Research proposal

New mate allocation strategies to accelerate genetic gain in agricultural species.

Student Name: Chensong Chen

Principle supervisor: Prof Ben Hayes

Co-supervisor: Dr Elizabeth Ross, Dr Owen Powell, Dr Eric Dinglasan

**Introduction**

With the critical human population increment and the climate changes, the strategies of agricultural breeding are considered as the key technique for ensuring global food supply in the further 30 years (He and Li, 2020). The genomic selection (GS) methods (Meuwissen et al., 2001), has been used as an established methodology for years in agricultural breeding for accelerating the breeding speed.

The major object of the breeding programs is improving the average genetic gain of the future generations. The strategy that evaluating the estimated breeding value (EBV) for certain species by the predicted overall genetic effects has been tested to be effective and viable (Goddard, 2009). The foundation of the GS is the genetic prediction. The core theory of the genetic prediction is the classical complex quantitative traits model(Bernardo, 2008; Meuwissen et al., 2013).

*Genomic prediction & Genomic selection*

By the description of the classical complex quantitative traits model, the phenotype traits are formed based on an infinite number of infinitesimal genetic effect and the interaction of non-genetic factors such as environmental factor (Bulmer, 1980). Genomic selection is a marker-based selection method which introduces genetic markers covering the whole genome so that all quantitative trait loci (QTL) are in linkage disequilibrium with at least one marker. This technology was raised by the discovery of the large amount of single nucleotide polymorphisms (SNP) by the next-gen whole genome sequencing and new efficient genotyping approaches. So far, the Best Linear Unbiased Prediction (BLUP) (Henderson, 1984) is commonly used to calculate breeding values. By fitting based on the given genotypes and phenotypes, a BLUP model could estimate the linkages from genotypes to traits. The standard BLUP model, and many extended GBLUP model (BLUP mixed with other effective matrix such as dominance, epistatic, and genome-wide heterozygosity effects) are tested well with many agricultural species. The upcoming challenges are the hidden and unknown associations in a genomic level would actually increase the difficulties of genetic prediction.

Previous studies have verified that performing during the breeding process can obviously increase the average gain per year in many crop species such as maize (Môro et al., 2019; Vivek et al., 2017) and soybean (Smallwood et al., 2019).

*Machine Learning approach*

As predicting complex traits from dense genomic data has been an engaging topic among breeder and bioinformaticians. While considering the large amount of the genomic data, some machine learning appro aches such as Random forests and Neural network could be well candidates as they are designed for solving problems with big data. In some traditional machine learning research fields like image/text classification, a large number of studies showed that the CNN approach has outstanding abilities in feature selection and reducing the risk of curse of dimensionality (Guo et al., 2016). Meanwhile, there is a common advantage for almost DL algorithms including multilayer perceptron and Convolutional Neural Network is the capacity to solve predictions that contain some unknown patterns while the conventional statistical methods usually require a designed incidence matrix, especially in the nowadays GS field. For instance, the statistical model needs to consider non-additive effects, and associate with an environment-interactive matrix, which could be challenging.

The Convolutional Neural Network (CNN), which is a subtype of the neural network, has been studied to solve genetic prediction problems with dense genotype markers in many plant and animal species (Abdollahi-Arpanahi et al., 2020; Ma et al., 2018; Montesinos-López et al., 2018; Waldmann, 2018).

[Random forest]

*Aim*

This project aims to import artificial intelligence to improve performance of the current genomic selection strategy.

*Methods and Materials*

Dense genotyping data for all recorded plant/animal samples

Recording of traits and environmental factors that response to each sample.

Convolutional neural network model based on python.

Random forest model coded by python.

GBLUP by R will be introduced as a competitor.

Hypothesis:

In some particular situation, such as traits effected by more non-additive genotypes, machine learning approaches could be competitive methods to improve GS performance.

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.

Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop science *48*, 1649-1664.

Bulmer, M. G. (1980). The mathematical theory of quantitative genetics: Clarendon Press.).

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. Genetica *136*, 245-257.

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., and Lew, M. S. (2016). Deep learning for visual understanding: A review. Neurocomputing *187*, 27-48.

He, T., and Li, C. (2020). Harness the power of genomic selection and the potential of germplasm in crop breeding for global food security in the era with rapid climate change. The Crop Journal *8*, 688-700.

Henderson, C. R. (1984). Applications of linear models in animal breeding.

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., Hayes, B., and Goddard, M. (2013). Accelerating improvement of livestock with genomic selection. Annu Rev Anim Biosci *1*, 221-237.

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.

Môro, G. V., Santos, M. F., and de Souza Junior, C. L. (2019). Comparison of genome-wide and phenotypic selection indices in maize. Euphytica *215*, 1-14.

Smallwood, C. J., Saxton, A. M., Gillman, J. D., Bhandari, H. S., Wadl, P. A., Fallen, B. D., Hyten, D. L., Song, Q., and Pantalone, V. R. (2019). Context‐Specific Genomic Selection Strategies Outperform Phenotypic Selection for Soybean Quantitative Traits in the Progeny Row Stage. Crop Science *59*, 54-67.

Vivek, B., Krishna, G., Vengadessan, V., Babu, R., Zaidi, P., Mandal, S., Grudloyma, P., Takalkar, S., Krothapalli, K., and Singh, I. (2017). Use of genomic estimated breeding values results in rapid genetic gains for drought tolerance in maize.

Waldmann, P. (2018). Approximate Bayesian neural networks in genomic prediction. Genetics Selection Evolution *50*, 1-9.