BMC Bioinformatics

RESEARCH Open Access

Ensemble learning for integrative prediction of genetic values with genomic variants

Lin-Lin Gu[1], Run-Qing Yang[2], Zhi-Yong Wang[1*], Dan Jiang[1*] and Ming Fang[1,3*]

*Correspondence:
zywang@jmu.edu.cn;
jiangdan666@163.com;
fangming618@126.com

[1] Key Laboratory of Healthy Mariculture for the East China Sea, Ministry of Agriculture and Rural Affairs and Fisheries College, Jimei University, Xiamen, People's Republic of China
[2] Research Center for Aquatic Biotechnology, Chinese Academy of Fishery Sciences, Beijing, People's Republic of China
[3] Life Science College, Heilongjiang Bayi Agricultural University, Daqing, People's Republic of China

Abstract

Background: Whole genome variants offer sufficient information for genetic prediction of human disease risk, and prediction of animal and plant breeding values. Many sophisticated statistical methods have been developed for enhancing the predictive ability. However, each method has its own advantages and disadvantages, so far, no one method can beat others.

Results: We herein propose an Ensemble Learning method for Prediction of Genetic Values (ELPGV), which assembles predictions from several basic methods such as GBLUP, BayesA, BayesB and BayesCπ, to produce more accurate predictions. We validated ELPGV with a variety of well-known datasets and a serious of simulated datasets. All revealed that ELPGV was able to significantly enhance the predictive ability than any basic methods, for instance, the comparison *p*-value of ELPGV over basic methods were varied from 4.853E−118 to 9.640E−20 for WTCCC dataset.

Conclusions: ELPGV is able to integrate the merit of each method together to produce significantly higher predictive ability than any basic methods and it is simple to implement, fast to run, without using genotype data. is promising for wide application in genetic predictions.

Keywords: Genomic prediction, Genomic selection, Ensemble learning, Machine learning

Background

Genome-wide distributed variants provide sufficient information for prediction of genetic value. In human studies, genetic value prediction is usually applied for prediction of complex traits such as disease risk and human height [1, 2]. In plants and animals, genetic prediction is important for genetic selection [3, 4]. So far, a variety of statistical analysis methods have been used to predict genetic values. Genomic best linear unbiased estimates (GBLUP) are the most common method, which uses genome-wide molecular markers to construct a kinship matrix between individuals and then uses BLUP techniques to predict individual genetic values [5]. Bayesian method is another popular method, which mainly includes BayesA, BayesB, BayesCπ and BayesLASSO, etc. [6–8]. These methods use Monte Carlo Markov chain techniques to estimate parameters. The main difference among them is the assignment of hyperparameters for variables, and each method has its own advantages and disadvantages. BayesA is mainly applicable for the traits controlled by genes with multiple tiny effects, whereas BayesB and BayesCπ are suitable for the traits controlled by a small number of main effect genes. In human disease risk prediction, "Clumping + Thresholding" (C + T) method has been developed and applied [9–12]. C + T method first identifies a set of markers with predictive power, and then uses these markers to predict disease risk by logistic regression [1, 13, 14], which is suitable for the disease controlled by several main effect genes. Although several methods exist, each has its own limitations, so far, there is no one method that always outperforms others.

Ensemble learning is a machine learning method, which integrates the predictions from multiple methods to obtain a new prediction through supervised or unsupervised learning methods [15]. As early as 20 years ago, it was found that ensemble learning can reduce generalization error [16] and ensemble methods that combine the output of multiple methods have been shown to achieve better generalizability than a single method [17]. So far, ensemble learning has independently made a substantial impact on the field of bioinformatics through their widespread applications [18]. One example is in predicting localization of long non-coding RNAs, where multiple sub-networks were used to integrate distinct feature sets to maximize method performance [19]. In another work, a CNN/RNN (Convolutional Neural Networks/Recurrent Neural Network) ensemble was used to integrate features and raw sequence data to predict different types of translation initiation sites [20], overcoming the generalizability issue of traditional methods that can only predict a specific type of translational initiation sites. Moreover, the stability and reproducibility offered by ensemble methods such as in feature selection are also making a substantial impact in biomarker discovery [21, 22]. To our best knowledge, the remarkable flexibility and adaptability characters of ensemble learning has led to the proliferation of their application in bioinformatics research [23].

We herein propose an ensemble learning method for Prediction of Genetic Values (ELPGV). ELPGV trains several different basic methods, such as GBLUP, BayesA, BayesB and BayesCπ, to produce more accurate prediction. The core of ELPGV uses the hybrids of differential evolution [24] and particle swarm optimization [25] to train the weight, by which the predictions of basic methods are weighted averaged to generate new prediction. A variety of dataset including WTCCC (Wellcome Trust Case Control Consortium), IBDGC (International Inflammatory Bowel Disease Genetics Consortium), cattle, wheat and computational simulations are employed to validate ELPGV.

Materials and methods

Basic methods

The prediction is based on a linear method according to Eq. (1):

$$\boldsymbol{y} = \boldsymbol{X\alpha} + \boldsymbol{Z\beta} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{y}$ is the phenotypes; $\boldsymbol{X}$ is design matrix for fixed effects; $\boldsymbol{\alpha}$ is the fixed effect; $\boldsymbol{Z}$ is genotypes of variants, coding with "0", "1" and "2" for genotypes "AA", "Aa" and "aa" respectively, or genotype dosages of SNPs; $\boldsymbol{\beta}$ is the SNP effects; and $\boldsymbol{e}$ is the residual errors, assumed to follow normal distribution,$\boldsymbol{e} \sim N\left(0, \boldsymbol{I}\sigma_e^2\right)$, where $\boldsymbol{I}$ is a vector of identity matrix and σ_e^2 is the residual variance.

In this study, four basic methods are used for genetic value predictions, BayesA, BayesB, BayesCπ and GBLUP. In BayesA, all SNPs are assumed to contribute to genetic variation, and the variance of the SNP effect is assumed to follow inverse chi-square distribution; BayesB and BayesCπ assumes a small fraction (π) of SNPs have non-zero effects [6, 8], where π is set as 0.1 in BayesB [26]. The Bayesian methods are implemented with the function "BGLR" in the R package "BGLR" [27]. In the GBLUP, the variances of all SNP effects are assumed to be equal, and then the genetic values are estimated with mixed model equation through kinship matrix constructed with SNPs [5]. The GBLUP is implemented using the function "emmreml" in the R package "EMMREML" [28].

ELPGV model construction

The ELPGV framework comprises two components, weight training and weighted prediction. First, it trains basic methods to get predictions; then, it trains the weight of basic methods with machine learning; finally, it generates new predictions by the weighted average of the predictions of basic methods. The schematic diagram of the study methodology is given in Fig. 1.

Suppose n basic methods are investigated, the prediction of ELPGV can be expressed as Eq. (2), where, $\boldsymbol{p}_j$ is the predicted values of the jth basic method, which is easily obtained from each basic method, and W_j is the weight of the jth basic method, respectively.

$$\boldsymbol{g}_{predicted} = \sum_{j=1}^{n} W_j \times \boldsymbol{p}_j \tag{2}$$

To train the weight $\boldsymbol{W}$, a fitness function is defined as the correlation coefficient between the predicted values $\boldsymbol{g}_{predicted}$ and observed values $\boldsymbol{y}_{observed}$ (Eq. 3), $\boldsymbol{g}_{predicted}$ is the predicted values of ELPGV based on Eq. (2).

$$f(\boldsymbol{W}) = \frac{\sum \left(\boldsymbol{y}_{observed} - \overline{\boldsymbol{y}}_{observed}\right)\left(\boldsymbol{g}_{predicted} - \overline{\boldsymbol{g}}_{predicted}\right)}{\sqrt{\sum \left(\boldsymbol{y}_{observed} - \overline{\boldsymbol{y}}_{observed}\right)^2}\sqrt{\sum \left(\boldsymbol{g}_{predicted} - \overline{\boldsymbol{g}}_{predicted}\right)^2}} \tag{3}$$

For testing population, phenotype $\boldsymbol{y}_{observed}$ is unknown, we therefore introduce reference genetic values to replace the unknown phenotypic values in Eq. (3). The genetic

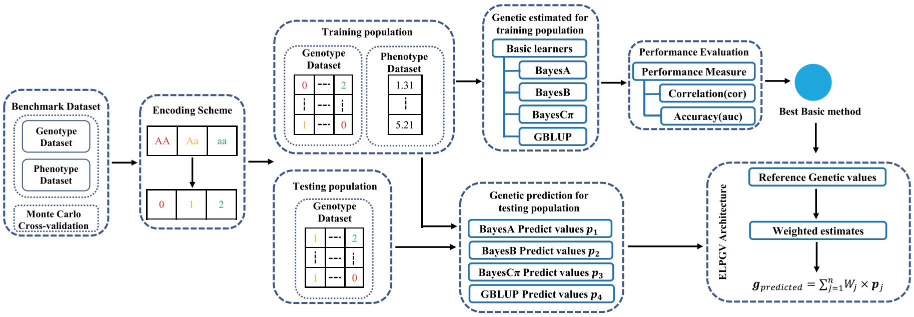

Fig. 1 Schematic diagram of the study methodology

predictions with the best fitness among basic methods was taken as the reference genetic values.

ELPGV uses a mixture of differential evolution (DE) algorithm and particle swarm optimization (PSO) algorithm to estimate the weight $\boldsymbol{W}$, which includes initialize, mutation, crossover and selection steps.

Step 1. Initialization: ELPGV randomly initializes the weights $\boldsymbol{W}_{i\cdot}(W_{i,1},\ldots,W_{i,j})$ and the optimization velocities $\boldsymbol{V}_{i\cdot}(V_{i,1},\ldots,V_{i,j})$, for $i=1,\ldots,m$ and $j=1,\ldots,n$, where m is the number of particles or the number of candidate weight; and n is the number of basic methods. The weight is initialized with Eq. (4) and the optimization velocity is initialized with Eq. (5).

$$W_{i,j}=rand(W_{min},W_{max}) \tag{4}$$

$$V_{i,j}=rand(V_{min},V_{max}) \tag{5}$$

First, the m group weights are replaced into Eq. (2) to obtain the ELPGV predictions of m groups, respectively; then the predictions are replaced into Eq. (3) to assess the corresponding fitness for each group. We then define the optimal weight $\boldsymbol{W}(0)$ as the best fitness one in all group weights,

$$\boldsymbol{W}(0)=argmax\big(f(\boldsymbol{W}_{i\cdot})\big) \tag{6}$$

Step 2. Mutation: In t th iteration, Eq. (4) is replaced with Eqs. (7) and (8) for updating the weight of each group, respectively.

$$\boldsymbol{P}_{i\cdot}(t)=\boldsymbol{W}_{i\cdot}(t-1)+\boldsymbol{V}_{i\cdot}(t-1) \tag{7}$$

$$\boldsymbol{H}_{i\cdot}(t)=\boldsymbol{W}_{k\cdot}(t-1)+F\times\big(\boldsymbol{W}_{p\cdot}(t-1)-\boldsymbol{W}_{q\cdot}(t-1)\big) \tag{8}$$

where F is scaling factor, controlling the effect of difference vector, the index $i\neq k\neq p$ $\neq q$.

Step 3. Crossover: The crossover operation switches the weight at current iteration (t) and last iteration ($t-1$) randomly with Eq. (9), where CR is crossover probability and $rand_{i\cdot}(0,1)$ is a random value between 0 and 1 of i th group weight.

$$\boldsymbol{U}_{i\cdot}(t)=\begin{cases}\boldsymbol{H}_{i\cdot}(t) & rand_{i\cdot}(0,1)\leq CR\\ \boldsymbol{W}_{i\cdot}(t-1) & else\end{cases} \tag{9}$$

Step 4. Selection: Last, the all the group weights are updated with Eqs. (10) and (11).

$$\boldsymbol{G}_{i\cdot}(t)=\begin{cases}\boldsymbol{U}_{i\cdot}(t) & f(\boldsymbol{U}_{i\cdot}(t))\geq f(\boldsymbol{P}_{i\cdot}(t))\\ \boldsymbol{P}_{i\cdot}(t) & else\end{cases} \tag{10}$$

$$\boldsymbol{W}_{i\cdot}(t)=\begin{cases}\boldsymbol{G}_{i\cdot}(t) & f(\boldsymbol{G}_{i\cdot}(t))\geq f(\boldsymbol{W}_{i\cdot}(t-1))\\ \boldsymbol{W}_{i\cdot}(t-1) & else\end{cases} \tag{11}$$

After t th iteration, each group weight has a velocity which are updated as Eq. (12), where ε is inertia weight, c_1 and c_2 are accelerated factors.

$$\begin{aligned} \boldsymbol{V}_{i\cdot}(t) &= \varepsilon * \boldsymbol{V}_{i\cdot}(t-1) + c_1 * rand(0,1) * (\boldsymbol{W}_{i\cdot}(t) - \boldsymbol{W}_{i\cdot}(t-1)) \\ &\quad + c_2 * rand(0,1) * (\boldsymbol{W}(t-1) - \boldsymbol{W}_{i\cdot}(t-1)) \end{aligned} \tag{12}$$

At the same time, ELPGV updates the fitness with new weights at t th updating with Eq. (3), the optimal weight can be expressed as Eq. (13) in t th iteration.

$$\boldsymbol{W}(t) = argmax\left(f(\boldsymbol{W}_{i\cdot}(t))\right) \tag{13}$$

After the fitness meets a certain criterion, or the iterations reach the maximum number, ELPGV returns the optimal weights $\boldsymbol{W}$ and the predictions with Eq. (2). To reduce sampling error and increase the estimate accurate of weights, the whole estimates are repeated for 100 times and the averaged weights are taken for ELPGV (Table 1).

Monte Carlo cross-validation

Cross-validation was employed to evaluate the prediction performance of GS methods. The individuals of each dataset were first randomly divided into two parts with ratio 9:1, and they were taken as training set and testing set, respectively. The cross-validation was repeated 100 times. In the prediction, the phenotypes of individuals in testing set were masked, and the genetic values were predicted with training set; then the Pearson's correlation coefficient between the predicted values and their true phenotypes were used to evaluate the predictive ability of each method.

Paired-sample t-test

Because all the methods are compared with the same replicated dataset, we were able to compare ELPGV with other basic methods using paired-sample t-test, which is expressed as $t = \overline{d}/s_{\overline{d}}$, with degree of freedom $n-1$, where n is the times of cross validation and d is the difference of the predictive ability between ELPGV and other methods.

WTCCC dataset

The WTCCC dataset was accessed from the Wellcome Trust Case Control Consortium (WTCCC1, https://www.wtccc.org.uk/) [29], including 14,000 cases and 2,938 shared controls, all were genotyped for ~ 450,000 SNPs. Six diseases were investigated, including

Table 1 Lists the hyper parameters used in above equations

Parameters	Value
W_{min} minimum weight	0
W_{max} maximum weight	1
V_{min} minimum update velocity	− 0.01
V_{max} maximum update velocity	0.01
m the weight size	20
F scaling factor	0.5
CR crossover probability	0.3
ε inertia weight	1
c_1 accelerated factor 1	2
c_2 accelerated factor 2	2
Max_iterations	25

bipolar disorder (BD), coronary artery disease (CAD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D). For this study, we removed SNPs using PLINK [30], with either minor allele frequency (MAF) < 0.01, or genotype call rate (CR) < 0.95, or p-value < 0.05 from Hardy–Weinberg equilibrium (HWE) test; then the SNPs were further pruned with PLINK [30] ($r^2 = 0.5$) for reducing computational burden. The number of cases and the SNPs of each disease are shown in Table 2.

Inflammatory bowel disease (IBD) dataset

The inflammatory bowel disease dataset was accessed from the International IBD Genetics Consortium (IBDGC), including 20,155 Crohn disease (CD), 15,191 ulcerative colitis disease (UC) and 34,257 controls of European ancestry. In total, genotypes were called using optiCall for 192,402 autosomal variants before quality control. A total of 161,681 SNPs was available after removing the SNPs with MAF < 0.02 and p-value < 10e−5 from the HWE test. The missing genotypes were imputed with impute2 using 1000 genome as a reference. (For details, see refs. [31]. To reduce computation burden, we further pruned SNPs for linkage disequilibrium with threshold $r^2 = 0.5$ using PLINK [30] and randomly sampled 1,000 individuals from Liege and Brussels batches.

Cattle dataset

German Holstein genomic prediction population was further employed to validate ELPGV, which comprised 5024 bulls [32], and all were genotyped with the Illumina Bovine SNP50 Beadchip [33]. After removing the SNPs with HWE p-value < 10 − 4, CR < 0.95 and MAF < 0.01, a total of 42,551 SNPs remained for the downstream analysis. The estimated breeding values of three traits milk fat percentage (mfp), milk yield (my), and somatic cell score (scs) were available and used in this study.

Wheat dataset

The wheat dataset was collected from CIMMYT's Global Wheat Program, the grain yields (GY) of the 599 wheat inbred lines were recorded for four places [34, 35]. Each wheat line was genotyped with 1447 Diversity Array Technology (DArT) by Triticarte Pty. Ltd, which had two genotypes coded with "0" or "1", to indicate its presence or absence, respectively, after filtering, 1279 markers were kept for analysis.

Table 2 Brief summary of the disease to WTCCC data sets

Disease	Case size	SNP size
BD	1868	373,369
CAD	1926	372,541
HT	1952	373,338
RA	1860	373,056
T1D	1963	372,964
T2D	1924	373,149

bipolar disorder (BD), coronary artery disease (CAD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D) and type 2 diabetes (T2D)

Simulations

We took advantage of the genotypes of wheat datasets for simulation. A number of QTL were simulated with effects sampled from gamma distribution with scale parameter 1.66 and shape parameter 0.4; the residual errors were sampled from normal distribution with variance set according to the heritability. We performed two simulation experiments to investigate the performance of ELPGV: (1) simulation of different number of QTLs, 5 and 1,000, respectively; and (2) simulation of different heritabilities, 0.5 and 0.2, respectively. This led to four sets of experiments (QTL5 and $h^2 = 0.2$, QTL5 and $h^2 = 0.5$, QTL1000 and $h^2 = 0.2$, QTL1000 and $h^2 = 0.5$).

Results

We used both real dataset and simulated dataset to validate the performance of ELPGV. In this study, four popular GS methods, GBLUP, BayesA, BayesB, and BayesCπ, were used for assembling with ELPGV, although ELPGV is able to assemble as many methods as possible. In addition, cross-validation was employed to evaluate the prediction performance of each method. Taken advantage of the fact that all the methods were compared with the same dataset, we used paired-sample *t*-test for significance comparison.

WTCCC dataset

We first illustrated the results of T2D. Four basic methods, GBLUP, BayesA, BayesB, and BayesCπ were applied for genetic prediction, BayesCπ performed the highest predictive ability ($r = 0.8471$) and GBLUP performed the lowest ($r = 0.4390$) (Table 3). We then used ELPGV to assemble the predictions of four Basic methods to generate new predictions. To this end, we first evaluated the fitting effect of four basic methods with train set, the basic methods with the best fitting effect was used to generate the reference genetic values. The fitting effect was defined as the correlation between the estimated genetic values and the phenotypes in train set. It was found that BayesCπ usually had the best n than other methods. With reference genetic values, ELPGV assembled four basic methods to obtain new predictions, the average predictive ability of ELPGV across 100 validations was $r = 0.8471$, significantly higher than any basic methods with comparison *p*-value ranged from 1.090E−112 (GBLUP) to 6.458E−31 (BayesCπ) Table 3). Because we compared each method with the same dataset, we were able to compare ELGPV with four basic methods in each of 100 experiments, separately. Figure 2a–f shows the prediction abilities in each of experiment, ELPGV is more accurate than other four basic methods, and the advantage of ELPGV over GBLUP is more obvious. We also compared ELPGV with four basic methods in dataset of BD, CAD, T1D, RA and HT (Table 3). For all diseases, ELPGV was obviously more accurate than four basic methods with *p*-values ranged from 4.853E−118 to 9.640E−20 (Table 3).

IBD dataset

We also applied ELPGV to predict disease risk for IBD dataset of European ancestry. The averaged predictive ability of 100 cross-validations of GBLUP, BayesA, BayesB

Table 3 The predictive ability of four basic methods and ELPGV, and the comparison *p*-value between ELPGV and others in T1D, T2D, BD, RA, CAD, HT with WTCCC dataset

Method	T1D		T2D		BD		RA		CAD		HT	
	Predictive ability	*p*-value	Predictive ability	*p*-value	Predictive ability	*p*-value	Predictive ability	*p*-value	Predictive ability	*p*-value	Predictive ability	*p*-value
ELPGV	0.8879 ± 0.0008	—	0.8471 ± 0.0009	—	0.9195 ± 0.0006	—	0.8843 ± 0.0008	—	0.8705 ± 0.0008	—	0.9132 ± 0.0006	—
BayesA	0.8664 ± 0.0009	9.037E−69	0.8022 ± 0.0014	9.330E−72	0.8984 ± 0.0007	5.191E−73	0.8628 ± 0.0010	1.454E−68	0.8459 ± 0.0010	6.011E−61	0.8879 ± 0.0007	2.525E−75
BayesB	0.8841 ± 0.0008	4.330E−27	0.8396 ± 0.0010	1.517E−33	0.9167 ± 0.0005	4.416E−32	0.8810 ± 0.0008	1.425E−27	0.8651 ± 0.0008	2.272E−28	0.9100 ± 0.0005	9.319E−29
BayesCπ	0.8863 ± 0.0008	1.827E−23	0.8447 ± 0.0010	6.458E−31	0.9168 ± 0.0006	9.400E−39	0.8827 ± 0.0008	9.640E−20	0.8688 ± 0.0008	1.720E−24	0.9111 ± 0.0006	9.426E−32
GBLUP	0.4957 ± 0.0030	1.597E−114	0.4390 ± 0.0032	1.090E−112	0.6311 ± 0.0023	2.207E−112	0.5337 ± 0.0025	4.853E−118	0.5459 ± 0.0028	7.565E−109	0.5596 ± 0.0028	3.421E−110

ELPGV is the ensemble learning based on BayesA, BayesB, BayesCπ and GBLUP

— Represents no explicit result was found in this method

type 1 diabetes (T1D), type 2 diabetes (T2D), bipolar disorder (BD), rheumatoid arthritis (RA), coronary artery disease (CAD) and hypertension (HT)

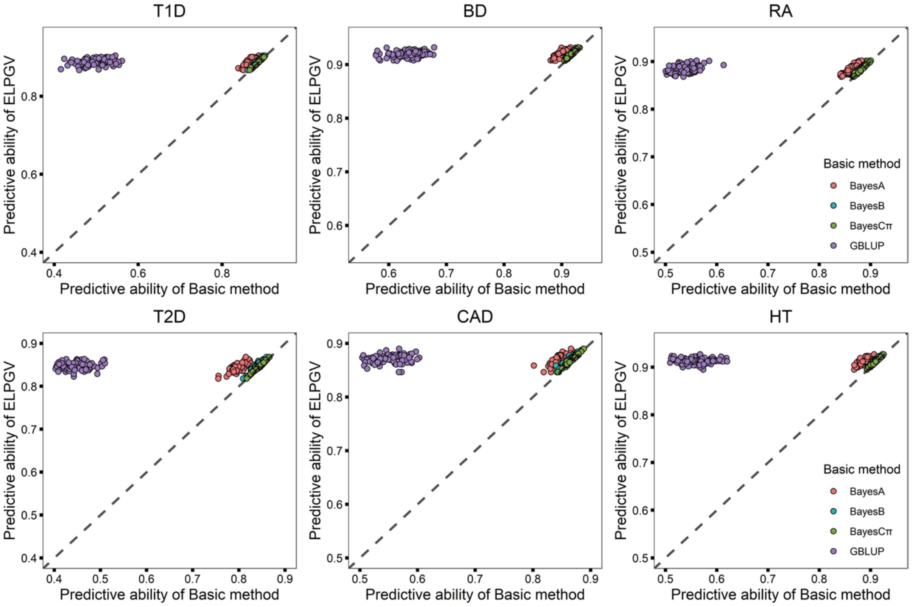

Fig. 2 Comparison of the predictive ability of ELPGV and the basic method. **a** T1D, **b** BD, **c** RA, **d** T2D, **e** CAD and **f** HT with WTCCC dataset; different method is denoted with different color, each dot represents single experiment

Table 4 The predictive ability of four basic methods and ELPGV, and the comparison *p*-value between ELPGV and others in CD, UC, IBD with IBDGC dataset

Method	CD		UC		IBD	
	Predictive ability	*p*-value	Predictive ability	*p*-value	Predictive ability	*p*-value
ELPGV	0.4516 ± 0.0065	—	0.7920 ± 0.0045	—	0.4253 ± 0.0071	—
BayesA	0.4359 ± 0.0067	1.058E−15	0.7817 ± 0.0045	7.365E−25	0.4128 ± 0.0075	2.470 E−13
BayesB	0.4338 ± 0.0063	4.932E−13	0.7831 ± 0.0047	4.845E−16	0.4052 ± 0.0067	1.911E−14
BayesCπ	0.4452 ± 0.0063	3.938E−07	0.7845 ± 0.0047	3.878E−13	0.4171 ± 0.0068	5.272E−10
GBLUP	0.3692 ± 0.0063	3.659E−34	0.6687 ± 0.0052	3.314E−56	0.3455 ± 0.0079	9.965E−35

ELPGV is the ensemble learning based on BayesA, BayesB, BayesCπ and GBLUP

— Represents no explicit result was found in this method

Crohn disease (CD) and ulcerative colitis disease (UC)

and BayesCπ of UC was 0.6687, 0.7817, 0.7831 and 0.7845, respectively. After assembled with ELPGV, the averaged predictive ability was 0.7920, significantly higher than four basic methods, the *p*-values were from ranged from 3.314E−56 to 3.878E−13 (Table 4). Similarly, the prediction abilities of CD of four basic methods were ranged from 0.3692 (GBLUP) to 0.4452 (BayesCπ), after assembled with ELPGV, the predictive ability was increased to 0.4516, significantly higher than four basic methods (*p*-value varied from 3.659E−34 to 3.938E−07, Table 4). We also show the comparison of each experiment individually, for vast majority of individual experiment, ELPGV outperformed four basic methods, among them, GBLUP performed the lower predictive ability (Fig. 3a–c).

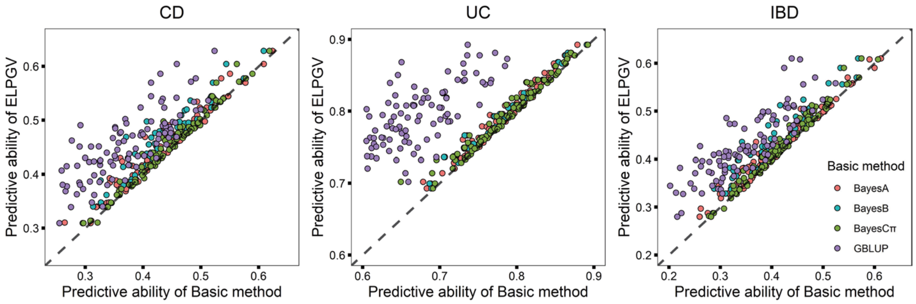

Fig. 3 Comparison of the predictive ability of ELPGV and the basic method. **a** CD, **b** UC and **c** IBD with IBDGC dataset; different method is denoted with different color, each dot represents single experiment

Cattle dataset

We further validated ELPGV with a cattle dataset of German Holstein, in which milk fat percent (mfp), milk yield (my) and somatic cell score (scs) were investigated. For genetic prediction of mfp, BayesCπ performed the highest predictive ability among four basic methods ($r=0.8632$), whereas GBLUP performed the lowest ($r=0.8259$) (Table 5). After assembled of four basic methods with ELPGV, the predictive ability was 0.8748, significantly higher than any basic methods (the comparison *p*-values ranged from 9.943E−80 to 2.356E−10, Table 5). The individual experiment showed that for vast majority of the predictions, ELPGV was obviously more accurate than four basic methods, especially than GBLUP (Fig. 4b). For my, ELPGV also outperformed the four basic methods

Table 5 The predictive ability of four basic methods and ELPGV, and the comparison *p*-value between ELPGV and others in mfp, my, scs with cattle dataset

Method	mfp		my		scs	
	Predictive ability	*p*-value	Predictive ability	*p*-value	Predictive ability	*p*-value
ELPGV	0.8748 ± 0.0009	—	0.7959 ± 0.0016	—	0.7523 ± 0.0019	—
BayesA	0.8713 ± 0.0010	2.665E−31	0.7935 ± 0.0017	5.726E−19	0.7496 ± 0.0019	1.242E−23
BayesB	0.8739 ± 0.0009	2.356 E−10	0.7948 ± 0.0017	1.335E−07	0.7503 ± 0.0020	5.614E−11
BayesCπ	0.8632 ± 0.0010	5.884E−52	0.7928 ± 0.0017	1.026E−26	0.7518 ± 0.0019	0.001E−00
GBLUP	0.8259 ± 0.0013	9.943E−80	0.7809 ± 0.0017	5.133E−52	0.7482 ± 0.0019	3.801E−29

ELPGV is the ensemble learning based on BayesA, BayesB, BayesCπ and GBLUP

— Represents no explicit result was found in this method

mfp, milk fat percentage; my, milk yield; scs, somatic cell score

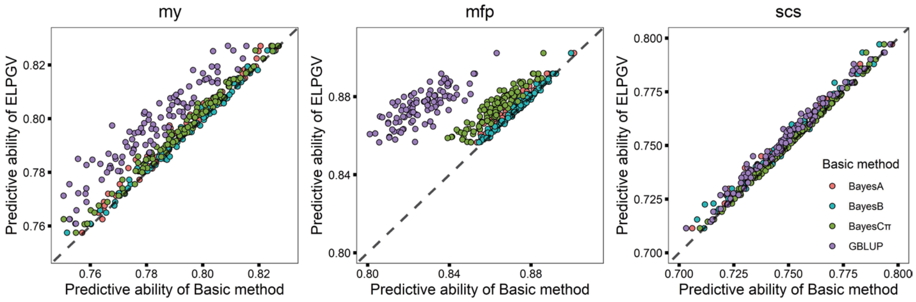

Fig. 4 Comparison of the predictive ability of ELPGV and the basic methods. **a** my, **b** mfp and **c** scs with cattle dataset; different method is denoted with different color each dot represents single experiments

(Fig. 4a) with the comparison *p*-values ranged from 5.133E−52 (GBLUP) to 1.335E−07 (BayesB). For scs, the advantage of ELPGV over four basic methods was also significant, and the *p*-values were ranged from 3.801E−29 (GBLUP) to 0.001 (BayesCπ) (Table 5). Figure 4 shows the accuracies of ELPGV and four basic methods in 100 individual experiments, which displays that for large proportion of predictions, ELPGV has higher prediction abilities than those of four basic methods for all the investigated traits.

Wheat dataset

Wheat yields measured under four places were investigated, which includes 599 individuals genotyped for 1279 SNPs. The averaged predictive ability across 100 cross validations of each place is shown in Table 6. For the first place, the prediction abilities of GBLUP, BayesA, BayesB and BayesCπ were 0.5251, 0.5231, 0.5080, and 0.5215, respectively, and the predictive ability of ELPGV was 0.5273, which was significantly higher than four basic methods (the comparison *p*-values ranged from 7.965E−19 to 2.297E−04 (Table 6). For other three places, the results also showed that the prediction accuracy of ELPGV was consistently higher than four basic methods (Table 6). All the predictions of 100 cross-validation are shown in Fig. 5a–d, and ELPGV outperforms four basic methods for majority of single experiments in four places.

Simulations

We finally performed simulation studies to further investigate the performance of ELPGV. For each group, 100 simulated datasets were generated. Each dataset was randomly divided into 5 parts evenly, and 4 of them were taken as train set and the left 1 part was taken as test set. We first ran four basic methods including GBLUP, BayesA, BayesB and BayesCπ; then assembled the predictions with ELPGV to produce new predictions. For all simulations, ELPGV performed significant higher prediction abilities than corresponding four basic methods, the comparison *p*-values were ranged from 3.553E−34 to 0.001E−00 (Table 7). The 100 replicated experiments also obviously revealed that for each of experiments, the prediction of ELPGV was more accurate than other basic methods (Fig. 6a–d) and the gain of ELPGV over GBLUP was more obvious when QTL number was 5 than 1,000.

We next investigated the effect of sample size of training set. We randomly sampled 100, 200, 300, 400, 500 and 599 individuals from wheat data, respectively, the QTL

Table 6 The predictive ability of four basic methods and ELPGV, and the comparison *p*-value between ELPGV and others in GY with wheat dataset

Method	The first place		The second place		The third place		The fourth place	
	Predictive ability	*p*-value	Predictive ability	*p*-value	Predictive ability	*p*-value	Predictive ability	*p*-value
ELPGV	0.5273 ± 0.0104	—	0.5092 ± 0.0101	—	0.4050 ± 0.0102	—	0.4722 ± 0.0101	—
BayesA	0.5231 ± 0.0105	1.550E−10	0.5057 ± 0.0101	1.238E−07	0.3953 ± 0.0103	3.711E−15	0.4676 ± 0.0102	7.265E−14
BayesB	0.5080 ± 0.0104	2.296E−17	0.4947 ± 0.0104	5.993E−12	0.3914 ± 0.0101	4.098E−06	0.4542 ± 0.0101	1.261E−18
BayesCπ	0.5215 ± 0.0105	7.965E−19	0.5046 ± 0.0101	6.695E−11	0.3947 ± 0.0102	2.264E−17	0.4672 ± 0.0102	1.200E−13
GBLUP	0.5251 ± 0.0105	2.297E−04	0.5058 ± 0.0100	7.877E−07	0.3954 ± 0.0104	2.938E−12	0.4698 ± 0.0102	5.013E−05

ELPGV is the ensemble learning based on BayesA, BayesB, BayesCπ and GBLUP

— Represents no explicit result was found in this method

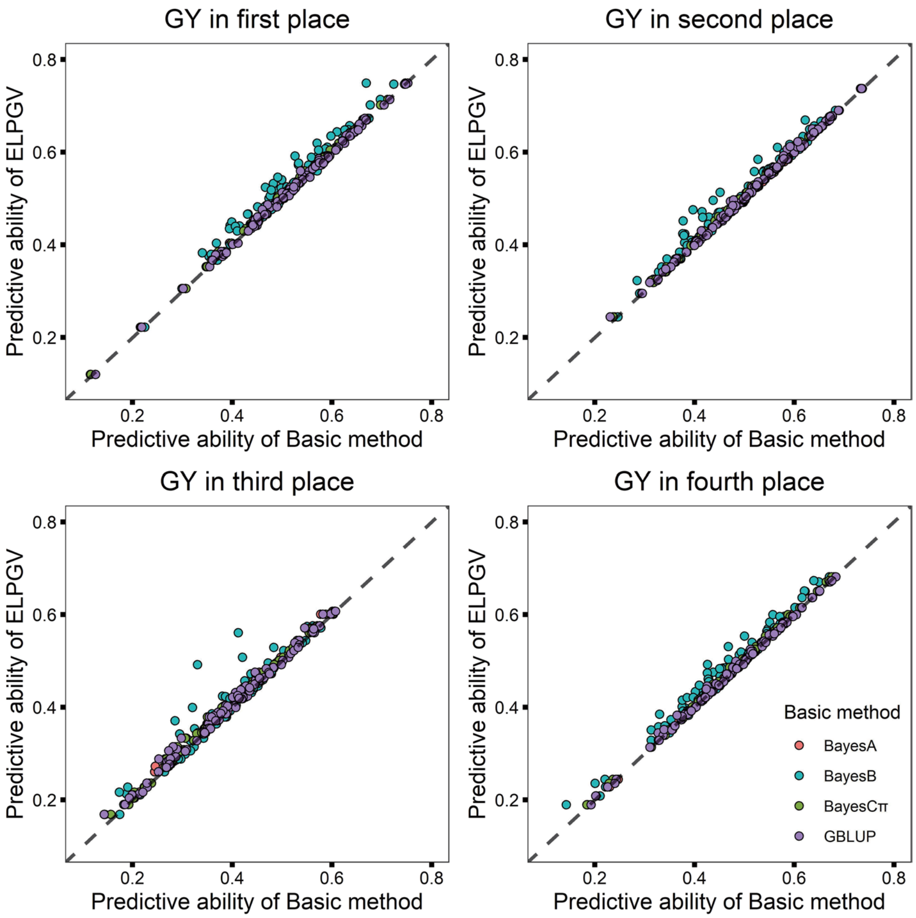

Fig. 5 Comparison of the predictive ability of ELPGV and the basic method. **a**–**d** GY (grain yield) under four places of CIMMYT wheat dataset; different method is denoted with different color; each dot represents single experiment

Table 7 The averaged predictive ability across 100 replications for different methods in 4 scenes of simulation

Method	0.2				0.5			
	5qtl		1000qtl		5qtl		1000qtl	
	Predictive ability	*p*-value	Predictive ability	*p*-value	Predictive ability	*p*-value	Predictive ability	*p*-value
ELPGV	0.4346 ± 0.0116	—	0.3079 ± 0.0109	—	0.7034 ± 0.0064	—	0.5851 ± 0.0085	—
BayesA	0.4042 ± 0.0122	3.317E−15	0.2961 ± 0.0110	1.565E−08	0.6865 ± 0.0068	1.057E−16	0.5765 ± 0.0085	2.652E−17
BayesB	0.4257 ± 0.0118	0.001E−00	0.2842 ± 0.0112	1.096E−08	0.6985 ± 0.0065	0.002E−00	0.5756 ± 0.0086	6.910E−09
BayesCπ	0.3741 ± 0.0140	4.577E−11	0.2956 ± 0.0108	2.828E−10	0.6955 ± 0.0064	1.712E−12	0.5768 ± 0.0086	2.048E−16
GBLUP	0.2952 ± 0.0144	1.032E−22	0.2959 ± 0.0110	6.161E−07	0.5505 ± 0.0096	3.553E−34	0.5727 ± 0.0084	7.160E−13

ELPGV is the ensemble learning based on BayesA, BayesB, BayesCπ and GBLUP

— Represents no explicit result was found in this method

number was set as 5 and the heritability was 0.5. For each of sample sizes, 100 independent datasets were generated. The cross validation was used to evaluate the prediction abilities. It revealed that the prediction abilities of ELPGV were higher than four

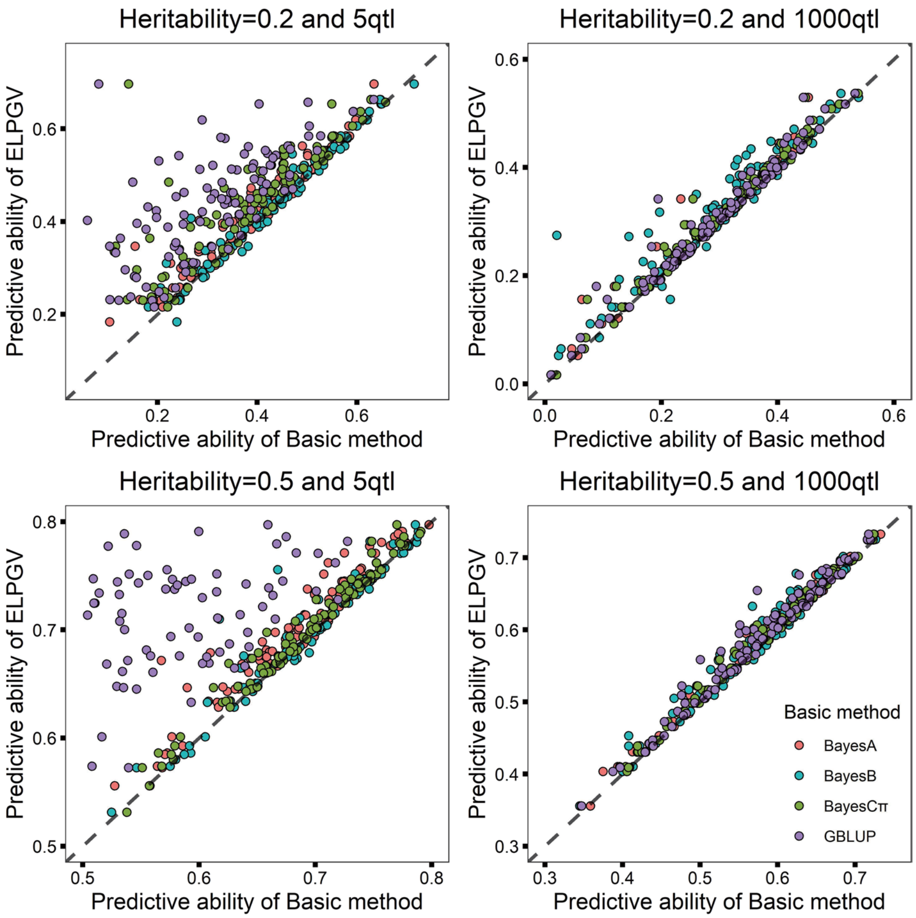

Fig. 6 Comparison of the predictive ability of ELPGV and the basic method in simulation. **a** 5 QTL with heritiability 0.2; **b** 1000 QTL with heritiability 0.2; **c** 5 QTL with heritiability 0.5 and **d** 1000 QTL with heritiability 0.5. Different method is denoted with different color, and each dot represents single experiment

Table 8 The maximum and minimum difference of the predictive ability between ELPGV and other methods in different sample size of simulation

Method	100	200	300	400	500	599
	Predictive ability	Predictive ability	Predictive ability	Predictive ability	Predictive ability	Predictive ability
ELPGV	0.6055 ± 0.0220	0.7175 ± 0.0159	0.7559 ± 0.0085	0.7189 ± 0.0070	0.7378 ± 0.0061	0.7034 ± 0.0064
BayesA	0.5486 ± 0.0256	0.6827 ± 0.0170	0.7200 ± 0.0085	0.7034 ± 0.0073	0.7231 ± 0.0069	0.6865 ± 0.0068
BayesB	0.6007 ± 0.0237	0.7075 ± 0.0170	0.7502 ± 0.0084	0.7158 ± 0.0069	0.7344 ± 0.0060	0.6985 ± 0.0065
BayesCπ	0.5334 ± 0.0215	0.6215 ± 0.0266	0.7271 ± 0.0100	0.7069 ± 0.0076	0.7354 ± 0.0060	0.6955 ± 0.0064
GBLUP	0.5266 ± 0.0216	0.4785 ± 0.0243	0.5655 ± 0.0151	0.5511 ± 0.0143	0.5934 ± 0.0095	0.5505 ± 0.0096
Maximum difference	0.0789 (15.0%)	0.2390 (49.9%)	0.1904 (33.7%)	0.1678 (30.4%)	0.1444 (24.3%)	0.1529 (27.7%)
Minimum difference	0.0048 (0.8%)	0.0100 (1.4%)	0.0057 (0.8%)	0.0031 (0.4%)	0.0024 (0.3%)	0.0049 (0.7%)

ELPGV is the ensemble learning based on BayesA, BayesB, BayesCπ and GBLUP

basic methods for all simulated sample size (Table 8). We next investigated if the advantage of ELPGV over other methods was dependent on the sample size. To do this, we summarized the maximum and minimum difference of the prediction abilities between

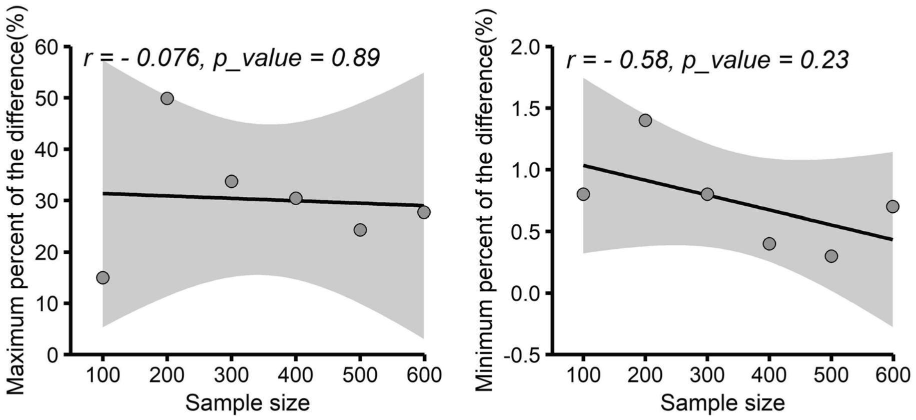

Fig. 7 The relationship between sample size and maximum difference or minimum difference of the methods

ELPGV and other methods, respectively (Table 8) and correlated the maximum (Fig. 7a) or minimum (Fig. 7b) differences to the corresponding sample sizes. But we did not find evidence of significant correlation ($r = -$ 0.58 and $-$ 0.076 with p-value 0.23 and 0.89), which implies that the gain of ELPGV over basic methods is not affected by sample size.

Discussion

We have presented an ensemble learning method, ELPGV to predict genetic values. The key feature of ELPGV is that it assembles predictions of other basic methods into more accurate predictions. Extensive datasets of human, cattle and wheat have been employed to validate the performance of ELPGV, all results consistently revealed that ELPGV was able to integrate the merit of each method together to produce significantly higher predictive ability than any basic methods. Based on these advantages, ELPGV is expected to be widely used for prediction in large data sets.

Ensemble learning has been widely utilized in genome selection, such as Ma et al. [36], who assembles two basic methods and trains the weights with PSO algorithm; however, it has several disadvantages, (1) it assumes the phenotypes of testing individuals have been known, so that it is only applicable for prediction with known phenotype, which is less meaningful in practice; (2) the performance of the traditional PSO greatly depends on its parameters, and it often suffers from being trapped in local optima [37, 38], which is consistent with the study of Cai et al. [39]. Liang et al. [40] construct a stacking ensemble learning framework (SELF), integrating three machine learning methods and an ordinary least square regression was chosen as the meta learner, to improve the genomic predictions. A lot of experiment indicated that SELF with the great potential to improve genomic predictions in other animal and plant populations. In actual analysis, SELF taken the genomic relationship matrix derived by genotypes as the inputs directly. But this might reduce the prediction accuracy of a single basic method. Additionally, Gianola et al. [41] was found that bagging can ameliorate predictive performance of GBLUP and make it more robust against over-fitting. However, because of predictive ability increases with training set size [42]. It is obvious that bagging may not be feasible for immense data sets.

DE algorithm is another kind of evolutionary algorithm, which has been applied to a series of problems arising in various fields of science, engineering, and management [43–45]. In our analysis, we found that DE algorithm is much more stable, and always converge to the same solution after repeated operations; furthermore, DE converges fast and is very accurate for high-dimensional problems, which has three main parameters (initialize solution size, scaling factor F, crossover probability CR), but it is not sensitive for parameter setups [46]. While DE algorithm has many advantages, the disadvantage of it is that it is difficult to update model parameters [46], but PSO does not have this problem. So, hybridization is an important modification in DE which is implemented to enhance its performance and convergence speed. Plenty of work can be found in the literature on the hybridization of DE. For instance, Pant et al. [47] proposed a hybrid version of DE with PSO and results show that the proposed DE-PSO is quite competent for solving the considered test functions as well as real-life problems. Zhang et al. [48] proposed a hybrid technique using DE with PSO for unconstrained optimization problems. Similarly, ELPGV is the hybrid of DE and PSO too, which not only inherits the high precision merit of DE algorithm, but also possesses the fast convergence character of PSO algorithm.

In the prediction of the disease risk for human, ELPGV exhibits greater advantages over four basic methods. In almost all of situations, ELPGV is more accurate than others, the gain is much more obvious when comparing with GBLUP, reflecting that GBLUP is not very suitable for human dataset, may be due to the fact that the relationships between individuals are quite limited and few information is available for GBLUP predictions. In contrast, the situation is quite different for cattle and wheat datasets. The reason may be that the aim of these datasets is for selection breeding and the individuals have extensive relationship, which is consistent with the literature [49]. Additionally, Heslot et al. [50], Azodi et al. [51] and Schrauf et al. [52] also compared GBLUP (or equivalent models) with other genomic prediction methods in a variety of plant datasets and have shown that the difference between GBLUP and other methods is negligible under large data sizes and polygenic architectures. Because the GBLUP efficiently predicts individual genetic values using the relationship information, and all markers are assumed in a sense to contribute equally to the construction of Kinship matrix.

It is shown that the performance of ELPGV is greatly affected by the method similarity, which is consistent with Granitto et al. [53] who concludes diverse basic methods is an essential characteristic of a good ensemble method. Therefore, one way to improve the performance of ELPGV is to increase the diversity of basic methods. For example, BayesB, BayesCπ and BayesR [54] are working well for major-effect QTL method, they often performed similar prediction abilities, so integrating them would not enhance the predictive ability of ELPGV too much; similarly, rrBLUP [55] is theoretically quite similar to GBLUP, both are based on polygenic method, it would not substantially increase the predictive ability by integrating them together.

We have proposed ELPGV method for optimizing the parameters, which greatly improves the precise of parameter estimates. It's versatility to allow for different and more complex criterion to be maximized. However, it still has room to improve, for example, combining DE or PSO with other optimization algorithms to form a better hybrid algorithm [46], or using other ensemble strategies, such as sequence integration methods such as boosting method [56].

Conclusions

We have presented an ensemble learning method, ELPGV, to predict genetic values. The key feature of ELPGV is that it assembles predictions of other basic methods into more accurate predictions. ELPGV is able to integrate the merit of each method together to produce significantly higher predictive ability than any basic methods and it is simple to implement, which uses only the predictions of basic methods as input without using genotype data. Therefore, ELPGV requires quite few computers RAM and can complete task even with PC computer; furthermore, ELPGV is computationally fast, which takes only several minutes to complete the assembling for tens thousands of individuals and is promising for wide application in genetic predictions.

Acknowledgements
Not applicable.

Author contributions
FM, JD and WZY designed and directed the work. Most of data analysis (e.g., writing the program of ELPGV, simulation and real data research) were conducted by GLL. WZY and JD corrected the grammar and format of the article. YRQ contributed to real dataset analysis. The original draft of the manuscript was written by GLL, JD and FM, and every author has reviewed the work.

Funding
The research was supported by the National Key Research & Development Program of China (2022YFD2401001), the Seed Industry Innovation and Industrialization Project of Fujian Province (2021FJSCZY01), the Natural Science Foundation of Fujian Province of China (2020J01672 and 2021J02045), China Agriculture Research System (CARS-47-G04).

Availability of data and materials
The ELPGV scripts are available in the release package on github (https://github.com/GuLinLin/ELPGV). The detailed user manual is available from https://github.com/GuLinLin/ELPGV/blob/main/ELPGV_User_Manual.pdf. This study makes use of data generated by the Wellcome Trust Case Control Consortium (WTCCC). A full list of the investigators who contributed to the generation of the WTCCC data is available from http://www.wtccc.org.uk/, where the dataset can be publicly accessed. Funding for the WTCCC project was provided by the Wellcome Trust under award 076113 and 085475. The following datasets can be directly downloaded from the links below: Cattle: https://academic.oup.com/g3journal/article/5/4/615/6025251. Wheat: http://cran.r-project.org/web/packages/BLR/index.html. The IBD dataset of this study are available from the international IBD Genetics Consortium (https://www.ibdgeneticsa.org/) by reasonable application.

Declarations

Ethics approval and consent to participate
Ethics approval was not required for this study.

Consent for publication
Not applicable.

Competing interests
The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 22 November 2022 Accepted: 26 February 2024
Published online: 21 March 2024

References

1. Lello L, Avery SG, Tellier L, Vazquez AI, de los Campos G, Hsu SDH. Accurate genomic prediction of human height. Genetics. 2018;210:477–97.
2. Yin L, Zhang H, Zhou X, Yuan X, Zhao S, Li X, et al. KAML: improving genomic prediction accuracy of complex traits using machine learning determined parameters. Genome Biol. 2020;21:146.
3. Schaeffer LR. Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet. 2006;123:218–23.
4. Desta ZA, Ortiz R. Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci. 2014;19:592–601.
5. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.

6. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.
7. Yi N, Xu S. Bayesian LASSO for quantitative trait loci mapping. Genetics. 2008;179:1045–55.
8. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the bayesian alphabet for genomic selection. BMC Bioinform. 2011;12:186.
9. The International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460:748–52.
10. Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park J-H. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat Genet. 2013;45:400–5.
11. Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLoS Genet. 2013;9:e1003348.
12. Privé F, Vilhjálmsson BJ, Aschard H, Blum MGB. Making the most of clumping and thresholding for polygenic scores. Am J Hum Genet. 2019;105:1213–21.
13. Okser S, Pahikkala T, Airola A, Salakoski T, Ripatti S, Aittokallio T. Regularized machine learning in the genetic prediction of complex traits. PLoS Genet. 2014;10:e1004754.
14. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. Am J Hum Genet. 2019;104:21–34.
15. Dietterich TG. Ensemble methods in machine learning. In: Multiple classifier systems. Berlin: Springer; 2000. pp. 1–15.
16. Hansen LK, Salamon P. Neural network ensembles. IEEE Trans Pattern Anal Mach Intell. 1990;12:993–1001.
17. Ju C, Bibaut A, van der Laan M. The relative performance of ensemble methods with deep convolutional neural networks for image classification. J Appl Stat. 2018;45:2800–18.
18. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. Next-generation machine learning for biological networks. Cell. 2018;173:1581–92.
19. Cao Z, Pan X, Yang Y, Huang Y, Shen H-B. The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier. Bioinformatics. 2018;34:2185–94.
20. Zhang S, Hu H, Jiang T, Zhang L, Zeng J. TITER: predicting translation initiation sites by deep learning. Bioinformatics. 2017;33:i234–42.
21. Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics. 2010;26:392–8.
22. Pusztai L, Hatzis C, Andre F. Reproducibility of research and preclinical validation: problems and solutions. Nat Rev Clin Oncol. 2013;10:720–4.
23. Cao Y, Geddes TA, Yang JYH, Yang P. Ensemble deep learning in bioinformatics. Nat Mach Intell. 2020;2:500–8.
24. Storn R, Price K. Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. J Global Optim. 1997;11:341–59.
25. Kennedy J, Eberhart R. Particle swarm optimization. In: Proceedings of ICNN'95—international conference on neural networks, vol 4. 1995. pp. 1942–8.
26. Liu G, Dong L, Gu L, Han Z, Zhang W, Fang M, et al. Evaluation of genomic selection for seven economic traits in yellow drum (Nibea albiflora). Mar Biotechnol. 2019;21:806–12.
27. Pérez P, de los Campos G. Genome-wide regression and prediction with the BGLR statistical package. Genetics. 2014;198:483–95.
28. Deniz A, Godfrey OU. EMMREML: fitting mixed models with known covariance structures. R package version 3.1. https://CRAN.R-project.org/package=EMMREML. 2015.
29. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447:661–78.
30. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
31. Huang H, Fang M, Jostins L, Umićević Mirkov M, Boucher G, Anderson CA, et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature. 2017;547:173–8.
32. Zhang Z, Erbe M, He J, Ober U, Gao N, Zhang H, et al. Accuracy of whole-genome prediction using a genetic architecture-enhanced variance-covariance matrix. G3 Genes Genom Genet. 2015;5:615–27.
33. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, et al. Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE. 2009;4:e5350.
34. Crossa J, de los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, et al. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics. 2010;186:713–24.
35. Gianola D, Okut H, Weigel KA, Rosa GJ. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genet. 2011;12:87.
36. Ma W, Qiu Z, Song J, Li J, Cheng Q, Zhai J, et al. A deep convolutional neural network approach for predicting phenotypes from genotypes. Planta. 2018;248:1307–18.
37. Angeline PJ. Evolutionary optimization versus particle swarm optimization: philosophy and performance differences. In: Porto VW, Saravanan N, Waagen D, Eiben AE, editors. Evolutionary programming VII. Berlin: Springer; 1998. p. 601–10.
38. Liu B, Wang L, Jin Y-H, Tang F, Huang D-X. Improved particle swarm optimization combined with chaos. Chaos Solitons Fractals. 2005;25:1261–71.
39. Cai J, Ma X, Li L, Haipeng P. Chaotic particle swarm optimization for economic dispatch considering the generator constraints. Energy Convers Manag. 2007;48:645–53.
40. Liang M, Chang T, An B, Duan X, Du L, Wang X, et al. A stacking ensemble learning framework for genomic prediction. Front Genet. 2021;12:600040.
41. Gianola D, Weigel KA, Krämer N, Stella A, Schön C-C. Enhancing genome-enabled prediction by bagging genomic BLUP. PLoS ONE. 2014;9:e91693.
42. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE. 2008;3:e3395.

43. Dragoi E-N, Curteanu S, Galaction A-I, Cascaval D. Optimization methodology based on neural networks and self-adaptive differential evolution algorithm applied to an aerobic fermentation process. Appl Soft Comput. 2013;13:222–38.
44. Arya R, Choube SC. Differential evolution based technique for reliability design of meshed electrical distribution systems. Int J Electr Power Energy Syst. 2013;48:10–20.
45. Li Y, Wang Y, Li B. A hybrid artificial bee colony assisted differential evolution algorithm for optimal reactive power flow. Int J Electr Power Energy Syst. 2013;52:25–33.
46. Bilal S, Pant M, Zaheer H, Garcia-Hernandez L, Abraham A. Differential evolution: a review of more than two decades of research. Eng Appl Artif Intell. 2020;90:103479.
47. Pant M, Thangaraj R, Grosan C, Abraham A. Hybrid differential evolution-particle swarm optimization algorithm for solving global optimization problems. In: 2008 Third international conference on digital information management. London: IEEE; 2008. pp. 18–24.
48. Zhang C, Ning J, Lu S, Ouyang D, Ding T. A novel hybrid differential evolution and particle swarm optimization algorithm for unconstrained optimization. Oper Res Lett. 2009;37:117–22.
49. Spiliopoulou A, Nagy R, Bermingham ML, Huffman JE, Hayward C, Vitart V, et al. Genomic prediction of complex human traits: relatedness, trait architecture and predictive meta-models. Hum Mol Genet. 2015;24:4167–82.
50. Heslot N, Yang H-P, Sorrells ME, Jannink J-L. Genomic selection in plant breeding: a comparison of models. Crop Sci. 2012;52:146–60.
51. Azodi CB, Bolger E, McCarren A, Roantree M, de los Campos G, Shiu S-H. Benchmarking parametric and machine learning models for genomic prediction of complex traits. G3 Genes Genom Genet. 2019;9:3691–702.
52. Schrauf MF, de los Campos G, Munilla S. Comparing genomic prediction models by means of cross validation. Front Plant Sci. 2021;12:734512.
53. Granitto PM, Verdes PF, Ceccatto HA. Neural network ensembles: evaluation of aggregation algorithms. Artif Intell. 2005;163:139–62.
54. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. PLoS Genet. 2015;11:e1004969.
55. Whittaker JC, Thompson R, Denham MC. Marker-assisted selection using ridge regression. Genet Res. 2000;75:249–52.
56. Bartlett P, Freund Y, Lee WS, Schapire RE. Boosting the margin: a new explanation for the effectiveness of voting methods. Ann Statist. 1998;26:1651–86.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.