Front. Agr. Sci. Eng. 2017, **4**(3): 268–278
DOI: 10.15302/J-FASE-2017164

Available online at http://engineering.cae.cn

REVIEW

Statistical considerations for genomic selection

Huimin KANG, Lei ZHOU, Jianfeng LIU (✉)

National Engineering Laboratory for Animal Breeding/Key Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture/College of Animal Science and Technology, China Agricultural University, Beijing 100193, China

Abstract Genomic selection is becoming increasingly important in animal and plant breeding, and is attracting greater attention for human disease risk prediction. This review covers the most commonly used statistical methods and some extensions of them, i.e., ridge regression and genomic best linear unbiased prediction, Bayesian alphabet, and least absolute shrinkage and selection operator. Then it discusses the measurement of the performance of genomic selection and factors affecting the prediction of performance. Among the measurements of prediction performance, the most important and commonly used measurement is prediction accuracy. In simulation studies where true breeding values are available, accuracy of genomic estimated breeding value can be calculated directly. In real or industrial data studies, either training-testing approach or *k*-fold cross-validation is commonly employed to validate methods. Factors influencing the accuracy of genomic selection include linkage disequilibrium between markers and quantitative trait loci, genetic architecture of the trait, and size and composition of the training population. Genomic selection has been implemented in the breeding programs of dairy cattle, beef cattle, pigs and poultry. Genomic selection in other species has also been intensively researched, and is likely to be implemented in the near future.

Keywords genomic estimated breeding value, genomic selection, linkage disequilibrium, statistical methods

1 Introduction

In the past, genetic evaluation of livestock was solely based on information of phenotype and pedigree. With the emerging genome genotyping technology, marker-assisted selection (MAS)[1], which indirectly selects individuals utilizing information of markers associated with quantitative trait loci (QTL), has attracted wide attention in livestock and plant breeding. However, practical application of MAS in breeding programs has not met the initial expectations[2]. This is because the traits of interest are usually controlled by a large number of genes or QTL with small effects, and only a small number of markers are available for MAS.

In 2001, Meuwissen et al.[3] proposed the methodology of genomic selection (GS), a variant of MAS, which applies predictions using markers spanning the whole genome. Markers are assumed to be in linkage disequilibrium (LD) with QTL, and therefore a larger proportion of additive genetic variance can be explained by markers. With this approach, accuracy of prediction in simulation studies reached 0.85. Schaeffer[4] demonstrated that such high accuracy can potentially double the genetic gain in progeny testing schemes and save 92% of the costs of progeny testing in dairy cattle. As a consequence, genomic selection is being gradually adopted in the genetic evaluation of dairy cattle in both developed and developing countries, and in other domestic animals and plants, and for human diseases. As illustrated in Fig. 1, implementation of GS commonly involves the following steps: (1) constructing the training/reference population, where individuals have both genotypic and phenotypic information; (2) genotyping candidate individuals; (3) using the appropriate statistical method to obtain genomic estimated breeding value (GEBV) of candidates; and (4) selecting individuals based on GEBV.

Since the proposal, appropriate statistical methods for GS have been extensively developed and GS has become more widely practiced. The purpose of this paper is to review the commonly used statistical methods and the implementation of GS in various fields.

2 Statistical methods

Except for the genomic best linear unbiased prediction (GBLUP) and single-step method, which is introduced in section ridge regression and GBLUP, other methods

Received March 14, 2017; accepted June 4, 2017

Correspondence: liujf@cau.edu.cn

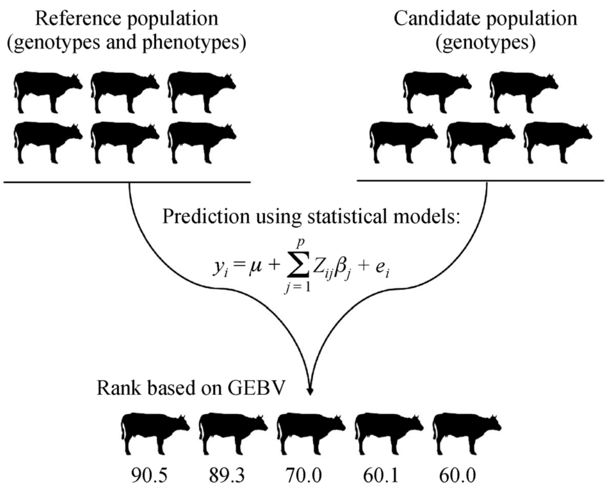
Reference population
(genotypes and phenotypes)
Candidate population
(genotypes)
Prediction using statistical models:
$y_i = \mu + \sum_{j=1}^{p} Z_{ij}\beta_j + e_i$
Rank based on GEBV
90.5
89.3
70.0
60.1
60.0

Fig. 1 Implementation of genomic selection. To implement genomic selection, a reference population should be constructed, in which individuals are genotyped and phenotyped. Based on the reference population, genomic estimated breeding values (GEBV) are obtained by using statistical methods for candidate individuals only having genotypic information. Individuals are selected according to the rank of their GEBV.

introduced here use the following basic linear regression model to predict marker effects:

$$y_i = \mu + \sum_{j=1}^{p} Z_{ij}\beta_j + e_i \tag{1}$$

where $i = 1, 2, \ldots, n, j = 1, 2, \ldots, p$, n is the total number of individuals, p is the total number of markers, y_i is the response variable, μ is the overall mean, Z_{ij} is the element in the incidence matrix corresponding to genotype of marker$j$ for individual i, β_j is the estimated effect of marker$j$, and e_i is the residual effect. Then, marker effects are multiplied with the genotype for each individual to obtain its GEBV, i.e., $\mathrm{GEBV}_i = \sum_{j=1}^{p} Z_{ij}\beta_j$.

2.1 Ridge regression and GBLUP

Ridge regression[5] was one of the first methods proposed for genomic selection[3]. The ridge regression estimator:

$$\hat{\boldsymbol{\beta}}_{\mathrm{RR}} = (\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{I}_p)^{-1}\mathbf{Z}'(\mathbf{y} - \mu\mathbf{1}_n) \tag{2}$$

Eq. (2) is obtained by minimizing the penalized sum of squares:

$$\sum_{i=1}^{n}\left((y_i - \mu) - \sum_{j=1}^{p} Z_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{3}$$

which additionally constrains the sum of the squared coefficients, compared to ordinary least squares. Therefore, ridge regression not only minimizes the residual sum of squares, but also has a penalty term, i.e., $\lambda\sum_{j=1}^{p}\beta_j^2$, on the estimated marker effects β. λ ($\geqslant 0$) is the penalty parameter which controls the strength of shrinkage. The parameter λ can either be fixed or estimated by different methods, such as cross-validation employed in Howard et al.'s study[6]. Ridge regression shrinks the coefficients of correlated effects equally toward zero, but does not force them to zero[7].

Unlike the properties of the ordinary least squares estimator, the ridge regression estimator is biased when $\lambda \neq 0$. However, the advantage of ridge regression over ordinary least squares is that ridge regression can be used when (1) the number of markers exceeds the number of observations, and (2) variables are correlated with each other.

The estimator of ridge regression BLUP (RR-BLUP) is

$$\hat{\boldsymbol{\beta}}_{\mathrm{RR-BLUP}} = \left(\mathbf{Z}'\mathbf{Z} + \frac{\sigma_e^2}{\sigma_\beta^2}\mathbf{I}_p\right)^{-1}\mathbf{Z}'(\mathbf{y} - \mu\mathbf{1}_n) \tag{4}$$

where σ_e^2 is the residual variance ($\mathrm{var}(\mathbf{e}) = \mathbf{I}_n\sigma_e^2$) and σ_β^2 is the variance of regression coefficients ($\mathrm{var}(\boldsymbol{\beta}) = \mathbf{I}_p\sigma_\beta^2$). $\hat{\boldsymbol{\beta}}_{\mathrm{RR-BLUP}}$ is equal to $\hat{\boldsymbol{\beta}}_{\mathrm{RR}}$ with $\lambda = \frac{\sigma_e^2}{\sigma_\beta^2}$ in Eq. (2). The variance parameters can be estimated by restricted maximum likelihood.

With RR-BLUP, marker effects are calculated first, then GEBV of individual i is calculated as $\sum_{j=1}^{p} Z_{ij}\hat{\beta}_j$. VanRaden[8] deduced GBLUP which can obtain the same results as RR-BLUP directly[9].

The GBLUP model is given as follows:

$$\mathbf{y} = \mu\mathbf{1}_n + \mathbf{Z}\mathbf{g} + \mathbf{e}$$

where $\mathbf{y}$ is a $n \times 1$ vector of the response variable; μ is the overall mean; $\mathbf{1}_n$ is a vector of n ones; g is the $n \times 1$ vector of additive genomic effects with distribution of $N(\mathbf{0},\mathbf{G}\sigma_g^2)$, $\mathbf{Z}$ is the corresponding incidence matrix; and $\mathbf{e}$ is the vector of random residuals with distribution of $N(\mathbf{0},\mathbf{D}\sigma_e^2)$. $\mathbf{D}$ is a diagonal matrix. The $\mathbf{G}$ is the genomic relationship matrix, which is usually constructed by the first method of VanRaden[8]:

$$\mathbf{G} = \frac{(\mathbf{M}-\mathbf{P})(\mathbf{M}-\mathbf{P})'}{2\sum_{j=1}^{n} p_j(1-p_j)}$$

where $\mathbf{M}$ is a matrix of single nucleotide polymorphism (SNP) genotypes for each individual, $\mathbf{P}$ is a matrix of 2 times the observed allele frequency of the second allele p at locus j (p_j). Ideally, allele frequencies in base population should be used in the construction of $\mathbf{G}$. However, they are not available in most practical situations. The observed allele frequencies of genotyped individuals are commonly used in studies and applications[10,11].

Legarra et al.[12] and Christensen and Lund[13] developed in parallel the basic theory for single-step genomic selection. They derived an extended relationship matrix ($\mathbf{H}$) and its inverse involving both genotyped and non-genotyped individuals:

$$\mathbf{H} = \begin{bmatrix} \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G}-\mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{12} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{12} & \mathbf{G} \end{bmatrix} \tag{5}$$

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1}-\mathbf{A}_{22}^{-1} \end{bmatrix} \tag{6}$$

where $\mathbf{G}$ is the aforementioned genomic relationship matrix for the genotyped individuals, $\mathbf{A}_{11}$, $\mathbf{A}_{12}$ and $\mathbf{A}_{22}$ are partitions of $\mathbf{A}$, the numerator relationship matrix based on pedigree, and subscripts 1 and 2 refer to non-genotyped and genotyped individuals, respectively.

Using $\mathbf{H}^{-1}$ in the mixed model equations, breeding values of both genotyped and non-genotyped individuals can be obtained simultaneously by solving the mixed model equations. The single-step method does not perform the works of multi-step methods, it is a unified approach to use phenotype, pedigree and genomic information simultaneously for genomic prediction. The single-step method overcomes drawbacks of multi-step methods, including loss of information, inaccuracy and bias[14].

For simplicity of implementation of the single-step method, $\mathbf{G}^{-1}$ and $\mathbf{A}_{22}^{-1}$ can be created explicitly. The corresponding computing time is proportional to n^3, where n is the number of individuals with genotypes. With this method, $\mathbf{G}^{-1}$ and $\mathbf{A}_{22}^{-1}$ can be computed for no more than perhaps 150000 individuals due to memory and computing time limitations[15]. However, the number of genotyped individuals is growing rapidly. In dairy cattle, more than 1.6 million Holsteins has been genotyped in the USA (Council On Dairy Cattle Breeding; https://www.uscdcb.com/Genotype/cur_freq.html). Misztal et al.[16] suggested calculating a sparse inverse of $\mathbf{G}$ using the algorithm for proven and young (APY) animals, which is gaining popularity. With APY, the genotyped individuals are divided into a group of core individuals and a group of noncore individuals. Then the approximate inverse of $\mathbf{G}$ is set up with formulas

$$\mathbf{G}_{\mathrm{APY}}^{-1} = \begin{bmatrix} \mathbf{G}_{cc}^{-1} + \mathbf{G}_{cc}^{-1}\mathbf{G}_{cn}M_{nn}^{-1}\mathbf{G}'_{cn}\mathbf{G}_{cc}^{-1} & -\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn}M_{nn}^{-1} \\ \mathbf{M}_{nn}^{-1}\mathbf{G}'_{cn}\mathbf{G}_{nn}^{-1} & \mathbf{M}_{nn}^{-1} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{G}_{\mathrm{APY}}^{cc} & \mathbf{G}_{\mathrm{APY}}^{cn} \\ \mathbf{G}_{\mathrm{APY}}^{cn} & \mathbf{M}_{nn}^{-1} \end{bmatrix}$$

and

$$\mathbf{M}_{nn} = \mathrm{diag}\{g_{ii} - \mathbf{g}'_{ci}\mathbf{G}_{cc}^{-1}\mathbf{g}'_{ci}\}$$

where subscript c refers to core individuals and subscript n refers to noncore individuals, g_{ii} is the ith diagonal element of $\mathbf{G}_{cc}$ and $\mathbf{g}_{ci}$ is the ith column of $\mathbf{G}_{cn}$ [17].

It was demonstrated that the approximate value of $\mathbf{G}^{-1}$ is very accurate when the number of core animals is 10k in a population of Holstein[17]. APY has a linear computing and memory cost for noncore animals[16]. Therefore, the computing requirements are dramatically lowered. Ostersen et al.[18] recommended that the core group should represent all generations and maximize the number of genotyped offspring in regards of prediction accuracy and convergence rate.

When the mixed model equations are solved with the commonly used preconditioned conjugate gradient, $\mathbf{A}_{22}^{-1}$ is only used in $\mathbf{A}_{22}^{-1}\mathbf{q}$, where $\mathbf{q}$ is a vector. As shown by Strandén and Mäntysaari[19]:

$$\mathbf{A}_{22}^{-1} = \mathbf{A}^{22} - \mathbf{A}^{21}(\mathbf{A}^{11})^{-1}\mathbf{A}^{12}$$

where $\mathbf{A}^{11}$, $\mathbf{A}^{21}$, $\mathbf{A}^{12}$ and $\mathbf{A}^{22}$ are partitions of $\mathbf{A}^{-1}$. Therefore,

$$\begin{aligned} \mathbf{A}_{22}^{-1}\mathbf{q} &= [\mathbf{A}^{22} - \mathbf{A}^{21}(\mathbf{A}^{11})^{-1}\mathbf{A}^{12}]\mathbf{q} \\ &= \mathbf{A}^{22}\mathbf{q} - \mathbf{A}^{21}(\mathbf{A}^{11})^{-1}\mathbf{A}^{12}\mathbf{q} \end{aligned}$$

In calculation of $\mathbf{t} = \mathbf{A}^{21}(\mathbf{A}^{11})^{-1}\mathbf{A}^{12}\mathbf{q}$, $\mathbf{t}_1 = \mathbf{A}^{12}\mathbf{q}$, $\mathbf{t}_2 = (\mathbf{A}^{11})^{-1}\mathbf{t}_1$ and $\mathbf{t} = \mathbf{A}^{21}\mathbf{t}_2$ can be calculated sequentially. The vector $\mathbf{t}_2$ is the solution of sparse equations$\mathbf{A}^{11}\mathbf{t}_2 = \mathbf{t}_1$, which can be easily solved with the sparse Cholesky decomposition of $\mathbf{A}^{11}$ [20].

2.2 Bayesian alphabet

Bayesian alphabet refers to a number of alphabets used to indicate various Bayesian regression models, which differ in their prior assumptions, while sharing the same phenotypic model as Eq. (1)[21]. Bayesian inference derives the posterior distribution from a prior and likelihood function according to the Bayes theorem. Estimates of marker effects are based on the posterior distribution. Bayesian inferences are usually realized by Gibbs sampling or Metropolis-Hasting algorithm.

Bayesian alphabet begins with BayesA and BayesB where the data are modeled at two levels, i.e., the level of data and the level of the variances of marker effects[3]. The models at the level of data (including fixed effects and random effects) are equal to that with RR-BLUP, except for assumption of the variances of marker effects. With RR-BLUP, markers share the same variance. However, with BayesA, variances of marker effects are different from each other, and they have the same prior distribution, i.e., the scaled inverted chi-square distribution, $\chi^{-2}(v,S)$, where S is the scale parameter and v is the degrees of freedom. BayesB further assumed a mixture prior distribution for variances of marker effects. That is, $\sigma^2_{\beta j}=0$ with probability π, and $\sigma^2_{\beta j}\sim\chi^{-2}(v,S)$ with probability $(1-\pi)$.

Gianola et al.[21] pointed out the drawbacks of BayesA and BayesB concerning the prior variances of marker effects. With BayesA and BayesB, the shrinkage of marker effects is strongly influenced by the hyperparameters. Habier et al.[22] improved them with the proposed BayesCπ and BayesDπ. BayesCπ assumes a common variance for markers having effects and in BayesDπ the scale parameter S is estimated instead of being specified by the users. The proportion of markers with nonzero effect, π, is estimated in both BayesCπ and BayesDπ.

After the proposal of single-step GBLUP, Fernando et al.[23] presented the single-step Bayesian regression models which combines genotyped and non-genotyped individuals as in single-step GBLUP. The model is as following:

$$\mathbf{y}=\mathbf{X}^*\boldsymbol{\beta}^*+\mathbf{W}\boldsymbol{\alpha}+\mathbf{U}\boldsymbol{\varepsilon}+\mathbf{e}$$

with

$$\mathbf{X}^*=\begin{bmatrix}\mathbf{X}_1 & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(-\mathbf{1})\\ \mathbf{X}_2 & -\mathbf{1}\end{bmatrix},\boldsymbol{\beta}^*=\begin{bmatrix}\boldsymbol{\beta}\\ \mu_g\end{bmatrix},$$

$$\mathbf{W}=\begin{bmatrix}\mathbf{Z}_1\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{M}_2\\ \mathbf{Z}_2\mathbf{M}_2\end{bmatrix}\text{ and }\mathbf{U}=\begin{bmatrix}\mathbf{Z}_1\\ \mathbf{0}\end{bmatrix}$$

where subscript 1 and 2 refer to non-genotyped and genotyped individuals, $\boldsymbol{\alpha}$ is the vector of partial-regression coefficients of the marker covariates, $\mathbf{M}_2$ is the matrix of marker covariates of genotyped individuals, $\mu_g=\mathbf{k}'\boldsymbol{\alpha}$ where $\mathbf{k}$ is the vector of expected values of marker covariates for a random individual in the absence of selection, and $\boldsymbol{\varepsilon}$ is used to account for deviations of imputed marker covariates $\hat{\mathbf{M}}_1=\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{M}_2$ from actual values for non-genotyped individuals. $\boldsymbol{\beta}$ is the vector of fixed effects, $\mathbf{X}_1$ and $\mathbf{X}_2$ are the corresponding design matrices of $\boldsymbol{\beta}$, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are design matrices, $\mathbf{A}_{12}$ and $\mathbf{A}_{22}$ are partitions of $\mathbf{A}$, and $\mathbf{e}$ is the vector of residuals.

The vector of predicted breeding values is

$$\hat{\mathbf{g}}=\begin{bmatrix}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}(-\mathbf{1})\\ -\mathbf{1}\end{bmatrix}\hat{\mu}_g+\begin{bmatrix}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{M}_2\\ \mathbf{M}_2\end{bmatrix}\hat{\boldsymbol{\alpha}}+\mathbf{U}\hat{\boldsymbol{\varepsilon}}$$

2.3 Least absolute shrinkage and selection operator

Least absolute shrinkage and selection operator (LASSO) was proposed by Tibshirani[24] to overcome the limitations of ordinary least squares. The only difference between LASSO and ridge regression is that LASSO uses the penalty function $\sum_{j=1}^{p}|\beta_j|$ instead of $\sum_{j=1}^{p}\beta_j^2$ in ridge regression. With this penalty function, LASSO sets some coefficients to zero, but shrinks nonzero coefficients less strongly than ridge regression. Therefore, LASSO performs both variable selection and shrinkage of regression coefficients. Several efficient algorithms have been developed for implementation of LASSO[7,25]. Park and Casella[26] presented a Bayesian version of LASSO and suggested its implementation using Gibbs sampling.

However, LASSO has some obvious shortcomings. First, in a high-dimensional case ($n<<p$), LASSO can select at most n nonzero regression coefficients[27]. Second, LASSO does not perform group selection, that is, when variables in the group are pairwise correlated, LASSO selects only one of them and this selection is arbitrary[27]. Third, LASSO fails to consistently select this variable in some circumstances[28]. For instance, LASSO gives different results with and without implementation of normalization of regression coefficients.

To overcome the first two drawbacks of LASSO, Zou and Hastie[27] proposed Elastic Net (EN) by using a weighted average of the penalty function in ridge regression and LASSO, i.e., $\alpha\sum_{j=1}^{p}|\beta_j|+(1-\alpha)\sum_{j=1}^{p}\beta_j^2$ with $0\leqslant\alpha\leqslant1$. Thus, EN involves two tuning parameters, λ and α, where λ is a penalty parameter as in Eq. (3). Zou[28] presented adaptive lasso with penalty function $\sum_{j=1}^{p}\hat{\omega}_j|\beta_j|$, where $\hat{\omega}_j$ represents the adaptive data-driven weights. Adaptive lasso overcomes the inconsistency of LASSO.

In addition to the statistical methods presented above, a variety of other methods have been developed for GS, including extensions of the above methods as well as new strategies proposed for GS. Some examples of extensions are TABLUP[29], pseudo single-step[30], single-step random regression test-day model[31], EM-BayesA[32], BayesD0, D1, D2 and D3[33], BayesR[34], and Bayesian multivariate antedependence model[35]. Other approaches include Nadaraya-Watson estimator[36], reproducing kernel Hilbert space[36], support vector machine[37] and neural network[38].

Among these models, GBLUP has been the most commonly used, followed by Bayesian alphabet. Other methods generally have limited advantage in terms of prediction accuracy, and they are more complicated to implement. As more and more individuals are genotyped, the APY is a useful approach to approximate the inverse of **G** in single-step GBLUP. In single-step Bayesian regression, the matrix inversion is not required. It also has the advantage that computing time and memory increase linearly with the number of observations and number of markers. If the strategy for parallelization presented by Fernando et al.[23] can be realized with computer clusters, single-step Bayesian regression can also be implemented in routine applications.

3 Prediction performance

3.1 Measures of prediction performance

Genomic selection was proposed to predict breeding values of individuals. Therefore, the most important measure of prediction performance is prediction accuracy. Prediction accuracy has a linear relationship with genetic response to selection. The higher the accuracy, the larger the response to selection will be obtained.

In simulation studies where true breeding values (TBV) are known, prediction accuracy is calculated as the correlation between estimated breeding value (EBV) and TBV. However, in empirical studies, TBV are not available. The most commonly used variables are phenotypes (original phenotype or phenotype adjusted for fixed effects)[11,39], averages of offspring performance[40] and EBV[39]. As these variables contain residual effects, the estimated correlation is commonly divided by the square root of heritability or the square root of the average reliability of the validation individuals. EBV is not recommended, because they are regressed toward the mean depending on their accuracy, whereas the other two variables are not[41]. When EBV have to be used, they can be divided by their reliability calculated as $r^2 = 1 - \frac{SEP^2}{\sigma_a^2}$, where SEP is the standard error of prediction and σ_a^2 is the additive genetic variance.

Apart from prediction accuracy, other commonly used measurements of prediction performance are mean squared error (MSE) of prediction, the area under the receiver operating curve (AUC) and bias. MSE is usually computed as the average square of the difference between TBV (or its alternatives) and EBV centered on zero. MSE assesses the overall quality of prediction. The small MSE indicates the estimator is precise and accurate. AUC is used in genomic prediction of binary/disease traits. The larger the AUC, the better the prediction. An AUC of 1 indicates the perfect prediction. Bias is measured as the regression coefficient of TBV (or its alternatives) on EBV. Patry and Ducrocq[42] found that breeding values of preselected young sires and their daughters were significantly underestimated in a simulation study. Bias is important when a proportion of GEBV is combined with other measurements to select individuals[41].

3.2 Validation strategies in GS

In simulation studies where TBV is available, the measurement of prediction, e.g., accuracy, is commonly the average of correlation of TBV and GEBV of a certain number of replicates[3,31,35].

In empirical studies, two alternative strategies are employed. When accuracies of some individuals are very high based on pedigree and phenotype, the training-testing approach is used. Those highly accurate EBV can be used as a perfect alternative to TBV. For example, in dairy cattle, it is common that elite bulls have accuracies of 0.99. In this situation, individuals are divided into training and test population, usually based on a specific year[40,43,44]. In validation studies, phenotypes collected after the specific year are masked.

When the whole data set is small, *k*-fold cross validation is a good choice[31,39,40]. In *k*-fold cross-validation, individuals are partitioned into *k* subsets with nearly equal size. One subset is retained as test set and the remainder is used as training sets. Phenotypes of individuals in the test set are masked. Breeding values of these individuals are predicted and then used to measure prediction performance. This process is then repeated *k* times, ensuring each subset is used only once as a test set. The value of *k* is usually 5 or 10. The number of subsets should be sufficient to limit the sampling variance of measures of prediction performance. Meanwhile, the size of training set should be large enough to provide a meaningful prediction[41].

To avoid inflated accuracy resulting from close relationship (e.g., family relationship) between training and test individuals[45,46], partitioning can be based on family, strains and lines[31,47]. Saatchi et al.[48] proposed a way of grouping individuals using *k*-mean clustering method based on elements of pedigree numerator relationship matrix (**A** matrix).

Besides *k*-fold cross-validation, repeated random subsampling validation has also been investigated[35]. This approach randomly splits the whole data set into training and test sets. Then genomic prediction is performed and estimates of statistics are calculated based on the repeated calculations.

3.3 Factors affecting accuracy of prediction

If a large number of QTL with small effects contribute to the trait, the following formula can be used to derive the upper bound of accuracy of prediction:

$$r = \sqrt{\frac{N_p}{N_p + Me/h^2}} \tag{7}$$

where N_p is the number of individuals in the training population, h^2 is the heritability of the trait, and Me is the number of independent chromosome segments[49–51]. Three formulas were used to estimate Me [46]. The first one was on the basis of Goddard et al.'s study[52]: $Me = \frac{1}{\mathrm{var}(\mathbf{G}-\mathbf{A})}$, where $\mathbf{G}$ is the genomic relationship matrix based on markers and $\mathbf{A}$ is the numerator relationship matrix based on pedigree. The second one is $Me = 2N_eL/\ln(4N_eL)$ [51] and the third one is $Me = 2N_eL$ [9], where N_e is the effective population size and L is the genome size. The Me estimates using the first and the second formula are similar to Me derived from Eq. (7) when accuracy was available. The third equation gives an inflated Me.

From Eq. (7), it has been demonstrated that the size of training population, Me, and heritability of the trait affect the accuracy of prediction. Equation (7) assumes that all the genetic variance can be captured by markers. However, this is not the case in real data applications. For example, Mehrban et al.[53] found that 65% and 66% genetic variances of backfat thickness and marbling scores were captured by the 50k SNP chip with GBLUP model in beef cattle. Meanwhile, genetic effects are assumed to be additive in Eq. (7). However, there may be dominant and epistatic effects. In addition, Eq. (7) does not take different family relationships between training and test sets into consideration. Therefore, the extent of LD between markers and QTL, and genetic basis of the traits and family relationship are also important factors influencing the accuracy.

3.3.1 Linkage disequilibrium between markers and QTL

The effect of LD between markers and QTL can be illustrated by the following facts:

(1) *Higher-density genotyping increases accuracy.* With higher density of markers, the LD between markers and QTL is expected to be greater. Therefore, higher accuracies were achieved with higher-density SNP chips[31,54,55]. Due to the relatively high cost of genotyping, many researchers focus on imputation of genotypes as an alternative[56–58].

(2) *Low accuracies in across-population prediction and small gain in accuracy from multi-population prediction.* In Hayes et al.'s study[59], accuracies of GS for Jersey cattle using a Holstein population as training data and vice versa ranged from − 0.06 to 0.23 for five traits. Moreover, they reported slightly increased accuracy of multi-population prediction compared to within population prediction. Hidalgo et al.[60] found similar results in across-population genomic selection in pigs. They also found that the effect of adding data from another population in the training set depends on traits. High genetic correlation of traits from different populations has a positive effect on prediction accuracy. Moghaddar et al.[61]. found zero or negative effect on accuracy of including distant breeds in the training population. These results may be explained by the difference in LD patterns, allele frequencies and QTL among different populations[62].

(3) *Decline in accuracy over generations.* Simulation and empirical studies show that accuracy of prediction declines with the number of generations between training and test population[11,31,63,64]. The decrease can be explained by the breakdown of LD between markers and QTL and less close family relationships between training and test sets. Moreover, different statistical methods were reported to have different persistencies in prediction.

The highest density of markers is in sequence data. It is anticipated that genomic selection with sequence data will directly use causal variants (no longer relying on LD) and prediction across breeds and generations will become more accurate. However, it cannot yet be unequivocally concluded that genomic selection using sequence data are a better choice. While some simulation studies showed the advantage of GS with sequence data[65–67], the benefit of sequence data was limited compared to medium/high density markers in other simulation studies and real data analyses[68–71]. Possible reasons for these limited benefits are: (1) noise is added when there are errors in genotypes resulted from sequencing[72] and imputation errors, especially imputation error for low-frequency variants[73]; (2) it is hard to observe rare alleles which may contribute to the trait in both training and test sets; (3) advantages of sequence data depend on genetic architecture, the largest increase was observed when all causal mutations were rare and had low LD with SNPs in chips[72,74]; and (4) the current GS models may not fit sequence data well, and more appropriate models and more information (e.g., biological information) for GS are needed[69].

3.3.2 Genetic architecture of a predicted trait

Genetic architecture herein refers to (1) heritability, (2) minimum allele frequency (MAF) of QTL, and (3) gene or QTL effects.

(1) *Heritability*. Heritability decides the amount of information the collected raw phenotypes can provide, then the accuracy of response variable of the training population, and finally the accuracy of prediction of test individuals. Heritability effects have been reported in many studies[40,75].

(2) *MAF of QTL*. SNPs with similar MAF can potentially have high LD, whereas LD between SNPs with different MAF is low. The extend of LD between SNPs and QTL depends on MAF of SNPs and QTL[76]. However, SNP chips are commonly designed to exclude SNPs with very low MAF. Therefore, if MAF of QTL is low, it is hard for SNPs on chips to have a strong LD. Decreased accuracy with low MAF of QTL has been observed in many studies[74,77,78].

(3) *Gene or QTL effects*. The commonly used models assume gene effects are additive. As dominance and epistatic interaction effects also contribute to genetic variance in some traits[79,80], including those effects in models may improve the prediction ability. Modeling dominance effects further improved prediction accuracy in some[79–81] but not all cases[79,82,83]. When information is sufficient to estimate marker effects, increase in accuracy may depend on the amount of genetic variance dominance effects taken into account[79]. Several approaches have been developed for modeling both additive and epistatic effects[84–86], and both increases and decreases in accuracy have been observed[87–89].

The effect of number of QTL on prediction accuracy is relatively small for GBLUP[49,77,78]. However, accuracy of Bayesian variable selection models decreases with the increase in number of QTL[50,90]. Moreover, as assumptions and information used vary with models, no model performs best in all situations. In general, Genomic BLUP models perform almost as well as Bayesian models in most real livestock scenarios and are simple to implement, and they are more commonly used in applications.

3.3.3 Training population

The following two factors for training population affect the prediction:

(1) *The size of training population*. Accuracy increases with the size of the training set[3,91]. The size of the training population should be large enough to accurately estimate the marker effects. It is better if individuals in training sets are less related to each other[92,93].

(2) *Relationship between training and test sets*. Many studies have proved that closer family relationships between training and test populations result in higher accuracies[46,92–94]. Moreover, the size of training sets affects the relative importance of family relationship and LD on prediction[46,95]. That is, family relationship is more important than LD in prediction with a small training set.

Rincent et al.[93] used a CDmean-based method to optimize the sampling of the training population, considering both the relatedness between individuals within the training population and the relationship between training and test individuals.

4 Applications of GS

Since the proposal of GS, tremendous studies have been performed to explore the theory and application of GS in animals, plants and humans. GS was first implemented in dairy cattle, and it has so far been most widely and successfully used for this. According to Garcia-Ruiz et al.'s study[96] and Taylor et al.'s study[97], the total number of Holsteins genotyped in the USA has reached 1.2 million by 2016, and rates of annual genetic improvement have increased by 50% to 100% for moderately heritable yield traits and by 300% to 400% for lowly heritable fitness traits (Fig. 2). In addition to the USA, many other countries have also implemented GS in dairy cattle, including Australia, Canada, China, Denmark, Germany, the Netherlands, New Zealand and Sweden. For other domestic animals, GS has already been implemented in pig (e.g., PIC, DanAvl, and Genesus) and poultry industries (e.g., Hy-Line, Cobb-Vantress and Aviagen). Moreover, genomic evaluation has now been implemented in sheep in Australia and New Zealand, dairy sheep in France, goats in France and the UK, and beef cattle (such as Angus, Charolais, Limousin, and Simmental) in France, North America and the UK.

Compared to domestic animals, GS is in its infancy in crop science and forestry, although a lot of studies have been conducted. The theory of GS can also be applied to human disease risk prediction. Case studies have been conducted on Celiac disease[98,99], type I diabetes[100], coronary heart disease[101], breast cancer[102], bladder cancer[103], skin cancer[104], and others. As Visscher[105] predicted, personalized genetics and genomics will become an integral part of health care and clinical practice in future.

5 Future prospects

In recent years, both the theory and the application of GS have been extensively explored. GS has been successfully implemented in domestic animals, and it is fully expected to be used for plant breeding and human disease risk prediction. With the declining cost of genotyping, more individuals and possibly all individuals in some populations will be genotyped. This may revolutionize GS methods. Moreover, other types of data are becoming more easily obtained, such as epigenome, transcriptome and proteome. Integrating data from multiple layers is expected to improve prediction performance as more

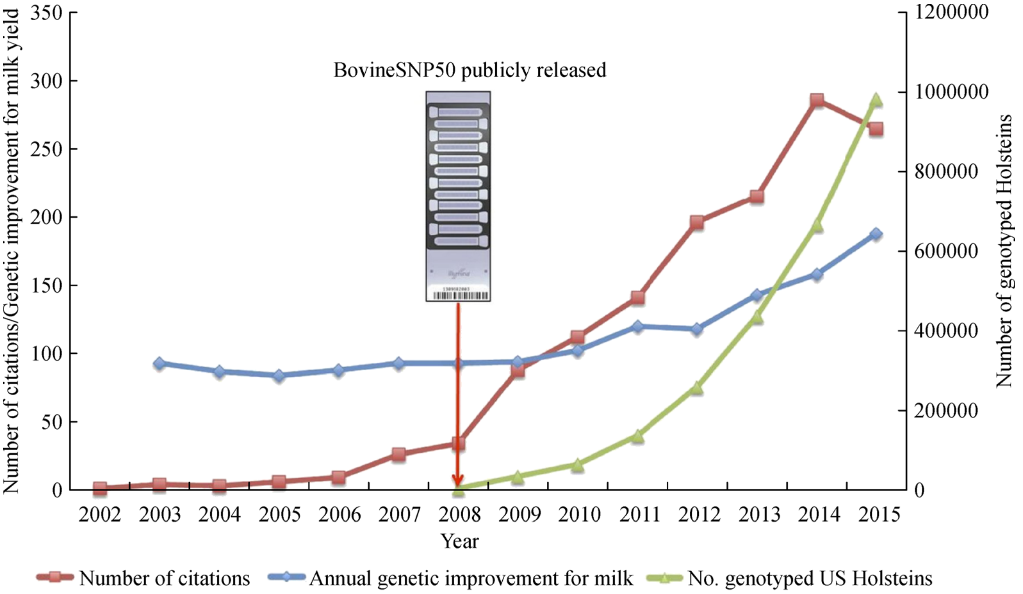

Fig. 2 Annual number of citations of reference[3], the rate of genetic improvement in milk production[96], and the number of Holstein cows chip-genotyped by December of each year from the Council for Dairy Cattle Breeding database (https://www.cdcb.us/Genotype/cur_density.html). Adapted from reference[97].

information is utilized. Methods modeling interactions of high-dimensional data need to be more fully developed to achieve improved genomic prediction.

Acknowledgements This work was supported by the National Natural Science Foundations of China (31272419, 31661143013), the National High Technology Research and Development Program of China (2013AA102503), China Agriculture Research System (CARS-36), and the Program for Changjiang Scholar and Innovation Research Team in University (IRT_15R62).

Compliance with ethics guidelines Huimin Kang, Lei Zhou, and Jianfeng Liu declare that they have no conflicts of interest or financial conflicts to disclose.

This article is a review and does not contain any studies with human or animal subjects performed by any of the authors.

References

1. Dekkers J C M, Hospital F. The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews Genetics*, 2002, **3**(1): 22–32
2. Dekkers J C. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. *Journal of Animal Science*, 2004, **82 E-Suppl**: E313–328
3. Meuwissen T H E, Hayes B J, Goddard M E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 2001, **157**(4): 1819–1829
4. Schaeffer L R. Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics*, 2006, **123**(4): 218–223
5. Hoerl A E, Kennard R W. Ridge regression- biased estimation for nonorthogonal problems. *Technometrics*, 1970, **12**(1): 55–67
6. Howard R, Carriquiry A L, Beavis W D. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3: Genes, Genomes, Genetics*, 2014, **4**(6): 1027–1046
7. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 2010, **33**(1): 1–22
8. VanRaden P M. Efficient methods to compute genomic predictions. *Journal of Animal Science*, 2008, **91**(11): 4414–4423
9. Hayes B J, Visscher P M, Goddard M E. Increased accuracy of artificial selection by using the realized relationship matrix. *Genetical Research*, 2009, **91**(1): 47–60
10. Christensen O F, Madsen P, Nielsen B, Ostersen T, Su G. Single-step methods for genomic evaluation in pigs. *Animal*, 2012, **6**(10): 1565–1571
11. Wolc A, Arango J, Settar P, Fulton J E, O'Sullivan N P, Preisinger R, Habier D, Fernando R, Garrick D J, Dekkers J C M. Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. *Genetics, Selection, Evolution*, 2011, **43**(1): 23
12. Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science*, 2009, **92**(9): 4656–4663
13. Christensen O F, Lund M S. Genomic prediction when some animals are not genotyped. *Genetics, Selection, Evolution*, 2010, **42**(1): 2
14. Legarra A, Christensen O F, Aguilar I, Misztal I. Single step, a general approach for genomic selection. *Livestock Science*, 2014, **166**: 54–65
15. Misztal I. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. *Genetics*, 2016, **202**(2): 401–409
16. Misztal I, Legarra A, Aguilar I. Using recursion to compute the

inverse of the genomic relationship matrix. *Journal of Dairy Science*, 2014, **97**(6): 3943–3952

17. Fragomeni B O, Lourenco D A L, Tsuruta S, Masuda Y, Aguilar I, Legarra A, Lawlor T J, Misztal I. Hot topic: Use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes. *Journal of Dairy Science*, 2015, **98**(6): 4090–4094
18. Ostersen T, Christensen O F, Madsen P, Henryon M. Sparse single-step method for genomic evaluation in pigs. *Genetics, Selection, Evolution*, 2016, **48**(1): 48
19. Strandén I, Mäntysaari E A. Comparison of some equivalent equations to solve single-step GBLUP. In: Proceedings of the 10th World Congress on Genetics Applied to Livestock Production 2014, Vancouver, Canada, 2015
20. Masuda Y, Misztal I, Tsuruta S, Legarra A, Aguilar I, Lourenco D A, Fragomeni B O, Lawlor T J. Implementation of genomic recursions in single-step genomic best linear unbiased predictor for US Holsteins with a large number of genotyped animals. *Journal of Animal Science*, 2016, **99**(3): 1968–1974
21. Gianola D, de los Campos G, Hill W G, Manfredi E, Fernando R. Additive genetic variability and the Bayesian alphabet. *Genetics*, 2009, **183**(1): 347–363
22. Habier D, Fernando R L, Kizilkaya K, Garrick D J. Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*, 2011, **12**(1): 186
23. Fernando R L, Dekkers J C, Garrick D J. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. *Genetics, Selection, Evolution*, 2014, **46**(1): 50
24. Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996, **58**(1): 267–288
25. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Annals of Statistics*, 2004, **32**(2): 407–451
26. Park T, Casella G. The Bayesian lasso. *Journal of the American Statistical Association*, 2008, **103**(482): 681–686
27. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 2005, **67**(2): 301–320
28. Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 2006, **101**(476): 1418–1429
29. Zhang Z, Liu J, Ding X, Bijma P, de Koning D J, Zhang Q. Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PLoS One*, 2010, **5**(9): e12648
30. Su G, Christensen O F, Ostersen T, Henryon M, Lund M S. Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. *PLoS One*, 2012, **7**(9): e45293
31. Kang H, Zhou L, Mrode R, Zhang Q, Liu J F. Incorporating single-step strategy into random regression model to enhance genomic prediction of longitudinal trait. *Heredity*, 2016
32. Sun X, Qu L, Garrick D J, Dekkers J C M, Fernando R L. A fast EM algorithm for BayesA-like prediction of genomic breeding values. *PLoS One*, 2012, **7**(11): e49157
33. Wellmann R, Bennewitz J. Bayesian models with dominance effects for genomic evaluation of quantitative traits. *Genetical Research*, 2012, **94**(1): 21–37
34. Erbe M, Hayes B J, Matukumalli L K, Goswami S, Bowman P J, Reich C M, Mason B A, Goddard M E. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of Dairy Science*, 2012, **95**(7): 4114–4129
35. Jiang J, Zhang Q, Ma L, Li J, Wang Z, Liu J F. Joint prediction of multiple quantitative traits using a Bayesian multivariate ante-dependence model. *Heredity*, 2015, **115**(1): 29–36
36. Gianola D, Fernando R L, Stella A. Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*, 2006, **173** (3): 1761–1776
37. Maenhout S, De Baets B, Haesaert G, Van Bockstaele E. Support vector machine regression for the prediction of maize hybrid performance. *Theoretical and Applied Genetics*, 2007, **115**(7): 1003–1013
38. Gianola D, Okut H, Weigel K A, Rosa G J M. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genetics*, 2011, **12**(1): 87
39. Fernandes G A, Rosa G J M, Valente B D, Carvalheiro R, Baldi F, Garcia D A, Gordo D G M, Espigolan R, Takada L, Tonussi R L, de Andrade W B F, Magalhaes A F B, Chardulo L A L, Tonhati H, de Albuquerque L G. Genomic prediction of breeding values for carcass traits in Nellore cattle. *Genetics, Selection, Evolution*, 2016, **48 (1)**: 1–8
40. Luan T, Woolliams J A, Lien S, Kent M, Svendsen M, Meuwissen T H E. The accuracy of Genomic Selection in Norwegian red cattle assessed by cross-validation. *Genetics*, 2009, **183**(3): 1119–1126
41. Daetwyler H D, Calus M P L, Pong-Wong R, de Los Campos G, Hickey J M. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics*, 2013, **193**(2): 347–365
42. Patry C, Ducrocq V. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. *Journal of Dairy Science*, 2011, **94**(2): 1011–1020
43. Li X, Wang S, Huang J, Li L, Zhang Q, Ding X. Improving the accuracy of genomic prediction in Chinese Holstein cattle by using one-step blending. *Genetics, Selection, Evolution*, 2014, **46**(1): 66
44. VanRaden P M, Van Tassell C P, Wiggans G R, Sonstegard T S, Schnabel R D, Taylor J F, Schenkel F S. Invited review: reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science*, 2009, **92**(1): 16–24
45. Habier D, Fernando R L, Dekkers J C M. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 2007, **177**(4): 2389–2397
46. Wientjes Y C J, Veerkamp R F, Calus M P L. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics*, 2013, **193**(2): 621–631
47. Legarra A, Robert-Granié C, Manfredi E, Elsen J M. Performance of genomic selection in mice. *Genetics*, 2008, **180**(1): 611–618
48. Saatchi M, McClure M C, McKay S D, Rolf M M, Kim J, Decker J E, Taxis T M, Chapple R H, Ramey H R, Northcutt S L, Bauck S, Woodward B, Dekkers J C M, Fernando R L, Schnabel R D, Garrick D J, Taylor J F. Accuracies of genomic breeding values in

American Angus beef cattle using K-means clustering for cross-validation. *Genetics, Selection, Evolution*, 2011, **43**(1): 40

49. Daetwyler H D, Villanueva B, Woolliams J A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One*, 2008, **3**(10): e3395
50. Daetwyler H D, Pong-Wong R, Villanueva B, Woolliams J A. The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, 2010, **185**(3): 1021–1031
51. Goddard M. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, 2009, **136**(2): 245–257
52. Goddard M E, Hayes B J, Meuwissen T H E. Using the genomic relationship matrix to predict the accuracy of genomic selection. *Journal of Animal Breeding and Genetics*, 2011, **128**(6): 409–421
53. Mehrban H, Lee D H, Moradi M H, IlCho C, Naserkheil M, Ibáñez-Escriche N. Predictive performance of genomic selection methods for carcass traits in Hanwoo beef cattle: impacts of the genetic architecture. *Genetics, Selection, Evolution*, 2017, **49**(1): 1
54. Habier D, Fernando R L, Dekkers J C M. Genomic selection using low-density marker panels. *Genetics*, 2009, **182**(1): 343–353
55. Solberg T R, Sonesson A K, Woolliams J A, Meuwissen T H E. Genomic selection using different marker types and densities. *Journal of Animal Science*, 2008, **86**(10): 2447–2454
56. Khatkar M S, Moser G, Hayes B J, Raadsma H W. Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. *BMC Genomics*, 2012, **13**(1): 538
57. Weng Z, Zhang Z, Ding X, Fu W, Ma P, Wang C, Zhang Q. Application of imputation methods to genomic selection in Chinese Holstein cattle. *Journal of Animal Science and Biotechnology*, 2012, **3**(1): 6
58. Weng Z, Zhang Z, Zhang Q, Fu W, He S, Ding X. Comparison of different imputation methods from low- to high-density panels using Chinese Holstein cattle. *Animal*, 2013, **7**(5): 729–735
59. Hayes B J, Bowman P J, Chamberlain A C, Verbyla K, Goddard M E. Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genetics, Selection, Evolution*, 2009, **41**(1): 51
60. Hidalgo A M, Bastiaansen J W M, Lopes M S, Harlizius B, Groenen M A M, de Koning D J. Accuracy of predicted genomic breeding values in purebred and crossbred pigs. *G3: Genes, Genomes, Genetics*, 2015, **5**(8): 1575–1583
61. Moghaddar N, Swan A A, van der Werf J H J. Comparing genomic prediction accuracy from purebred, crossbred and combined purebred and crossbred reference populations in sheep. *Genetics, Selection, Evolution*, 2014, **46**(1): 58
62. de Roos A P W, Hayes B J, Goddard M E. Reliability of genomic predictions across multiple populations. *Genetics*, 2009, **183**(4): 1545–1553
63. Akanno E C, Schenkel F S, Sargolzaei M, Friendship R M, Robinson J A B. Persistency of accuracy of genomic breeding values for different simulated pig breeding programs in developing countries. *Journal of Animal Breeding and Genetics*, 2014, **131**(5): 367–378
64. Muir W M. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *Journal of Animal Breeding and Genetics*, 2007, **124**(6): 342–355
65. Meuwissen T, Goddard M. Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics*, 2010, **185**(2): 623–631
66. Clark S A, Hickey J M, van der Werf J H J. Different models of genetic variation and their effect on genomic evaluation. *Genetics, Selection, Evolution*, 2011, **43**(1): 18
67. Iheshiulor O O M, Woolliams J A, Yu X, Wellmann R, Meuwissen T H E. Within- and across-breed genomic prediction using whole-genome sequence and single nucleotide polymorphism panels. *Genetics, Selection, Evolution*, 2016, **48**(1): 15
68. van Binsbergen R, Calus M P L, Bink M C A M, van Eeuwijk F A, Schrooten C, Veerkamp R F. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. *Genetics, Selection, Evolution*, 2015, **47**(1): 71
69. Pérez-Enciso M, Rincón J C, Legarra A. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. *Genetics, Selection, Evolution*, 2015, **47**(1): 43
70. Heidaritabar M, Calus M P L, Megens H J, Vereijken A, Groenen M A M, Bastiaansen J W M. Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. J*ournal of Animal Breeding and Genetics*, 2016, **133**(3): 167–179
71. Ni G, Cavero D, Fangmann A, Erbe M, Simianer H. Whole-genome sequence-based genomic prediction in laying chickens with different genomic relationship matrices to account for genetic architecture. *Genetics, Selection, Evolution*, 2017, **49**(1): 8
72. MacLeod I M, Hayes B J, Goddard M E. The effects of demography and long-term selection on the accuracy of genomic prediction with sequence data. *Genetics*, 2014, **198**(4): 1671–1684
73. Daetwyler H D, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum R F, Liao X, Djari A, Rodriguez S C, Grohs C, Esquerré D, Bouchez O, Rossignol M N, Klopp C, Rocha D, Fritz S, Eggen A, Bowman P J, Coote D, Chamberlain A J, Anderson C, VanTassell C P, Hulsegge I, Goddard M E, Guldbrandtsen B, Lund M S, Veerkamp R F, Boichard D A, Fries R, Hayes B J. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature Genetics*, 2014, **46**(8): 858–865
74. Druet T, Macleod I M, Hayes B J. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. *Heredity*, 2014, **112**(1): 39–47
75. Guo G, Lund M S, Zhang Y, Su G. Comparison between genomic predictions using daughter yield deviation and conventional estimated breeding value as response variables. *Journal of Animal Breeding and Genetics*, 2010, **127**(6): 423–432
76. Yang J, Benyamin B, McEvoy B P, Gordon S, Henders A K, Nyholt D R, Madden P A, Heath A C, Martin N G, Montgomery G W, Goddard M E, Visscher P M. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 2010, **42**(7): 565–569
77. Wientjes Y C J, Calus M P L, Goddard M E, Hayes B J. Impact of QTL properties on the accuracy of multi-breed genomic prediction. *Genetics, Selection, Evolution*, 2015, **47**(1): 42
78. Uemoto Y, Sasaki S, Kojima T, Sugimoto Y, Watanabe T. Impact of QTL minor allele frequency on genomic evaluation using real genotype data and simulated phenotypes in Japanese Black cattle. *BMC Genetics*, 2015, **16**(1): 134

79. Sun C, VanRaden P M, Cole J B, O'Connell J R. Improvement of prediction ability for genomic selection of dairy cattle by including dominance effects. *PLoS One*, 2014, **9**(8): e103934

80. Guo X, Christensen O F, Ostersen T, Wang Y, Lund M S, Su G. Genomic prediction using models with dominance and imprinting effects for backfat thickness and average daily gain in Danish Duroc pigs. *Genetics, Selection, Evolution*, 2016, **48**(1): 67

81. Nishio M, Satoh M. Including dominance effects in the genomic BLUP method for genomic evaluation. *PLoS One*, 2014, **9**(1): e85792

82. Santos V S, Martins Filho S, Resende M D, Azevedo C F, Lopes P S, Guimarães S E, Silva F F. Genomic prediction for additive and dominance effects of censored traits in pigs. *Genetics and Molecular Research*, 2016, **15**(4)

83. Ertl J, Legarra A, Vitezica Z G, Varona L, Edel C, Emmerling R, Götz K U. Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. *Genetics, Selection, Evolution*, 2014, **46**(1): 40

84. Xu S. An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics*, 2007, **63**(2): 513–521

85. Wang D, Salah El-Basyoni I, Stephen Baenziger P, Crossa J, Eskridge K M, Dweikat I. Prediction of genetic values of quantitative traits with epistatic effects in plant breeding populations. *Heredity*, 2012, **109**(5): 313–319

86. Cai X, Huang A, Xu S. Fast empirical Bayesian LASSO for multiple quantitative trait locus mapping. *BMC Bioinformatics*, 2011, **12**(1): 211

87. Hu Z, Li Y, Song X, Han Y, Cai X, Xu S, Li W. Genomic value prediction for quantitative traits under the epistatic model. *BMC Genetics*, 2011, **12**(1): 15

88. Lorenzana R E, Bernardo R. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theoretical and Applied Genetics*, 2009, **120**(1): 151–161

89. Jiang Y, Reif J C. Modeling epistasis in genomic selection. *Genetics*, 2015, **201**(2): 759–768

90. Coster A, Bastiaansen J W M, Calus M P L, van Arendonk J A M, Bovenhuis H. Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. *Genetics, Selection, Evolution*, 2010, **42**(1): 9

91. Zhong S, Dekkers J C M, Fernando R L, Jannink J L. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a Barley case study. *Genetics*, 2009, **182**(1): 355–364

92. Pszczola M, Strabel T, Mulder H A, Calus M P L. Reliability of direct genomic values for animals with different relationships within and to the reference population. *Journal of Dairy Science*, 2012, **95**(1): 389–400

93. Rincent R, Laloë D, Nicolas S, Altmann T, Brunel D, Revilla P, Rodríguez V M, Moreno-Gonzalez J, Melchinger A, Bauer E, Schoen C C, Meyer N, Giauffret C, Bauland C, Jamin P, Laborde J, Monod H, Flament P, Charcosset A, Moreau L. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). *Genetics*, 2012, **192**(2): 715–728

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

95. Clark S A, Hickey J M, Daetwyler H D, van der Werf J H J. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genetics, Selection, Evolution*, 2012, **44**(1): 4

96. García-Ruiz A, Cole J B, VanRaden P M, Wiggans G R, Ruiz-López F J, Van Tassell C P. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proceedings of the National Academy of Sciences of the United States of America*, 2016, **113** (28): E3995–E4004

97. Taylor J F, Taylor K H, Decker J E. Holsteins are the genomic selection poster cows. *Proceedings of the National Academy of Sciences of the United States of America*, 2016, **113**(28): 7690–7692

98. Romanos J, Rosén A, Kumar V, Trynka G, Franke L, Szperl A, Gutierrez-Achury J, van Diemen C C, Kanninga R, Jankipersadsing S A, Steck A, Eisenbarth G, van Heel D A, Cukrowska B, Bruno V, Mazzilli M C, Núñez C, Bilbao J R, Mearin M L, Barisani D, Rewers M, Norris J M, Ivarsson A, Boezen H M, Liu E, Wijmenga C, Prevent C D G. Improving coeliac disease risk prediction by testing non-HLA variants additional to HLA variants. *Gut*, 2014, **63**(3): 415–422

99. Abraham G, Tye-Din J A, Bhalala O G, Kowalczyk A, Zobel J, Inouye M. Accurate and robust genomic prediction of celiac disease using statistical learning. *PLoS Genetics*, 2014, **10**(2): e1004137

100. Wei Z, Wang K, Qu H Q, Zhang H, Bradfield J, Kim C, Frackleton E, Hou C, Glessner J T, Chiavacci R, Stanley C, Monos D, Grant S F, Polychronakos C, Hakonarson H. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genetics*, 2009, **5**(10): e1000678

101. Abraham G, Havulinna A S, Bhalala O G, Byars S G, De Livera A M, Yetukuri L, Tikkanen E, Perola M, Schunkert H, Sijbrands E J, Palotie A, Samani N J, Salomaa V, Ripatti S, Inouye M. Genomic prediction of coronary heart disease. *European Heart Journal*, 2016, **37**(43): 3267–3278

102. Vazquez A I, Veturi Y, Behring M, Shrestha S, Kirst M, Resende M F Jr, de Los Campos G. Increased proportion of variance explained and prediction accuracy of survival of breast cancer patients with use of whole-genome multiomic profiles. *Genetics*, 2016, **203**(3): 1425–1438

103. de Maturana E L, Chanok S J, Picornell A C, Rothman N, Herranz J, Calle M L, García-Closas M, Marenne G, Brand A, Tardón A, Carrato A, Silverman D T, Kogevinas M, Gianola D, Real F X, Malats N. Whole genome prediction of bladder cancer risk with the Bayesian LASSO. *Genetic Epidemiology*, 2014, **38**(5): 467–476

104. Vazquez A I, de los Campos G, Klimentidis Y C, Rosa G J, Gianola D, Yi N, Allison D B. A comprehensive genetic approach for improving prediction of skin cancer risk in humans. *Genetics*, 2012, **192**(4): 1493–1502

105. Visscher P M. Human complex trait genetics in the 21st century. *Genetics*, 2016, **202**(2): 377–379