Running head: Genomic prediction for feed efficiency

Accuracy of genomic predictions for feed efficiency traits of beef cattle using 50K and imputed HD genotypes[1]

D. Lu[*†], E. C. Akanno[*], J. J. Crowley[*‡] , F. Schenkel[§], H. Li[§], M. De Pauw[*], S. S. Moore[*¥], Z. Wang[*], C. Li[*¶], P. Stothard[*], G. Plastow[*], S. P. Miller[*†§] and J. A. Basarab[*ǂ 2]

[*]Livestock Gentec, Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, AB, Canada; [†]AgResearch, Invermay Agricultural Centre, Post Box 50034, Mosgiel 9053, New Zealand; [‡]Canadian Beef Breeds Council, Calgary, Alberta T2E 7H7, Canada; [§]Centre for Genetic Improvement of Livestock, Department of Animal and Poultry Sciences, University of Guelph, ON, Canada; [¥]Centre for Animal Science, Queensland Alliance for Agriculture and Food Innovation, University of Queensland, St Lucia, Australia; [¶]Lacombe Research Centre, Agriculture and Agri-Food Canada, 6000 C & E Trail, Lacombe, AB, Canada; [ǂ]Lacombe Research Centre, Alberta Agriculture and Forestry, 6000 C & E Trail, Lacombe, AB, Canada

---

[1] Acknowledgement: The authors gratefully acknowledge funding support from Alberta Livestock and Meat Agency Ltd., Alberta Innovates Bio Solutions, and Alberta Agriculture and Rural Development (ARD) and in-kind contribution in animals, facilities and people received from Agriculture and Agri-Food Canada (AAFC), Lacombe Research Centre, with special thanks to Cletus Sehn, Ken Grimson, and their staff for animal care and management. Special thanks are also extended to Cathy Bryant and Sheldon Johnston of ARD and Lisa McKeown (Livestock Gentec) for project coordination, data collection and data base

management and to Barry Irving and staff at the Kinsella ranch, Neil French and staff at Olds College (Alberta, Canada), Stuart Thiessen and staff at Namaka Farms near Strathmore (Alberta, Canada), Todd McKinnon and staff at Three Cross Ranches near Airdire (Alberta, Canada), Beefbooster Ltd., (Calgary, Alberta, Canada) and Deseret Ranches of Alberta (near Raymond, Alberta) for in-kind contribution of animals, facilities and data. GrowSafe Systems Ltd is also acknowledged for their continual contributions of technical expertise, hardware and software. Assistance with parentage analysis provided by Drs. Ricardo Ventura (BIO) and Mehdi Sargolzaei (University of Guelph) is gratefully acknowledged. The management of the Angus and Charolais herds and feed efficiency data collection were supported by the AAFC livestock and genomics program fund and the AAFC peer-reviewed A-base project fund (RBPI-1139). Other funding supports from Ontario Cattlemen's Association, Canadian Beef Cattle Research Council, Agriculture and Agri-Food Canada Science Cluster, Canadian Simmental Association, Agriculture & Agri-Food Canada's Growing Forward Program and Ontario Ministry of Agriculture and Food are hereby acknowledged. Thanks to University of Guelph colleagues Drs. Kendall Swanson and Ira Mandell for supporting the beef database, Dr. Gordon VanderVoort for database support, Charlie Watson and staff at the Elora Beef Research Centre and Tim Caldwell (University of Guelph) for technical assistance.

[2]Corresponding author: John.Basarab@gov.ab.ca

ABSTRACT

The accuracy of genomic predictions can be used to assess the utility of dense marker genotypes for genetic improvement of beef efficiency traits. This study was designed to test the impact of genomic distance between training and validation populations, training population size, statistical methods and density of genetic markers on prediction accuracy for feed efficiency traits in multi-breed and crossbred beef cattle. A total of 6,794 beef cattle data collated from various projects and research herds across Canada were used. Illumina BovineSNP50 (50K) and imputed Axiom Genome-Wide BOS 1 Array (HD) genotypes were available for all animals. The traits studied were dry matter intake (DMI), average daily gain (ADG) and residual feed intake (RFI). Four validation groups of 150 animals each, including Angus (AN), Charolais (CH), Angus-Hereford crosses (ANHH), and a Charolais-based composite (TX) were created by considering the genomic distance between pairs of individuals in the validation groups. Each validation group had seven corresponding training groups of increasing sizes (n = 1000; 1999; 2999; 3999; 4999; 5998 and 6644), which also represent increasing average genomic distance between pairs of individuals in the training and validations groups. Prediction of genomic breeding values (GEBV) was carried out using genomic best linear unbiased prediction (GBLUP) and Bayesian method C (BayesC). The accuracy of genomic predictions was defined as the Pearson's correlation between adjusted phenotype and GEBV (*r*), unless otherwise stated. Using 50K genotypes, the highest average *r* achieved in purebreds (AN, CH) was 0.41 for DMI, 0.34 for ADG and 0.35 for RFI, while in crossbreds (ANHH, TX) it was 0.38 for DMI, 0.21 for ADG and 0.25 for RFI. Similarly, when imputed HD genotypes were applied in purebreds (AN, CH), the highest average *r* was 0.14 for DMI, 0.15 for ADG and 0.14 for RFI, while in crossbreds (ANHH, TX) it was 0.38 for DMI, 0.22 for ADG, 0.24 for RFI. The *r* of GBLUP predictions were greatly reduced with increasing genomic average distance as compared to those from BayesC predictions. The results indicate that 50K genotypes, used with BayesC, were more effective for predicting GEBV in purebred cattle. Imputed HD genotypes found utility when dealing with composites and crossbreds. Formulation of a fairly large training set for genomic predictions in beef cattle should consider the genomic distance between the training and target population.

INTRODUCTION

The availability of affordable high density genotyping services for cattle provides an opportunity for the application of genomic selection (GS) for genetic improvement of economically important traits in beef cattle. These genotypes can be used to produce genomic estimated breeding values (GEBV) for a group of selection candidates without phenotypes as proposed by Meuwissen et al. (2001). The accuracy of genomic predictions is the key to successful application of GS and largely depends on the marker-QTL linkage disequilibrium (LD) and the genetic relationship among animals in the training and validation groups (Habier et al., 2007). Because accuracy cannot be assessed in the population used for training the SNP effects, care is required in choosing an informative training population for beef cattle where many breeds and distantly related animals are used to produce commercial cattle. In addition, accuracy of GS can be greatly reduced in multi-breed and crossbred populations due to inconsistent LD across multiple populations (de Roos et al. 2009). The use of high density markers and large training sets was proposed by Goddard and Hayes (2007) as a way to improve accuracy of GS in crossbred populations. A low cost solution called genotype imputation (Howie et al., 2009; Sargolzaei et al., 2014) is currently available for increasing the density of markers. Apart from reports by Chen et al. (2013) and Khansefid et al. (2014), research into the accuracy of genomic predictions for feed efficiency using genotypes from the BovineSNP50 BeadChip (50K; Illumina Inc. San Diego, CA, USA) and the Axiom Genome-Wide BOS 1 Array (HD; Affymetrix Inc., Santa Clara, CA) are limited in literature. The objective of the present study was to test the impact of genomic distance between training and validation populations, size of reference population, statistical approaches and marker density on prediction accuracy for feed efficiency traits in multi-breed and crossbred beef cattle.

MATERIALS AND METHODS

All management and procedures involving live animals where applicable, conformed to the guidelines outlined in the Canadian Council on Animal Care (CCAC, 1993), otherwise, existing datasets from the various Canadian research herds was used.

Animals and Phenotypic Records

A total of 6,796 beef cattle data were collated from various projects and research herds across Canada, including 3,692 from the Phenomic Gap Project (PG1) based at Lacombe Research Centre (LRC), Lacombe, AB; 875 Angus (AN), 569 Charolais (CH) and 906 beef-dairy hybrids (HYB) from the University of Alberta's Roy Berg Kinsella Research Ranch (KRR), Kinsella, AB; and 754 multi-breed and crossbred cattle mainly Angus-based with various proportions of Simmental (SM), Piedmontese (PI), Gelbvieh (GV), CH and Limousin (LM) from the University of Guelph's Elora Beef Cattle Research Station (ERS), Elora, ON. The PG1 animals which represent over 50% of the dataset included 1,225 Angus-Hereford (ANHH) and 353 Charolais-Red Angus (CHAR) crosses from LRC, 272 HYB from KRR, 1,526 crossbreds from three commercial herds and 316 Hereford (HH) cattle from various seed stock producers. More details on each of these herds and datasets were reported by Chen et al. (2013), Lu et al. (2013), López-Campos et al. (2013) and Akanno et al. (2014a). In terms of breeds, the whole dataset consisted of 968 AN, 572 CH, 316 HH, 17 SM, 17 LM, 1,225 ANHH, 484 ANSM, 353 CHAR, 1,105 TX (Beefbooster composite that are heavily influenced by CH with infusion of Holstein, Maine Anjou, and Chianina; http://www.beefbooster.com), 1,178 HYB and 561 animals of other breed combinations.

Phenotypic records, including dry matter intake (DMI), average daily gain (ADG) and residual feed intake (RFI) were available for all of the 6,796 animals. Phenotype collection were described in details by Basarab et al. (2011), Chen et al. (2013), and Lu et al. (2013). Briefly, feed intake (FI) and body weight (BW) were collected in post-weaning performance tests. For the KRR animals, performance test were approximately 120 days with FI measured daily and BW recorded every other week. The PG1 animals had test periods varying between 76 and 112 days, with FI measured daily, and BW recorded on two consecutive days at the beginning and the end of test, and around 28 day intervals during the test. The ERS animals had an average test length of 111 days with daily FI measurement and 28 day weight recording. Residual feed intake was the difference between observed DMI and expected DMI being modelled on ADG, $BW^{0.75}$ and ultrasound backfat (BFT) measured at end of test. The data was collated and adjusted for variation among the datasets (Crowley et al., 2014). Briefly, animals were filtered out based on the following criteria: 1) missing observation of any of the traits or model effects of interest; 2) animals older than 450d at the start of test; 3) any record with greater than 3 standard deviations from the mean estimated within dataset of any or all of ADG, DMI, $BW^{0.75}$ and BFT; and 4) animals belonging to a contemporary group (CG) with less than five individuals. The CG was defined as data source, herd, year, group, and pen. Feeding trials for ERS animals were included in their group.

Genotype data

All animals with phenotypes were genotyped with the 50K beadchip version 1 or 2. Genotypes from the various Canadian research sources were corrected for any discrepancy in the strands and allele designation using guidelines provided by Illumina (2006) before merging into a single genotype file. For the 50K genotypes, quality control (QC) was carried out to remove SNPs if one of the following was true: SNP with minor allele frequency (MAF) < 0.01, call rate < 0.90, and heterozygosity excess > 0.15. A total of 42,610 SNPs passed the QC and entered into subsequent analyses. Animals with HD genotypes (n=4,522), from different Canadian cattle breeds, included AN (469), CH (474), HH (476), Holstein (447), LM (461), SM (417), GV (417), Beefbooster composite (478), ERS crossbreds (504) and Alberta crossbreds (379) were used as multi-breed reference dataset for imputing from 50K to HD genotypes.

The 6796 50K genotypes collated from various Canadian research herds were coded in two formats: Illumina A/B and FORWARD/FORWARD, while the Affymetrix HD genotypes were coded using +/+ format. Then, as a first step, all 50K genotypes were accordingly converted to the +/+ format prior to imputation based on the DNA strand designation and allele determination in each coding format.

Single nucleotide polymorphisms in the HD chip that did not map to the *Bos taurus* UMD 3.1 reference assembly, SNPs located on sex chromosomes, and SNPs not present in the Run 4.0 of the 1,000 bull genomes project were excluded, resulting in 508,868 SNPs in the reference HD genotypes. The software FImpute v2.2 (Sargolzaei et al., 2014) was used for imputing the HD genotypes of all 6796 beef cattle, using default parameters and population-based imputation. Quality control criteria applied to the HD genotypes were the same as previously described for 50K genotypes, leaving 468112 SNPs on 29 autosomes for subsequent analyses.

Statistical Model and Analysis

Two of the 6796 animals were removed from the dataset due to inconsistent pedigree information. The final number of animals used for this study was 6794. The first analysis was to investigate population stratification among the animals using a classical multidimensional scaling (MDS) approach and all 42,610 SNPs to obtain the first six dimensions of genetic dissimilarity among the animals (Purcell et al., 2007). The six dimensions of the MDS were fitted as covariates in model [1] used to produce the adjusted ADG and DMI. Adjusted RFI was produced from model [1] without backfat as a covariate.

$$y_{ijkm} = \mu + \gamma_1(age_i) + \gamma_2(bf_i) + cg_k + \sum_{j=1}^{6} \beta_j b_j + e_{ijkm} \qquad [1]$$

where y_{ijkm} is the phenotype of animal; μ the overall mean; γ_1 and γ_2 the regression coefficients for fixed effects of age and backfat, respectively; *cg* the *k*th contemporary group that consisted of sex, herd-year, and data source; β_j the linear regression coefficient of the *j*th dimension and b_j the coordinate of the *j*th dimension; and e_{ijkm} the residual. The residual was used as adjusted phenotype to compute GEBV in both genomic best linear unbiased prediction (GBLUP) and BayesC approaches. In addition, model [1] was expanded into a three-trait multi-variate model that included ADG, DMI, and RFI as response variables, and a random animal effect that uses pedigree information for estimating genetic parameters of studied traits.

The GBLUP approach was applied to the following statistical mixed model,

$$y = 1\mu + Zu + e \qquad [2]$$

where *y* is the vector of the adjusted phenotype values from model [1], *Z* the incidence matrix for all animals with genotype, *u* the vector of additive effect of individual SNP, and *e* the vector of random error. The mixed model equation was:

$$\begin{bmatrix} 1'_n 1 & 1'_n Z \\ Z' 1_n & Z'Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} 1'_n y \\ Z' y \end{bmatrix} \qquad [3]$$

where *G* in equation [3] represents the genomic relationship matrix that follows the formula by VanRaden et al. (2009). Pedigree information was not used. Phenotypic data of validation animals were assumed unknown, and their GEBV were obtained by solving equation [3]. The GBLUP approach was implemented using the GEBV software by Sargolzaei et al. (2009).

In the Bayesian approach, the fraction of loci with no effect, π, was estimated using method BayesCπ to be approximately 0.77, 0.85, and 0.95 for RFI, ADG and DMI, respectively, with the 50K genotypes, and 0.99 for the 3 traits with the HD genotypes.

Thereafter, method BayesC was used with corresponding π value to simultaneously estimate SNP effects across the entire genome using the following mixed model

$$y_i = \mu + \sum_{j=1}^{k} X_{ij} m_j + e_i \qquad [4]$$

where y_i represents the adjusted phenotype of individual i from model 1, X_{ij} is the vector of indicator variables representing the genotypes of the j^{th} SNP for individual i, m_j is the random effect for j^{th} SNP, k is the total number of SNPs, and $e_i \sim N(0, \sigma_e^2)$ is the random residual. The prior for m_j depends on the variance $\sigma_{m_j}^2$ and the prior probability π as follows

$$m_j \middle| \pi, \sigma_{m_j}^2 = \begin{cases} 0 & given\ \pi, \\ \sim N\left(0, \sigma_{m_j}^2\right) & given\ (1-\pi) \end{cases} \qquad [5]$$

$$\sigma_m^2 | \upsilon_m,\ S_m^2 \sim \upsilon_m S_m^2 \chi_{\upsilon_m}^{-2}\ ,$$

where $S_m^2 = \frac{\tilde{\sigma}_m^2(\upsilon_m - 2)}{\upsilon_m}$ and $\tilde{\sigma}_m^2 = \frac{\tilde{\sigma}_s^2}{(1-\pi)\sum_{j=1}^{k} 2p_j(1-p_j)}$, with $\tilde{\sigma}_s^2$ being the genetic variance explained by all markers, υ_m the degree of freedom of 4 and p_j the allele frequency of j^{th} SNP. The BayesC method uses a common σ_m^2 for all markers (Habier et al., 2011). Markov Chain Monte Carlo methods with 50,000 iterations were used to generate posterior mean estimates of SNP effects after discarding 5,000 iterations as burn-ins. The Bayesian analyses were carried out using software GenSel v4.58R of Fernando and Garrick (2013).

Pearson's correlation between adjusted phenotype and GEBV (r) was used to evaluate the accuracy of predictions for various reference and validation populations tested, unless otherwise stated. Realized accuracy (equivalent to $\frac{r}{\sqrt{trait\ heritability}}$, (Hayes et al., 2010)) was used only to compare results from this study with documented findings.

Training and Validation Scenarios Investigated

Genomic distance was computed for pairs of animals using Euclidean metric and the six MDS coordinates. Validation groups of 150 animals each were created for AN, CH, ANHH, and TX breed groups. Animals in each validation group were chosen to minimize genomic distance between pairs of animals in the group. This approach is based on our observation that a given group of prediction animals could be split into subsets of individuals that are genomically closely related and therefore might be best predicted by different groups of training individuals. Each animal chosen for validation appeared in only one validation group. There were three validation groups for CH animals, and five groups for each of AN, ANHH, and TX breed groups. Once a validation group was formed, seven training groups of increasing sizes were created from the remaining 6644 animals. The first training group consisted of 1000 animals, each of which had the shortest average genomic distance with animals in the validation group. The second training group included 1000 animals in the first training group, in addition to 999 animals chosen from the remaining individuals based on shortest average genomic distance with animals in the validation group. This process was repeated for training groups 3, 4, 5, and 6. Training group 7 contained all 6644 animals.

RESULTS

Descriptive statistics and genetic parameters of studied traits

Details on animal performance and feed efficiency traits are presented in Table 1, which was adopted from Crowley et al. (2014). For 6794 animals used in this study, phenotypic means (±SD) for ADG, DMI, and RFI were 1.45±0.39 kg/d, 9.23±1.59 kg/d and 0.00±0.63 kg/d, respectively. Heritability estimates (±SE), using the pedigree relationship matrix, were 0.38±0.04, 0.48±0.04, and 0.38±0.04 for ADG, DMI and RFI, respectively, while the genetic correlations between ADG and DMI, ADG and RFI, DMI and RFI were 0.69, 0.01, and 0.56, respectively.

Genomic distance between training and validation populations

Table 2 shows the average genomic distance between pairs of individuals in the training and validation groups. Within a given validation group, average genomic distance between pairs of individuals in the training and validation groups increases as individuals that are less related to the validation population are included in the training population. To assist with visualizing the genomic distance between training and validation animals, genomic distance was compared to the proportion of the genome being different between two individuals (Figure 1). The genomes of two individuals for genotypes coded as 0, 1, and 2 was 100% different when genotype difference at every single locus was 2. Linear regression of proportion of genome difference on genomic distance, both based on the 50K genotypes, was carried out for each validation and their training groups and the result is embedded in Figure 1. The coefficients of determination (R^2) for all validation groups ranged from 0.90 to 0.99, implying that most of the variations in the genome difference around the mean were explained by the genomic distance. The intercepts of the regression equation showed slightly greater genome difference between the crossbred validation group (ANHH and TX; 27.64 and 28.23) and their training groups than between the purebred validation groups (AN and CH; 26.14 and 27.22) and their training groups. However, the slopes of the regression equation for AN and CH (227.11 and 163.10, respectively) were larger than those for ANHH and TX (130.69 and 100.46, respectively), indicating faster increases in genome difference as genomic distance increases in the AN and CH than in the ANHH and TX validation groups. This reflects the fact that the AN and CH animals very different genomically to the crossbred ones, therefore genome differences between AN, CH validations and their training groups increased rapidly as the training groups expanded to include the crossbred individuals.

Average genomic distance between pairs of training and validation animals was also computed based on the imputed HD genotypes, and presented in Table 2. Apart from the relationship between genomic distance and number of animals in the training groups already observed with the 50K genotypes, the average genomic distance appeared to be shortened when the imputed HD genotypes were used. Validation animals therefore appeared to be more closely related to individuals in training groups.

Accuracy of genomic predictions using 50K and imputed HD

The correlation between adjusted phenotype and GEBV (*r*) in AN, CH, ANHH, and TX validation groups across the studied traits using GBLUP and BayesC are presented in Table 3 for 50K genotypes and Table 4 for imputed HD genotypes. On average, when using 50K and imputed HD genotypes, BayesC showed slightly greater *r* across the studied traits compared to GBLUP (Tables 3 and 4). Within a given trait and validation population for the 50K genotypes (Table 3), the *r* tended to decrease with increasing size of training population which represented an increasing average genomic distance between pairs of individuals in the training and validation groups (Table 2). The *r* decreased faster with increasing genomic distance when using the GBLUP method compared to BayesC, which tended to be more stable (Table 3).

Figure 2 shows the relationship between *r* and genomic distance across the studied traits and validation groups. For each 0.0001 increment in genomic distance, *r* changed by 0.017, 0.022, 0.023 and 0.049 in AN, CH, ANHH and TX validation groups, respectively, when using GBLUP method to predict RFI. While the correlation *r* for all traits in the AN group dropped consistently when more animals were added to the initial training group, this trend was not observed in the BayesC predictions for the CH animals. The correlation *r* for their predictions remained relatively stable as the training group increased in size, and also observed in BayesC predictions of RFI in the ANHH animals, as well as RFI and DMI in the TX animals. Nevertheless the correlation *r* of ADG, DMI predictions for the ANHH animals, as well as ADG prediction for the TX animals appeared to increase slightly when their training group size increased from 1000 to 3999 or 4999, and remain relatively stable onwards. In general the highest *r* were 0.35 for RFI, 0.34 for ADG and 0.41 for DMI, on average, while the highest *r* in crossbred cattle (ANHH and TX) were 0.25 for RFI, 0.21 for ADG and 0.38 for DMI, on average (Table 3). When the imputed HD was used, the highest *r* was 0.14 for RFI, 0.15 for ADG and 0.14 for DMI in purebreds (AN and CH), on average, while the highest *r* in crossbred cattle (ANHH and TX) were 0.24 for RFI, 0.22 for ADG and 0.38 for DMI, on average (Table 4). Crossbred validation groups (ANHH and TX) showed greater *r* across the studied traits and statistical methods, on average, than purebred validation groups (AN and CH).

Because the accuracy of GS should be the correlation between GEBV and the true breeding value which is assumed unknown, Table 5 presents a realised accuracy computed for AN and TX validation populations across traits and for 50K genotypes. Using GBLUP gave realised accuracies in the range of 0.36 – 0.49 for RFI, 0.29 – 0.37 for ADG, and 0.51 – 0.63 for DMI, while BayesC gave generally higher realised accuracies of 0.49 – 0.55 for RFI, 0.37 – 0.43 for ADG, and 0.58 – 0.63 for DMI in the Angus validation population. Similarly, in the Beefbooster composite validation population, realised accuracies from GBLUP ranged from 0.20 – 0.33 for RFI, 0.16 – 0.19 for ADG, and 0.30 – 0.49 for DMI, while BayesC realised accuracies ranged from 0.31 – 0.38 for RFI, 0.23 – 0.27 for ADG, and 0.45 – 0.54 for DMI. Table 5 also shows the regression coefficient in brackets for regressing adjusted phenotypes on GEBV across the various scenarios and methods studied. The coefficient for all traits is expected to be equal to one where values greater or lower than one reflects an under or over estimation of GEBV, respectively. The GBLUP predictions were all overestimated with levels of biasness going up with increasing size of the reference population, which coincides with increasing genomic distance between training and validation groups. On the contrary, the BayesC predictions were underestimated though not as severely as the GBLUP predictions were over overestimated. The degree of over-prediction with GBLUP was greatly reduced by replacing 50K genotypes with HD genotypes; however, this replacement slightly increased under-prediction with BayesC (Figure 3).

DISCUSSION

This study applied a GS approach based on bovine 50K and imputed HD genotypes to determine the accuracy of GEBV for DMI, ADG, and RFI in a multi-breed and crossbred beef cattle validation population that was created by considering the genomic distance between pairs of individuals in the training and validation groups. The mean performance and estimated genetic parameters for the studied traits were typical of beef cattle in North America and were in agreement with previous reports (Arthur et al. 2001; Nkrumah et al. 2006; Berry and Crowley 2012).

Genomic predictions were carried out using GBLUP and BayesC statistical methods. When comparing the results from these two methods, it is important to consider their fundamental differences in approach and assumptions. The GBLUP approach uses a genomic relationship matrix of which covariance between pairs of individuals was estimated and expected to be deviated from a numerator relationship matrix based on pedigree due to allele segregation at QTL (Goddard et al., 2011; Habier et al., 2013), and sampling error associated with genomic position (Goddard et al., 2011). Though the true position of a QTL is unknown, allele segregation at the QTL can be inferred by segregation of SNPs surrounding it, which depends on LD among the SNPs. This inherent LD is affected by 1) traits of interest which are generally assumed to be controlled by different number of QTL with various effect sizes (Shrimpton and Robertson, 1988; Hayes and Goddard, 2001); 2) population structure such that homogeneous populations (small effective population size, N_e) possess higher LD than admixed or crossbred populations (Meuwissen et al., 2002; Sargolzaei et al., 2008; de Roos et al., 2008; Lu et al., 2012); and 3) small physical distance between SNPs and QTL which ensures higher LD between them as observed in LD studies (for e.g, Dunning et al., 2000; Hayes et al., 2003; Laido et al., 2014), that is, higher LD is achieved with higher SNP density. These three elements also contribute to GEBV predictions using a Bayesian approach.

We found an advantage in the accuracy of GEBV predicted for DMI, ADG and RFI using BayesC over GBLUP. This finding agrees in principle with reports by Habier et al. (2010) and Gunia et al. (2014), but disagrees with the results by Lee et al. (2014). BayesC detects QTL and estimates SNP effects with a small proportion of SNPs having large effects on traits (Habier et al., 2011). Because QTL detection is involved, LD between QTL and its surrounding SNPs becomes important and the BayesC method exploits this LD advantage (Habier et al., 2007; Habier et al., 2011). The SNPs surrounding large QTL, such as those for DMI on chromosome 6 (Lu et al., 2013; Saatchi et al., 2014), have stronger LD with the QTL, and thus their effect is more accurately estimated than those SNPs around small QTL for RFI (Lu et al., 2013; Saatchi et al., 2014), therefore, this could be a reason why BayesC predicted GEBV for DMI much better than it did for RFI. On the contrary, in a GBLUP approach, traits are assumed to be controlled by an infinite number of genes, each with very small effect (Fisher, 1918), which could explain the slightly lower accuracy of GEBV for DMI and RFI. In addition, the coefficients of the genomic relationship matrix do not reflect genetic covariance between two individuals at a QTL in the case of no LD between the QTL and the surrounding SNPs (Habier et al., 2013) which may have contributed to lower prediction accuracy for RFI using the GBLUP method. The implication is that a prior knowledge of genetic architecture of traits being analysed may be more important for choosing the right statistical approach, although different approaches for different traits may be problematic for routine evaluations for a given situation.

The ability to predict genomic breeding values within and between populations partly depends on the extent of LD in the population (Goddard et al., 2011; Habier et al., 2013). More extensive LD means more variation in genomic relationship, and thus, requires fewer SNP for the prediction of these relationships (Goddard et al., 2011). The LD in a crossbred population extends over shorter ranges compared to purebred populations due to recombination of chromosome segments. Therefore, variation in relationship is small, and requires a larger number of SNPs to predict these relationships accurately (Goddard et al., 2011). A larger number of SNPs is also needed to reduce the error caused by SNP positions being sampled across the genome (Goddard et al., 2011). However, Su et al. (2012) reported no gain in prediction accuracy when using imputed 777K genotypes *versus* the 50K in Nordic Holstein and Red Dairy cattle. On the contrary, Gunia et al. (2014) reported a very slight reduction in GEBV accuracy when SNP density was increased from 50K to 777K, using a GBLUP approach, though little improvement in GEBV accuracy was observed when BayesC was applied in French Charolais. Our results showed a large reduction in *r* in the purebred validation group (AN and CH) when imputed HD genotypes were used for GBLUP and BayesC predictions. The HD genotypes in this study were inferred from the 50K genotypes, using a population imputation approach with a multi-breed and crossbred reference population. Table 2 shows that the HD genotypes imputed in this study made genomic distance between pairs of individuals shorter than it appeared as estimated with the 50K genotypes. This might not have reflected true relationship among the animals, especially between the pure AN, CH and the crossbreds. Allele frequencies (*p*) at imputed loci in the AN and CH may have been suppressed by those from other breeds and crossbreds in the reference population, such that the scalar ($2\sum p_i(1-p_i)$) in VanRaden's genomic relationship formula (VanRaden et al., 2009) applied in the GBLUP method may well accurately represent the crossbred animals, for instance, the ANHH and TX crossbreds, leading to improved prediction accuracy when using imputed HD genotypes in the crossbred validation groups (ANHH and TX). Similarly for the BayesC method, estimation of SNP effects in the training population may have been driven by the crossbred allele frequency leading to a reduction in prediction accuracy when using the imputed HD genotypes in purebred cattle (AN and CH), however, small improvement in *r* for crossbred cattle (ANHH and TX) were observed. Moghaddar et al. (2015) reported a somewhat similar result for Merino sheep, where 50K genotypes were imputed from a 12K SNP panel using various reference groups. The researchers found that the 50K genotypes, which were imputed from a reference population of mixed crossbred Merino or non-Merino purebreds, gave lower prediction accuracy than the real 12K genotypes.

The prediction accuracy of GEBV also depends among other factors on the size of the training dataset and the strength of genomic relationships between all pair-wise combinations of individuals in the training and validation groups (Goddard 2009; Daetwyler et al., 2008). The greater the size of the training set and the higher the level of genomic relationship among individuals across the training and validation groups, the more likely the GEBV accuracy can be improved. The present study expressed the degree of relationship between pairs of individuals in the training and validation group as genomic distance between them, which eroded as more animals unrelated to the validation group were added to the training group. This created some confounding between increasing size of the reference population and increasing genetic distance. The genomic distance as calculated in the present study is synonymous with genetic distance which measures the degree of genetic divergence between species or between populations within a species (Nei, 1987). Populations with many similar genes have small genetic distances which indicate that they are closely related and have a recent common ancestor. The reduction in *r* as training-validation genomic relationship decays or genomic distance increases has been documented (Habier et al., 2010; Akanno et al., 2014b; Ventura et al., 2014), and was observed for most of GBLUP predictions in the present study. The GBLUP prediction accuracy for RFI, for example, reduced faster than those for DMI as genomic distance increased. This supports the views of Clark et al. (2011) that traits controlled by a large number of genes with small effects are more sensitive to variation in genetic relationship between training and validation groups than traits controlled by large QTL. On the contrary, BayesC predictions across the studied traits showed a small reduction in prediction accuracy as genomic distance increased and in some instance an improvement in *r* was observed (for e.g ANHH and TX validation groups). This could indicate that BayesC predictions are less sensitive and more robust to training-validation genomic distance than GBLUP predictions.

Chen et al. (2013) used a group of 522 AN and 395 CH, which is a subset of the animals used in the present study, to predict GEBV for RFI in the AN animals, using BayesB approach with the 50K genotypes, and found that within-breed predictions for AN had the highest realised accuracy of 0.53. This accuracy is comparable to our highest realised accuracy of 0.55 for AN being trained by a group of 1000 animals, using the 50K genotypes and BayesC method. When Chen et al. (2013) combined both AN and CH to predict RFI of the AN animals, using the same set of genotypes and BayesB method, they observed a realised accuracy of 0.53 for RFI prediction, which was the same as the realised accuracy for within AN prediction. In the present study, adding more animals to the initial training group made the realised accuracy drop slightly to 0.52 – 0.54, whereas using all 6644 animals to train the AN made the realised accuracy drop even further. Theoretically, an increase in number of training individuals should increase predictive ability (Hayes et al., 2009; Garrick 2011), especially where effective population size is large as in beef cattle. However, in this study, adding animals from various research populations to the reference coincided with adding animals that were less related, increasing the average genomic distance between animals in the training and validation groups. This could be a result particular to our dataset. The implication of this finding in beef cattle is that prediction accuracy does not depend only on having a large training population but also on including training individuals that are closely related to the validation or target population when 50K genotypes are used. This is not the case in dairy cattle where half-sib families are large and the phenotypes used are often sire proofs with high accuracy.

Additionally, the imputed HD genotypes showed a small improvement in prediction accuracy with increasing genomic distance/training size and demonstrated more usefulness in crossbreds (ANHH and TX) than in purebreds (AN and CH), while 50K genotypes showed greater prediction accuracy in purebreds (AN and CH) than in crossbreds (ANHH and TX), across the studied traits. This finding supports the expectation of Goddard and Hayes (2007) that high density markers and large training sets are required to improve prediction accuracy in crossbreds because high density markers will ensure that LD is consistent across multi-breed and crossbred populations. As noted earlier, in purebreds the LD between QTL and markers are likely to be conserved in larger distances so a lower marker density is sufficient to predict GEBVs with moderate accuracy. On the other hand, genomic prediction in crossbreds exploits inherent LD in parental breeds and new LD due to recent crosses, thus, higher density markers are required to exploit both sources of LD for GEBV prediction (Akanno et al. 2014b). Therefore, further investigation into the utility of higher SNP density for genomic prediction in crossbreds is warranted.

CONCLUSION

This study demonstrated the utility of the Illumina BovineSNP50 BeadChip and imputed Axiom Genome-Wide BOS 1 Array genotypes for genomic prediction of DMI, ADG and RFI in a beef cattle validation population that was created by considering the genomic distance between pairs of individuals in the training and validation groups. The results indicate that 50K genotypes, in conjunction with Bayesian methods was a more effective tool for predicting GEBV in purebreds. Imputed HD genotypes found utility when dealing with composite and crossbred populations. Moderate to high accuracy of genomic predictions were realised for DMI, ADG and RFI in purebred and crossbred beef cattle. In addition, formulation of a fairly large training population for estimating SNP effects in beef cattle should take into account the relationship between pairs of individuals in the training and target population.

LITERATURE CITED

Akanno, E. C., G. S. Plastow, C. Li, S. P. Miller, and J. A. Basarab. 2014a. Accuracy of molecular breeding values for production and efficiency traits of Canadian crossbred beef cattle using a cross-validation approach. In: Proc. 10th World Congr. Genet. Appl. Livest. Prod., Vancouver, Canada, p. 105.

Akanno, E. C., F. S. Schenkel, M. Sargolzaei, R. M. Friendship, and J. A. Robinson. 2014b. Persistency of accuracy of genomic breeding values for different simulated pig breeding programs in developing countries. J. Anim. Breed. Genet. Doi:10.1111/jbg.12085, 1-12.

Arthur, P. F., J. A. Archer, D. J. Johnson, R. M. Herd, E. C.Richardson, and P. F. Parnell. 2001. Genetic and phenotypic variance and covariance components for feed intake, feed efficiency, and other post-weaning traits in Angus cattle. J. Anim. Sci. 79:2805-2811

Basarab, J. A., M. G. Colazo, D. J. Ambrose, S. Novak, D. McCartney, V. S. Baron. 2011. Residual feed intake adjusted for backfat thickness and feeding frequency is independent of fertility in beef heifers. Can. J. Anim. Sci. 91:573-584.

Berry, D. P., and J. J. Crowley. 2012. Residual intake and body weight gain: A new measure of efficiency in growing cattle. J. Anim. Sci. 90:109-115.

Canadian Council on Animal Care (CCAC). 1993. Guide to the care and use of experimental animals. In: Olfert, E. D., B. M. Cross, A. A. McWilliams. (Eds.), Canadian Council on Animal Care, Vol. 1. Ottawa ON.

Chen, L., F. Schenkel, M. Vinsky, D. H. Crews (Jr), C. Li. 2013. Accuracy of predicting genomic breeding values for residual feed intake in Angus and Charolais beef cattle. J. Anim. Sci. Aug. 26, 2013.

Clark, S. A., J. M. Hickey, and J. H. van der Werf. 2011. Different models of genetic variation and their effect on genomic evaluation. Genetics, selection, evolution: GSE 43:18.

Crowley, J. J., P. Stothard, J. Basarab, S. P. Miller, C. Li, Z. Wang, G. Plastow, M. de Pauw, S. M. Moore, D. B. Lu. 2014. Collation of data and genetic parameter estimation in different experimental Canadian beef cattle populations measured for feed efficiency. In: Proc. 10th World Congr. Genet. Appl. Livest. Prod., Vancouver, Canada, p. 117.

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PloS one 3(10):e3395.

de Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of genomic predictions across multiple populations. Genetics. 183(4): 1545.

de Roos, A. P., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179(3):1503-1512.

Dunning, A. M., F. Durocher, C. S. Healey, M. D. Teare, S. E. McBride, F. Carlomagno, C. F. Xu, E. Dawson, S. Rhodes, S. Ueda, E. Lai, R. N. Luben, E. J. Van Rensburg, A. Mannermaa, V. Kataja, G. Rennart, I. Dunham, I. Purvis, D. Easton, and B. A. Ponder. 2000. The extent of linkage disequilibrium in four populations with distinct demographic histories. American journal of human genetics 67(6):1544-1554.

Fernando, R. L. and D. Garrick. 2013. Bayesian methods applied to GWAS. Methods Mol Biol 1019:237-274.

Fisher, R.A. The correlation between relatives on the supposition of mendelian inheritance. Trans R Soc Edin 1918, 52:399-433.

Garrick, D. J. 2011. The nature, scope and impact of genomic prediction in beef cattle in the United States. Genetics, selection, evolution: GSE 43:17.

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136(2):245-257.

Goddard, M. E., and B. J. Hayes. 2007. Genomic selection. J. Anim. Breed. Genet. 124(6):323-

330.

Goddard, M.E., B.J. Hayes, and T.H. Meuwissen. 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. J. Anim. Breed. Genet. 128(6):409-421. doi: 10.1111/j.1439-0388.2011.00964.x.

Gunia, M., R. Saintilan, E. Venot, C. Hoze, M. N. Fouilloux, and F. Phocas. 2014. Genomic prediction in French Charolais beef cattle using high-density single nucleotide polymorphism markers. Journal of animal science 92(8):3258-3269.

Habier, D., J. Tetens, F. R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics, selection, evolution: GSE 42:5.

Habier, D., R. L. Fernando, and D. J. Garrick. 2013. Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics 194(3):597-607.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4): 2389.

Habier, D., R.L. Fernando, K. Kizilkaya, and D.J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics. 12: 186.

Hayes, B. and M. E. Goddard. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genetics, selection, evolution: GSE 33(3):209-229.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: progress and challenges. Journal of dairy science 92(2):433-443.

Hayes, B. J., P. M. Visscher, H. C. McPartlan, and M. E. Goddard. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome research 13(4):635-643.

Hayes, B. J., J. Pryce, A. J. Chamberlain, P. J. Bowman, and M. E. Goddard. 2010. Genetic architecture of complex traits and accuracy of genomic prediction: Coat locour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS Genetics 6:E1001139. Doi:10.1371/journal.pgen.1001139.

Howie, B. N., P. Donnelly, and J. Marchini. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics 5(6): e1000529. doi:10.1371/journal.pgen.1000529.

Illumina, Inc. 2006. "Top/Bot" strand and "A/B" allele. Tech Note. Illumina Inc., 5200 Research Way, San Diego, CA, USA.

Khansefid, M., J. E. Pryce, S. Bolormaa, S. P. Miller, Z. Wang, C. Li, M. E. Goddard. 2014. Estimation of genomic breeding values for residual feed intake in a multibreed cattle population. J Anim Sci. 92(8):3270-83. doi: 10.2527/jas.2014-7375.

Laido, G., D. Marone, M. A. Russo, S. A. Colecchia, A. M. Mastrangelo, P. De Vita, and R. Papa. 2014. Linkage disequilibrium and genome-wide association mapping in tetraploid wheat (Triticum turgidum L.). PloS one 9(4):e95211.

Lee, J., M. Saatchi, H. Su, R.L. Fernando, and D.J. Garrick. 2014. "Genomic Prediction using Single or Multi-Breed Reference Populations in US Maine-Anjou Beef Cattle," Animal Industry Report: AS 660, ASL R2856. Available at: http://lib.dr.iastate.edu/ans_air/vol660/iss1/21

López-Campos, O., J.L. Aalhus, E.K. Okine, V.S. Baron, and J.A.Basarab. 2013. Effects of calf- and yearling-fed beef production systems and growth promotants on production and profitability. Can. J. Anim. Sci. 93:171-184.

Lu, D., M. Sargolzaei, M. Kelly, C. Li, G. Vander Voort, Z. Wang, G. Plastow, S. Moore, and S. P. Miller. 2012. Linkage disequilibrium in Angus, Charolais, and Crossbred beef cattle. Frontiers in genetics 3:152.

Lu, D., S. Miller, M. Sargolzaei, M. Kelly, G. Vander Voort, T. Caldwell, Z. Wang, G. Plastow, and S. Moore. 2013. Genome-wide association analyses for growth and feed efficiency traits in beef cattle. J. Anim. Sci. 91:3612-3633.

Meuwissen, T. H. E., B. J. Hayes, M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157:1819-1829.

Meuwissen, T. H., A. Karlsen, S. Lien, I. Olsaker, and M. E. Goddard. 2002. Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics 161(1):373-379.

Nei, M. 1987. Molecular Evolutionary Genetics. (Chapter 9). New York: Columbia University Press.

Nkrumah, D. J., E.K. Okine, G.W. Mathison, K. Schnid, C. Li, J.A. Basarab, M.A. Price, Z. Wang, and S.S. Moore. 2006. Relationships of feedlot feed efficiency, performance, and feeding behavior with metabolic rate, methane production, and energy partitioning in beef cattle. J. Anim. Sci. 84:145-153.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly, and P.C. Sham. 2007. PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81.

Saatchi, M., J. E. Beever, J. E. Decker, D. B. Faulkner, H. C. Freetly, S. L. Hansen, H. Yampara-Iquise, K. A. Johnson, S. D. Kachman, M. S. Kerley, J. Kim, D. D. Loy, E. Marques, H. L. Neibergs, E. J. Pollak, R. D. Schnabel, C. M. Seabury, D. W. Shike, W. M. Snelling, M. L. Spangler, R. L. Weaber, D. J. Garrick, and J. F. Taylor. 2014. QTLs associated with dry matter intake, metabolic mid-test weight, growth and feed efficiency have little overlap across 4 beef cattle studies. BMC genomics 15:1004.

Sargolzaei M., J. P. Chesnais, and F. S. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 2014, 15:478. doi:10.1186/1471-2164-15-478

Sargolzaei, M., F. S. Schenkel, G. B. Jansen, and L. R. Schaeffer. 2008. Extent of linkage disequilibrium in Holstein cattle in North America. Journal of dairy science 91(5):2106-2117.

Sargolzaei, M., F.S. Schenkel, P.M. VanRaden. 2009. gebv: Genomic breeding value estimator for livestock. In Technical report to the Dairy Cattle Breeding and Genetics Committee. University of Guelph; 2009.

Shrimpton, A. E. and A. Robertson. 1988. The Isolation of Polygenic Factors Controlling Bristle Score in Drosophila Melanogaster. II. Distribution of Third Chromosome Bristle Effects within Chromosome Sections. Genetics 118(3):445-459.

Su, G., R.F. Brøndum, P. Ma, B. Guldbrandtsen, G.P. Aamand, and M.S. Lund. 2012. Comparison of genomic predictions using medium-density (∼54,000) and high-density (∼777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. Journal of Dairy Science 95(8): 4657–4665.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: reliability of genomic predictions for North American Holstein bulls. Journal of dairy science 92(1):16-24.

Ventura, R.V., D. Lu, F.S. Schenkel, Z. Wang, C. Li, and S.P. Miller. 2014. Impact of reference population on accuracy of imputation from 6k to 50K SNP chips in multi-breed beef cattle populations. J. Anim. Sci. 92:1433–1444. doi:10.2527/jas.2013-6638.

Moghaddar, N., K.P. Gore, H.D. Daetwyler, B.J. Hayes, and J.H.J. van der Werf. 2015. Accuracy of genotype imputation based on random and selected reference sets in purebred and crossbred sheep populations and its effect on accuracy of genomic prediction. Genet Sel Evol (2015) 47:97. DOI 10.1186/s12711-015-0175-8.

Table 1. Least square means of performance and feed efficiency traits^{1} among different data sources^{2}

	Mean	AN, CH	HYB	PG1	ERS	SE^{3}
n^{4}	-	1599	907	3881	930	-
Start age (d)	299	312^{a}	301^{b}	297^{bc}	284^{c}	2.82
DMI (kg/d)	9.22	9.31^{a}	9.98^{b}	8.72^{c}	10.39^{d}	0.07
ADG (kg/d)	1.46	1.40^{a}	1.62^{b}	1.33^{c}	1.96^{d}	0.02
BW (kg)	430	430^{a}	454^{b}	420^{c}	457^{b}	3.32
BFAT (mm)	8.03	9.46^{a}	6.24^{b}	6.13^{b}	14.66^{c}	0.23
RFI (kg/d)	0	0.10	-0.02	-0.02	-0.06	0.05

This result is adopted from Crowley et al. (2014). ^{1}DMI = average dry matter intake, ADG = average daily gain, BW = mid-test bodyweight, BFAT = final ultrasound backfat, RFI = residual feed intake; ^{2}AN = Angus, CH = Charolais, HYB = beef-dairy hybrids, PG1 = Phenomic Gap Project, ERS = Elora Research Station; ^{3}Pooled standard error; ^{4}Total number of animals was 7317, of which 6794 were used in the present study.

Table 2. Average genomic distance (×10^{-3}) between a pair of individuals in the training and validation groups

Validation		Training group						
group^{1} (n=150)		n = 1000	n = 1999	n = 2999	n = 3999	n = 4999	n = 5998	n = 6644
AN	50K	0.63±0.53	2.27±1.80	3.77±2.63	5.50±3.81	7.36±5.08	9.45±6.63	11.60±9.45
	HD	0.52±0.91	1.58±1.39	2.36±1.66	3.26±2.18	4.20±2.75	5.25±3.51	6.25±4.64
CH	50K	1.94±1.99	4.51±2.97	6.28±3.53	7.94±4.21	9.59±5.04	11.50±6.39	13.00±7.57
	HD	1.33±1.42	2.29±1.85	3.81±2.17	4.69±2.46	5.52±2.79	6.51±3.44	4.19±3.86
ANHH	50K	2.85±0.65	3.71±1.02	4.40±1.29	4.96±1.49	5.57±1.81	6.58±2.93	7.69±4.42
	HD	1.67±1.09	2.20±1.24	2.57±1.34	2.86±1.41	3.14±1.50	3.65±1.95	4.19±2.59
TX	50K	1.01±0.40	2.15±1.26	3.15±1.77	4.02±1.77	4.86±2.16	6.13±2.58	7.12±3.72
	HD	0.57±0.34	1.33±1.06	1.89±1.30	2.35±1.45	2.74±1.57	3.39±2.12	3.85±2.50

[1]AN = Angus; CH = Charolais; ANHH = Angus-Hereford crosses; TX = Beefbooster composite. Increasing size of training groups coincided with including individuals less related with validation animals.

Table 3. Correlation between adjusted phenotype and GEBV for RFI, ADG and DMI using the 50K genotypes and two statistical methods[1]

Validation	Training	RFI		ADG		DMI	
group[2]	size	GBLUP	BayesC	GBLUP	BayesC	GBLUP	BayesC
AN	1000	0.31±0.02	0.35±0.05	0.24±0.10	0.27±0.11	0.44±0.04	0.44±0.03
	1999	0.27±0.03	0.33±0.04	0.18±0.09	0.24±0.12	0.41±0.07	0.44±0.03
	2999	0.29±0.06	0.34±0.03	0.19±0.08	0.23±0.12	0.39±0.06	0.43±0.04
	3999	0.27±0.07	0.32±0.05	0.20±0.08	0.24±0.13	0.38±0.08	0.41±0.05
	4999	0.25±0.08	0.31±0.05	0.21±0.08	0.25±0.13	0.38±0.09	0.41±0.06
	5998	0.24±0.07	0.31±0.05	0.20±0.09	0.25±0.13	0.37±0.07	0.41±0.05
	6644	0.23±0.07	0.31±0.05	0.20±0.08	0.24±0.13	0.36±0.07	0.40±0.06
	Mean	0.26	0.32	0.20	0.24	0.39	0.41
CH	1000	0.38±0.09	0.36±0.07	0.28±0.01	0.33±0.02	0.38±0.07	0.39±0.05
	1999	0.36±0.09	0.35±0.07	0.30±0.03	0.35±0.03	0.36±0.06	0.39±0.05
	2999	0.33±0.07	0.35±0.06	0.31±0.10	0.36±0.07	0.34±0.07	0.39±0.08
	3999	0.27±0.14	0.34±0.09	0.29±0.11	0.35±0.06	0.29±0.11	0.40±0.07
	4999	0.24±0.18	0.34±0.11	0.28±0.07	0.36±0.04	0.27±0.11	0.40±0.08
	5998	0.25±0.14	0.36±0.11	0.24±0.08	0.35±0.03	0.25±0.10	0.39±0.08
	6644	0.25±0.13	0.37±0.11	0.24±0.09	0.33±0.04	0.24±0.10	0.40±0.08
	Mean	0.29	0.35	0.27	0.34	0.30	0.39
ANHH	1000	0.20±0.12	0.21±0.13	0.15±0.06	0.20±0.09	0.23±0.14	0.27±0.10
	1999	0.15±0.11	0.22±0.10	0.15±0.10	0.20±0.09	0.18±0.11	0.26±0.08
	2999	0.14±0.10	0.21±0.10	0.14±0.10	0.21±0.09	0.21±0.11	0.27±0.10
	3999	0.15±0.10	0.20±0.10	0.14±0.09	0.21±0.08	0.23±0.09	0.31±0.10
	4999	0.15±0.10	0.21±0.10	0.14±0.08	0.23±0.08	0.23±0.09	0.32±0.09
	5998	0.14±0.09	0.20±0.09	0.14±0.07	0.24±0.07	0.25±0.06	0.32±0.08
	6644	0.12±0.09	0.19±0.09	0.14±0.06	0.23±0.08	0.23±0.06	0.31±0.08
	Mean	0.15	0.20	0.14	0.21	0.22	0.29
TX	1000	0.21±0.09	0.27±0.11	0.12±0.05	0.11±0.05	0.35±0.07	0.39±0.06
	1999	0.18±0.09	0.26±0.14	0.12±0.05	0.16±0.05	0.32±0.07	0.38±0.06
	2999	0.17±0.12	0.25±0.15	0.12±0.06	0.18±0.08	0.30±0.07	0.39±0.06
	3999	0.17±0.11	0.26±0.13	0.12±0.07	0.19±0.08	0.28±0.07	0.40±0.06
	4999	0.16±0.11	0.26±0.12	0.10±0.08	0.20±0.09	0.23±0.04	0.39±0.06
	5998	0.14±0.09	0.24±0.12	0.11±0.10	0.21±0.10	0.23±0.06	0.39±0.07
	6644	0.13±0.08	0.24±0.12	0.11±0.09	0.20±0.11	0.21±0.08	0.39±0.07
	Mean	0.16	0.25	0.11	0.17	0.27	0.38

[1]Within a given validation group, increasing training size represents increasing genomic distance between pairs of individuals in the training and validation groups

[2]AN = Angus; CH = Charolais; ANHH = Angus-Hereford crosses; TX = Beefbooster composite

Table 4. Correlation between adjusted phenotype and GEBV for RFI, ADG and DMI using the HD genotypes and two statistical methods[1]

Validation	Training	RFI		ADG		DMI	
group[2]	size	GBLUP	BayesC	GBLUP	BayesC	GBLUP	BayesC
AN	1000	0.10±0.06	0.08±0.05	0.07±0.10	0.01±0.09	0.15±0.07	0.10±0.01
	1999	0.09±0.07	0.09±0.03	0.07±0.11	0.02±0.07	0.15±0.06	0.11±0.06
	2999	0.08±0.07	0.09±0.04	0.07±0.12	0.04±0.10	0.13±0.08	0.12±0.06
	3999	0.08±0.08	0.09±0.04	0.08±0.12	0.05±0.10	0.14±0.07	0.14±0.05
	4999	0.10±0.07	0.11±0.05	0.08±0.12	0.05±0.10	0.16±0.07	0.16±0.05
	5998	0.10±0.08	0.11±0.05	0.08±0.11	0.04±0.09	0.16±0.07	0.17±0.04
	6644	0.10±0.07	0.12±0.05	0.09±0.11	0.04±0.08	0.17±0.06	0.18±0.04
	Mean	0.09	0.10	0.07	0.04	0.14	0.14
CH	1000	0.11±0.03	0.13±0.02	0.18±0.21	0.14±0.17	0.11±0.14	0.10±0.13
	1999	0.10±0.07	0.13±0.05	0.18±0.20	0.12±0.14	0.10±0.15	-0.01±0.16
	2999	0.09±0.05	0.14±0.05	0.17±0.18	0.13±0.13	0.08±0.13	0.08±0.12
	3999	0.08±0.10	0.14±0.04	0.16±0.14	0.14±0.11	0.08±0.08	0.08±0.10
	4999	0.11±0.10	0.15±0.05	0.15±0.15	0.13±0.12	0.11±0.08	0.09±0.10
	5998	0.10±0.06	0.15±0.04	0.13±0.15	0.12±0.13	0.12±0.11	0.11±0.11
	6644	0.12±0.06	0.17±0.07	0.13±0.15	0.13±0.12	0.13±0.11	0.12±0.12
	Mean	0.10	0.14	0.15	0.13	0.10	0.08
ANHH	1000	0.20±0.13	0.19±0.13	0.17±0.07	0.17±0.12	0.26±0.11	0.26±0.09
	1999	0.15±0.09	0.21±0.10	0.16±0.10	0.20±0.09	0.23±0.09	0.27±0.09
	2999	0.12±0.11	0.21±0.11	0.17±0.10	0.22±0.08	0.23±0.11	0.29±0.10
	3999	0.13±0.09	0.20±0.10	0.16±0.11	0.21±0.07	0.25±0.10	0.31±0.11
	4999	0.15±0.12	0.20±0.10	0.18±0.10	0.24±0.06	0.27±0.10	0.34±0.10
	5998	0.16±0.10	0.20±0.10	0.16±0.08	0.24±0.07	0.28±0.11	0.33±0.10
	6644	0.15±0.09	0.19±0.09	0.17±0.07	0.25±0.07	0.28±0.12	0.33±0.11
	Mean	0.15	0.20	0.16	0.22	0.25	0.31
TX	1000	0.23±0.09	0.25±0.11	0.15±0.05	0.10±0.04	0.38±0.09	0.38±0.07
	1999	0.24±0.10	0.25±0.13	0.17±0.05	0.19±0.05	0.37±0.07	0.39±0.06
	2999	0.22±0.10	0.24±0.13	0.16±0.08	0.22±0.07	0.35±0.08	0.38±0.05
	3999	0.21±0.12	0.24±0.13	0.16±0.09	0.23±0.07	0.33±0.07	0.38±0.04
	4999	0.22±0.10	0.24±0.12	0.17±0.09	0.24±0.08	0.32±0.06	0.38±0.05
	5998	0.20±0.09	0.23±0.12	0.16±0.09	0.25±0.08	0.32±0.04	0.38±0.04
	6644	0.19±0.07	0.23±0.11	0.17±0.09	0.25±0.09	0.32±0.05	0.38±0.05
	Mean	0.21	0.24	0.16	0.21	0.34	0.38

[1] Within a given validation group, increasing training size represents increasing genomic distance between pairs of individuals in the training and validation groups

[2] AN = Angus; CH = Charolais; ANHH = Angus-Hereford crosses; TX = Beefbooster composite

Table 5. Accuracy[1] of genomic estimated breeding values predicted with 50K panel for RFI, ADG and DMI using GBLUP and BayesC for Angus (AN) and Beefbooster composite (TX) validation groups. Regression coefficient of adjusted phenotype on predicted GEBV in brackets ()[2]

		Training group[3]						
Traits	Methods	n = 1000	n = 1999	n = 2999	n = 3999	n = 4999	n = 5998	n = 6644
	AN							
RFI	GBLUP	0.49 (0.47)	0.44 (0.38)	0.46 (0.38)	0.44 (0.34)	0.39 (0.30)	0.37 (0.28)	0.36 (0.26)
	BayesC	0.55 (1.17)	0.52 (1.49)	0.54 (1.53)	0.52 (1.50)	0.50 (1.45)	0.49 (1.27)	0.50 (1.24)
ADG	GBLUP	0.37 (0.36)	0.29 (0.23)	0.30 (0.22)	0.31 (0.22)	0.33 (0.23)	0.31 (0.20)	0.31 (0.20)
	BayesC	0.43 (1.90)	0.38 (1.26)	0.37 (0.98)	0.38 (0.93)	0.40 (0.86)	0.40 (0.85)	0.39 (0.86)
DMI	GBLUP	0.63 (0.67)	0.58 (0.50)	0.56 (0.46)	0.55 (0.44)	0.54 (0.42)	0.53 (0.40)	0.51 (0.38)
	BayesC	0.63 (1.06)	0.62 (1.22)	0.61 (1.18)	0.59 (1.11)	0.58 (1.05)	0.58 (1.05)	0.58 (1.06)
	TX							
RFI	GBLUP	0.33 (0.33)	0.29 (0.24)	0.27 (0.20)	0.27 (0.19)	0.26 (0.17)	0.21 (0.13)	0.20 (0.12)
	BayesC	0.37 (1.32)	0.38 (1.26)	0.35 (1.55)	0.33 (1.40)	0.35 (1.29)	0.31 (1.15)	0.31 (1.11)
ADG	GBLUP	0.19 (0.23)	0.18 (0.19)	0.18 (0.17)	0.19 (0.17)	0.16 (0.13)	0.17 (0.13)	0.17 (0.12)
	BayesC	0.23 (12.81)	0.26 (0.80)	0.26 (0.78)	0.25 (0.85)	0.26 (0.81)	0.26 (0.91)	0.27 (0.93)
DMI	GBLUP	0.49 (0.57)	0.45 (0.47)	0.43 (0.39)	0.39 (0.33)	0.32 (0.25)	0.32 (0.24)	0.30 (0.22)
	BayesC	0.54 (1.18)	0.53 (1.07)	0.50 (1.17)	0.47 (1.20)	0.45 (1.13)	0.46 (1.15)	0.45 (1.15)

[1]Accuracy is measured by correlation between adjusted phenotype and predicted genomic estimated breeding values in the validation group divided by the square root of estimated heritability. [2]A coefficient of 1 is expected. [3]Increasing training size represents increasing genomic distance between pairs of individuals in the training and validation groups

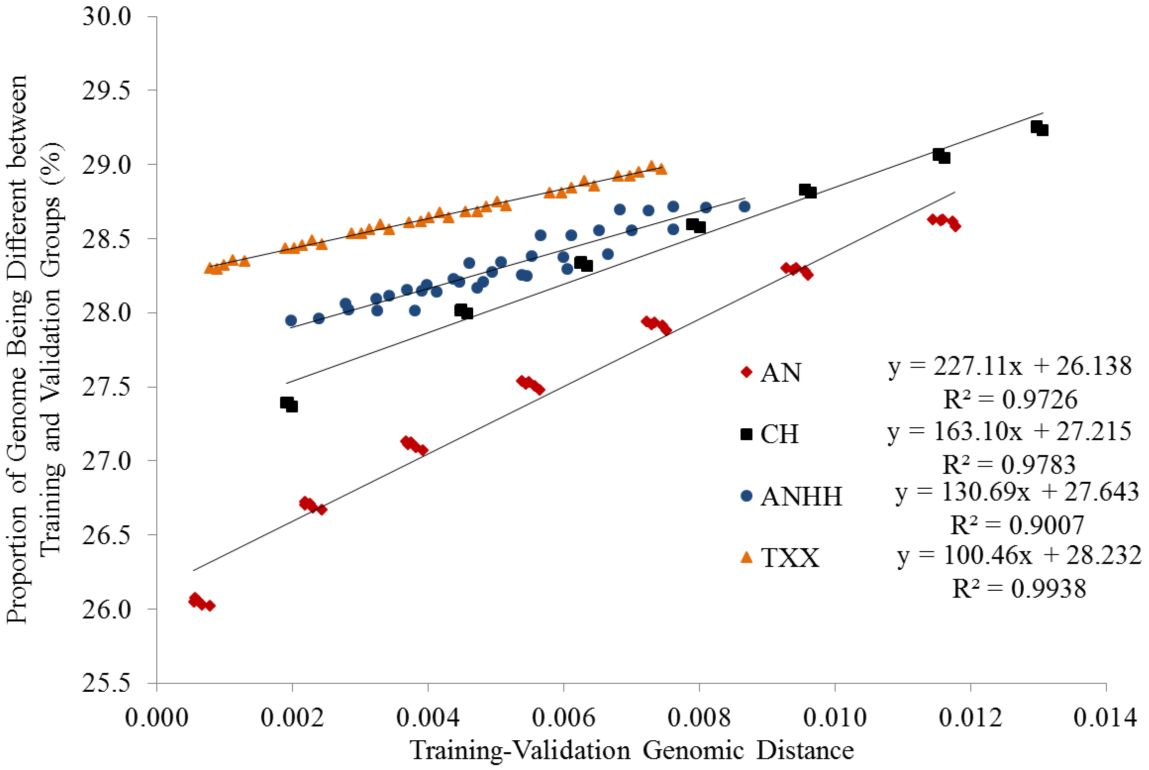
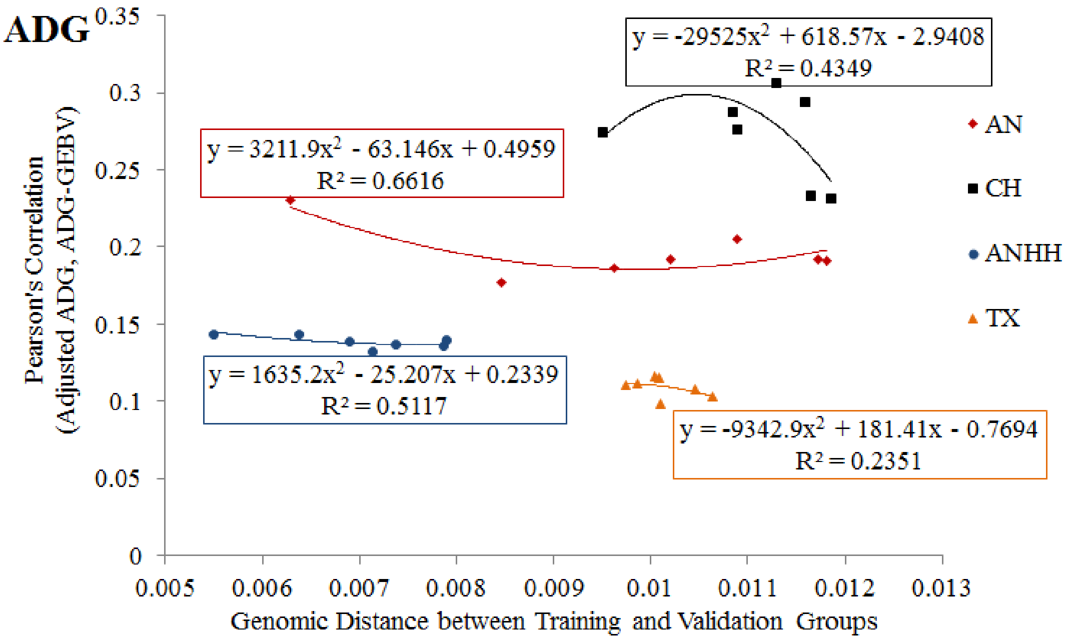
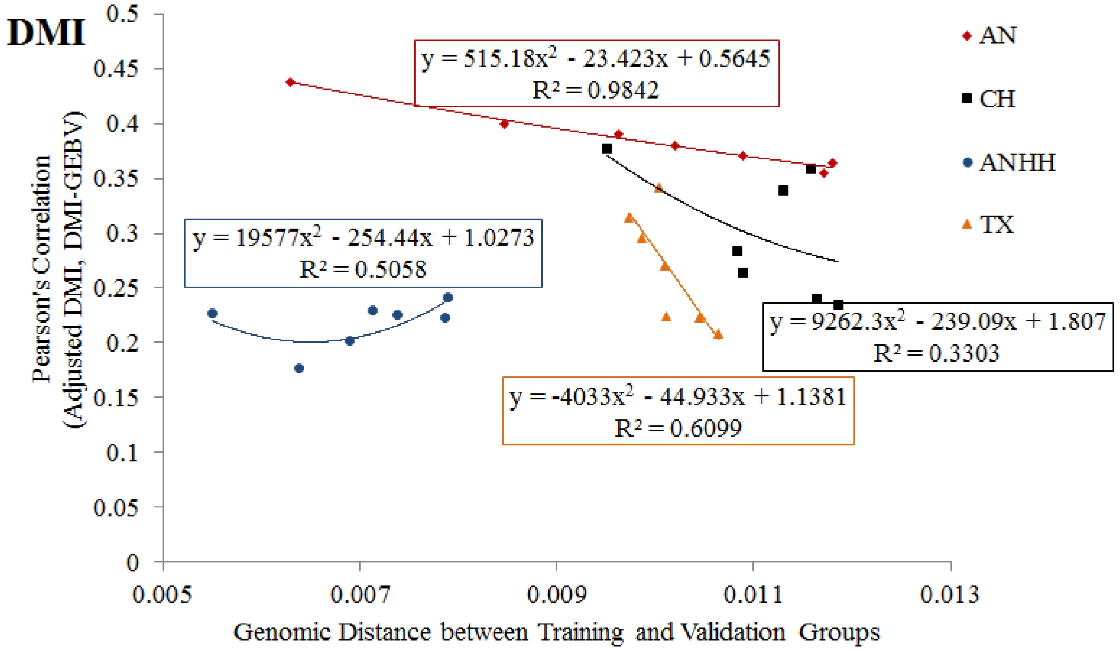


Figure 1. Genomic distance versus proportion of the genome being different between training and validation groups. AN = Angus; CH = Charolais; ANHH = Angus-Hereford cross; TX = Beefbooster composite


ADG
Pearson's Correlation (Adjusted ADG, ADG-GEBV)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.005
0.006
0.007
0.008
0.009
0.01
0.011
0.012
0.013
Genomic Distance between Training and Validation Groups
y = -29525x² + 618.57x - 2.9408
R² = 0.4349
y = 3211.9x² - 63.146x + 0.4959
R² = 0.6616
y = 1635.2x² - 25.207x + 0.2339
R² = 0.5117
y = -9342.9x² + 181.41x - 0.7694
R² = 0.2351
◆ AN
■ CH
● ANHH
▲ TX


DMI
Pearson's Correlation (Adjusted DMI, DMI-GEBV)
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.005
0.007
0.009
0.011
0.013
Genomic Distance between Training and Validation Groups
y = 515.18x² - 23.423x + 0.5645
R² = 0.9842
y = 19577x² - 254.44x + 1.0273
R² = 0.5058
y = 9262.3x² - 239.09x + 1.807
R² = 0.3303
y = -4033x² - 44.933x + 1.1381
R² = 0.6099
◆ AN
■ CH
● ANHH
▲ TX

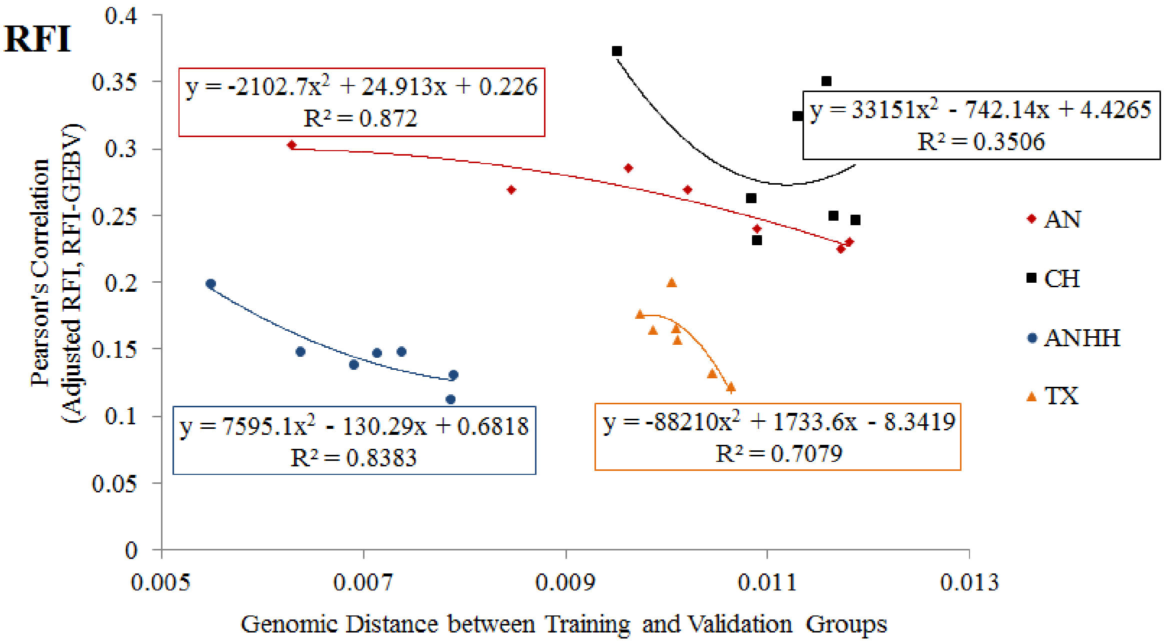

RFI
Pearson's Correlation (Adjusted RFI, RFI-GEBV)
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.005
0.007
0.009
0.011
0.013
Genomic Distance between Training and Validation Groups
y = -2102.7x² + 24.913x + 0.226
R² = 0.872
y = 33151x² - 742.14x + 4.4265
R² = 0.3506
y = 7595.1x² - 130.29x + 0.6818
R² = 0.8383
y = -88210x² + 1733.6x - 8.3419
R² = 0.7079
◆ AN
■ CH
● ANHH
▲ TX

Figure 2. Relationship between correlations (r) and the genomic distance between pairs of individuals in the training and validation groups. AN = Angus; CH = Charolais; ANHH = Angus-Hereford cross; TX = Beefbooster composite

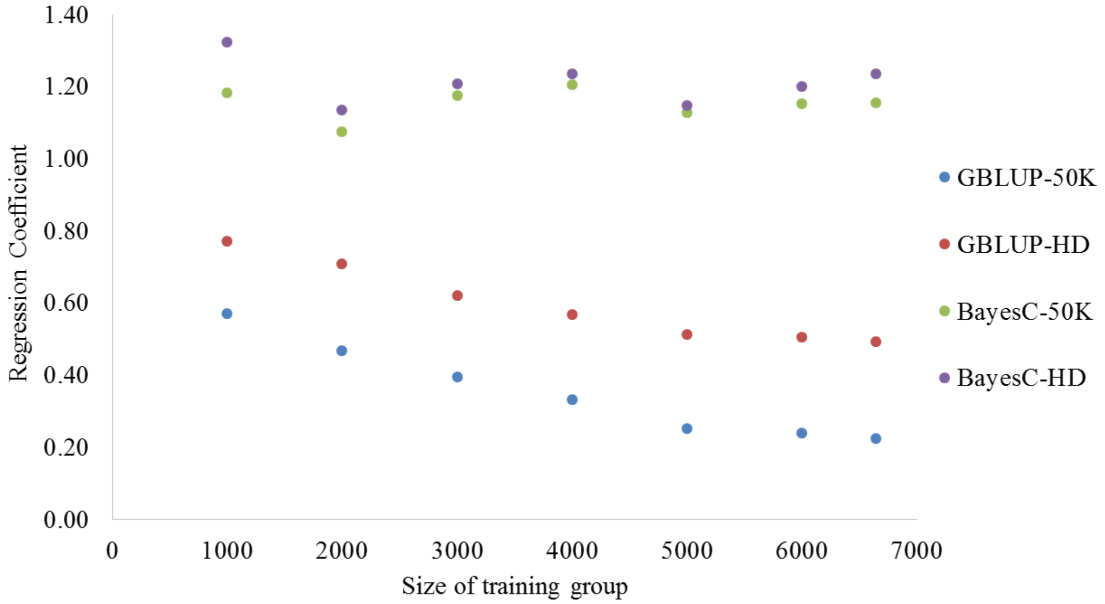


Figure 3. Regression coefficients of adjusted DMI on DMI-GEBV when using 50K and imputed HD genotypes in Beefbooster composite