

A fast genomic selection approach for large genomic data

Hailan Liu¹ · Guo-Bo Chen²

Received: 26 September 2016 / Accepted: 27 February 2017 / Published online: 7 April 2017
© Springer-Verlag Berlin Heidelberg 2017

Abstract

Key message We propose a novel computational method for genomic selection that combines identical-by-state (IBS)-based Haseman–Elston (HE) regression and best linear prediction (BLP), called HE-BLP.

Abstract Genomic best linear unbiased prediction (GBLUP) has been widely used in whole-genome prediction for breeding programs. To determine the total genetic variance of a training population, a linear mixed model (LMM) should be solved via restricted maximum likelihood (REML), whose computational complexity is the cube of the sample size. We proposed a novel computational method combining identical-by-state (IBS)-based Haseman–Elston (HE) regression and best linear prediction (BLP), called HE-BLP. With this method, the total genetic variance can be estimated by solving a simple HE linear regression, which has a computational complex of the sample size squared; therefore, it is suitable for large-scale genomic data, except those with which environmental effects need to be estimated simultaneously, because it does

not allow for this estimation. In Monte Carlo simulation studies, the estimated heritability based on HE was identical to that based on REML, and the prediction accuracy via HE-BLP and traditional GBLUP was also quite similar when quantitative trait loci (QTLs) were randomly distributed along the genome and their effects followed a normal distribution. In addition, the kernel row number (KRN) trait in a maize IBM population was used to evaluate the performance of the two methods; the results showed similar prediction accuracy of breeding values despite slightly different estimated heritability via HE and REML, probably due to the underlying genetic architecture. HE-BLP can be a future genomic selection method choice for even larger sets of genomic data in certain special cases where environmental effects can be ignored. The software for HE regression and the simulation program is available online in the **Genetic Analysis Repository (GEAR; <https://github.com/gc5k/GEAR/wiki>)**.

Introduction

Given various possible approaches, marker-assisted selection (MAS) and whole-genome selection are widely employed for improving agronomic traits of economic importance. Since it was proposed in the 1990s, MAS has been used for genetic improvement in crops such as rice, maize, and wheat (Lande and Thompson 1990; Jena et al. 2006; Xiao et al. 2012; Hao et al. 2014). However, MAS only transfers detected large-effect quantitative trait loci (QTL) to recipient varieties. It seems to be inadequate for improving many important agronomic traits, such as grain yield, plant height, kernel quality, and disease resistance (Jannink et al. 2010). This may be because QTL are not detected due to a lack of statistical

Communicated by Mikko J. Sillanpää.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-017-2887-3) contains supplementary material, which is available to authorized users.

✉ Hailan Liu
lhlzju@hotmail.com

✉ Guo-Bo Chen
chenguobo@gmail.com

¹ Maize Research Institute, Sichuan Agricultural University, Chengdu, Sichuan Province 611130, China

² Evergreen Landscape and Architecture Studio, Xixi Road 562, Hangzhou, Zhejiang Province 310026, China

power; consequently, no efficient markers are available for MAS.

Recent progress in molecular marker techniques and statistical methods has overcome these conventional MAS deficiencies. Meuwissen et al. (2001) pioneered the development of genomic selection (GS) technology, utilizing all quantitative trait loci information to predict breeding values on the basis of high-density markers, such as single nucleotide polymorphisms (SNPs). According to their simulation results, GS technology predicted genomic estimated breeding values (GEBV) with an accuracy of 0.85 in terms of the correlation between the true and predicted breeding values. Therefore, this promising technology may revolutionize both animal and plant genetic improvement and accelerate the breeding process (Heffner et al. 2009; Jannink et al. 2010). GS technology has been successfully implemented in many countries in species such as dairy cattle, maize, rice, and loblolly pine (Bernardo and Yu 2007; Hayes et al. 2009; Riedelsheimer et al. 2012; Resende et al. 2012; Xu et al. 2014; Spindel et al. 2015).

To date, many statistical prediction methods, including genomic best linear unbiased prediction (GBLUP), trait-specific relationship matrix best linear unbiased prediction (TABLUP), BayesA, BayesB, and least absolute shrinkage and selection operator (LASSO), have been developed to implement genomic prediction in animal and plant breeding (Meuwissen et al. 2001; Usai et al. 2009; Zhang et al. 2010; Xu et al. 2014). Among them, GBLUP is the most popular for its simple implementation and serves as the benchmark for methodological comparison. Under the assumption that all QTLs follow a normal distribution, GBLUP is very robust under a broad range of scenarios (Xu et al. 2014). In order to implement GBLUP, a restricted maximum likelihood (REML) algorithm must first be employed to estimate the total genetic variance. However, with the increasing amount of individuals and markers, the computational burden for solving a linear mixed model (LMM) will be very intensive, approximately in the order of N^3 , where N is the sample size.

Recently, identical-by-state (IBS)-based Haseman–Elston (HE) regression, an alternative for REML when environmental effects can be ignored, was proposed to estimate heritability via least squares (Chen 2014; Golan et al. 2014; Hu and Yang 2014). In this study, we developed a new GS method combining HE regression and best linear prediction (HE-BLP), and investigated the properties of HE-BLP on the prediction accuracy of genomic selection in simulation studies and in real data [i.e., the kernel row number (KRN) trait in the maize IBM population].

Materials and methods

The maize IBM population

Phenotype and genotype data for maize IBM were derived from a cross between B73 and Mo17 and were downloaded from Bommert et al. (2013) and MaizeGDB (http://www.maizegdb.org/data_center/qtl-data). The KRN trait, based on a total of 302 recombinant inbred lines (RILs), in this population was evaluated over 4 years (2000, 2001, 2002, and 2003) at the University of Illinois Research and Education Center. A total of 1339 molecular markers were genotyped. For more IBM population details please refer to Bommert et al. (2013).

Genetic models

Without loss of generality, assuming additive effects only and a standardized phenotype: $y_i = \frac{y'_i - \bar{y}}{\sigma_y}$, y'_i represents the raw phenotypic value; \bar{y} represents the mean value of the phenotypic values; and σ_y represents the standard error of the phenotypic values. In addition, when the target phenotype is to be standardized, its fixed effects should first be removed by regression. The linear model of a quantitative trait can be written as

$$\mathbf{y} = \mathbf{Z}\mathbf{u} + \mathbf{e} \quad (1)$$

in which \mathbf{y} is the vector for the phenotypic value of n individuals; \mathbf{Z} is the standardized genotype matrix of $n \times m$ dimensions; m is the number of markers; \mathbf{u} is the additive effects for the causal loci, following $N(0, \mathbf{I}\sigma_u^2)$; and \mathbf{e} is the residual following $N(0, \mathbf{I}\sigma_e^2)$. Let $\mathbf{g} = \mathbf{Z}\mathbf{u}$, be the vector for total genetic effects, and $g_i \sim N(0, \sigma_g^2 = m\sigma_u^2)$. Then, $\text{var}(\mathbf{y}) = \mathbf{\Omega}\sigma_g^2 + \mathbf{I}\sigma_e^2$, in which $\mathbf{\Omega} = \frac{\mathbf{Z}\mathbf{Z}'}{m}$ is the genetic relationship matrix; for a pair of individuals i and j , $\Omega_{ij} = \frac{Z_i Z_j'}{m}$. In this study, we assumed every marker was causal QTL.

We used HE to estimate total additive genetic variance (σ_g^2) in the training population (Chen 2014)

$$\mathbf{Y} = b_0 + b_1 \boldsymbol{\omega} + \boldsymbol{\varepsilon} \quad (2)$$

in which $\mathbf{Y} = (y_i - y_j)^2$, a $1 \times \mathcal{N}$ vector $\left[\mathcal{N} = n \times \frac{n-1}{2} \approx \frac{n^2}{2} \right]$,

is the squared difference of the phenotypic values between a pair of individuals from the training population. $\boldsymbol{\omega}$, a $1 \times \mathcal{N}$ vector, is the genetic relatedness between a pair of individuals i and j ($\omega_{ij} = \Omega_{ij}$; $\boldsymbol{\omega}$ takes the off-diagonal elements of $\mathbf{\Omega}$); $\boldsymbol{\varepsilon}$ is the residual error; b_0 is the intercept; and b_1 is the regression coefficient. The least-squares estimate for the regression coefficient is $b_1 = \frac{\text{cov}(\mathbf{Y}, \boldsymbol{\omega})}{\text{var}(\boldsymbol{\omega})}$. For various experimental popula-

tions used in breeding programs, such as a backcross population (BC), F2, a recombinant inbred line (RIL), and double

haploid (DH), and when environmental effects can be ignored, the HE regression provides an unbiased estimate of the additive variance component and the narrow-sense heritability h^2 . For RIL, DH, and BC, $E(b_1)_{RIL} = E(b_1)_{DH} = E(b_1)_{BC} = -\rho_{kl}^2 h_l^2$, and for F2, $E(b_1)_{F2} = -2\rho_{kl}^2 h_l^2$, in which h_l^2 is the heritability of the l th QTL and ρ_{kl}^2 is the squared correlation linkage disequilibrium between the k th marker and the l th QTL (see the **Supplementary Notes** for its detailed mathematical derivations; a more general derivation can be found in Chen 2014, 2016). In the original HE regression, the regression coefficient is $(1 - 2c_{kl})^2 h_l^2$, in which c_{kl} is the linkage recombination fraction between the k th marker and the l th QTL. Therefore, under the context of this association, which uses IBS information, the linkage disequilibrium (LD) metric replaces the recombination in the original HE regression.

REML and least squares use different statistical mechanisms to estimate the parameter of interest; additive genetic variance and their equivalence can be established when the QTLs are randomly distributed along the genome (Chen 2016). We briefly discussed this in the Appendix.

Best linear prediction (BLP)

Given the estimated genetic variance, $\hat{\sigma}_g^2$, BLP was subsequently used to predict the breeding value of each line of the candidate population. Because environmental effects are ignored, or possibly pre-corrected, the breeding values are not BLUP, but rather BLP:

$$\hat{g}_2 = \hat{\sigma}_g^2 \Omega_{21} V^{-1} y_1 \quad (3)$$

in which \hat{g}_2 is the predicted breeding values in the future population; y_1 is the phenotypic values in the training population, as used in Eq. 1; Ω_{21} is the genetic relationship matrix between the candidate population and the training population; $\hat{\sigma}_g^2$ is estimated from HE; the inverse of the V matrix is computed using $V^{-1} = (\hat{\sigma}_g^2 \Omega_{11} + \hat{\sigma}_e^2 I)^{-1}$ where the Ω_{11} is the genetic relationship matrix for the training population.

In practice, it is not always possible to construct Ω_{12} if a future generation has not been genotyped. Alternatively, it is practical to use the equation $\hat{g}_2 = Z_2 \hat{u}$, in which Z_2 is a genotypic matrix of the future population and $\hat{u} = \hat{\sigma}_g^2 z_1' v^{-1} y_1 / m$ is the BLP estimation for the markers in the training population.

Results

Monte Carlo simulation studies

To compare the predictability of the proposed HE-BLP and the traditional GBLUP, we simulated four kinds of

experimental populations, including BC, DH, F2, and RIL, in which 1,000 equally frequent biallelic markers were evenly spaced [the recombination rate was c between the i th and the $(i + 1)$ th markers]. These 1000 markers were defined as QTLs, having their effects sampled from a normal distribution. Each simulated scenario included 100 replications.

First, we evaluated the performance of HE and REML in estimating heritability for the simulated data under a fixed population size ($n = 500$). The simulated parameters were set at $h^2 = 0.25$ or 0.5 , $c = 0.1$ or 0.2 , giving four generic scenarios for each population. As the QTLs were randomly distributed along the genome, equivalence between HE and REML was satisfied. For example, when $c = 0.1$ and $h^2 = 0.25$, $\hat{h}^2 = 0.252 \pm 0.063$ and $\hat{h}^2 = 0.252 \pm 0.055$ in the BC population via HE and REML, respectively, which were unbiased estimates. When $c = 0.2$ and $h^2 = 0.25$, $\hat{h}^2 = 0.240 \pm 0.071$ and $\hat{h}^2 = 0.243 \pm 0.062$ in the BC population via HE and REML, respectively (Fig. 1). Under the other three experimental populations, unbiased heritability was estimated under all simulated scenarios. The standard error of the estimated heritability from HE was higher than that generated by REML (Fig. 1).

Furthermore, we assessed the prediction accuracy of HE-BLP and GBLUP. The prediction accuracy was measured as the squared correlation coefficient (r^2) between the predicted breeding values and the simulated phenotypes. The parameters of the simulation were as follows: the population size of the training data was 5,000, and the candidate population size was 100 (Fig. 2). As expected, HE-BLP and the traditional GBLUP showed similar prediction accuracies. For example, when $c = 0.1$ and $h^2 = 0.25$, r^2 was 0.2074 ± 0.058 and 0.2075 ± 0.058 in the BC population via HE-BLP and GBLUP, respectively. In addition, owing to its continuous multi-generation selfing, the RIL population has decayed LD [for RIL, the LD between marker i and $i + 1$ is $\frac{1}{1+2c}$ for the coupling phases (haplotype AB or ab), and $\frac{2c}{1+2c}$ for the repulsion phase (haplotype Ab or aB); c is the recombination fraction between a pair of loci] than BC, DH, and F2 populations; therefore, its prediction accuracy was slightly lower than that of the other three populations (Fig. 2).

We also investigated the performance of HE-BLP under a much smaller, but perhaps more realistic, sample size because plant breeders often have a couple hundred elite/germplasm lines (Spindel et al. 2016; Yu et al. 2016; Riedelsheimer et al. 2012). The population size of the training data was set at 500, and the candidate population was 100. In the BC population, when $c = 0.1$ and $h^2 = 0.25$, the prediction accuracy was 0.086; and it increased to 0.251 given $c = 0.1$ and $h^2 = 0.25$ (Fig. 3a). When $c = 0.2$ and $h^2 = 0.5$, the prediction accuracy was 0.188 in the BC population

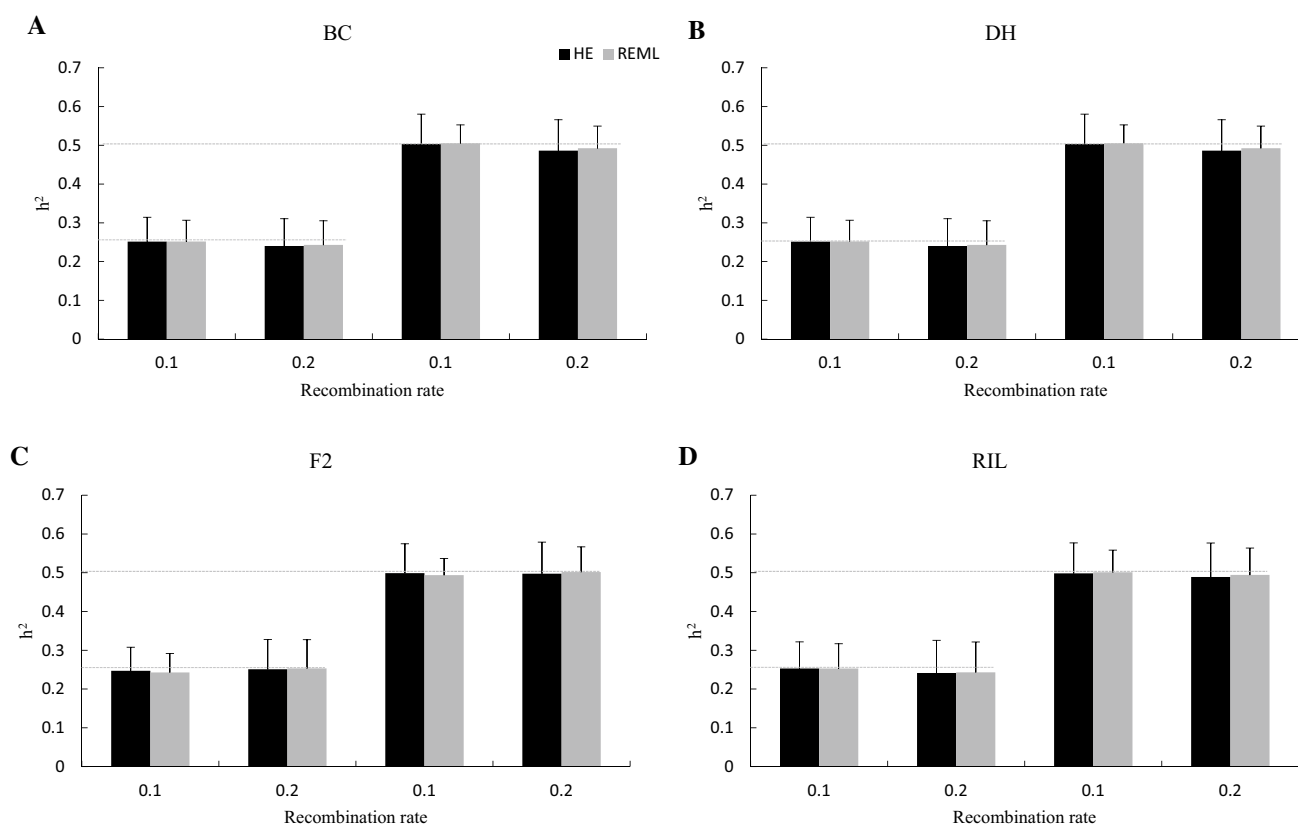


Fig. 1 The estimated heritability based on a fixed sample size (500) using Haseman–Elston (HE) and restricted maximum likelihood (REML) methods in 100 simulations when heritability (h^2) was set at 0.25 and 0.5, respectively, and recombination rates (c) were set as

0.1 and 0.2, respectively. **a** Backcross (BC) population; **b** double haploid (DH) population; **c** F2 population; and **d** recombinant inbred line (RIL) population. The dotted lines represent the theoretical heritability. The vertical bar indicates the standard error for 100 simulations

(Fig. 3b), which indicated that the predictability decreased with the increase of the recombination rate, which led to reduced LD. Moreover, when the size of the training population was increased to 1000, the prediction accuracy was 0.338 in the BC population given $h^2 = 0.5$ and $c = 0.1$. This indicated that the predictability was proportional to the population size (Fig. 4). In addition, the predicted breeding values from HE-BLP and GBLUP for the F2 population (training population size=500, candidate population size=100, $c = 0.1$ and $h^2 = 0.5$) was highly correlated ($r^2 = 0.996 \pm 0.0053$) based on 100 simulations.

Reduced computational cost in estimating the additive variance

In HE, the least-squares estimate for the regression coefficient is $b_1 = \text{cov}(Y, \omega) / \text{var}(\omega)$. The computational complexity of matrix algebraic operations is well documented (https://en.wikipedia.org/wiki/Computational_complexity_of_mathematical_operations). The operations for covariance/variance are related to multiplication of Y , a $1 \times \mathcal{N}$ vector, to ω' , an $\mathcal{N} \times 1$ vector, having a

computational complexity of $\mathcal{N} \approx n^2/2$. In contrast, for REML, the computational complexity is at least proportional to tn^3 , in which t is the rounds of iterations upon the stop rule adopted for the algorithm. We evaluated the computation time of estimating heritability via HE and REML for a simulated F2 population, which had population sizes of 5000, 10,000, and 20,000 ($h^2 = 0.5$ and marker number=1000). To complete the estimation of the additive variance component for 100 replications of the simulated populations, HE took about 1, 6, and 25 min, respectively, to complete the estimation of the additive variance components, whereas REML took 1.5 min, 16.33 h, and 125 h, respectively. HE had significantly reduced computational cost. Therefore, HE had a computational advantage over REML when the estimation of environmental effects is ignored.

Prediction of the KRN in the maize IBM population

We utilized a maize IBM population to evaluate the predictability for the KRN trait in different years with HE-BLP

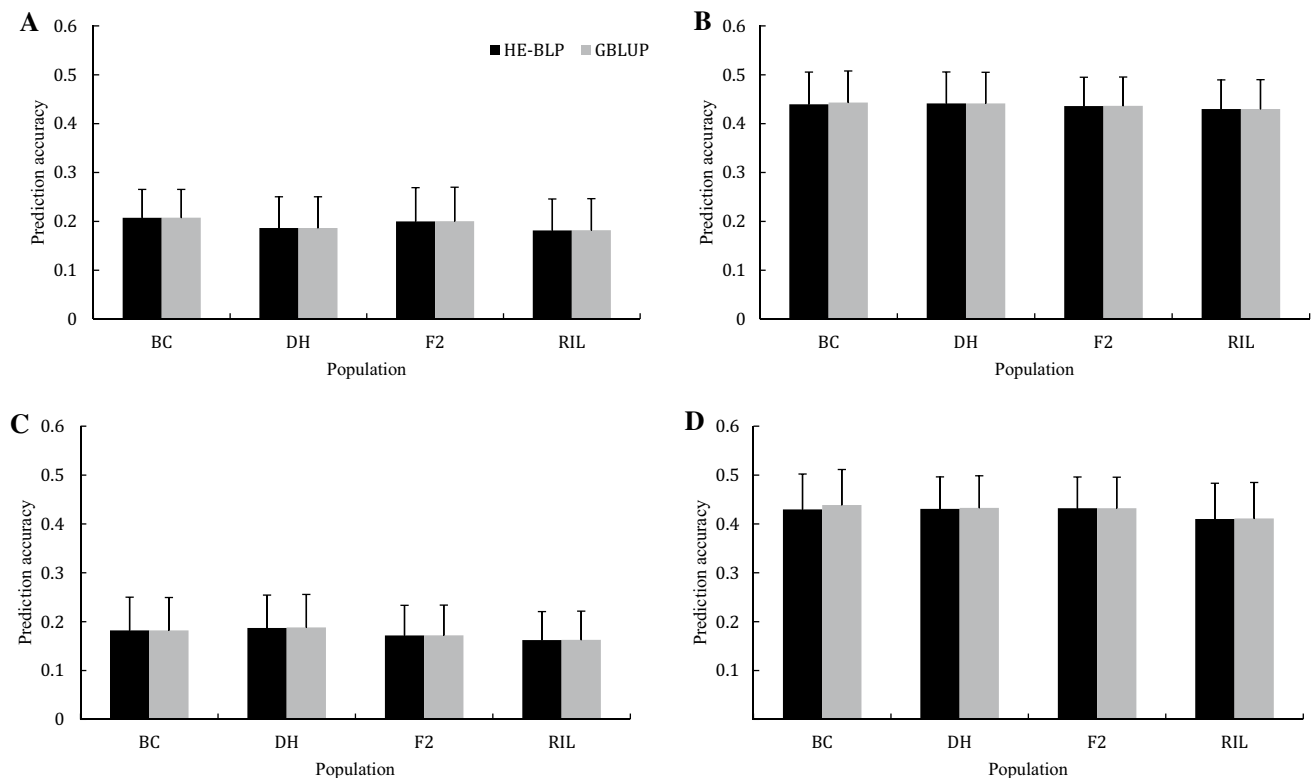


Fig. 2 Predictability based on a training population with a fixed sample size (5,000) and a candidate population with a fixed sample size (100) using Haseman–Elston regression and best linear prediction (HE-BLP) and genomic best linear unbiased prediction (GBLUP) methods in 100 simulations. **a** When $h^2 = 0.25$ and $c = 0.1$; **b** when

$h^2 = 0.5$ and $c = 0.1$; **c** when $h^2 = 0.25$ and $c = 0.2$; and **d** when $h^2 = 0.5$ and $c = 0.2$. Predictability was measured as the squared correlation coefficient (r^2) between the phenotypes and the predicted breeding values

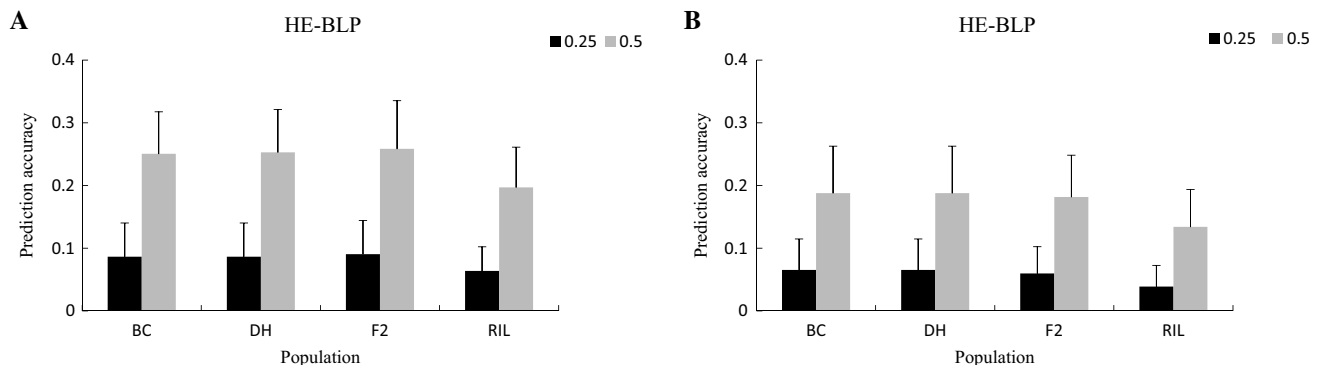


Fig. 3 Predictability based on a training population with a fixed sample size (500) and a candidate population with a fixed sample size (100) using the HE-BLP method in 100 simulations when h^2 was 0.25

and 0.5, and c was 0.1 and 0.2. **a** When $c = 0.1$ and $h^2 = 0.25$ and 0.5; and **b** when $c = 0.2$ and $h^2 = 0.25$ and 0.5. Predictability was measured as the r^2 between phenotypes and predicted breeding values

and traditional GBLUP methods. The maize IBM population of 1 year was used as a training population to predict the breeding values of the KRN in the other 3 years.

First, we estimated the heritability of KRN every year, and found that the estimated heritability based on HE

varied from 0.411 to 0.553; those based on REML varied from 0.302 to 0.513 (Fig. 5a). To compare the prediction accuracy of HE-BLP and traditional GBLUP, the maize IBM population of each year was randomly split into a training population and a candidate population. We

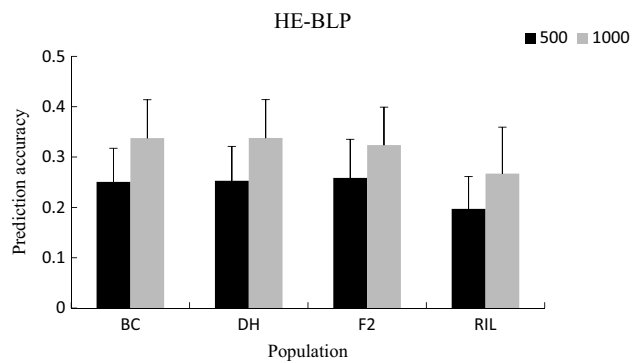


Fig. 4 Predictability based on the training population sizes (500 and 1000) and a candidate population with the fixed sample size (100) using the HE-BLP method in 100 simulations when $c = 0.1$ and $h^2 = 0.5$. Predictability was measured as the r^2 between the phenotypes and the predicted breeding values

compared the accuracy of predicting the KRN with HE-BLP and traditional GBLUP, and found that the predictability of the two methods was similar. The prediction accuracies were 0.168 ± 0.054 and 0.165 ± 0.055 in the year

2000 via HE-BLP and GBLUP, respectively, 0.089 ± 0.041 and 0.089 ± 0.039 in the year 2001, 0.115 ± 0.038 and 0.115 ± 0.037 in the year 2002, and 0.123 ± 0.036 and 0.121 ± 0.036 in the year 2003 (Fig. 5b). Moreover, data of one of the four years were utilized to predict the phenotypic values of the other three years, and the result indicated that the predictability of the two methods was similar. For example, when data from the year 2000 were set as the training population and that from the year 2001 were as the candidate population, the prediction accuracy of the two methods was 0.281 (Table 1).

Discussion

Among various genomic selection implementations, a combination of REML and BLUP is considered the benchmark. However, as large sample sizes are emerging for many species (e.g., ~1 million records for dairy cattle), reducing computational burden in GBLUP becomes necessary. A close scrutiny of the current GBLUP finds actually two steps involved: (1) estimation of the variance component,

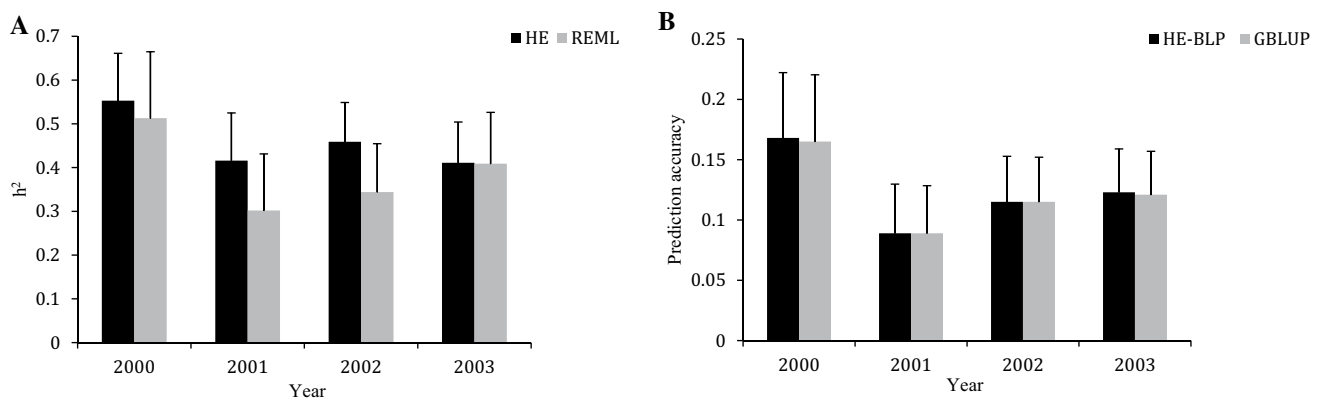


Fig. 5 Comparison between HE-BLP and GBLUP based on the kernel row number (KRN) trait of the maize IBM population. **a** Estimated heritability using the HE and REML methods for the KRN trait in the maize IBM population. **b** Predictability of 100 replicates via HE-BLP and GBLUP for the KRN of 4 years based on the maize

IBM population when the training sample size was set at 100 and the remaining sample was set as the candidate population. Predictability was measured as the r^2 between the phenotypes and the predicted breeding values

Table 1 Predictability based on maize kernel row number (KRN) from different years

Training	Candidate							
	2000		2001		2002		2003	
	HE-BLP	GBLUP	HE-BLP	GBLUP	HE-BLP	GBLUP	HE-BLP	GBLUP
2000			0.281	0.281	0.358	0.358	0.321	0.320
2001	0.307	0.309			0.314	0.311	0.295	0.290
2002	0.448	0.432	0.385	0.370			0.534	0.503
2003	0.399	0.399	0.357	0.357	0.510	0.510		

Predictability was measured as the squared correlation coefficient (r^2) between the phenotypes and the predicted breeding values

and (2) BLUP prediction for breeding values. In traditional implementation, both steps involve inversion of the matrix, an operation that is the sample size cubed. In order to reduce computing cost, the proven and young (APY) algorithm was applied to genomic selection studies (Misztal 2016). In the APY, only part of the dataset (core individuals in the sample) is sampled for inversion; then the inversion of the whole data matrix is leveraged by Cholesky decomposition (Henderson 1976). For example, if k and the $1 - k$ proportion of the sample are core and non-core individuals in the whole dataset, respectively, APY can reduce the computational cost to $(kn)^3$, compared with n^3 , using all samples directly.

In this study, we reported the development of HE-BLP, which combines IBS-based HE regression and BLP. HE is a simple linear regression model, and the least square analysis can be used to compute the total genetic variance with a cost of n^2 , compared with a cost of n^3 , using REML. The results based on simulation studies and real data indicated that the predictions were similar between HE-BLP and traditional GBLUP; therefore, its simplicity and efficiency is advantageous for harnessing large-scale genomic data. However, HE-BLP is a simplification that ignores environmental effects and therefore is mostly appropriate for use in single-environmental trials, and possibly in completely balanced multi-environmental trials. Of note, Ritland proposed to estimate heritability in natural populations by regressing the cross-product of a pair of individuals to their identical-by-descent (IBD), which is estimated from genetic markers (Ritland 1996, 2000). His method is closely related to the logic underlying HE regression (Sillanpää 2011), while our HE is directly based on IBS, which is most commonly used in genome-wide association studies.

Interestingly, regardless of the differences in estimated heritability, as shown for the KRN in the IBM population, HE-BLP and GBLUP demonstrated very similar prediction accuracies. The equivalence between HE and REML held under neutral genetic architecture, under which QTL were randomly allocated along the genome. However, as demonstrated in the IBM data, there was a slight difference between the estimates of heritability from HE and REML. The reason may be two-fold: first, the genetic architecture of the KRN in the IBM population is not neutral; second, there is a subtle difference in the statistical mechanism between HE and REML. It is documented that a selected/ascertained population could disturb the variance component estimation (Van der Werf and de Boer 1990). In this study, we did not consider selection that can disturb genetic architecture. Under the framework of maximum likelihood, such as REML, it is unclear how to predict the bias. However, under HE, it has been proved that the bias could be predicted as long as the genetic architecture is known (Chen 2016). Although whether there is an analytical solution

is not of primary interest for a direct breeding program, it will help to elucidate how fundamental genetic architecture facilitates breeding in the future. In addition, for broader application of HE-BLP, we will incorporate non-additive effects, including dominance and epistasis, into this model in future work.

Appendix: Conditions for the equivalence between REML and HE for the estimation of heritability

The additive genetic variance is defined as (Lynch and Walsh 1998, page 102)

$$\sigma_g^2 = \mathbf{u}'\mathbf{M}\mathbf{u}$$

$$\text{in which } \mathbf{M} = \frac{\mathbf{Z}'\mathbf{Z}}{m} = \begin{pmatrix} \rho_1^2 & \rho_{1,2} & \rho_{1,3} & \cdots & \rho_{1,m} \\ \rho_{2,1} & \rho_2^2 & \rho_{2,3} & \cdots & \rho_{2,m} \\ \rho_{3,1} & \rho_{3,2} & \rho_3^2 & \cdots & \rho_{3,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{m,1} & \rho_{m,2} & \rho_{m,3} & \cdots & \rho_m^2 \end{pmatrix}, \text{ the vari-}$$

ance-covariance matrix for QTLs. Alternatively written, $\sigma_g^2 = \sum_{i=1}^m u_i^2 + \sum_{j \neq k} \sum_{k=1}^m \rho_{j,k} u_j u_k$, the first summation for intra-locus variance and the second summation for inter-locus covariance.

In contrast under HE (Chen 2014, 2016), for standardized phenotype,

$$E(\sigma_{g|HE}^2) = -\frac{b_1}{2} = m \frac{\mathbf{u}'\tilde{\mathbf{M}}\mathbf{u}}{\mathbf{1}'\mathbf{M}\mathbf{1}}$$

$$\text{in which } \mathbf{1} \text{ is a } 1 \times m \text{ vector of } 1. \tilde{\mathbf{M}} = \sum_{i=1}^m m_i, \text{ in which } m_i = \begin{pmatrix} \rho_{1,i}^2 & \rho_{1,i}\rho_{2,i} & \rho_{1,i}\rho_{3,i} & \cdots & \rho_{1,i}\rho_{m,i} \\ \rho_{2,i}\rho_{1,i} & \rho_{2,i}^2 & \rho_{2,i}\rho_{3,i} & \cdots & \rho_{2,i}\rho_{m,i} \\ \rho_{3,i}\rho_{1,i} & \rho_{3,i}\rho_{2,i} & \rho_{3,i}^2 & \cdots & \rho_{3,i}\rho_{m,i} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{m,i}\rho_{1,i} & \rho_{m,i}\rho_{2,i} & \rho_{m,i}\rho_{3,i} & \cdots & \rho_{m,i}^2 \end{pmatrix} - \text{the vari-}$$

ance-covariance matrix for all pair of QTLs, including to itself, but via the i th marker.

When $\sum_{j \neq k} \sum_{k=1}^m \rho_{j,k} u_j u_k = 0$, HE and REML are equivalent in estimating the additive variance. $\sum_{j \neq k} \sum_{k=1}^m \rho_{j,k} u_j u_k = 0$ can be interpreted that QTLs are randomly allocated along the genome; in opposite, QTLs of similar effects are clustered together, leading to $\sum_{j \neq k} \sum_{k=1}^m \rho_{j,k} u_j u_k \neq 0$.

In a simplified example, if QTLs are in linkage equilibrium between each other, \mathbf{m}_i becomes a sparse matrix with only the i th element at the diagonal is 1 and zero everywhere. $\tilde{\mathbf{M}} = \sum_{i=1}^m m_i = \mathbf{I}$, then the numerator $\mathbf{m}\mathbf{u}'\tilde{\mathbf{M}}\mathbf{u} = m^2\sigma_u^2$. In the denominator, $\mathbf{M} = \mathbf{I}$ and $\mathbf{1}'\mathbf{M}\mathbf{1} = m$. $\hat{\sigma}_{g|HE}^2 = m\sigma_u^2 = \sigma_g^2$. However, discrepancy

between HE and REML may occur when the QTLs of the same effect direction are clustered together, and see more detailed discussion in Chen's study (2016).

Author contribution statement HL and GBC conceived and performed the study as well as wrote the manuscript.

Acknowledgements The authors are grateful to the editor and the two anonymous reviewers for their constructive comments, and Peter Bommert for providing information for the maize IBM population.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

References

- Bernardo R, Yu JM (2007) Prospects for genome wide selection for quantitative traits in maize. *Crop Sci* 47:1082–1090
- Bommert P, Nagasawa NS, Jackson D (2013) Quantitative variation in maize kernel row number is controlled by the FASCIATED EAR2 locus. *Nat Genet* 45:334–337
- Chen GB (2014) Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman-Elston regression. *Front Genet* 5:107
- Chen GB (2016) On the reconciliation of missing heritability for genome-wide association studies. *Eur J Hum Genet* 24:1810–1816
- Golan D, Lander ES, Rosset S (2014) Measuring missing heritability: Inferring the contribution of common variants. *Proc Natl Acad Sci U S A* 111:E5272–E5281
- Hao XM, Li XW, Yang XH, Li JS (2014) Transferring a major QTL for oil content using marker-assisted backcrossing into an elite hybrid to increase the oil content in maize. *Mol Breeding* 34:739–748
- Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: Genomic selection in dairy cattle: Progress and challenges. *J Dairy Sci* 92:433–443
- Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. *Crop Sci* 49:1–12
- Henderson CR (1976) A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* 32:69–83
- Hu Z, Yang RC (2014) Marker-based estimation of genetic parameters in genomics. *PLoS ONE* 9:e102715
- Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. *Brief Funct Genom* 9:166–177
- Jena KK, Jeung JU, Lee JH, Choi HC, Brar DS (2006) High-resolution mapping of a new brown planthopper (BPH) resistance gene, Bph18(t), and marker-assisted selection for BPH resistance in rice (*Oryza sativa* L.). *Theor Appl Genet* 112:288–297
- Lande R, Thompson (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124:743–756
- Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland
- Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
- Misztal I (2016) Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. *Genetics* 202:401–409
- Resende MF Jr, Munoz P, Resende MD, Garrick DJ, Fernando RL, Davis JM, Jokela EJ, Martin TA, Peter GF, Kirst M (2012) Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). *Genetics* 190:1503–1510
- Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisek J, Technow F, Sulpice R, Altmann T, Stitt M, Willmitzer L, Melchinger AE (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat Genet* 44:217–220
- Ritland K (1996) A marker-based method for inference about quantitative inheritance in natural population. *Evol Int J org Evol* 50:1062–1073
- Ritland K (2000) Marker-inferred relatedness as a tool for detecting heritability in nature. *Mol Ecol* 9:1195–1204
- Sillanpää MJ (2011) On statistical methods for estimating heritability in wild populations. *Mol Ecol* 20:1324–1332
- Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redoña E, Atlin G, Jannink JL, McCouch SR (2015) Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS. Genet* 11:e1004982
- Spindel JE, Begum H, Akdemir D, Collard B, Redona E, Jannink J-L, McCouch S (2016) Genome-wide prediction models that incorporate *de novo* GWAS are a powerful new tool for tropical rice improvement. *Heredity* 116:395–408
- Usai MG, Goddard ME, Hayes BJ (2009) LASSO with cross-validation for genomic selection. *Genet Res* 91:427–436
- Van der Werf JHJ, de Boer IJM (1990) Estimation of additive genetic variance when base populations are selected. *J Anim Sci* 68:3124–3132
- Xiao SH, Zhang HP, You GX, Zhang XY, Yan CS, Chen X (2012) Integration of marker-assisted selection for resistance to pre-harvest sprouting with selection for grain-filling rate in breeding of white-kernelled wheat for the Chinese environment. *Euphytica* 188:85–88
- Xu S, Zhu D, Zhang Q (2014) Predicting hybrid performance in rice using genomic best linear unbiased prediction. *Proc Natl Acad Sci U S A* 111:12456–12461
- Yu X, Li X, Guo T, Zhu C, Wu Y, Mitchell SE, Roozeboom KL, Wang D, Wang ML, Pederson GA, Tesso TT, Schnable PS, Bernardo R, Yu J (2016) Genomic prediction contributing to a promising global strategy to turbocharge gene banks. *Nat Plants*. doi:10.1038/NPLANTS.2016.150
- Zhang Z, Liu J, Ding X, Bijma P, de Koning D-J, Zhang Q (2010) Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PLoS ONE* 5:e12648