Genomic Selection in Chickpea using Whole Genome Re-sequencing Data

Yongle Li[1]*, Pradeep Ruperao[2], Jacqueline Batley[2], David Edwards[2], Kristy Hobson[3], Tim Sutton[4,1]

1 School of Agriculture, Food and Wine, The University of Adelaide, SA 5064
2 School of Plant Biology, The University of Western Australia, WA 6009
3 NSW Department of Primary Industries, NSW 2340
4 South Australian Research and Development Institute, SA 5001

***Contact: yongle.li@adelaide.edu.au**

Introduction

Next-generation sequencing technology offers a relatively cheap and high-throughput genotyping option to discover genome variation and identify selection signatures in crop species such as chickpea. These information can be used in genomic selection, a new approach for selecting quantitative traits. It uses DNA markers distributed across the whole genome to predict the breeding value of an untested line.

Results

Whole genome re-sequencing (WGRS) data were obtained from 315 Australian chickpea advanced breeding lines. More than half a million SNPs were discovered using Illumina paired-end sequencing with genome coverage of 5-10X. Phylogenetic analysis showed that the desi breeding lines have much higher genetic diversity than the desi varieties released before 2005 (Fig. 1). A cluster with high level of Phytophthora root rot (PRR) resistance, consisting of 75 lines with wild species *C. echinospermum* background from Stage 1 and Stage 2, separated clearly from the rest of the lines.

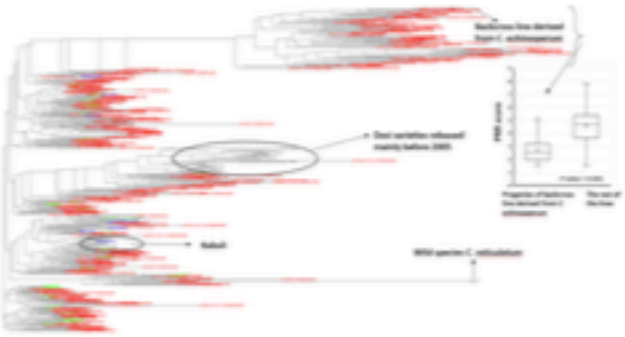

Fig.1 Phylogenetic tree of 315 breeding lines and 48 released varieties. The 48 released Australian varieties are coloured in black. Stage 1 and Stage 2 lines are coloured in red. Stage 3 lines are coloured in green. The PBA varieties are coloured in blue.

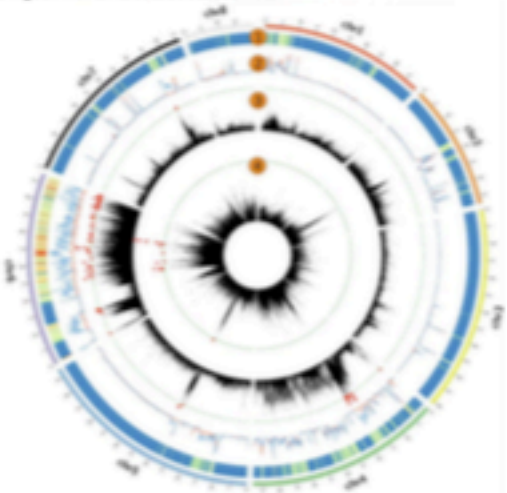

Fig.2 Genome variation and GWAS based on whole genome re-sequencing data
1) SNPs density 2) Nucleotide diversity (θπ) of the top 50 PRR resistant lines (π-red) and the top 50 PRR susceptible lines (π-blue). 3) Circular Manhattan plots displaying the GWAS-PRR result. 4) Circular Manhattan plots displaying the GWAS-yield result. The SNP are represented by black dot, while the SNP with significant association are represented by red triangle (p-values lower than 1.67E-07).

References:

1. Amalraj A, Taylor J, Billakurthi K, Li Y, Muzaid E, Hobson K, Sutton T (2019) Mapping resistance loci to Phytophthora root rot (caused by *Phytophthora medicaginis*) in chickpea (*Cicer arietinum* L.) and wild (*Cicer echinospermum* P.H. Davis) relatives. *Mol Breed* 39:57. https://doi.org/10.1007/s11032-019-0961-3

Genome-wide association study identified 118 and 20 SNPs significantly associated with PRR resistance and grain yield, respectively (Fig. 2). Two of the SNPs on Chr4 were in close proximity (~600Kb away) of two PRR QTL identified in a previous study (Amalraj et al., 2019). All regions surrounding the SNP significantly associated with PRR were co-located with the regions with reduced level of nucleotide diversity in PRR tolerant lines and of long extent level of linkage disequilibrium. This suggested these regions may have gone through selective sweeps caused by selective breeding of PRR resistance. A number of candidate genes, flanking the significant SNP, have functional annotation involved in biotic stress response such as LRR receptor-like kinase, cysteine-rich receptor-like kinases, serine/threonine kinase, ethylene-responsive transcription factor, auxin-induced protein 5NG4.

Table 1 Prediction accuracies of yield for the new advanced lines using their close relatives as training sets

Training sets (No. of lines)	Target sets (No. of lines)	Prediction models: RR-BLUP	Bayes(ian)-LASSO	Bayesian ridge regression
2012BB (315)	2013BB (314)	0.33***	0.34***	0.36***
2012BB (315)	2013HM (314)	0.06	0.10	0.06
2012BB (315)	2013MO (314)	0.36***	0.39***	0.38***
2012HM (315)	2013HM (314)	-0.02	-0.01	-0.02
2012MO (315)	2013BB (314)	-0.32***	-0.06	-0.28***
2012HM (315)	2013MO (314)	-0.34***	-0.04	-0.28***
2012MO (315)	2013MO (314)	0.22***	0.17***	0.19***
2012MO (315)	2013BB (314)	0.19*	0.14*	0.17*
2012MO (315)	2013HM (314)	0.10	0.12	0.08

Note: 1) BB, HM, and MO represent yield trials at Billa Billa (QLD), Hermitage (QLD), and Merredin (WA), respectively in 2012 and 2013 are the years that the yield trials were conducted. 2) The training sets consist of 315 breeding lines from 2012 stage 1 and 2, the target sets consist of 187 new breeding lines from 2013 stage 1.

The 315 advanced lines were measured for grain yield in three locations in 2012 and 2013. To mimic a common practice of evaluating advanced lines in plant breeding programs, the earlier datasets were used as training sets to predict the latter datasets. The 2012BB and 2012MO datasets (training sets) predicted the 2013BB and 2013MO datasets (target sets) with an acceptable accuracy ranging from r = 0.14 to 0.39 whereas they predicted the 2013HM dataset poorly with negatively correlated or no correlation (Table 1). The 2012HM dataset predicted the 2013 datasets poorly and even for the 2013HM dataset. This may be due to the fact that BB and MO were rainfed trials whereas HM was irrigated.

Table 2 Effects of SNP functions on prediction accuracy based on cross-validation

Traits	Test	SNPs locations in the genome (No. of SNPs)				
		Whole genome (300K)	Exome+5'&3' UTR (113K)	Exome (29K)	Missense+alternative splicing in exome (11K)	Random SNP (29K)
2013BB	Yield	0.57±0.01	0.57±0.01	0.58±0.00	0.56±0.01	0.61±0.01
2013HM	Yield	0.30±0.01	0.27±0.01	0.32±0.01	0.29±0.01	0.29±0.01
2013MO	Yield	0.56±0.01	0.57±0.01	0.58±0.01	0.55±0.00	0.64±0.00
2013HM nursery	PRR	0.63±0.01	0.63±0.00	0.62±0.00	0.63±0.01	0.65±0.00

To test the effect of SNP functions on prediction accuracy, the 300K SNP were divided into different classes: Exome+3'&5'UTR+MNase, Exome, Missense + alternative splicing in exome, 29K random SNP (Table 2). The four different classes did not decrease prediction accuracies of yield and PRR resistance significantly despite consisting of a much smaller number of markers.

Conclusion

This study demonstrates the power of combining whole genome re-sequencing data and GWAS to rapidly identify candidate genes for yield and disease resistance. The results have important implication for implementing genomic selection into plant breeding programs.