Plant Biotechnology Journal

Open Access

doi: 10.1111/pbi.14384

Brief Communication

DeepCCR: large-scale genomics-based deep learning method for improving rice breeding

Xiaoding Ma[1,†], Hao Wang[1,†], Shengyang Wu[2,†], Bing Han[1], Di Cui[1], Jin Liu[3], Qiang Zhang[4], Xiuzhong Xia[5], Peng Song[6], Cuifeng Tang[7], Leiyue Geng[8], Yaolong Yang[9], Shen Yan[1,]*, Kunneng Zhou[10,]* and Longzhi Han[1,]*

[1] *State Key Laboratory of Crop Gene Resources and Breeding, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China*
[2] *Bokaisen Biotechnology Ltd., Beijing, China*
[3] *Jiangxi Research Center of Crop Germplasm Resources, Rice Research Institute, Jiangxi Academy of Agricultural Sciences, Nanchang, China*
[4] *Jilin Provincial Laboratory of Crop Germplasm Resources, Rice Research Institute, Jilin Academy of Agricultural Sciences, Changchun, China*
[5] *Guangxi Key Laboratory of Rice Genetics and Breeding, Rice Research Institute, Guangxi Academy of Agricultural Sciences, Nanning, China*
[6] *National Key Laboratory of Crop Genetic Improvement, College of Plant Science & Technology, Huazhong Agricultural University, Wuhan, China*
[7] *Key Lab of Southwestern Crop Gene Resources and Germplasm Innovation, Biotechnology and Germplasm Resources Institute, Kunming, China*
[8] *Institute of Coastal Agriculture, Hebei Academy of Agriculture and Forestry Sciences, Tangshan, China*
[9] *State Key Laboratory of Rice Biology and Breeding, China National Rice Research Institute, Hangzhou, China*
[10] *Anhui Province Key Laboratory of Rice Germplasm Innovation and Molecular Improvement, Rice Research Institute, Anhui Academy of Agricultural Sciences, Hefei, China*

Received 20 February 2024;
revised 31 March 2024;
accepted 7 May 2024.

**Correspondence* (Tel +86-10-62176784; fax +86-10-62156875; email hanlongzhi@caas.cn (L.H.); Tel +86-551-62160188; fax +86-551-62160125; email zhoukunneng1986@163.com (K.Z.); Tel +86-10-62186693; fax +86-10-62156875; email yanshen@caas.cn (S.Y.))
[†] These authors contributed equally to this work.

Keywords: genomic selection, convolutional neural network, Chinese cultivated rice population, yield trait.

Rice is the staple food crop for half of the world's population. Traditional phenotype-based and marker-assisted selection methods have been used in rice improvement, but they are time-consuming, costly and labor-intensive. Therefore, research and implementation of novel breeding strategies to improve rice yield is a high priority. Genomic selection (GS) has paved the way for overcoming these limitations (Yu *et al*., 2016). The major factor in the effective application of GS breeding models is the construction of a large-scale training population with genomic diversity covering the target selection materials (Fu *et al*., 2022). However, the practical implementation of the general population in applied rice breeding programmes is still at a nascent stage, and a comprehensive assessment of genomic predictability for various traits has likewise not yet been undertaken.

To construct a generally representative training population, we compiled the first Chinese cultivated rice population (CCRP), which consisted of 4015 rice accessions from 25 Chinese provinces covering five major rice-growing regions that account for more than 99% of the total annual rice-growing area in China (Figure 1a; Tables S1 and S2). These accessions included 1943 indica and 2072 japonica rice accessions, more than 96% of which were cultivars and breeding lines (Figure 1b; Tables S1 and S2). Cluster analysis revealed that CCRP was quite different from the 3 K population (Figure S1) (Wang *et al*., 2018); we believe that CCRP represents the characteristics and genetic diversity of rice varieties from almost all rice-growing regions in China (Figure 1c,d). To accurately and systematically investigate the phenotypes of CCRP, we selected seven representative sites (Nanning city (NN), Guangxi Province; Wuhan city (WH), Hubei Province; Nanchang city (NC), Jiangxi Province; Hefei city (HF), Anhui Province; Kunming city (KM), Yunnan Province; Tanghai city (TH), Hebei Province; and Gongzhuling city (GZL), Jilin Province) in five rice-growing regions in China for two consecutive years (Figure 1e). Yield traits have consistently been a primary focus in rice breeding. Hence, the key traits of interest in this study included heading date (HD), plant height (PH), panicle length (PL), tiller number (TN), grain per panicle (GP), seed set rate (SST), grain length (GL), grain width (GW), thousand-grain weight (TGW) and yield (Y) (Figures S2 and S3), and we gathered phenotypic data over two consecutive years to assess repeatability and rectify systematic biases within the data set (Figure 1f,g; Figure S3).

To meet the need for genome prediction in rice breeding, we resequenced 4015 accessions (Figure 1h–k; Figure S4) and proposed DeepCCR, a deep learning method based on a convolutional neural network combined with bidirectional long short-term memory, to predict phenotypic values at different planting sites (Figure 1l,m). To evaluate the predictive performance of DeepCCR, we compared it to four state-of-the-art methods (XGBoost, LightGBM, DNNGP, and GBLUP) at seven sites. The results of the 10-fold cross-validation show that DeepCCR achieved the best performance among all compared methods. Specifically, at the GZL site, the prediction accuracies of DeepCCR for the rice traits Y, HD, PH, PL, TN, GP, SSR, GL, GW and TGW are 79.7%, 67.5%, 75.3%, 72.5%, 66.9%, 77.0%, 73.2%, 70.6%, 64.3% and 74.0%, respectively. DeepCCR outperforms the runner-up model by 17.2%, 11.7%, 19.9%, 12.8%, 9.6%, 12.6%, 6.6%, 12.8%,

Please cite this article as: Ma, X., Wang, H., Wu, S., Han, B., Cui, D., Liu, J., Zhang, Q., Xia, X., Song, P., Tang, C., Geng, L., Yang, Y., Yan, S., Zhou, K. and Han, L. (2024) DeepCCR: large-scale genomics-based deep learning method for improving rice breeding. *Plant Biotechnol. J.*, https://doi.org/10.1111/pbi.14384.

10.3% and 12.6%, respectively (Figure S5a). Furthermore, DeepCCR also exhibited superior performance in mean square error (MSE) compared with existing advanced methods (Figure S5b). The computational time results demonstrated that the computational efficiency of DeepCCR is comparable to other models (Figure S6).

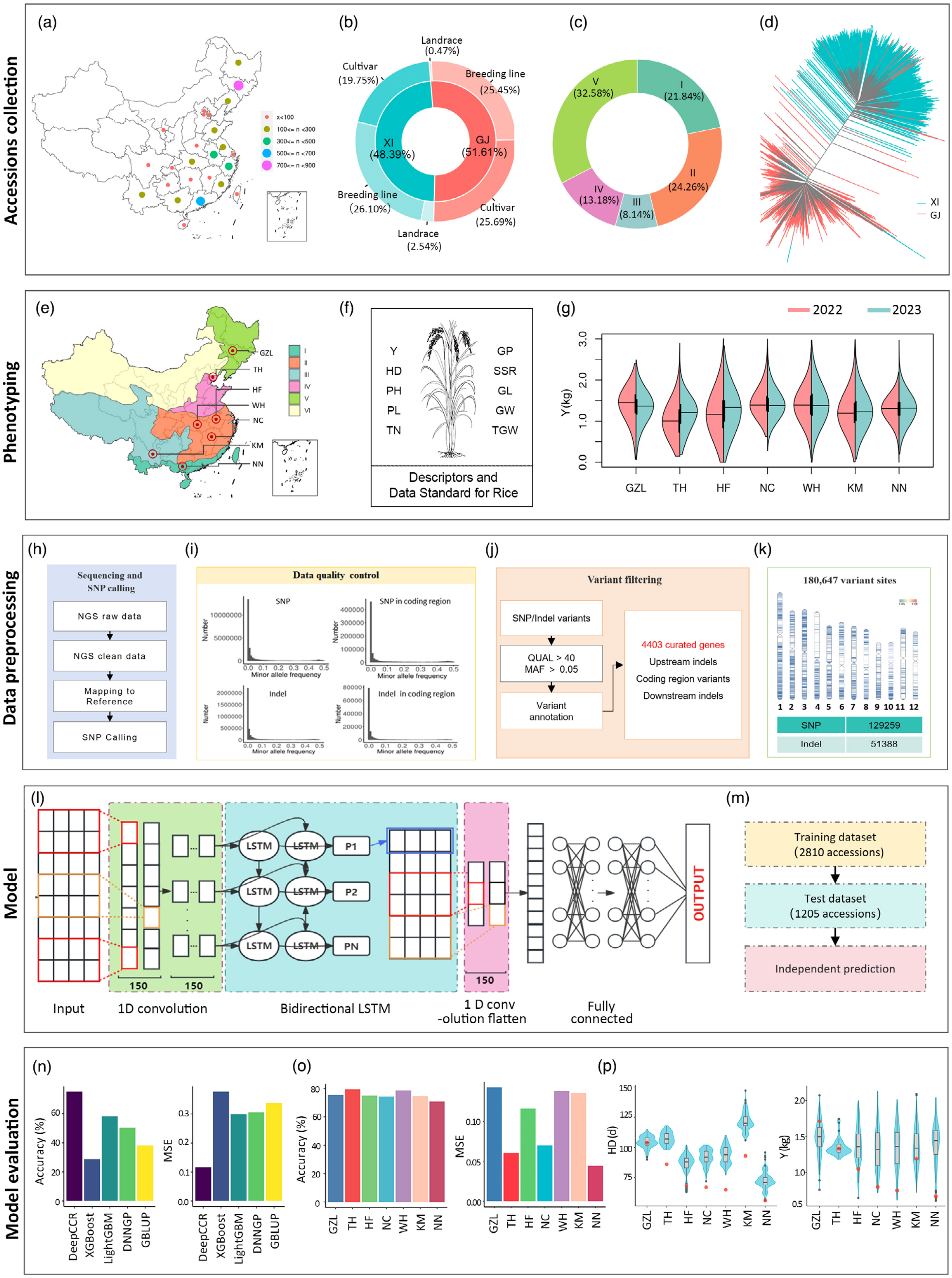

Figure 1 The proposed DeepCCR framework. (a) Geographical distribution of the 4015 rice accessions of the CCRP. (b) Composition of CCRP. (c) The distribution of 4015 accessions in China's five major rice-growing regions (Tables S1 and S2). (d) Neighbour-joining tree of CCRP. (e) Sites selected for planting CCRP. (f) 'Descriptors and Data Standard for Rice' were used to measure yield-related traits. (g) Yield distribution of CCRP in seven sites in two consecutive years. (h) Workflow for sequencing and SNP calling. (i) Distribution of MAF values for SNPs and Indels. (j) Workflow for variant filtering. (k) Visualization of variant density across the genome (bin size = 10 kb). (l) The algorithmic framework used by DeepCCR. (m) The CCRP data sets were divided randomly into a training data set and a test data set. Jigeng 816 was selected for independent prediction. (n) Predictive performance of different algorithms for Y trait information in the test data set in the GZL site. (o) Performance of DeepCCR in predicting Y trait density in the test data set in seven planting sites. (p) Performance of DeepCCR on the external validation data set Jigeng 816. The violin plot illustrates the distribution of values for the HD (left) and Y (right) traits among local varieties. The predicted values for Jigeng 816 are marked in red.

Next, we explored the predictive performance of the models on the test dataset and the results of the comparisons among the 10 traits at the seven sites also demonstrated the excellent performance of DeepCCR (Figure 1n; Figure S7). In the HF site, DeepCCR had higher genomic predictability (63.3%–78.2%) for the traits Y, HD, PH, PL, GP, SSR, GL and TGW, with lower predictive performance for TN and GW (Figure 1o; Figure S7a). The DeepCCR predictor results in the GZL, TH, NC, WH, KM and NN sites also exhibited high accuracy (Figure S7a). To comprehensively benchmark the predictive performance of DeepCCR, we calculated the MSE of the model at seven sites and obtained satisfactory results (Figure S7b). This suggested that our method is better at making genomic predictions for Chinese cultivated rice.

We next performed an external validation of the predictive ability of DeepCCR using Jigeng 816, the main cultivar in Jilin Province. DeepCCR demonstrated outstanding performance in predicting the 10 traits in the Jigeng 816 dataset (Table S3). Specifically, the predicted Y of 1.71 kg (converted to 564.3 kg/mu) and the predicted HD of 102 days are consistent with those of actual field plantings (https://www.ricedata.cn/). Considering that HD and Y are key indicators for assessing the ecological adaptability of an accession, we proceeded with the validation and prediction of the performance of Jigeng 816 at the remaining six planting sites. The HD of Jigeng 816 was below the 25% quantile for local varieties; however, the predicted Y in TH and KM reached the median and 25% quantile of local varieties, respectively (Figure 1p; Table S3). These results suggested that Jigeng 816 exhibits relatively better adaptability to TH and KM and can be used as a superior breeding line to assist in variety improvement. The results also demonstrated that the DeepCCR model excels in predicting the traits of new rice varieties. Additionally, the model can assess the most suitable planting sites for a given variety of rice.

To facilitate the use of the model by breeders, we constructed a web server (www.ai-breeder.com) containing the DeepCCR model. Users need to submit only standard FASTQ or VCF files, and the system automatically provides prediction results for the 10 traits in different sites (Figure S8).

In this study, we constructed the first large-scale Chinese rice population data set for rice genomic selection. We also conducted a comprehensive multiyear, multisite phenotypic survey and developed a companion deep neural network model to predict phenotypes and the ecological regions adapted for planting, as well as an easy-to-use online web server. The data set and results presented in this study offer a framework for breeders to quickly and efficiently breed superior rice varieties to address global food security issues. Additionally, with the increased number of materials in the data set and more comprehensive collection of multi-omics data (Wu *et al*., 2024), the predictive performance of DeepCCR will be further improved to enhance crop improvement programmes.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2021YFD1200500), the STI 2030-Major Projects (2022ZD04017) and the Agricultural Science and Technology Innovation Program of the Chinese Academy of Agricultural Sciences.

Conflict of interest

The authors declare no conflicts of interest.

Author contributions

L.H., K.Z. and S.Y. designed the experiments; X.M., B.H., D.C., J.L., Q.Z., X.X., P.S., C.T., L.G. and Y.Y. performed the field experiments; X.M. and S.W. performed the statistical analysis and all bioinformatic analyses; and X.M., H.W., S.W. and S.Y. wrote the manuscript.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author upon reasonable request.

References

Fu, J., Hao, Y., Li, H., Reif, J.C., Chen, S., Huang, C., Wang, G. *et al*. (2022) Integration of genomic selection with doubled-haploid evaluation in hybrid breeding: from GS 1.0 to GS 4.0 and beyond. *Mol. Plant*, **15**, 577–580.

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M. *et al*. (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, **557**, 43–49.

Wu, C., Luo, J. and Xiao, Y. (2024) Multi-omics assists genomic prediction of maize yield with machine learning approaches. *Mol. Breed*. **44**, 14.

Yu, X., Li, X., Guo, T., Zhu, C., Wu, Y., Mitchell, S.E., Roozeboom, K.L. *et al*. (2016) Genomic prediction contributing to a promising global strategy to turbocharge gene banks. *Nat. Plants*, **2**, 16150.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Tables S1–S3 Supplementary tables.
Figures S1–S8 Supplementary figures.
Data S1 Materials and methods.