



Contents lists available at ScienceDirect

The Crop Journal

journal homepage: www.keaipublishing.com/en/journals/the-crop-journal/

A novel genomic prediction method combining randomized Haseman-Elston regression with a modified algorithm for Proven and Young for large genomic data

Hailan Liu ^{a,*}, Guo-Bo Chen ^{b,c,*}^a Maize Research Institute, Sichuan Agricultural University, Chengdu 611130, Sichuan, China^b Phase I Clinical Research Institute, Zhejiang Provincial People's Hospital, Affiliated People's Hospital, Hangzhou Medical College, Hangzhou 310014, Zhejiang, China^c Key Laboratory of Endocrine Gland Diseases of Zhejiang Province, Hangzhou 310014, Zhejiang, China

ARTICLE INFO

Article history:

Received 2 March 2021

Revised 13 September 2021

Accepted 16 September 2021

Available online 23 October 2021

Keywords:

Genomic prediction

GBLUP

Randomized HE-regression

Modified algorithm for Proven and Young

ABSTRACT

Computational efficiency has become a key issue in genomic prediction (GP) owing to the massive historical datasets accumulated. We developed hereby a new super-fast GP approach (SHEAPY) combining randomized Haseman-Elston regression (RHE-reg) with a modified Algorithm for Proven and Young (APY) in an additive-effect model, using the former to estimate heritability and then the latter to invert a large genomic relationship matrix for best linear prediction. In simulation results with varied sizes of training population, GBLUP, HEAPY|A and SHEAPY showed similar predictive performance when the size of a core population was half that of a large training population and the heritability was a fixed value, and the computational speed of SHEAPY was faster than that of GBLUP and HEAPY|A. In simulation results with varied heritability, SHEAPY showed better predictive ability than GBLUP in all cases and than HEAPY|A in most cases when the size of a core population was 4/5 that of a small training population and the training population size was a fixed value. As a proof of concept, SHEAPY was applied to the analysis of two real datasets. In an *Arabidopsis thaliana* F₂ population, the predictive performance of SHEAPY was similar to or better than that of GBLUP and HEAPY|A in most cases when the size of a core population (200) was 2/3 of that of a small training population (300). In a sorghum multiparental population, SHEAPY showed higher predictive accuracy than HEAPY|A for all of three traits, and than GBLUP for two traits. SHEAPY may become the GP method of choice for large-scale genomic data.

© 2021 Crop Science Society of China and Institute of Crop Science, CAAS. Production and hosting by Elsevier B.V. on behalf of KeAi Communications Co., Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Application of conventional marker-assisted selection (MAS) to animal and plant breeding has improved trial-and-error breeding based on phenotypic selection [1–4], but it is valid only for qualitative traits and quantitative traits controlled by large-effect quantitative-trait loci (QTL) and fails for polygenic traits controlled by many small-effect QTL [5,6]. With the development of high-throughput genotyping technology and statistical models, a technology called genomic prediction (GP) has been proposed and successfully applied to the improvement of quantitative traits [7–10].

In recent years, a variety of GP methods including parametric methods (such as GBLUP, BayesA, and LASSO) and nonparametric

methods such as support vector machine (SVM), reproducing kernel Hilbert space (RKHS), and random forest (RF) have been developed [11]. GBLUP is a benchmark method whose computational process consists of two steps: (I) estimating heritability via restricted maximum likelihood (REML) and (II) computing genomic estimated breeding values (GEBV) by best linear unbiased prediction (BLUP). Both steps of GBLUP involve inverting a large genomic relationship matrix. To increase the predictive accuracy of GP, a very effective strategy is to increase training population sizes, but it is difficult or impracticable for conventional GBLUP to invert a genomic relationship matrix (GRM) for a large-sized training population [12]. For example, GBLUP failed to handle a massive set of greater than 3 million genotypes of U.S. Holstein cattle [13]. The challenge is maintaining both predictive accuracy and computational efficiency in calculating with large datasets. Some efficient computational strategies have been developed in response to this challenge. (I) Miszta et al. [14] developed the Algorithm for

* Corresponding authors.

E-mail addresses: chenguobo@gmail.com (G.-B. Chen), hlzju@hotmail.com (H. Liu).

Proven and Young (APY) to invert a GRM of millions of individuals. It greatly improves computational efficiency by obtaining the inverse of the GRM of a large training population using a small one randomly selected from it: the computational time of the inverse of GRM is only cubic in the size of the small population. Single-step GBLUP (ssGBLUP) based on APY has successfully performed GP of over seven million U.S. Holstein cattle [15,16]. (II) Liu and Chen [17,18] used Haseman-Elston (HE) regression based on genome-wide identity by state (IBS) to estimate heritability. The computational cost of HE is quadratic and that of REML is cubic in the training population size. Recently, Liu and Chen [12] developed a rapid genomic prediction method (HEAPY|A) combining the HE model with APY in an additive-effect model.

We present a super-fast GP method (SHEAPY) combining a modified APY with a scalable randomized algorithm for HE regression (RHE-reg), and evaluate its predictive accuracy and computing time.

2. Materials and methods

2.1. Genetic model

The basic model focuses only on additive effects, and is described as

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

in which \mathbf{y} is an $n \times 1$ vector of the standardized phenotype of interest; μ is the overall mean of the model; $\mathbf{1}$ is an $n \times 1$ vector of ones; \mathbf{Z} is an $n \times m$ matrix of standardized genotype coding values; \mathbf{u} is the $m \times 1$ vector of the additive effects following $N(0, \mathbf{I}\sigma_u^2)$; and \mathbf{e} is an $n \times 1$ vector of residuals error following $N(0, \sigma_e^2)$.

2.2. Estimating heritability via randomized HE-reg

IBS-based RHE-reg is used to estimate heritability [19] as follows:

$$y_i y_j = b_0 + b_1 \mathbf{G}_{ij} + e_{ij} \quad (2)$$

in which y_i and y_j denote the phenotypes of individuals i and j ($i, j = 1 \dots n$, and $i \neq j$), b_0 is the intercept, b_1 is the regression coefficient, \mathbf{G}_{ij} is the genetic relatedness ($\mathbf{G}_{ij} = \frac{\mathbf{Z}_i \mathbf{Z}_j^T}{m}$) between individuals i and j , and e_{ij} is residual error following $N(0, \sigma_e^2)$. For a trait, its additive genetic variance is $\hat{\sigma}_g^2 = \hat{b}_1$, and its $\hat{\sigma}_e^2 = \hat{\sigma}_y^2 - \hat{\sigma}_g^2$. After some matrix algebra and numerical optimization, ordinary least squares is used to estimate $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$. The computational equation is described as

$$\begin{bmatrix} \text{tr}[\mathbf{G}^2] & \text{tr}[\mathbf{G}] \\ \text{tr}[\mathbf{G}] & n \end{bmatrix} \begin{bmatrix} \hat{\sigma}_g^2 \\ \hat{\sigma}_e^2 \end{bmatrix} = \begin{bmatrix} \mathbf{y}^T \mathbf{G} \mathbf{y} \\ \mathbf{y}^T \mathbf{y} \end{bmatrix} \quad (3)$$

in which $\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}^T}{m}$ corresponds to the genetic relationship matrix between individuals and $\text{tr}[\mathbf{G}] = n$. In Eq. (3), $\mathbf{G}^2 = \mathbf{G}\mathbf{G}^T$ can be computationally expensive if n and m are large. To increase computational efficiency, the property of the trace operation of a matrix is used to estimate $\text{tr}[\mathbf{G}^2]$ with the equation $\text{tr}[\mathbf{G}^2] = \sum_{i,j} \mathbf{G}_{ij}^2 = E(\mathbf{w}_s^T \frac{\mathbf{Z}\mathbf{Z}^T}{m} \mathbf{w}_s) = E(\frac{\mathbf{Z}^T \mathbf{Z}}{m} \mathbf{w}_s \mathbf{w}_s^T)$, in which each entry of \mathbf{w}_s comes from a standard normal distribution $N(0, 1)$. Because in comparison with the computational cost of direct calculation of

\mathbf{G}^2 , which is $\mathcal{O}(n^2 m)$, $\mathbf{w}_s^T \frac{\mathbf{Z}\mathbf{Z}^T}{m}$ can be calculated very quickly, this formulation alleviates the computational burden. $\text{tr}[\mathbf{G}^2]$ is calculated via a randomized estimation [20]. The equation is as follows:

$$\text{tr}[\mathbf{G}^2] = \frac{1}{S} \frac{1}{m^2} \sum_{s=1}^S (\mathbf{w}_s^T \mathbf{Z}\mathbf{Z}^T) (\mathbf{Z}\mathbf{Z}^T \mathbf{w}_s) \quad (4)$$

Here S represents the rounds of randomization implemented, and was set as 5 throughout the study. The sampling variance of the randomized estimator in Eq. (4) is $\frac{1}{S} \text{tr}(\mathbf{G}^4)$ according to matrix algebra. Under Eq. (3), it is straightforward to see that $\hat{\sigma}_g^2 = \frac{\mathbf{y}^T (\mathbf{G} - \mathbf{I}) \mathbf{y}}{\text{tr}(\mathbf{G}^2) - n}$,

in which $\text{tr}(\mathbf{G}^2)$ is a plug-in estimate from Eq. (4). By Taylor expansion, it can be proven that $E(\hat{\sigma}_g^2) = \sigma_g^2 + \frac{1}{S} \frac{\text{tr}(\mathbf{G}^4)}{[\text{tr}(\mathbf{G}^2) - n]^2} \sigma_g^2$, and the bias vanishes when S is large enough. In particular, $E[\text{tr}(\mathbf{G}^2)] = \frac{n^2}{m_e} + n$, in which m_e is the effective number of markers for the population [18], and the bias is upon the realized value of $\frac{\sigma_g^2}{S} \left\{ \frac{\text{tr}(\mathbf{G}^4)}{[\text{tr}(\mathbf{G}^2) - n]^2} \right\}$. For breeding populations, m_e is often smaller than 100, so the randomized estimate is nearly unbiased.

2.3. Best linear prediction

Best linear prediction is used to compute genomic estimated breeding values (GEBV) for candidate populations:

$$\hat{\mathbf{y}}_2 = \hat{\mu}_1 \mathbf{1} + \hat{\sigma}_g^2 \mathbf{G}_{21} (\hat{\sigma}_g^2 \mathbf{G}_{11} + \hat{\sigma}_e^2 \mathbf{I})^{-1} (\mathbf{y}_1 - \hat{\mu}_1 \mathbf{1}) \quad (5)$$

in which $\hat{\mathbf{y}}_2$ represents the vector of the GEBV in the candidate populations, \mathbf{y}_1 represents the vector of the phenotypic values in the training populations, $\hat{\mu}_1$ represents the estimated mean value ($\hat{\mu}_1 = (\mathbf{1}^T \mathbf{V}^{-1} \mathbf{1})^{-1} \mathbf{1}^T \mathbf{V}^{-1} \mathbf{y}_1$; $\mathbf{V}^{-1} = (\hat{\sigma}_g^2 \mathbf{G}_{11} + \hat{\sigma}_e^2 \mathbf{I})^{-1}$), \mathbf{G}_{11} represents the additive genetic relationship matrix for the training populations, and \mathbf{G}_{21} represents the additive genetic relationship matrix between the candidate and training populations.

2.4. A modified algorithm for Proven and Young

If there exists an additive relationship between entries, as is often the case in pedigree-based GRM in breeding programs, a numerical shortcut may be applied to reduce the computational cost of matrix inversion. An early realization of such a fast GRM inversion was discovered and implemented by Henderson [21]. After decomposing a GRM into triangular and diagonal matrices, Henderson radically reduced the computational burden for the inversion of a pedigree-based GRM. Applying Henderson's method to a single nucleotide polymorphism (SNP)-based GRM, Misztal et al. [14,22] proposed the Algorithm for Proven and Young (APY) to invert the GRM. The APY inverse is:

$$\mathbf{V}^{-1} = \begin{bmatrix} \mathbf{I} & -\mathbf{P}_{cn} \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{V}_{cc}^{-1} & 0 \\ 0 & \mathbf{M}_{nn}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & 0 \\ -\mathbf{P}_{nc} & \mathbf{I} \end{bmatrix} \quad (6)$$

where \mathbf{V}_{cc} indicates the relationship matrix between the core individuals (a randomly selected subset of the training populations); $\mathbf{P}_{nc} = \mathbf{P}_{cn} = \mathbf{V}_{cc}^{-1} \mathbf{V}_{cn}$, in which \mathbf{V}_{cn} represents the relationship matrix between the core and noncore individuals; and \mathbf{M}_{nn} is a diagonal matrix of genomic Mendelian sampling for noncore individuals. Here \mathbf{M}_{nn} is as follows:

$$\mathbf{M}_{nn} = \text{diag}(\mathbf{V}_{nn}) - \text{diag}(\mathbf{P}_{cn}^T \mathbf{V}_{cn}) \quad (7)$$

in which \mathbf{V}_{nn} indicates the relationship matrix between the non-core individuals.

$$\mathbf{G}_{cn[k,k]} = \text{diag}(\mathbf{P}_{cn}^T \mathbf{V}_{cn})_{kk} = \mathbf{P}_{cn[k,k]}^T \mathbf{V}_{cn[k,k]} = \sum_{m=1}^c \mathbf{P}_{cn[k,m]}^T \mathbf{V}_{cn[m,k]} \quad (8)$$

Simplifying \mathbf{M}_{nn} , we propose a modified APY that further reduces the computational cost of inverting the large matrix. When the core and noncore individuals are random samples, $E(\mathbf{V}_{cn[m,k]}) = 0$. Therefore $\mathbf{G}_{cn[k,k]} = 0$ and $\mathbf{M}_{nn} = \text{diag}(\mathbf{V}_{nn}) = \mathbf{I}_{nn}$, where \mathbf{I}_{nn} is an identity matrix. When $\mathbf{M}_{nn} = \mathbf{I}_{nn}$, the computational time can be further reduced without loss of accuracy of APY. The modified APY inverse is consequently

$$\mathbf{V}_i^{-1} = \begin{bmatrix} \mathbf{I} & -\mathbf{P}_{cn} \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{V}_{cc}^{-1} & 0 \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} & 0 \\ -\mathbf{P}_{nc} & \mathbf{I} \end{bmatrix} \quad (9)$$

2.5. Simulated data

An F_2 population was simulated to compare the performance and computational time of GBLUP, HEAPY|A, and SHEAPY. A chromosome with a length of 5000 or 3000 centiMorgans [the recombination rate is c ($c = 0.01$ for approximately 1 cM under the Haldane mapping function) between the i^{th} and $(i+1)^{\text{th}}$ markers] was simulated. All markers on the chromosome were defined as QTL whose effects followed a standard normal distribution. Two conditions in the F_2 population were simulated: (1) multiple training population sizes (4000, 6000, 8000, and 10 000) with a fixed heritability ($h^2 = 0.75$); (2) multiple levels of heritability ($h^2 = 0.1, 0.3, 0.5, 0.7$ and 0.9) with a fixed training population size (1000). Each simulation scenario was replicated 10 times.

2.6. Real data

Two datasets, for *Arabidopsis thaliana* and sorghum, were used to evaluate the predictive accuracy of GBLUP, HEAPY|A, and SHEAPY. (1) Phenotype and genotype data of an *Arabidopsis thaliana* F_2 population (P169) derived from a cross between Ts-1 and Tsu-1 were obtained from Salomé et al. [23]. This F_2 population consisted of 447 plants, and 240 SNP markers were genotyped. Seven traits, including DTF1 (days until visible flower buds in the center of the rosette), DTF2 (days until inflorescence stem reached 1 cm in height), DTF3 (days until first open flower), RLN (rosette leaf number), CLN (cauline leaf number), TLN (total leaf number: sum of RLN and CLN), and LIR1 (leaf initiation rate (RLN/DTF1)), were used. (2) The phenotype and genotype data of a sorghum multiparental population were obtained from Higgins et al. [24]. This population consisted of 5 biparental $F_{2:3}$ families including 724 lines and 9139 SNP genotyped markers. The adjusted phenotypic values of plant height in Urbana, Illinois, USA (HT-IL), plant height in Mexico (HT-MX), and flowering time in Urbana, Illinois, USA (FL-IL) were used.

3. Results

3.1. Monte Carlo simulation

To assess the estimated heritability, predictability, and computational time of GBLUP, HEAPY|A, and SHEAPY, Monte Carlo simulations were performed for F_2 populations with four training population sizes (4000, 6000, 8000, and 10,000), candidate population size (100), heritability ($h^2 = 0.75$), 5001 equal-frequency biallelic markers, and recombination rate ($c = 0.01$) (Table 1). For HEAPY|A and SHEAPY, 1/2 of the training populations were ran-

domly selected as core populations. All methods showed nearly unbiased estimates of the true heritability of 0.75. All methods showed similar predictive accuracy. For example, when the size of a training population was 10,000, $\hat{r}_{\text{GBLUP}}^2 = 0.704 \pm 0.010$, $\hat{r}_{\text{HEAPY|A}}^2 = 0.698 \pm 0.012$, and $\hat{r}_{\text{SHEAPY}}^2 = 0.693 \pm 0.010$. SHEAPY greatly reduced computational time in comparison with GBLUP and HEAPY|A. When the size of a training population was 10,000, $T_{\text{GBLUP}} = 24972.2\text{s}$, $T_{\text{HEAPY|A}} = 2558.9\text{s}$, and $T_{\text{SHEAPY}} = 1005.8\text{s}$.

The predictive accuracy of the three methods was also compared for multiple levels of heritability when the simulation parameters were set as training population size (1000), core population size (800), candidate population size (100), heritability ($h^2 = 0.1, 0.3, 0.5, 0.7$, and 0.9), 3001 equal-frequency biallelic markers, and recombination rate ($c = 0.01$). SHEAPY showed better predictive ability than GBLUP in all cases and than HEAPY|A in most cases Table 2.

3.2. Genomic prediction of the traits in an *Arabidopsis thaliana* F_2 population and a sorghum population

For the *Arabidopsis thaliana* F_2 population, the size of the training population was 300 for GBLUP, HEAPY|A and SHEAPY and that of the core population was 200 for HEAPY|A and SHEAPY (Table 3). For all seven traits, GBLUP and HEAPY|A showed similar predictive accuracy. For DTF1, RLN, CLN, and TLN the three methods performed similarly, and for DTF2 and DTF3, SHEAPY slightly outperformed the other two. Only for LIR1 did GBLUP and HEAPY|A show an advantage over SHEAPY (0.406 ± 0.018 for GBLUP, 0.402 ± 0.019 for HEAPY|A, and 0.340 ± 0.022 for SHEAPY). Thus, the predictive ability of SHEAPY was similar to or better than that of GBLUP and HEAPY|A in most cases.

For the sorghum multiparental population, the size of the training population was 500 for GBLUP, HEAPY|A, and SHEAPY, and the size of the core population was 400 for HEAPY|A and SHEAPY (Table 4). GBLUP was superior to HEAPY|A for all three traits, and showed predictive accuracy similar to that of SHEAPY for HT-IL (0.842 ± 0.015 for GBLUP and 0.845 ± 0.011 for SHEAPY). GBLUP also outperformed SHEAPY for HT-MX and FL-IL. But the results were inconsistent with those from simulations, and investigating possible causes revealed that the relative frequencies of most of the reference alleles were between 0.6 and 1.0 at each marker locus and that this distribution held for 77% of the markers. There were no or only a few non-reference genotypes at some marker loci in the core population when the core population was randomly selected from the training population, and this deficiency severely impaired the predictive accuracy of HEAPY|A and SHEAPY. Accordingly, the predictive accuracies of GBLUP, HEAPY|A, and SHEAPY were evaluated again with 2081 markers whose non-reference allele proportions were greater than 0.415, to exclude the impact of invalid markers. SHEAPY then showed higher predictive accuracy than HEAPY|A for all three traits, and than GBLUP for HT-IL and FL-IL (Table 5).

4. Discussion

GP has increased the rate of genetic gain and reduced the generation interval. With its wide application, a massive quantity of historical datasets have been accumulated, with which large training populations can be established to achieve high predictive accuracy. But the dimension of GRMs will increase geometrically when multi-trait and/or multi-environment data are integrated into the GP model. Thus computational efficiency in predicting individuals with large-scale data emerges as a necessity.

Inversion of the GRM, an indispensable step in GBLUP, becomes very difficult when the size of the training population reaches

Table 1

Comparison of the estimated heritability, predictive accuracy, and computational time of GBLUP, HEAPY|A, and SHEAPY for multiple sizes of training population based on 10 simulation in an F_2 population.

Training size	$\widehat{h^2}$			Prediction accuracy ($\widehat{r^2}$)			Average time of each simulation		
	GBLUP	HEAPY A	SHEAPY	GBLUP	HEAPY A	SHEAPY	GBLUP	HEAPY A	SHEAPY
4000	0.790 ± 0.036	0.739 ± 0.035	0.733 ± 0.041	0.680 ± 0.016	0.668 ± 0.014	0.684 ± 0.016	3739.7	275.6	97.6
6000	0.777 ± 0.031	0.728 ± 0.024	0.758 ± 0.036	0.693 ± 0.015	0.673 ± 0.012	0.678 ± 0.015	10857.9	719.5	260.9
8000	0.775 ± 0.035	0.733 ± 0.030	0.743 ± 0.026	0.714 ± 0.013	0.699 ± 0.014	0.716 ± 0.012	18462.5	1453.5	548.6
10,000	0.762 ± 0.034	0.733 ± 0.029	0.744 ± 0.044	0.704 ± 0.010	0.698 ± 0.012	0.693 ± 0.010	24972.2	2558.9	1005.8

The training population sizes were 4000, 6000, 8000, and 10 000. The core population size for HEAPY|A and SHEAPY was half that of the training population. The candidate population size was a fixed number (100), heritability (h^2) was set as 0.75, and the recombination rate (c) was set as 0.01. The squared correlation coefficient (r^2) between the phenotypes and the predicted genotypic values was defined as the prediction accuracy, and the values after ± symbols represent corresponding standard errors.

Table 2

Comparison of the estimated heritability and predictive accuracy of GBLUP, HEAPY|A, and SHEAPY at multiple levels of heritability based on 10 simulations with a simulated F_2 population.

h^2	GBLUP		HEAPY A		SHEAPY	
	$\widehat{h^2}$	Prediction accuracy ($\widehat{r^2}$)	$\widehat{h^2}$	Prediction accuracy ($\widehat{r^2}$)	$\widehat{h^2}$	Prediction accuracy ($\widehat{r^2}$)
0.1	0.099 ± 0.007	0.062 ± 0.020	0.106 ± 0.010	0.062 ± 0.020	0.104 ± 0.010	0.103 ± 0.022
0.3	0.309 ± 0.013	0.220 ± 0.030	0.296 ± 0.020	0.221 ± 0.029	0.318 ± 0.023	0.273 ± 0.029
0.5	0.527 ± 0.022	0.403 ± 0.029	0.482 ± 0.027	0.403 ± 0.028	0.530 ± 0.038	0.446 ± 0.028
0.7	0.763 ± 0.036	0.604 ± 0.023	0.666 ± 0.033	0.603 ± 0.022	0.742 ± 0.054	0.619 ± 0.022
0.9	0.945 ± 0.035	0.789 ± 0.017	0.835 ± 0.034	0.819 ± 0.011	0.882 ± 0.047	0.807 ± 0.012

The training population size was 1000. The core population size was 800. The candidate population size was 100. The number of markers was 3001. The heritability (h^2) was set as 0.1, 0.3, 0.5, 0.7 and 0.9. The recombination rate (c) was set as 0.01. The squared correlation coefficient (r^2) between phenotypes and the predicted genotypic values was defined as the prediction accuracy and values following ± symbols represent the corresponding standard errors.

Table 3

Comparison of predictive accuracy among GBLUP, HEAPY|A, and SHEAPY for seven traits in an *Arabidopsis thaliana* F_2 (P169) population based on 10 simulations.

Trait	Training	Candidate	GBLUP	HEAPY A	SHEAPY
DTF1	300	146	0.321 ± 0.023	0.319 ± 0.024	0.312 ± 0.036
DTF2	300	147	0.606 ± 0.015	0.605 ± 0.014	0.637 ± 0.019
DTF3	300	147	0.678 ± 0.014	0.677 ± 0.015	0.701 ± 0.014
RLN	300	147	0.683 ± 0.029	0.686 ± 0.030	0.686 ± 0.017
CLN	300	147	0.532 ± 0.016	0.538 ± 0.014	0.522 ± 0.015
TLN	300	147	0.707 ± 0.024	0.711 ± 0.022	0.703 ± 0.016
LIR1	300	147	0.406 ± 0.018	0.402 ± 0.019	0.340 ± 0.022

The core population size of HEAPY|A and SHEAPY was 200. The squared correlation coefficient (r^2) between the phenotypes and the predicted genotypic values was defined as the prediction accuracy and the values following ± symbols represent the corresponding standard errors. DTF1, days until visible flower buds in the center of the rosette; DTF2, days until inflorescence stem reached 1 cm in height; DTF3, days until first open flower; RLN, rosette leaf number; CLN, cauline leaf number; TLN, total leaf number; sum of RLN and CLN; LIR1, leaf initiation rate (RLN/DTF1).

Table 4

Comparison of predictive accuracy among GBLUP, HEAPY|A, and SHEAPY for three traits in a sorghum $F_{2:3}$ population based on 10 simulations with 9139 markers.

Traits	Training	Candidate	GBLUP	HEAPY A	SHEAPY
HT-IL	500	212	0.842 ± 0.015	0.610 ± 0.058	0.845 ± 0.011
HT-MX	500	207	0.850 ± 0.004	0.220 ± 0.070	0.723 ± 0.081
FL-IL	500	212	0.861 ± 0.003	0.361 ± 0.069	0.682 ± 0.072

The core population size for HEAPY|A and SHEAPY was 400. The squared correlation coefficient (r^2) between the phenotypes and the predicted genotypic values was defined as the prediction accuracy and the values following ± symbols represent the corresponding standard errors. HT-IL, plant height in Urbana, Illinois, USA; HT-MX, plant height in Mexico; FL-IL, flowering time in Urbana, Illinois, USA.

Table 5

Comparison of predictive accuracy among GBLUP, HEAPY|A and SHEAPY for three traits of a sorghum $F_{2:3}$ population based on 10 simulations with 2081 markers.

Traits	Training	Candidate	GBLUP	HEAPY A	SHEAPY
HT-IL	500	212	0.719 ± 0.072	0.696 ± 0.044	0.850 ± 0.016
HT-MX	500	207	0.851 ± 0.006	0.433 ± 0.091	0.668 ± 0.086
FL-IL	500	212	0.603 ± 0.081	0.389 ± 0.061	0.674 ± 0.083

The core population size of HEAPY|A and SHEAPY was 400. The number of markers whose proportion of the non-reference allele is larger than 0.415 in the multiparental sorghum population is 2081. The squared correlation coefficient (r^2) between the phenotypes and the predicted genotypic values was defined as the prediction accuracy and the values following ± symbols represent the corresponding standard errors. HT-IL, plant height in Urbana, Illinois, USA; HT-MX, plant height in Mexico; FL-IL, flowering time in Urbana, Illinois, USA.

150,000 [25], and new methods such as APY, updating the inverse, and a recursive algorithm have been presented accordingly [22,26–27]. Chen [28] proposed HE regression based on IBS to estimate heritability, markedly reducing computational time in comparison with GBLUP by adoption of least-squares instead of REML estimation [17–18]. Combining the merits of APY and HE, we have previously developed HEAPY[A] [12], whose computational speed was over 13 times that of GBLUP when the size of a training population was 4000 and that of the core population was half that.

In this study, we developed a new computational approach (SHEAPY) combining RHE-reg with a modified APY in an additive-effect model. Removing the computation of a GRM between noncore individuals via replacement of M_{nn} with I_{nn} greatly accelerated the modified APY. When the size of a training population was 10,000 and that of the core population was half of it, the computational speed of SHEAPY was over 20 times of GBLUP and over 2 times of HEAPY[A] with similar predictive accuracy. When the size of a training population was relatively small, as it was in this study, the size of the core population should be increased to obtain predictive accuracy similar to or better than GBLUP. For example, the size of the training population was set as 300 and the core population as 2/3 of that (200) in the *Arabidopsis thaliana* F₂ population. The prediction results from the sorghum multiparental population showed that rare markers severely impaired the predictive performance of the SHEAPY owing to their low representation in the core population and accordingly should be excluded.

How to choose the core population is crucial but remains an open question. Pocrnic et al. [25] found that the optimal size of a core population depended on effective population size (for example, the optimal number is about 14,000 for Holstein and Angus cattle, 12,000 for Jersey cattle, and 6000 for pigs and broiler chickens). It remains to be investigated which individuals need to be selected in order to further reduce the computational cost and maintain predictive accuracy in practice.

5. Conclusions

Integrating I) a randomized algorithm for estimating variance components in RHE-reg and II) a modified APY, the proposed SHEAPY algorithm showed marked superiority to GBLUP and HEAPY[A] in computational efficiency. A computational framework was developed under the additive-effect model. A computer program (SHEAPY) is available from the authors.

CRedit authorship contribution statement

Hailan Liu: Conceptualization, Investigation, Formal analysis, Methodology, Writing - original draft, Writing -review & editing, Project administration. **Guo-Bo Chen:** Conceptualization, Investigation, Methodology, Writing - original draft, Writing - review & editing, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by the National Natural Science Foundation of China to Guo-Bo Chen (31771392) and Zhejiang Provin-

cial People's Hospital Research Startup to Guo-Bo Chen (ZRY2018A004). The authors are grateful to Prof. James C. Nelson for English editing.

References

- [1] R. Lande, R. Thompson, Efficiency of marker-assisted selection in the improvement of quantitative traits, *Genetics* 124 (1990) 743–756.
- [2] C. Schrooten, H. Bovenhuis, J.A.M. van Arendonk, P. Bijma, Genetic progress in multistage dairy cattle breeding schemes using genetic markers, *J. Dairy Sci.* 88 (2005) 1569–1581.
- [3] X. Zhao, G. Tan, Y. Xing, L. Wei, Q. Chao, W. Zuo, T. Lübberstedt, M. Xu, Marker-assisted introgression of *qHSR1* to improve maize resistance to head smut, *Mol. Breed.* 30 (2012) 1077–1088.
- [4] X. Hao, X. Li, X. Yang, J. Li, Transferring a major QTL for oil content using marker-assisted backcrossing into an elite hybrid to increase the oil content in maize, *Mol. Breed.* 34 (2014) 739–748.
- [5] E.L. Heffner, A.J. Lorenz, J.L. Jannink, M.E. Sorrells, Plant breeding with genomic selection: gain per unit time and cost, *Crop Sci.* 50 (2010) 1681–1690.
- [6] J.L. Jannink, A.J. Lorenz, H. Iwata, Genomic selection in plant breeding: from theory to practice, *Brief. Funct. Genomics* 9 (2010) 166–177.
- [7] T.H. Meuwissen, B.J. Hayes, M.E. Goddard, Prediction of total genetic value using genome-wide dense marker maps, *Genetics* 157 (2001) 1819–1829.
- [8] M.E. Goddard, B.J. Hayes, Genomic selection, *J. Anim. Breed. Genet.* 124 (2007) 323–330.
- [9] M. Georges, C. Charlier, B. Hayes, Harnessing genomic information for livestock improvement, *Nat. Rev. Genet.* 20 (2019) 135–156.
- [10] K.P. Voss-Fels, M. Cooper, B.J. Hayes, Accelerating crop genetic gains with genomic selection, *Theor. Appl. Genet.* 132 (2019) 669–686.
- [11] X. Wang, Y. Xu, Z. Hu, C. Xu, Genomic selection methods for crop improvement: current status and prospects, *Crop J.* 6 (2018) 330–340.
- [12] H. Liu, G.B. Chen, A rapid genomic selection method combining Haseman-Elston (HE) model and algorithm for proven and young (APY), *Mol. Breed.* 40 (2020) 12.
- [13] I. Misztal, D. Lourenco, A. Legarra, Current status of genomic evaluation, *J. Anim. Sci.* 98 (2020) 1–14.
- [14] I. Misztal, A. Legarra, I. Aguilar, Using recursion to compute the inverse of the genomic relationship matrix, *J. Dairy Sci.* 97 (2014) 3943–3952.
- [15] B.O. Fragomeni, D.A.L. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, A. Legarra, T.J. Lawlor, I. Misztal, Hot topic: use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes, *J. Dairy Sci.* 98 (2015) 4090–4094.
- [16] Y. Masuda, I. Misztal, S. Tsuruta, A. Legarra, I. Aguilar, D.A.L. Lourenco, B.O. Fragomeni, T.J. Lawlor, Implementation of genomic recursions in single-step genomic best linear unbiased predictor for US Holsteins with a large number of genotyped animals, *J. Dairy Sci.* 99 (2016) 1968–1974.
- [17] H. Liu, G.B. Chen, A fast genomic selection approach for large genomic data, *Theor. Appl. Genet.* 130 (2017) 1277–1284.
- [18] H. Liu, G.B. Chen, A new genomic prediction method with additive-dominance effects in the least-squares framework, *Heredity* 121 (2018) 196–204.
- [19] Y. Wu, S. Sankaranarayanan, A scalable estimator of SNP heritability for biobank-scale data, *Bioinformatics* 34 (2018) i187–i194.
- [20] N. Halko, P.G. Martinsson, J.A. Tropp, Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions, *SIAM Rev.* 53 (2011) 217–288.
- [21] C.R. Henderson, A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values, *Biometrics* 32 (1976) 69–83.
- [22] I. Misztal, Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size, *Genetics* 202 (2016) 401–409.
- [23] P.A. Salome, K. Bomblies, R.A.E. Laitinen, L. Yant, R. Mott, D. Weigel, Genetic architecture of flowering-time variation in *Arabidopsis thaliana*, *Genetics* 188 (2011) 421–433.
- [24] R.H. Higgins, C.S. Thurber, I. Assaranurak, P.J. Brown, Multiparental mapping of plant height and flowering time QTL in partially isogenic sorghum families, *G3-Genes Genomes Genet.* 4 (2014) 1593–1602.
- [25] I. Pocrnic, D.A.L. Lourenco, Y. Masuda, I. Misztal, Dimensionality of genomic information and performance of the Algorithm for Proven and Young for different livestock species, *Genet. Sel. Evol.* 48 (2016) 82.
- [26] P. Faux, N. Gengler, I. Misztal, A recursive algorithm for decomposition and creation of the inverse of the genomic relationship matrix, *J. Dairy Sci.* 95 (2012) 6093–6102.
- [27] K. Meyer, B. Tier, H.U. Graser, Technical note: updating the inverse of the genomic relationship matrix, *J. Anim. Sci.* 91 (2013) 2583–2586.
- [28] G.B. Chen, Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman-Elston regression, *Front. Genet.* 5 (2014) 107.