

Predicting a path to increased genetic gain using artificial intelligence

Kai Voss-Fels¹, Eric Dinglasan¹, Lee Hickey¹, Bertus Jacobs², Ben Hayes¹

¹ Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St Lucia, QLD, 4072, Brisbane, Australia

² LongReach Plant Breeders, Landsdale, SA, Australia

Introduction

One of the most critical steps in a breeding program is parent selection for crossing. Conventionally, selections are made based on performance *per se* or on breeding values. This limits the inference that can be made about how complementary sets of favorable chromosome segments of the parents are. Maximizing the number of good alleles in the offspring is particularly challenging when a combination of quantitative traits (grain yield) and mono- or oligogenic traits (e.g. disease resistance, quality) are considered simultaneously. To overcome this problem, we use AI algorithms inspired by biological evolution to find parental combinations that maximize the number of favorable alleles in the offspring. Breeding lines are advanced using 'speed breeding' to rapidly stack up favorable chromosome segments in the shortest possible time.

"The time is to shorten several years of cross-breeding and selection"

Materials & Methods

Genotype & phenotype data for >34,000 spring wheat lines

- Genotype profiles for the entire panel were generated using DArTseq genotyping (DArT Pty Ltd (Canberra), resulting in 5,112 high quality SNPs
- Phenotype data from 694 individual trials run in Australia between 2012-2018, including 44 locations and 29 trial types
- Traits measured: Grain yield, resistances, quality

Local genomic estimated breeding values (local GEBVs)

Instead of considering GEBVs for breeding lines we use local GEBVs for haplotypes of chromosomal segments that are conserved in the population and show strongly reduced allelic recombination (Fig. 1). Briefly, markers that are in strong linkage disequilibrium (LD) are grouped into LD blocks. For each block, haplotype effects are obtained by summing up allele effects.

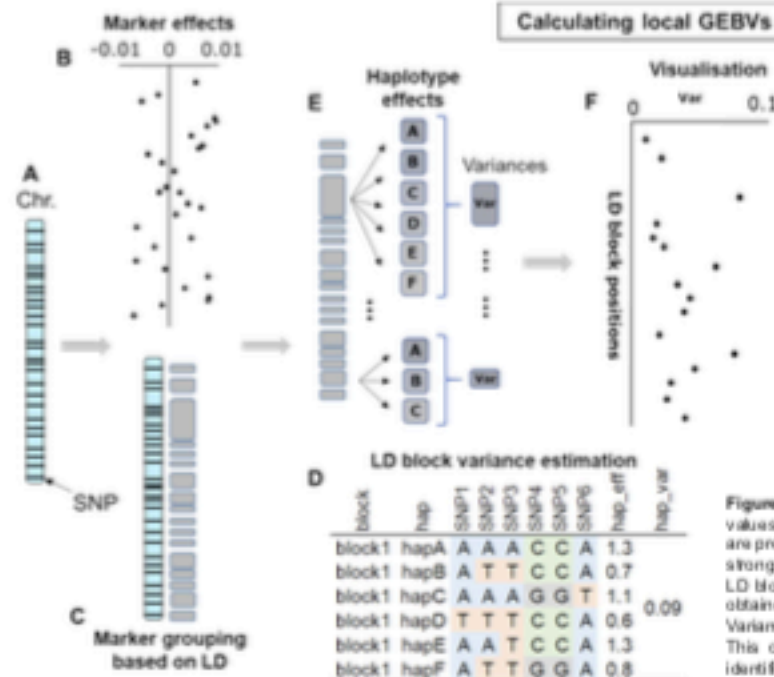


Figure 1: Local genomic estimated breeding values (local GEBVs). (A, B) Single marker effects are predicted (SNP-BLUP). (C) Markers that are in strong linkage disequilibrium (LD) are grouped into LD blocks. (D, E) Local GEBVs for haplotypes are obtained by summing up allele effects. (F) Variances across local GEBVs can be calculated. This can be an alternative to conventional QTL identification methods.

Parent selection using genetic algorithms

To select parents for crosses we use 'genetic algorithms' that identify the most optimal parental combination for maximising the number of favourable haplotypes in the offspring (Fig. 2). The algorithm searches the optimal set of parental lines that maximises the number of complementary haplotypes with high local GEBVs. The breeding lines are advanced using 'speed breeding' which enables up to six generations of spring wheat per year.

Results

Since genotype-by-environment interaction is very prevalent in Australia, the data set has been split up into two macro regions (MR). In total, ~15.7K and 17.7K lines have been tested in MR1 and MR2. When fitting a SNP-BLUP model and predicting the performance of both subsets in the respective MR, correlations between predicted and actual performance were 0.55 and 0.62 for MR1 and MR2.

When predicting new genotypes in new years based on data from previous years, prediction accuracy ranged between 0.11 and 0.41, depending on i) the size of the training population and ii) the test region (Table 1).

Table 1: Prediction accuracies for two macro regions (MR1 & MR2) using a SNP-BLUP model.

Data span: 2012-2018	predict 2013		predict 2017		predict 2018	
	MR1	MR2	MR1	MR2	MR1	MR2
trPop (n)	996	1,662	9,726	10,451	12,473	12,816
valPop (n)	1,711	1,798	3,083	2,712	3,191	4,923
Accuracy (r)	0.11	0.27	0.14	0.26	0.24	0.41

Genotypes that were to be predicted were fully removed from the training population. Accuracy (r) = cor(GEBV, BLUE)

The single marker effect estimates were used to obtain local GEBVs for haplotypes. Future predictions based on local GEBVs resulted in a potential increase of 2.1 to 2.6-fold for grain yield if the best haplotypes across the entire genome were stacked up.

"The Australian wheat gene pool holds the potential for yield increases of 2.1 to 2.6-fold"

The best potential offspring that could be generated from combining breeding lines differed depending on how many parents were selected (Fig. 2). The best potential individual had a 1.8 to 2.0-fold higher GEBV than the best tasted breeding line in MR1 and a 1.4 to 1.6-fold higher GEBV than the best line in MR2.

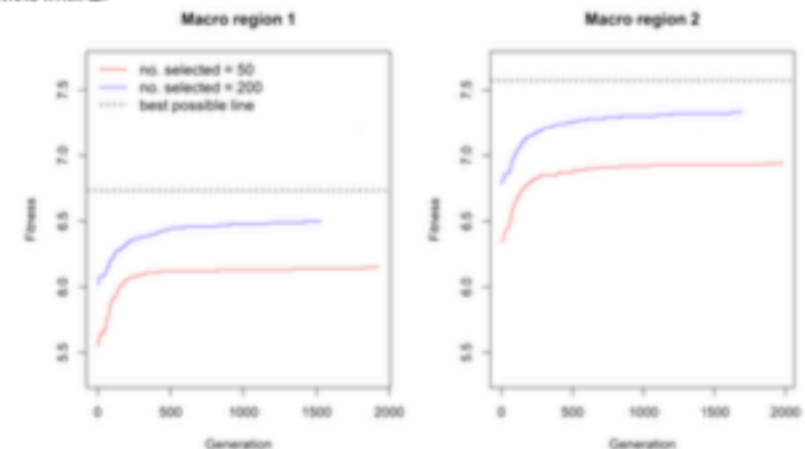


Figure 2: The genetic algorithm identifies parental combinations in the breeding pool that maximise the number of favourable haplotypes in the offspring. The dashed line represents the best possible breeding line that could be created if all favourable haplotypes across the genome were stacked up.

Conclusions

With local GEBVs, effects of chromosome segments that are formed by blocks of DNA markers can be estimated. A fundamental difference of our approach to conventional genomic selection strategies is that it explores complementarity of chromosome segments between genotypes, rather than considering their absolute genome-wide breeding values. Therefore, breeding lines that carry interesting haplotypes can be considered in the mating design, even if they have a lower total genome-wide GEBV than the top performers. Furthermore, local allelic interactions that might cause the formation of LD blocks are potentially captured by our method.

New breeding lines generated by from genotypes that were selected by the genetic algorithm are currently advanced and intercrossed under 'speed breeding' conditions. This approach holds the potential to fast-track the development of germplasm with significantly increased numbers of favourable chromosome segments and thereby rapidly accelerate genetic gain in wheat.

Acknowledgements

This work is supported by the Australian Research Council. Project title: 'FastStack - evolutionary computing to stack desirable alleles in wheat' (LP170100317). We thank Jose Crossa (CIMMYT) and Sandra Dunckel (KWS UK Ltd, formerly LongReach) for their support with data collection and analysis.