Introduction

Sugarcane Production & Breeding background

As one of the global main crops, the sugarcane provided about 1.9 million tons yield in 2019, and accounted for approximate 79% of total global sugar production in the same year (Service, 2021). However, due to the complex genome structure, the breeding outcomes in sugarcane including the rate of selection cycle decrease and genetic gain in the past 20 years were slow. The typical breeding cycle of sugarcane by clonal selections which rely on the end-of-season yields is 12-14 years (Wei and Jackson, 2017).

Genomic Selection

The Genomic selection (GS) (Goddard and Hayes, 2007; Meuwissen et al., 2001) was determined as an outstanding approach in modern agricultural breeding, and developed from the classical marker-assisted selection (MAS) (Dekkers, 2004). The GS has an inherent ability that the GS can estimate breeding values before making crosses, which provides opportunities to decrease breeding cycles (Jonas and de Koning, 2013). Additionally, after introducing high-dense genome-wide genotyping technologies, the capacities and accuracy of GS have been widely extended by the marker-rich array (Meuwissen et al., 2001). In crop breeding fields, several GS project which focus on major crop species including rice (Xu et al., 2021), wheat (Juliana et al., 2020; Lozada and Carter, 2020), maize (Beyene et al., 2021; Krishnappa et al., 2021) achieved remarkable improvements.

The overall goal for almost breeding methods including GS is maximizing the average genetic gain, which require the programs can correctly estimate the sum of breeding values via a genetic model. Currently, various statistical GS models have been developed, including BLUP (Best Linear Unbiased prediction) methods such as GBLUP (VanRaden, 2008) with its extended versions (Yadav et al., 2021), and Bayesian based models (Meuwissen et al., 2001).

For sugarcane breeding, there are several challenges while performing GS on sugarcane genome. First, the sugarcane genome is highly polyploid, which increases the difficulty of precise genotyping (Aitken et al., 2016). Up to date, the compromise genotyping methods for diploid studies is the single-dose markers (SDMs) (Wu et al., 1992), which gives an approximate distribution for polyploid genomes. Under the above condition, analysers are required to generate a specific genetic model for reducing deviation between the limited detection and the actual complex genome. Meanwhile, the agricultural traits in sugarcane tend to be affected by a mixture of additive effects and non-additive effects, the latter contains both dominance and epistasis factors.

For conventional statistical methods, a study extended the basic additive GBLUP model with dominance genomic relationship matrix and epistatic genomic relationship matrix, the extended model was considered to have ability to estimate breeding values based on both additive and non-additive factors (Yadav et al., 2021). This study successfully increased the accuracy (measured as Pearson’s correlation between predictions and observations) of predictions in 3 sugarcane commercial traits. Another study performed GenomicSS and BayesR, both two methods can estimate non-additive effects in a genomic level then finally increase the performance of complex sugarcane traits prediction (Hayes et al., 2021).

Recently there were some groups also tried using artificial intelligence to perform GS in agricultural breeding. While the conventional statistical methods normally require strong assumption before training, using ML approaches in GS can be determined as assumption-free. Up to date, several studies have tested performance for some popular ML methods in GS. For example, a Deep Learning (DL) based workflow which named “DeepGS” (Ma et al., 2018) was tested in predicting traits of wheat, boosting method and Random Forest (RF) were tested in another study and got some positive achievements in bull predictions (Abdollahi-Arpanahi et al., 2020). However, in the above studies, the ML approaches could only surpass the conventional statistical methods in some limited conditions (excluding genotype-by-environment-by-trait interactions) (Montesinos-López et al., 2018). Another limitation of ML especially for Deep Learning is its black-box effect, the whole training of DL is formed purely by mathematical processing, and both the model and trained parameters are lacking biological meanings. Currently scientists can only evaluate the model capacities by measuring its accuracy from predictions to observations. To improve the performance of ML methods in GS field, developers should optimize both usability and interpretability of these ML models.

The objective of this study is to develop, assess, optimize several popular ML methods (RF, MLP, CNN) particularly for GS in sugarcane genome. The second goal for the generated training output, the study would perform several downstream analysis workflows for deconstructing the black-hive-box during a well-designed ML training process and discover the potential hidden genetic pattern that affect these complex traits.

Materials and methods

Reference population and genotyping

Phenotype collection and statistical analyses

Construction of genomic relationship matrices (GRM)

Random Forest modelling (RF)

Multilayer perceptron modelling (MLP)

Convolutional neural network modelling (CNN)

Results

Discuss

Reference

Abdollahi-Arpanahi, R., Gianola, D., and Peñagaricano, F. (2020). Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. Genetics Selection Evolution *52*, 12.

Aitken, K., Farmer, A., Berkman, P., Muller, C., Wei, X., Demano, E., Jackson, P., Magwire, M., Dietrich, B., and Kota, R. (2016). Generation of a 345K sugarcane SNP chip. Paper presented at: Proc. Aust. Soc. Sugar Cane Technol.

Beyene, Y., Gowda, M., Pérez-Rodríguez, P., Olsen, M., Robbins, K. R., Burgueño, J., Prasanna, B. M., and Crossa, J. (2021). Application of Genomic Selection at the Early Stage of Breeding Pipeline in Tropical Maize. Frontiers in Plant Science *12*.

Dekkers, J. C. (2004). Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. J Anim Sci *82 E-Suppl*, E313-328.

Goddard, M. E., and Hayes, B. J. (2007). Genomic selection. Journal of Animal Breeding and Genetics *124*, 323-330.

Hayes, B. J., Wei, X., Joyce, P., Atkin, F., Deomano, E., Yue, J., Nguyen, L., Ross, E. M., Cavallaro, T., Aitken, K. S., and Voss-Fels, K. P. (2021). Accuracy of genomic prediction of complex traits in sugarcane. Theoretical and Applied Genetics *134*, 1455-1462.

Jonas, E., and de Koning, D.-J. (2013). Does genomic selection have a future in plant breeding? Trends in Biotechnology *31*, 497-504.

Juliana, P., Singh, R. P., Braun, H.-J., Huerta-Espino, J., Crespo-Herrera, L., Govindan, V., Mondal, S., Poland, J., and Shrestha, S. (2020). Genomic Selection for Grain Yield in the CIMMYT Wheat Breeding Program—Status and Perspectives. Frontiers in Plant Science *11*.

Krishnappa, G., Savadi, S., Tyagi, B. S., Singh, S. K., Mamrutha, H. M., Kumar, S., Mishra, C. N., Khan, H., Gangadhara, K., Uday, G.*, et al.* (2021). Integrated genomic selection for rapid improvement of crops. Genomics *113*, 1070-1086.

Lozada, D. N., and Carter, A. H. (2020). Genomic Selection in Winter Wheat Breeding Using a Recommender Approach. Genes (Basel) *11*, 779.

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J., and Ma, C. (2018). A deep convolutional neural network approach for predicting phenotypes from genotypes. Planta *248*, 1307-1318.

Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics *157*, 1819-1829.

Montesinos-López, A., Montesinos-López, O. A., Gianola, D., Crossa, J., and Hernández-Suárez, C. M. (2018). Multi-environment genomic prediction of plant traits using deep learners with dense architecture. G3: Genes, Genomes, Genetics *8*, 3813-3828.

Service, U. S. D. o. A. F. A. (2021). Sugar: World Markets and Trade. World Production, Markets, and Trade Report.

VanRaden, P. M. (2008). Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science *91*, 4414-4423.

Wei, X., and Jackson, P. (2017). Addressing slow rates of long-term genetic gain in sugarcane. Pakistan Sugar Journal *32*, 23.

Wu, K. K., Burnquist, W., Sorrells, M. E., Tew, T. L., Moore, P. H., and Tanksley, S. D. (1992). The detection and estimation of linkage in polyploids using single-dose restriction fragments. Theoretical and Applied Genetics *83*, 294-300.

Xu, Y., Ma, K., Zhao, Y., Wang, X., Zhou, K., Yu, G., Li, C., Li, P., Yang, Z., Xu, C., and Xu, S. (2021). Genomic selection: A breakthrough technology in rice breeding. The Crop Journal *9*, 669-677.

Yadav, S., Wei, X., Joyce, P., Atkin, F., Deomano, E., Sun, Y., Nguyen, L. T., Ross, E. M., Cavallaro, T., Aitken, K. S.*, et al.* (2021). Improved genomic prediction of clonal performance in sugarcane by exploiting non-additive genetic effects. Theoretical and Applied Genetics *134*, 2235-2252.