Evaluation of GBLUP and Bayes-Alphabet Based on Different Marker Density For Genomic Prediction in Alpine Merino Sheep

Shaohua Zhu
Chinese Academy of Agricultural Sciences Lanzhou Institute of Husbandry and Pharmaceutical Sciences
Tingting Guo
Chinese Academy of Agricultural Sciences Lanzhou Institute of Husbandry and Pharmaceutical Sciences
Chao Yuan
Chinese Academy of Agricultural Sciences Lanzhou Institute of Husbandry and Pharmaceutical Sciences
Jianbin Liu
Chinese Academy of Agricultural Sciences Lanzhou Institute of Husbandry and Pharmaceutical Sciences
Jianye Li
Chinese Academy of Agricultural Sciences Lanzhou Institute of Husbandry and Pharmaceutical Sciences
Mei Han
Chinese Academy of Agricultural Sciences Lanzhou Institute of Husbandry and Pharmaceutical Sciences
Hongchang Zhao
Chinese Academy of Agricultural Sciences Lanzhou Institute of Husbandry and Pharmaceutical Sciences
Yi Wu
Chinese Academy of Agricultural Sciences Lanzhou Institute of Husbandry and Pharmaceutical Sciences
Weibo Sun
Chinese Academy of Agricultural Sciences Lanzhou Institute of Husbandry and Pharmaceutical Sciences
Xijun Wang
Gansu provincial Sheep Breeding Technology Extension Station
Tianxiang Wang
Gansu provincial Sheep Breeding Technology Extension Station

Jigang Liu
Gansu Provincial Sheep Breeding Technology Extension Station
Christian Keambou Tiambo
Centre for Tropical Livestock Genetics and Health, International Livestock Research Institute
Yaojing Yue
Chinese Academy of Agricultural Sciences Lanzhou Institute of Husbandry and Pharmaceutical Sciences
Bohui Yang (yangbh2004@163.com)
Chinese Academy of Agricultural Sciences Lanzhou Institute of Husbandry and Pharmaceutical Sciences

---

Methodology article

Keywords: Genomic prediction, Alpine Merino sheep, Wool traits, GBLUP, Bayes-Alphabet, Marker density

Posted Date: February 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-167352/v1

Evaluation of GBLUP and Bayes-Alphabet Based on Different Marker Density for Genomic Prediction in Alpine Merino Sheep

Shaohua Zhu[1,2], Tingting Guo[1,2], Chao Yuan[1,2], Jianbin Liu[1,2], Jianye Li[1,2], Mei Han[1,2], Hongchang Zhao[1,2], Yi Wu[1,2], Weibo Sun[1,2], Xijun Wang[3], Tianxiang Wang[3], Jigang Liu[3], Christian Keambou Tiambo[4], Yaojing Yue[2]* and Bohui Yang[1]*

[1].*Animal Science Department, Lanzhou Institute of Husbandry and Pharmaceutical Sciences, Chinese Academy of Agricultural Sciences, Lanzhou, 730050, China.*

[2].*Sheep Breeding Engineering Technology Center, Chinese Academy of Agricultural Sciences, Lanzhou, 730050, China.*

[3].*Gansu Provincial Sheep Breeding Technology Extension Station, Sunan, 734400, China.*

[4]. *Centre for Tropical Livestock Genetics and Health (CTLGH), International Livestock Research Institute, Nairobi, 00100, Kenya.*

*** Correspondence:**

Yaojing Yue: Tel: +86-0931-2115273; Fax: +86-0931-2115272; E-mail: yueyaojing@126.com

Bohui Yang: Tel: +86-0931-2115272; Fax: +86-0931-2115272; E-mail: yangbh2004@163.com

Email addresses:

Shaohua Zhu: zhu87932890@126.com

Tingting Guo: guotingting@caas.cn

Chao Yuan: yuanchao@caas.cn

Jianbin Liu: liujianbin@caas.cn

Jianye Li: lijianye218@163.com

Mei Han: 82101186189@caas.cn

Hongchang Zhao: 18837101296@163.com

Yi Wu: 1593050417@qq.com

Weibo Sun: swb887246@126.com

Xijun Wang: 1091606575@qq.com

Tianxiang Wang: zhangyesyz@163.com

Jigang Liu: hcycwtx@sina.com

Christian Keambou Tiambo: C.tiambo@cgiar.org

Abstract

Background

The marker density, the heritability level of trait and the statistical models adopted are critical to the accuracy of genomic prediction (GP) or genomic selection (GS). The studies on the impact of the above factors on accuracy of GP are usually focused on the comparison and discussion of simulated datasets. If the potential of GS is to be fully utilized to optimize the effect of breeding and selection, it is essential to incorporate these factors into real data for understanding their impact on GP accuracy, more clearly and intuitively. Herein, we studied the genomic prediction of six wool traits of sheep by two different models, including genomic best linear unbiased prediction (GBLUP), and Bayes-Alphabet. We adopted 5-fold cross-validation to perform the accuracy evaluation based on the genotyping data of Alpine Merino sheep (n=821).

Results

The GP accuracy of the six traits was found to be between 0.28 and 0.60, as demonstrated by the cross-validation results. We showed that the accuracy of GP could be improved by increasing the marker density, which is closely related to the model adopted and the heritability level of the trait. Moreover, based on two different marker densities, it was derived that the prediction effect of GBLUP model for traits with low heritability was better (GBLUP has the highest accuracy of 28.57% higher than Bayes-Alphabet); while with the increase of heritability level, the advantage of Bayes-Alphabet would be more obvious, therefore, different models of GP are appropriate in different traits.

Conclusion

This is the first study of optimization of GP has been applied to the domesticated Alpine Merino sheep populations. The main aim was to study the influence and interaction of different models and marker densities on GP accuracy. These findings indicated the significance of applying appropriate

models for GP which would assist in further exploring the optimization of GP.

Keywords: Genomic prediction; Alpine Merino sheep; Wool traits; GBLUP; Bayes-Alphabet; Marker density

Background

The advancement in the field of quantitative genetics and molecular biology has improved the selection and breeding methods of domestic animals [1]. Meuwissen et al. 2001 proposed a more advantageous selection method, known as genomic selection (GS) or genomic prediction (GP) [2]. This method combines the genome-wide single nucleotide polymorphism (SNP) with phenotypic data and implicates them for genetic evaluation [3-5]. It was first applied to the dairy cows [6] and is now widely used in other model animals such as beef cattle [7], pigs [8], goats [9], and sheep [10], aquatic animals like Atlantic salmon[11], rainbow trout[12], and plants [13, 14], such as wheat [15] and alfalfa [16]. GS has made a substantial contribution to the modern breeding process, as compared to traditional methods; the main advantages of this method include improved estimation accuracy of breeding value (BVs) [17, 18], increased genetic progress, and reduced breeding costs [19, 20]. With the successive publish of various livestock genome sequences and the continuous upgrade of commercial SNP microarrays, different types and densities of microarrays have been adopted in the GP of different livestock [21]. Accuracy and cost are generally the most critical factors in GP, compared to low-density SNP microarrays, the high-density SNP microarrays could accommodate more SNP sites that may lead to higher coverage of the genotype data [22]. However, the cost of the high-density microarray was comparatively higher. In contrast, although the low-density SNP microarrays has fewer SNP sites, it is more applicable in population breeding with a huge dataset due to its lower cost. Both the methods have their own pros and cons and therefore, it is difficult to conclude which density microarray is best suitable for GP.

For the first time, Meuwissen et al. 2001 proposed a GS based on Bayes method, which includes BayesA and BayesB [2]. Based upon this approach, several other methods were also derived such as BayesCπ method [23], Bayesian least absolute shrinkage and selection operator (Bayesian LASSO) method [24]. Subsequently, in 2013, Gianola summarized these methods as the Bayes-Alphabet method [25]. In fact, the assumptions and strategies adopted by these methods are different. The BayesA assumes that all SNPs have genetic effects and the variance of marker effects should obey the t-distribution, whereas BayesB assumes that only a small proportion of SNPs have an effect. Furthermore, the BayesCπ is similar to BayesB, and estimates the proportion of sites with no effect of π in the model. The Bayesian LASSO method assumes that all markers have effects, and the variance of marker effects obeys the double exponential distribution also known as Laplace distribution [25]. VanRaden et al. 2008, proposed another calculation method for GP and named it as genomic best linear unbiased prediction (GBLUP). It calculates the relationship matrix of individuals via genome-wide genotype information instead of traditional pedigree information. Herein, the matrix denoted as G is applied to replace the A matrix in BLUP, to estimate the BVs according to the BLUP method [26]. Another novel approach known as single-step GBLUP (SSGBLUP or HBLUP) has been developed based on GBLUP [27]. This method integrates the phenotype, pedigree and genomic information into a model, and combines the traditional kinship matrix A with the genome relationship matrix G according to different weights to construct a new relationship matrix H, then simultaneously estimate the genetic effects of all individuals (including individuals with and without genotypes). Although there are various GP methods available, no method could be suitable for all traits. Therefore, in this study, two methods based on Bayes and GBLUP models were adopted to study the prediction accuracy of real data for different wool traits,

aiming to screen ideal GP models.

As an important domestic animal, sheep is one of the earliest domestic animals reared by humans [28] and provides diverse resources such as mutton, wool, skin, and milk. Merino and Merino-derived sheep breeds are distributed globally [29]. As the object of the current study, the Alpine Merino sheep has Australian Merino and Tibetan sheep lineage. Thanks to their adaptation in high-altitude hypoxia and excellent wool quality, they quickly adapted to the freezing Qinghai-Tibet Plateau, living in high altitude and cold conditions for generations [30]. The length and strength of the staple and fiber diameter are closely related to the wool quality and are the important economic traits of fine-wool sheep. Therefore, adopting genome analysis to explore wool traits is crucial for the selection and development of this population. However, the application of GP for this population has just started obtain their genomic information through SNP microarray, and combined with phenotypic datasets closely related to wool traits, then, adopted different methods of BVs estimation and compare the results, in particular, genome analysis could be performed from two aspects include genetic effects of markers and methods of GP. This could make a great contribution to the application of GP and GS in Alpine Merino sheep population.

In the current study, two different densities of SNPs including low (50K) and high (630K) were applied to estimate the genetic variance components of the Alpine Merino sheep datasets. Further, based upon the SNP genotypes data, different models were adopted for GP and cross-validated to compare the accuracy of different GP methods. The main purpose of this study is to investigate the impact of different densities of SNP genotypes and different GP methods on the accuracy and optimization methods of GP in Alpine Merino sheep populations.

Results

Statistics and processing of phenotypic data

A total of 6 wool traits were collected and the descriptive statistics of individual wool phenotype data was presented in Table 1, including the abbreviation of each trait, the corresponding standard error (S.E), the average value (represented by mean ± S.D), and the number of individuals that were effectively recorded (Numbers). For the wool traits, the standard deviation (S.D) ranged from 2.11 (FD) to 13.16 (SL), and the standard error (SE) ranged from 0.07 (FD) to 0.46 (SL).

The polygenic heritability and the GP accuracy

The phenotypic variance and the additive variance of the 6 wool traits based on L- and H-datasets were estimated to calculate polygenic heritability (h^2). For L-datasets, heritability ranged from 0.37 (FER) to 0.70 (SL); and for H-datasets, heritability ranged from 0.29 (FER) to 0.68 (SL). The estimated results of heritability (expressed as the proportion of additive variance in phenotypic variance) shown in Table 3, states that SL was the highest and the FER was the lowest irrespective of the L- or H-datasets. Moreover, the heritability estimated by L-datasets was slightly higher than that of H-datasets for these 6 wool traits.

The GP accuracy was calculated using 5 methods based on two marker density datasets (Table 4). For L-datasets, the GP accuracy of SL was the highest (0.59 for Bayesion LASSO model); and the GP accuracy of FER was the lowest (0.28 for BayesA model). Correspondingly, for H-datasets, the trait with the highest GP accuracy was also SL (0.58 for BayesA, BayesB, and Bayesion LASSO model), and the trait with the lowest GP accuracy was FER (0.31 for BayesA model).

Discussion

Genomic information and individual relationship matrix

The analyses involved in this study are all based on genomic information obtained from genotyping

142 through microarrays, GP has replaced the traditional phenotype and pedigree information with the
143 dense markers, providing a new method to estimate genetic variance, which improves the accuracy
144 of prediction and selection [31]. Genomic information is not only suitable for a population with
145 pedigree information, but can also be applied to populations without pedigree information or
146 incorrect, incomplete and even missing genealogical records [32, 33]. In the GBLUP model, the
147 traditional individual relationship matrix A constructed by pedigree was replaced by the genome
148 matrix G, which represents the relationship between individuals more accurately, as it is based on
149 a dense genome-wide marker. More importantly, this may capture the genetic connections from
150 unknown common ancestors, because it represents confirmed gene sharing, and has advantages over
151 presumed or conceptualized ancestral sharing [4]. In GBLUP model, it was assumed that each SNP
152 has an effect, and the cumulative effect of SNPs obey a normal distribution [34], the assumption
153 might only be applicable to certain specific groups or traits. According to the hypothesis of Habier
154 et al., for some traits, only a few markers have a larger effect, while most markers have little or no
155 effect [23, 35]. Therefore, GBLUP may not be suitable for such trait, in other words, the GP accuracy
156 of GBLUP will be lower than other models, like the FD trait in current study, the GP accuracy (0.56
157 based on L-datasets) of the Bayesion LASSO model was higher than that (0.52 based on L-datasets)
158 of the GBLUP model. From the above results, GBLUP may not be applicable to FD traits and its
159 predictive ability may not achieve satisfactory results. Hence, it is necessary to adopt different GP
160 models. In the Bayes-Alphabet method, models such as BayesB and BayesCπ assume that most of
161 the SNPs in the genome are located in regions without quantitative trait locus (QTL) and have no
162 effect [24]. while a small number of other SNPs existed in linkage disequilibrium (LD) together
163 with QTL, and accounts for most of the effect [34, 36]. According to reports, different Bayes-

Alphabet methods put forward a variety of prior hypotheses on the distribution of SNP effects (Table 2) [34]. In the current study, in addition to the GBLUP method, 4 typical Bayes-Alphabet methods (BayesA, BayesB, BayesCπ and Bayesion LASSO) were also used to compare the GP accuracy of the 6 wool traits.

In most cases, GP suffers limitations while adopting the high-density or low-density SNP genomic information, i.e., the number of marker effects that need to be estimated is often greater than the number of individuals to be recorded. In this study, both the L-and the H-datasets showed that the number (35,379 and 460,656) of markers was much larger than the number (821) of individuals. Although many advanced statistical methods [37, 38] have been proposed to overcome this challenge, the true distribution of QTL and SNP effects were unclear for many quantitative traits [34]. Moreover, in contrast to L-datasets, the H-datasets microarrays contain more genomic information, but it also involves more complex matrices and larger computation, which will undoubtedly increase the cost of time and economy [36].

Phenotypic statistics and estimation of heritability

In the current study, the collected phenotypic statistics of wool traits were compared with the results in previous reports: Moghaddar et al. collected 3000-8000 phenotypic records of various wool traits from different breeds of sheep in 2014, including the Poll Dorset, White Suffolk and Border Leicester. In their report, the statistical mean values of FD and FD_CV were 19.93±5.39 and 19.26±2.86 (mean ± S.D) respectively. The statistical mean of SS and SL was 33.82±9.82, 80.93±13.06, respectively [39]. In addition, according to the study by Hamadani (2019) et al. on Rambouillet sheep [40], where they collected and recorded the wool traits of 4,108 samples from 1998 to 2007, the statistical mean value of FD and SL was 21.26±0.03 (mean ± S.E), 56.1±0.05

186 respectively. The above comparison showed that the phenotypic statistics of the current study were
187 consistent with the earlier studies. It could be suggested that although the number of phenotypes
188 collected in this study was not as large as, the statistical values of phenotype measurement were still
189 reliable.
190 The additive and residual variance, and the heritability of the 6 wool traits of the Alpine Merino
191 sheep population were estimated. Daetwyler (2010) and Moghaddar (2014) et al. conducted the
192 genetic parameter estimation and GP studies on the sheep of multiple breeds including Merino,
193 Border Leicester, and White Suffolk. The results showed that the weighted average heritability of
194 SS and SL was in the range from 0.37 to 0.55 and 0.56 to 0.67, respectively. The weighted average
195 heritability of FD and FD_CV was between 0.62-0.75 and 0.47-0.57, respectively [39, 41]. In
196 addition, Safari (2005) and Fogarty (1995) et al. collected and summarized the genetic parameters
197 of 9 wool traits [42, 43]. Their results showed that the weighted average heritability of SS, SL,
198 CFWR, FD, FD_CV were 0.34, 0.46-0.48, 0.34-0.51 0.51-0.59 and 0.52, respectively. In the current
199 study, except for the slightly lower estimated value of FD (0.42-0.47), the other four wool traits
200 (Table 3) were close to the results reported in the previous literature. Especially, the SS (0.33-0.46)
201 was very close to them. The comparison with the previous literature suggested that the heritability
202 results estimated from the Alpine Merino dataset in the current study were reliable.
203 **GP results and accuracy of prediction**
204 If breeding scientists are to effectively apply genomic selection in their breeding programs, they
205 need to have a full understanding of the factors that affect the accuracy of the dataset predictions.
206 For effective application of GS and GP on sheep breeding programs, there should be a thorough
207 understanding of the factors affecting the accuracy of the dataset predictions. We collected 821

samples from the breeding program to investigate the influence and interaction of marker density and GP on the accuracy of prediction. Previous studies suggested that the density of markers has an essential impact on the accuracy of GP [44, 45]. Solberg and his collaborators (2008) adopted the simulated data to analyze the correlation between accuracy and marker density, their results showed that increasing the density of SNPs from 1 to 8 per centimorgan (cM) could improve the accuracy of GP by 25 %[46]. but this did not mean that the accuracy could always improve with the increase of marker density, in other words, there is a limit to this improvement. Heffner et al. (2011) conducted a study using a wheat dataset and showed that with the increased density from 192 to 1,158 markers, the accuracy of GP could be improved by 10 %. However, when the marker density increased from 192 to 384,it caused only a small increase in accuracy [47]. Most of the 10 % improvement mentioned above occurred in the interval from 192 to 384 markers, and the increase of the remaining markers did not significantly affect the accuracy. These results indicate that marker density has a positive effect on the accuracy of GP, while the response of accuracy to density will eventually stabilize [48].

Herein, we adopted the genome datasets based on the level of 50K and 630K microarray, respectively. Table 1 shows that with the marker density increases, the improved accuracy of GP for most traits, especially in FBS and FER, model Bayesion LASSO and BayesA increased by 12 and 11 %, respectively, while in other traits the accuracy was not significantly improved, such as CFWR and FD_CV, the accuracy of GBLUP and BayesB increased only by 1 %; FBS and FER benefited more from the increase in marker density than other traits, which could be explained by the fact that quantitative genetic characteristics require more markers to accurately estimate their many small effects of QTL[49]. Interestingly, there are exceptions in this study, for some traits, the accuracy

may even decrease: in FD trait, the accuracy of BayesA and Bayesion LASSO models were reduced by 3 % and 5 %, respectively. Two reasons that may explain why increasing number of markers on each chromosome led to a decrease in GP accuracy. Firstly, the number of markers in the microarray is much larger than the number of samples, which may be due to excessively high density of markers leading to the model overfitting [50]. Secondly, the increases in the number of markers will lead to the addition of more unknown variables (marker effects) and a lack of accurate estimation. The study from Fatemeh Alanoshahr et al. also showed that with the number of SNPs increased from 2000 to 3000, both BayesA and GBLUP model indicated a decrease in the accuracy of GP [51]. Our results suggest that increasing the density of markers could indeed improve the GP accuracy, but it is closely related to the trait itself. For traits with low heritability levels (FER and FBS), a small part of the phenotypic variation was explained by additive effects[52], and the increase of marker density may improve the accuracy of GP more obviously; correspondingly, for those traits with high heritability levels (CFWR and FD), increasing the marker density has little benefit on the GP accuracy, sometimes even has a negative impact on accuracy.

Among the 6 wool traits studied here, SL and FD_CV had the highest heritability (h^2=0.53 and h^2=0.58, respectively), and their corresponding accuracy of GP was also the highest, which ranged from 0.53 to 0.60 and 0.45 to 0.55, respectively. While for two traits with the lowest heritability, FBS (h^2=0.33) and FER (h^2=0.28), the accuracy was 0.29 to 0.38 and 0.28 to 0.36, respectively, which was lower than SL and FD_CV. For those traits with lower heritability, the correlation between phenotypic value and genetic value will be lower, the effect value of markers distributed across the genome may be estimated with lower accuracy [23], it suggested that higher heritability has a positive effect on the accuracy of GP. Bolormaa et al. (2013) also reported that the prediction

of the trait with the highest heritability was more accurate [53], and also several studies have shown that the accuracy of GP increases with the improved heritability[54, 55], the results of the current study agreed with them. In addition, we found that for traits with low heritability, GBLUP had a better prediction effect, whether it is adopting L- or H-datasets, but with the increase of heritability, the advantage of GBLUP is not obvious. From Table 4, it could be observed that for the trait SL with high heritability, the estimation accuracy of BayesB (0.58-0.60) and Bayesion LASSO (0.58-0.59) models performed better, this may indicate that for some traits with high heritability, BayesB and Bayesion LASSO assumes more reasonable distribution in marker effect, which leads to higher prediction accuracy. Similar results were obtained in the study of Honarvar and his coworkers, based on the simulation data of three different levels of heritability, they compared the accuracy of the RRBLUP and bayesion-LASSO models, and the results showed that the GP accuracy of the bayesion-LASSO model is higher than that of the RRBLUP model for these traits, but the former has a more obvious advantage in traits with high heritability [56], and it should be noted that GBLUP was equivalent to RRBLUP. In addition, the accuracy of GP was also related to the size and structure of the reference group [57, 58]. We will collect and organize a larger dataset in future and try to take the above factors into consideration in subsequent studies for better conclusive results.

Conclusions

To summarize, this study was based on two different densities of microarray genotyping data, using Bayes-Alphabet (including BayesA, BayesB, BayesCπ, Bayesion LASSO) and GBLUP model to perform the GP. The heritability of 6 wool traits of Alpine Merino sheep was estimated, and the accuracy of the BVs prediction of these traits under different conditions was evaluated through five-fold cross-validation. To the best of our knowledge, this was the first study of optimization of GP

has been applied to the domesticated Alpine Merino sheep populations. We have observed that for traits with low heritability (SS and FER), increasing the density of markers could improves the GP accuracy, but it has little impact on traits with high heritability (SL), and even decreases the accuracy (FD). The accuracy of the GBLUP model is generally higher than that of the Bayes-Alphabet model for SS and FER, while with the improvement of heritability, the advantage of GBLUP is no longer obvious. Therefore, from this study, we conclude that GBLUP is more suitable for traits with lower heritability (FER and FBS), and Bayes-alphabet, especially BayesB and Bayesion LASSO, have better GP effects for traits with high heritability (FD and SL), different GP models are applicable to different traits.

Methods

Animal resources and phenotypic data

The original phenotypic dataset was obtained from the Sheep Breeding Technology Extension Station of Gansu Province. These datasets consisted of 11,500 individuals based on 7 different herds with information such as region (herd), gender, and date of birth. The individuals in the current study included 821 Alpine Merino sheep (563 ewes and 258 rams) from HuangCheng pasture in Gansu Province, China, all born between year, 2014 to 2018. This pasture was under the jurisdiction of the Gansu Sheep Breeding Technology Extension Station which has a rigorously standardized system of breeding and management, to ensure that all the individuals have unified feeding and management conditions. The average age of each individual with phenotypic data was about 14 months. The wool traits involved in the current study were staple length (SL), clean fleece weight rate (CFWR), average fiber diameter (FD), coefficient of variation of average fiber diameter (FD_CV), staple strength (SS) and fleece extension rate (FER). The wool from individuals was

collected and evaluated according to the Agricultural Industry Standards of the People's Republic of China (NO. NY/T 1236-2006). Wool samples (~150-200 grams) collected from the abdomen of each individual, were weighed and stored in ziplock bags (Xingdeli Packaging Material Company Ltd., Shenzhen, China). Within one week, the samples were sent to the National Animal and Rural Ministry of Animal and Fur Quality Supervision and Inspection Center (Lanzhou, China) for weighing, screening and quality identification of wool. Blood samples (~5 mL) were also collected from each sheep from the jugular vein and immediately transferred to the vacutainer blood collection tube (Yuli Medical Equipment Company Ltd., Jiangsu Province, China). Blood samples were stored at -20°C for further genotyping [59]. The statistics used to estimate variance components and GP of each wool trait are presented in Table 1.

Genotypic data and quality control

The customized Affymetrix HD 630K microarray was employed as the datasets for the genotype of high-density SNP genotypes (H-datasets) for the Alpine Merino sheep. The genotyping platform for analysis was based on the array plate processing workflow of GeneTitan system (Santa Clara, California, USA) from Thermo Fisher (Affymetrix). The sites in the Illumina Ovine SNP 50K microarray were screened out from the Affymetrix HD 630K microarray and used as the datasets of low-density SNP genotypes (L-datasets). The H-and L-datasets were pre-processed using PLINK v1.9b4 software prior to the statistical analysis and variance component estimation [60]. The SNPs were eliminated with call rate (geno) below 95 %, minor allele frequency (MAF) below 0.01, which seriously deviated from the Hardy Weinberg Equilibrium with a P-value below 10^{-6}. Here, the X, Y chromosomes and mitochondrial chromosomes were excluded from the analysis. In addition, Beagle software (version number; 12Jul19.0df) was used to impute the missing SNPs [61]. After quality

control and impute, a total of 821 individuals with 460,656 autosomal SNPs were retained for H-datasets, and a total of 821 individuals with 35,379 autosomal SNPs for L-datasets.

Statistical methods for GP

We explored the application of SNP datasets of different densities in genome evaluation and further compared the accuracy of GP adopting 5 different models, including Bayes-Alphabet (BayesA, BayesB, BayesCπ, Bayesian LASSO) and GBLUP. Six wool traits from 821 samples were used to first, estimate the variance of each component, including the additive and residual variance; second, five different models were adopted to perform GP, and its accuracy was compared via 5-fold cross-validation, and all these models were evaluated in SNP datasets of H-and L-datasets. Replicate measurements were not available for the individuals so that the effects of permanent environmental were not modeled. The samples involved were from different herds and genders. These factors altered the phenotype in a fixed pattern, and hence the system environmental effects were added to the framework.

The statistical methods of Bayes-Alphabet involved can be written as:

$$\boldsymbol{y} = \boldsymbol{X}b + \sum_j^n \boldsymbol{Z}_{ij}\alpha_j + e \qquad (1)$$

Here, $\boldsymbol{y}$ represents the corrected phenotypic value of individuals, $\boldsymbol{X}b$ refers to a fixed term, and b contains a vector of 3 effects, including herds, genders, and mean of population. $\boldsymbol{Z}_{ij}$ represents the genotype of individual i at site j, and α_j represents the effect value of site j, and therefore $\sum_j^n Z_{ij}\alpha_j$ refers to the BV corresponding to individual i, e to the vector of residual effects. According to the method from Meuwissen et al. and Habier et al [2, 23], adopted the R package "BGLR" (https://github.com/gdlc/BGLR-R) to estimate the effect of markers [62]. The hypothetical distribution of all markers' effects in different Bayes methods and the formula of effect distribution

are shown in Table 2.

The methods of GBLUP involved in the current study confirms to a linear model.

$$\boldsymbol{y} = \boldsymbol{X}b + \boldsymbol{Z}u + e \qquad (2)$$

In Bayes-Alphabet model, in equation 2, $\boldsymbol{y}$, b, e and $\boldsymbol{X}$ represent the same parameters as those defined in equation 1, u is the vector of individuals breeding value, $\boldsymbol{Z}$ is the design matrix corresponding to the breeding value. The covariance matrix of additive effects is represented by $Var(u) = G\sigma_a^2$, where G is the matrix of relationships between individuals obtained from genomic information, calculated according to the approach of VanRaden [26] (equation 3) and also implemented through the R package "BGLR" (https://github.com/gdlc/BGLR-R) [62].

$$\boldsymbol{G} = \frac{\boldsymbol{W_a W_a^T}}{2\sum_{f=1}^{m} p_f(1-p_f)} \qquad (3)$$

where $\boldsymbol{W_a}$ represented the matrix of additive genetic effect markers, with dimension of the number of individuals (n) by the number of loci (m), and p_f is the minor allele frequency (MAF) value of locus f.

Accuracy of GP by K-fold cross-validation

Five-fold cross-validation was performed to compare the accuracy of different methods of GP. During K-fold cross validation, the population should be divided randomly [34]. The datasets consisting of 821 individuals were divided into five approximately equally-sized subgroups (each subgroup contained around 165 individuals). For 5-fold cross-validation, four subgroups which retain the phenotype and genotype, were regarded as training population (reference population) to estimate the parameters. The remaining subgroup i.e., candidate population was used to verify the samples, and correspondingly, the phenotype of this group of samples was set as missing (Not applicable, NA).

According to the above mentioned five models, the cross-validation was performed based on two types of genotypic data (H- and L-datasets), with different densities and the BVs of the validation group (candidate population) were predicted. In addition, the above cross-validation was performed in triplicates in order to ensure the randomness of individuals in the validation group. Finally, the GP accuracy values were calculated for each validation, averaged and then recorded as the final accuracy.

Abbreviations

GP: genomic prediction; GS: genomic selection; SNP: Single nucleotide polymorphism; GBLUP: genomic best linear unbiased prediction; BV: breeding value; Bayesian LASSO: Bayesian least absolute shrinkage and selection operator; SSGBLUP: single-step GBLUP; S.E: standard error; S.D: standard deviation; CFWR: clean fleece weight rate; SS: staple strength; FER: Fleece extension rate; FD: mean fiber diameter; FD_CV: Coefficient of variation of FD; SL: staple length; h^2: polygenic heritability; QTL: quantitative trait loci; LD: linkage disequilibrium; MAF: minor allele frequency

Declarations

Acknowledgements

We acknowledge Professor Jianlin Han and Professor Christian Keambou Tiambo (International Livestock Research Institute, Nairobi, 00100, Kenya) for reading through the manuscript and giving valuable comments.

Authors' contribution

BY and YY conceived and designed the experiments and explained the data. SZ analyzed the content of the data with the help of TG, CY, JL, MH, HZ and Christian. YW, WS, XW, TW and JL provided assistance with sample and data collection. SZ drafted the manuscript with the help of BY and YY.

Funding

This research: including experimental design, sample collection, data analysis, and manuscript writing, was funded by the Agricultural Science and Technology Innovation Program of China (CAAS-ASTIP-2015-LIHPS), the Selection of Scientific Research Topics for Significant Production of the Chinese Academy of Agricultural Sciences (CAAS-ZDXT2018006), and the Modern China Wool Cashmere Technology Research System (CARS-39-02).

Availability of data and materials

All analysis results data generated during this study are included in this manuscript. Requests for the raw data should be made to the corresponding authors.

Ethics approval and consent to participate

All animal work carried out in the current study was performed per the guidelines for the care and use of laboratory animals promulgated by the State Council of the People's Republic of China. The study was approved (License Number: 2019-008) by the Animal Management and Ethics Committee of Lanzhou, Institute of Animal Husbandry and Veterinary Sciences, Chinese Academy of Agricultural Sciences.

Consent to publish

Not applicable

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

References

1. Rabier C-E, Barre P, Asp T, Charmet G, Mangin B: **On the Accuracy of Genomic Selection**. *PLoS ONE* 2016, **11**.
2. Meuwissen THE, Hayes BJB, Goddard MEM: **Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps**. *Genetics* 2001, **157**:1819-1829.
3. Goertzel B, Pennachin C, Coelho L, Gurbaxani B, Maloney E, Jones J: **Combination of single nucleotide polymorphisms in neuroendocrine effector and receptor genes predict chronic fatigue syndrome**. *Pharmacogenomics* 2006, **7**:475-483.
4. Su G, Christensen O, Ostersen T, Henryon M, Lund M: **Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers**. *PloS one* 2012, **7**:e45293.
5. Iwata H, Jannink J-L: **Accuracy of Genomic Selection Prediction in Barley Breeding Programs: A Simulation Study Based On the Real Single Nucleotide Polymorphism Data of Barley Breeding Lines**. *Crop Science* 2011, **51**:1915.
6. Taylor J, Taylor K, Decker J: **Holsteins are the genomic selection poster cows**. *Proceedings of the National Academy of Sciences* 2016, **113**:201608144.
7. Taylor J, McKay S, Rolf M, Ramey H, Decker J, Schnabel R: **Genomic Selection in Beef Cattle**. In.; 2012: 211-233.
8. Cleveland M, Hickey J: **Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation**. *Journal of animal science* 2013, **91**.
9. Carillier-Jacquin C, Larroque H, Robert-Granié C: **Toward genomic selection in dairy goats**. *Productions Animales* 2017, **30**:19-30.
10. Werf JHJVD: **Potential benefit of genomic selection in sheep**. In: *Matching Genetics & Environment: A New Look at An Old Topic Conference of the Association for the Advancement of Animal Breeding & Genetics: 2009*.
11. Tsai HY, Hamilton A, Tinch AE, Guy DR, Gharbi K, Stear MJ, Matika O, Bishop SC, Houston RD: **Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array**. *Bmc Genomics* 2015, **16**(1):969.
12. Vallejo R, Leeds T, Fragomeni B, Gao G, Hernandez A, Misztal I, Welch T, Wiens G, Palti Y: **Evaluation of Genome-Enabled Selection for Bacterial Cold Water Disease Resistance Using Progeny Performance Data in Rainbow Trout: Insights on Genotyping Methods and Genomic Prediction Models**. *Frontiers in Genetics* 2016, **1**.
13. Legarra A, Robert-Granié C, Manfredi E: **Performance of Genomic Selection in Mice**. *Genetics* 2008, **180**:611-618.
14. Desta Z, Ortiz R: **Genomic selection: Genome-wide prediction in plant improvement**. *Trends in plant science* 2014, **19**.
15. Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M *et al*: **Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing**. *The Plant Genome* 2012, **5**.
16. Jia C, Zhao F, Wang X, Jianlin H, Zhao H, Liu G, Wang Z: **Genomic Prediction for 25 Agronomic and Quality Traits in Alfalfa (Medicago sativa)**. *Frontiers in Plant Science* 2018, **9**.
17. Sun C, VanRaden PM, Cole JB, O'Connell JR: **Improvement of Prediction Ability for Genomic Selection of Dairy Cattle by Including Dominance Effects**. *PLOS ONE* 2014, **9**(8):e103934.
18. Weller J, Ezra E, Ron M: **Invited review: A perspective on the future of genomic selection in**

dairy cattle. *Journal of Dairy Science* 2017, **100**.

19. Miglior F, Fleming A, Malchiodi F, Brito L, Martin P, Baes C: **A 100-Year Review: Identification and genetic selection of economically important traits in dairy cattle**. *Journal of Dairy Science* 2017, **100**:10251–10271.
20. Wiggans GR, Cole J, Hubbard S, Sonstegard T: **Genomic Selection in Dairy Cattle: The USDA Experience**. *Annual review of animal biosciences* 2016, **5**.
21. Singh B, Mal G, Gautam S, Mukesh M: **Whole-Genome Selection in Livestock**. In.; 2019: 349-364.
22. Di X, Matsuzaki H, Webster T, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G *et al*: **Dynamic Model Based Algorithms for Screening and Genotyping over 100 K Snps on Oligonucleotide Microarrays**. *Bioinformatics (Oxford, England)* 2005, **21**:1958-1963.
23. Habier DJBB: **Extension of the bayesian alphabet for genomic selection**. 2011, **12**(1):186.
24. Park T, Casella G: **The Bayesian Lasso**. *Journal of the American Statistical Association* 2008, **103**:681-686.
25. Gianola D: **Priors in Whole-Genome Regression: The Bayesian Alphabet Returns**. *Genetics* 2013, **194**.
26. Vanraden PM: **Efficient Methods to Compute Genomic Predictions**. *Journal of Dairy Science* 2008, **91**(11):4414-4423.
27. Aguilar I, Misztal I, Johnson D, Legarra A, Tsuruta S, Lawlor T: **Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score**. *Journal of dairy science* 2010, **93**:743-752.
28. Wang Z, Zhang H, Yang H, Wang S, Rong E, Pei W, Li H, Wang N: **Genome-Wide Association Study for Wool Production Traits in a Chinese Merino Sheep Population**. *PLOS ONE* 2014, **9**.
29. Ciani E, Emiliano L, D'Andrea M, Alloggio I, Marroni F, Ceccobelli S, Delgado JV, Sarti F, Kijas J, Lenstra J *et al*: **Merino and Merino-derived sheep breeds: A genome-wide intercontinental study**. *Genetics, selection, evolution : GSE* 2015, **47**:64.
30. Zhu S, Tingting G, Zhao H, Qiao G, Han M, Jianbin L, Yuan C, Wang T, Li F, Yaojing Y *et al*: **Genome-Wide Association Study Using Individual Single-Nucleotide Polymorphisms and Haplotypes for Erythrocyte Traits in Alpine Merino Sheep**. *Frontiers in Genetics* 2020, **11**.
31. Daetwyler HD, Swan A, Werf J, Hayes B: **Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multi-breed data assessed by cross-validation**. *Genetics, selection, evolution : GSE* 2012, **44**:33.
32. Visscher P, Yang J, Goddard M: **A Commentary on 'Common SNPs Explain a Large Proportion of the Heritability for Human Height' by Yang et al. (2010)**. *Twin research and human genetics : the official journal of the International Society for Twin Studies* 2010, **13**:517-524.
33. Yang J, Benyamin B, McEvoy B, Gordon S, Henders A, Nyholt D, Madden P, Heath A, Martin N, Montgomery G *et al*: **Common SNPs Explain a Large Proportion of the Heritability for Human Height**. *Nature genetics* 2010, **42**:565-569.
34. Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes Torres JM: **Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree**. *Genetics* 2009, **182**:375-385.
35. Liu Y, Lu S, Liu F, Shao C, Zhou Q, Wang N, Li Y, Yang Y, Zhang Y, Sun H *et al*: **Genomic Selection Using BayesCπ and GBLUP for Resistance Against Edwardsiella tarda in Japanese Flounder (Paralichthys olivaceus)**. *Marine Biotechnology* 2018, **20**.

36. Hayes B, Bowman P, Chamberlain AJ, Goddard M: **Invited Review: Genomic selection in dairy cattle: Progress and challenges**. *Journal of dairy science* 2009, **92**:433-443.
37. Cheng H, Kizilkaya K, Zeng J, Garrick D, Fernando R: **Genomic Prediction from Multiple-Trait Bayesian Regression Methods Using Mixture Priors**. *Genetics* 2018, **209**:genetics.300650.302018.
38. Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman P, Reich C, Mason BA, Goddard ME: **Erratum to "Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels" (J. Dairy Sci. 95:4114–4129)**. *Journal of dairy science* 2012, **95**:4114-4129.
39. Moghaddar N, Swan A, van der Werf J: **Genomic prediction of weight and wool traits in a multi-breed sheep population**. *Animal production science* 2014, **54**(5):544-549.
40. Hamadani A, Ganai N, Khan NN, Shanaz S, Ahmad T: **Estimation of Genetic, Heritability, and Phenotypic Trends for Weight and Wool Traits in Rambouillet Sheep**. *Small Ruminant Research* 2019, **177**.
41. Daetwyler HD, Hickey J, Henshall J, Dominik S, Gredler-Grandl B, Werf J, Hayes B: **Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population**. *Animal Production Science* 2010, **50**:1004.
42. Safari E, Fogarty N, Gilmour AR: **A review of genetic parameter estimates for wool, growth, meat and reproduction traits in sheep**. *Livestock Production Science* 2005, **92**:271-289.
43. Fogarty N: **Genetic parameters for live weight, fat and muscle measurements, wool production and reproduction in sheep: a review**. *Anim Breed Abstr* 1995, **63**:101-143.
44. Calus MPL, Meuwissen THE, Roos APW, De, Veerkamp RF, %J Genetics: **Accuracy of genomic selection using different methods to define haplotypes**. 2008, **178**(1):553.
45. Boustan A, Nejati Javaremi A, Moradi Shahrbabak M, Saatchi M: **Effect of using different number and type of records from different generations as reference population on the accuracy of genomic evaluation**. *Archiv fur Tierzucht* 2013, **56**:684-690.
46. Solberg T, Sonesson AK, Woolliams JA, Meuwissen T: **Genomic Selection Using Different Marker Types and Density**. *Journal of animal science* 2008, **86**:2447-2454.
47. Heffner E, Jannink J-L, Sorrells M: **Genomic Selection Accuracy using Multifamily Prediction Models in a Wheat Breeding Program**. *The Plant Genome Journal* 2011, **4**:65.
48. Campos G, Pong-Wong R, Daetwyler HD, Calus M: **Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding**. *Genetics* 2012, **193**.
49. Zhang Z, Erbe M, Jinlong H, Ober U, Gao N, Zhang H, Simianer H, Li J: **Accuracy of Whole Genome Prediction Using a Genetic Architecture Enhanced Variance-Covariance Matrix**. *G3-Genes Genomes Genetics* 2015, **5**.
50. Heslot N, Yang H-P, Sorrells M, Jannink J-L: **Genomic Selection in Plant Breeding: A Comparison of Models**. *Crop Science* 2012, **56**:146-160.
51. Ala Noshahr F, Rafat S, Imany-Nabiyyi R, Alijani S, Robert-Granié C: **Effects of Marker Density, Number of Quantitative Trait Loci and Heritability of Trait on Genomic Selection Accuracy**. *Iranian Journal of Applied Animal Science* 2017, **7**.
52. Medeiros de Oliveira Silva R, Fragomeni B, Lourenco D, Magalhães A, Irano N, Carvalheiro R, Canesin RC, Mercadante M, Boligon AA, Baldi F *et al*: **Accuracies of genomic prediction of feed efficiency traits using different prediction and validation methods in an experimental Nelore cattle population**. *Journal of Animal Science* 2016, **94**:3613-3623.

53. Bolormaa S, Pryce J, Kemper KE, Savin K, Hayes B, Barendse W, Zhang Y, Reich C, Mason BA, Bunch R *et al*: **Accuracy of prediction of genomic breeding values for residual feed intake and carcass and meat quality traits in Bos taurus, Bos indicus, and composite beef cattle**. *Journal of animal science* 2013, **91**.
54. Daetwyler HD, Hickey JM, Henshall JM, Dominik S, Gredler B, Werf JHJVD, Hayes BJJAPS: **Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population**. 2010, **50**(12):1004.
55. Daetwyler HD, Villanueva B, Woolliams J: **Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach**. *PLoS One 3 (2009) 10* 2008, **3**.
56. Honarvar M: **Accuracy of genomic prediction using RR-BLUP and Bayesian LASSO**. *European Journal of Experimental Biology* 2013, **2013**:42-47.
57. Heffner E, Jannink J-L, Iwata H, Souza E, Sorrells M: **Genomic Selection Accuracy for Grain Quality Traits in Biparental Wheat Populations**. *Crop Science* 2011, **51**:2597.
58. Dreisigacker S, Crossa J, Hearne S, Babu R, Prasanna B, Grondona M, Zambelli A, Windhausen V, Mathews K, Gorjanc G: **Evaluation of Genomic Selection Training Population Designs and Genotyping Strategies in Plant Breeding Programs Using Simulation**. *Crop Science* 2014, **54**:1476-1488.
59. Xiaoming M, Jia C, Fu D, Chu M, Ding X, Wu X, Guo X, Pei J, p.j B, Liang C *et al*: **Analysis of Hematological Traits in Polled Yak by Genome-Wide Association Studies Using Individual SNPs and Haplotypes**. *Genes* 2019, **10**:463.
60. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, Maller J, Sklar P, Bakker P, Daly M *et al*: **Plink: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses**. *American journal of human genetics* 2007, **81**:559-575.
61. Browning B, Browning S: **A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals**. *American journal of human genetics* 2009, **84**:210-223.
62. Perez P, De Los Campos G: **Genome-Wide Regression and Prediction with the BGLR Statistical Package**. *Genetics* 2014, **198**(2):483-495.

Figure Legends

Figure 1. Comparison of GP accuracy based on different density genotype datasets. The six traits were clean fleece weight rate (CFWR); staple strength (SS); fleece extension rate (FER); mean fiber diameter (FD); Coefficient of variation of FD (FD_CV); staple length (SL).

Figure 2. Based on genotype datasets of different densities, the GP accuracy of 5 models in different heritability level. On the left is the result for the H-datasets, and on the right is the result for the L-datasets. The six traits were clean fleece weight rate (CFWR); staple strength (SS); fleece extension rate (FER); mean fiber diameter (FD); Coefficient of variation of FD (FD_CV); staple length (SL). The five models were: BayesA (BA); BayesB (BB); BayesCπ (BC); Bayesion LASSO (BL); and GBLUP (GB).

Tables

Table 1. Descriptive statistics of phenotypic values of traits. [1] S.E, standard error; [2] S.D, standard deviation.

Table 2. Different GS methods and effects distribution.

Table 3. Estimates of additive and residual components of variance obtained adopting 'BGLR' for different datasets. [a] CFWR: clean fleece weight rate; SS: staple strength; FER: fleece extension rate; FD: mean fiber diameter; FD_CV: Coefficient of variation of FD; SL: staple length; [b] Polygenic heritability, the proportion of the additive effect variance to the total phenotypic variance.

Table 4. Comparison of prediction accuracies of 6 traits based on 2 datasets via 5 models. [a] Abbreviations of traits explained in Table 3; [b] S.E are in parenthesis; [c] BA: BayesA; BB; BayesB; BC: BayesCπ; BL: Bayesion LASSO; GB: genomic best linear unbiased prediction, GBLUP.

Figures

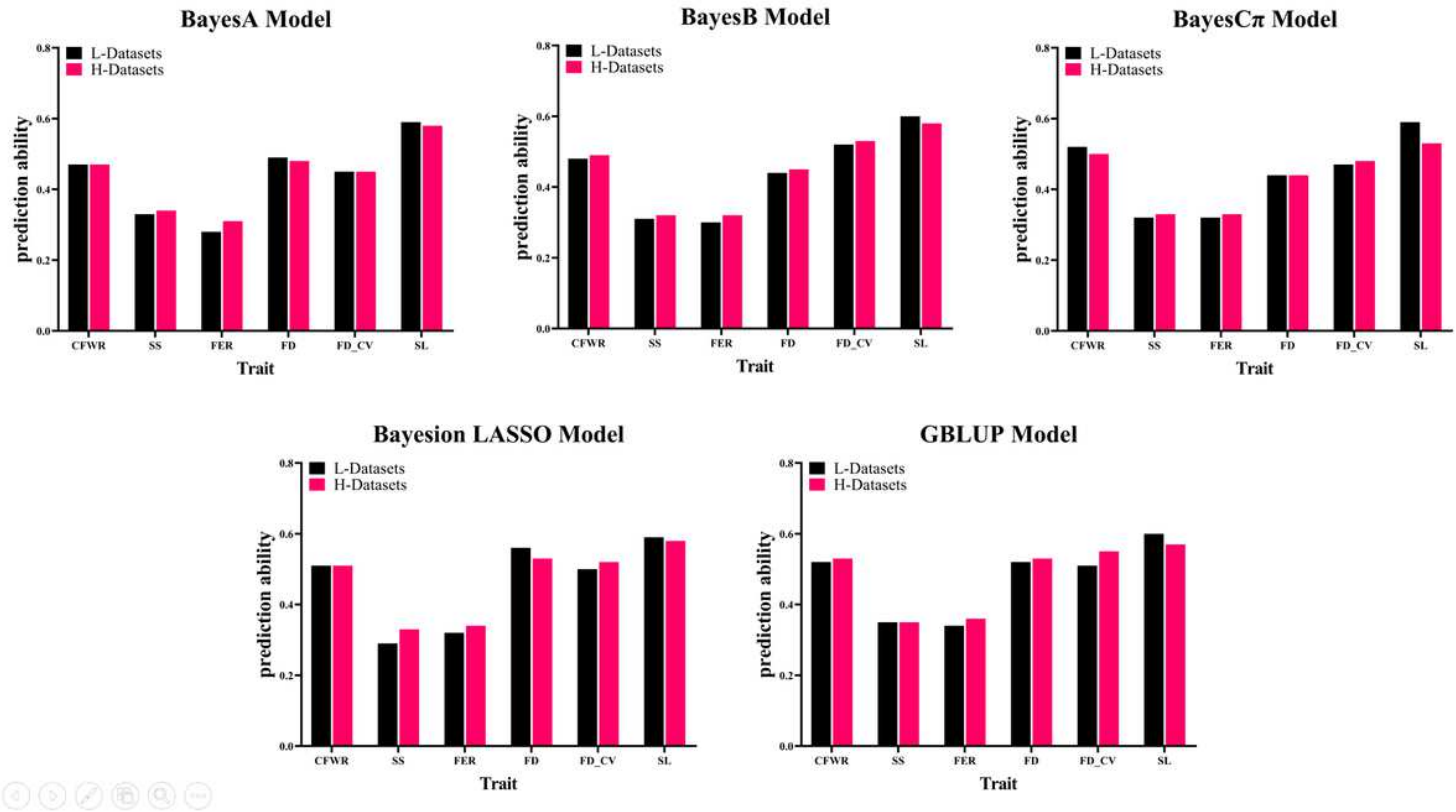

Figure 1

Comparison of GP accuracy based on different density genotype datasets. The six traits were clean fleece weight rate (CFWR); staple strength (SS); fleece extension rate (FER); mean fiber diameter (FD); Coefficient of variation of FD (FD_CV); staple length (SL).

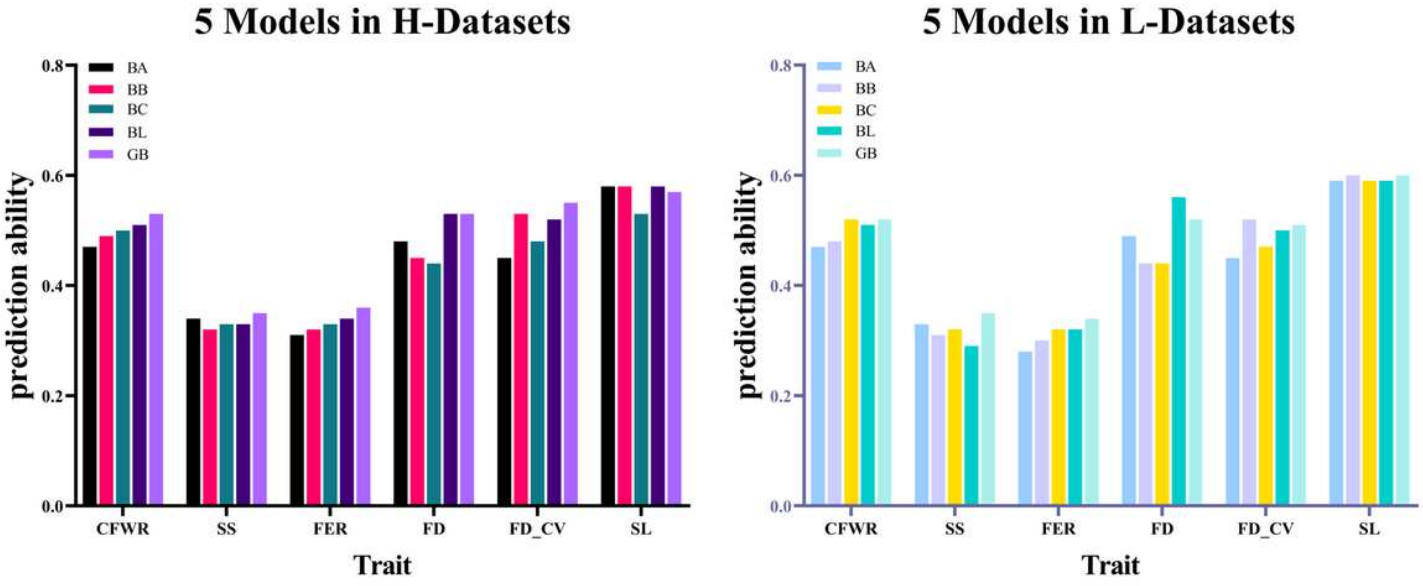

Figure 2

Based on genotype datasets of different densities, the GP accuracy of 5 models in different heritability level. On the left is the result for the H-datasets, and on the right is the result for the L-datasets. The six traits were clean fleece weight rate (CFWR); staple strength (SS); fleece extension rate (FER); mean fiber diameter (FD); Coefficient of variation of FD (FD_CV); staple length (SL). The five models were: BayesA (BA); BayesB (BB); BayesCπ (BC); Bayesion LASSO (BL); and GBLUP (GB).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.docx
- Table2.docx
- Table3.docx
- Table4.docx