and \mathbf{Z}_2 relate phenotypic observations in the vector \mathbf{y}_1 to fixed (β), residual polygenic (a^*), and SNP (g) effects, with elements $\mathbf{Z}_{2ij} = 0, 1, \text{ or } 2$, corresponding to the genotype of animal i at locus j , with g normally distributed with mean 0, and variance σ_{gj}^2 , for RR-BLUP, and drawn from an inverse χ^2 distribution with probability π in BayesB. The variance $\sigma_{gj}^2 = \sigma_a^2/n$ in RR-BLUP, and was estimated for each instance of j in BayesB. The vector \mathbf{e} contains random residual terms specific to animals. The parameters a^* and \mathbf{e} were treated as random. The matrix \mathbf{I}_n is an identity matrix of order equal to the number of animals with trait observations, whereas \mathbf{A} is the additive relationship matrix, $\sigma_{a^*}^2$ is the random residual polygenic effect variance, and σ_e^2 is the residual variance. Fixed effects fitted included contemporary group (breed-batch-test group combinations), whereas age at the start of test was used as a covariate.

The first 20,000 iterations from the total 100,000 iterations were discarded as burn-in. Mean SNP substitution effects were obtained from the posterior samples for each trait, and SNP ranked from greatest to least based on the magnitude of the allele substitution effect. From this ranking, the top 200 SNP were selected for further analysis. Allele substitution effects for the selected SNP were reestimated in each of the 5 replicates of the training data, with the first 5,000 iterations of the total of 20,000 discarded as burn in. For this analysis, π was set to 0.0005 so that estimates for all 200 SNP could be obtained.

Genomic Value Estimation

Trait-specific marker panels were obtained from analysis using the various methods outlined above. The SNP were subsequently used to derive marker scores. Marker scores were calculated as a weighted sum of the number of copies of the more frequent allele at each SNP locus, with the weights being the allele substitution effects (β) estimated. The summation of all MS for each individual yielded a GEBV:

$$GEBV = \sum_{j=1}^{N_m} T_{ij}\hat{g}_j,$$

where T_{ij} represents the marker genotype of animal i at SNP j , coded 0, 1, and 2 as described previously; \hat{g}_j is the estimate of SNP effect j ; and N_m is the number of SNP. The following nomenclature $GEBV_{No.SNP}^{Trait}$ was used

for clarity. The GEBV were derived for panels with all 37,959 markers as well as the top 200 SNP for each trait.

Genomic Predictions

The accuracy of prediction for the GEBV was assessed as the correlation between GEBV and the phenotype both within and across sire breeds.

Candidate Gene Analysis for RFI

For the trait of RFI, the 1:2 ratio of validation to training records was randomly replicated 5 times, and each replicate was analyzed using both RR-BLUP and BayesB methods so as to obtain SNP that consistently ranked within the top 200 because these were likely viable candidate genes for RFI. The number of times that an SNP was ranked within the top 200 after the 5 analyses yielded the “detection” frequency, expressed as a percentage. The positions of SNP with the greatest detection frequency were used to search for gene annotations and associated publications in Entrez Gene, HomoloGene, and PubMed.

RESULTS

Genetic Parameters and Variance Components

Phenotypic and genetic correlations between the 3 traits analyzed are shown in Table 1. Correlations were greatest between ADG and DMI and were least between ADG and RFI. There were significantly high phenotypic and genetic correlations for DMI with both RFI and ADG.

Table 2 gives variance components and genetic parameters for the traits evaluated. Estimates of phenotypic and genetic variance were greatest for DMI and least for ADG. Subsequently, single-trait heritability estimates for RFI and ADG were moderate to low, whereas DMI heritability was in the medium range.

Accuracy of GEBV Prediction

Table 3 shows trait-specific as well as between-trait correlations for GEBV with RFI, DMI, and ADG. For both BayesB and RR-BLUP with the 200 SNP panel,

Table 1. Genetic (below diagonal) and phenotypic (above diagonal) correlations between feed intake and efficiency traits¹

Item	RFI	ADG	DMI
RFI			
ADG	-0.03 ± 0.30		
DMI	0.51 ± 0.18	0.53 ± 0.18	

¹RFI = residual feed intake.

*Not significantly different from zero; all other phenotypic correlations were significant ($P < 0.001$).

Table 2. Variance components and parameter estimates for feed intake and efficiency traits

Model item ¹	ADG	DMI	RFI
Variance component			
Var(P)	0.08	2.09	0.85
Var(G)	0.02	0.86	0.25
Var(E)	0.05	1.23	0.61
Parameter			
h^2	0.28 ± 0.11	0.41 ± 0.12	0.29 ± 0.12

¹RFI = residual feed intake; Var(P) = phenotypic variance; Var(G) = direct genetic variance; Var(E) = residual variance; h^2 = direct heritability.

the highest correlation was observed between RFI and $GEBV_{200}^{RFI}$, whereas the lowest correlation was observed between DMI and $GEBV_{200}^{DMI}$. Accuracies between ADG with $GEBV_{200}^{RFI*}$ (GEBV obtained from estimates for association with ADG, but using SNP identified by training on RFI) were very low, whereas association between DMI and $GEBV_{200}^{RFI**}$ (GEBV obtained from estimates for association with DMI but using SNP identified by training on RFI) yielded higher correlations than trait-specific values. Correlations between traits and GEBV with all 37,959 markers included yielded lower correlations than those using only a subset of the top 200 SNP for both BayesB and RR-BLUP (Table 3). Generally, the RR-BLUP method yielded greater prediction accuracies than did BayesB, whereas prediction accuracy for RFI was greater than for DMI and ADG.

In Table 4, trait-specific correlations for different sire breeds are shown, for panels trained using both BayesB and RR-BLUP. The correlation of GEBV and RFI was slightly different within sire breed compared with the value obtained in across-breed comparisons. Further, for RR-BLUP, there was a pattern of differential accuracy within sire breed, with differences observed depending on what trait was being evaluated. For ADG, the Hybrid and Angus breeds tended to differ from each other, whereas for RFI, the Charolais sire breed tended to have a correlation pattern different from the Hybrid and Angus breeds (Table 4).

Candidate Genes for RFI

Eleven SNP associated with RFI were consistently ranked within the top 200 in 3 of 5 replicates (detection frequency of 60%) when the training data were analyzed using the RR-BLUP model. The greatest detection frequency obtained using the BayesB method was 40% (a total of 28 SNP had this detection frequency), whereas 92 SNP had a detection frequency of 40% or greater with the RR-BLUP method. Seven of the 11 SNP with detection frequency 60% were located either within a gene or close to a gene whose function could affect feed intake or feed efficiency (Table 5). Further, 4 of the 11 SNP were identified with a 40% detection frequency when using the BayesB method, whereas all 92 SNP from RR-BLUP had a detection frequency of at least 20% with the BayesB method. A total of 6 SNP

Table 3. Correlations of $GEBV_{200}$ and $GEBV_{37959}$ with trait phenotypes for BayesB and RR-BLUP analyses¹

Trait	GEBV	Method	Replication					Average
			1 (n = 203)	2 (n = 194)	3 (n = 255)	4 (n = 203)	5 (n = 198)	
ADG	$GEBV_{200}^{ADG}$	BAYESB ₂₀₀	0.12	0.34	0.26	0.12	0.28	0.22 ± 0.05
		RRBLUP ₂₀₀	0.00	0.52	0.46	0.42	0.46	0.37 ± 0.09
	$GEBV_{37959}^{ADG}$	BAYESB ₃₇₉₅₉	0.15					
		RRBLUP ₃₇₉₅₉	0.13					
DMI	$GEBV_{200}^{DMI}$	BAYESB ₂₀₀	-0.03	0.29	0.29	0.08	0.35	0.20 ± 0.07
		RRBLUP ₂₀₀	0.27	0.38	0.35	0.38	0.55	0.39 ± 0.05
	$GEBV_{37959}^{DMI}$	BAYESB ₃₇₉₅₉	0.24					
		RRBLUP ₃₇₉₅₉	0.25					
RFI	$GEBV_{200}^{RFI}$	BAYESB ₂₀₀	0.15	0.57	0.47	0.45	0.53	0.43 ± 0.07
		RRBLUP ₂₀₀	0.18	0.57	0.50	0.61	0.53	0.48 ± 0.08
	$GEBV_{37959}^{RFI}$	BAYESB ₃₇₉₅₉	0.12					
		RRBLUP ₃₇₉₅₉	0.11					
ADG	$GEBV_{200}^{RFI*}$	BAYESB_RFI ₂₀₀	0.06	0.06	0.00	0.06	-0.02	0.03 ± 0.02
DMI	$GEBV_{200}^{RFI**}$	RRBLUP_RFI ₂₀₀	-0.02	0.07	-0.06	-0.22	-0.12	-0.07 ± 0.05
		BAYESB_RFI ₂₀₀	0.29	0.47	0.31	0.33	0.22	0.32 ± 0.04
		RRBLUP_RFI ₂₀₀	0.41	0.42	0.25	0.43	0.48	0.40 ± 0.04

¹RFI = residual feed intake; BAYESB = Bayesian estimation using an algorithm called BayesBFast implemented in AlphaBayes (Hickey and Tier, 2009); RR-BLUP = random regression BLUP; GEBV = genomic breeding value. Standard errors for the average calculated as $SD/\sqrt{5}$. $GEBV_{200}^{RFI*}$ = GEBV obtained from ADG effects, with SNP selected for being associated with RFI; $GEBV_{200}^{RFI**}$ = GEBV obtained from DMI effects, with SNP selected for being associated with RFI.

were common between the 92 from RR-BLUP and the 28 from BayesB.

DISCUSSION

Diagnosis of convergence for the posterior estimates of SNP effects obtained after burn in was not carried out in this study. This is because the software program used for this analysis did not lend itself to such interrogation. However, AlphaBayes has been extensively tested for convergence in several simulated and real data sets. For data sets with 60,000 SNP, 60,000 Markov chain Monte Carlo samples with a burn-in of 10,000 samples was always enough to obtain convergence (J. Hickey, Animal Genetics and Breeding Unit, University of New England, Armidale, New South Wales, Australia, personal communication). In this study, we used 100,000 Markov chain Monte Carlo samples, with the

first 20,000 samples discarded as burn in to ensure that convergence was likely to be reached. As a further confirmation, 2 independent runs with different starting values were applied, and the estimates of SNP effects obtained thereafter (data not shown) had negligible differences between the runs. This gave an indication that the number of iterations chosen and the burn-in threshold were sufficient.

The strategy used in this analysis, to limit the number of SNP used for GEBV estimation to the top 200, was to maximize the chance of capturing a large number of SNP in greater linkage disequilibrium (**LD**) with underlying QTL as well as to reduce the number of redundant markers. Studies by Kizilkaya et al. (2010) and Zhong et al. (2009) have shown that panels that include QTL or markers in greater LD with QTL perform better when predicting across breeds or across multiple generations. The foregoing assumption is that markers

Table 4. Correlations (\pm SE, as the average of 5 replications) between $GEBV_{200}$ and trait phenotypes by sire breed for GEBV trained using BayesB and RR-BLUP¹

Method	Breed	ADG	DMI	RFI
Bayes	Across	0.22 ± 0.05	0.196 ± 0.07	0.43 ± 0.07
	Angus	0.25 ± 0.05	0.333 ± 0.07	0.55 ± 0.04
	Charolais	0.28 ± 0.13	0.200 ± 0.10	0.30 ± 0.12
	Hybrid	0.35 ± 0.10	0.261 ± 0.08	0.45 ± 0.08
	Undefined ²	0.17 ± 0.06	0.291 ± 0.08	0.31 ± 0.14
RR-BLUP	Across	0.37 ± 0.10	0.385 ± 0.05	0.48 ± 0.08
	Angus	0.36 ± 0.11	0.514 ± 0.04	0.54 ± 0.04
	Charolais	0.45 ± 0.13	0.319 ± 0.17	0.31 ± 0.08
	Hybrid	0.51 ± 0.08	0.495 ± 0.08	0.53 ± 0.09
	Undefined ²	0.39 ± 0.12	0.362 ± 0.11	0.44 ± 0.13

¹RFI = residual feed intake; RR-BLUP = random regression BLUP; GEBV = genomic breeding value.

²Sire breed not known.

Table 5. Locations, closest genes, and associated gene functions for SNP that ranked within the top 200 in 3 of 5 replicates of the training data analyzed using the random regression BLUP (RR-BLUP) method¹

SNP ID	Detection frequency, %	Position, bp	BTA	Distance to gene	Gene name	Gene function
ss86322201	60	147355780	1	21,611	ES 1 protein	Inhibition of cellular growth
ss86274038	60	45908516	24	51,911	SET binding protein 1	SET binding protein
ss86285204	60 ²	14738309	19	121,112	Chaperonin containing TCP1, subunit 6B	Mediates protein folding in the cytosol; folding of actin and tubulin
rs41641502	60 ²	14541593	19	5,326	Caspase regulator (CARP2)	Ubiquitin ligase/protein metabolism
rs42316404	60 ²	8999286	17	179,149	Endonuclease reverse transcriptase	Endonuclease reverse transcriptase
rs43557189	60	53208327	8	0	Transient receptor potential cation subfamily M, member 6 (TRPM6)	Ion exchange/Mg ⁺⁺ transport
rs42142693	60 ²	24107627	28	0	Bovine homolog of SLC25A16 solute carrier family (mitochondrial solute carrier)	Binding in transmembrane transport
rs41636768	60	55150035	18	NA	No gene annotation found	
ss105256889	60	44671099	21	NA	No gene annotation found	
rs41579807	60	14667205	19	NA	No gene annotation found	
rs41663853	60	14379998	28	NA	No gene annotation found	

¹SNP ID = National Center for Biotechnology Information rsSNP identification number; detection frequency = number of times an SNP ranks in the top 200 in 5 replicates for the RR-BLUP method; BTA = chromosome number; distance to closest gene (bases); rs = reference; ss = submitted; NA = no genes identified.

²Single nucleotide polymorphism also detected using the BayesB method with frequency 40%.

with a large effect signify markers in greater LD with the trait, and thus account for a larger portion of the trait variance. This strategy in itself has a practical implication in that by using a subset of SNP instead of the whole range of markers available in the analysis, equivalent prediction accuracy can be achieved without incurring the costs of genotyping associated with high-density SNP chips when used in a commercial application. In any case, it is very probable that for the 50K bovine SNP chip, only a subset of markers are useful for prediction purposes for various traits, and inclusion of additional SNP increases noise without a substantial change in prediction accuracy. This has been demonstrated in several studies (Luan et al., 2009; Kizilkaya et al., 2010) in which smaller subsets of markers have achieved accuracies equivalent to or greater than those of larger sets.

In this study, for all traits with 200 SNP markers, the BayesB method performed marginally less well than the RR-BLUP method. When allele substitution effects of SNP selected using RFI were reestimated using ADG as the training phenotype, the resulting GEBV ($GEBV_{200}^{RFI*}$) could not predict ADG for either BayesB or RR-BLUP. However, the same process with DMI resulted in a greater predictive accuracy than when using trait-specific GEBV ($GEBV_{200}^{DMI}$). The RFI SNP panel was able to achieve greater accuracies with DMI than when using the within-trait panel. This offers the prospect of a multitrait panel that can be used for both DMI and RFI. When using all available SNP (37,959), the predictive accuracy was much less than that observed with a smaller subset of 200 SNP. This informed the decision not to evaluate all 5 replicates with the full SNP panel (37,959), but rather to concentrate on the top 200 SNP.

Differences Between Methods

The performances of BayesB and RR-BLUP were quite varied, given the differences in assumptions for the Bayesian and BLUP methods. In the Bayesian methods, posterior estimates are influenced to a large extent by the choice of parameters given by the prior distribution. The biggest difference between the methods is in the assumptions associated with SNP variances. Typically, the genetic variance associated with each SNP in RR-BLUP is assumed to be small, and a uniform value of $\sigma_g^2 = \sigma_a^2/n$ is often used (as in this study because it is the one implemented in Alpha-Bayes), where σ_a^2 is the total genetic variance estimated by REML, σ_g^2 the variance associated with each SNP, and n is the number of loci. This SNP variance structure has been deemed unrealistic because many of the SNP are believed to have a small or no effect on trait variance, and many effects are fitted compared with the number of records present (Xu, 2003). An alternative definition, $\sigma_g^2 = \sigma_a^2/2\sum_j p_j(1-p_j)$, has been proposed (with p_j being the frequency of an allele at locus j), under assumptions of Hardy-Weinberg equilibrium and linkage equilibrium between QTL (Fernando et al., 2007).

Given that RR-BLUP fits all marker effects in the model, with marker variances obtained as a fraction of the total genetic variance, a larger number of markers would be needed to account for substantial genetic variance, especially for traits with low genetic variance. This means that for the RR-BLUP method, to achieve equivalent prediction accuracy compared with the Bayesian methods, larger SNP panels would be necessary, especially for ADG and RFI, whose trait variance is small compared with DMI, whereas n is the

same. Therefore, the results obtained in this study run contrary to that expectation. Such a result as seen in this study is possible where the SNP selected actually capture a reasonable proportion of QTL underlying the traits, which in turn reduces the number of SNP markers required in the prediction panel. The ability of the selected markers to be effective in prediction can be tested only by validation in an independent population.

Further, based on the suggestion by Meuwissen et al. (2001) that large QTL are heavily regressed back to the mean in RR-BLUP, the effects estimated by RR-BLUP will typically be small in comparison with those from Bayesian analyses, which fit only a fraction $(1 - \pi)$ of the total numbers of SNP available. This means that given that the SNP selection was accomplished by ranking SNP from greatest to least based on the magnitude of the allelic substitution effect, such regression would lower the rank of erstwhile larger QTL.

The use of a Bayesian model that includes a polygenic effect is expected to aid in effect estimation by properly partitioning the phenotypic variance to the various components. However, some studies (e.g., Calus and Veerkamp, 2007) have alluded to the minimal influence of including polygenic effects on accuracy in genomic selection analyses.

In all instances, the RR-BLUP method obtained greater correlations than the BayesB method. This difference may be related to the underlying genetic architecture of the traits. The infinitesimal model applied by RR-BLUP may fit the RFI and DMI data quite well compared with the notion of a few key QTL underlying the traits, as implemented in BayesB. Given that the range of metabolic processes that underlie RFI is quite large (Richardson and Herd, 2004) and considering recent discoveries suggesting that many putative genes may be associated with feed intake (Barendse et al., 2007; Chen et al., 2009), there is increasing evidence to suggest that a larger portion of the trait variance is under the influence of many QTL of small effect. This lends support to assertions that the assumptions

underpinning RR-BLUP may closely approximate the genetic architecture for RFI and DMI compared with Bayesian models. Still, a substantial number of QTL of large effect may be affecting these 2 traits.

On the other hand, given that little variation typically exists in ADG between animals both in this study and in similar studies, it is logical to assume that the genetic contribution toward this trait may be limited to a smaller number of QTL compared with RFI and DMI. Thus, the assumptions of the Bayesian model would be expected to favor a trait such as ADG. It is not immediately clear why this is not the case in this study, and further analysis with a larger data set will be necessary to verify this result. Estimates of variance components obtained from the 5 replicates of the training data are shown in Tables 6 and 7. Estimates obtained with the BayesB method were substantially greater than those obtained for RR-BLUP, and the proportion of the variance attributable to the SNP in BayesB was quite high. However, the correlations observed using both BayesB and RR-BLUP were less than those observed for the polygenic EBV (0.575, 0.504, and 0.602 for ADG, DMI, and RFI, respectively).

Within-Breed Correlations

The admixed population of crossbred animals used in this analysis consisted of steers sired by bulls of various breeds. Accuracy of prediction within sire breed showed greater variation between breeds when using the RR-BLUP method that when using the BayesB method. There was also greater prediction accuracy within breed than across breed.

This pattern of greater within-breed accuracy with RR-BLUP was clearly different from that observed using BayesB, for which the within-breed correlations were closer to the across-breed estimates. A possible reason for this may be the possibility that SNP selected using RR-BLUP may trace breed differences (SNP are optimized to capture breed differences) such that the

Table 6. Estimates of variance components for ADG, DMI, and residual feed intake (RFI) obtained in the 5 replicates of the training data with the random regression BLUP (RR-BLUP) method

Trait ¹	Parameter	Replicate					Average
		1	2	3	4	5	
ADG	ResVar	0.004 ± 0.003	0.021 ± 0.003	0.021 ± 0.003	0.021 ± 0.003	0.023 ± 0.004	0.018 ± 0.003
	GenVar	0.012 ± 0.004	0.005 ± 0.003	0.006 ± 0.004	0.006 ± 0.004	0.004 ± 0.005	0.006 ± 0.004
	SNPVar	0.036 ± 0.001	0.036 ± 0.002	0.038 ± 0.002	0.038 ± 0.002	0.030 ± 0.002	0.035 ± 0.002
DMI	ResVar	0.725 ± 0.088	0.720 ± 0.247	0.954 ± 0.220	1.040 ± 0.121	0.702 ± 0.177	0.828 ± 0.171
	GenVar	0.111 ± 0.095	0.734 ± 0.331	0.481 ± 0.228	0.193 ± 0.148	0.524 ± 0.226	0.408 ± 0.206
	SNPVar	0.069 ± 0.005	0.034 ± 0.005	0.032 ± 0.004	0.035 ± 0.003	0.034 ± 0.004	0.041 ± 0.004
RFI	ResVar	0.306 ± 0.025	0.349 ± 0.062	0.339 ± 0.118	0.385 ± 0.121	0.320 ± 0.090	0.340 ± 0.083
	GenVar	0.022 ± 0.025	0.166 ± 0.067	0.153 ± 0.137	0.200 ± 0.125	0.164 ± 0.106	0.141 ± 0.092
	SNPVar	0.067 ± 0.004	0.045 ± 0.004	0.044 ± 0.006	0.042 ± 0.008	0.046 ± 0.006	0.049 ± 0.006

¹Trait units are kilograms per day for ADG and DMI and kilograms of DM per day for RFI. ResVar = residual variance; GenVar = genetic variance; SNPVar = variance attributed to SNP as the difference between ResVar and ResVar + SNP (residual variance when SNP are included in the model as fixed effects).

Table 7. Estimates of variance components for ADG, DMI, and residual feed intake (RFI) obtained in the 5 replicates of the training data with the BayesB method

Trait ¹	Parameter	Replicate					Average
		1	2	3	4	5	
ADG	ResVar	0.017 ± 0.005	0.020 ± 0.010	0.023 ± 0.013	0.031 ± 0.009	0.019 ± 0.009	0.022 ± 0.009
	GenVar	0.007 ± 0.006	0.031 ± 0.015	0.026 ± 0.016	0.016 ± 0.010	0.023 ± 0.012	0.021 ± 0.012
	SNPVar	0.081 ± 0.006	0.083 ± 0.012	0.102 ± 0.012	0.096 ± 0.008	0.088 ± 0.010	0.090 ± 0.009
DMI	ResVar	0.582 ± 0.150	0.662 ± 0.151	0.720 ± 0.120	0.870 ± 0.244	0.771 ± 0.174	0.720 ± 0.184
	GenVar	0.143 ± 0.169	0.596 ± 0.163	0.564 ± 0.210	0.326 ± 0.234	0.289 ± 0.180	0.384 ± 0.191
	SNPVar	0.599 ± 0.081	0.523 ± 0.064	0.483 ± 0.043	0.507 ± 0.096	0.492 ± 0.049	0.521 ± 0.067
RFI	ResVar	0.247 ± 0.033	0.274 ± 0.076	0.193 ± 0.119	0.188 ± 0.097	0.198 ± 0.107	0.220 ± 0.086
	GenVar	0.048 ± 0.034	0.242 ± 0.085	0.339 ± 0.174	0.362 ± 0.126	0.310 ± 0.142	0.260 ± 0.112
	SNPVar	0.410 ± 0.049	0.364 ± 0.046	0.360 ± 0.039	0.390 ± 0.035	0.321 ± 0.041	0.369 ± 0.042

¹Trait units are kilograms per day for ADG and DMI and kilograms of DM per day for RFI. ResVar = residual variance; GenVar = genetic variance; SNPVar = variance attributed to SNP as the difference between ResVar and ResVar + SNP (residual variance when SNP are included in the model as fixed effects).

accuracy observed across breeds is confounded and not purely attributable to LD between SNP and underlying QTL.

Given that varying amounts of shrinkage are applied to SNP on the basis of differences in allele frequencies (the shrinkage term is the same for all SNP for the RR-BLUP method), any differences in allele frequencies between breeds for any locus will affect the size of the allele substitution effect and, by extension, the prediction accuracy. Habier et al. (2007) showed that for RR-BLUP, genetic relationships captured by the genetic markers affect prediction accuracy to a larger extent than in Bayesian methods because more markers are fit in the model. The consequence of this is that there would be an increase in prediction accuracy if validation animals became more related to training animals, especially if the markers were able to resolve relatedness more than the average relationship matrix.

A key issue in genomic selection of RFI is the utility of GEBV in selecting unphenotyped animals. In this study, the accuracies obtained were low compared with those seen in studies using dairy breeds, for which more accurate phenotypes are used to train SNP. A framework that allows incorporation of EPD and GEBV into a single unit of merit after appropriate weighting will be useful. The weights used could be derived from the reliability of the polygenic EBV and the percentage of genetic variance accounted for by the marker panels (VanRaden, 2001; Dekkers, 2007; Cerón-Rojas et al., 2008; Moser et al., 2009). A model that uses BLUP (Kachman, 2008) has also been proposed. Such a combined index for selection seems to be the best option, especially for beef cattle until such a time when large populations of animals have been tested for feed intake and GEBV accuracies are greater than the EBV accuracies obtained using traditional BLUP evaluations.

The number of animals in the training set also has a bearing on the accuracy of GEBV (Hayes et al., 2009). For RFI, a need therefore exists for increased testing of feed intake, despite the cost associated with such an undertaking. This is a priority for several Canadian

collaborations involving the Universities of Alberta and Guelph, Alberta Agriculture and Rural Development, and Agriculture and Agri-Food Canada.

Candidate Genes for RFI

Several studies have attempted to characterize the molecular basis of RFI. Barendse et al. (2007) and Sherman et al. (2008, 2010) describe a series of polymorphisms associated with RFI, but the usefulness of these SNP and associated genes in explaining the total RFI variance has yet to be determined. In this study, several SNP with a high detection frequency were in close proximity to genes that may be useful in controlling feed efficiency. Despite the fact that these SNP are associated with some genes of interest, their individual contribution was small. So far, no study involving RFI has shown a gene(s) with a significantly large effect, such that a candidate gene approach may not be the best strategy in characterizing the molecular basis of RFI. The SNP identified in this study may be more useful when seen as key elements of a gene network controlling RFI because the contribution of individual genes is likely to be small. Further research and analysis of gene networks for RFI is therefore warranted and is currently at an advanced stage in our laboratory.

Conclusions

In this study, accuracy of prediction, defined as the correlation between ADG, DMI, and RFI and trait-specific GEBV, was compared between SNP panels derived using 2 genomic selection methods, namely, BayesB and RR-BLUP. The RR-BLUP-derived GEBV achieved greater correlations with trait phenotypes, with accuracy being greatest for RFI. Differences in accuracy between sire breeds were observed with the RR-BLUP method. This may imply that significant differences may exist in SNP associated with RFI between the component breeds in the study population, and the SNP selected are consensus SNP that seem

to be inadequate for some breeds that are part of the composite population used. The accuracies obtained for all 3 traits were low, signaling a need for continued feed intake testing to acquire a large number of phenotyped animals, which may aid in better selection of SNP markers to be used for prediction as well as the continued evaluation of whether an admixed population such as ours can be useful in providing an across-breed prediction panel for RFI.

LITERATURE CITED

- Barendse, W., A. Reverter, R. J. Bunch, B. E. Harrison, W. Baris, and M. B. Thomas. 2007. A validated whole-genome association study of efficient food conversion in cattle. *Genetics* 176:1893–1905.
- Basarab, J. A., M. A. Price, J. L. Aalhus, E. K. Okine, W. M. Snelling, and K. L. Lyle. 2003. Residual feed intake and body composition in young growing cattle. *Can. J. Anim. Sci.* 83:189–204.
- Calus, M. P. L., and R. F. Veerkamp. 2007. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *J. Anim. Breed. Genet.* 124:362–368.
- Canadian Council on Animal Care. 1993. Guide to the Care and Use of Experimental Animals. Vol. I. 2nd ed. Can. Counc. Anim. Care, Ottawa, Ontario, Canada.
- Cerón-Rojas, J. J., F. Castillo-González, J. Sahagún-Castellanos, A. Santacruz-Varela, I. Benítez-Riquelme, and J. Crossa. 2008. A molecular selection index method based on eigenanalysis. *Genetics* 180:547–557.
- Chen, Y., C. Gondro, K. Quinn, B. Vanselow, P. F. Parnell, and R. M. Herd. 2009. Global gene expression profiling of Angus cattle selected for low and high net feed intake. Pages 30–33 in Proc. 18th Conf. Assoc. Advance. Anim. Breed. Genet. (Barossa Valley, South Australia, Australia).
- Crews, D. H., Jr. 2008. Genetic prediction of feed efficiency and input components. Pages 11–20 in Prediction of Genetic Merit of Animals for Selection, 9th Genetic Prediction Workshop, Beef Improve. Fed., Kansas City, MO. Beef Improvement Federation (BIF), Raleigh, NC.
- Dekkers, J. C. M. 2007. Prediction of response to marker-assisted and genomic selection using selection index theory. *J. Anim. Breed. Genet.* 124:331–341.
- Fernando, R. L., D. Habier, C. Stricker, J. C. M. Dekkers, and L. R. Totir. 2007. Genomic selection. *Acta Agric. Scand. A* 57:192–195.
- Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson. 2008. ASReml User Guide Release 3.0. VSN Int. Ltd., Hemel Hempstead, UK.
- Goonewardene, L. A., Z. Wang, M. A. Price, R.-C. Yang, R. T. Berg, and M. Makarechian. 2003. The effect of udder type and calving assistance on weaning traits of beef and dairy × beef calves. *Livest. Prod. Sci.* 81:47–56.
- Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177:2389–2397.
- Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard. 2009. Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet. Sel. Evol.* 41:51.
- Hickey, J. M., and B. Tier. 2009. AlphaBayes (Beta): Software for Polygenic and Whole Genome Analysis. User Manual. University of New England, Armidale, Australia.
- Johnston, D., H. Graser, and B. Tier. 2008. Integration of DNA markers into a BREEDPLAN tenderness EBV. Pages 83–87 in Prediction of Genetic Merit of Animals for Selection, 9th Genetic Prediction Workshop, Beef Improve. Fed., Kansas City, MO. Beef Improvement Federation (BIF), Raleigh, NC.
- Kachman, S. 2008. Incorporation of marker scores into national genetic evaluations. Pages 92–98 in Prediction of Genetic Merit of Animals for Selection, 9th Genetic Prediction Workshop, Beef Improve. Fed., Kansas City, MO. Beef Improvement Federation (BIF), Raleigh, NC.
- Kizilkaya, K., R. L. Fernando, and D. J. Garrick. 2010. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. *J. Anim. Sci.* 88:544–551.
- Luan, T., J. A. Woolliams, S. Lien, M. Kent, M. Svendsen, and T. H. E. Meuwissen. 2009. The accuracy of genomic selection in Norwegian Red cattle assessed by cross-validation. *Genetics* 183:1119–1126.
- MacNeil, M. D., J. D. Nkrumah, B. W. Woodward, and S. L. Northcutt. 2010. Genetic evaluation of Angus cattle for carcass marbling using ultrasound and genomic indicators. *J. Anim. Sci.* 88:517–522.
- Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829.
- Moser, G., B. Tier, R. E. Crump, M. S. Khatkar, and H. W. Raadsma. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genet. Sel. Evol.* 41:56.
- Richardson, E. C., and R. M. Herd. 2004. Biological basis for variation in residual feed intake in beef cattle. 2. Synthesis of results following divergent selection. *Aust. J. Exp. Agric.* 44:431–440.
- Scheet, P., and M. A. Stephens. 2006. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 78:629–644.
- Sherman, E. L., J. D. Nkrumah, and S. S. Moore. 2010. Whole genome SNP associations with feed intake and feed efficiency in beef cattle. *J. Anim. Sci.* 88:16–22.
- Sherman, E. L., J. D. Nkrumah, B. M. Murdoch, and S. S. Moore. 2008. Identification of polymorphisms influencing feed intake and efficiency in beef cattle. *Anim. Genet.* 39:225–231.
- VanRaden, P. M. 2001. Methods to combine estimated breeding values obtained from separate sources. *J. Dairy Sci.* 84(E. Suppl.):E47–E55.
- Xu, S. 2003. Estimating polygenic effects using markers of the entire genome. *Genetics* 163:789–801.
- Zhong, S., J. C. Dekkers, R. L. Fernando, and J. L. Jannink. 2009. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. *Genetics* 182:355–364.