Received: 26 May 2021 | Accepted: 14 September 2021

DOI: 10.1002/tpg2.20166

The Plant Genome

ORIGINAL RESEARCH

Genomic prediction of preliminary yield trials in chickpea: Effect of functional annotation of SNPs and environment

Yongle Li[1] | **Pradeep Ruperao**[2] | **Jacqueline Batley**[3] | **David Edwards**[3] | **William Martin**[4] | **Kristy Hobson**[5] | **Tim Sutton**[1,6]

[1] School of Agriculture, Food and Wine, The Univ. of Adelaide, Adelaide, SA 5064, Australia

[2] Statistics, Bioinformatics and Data Management, ICRISAT, Hyderabad 502324, India

[3] School of Biological Sciences, The Univ. of Western Australia, Perth, WA 6001, Australia

[4] Dep. of Agriculture and Fisheries, Warwick, Qld 4370, Australia

[5] NSW Dep. of Primary Industries, Tamworth, NSW 2340, Australia

[6] South Australian Research and Development Institute, Adelaide, SA 5064, Australia

Correspondence
Yongle Li, School of Agriculture, Food and Wine, The Univ. of Adelaide, Adelaide, SA 5064, Australia.
Email: yongle.li@adelaide.edu.au

Associate Editor: Francois Belzile.

Funding information
Australia-India Strategic Research Fund; Grains Research and Development Corporation; NSW Department of Primary Industries

Abstract

Achieving yield potential in chickpea (*Cicer arietinum* L.) is limited by many constraints that include biotic and abiotic stresses. Combining next-generation sequencing technology with advanced statistical modeling has the potential to increase genetic gain efficiently. Whole genome resequencing data was obtained from 315 advanced chickpea breeding lines from the Australian chickpea breeding program resulting in more than 298,000 single nucleotide polymorphisms (SNPs) discovered. Analysis of population structure revealed a distinct group of breeding lines with many alleles that are absent from recently released Australian cultivars. Genome-wide association studies (GWAS) using these Australian breeding lines identified 20 SNPs significantly associated with grain yield in multiple field environments. A reduced level of nucleotide diversity and extended linkage disequilibrium suggested that some regions in these chickpea genomes may have been through selective breeding for yield or other traits. A large introgression segment that introduced from *C. echinospermum* for phytophthora root rot resistance was identified on chromosome 6, yet it also has unintended consequences of reducing yield due to linkage drag. We further investigated the effect of genotype by environment interaction on genomic prediction of yield. We found that the training set had better prediction accuracy when phenotyped under conditions relevant to the targeted environments. We also investigated the effect of SNP functional annotation on prediction accuracy using different subsets of SNPs based on their genomic locations: regulatory regions, exome, and alternative splice sites. Compared with the whole SNP dataset, a subset of SNPs did not significantly decrease prediction accuracy for grain yield despite consisting of a smaller number of SNPs.

Abbreviations: BB, Billa Billa, QLD, Australia; BL, Bayesian least absolute shrinkage and selection operation; BLUE, best linear unbiased estimated; BRR, Bayesian ridge regression; Fst, fixation index; G × E, genotype × environment; GBS, genotyping by sequencing; GGE, genotype main effect plus genotype × environment interaction; GS, genomic selection; GWAS, genome-wide association studies; HM, Hermitage, QLD, Australia; LD, linkage disequilibrium; MAF, minor allele frequency; METs, multi-environment trials; MLM, mixed linear model; MNase-HS, micrococcal nuclease hypersensitive regions; MO, Moree, NSW, Australia; RR-BLUP, ridge regression best linear unbiased estimation; SNP, single nucleotide polymorphism; UTR, untranslated region.

1 | INTRODUCTION

Chickpea (*Cicer arietinum* L.) is an important pulse, rich in protein and essential micronutrients, such as magnesium, iron, zinc, and vitamins (Wallace et al., 2016). Compared with meat-derived protein, pulse protein is more efficient and sustainable in terms of resource use. Australia was the second-largest chickpea producer after India in 2017 according to FAOSTAT, although the average yield of chickpea in Australia is low (~1.4 t ha^{-1}) due to production constraints such as drought (Y. Li et al., 2018; Pang et al., 2016), salinity (Atieno et al., 2017; Khan et al., 2016), chilling stress (Kiran et al., 2019), and plant diseases (Amalraj et al., 2019; Y. Li et al., 2017).

Genomic selection (GS) is a cost-effective approach to incorporate quantitative traits in breeding programs (Crossa et al., 2014; Meuwissen et al., 2001). It uses molecular markers distributed across the whole genome to predict the breeding value of an untested line, and thus shifts the focus of marker identification to line selection. This approach assists in selecting the best parents for crossing, reduces the cost and time of the breeding cycle, and thus has been adopted rapidly by many animal and plant breeding programs (Asoro et al., 2013; Crossa et al., 2014; Hayes et al., 2009; Y. Li et al., 2018; Weigel et al., 2010). An opportunity to implement GS in plant breeding programs is at preliminary yield trial stages where GS can assist in selecting lines to be progressed to the next stage. Here, there are several questions to be considered: What germplasm should be included in the training set? To predict the yield of a target set at a particular environment, should the training set be evaluated in a similar environment or multiple environments with different conditions (e.g., rainfed and irrigated)? Is it possible to predict the grain yield of a line in a new environment based on similar or different environments by borrowing information from relevant environments? The first objective of this study was to address these questions using genotypic data and grain yield from a set of 315 advanced chickpea lines.

The factors affecting prediction accuracy of agronomic traits in crop species include training population size (Lorenz, 2013; Norman et al., 2017), training population composition (Berro et al., 2019; Crossa et al., 2014; Hoffstetter et al., 2016), marker density (Hickey et al., 2014; Norman et al., 2018), and the specific characteristics of prediction models used (de Los Campos et al., 2013; Heslot et al., 2012). A large proportion of genomic selection papers published to date are based on analysis using genotypic data from genotyping × sequencing (GBS) or single nucleotide polymorphism (SNP) array. Reduced representation GBS is a next-generation sequencing-based method. To reduce sequencing costs, only a fraction of the genome is sequenced by targeting low-copy genomic regions using restriction enzymes. SNP arrays are a hybridization-based genotyping method with a fixed number of SNP markers. One disadvantage of the two genotyping platforms is that they generate a relatively small number of markers, which limits the possibility of estimating the "true" upper boundary of prediction accuracy. By contrast, whole-genome resequencing can generate a large amount of SNP data in a very cost-effective manner, particularly for crop species with a relatively small genome such as chickpea and rice (*Oryza sativa* L.). This opens an exciting opportunity to study the effect of SNPs' density on prediction accuracy. In addition, it has been shown that SNPs located in coding and regulatory sequences explain a much larger proportion of genetic variance than those from random genomic locations (Koufariotis et al., 2018; Rodgers-Melnick et al., 2016). The data acquired here provided the opportunity to study the effect of SNP on prediction accuracy with regards to the context of their genomic location (i.e., coding or regulatory regions). The second objective of this study was to investigate the prediction accuracy of yield based on high-density SNPs, and their subsets (based on SNP location) generated from whole-genome resequencing method.

Core Ideas

- We recommend updating the training set with phenotypes from relevant environments for genomic selection.
- Subsetting SNP based on its functional annotations did not affect prediction accuracy for yield.
- An introgression segment for disease resistance has unintended consequences of reducing yield.

2 | MATERIALS & METHOD

2.1 | Plant materials, sequencing, and SNP discovery

Plant materials consisted of 315 advanced desi chickpea lines from the Australian chickpea breeding program of Breeding Stage 1, 2, and 3 in 2013. We extracted DNA from the young leaf of the plant using the Qiagen DNeasy Plant Mini Kit following the manufacturer's instructions. Pair-end sequencing libraries were constructed for each line with insert sizes of ~500 bp using Illumina's TruSeq library kit. The Illumina HiSeq 2000 platform was used to generate paired-end short reads (150 bp) with a target sequencing depth of 5–10× per line. Sequencing reads for each line were trimmed, filtered, and mapped to the CDC Frontier reference genome

2.6.3 (http://www.cicer.info/databases.php) using Trimmomatic and SOAP2. Bam files were processed to filter out reads with more than two base-pair mismatches. Homozygous SNPs were identified using SAMtools and BCFtools (H. Li et al., 2009). Nucleotide diversity ($\theta\pi$) and fixation index (Fst) of the 50 greatest and 50 least yielding lines were calculated using BCFtools. The raw sequences were deposited in the National Center for Biotechnology Information under accession number PRJNA743728.

2.2 | Phylogenetic analysis and linkage disequilibrium

To compare the genetic relationship and diversity of the 315 advanced desi lines with other Australian genotypes, a previous SNP dataset from Y. Li et al. (2017) was merged with the SNP dataset generated in this study, resulting in 55,195 SNP markers with minor allele frequency (MAF) >5% and a missing rate <20%. A phylogenetic tree was constructed based on the distance matrix estimated from this merged SNP dataset with the neighbour-joining clustering method implemented in TASSEL 5.0. The distance matrix was calculated using a Euclidean distance, where a homozygous locus is 100% similar to itself, but a heterozygous locus is only 50% similar to itself (due to the two different alleles present). The resulting tree was visualized with the Archaeopteryx Tree function in TASSEL 5.0. Linkage disequilibrium (LD) was measured by r^2 using 55,255 high confidence SNPs (minimum five reads per genotype, MAF >5%, missing rate <20%). The LD-decay curve under the mutation-drift-equilibrium model was estimated as described in in Li et al. (2011).

2.3 | Field trials and phenotypic analysis

The advanced lines were measured for grain yield at three locations: Billa Billa, Queensland (BB, 28.15° S lat; 150.30° E long), DAF Hermitage Research Facility, Queensland (HM, 28.21° S lat; 152.10° E long), and Moree, NSW (MO, 29.47° S lat; 149.83° E long) in 2012 (128 lines) and 2013 (315 lines). Randomized complete block design with three replicates was used for all trials. Plots comprised of two rows with an area of 11 m^2.

Grain yield data were first analyzed separately for single-environment analysis by fitting a mixed linear model (MLM) in which spatial effects such as row and column were considered. The resulting best linear unbiased estimated (BLUE) values for each genotype were used subsequently for genome-wide association studies (GWAS) and GS analysis. In GS, some training and target sets consist of multiple environments (i.e., 2012_BB_HM_MO). The BLUE values from single environment analysis were used to fit a multiple-environment MLM in which environments were treated as random effects. Statistical significance of fixed and random effects was assessed using Wald's test (Wald, 1943) and the likelihood ratio test, respectively (Van Belle et al., 2004). To visualize genotype × environment (G × E) effect, biplot analysis of the six yield trials was performed using the genotype main effect plus genotype × environment interaction (GGE) Biplot function of GeneStat v19 with the BLUE values generated from the single-environment analysis.

2.4 | Genome-wide association study

To minimize false-positive association, population structure was estimated based on 298,154 SNPs (MAF > 0.05, missing rate < 50%) using ADMIXTURE (v1.23) software, applying a model-based method to calculate ancestry of unrelated individuals (Alexander et al., 2011; Alexander et al., 2009). The number of groups (K) in the population was estimated from 1 to 10 and the most likely number of groups (k = 5) was determined by the cross-validation error and knowledge of the germplasm's breeding history. Genome-wide association study was conducted using the adjusted entry means of the 315 lines with the dataset of grain yield and 298,154 SNPs (MAF > 0.05, missing rate <50%). For grain yield, adjusted entry means of the 315 lines resulting from fitting a multiple-environment MLM were used for association study. The environments were restricted to 2012BB, 2012MO, 2013BB, and 2013MO due to large G × E effect resulting from inclusion of the environments 2012HM and 2013HM. The SUPER-GWAS algorithm, implemented in the GAPIT 3.0 software, was used to calculate the SNP effect, adjusting for the confounding effects of population structure and kinship (Lipka et al., 2012; Tang et al., 2016; Wang et al., 2014). The kinship matrix was calculated using the VanRaden method and compressed to its optimum groups subsequently based on the P3D method to accelerate computation time (Z. Zhang et al., 2010). To increase statistical power, the SUPER-GWAS method calculated the kinship matrix with a subset of markers that were not in LD with the testing markers (Wang et al., 2014). The parameters of the SUPER-GWAS model were Sangwich.bottom = "SUPER", LD = 0.1, Sangwich.top = "MLM." The significant p-value threshold was p-value = 1.61×10^{-7}, equivalent to the α level of .05 after Bonferroni correction.

2.5 | Genomic predictions

Three different models were used for genomic prediction with 298,154 SNPs (MAF > 0.05, missing rate <50%):

TABLE 1 Summary of single nucleotide polymorphisms (SNPs) and linkage disequilibrium (LD)

Chromosome	No. SNPs	Density of SNPs (no. SNPs/10 Kb)	No. SNPs used to estimate LD[a]	Mean/median r^2	Distance of LD extent (r^2 cut-off: 0.2 or 0.4)
Chr1	26,369	5.33	5,886	0.15/0.02	250/50 Kb
Chr2	18,115	4.85	5,802	0.10/0.02	100/20 Kb
Chr3	13,636	2.03	5,449	0.23/0.03	500/80 Kb
Chr4	52,315	8.83	10,689	0.27/0.04	1,000/180 Kb
Chr5	27,806	4.07	8,701	0.14/0.03	350/50 Kb
Chr6	129,328	19.52	9,377	0.43/0.36	7,000/900 Kb
Chr7	25,051	4.43	7,423	0.13/0.02	450/80 Kb
Chr8	5,534	2.74	1,928	0.14/0.01	150/20 Kb
Total	298,154		55,255		

[a]SNPs with high confidence (minimum five reads per genotype, minor allele frequency >5%).

ridge regression best linear unbiased estimation (RR-BLUP), Bayesian least absolute shrinkage and selection operation (BL), and Bayesian ridge regression (BRR). The detailed formulation of these models was described in previous papers (de Los Campos et al., 2013; Y. Li et al., 2018) In brief, RR-BLUP is a widely used prediction model that uses a flexible MLM formulation. It employs a shrinkage process that shrinks each marker effect equally toward zero with the "infinitesimal" assumption that complex traits are controlled by a large number of loci with very small and equal effect (Barton et al., 2017; Hill, 2014). In contrast, BL is a Bayesian shrinkage model with the feature of greater shrinkage of markers with small effects and less shrinkage of markers with large effects (Tibshirani, 1996), compared with RR-BLUP. BRR is a Bayesian version of RR-BLUP that performs homogenous shrinkage across markers. The gpMod function in the R package 'synbreed' was used to fit the three models and predict the target sets (https://synbreed.r-forge.r-project.org/). Prediction accuracies were calculated as Pearson's correlation coefficient between the predicted values and observed phenotypic values based on BLUEs.

In five-fold cross-validation, the dataset was divided into five subsets; one of which was used as a target set while the other four were used as a training set to train the RR-BLUP model. This process was repeated 10 times resulting in 50 cross-validations. Prediction accuracy was then estimated as Pearson's correlation coefficient between the predicted values and observed phenotypic values of the target set. To study the effect of SNP function on prediction accuracy, SNPs were filtered out based on their locations such as exome, 3′ and 5′ untranslated region (UTR), and micrococcal nuclease hypersensitive regions (MNase-HS) (Rodgers-Melnick et al., 2016; W. Zhang et al., 2012). Single nucleotide polymorphisms that lead to missense mutation (changing amino acids) and alternative splicing were predicted using SnpEff 4.3 (Cingolani et al., 2012).

3 | RESULTS

3.1 | Genome variation

In total, 315 advanced desi lines from Breeding Stage 1, 2, and 3 of the Australian chickpea breeding program were resequenced with a sequencing depth of 6.2× on average. After mapping to the CDC Frontier reference genome 2.6.3, a total of 298,154 homozygous SNPs with MAF >0.05 and missing genotype <50% were discovered. The SNP density among the eight chromosomes ranged from 19.5 SNPs/10 kB on chromosome 6 to 2.03 SNPs/10 kB in chromosome 3 (Table 1, Figure 1). Genetic diversity in this material measured as θ_π was 0.93×10^{-4}, which was higher than a previous study comprising 47 released Australian chickpea cultivars (Y. Li et al., 2017). Two regions (from 5,273,564 to 11,639,389 and from 18,536,446 to 61,765,087) on chromosome 6 have a much higher SNP density (21.2 SNPs/10 kB and 25.1 SNPs/10 kB) compared with the rest of chromosome 6 (3.7 SNPs/10 kB). Linkage disequilibrium, as measured by the mean of r^2 on each chromosome, ranged from 0.10 on chromosome 2 to 0.43 on chromosome 6 (Table 1 and Figure 2). Setting an r^2 threshold of 0.2, the extent of LD, ranging from 100 kb on chromosome 2 to 7,000 kb on chromosome 6, is much smaller (excepted for chromosome 6) than the previous study (Y. Li et al., 2017), suggesting a higher genetic diversity in this breeding material.

3.2 | Population structure

To facilitate a comparison between Australian chickpea released cultivars and the 315 S1, S2, and S3 advanced breeding lines, SNP data from Y. Li et al. (2017) consisting of Australian chickpea cultivars and breeding lines and an accession (PI4899777) of the wild chickpea species *Cicer reticulatum*

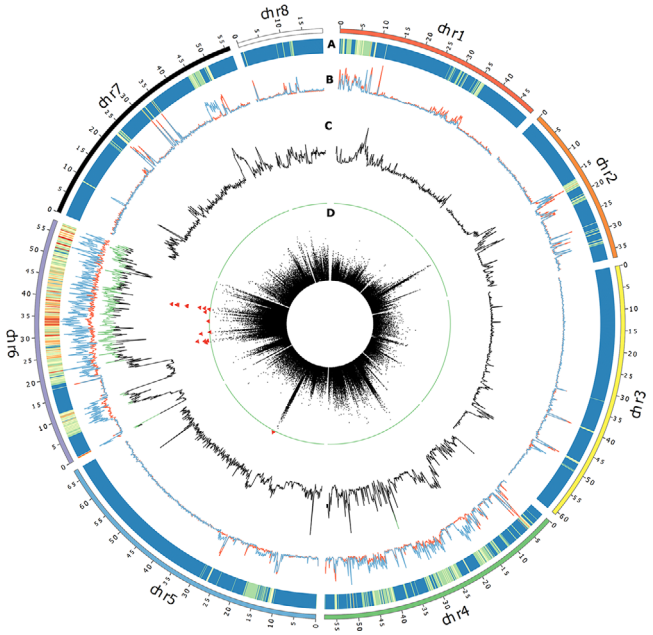

FIGURE 1 Genome variation and genome-wide association studies (GWAS) based on whole genome resequencing data. A: Single nucleotide polymorphisms (SNPs) density. B: Nucleotide diversity ($\theta\pi$) of the 50 greatest yielding lines (in red) and the 50 least yielding lines (in blue). C: Fixation index (Fst) genome scan of yield based on the 50 greatest and 50 least yielding lines. Extreme Fst values larger than 0.4 is highlighted in green. D: Circular Manhattan plots displaying the GWAS-yield result. The SNP are represented by black dot, while the SNP with significant association is represented by a red triangle (p-values lower than 1.61×10^{-5})

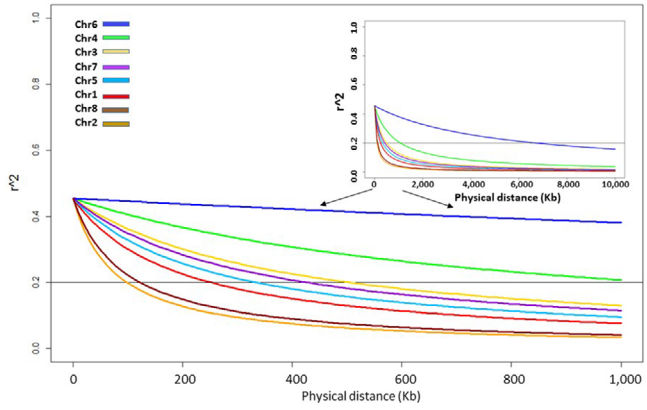

FIGURE 2 The extent of linkage disequilibrium (LD) in the eight chickpea chromosomes. The horizontal axis indicates the physical distance (kb), the vertical axis shows the LD between single nucleotide polymorphisms markers measured as r^2. The black horizontal lines represent the LD extent threshold ($r^2 = 0.2$). The curves are the trend of LD extent in each chromosome fitted using the mutation-drift-equilibrium model described in Li et al. (2011). The extended view of LD extent is shown in the top right

were included in the current analysis. As expected, PI4899777 separated clearly from the cultivated species *Cicer arietinum* (Figure 3, Supplementary Table S2). Most desi cultivars released before 2005 clustered together whereas the modern desi cultivars released by Pulse Breeding Australia after 2005 were more diverse. The 315 S1, S2, and S3 advanced breeding lines spread across different clusters indicating the chickpea breeding program has made significant progress recently on increasing the genetic diversity within the breeding program. A cluster of 70 lines from S1 and S2 clustered with the line 04067-81-2-1-1_B, a backcross derivative from the wild chickpea species *C. echinosperum*. This line is resistant to phytophthora root rot—a major root disease of chickpea in Australia (Amalraj et al., 2019). These 70 lines are progeny of 04067-81-2-1-1_B or its sister lines and have higher phytophthora root rot resistance levels and yet slightly lower yield than other lines in the analysed collection (Figure 3, Supplementary Table S1).

3.3 | Selection signatures of grain yield

The average yield of the 315 advanced lines in different year and location combinations ranged from 0.48 t/ha in the rainfed trial 2012MO to 4.13 t ha^{-1} in the irrigated trial 2012HM (Table 2, BB: Billa Billa; HM: Hermitage; MO: Moree). The genotypic effect for all trials was highly significant and correlations of yield among different trials were mostly positive and significant except for trials 2012HM and 2013HM (Table 3). The GGE biplot analysis of multi-environment trials (METs) also showed the 2012HM and 2013HM trials clearly separated from other trials. The 2012HM trial had a negative correlation (opposite direction) with the rest of the trials (Supplementary Figure S1), pointing to a strong G × E effect.

Genome-wide association study was performed for adjusted entry means of grain yield obtained only from the rainfed yield trials 2012BB, 2013BB, 2012MO, and 2013MO to avoid a strong G × E effect introduced by irrigated yield trials 2012HM and 2013HM. We identified 20 significant SNPs in total; 19 on chromosome 6 and one on chromosome 5 (Figure 1, Supplementary Table S3). Natural and human selection can shape the genome and often leave signals of selection. The nucleotide diversity (θ_π) of the 50 greatest yielding lines was lower than that of the 50 least yielding lines in specific regions of the genome (Figure 1). One representative region is a 43.8 Mb region on chromosome 6 spanning chromosome 6:18,00,000 to 61,800,000 with high SNP density. This region was further confirmed by the Fst genome scan, an allele frequency approach to identify selection signatures, showing divergence between high and low yielding groups. Haplotype analysis revealed two major

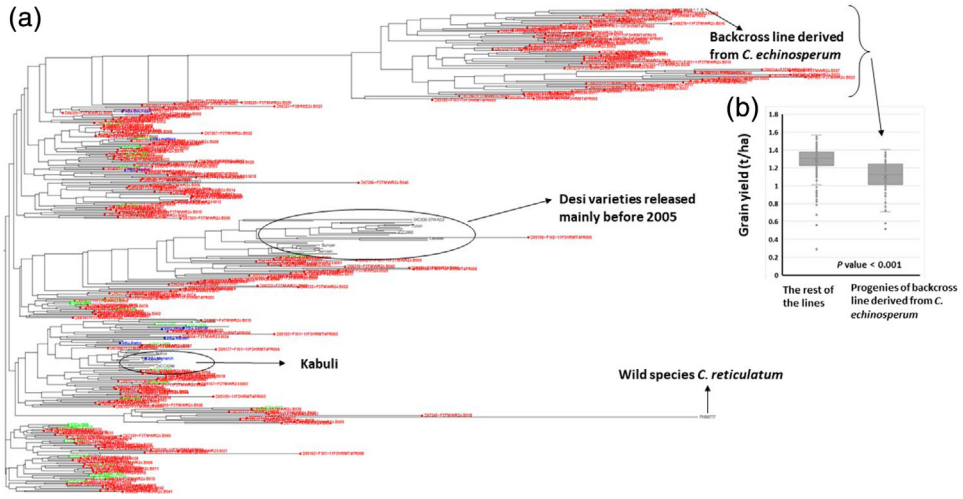

FIGURE 3 (a) Phylogenetic tree of the 315 lines and the Australian chickpea cultivars from Li et al. (2017) based on 57K high-quality single nucleotide polymorphisms. (b) Boxplot of grain yield in the progenies of the backcross line derived from *C. echinosperum* and the rest of the lines. The 58 released Australian cultivars and wild species *C. reticulatum* are colored in black. Stage 1 and Stage 2 lines are colored in red. Stage 3 lines are colored in green. The Pulse Breeding Australia cultivars are colored in blue. The name of the lines can be found in Supplemental Table S2

TABLE 2 Summary of phenotypic data

Trials	Traits	No. of lines	Mean (median)	Min.	Max.	Genotypic effect
2012BB	Yield (t ha^{-1})	128	1.98 (1.94)	1.64	2.45	<0.001
2012HM	Yield (t ha^{-1})	128	4.13 (4.18)	3.40	4.59	<0.001
2012MO	Yield (t ha^{-1})	128	0.48 (0.46)	0.16	0.78	<0.001
2013BB	Yield (t ha^{-1})	315	1.53 (1.57)	0.42	1.90	<0.001
2013HM	Yield (t ha^{-1})	315	1.94 (1.88)	0.64	3.41	<0.001
2013MO	Yield (t ha^{-1})	315	1.15 (1.19)	0.35	1.58	<0.001

Note. BB, HM, and MO represent yield trials at Billa Billa (QLD), Hermitage (QLD), and Moree (NSW), respectively. 2012 and 2013 are the years that the yield trials were conducted.

haplotypes in this chromosome 6 region; one haplotype shared by the wide species *C. echinospermum* backcross-derivative 04067-81-2-1-1_B, the other representing the *C. arietinum* lines in the collection. Approximately, two thirds of the 50 least yielding lines contain the wild-species haplotype. The wild species *C. echinospermum* is known for providing sources of resistance to Phytophthora root rot for the Australian chickpea breeding program. This 43.8 Mb region is likely an introgression segment that introduced from *C. echinospermum* for Phytophthora root rot resistance, yet it also has the unintended consequence of reducing yield due to linkage drag. Previous studies have identified molecular markers that can help in selecting and introgressing Phytophthora root rot resistance more precisely in chickpea (Amalraj et al., 2019). It remains to be seen whether maker-assisted selection can break the linkage drag in this region.

3.4 | Genomic selection of grain yield

Because of the polygenic nature of yield, a genomic selection/prediction approach was used to better understand the underlying mechanisms of inheritance. To mimic a common practice of evaluating advanced lines in plant breeding programs, earlier datasets were used as training sets to predict latter datasets. The 2012BB and 2012MO yield datasets (training sets) predicted the 2013BB and 2013MO datasets (target sets) with accuracies ranging from $r = 0.14$ to 0.39, whereas they predicted the 2013HM dataset poorly ($r = -0.02$ to 0.12, Table 4). The 2012HM dataset predicted the 2013 datasets poorly, reiterating the Biplot GGE analysis that showed the G × E effect among BB and MO trials to be much smaller than that of the 2012HM and 2013HM trials, an effect most likely due to the rainfed vs. irrigated nature of the trials.

TABLE 3 Correlations of grain yield in six different trials with 128 overlapping lines

	2012BB	2012HM	2012MO	2013BB	2013HM	2013MO
2012BB	–					
2012HM	−0.36***	–				
2012MO	0.38***	−0.36***	–			
2013BB	0.42***	−0.10	0.18*	–		
2013HM	0.43***	−0.24**	0.33***	−0.01	–	
2013MO	0.51***	−0.08	0.25**	0.50***	0.21*	–

Note. BB, HM, and MO represent yield trials at Billa Billa (QLD), Hermitage (QLD), and Moree (NSW), respectively. 2012 and 2013 are the years that the yield trials were conducted. Two-sided test of correlations different from zero.
*$p < .05$.
**$p < .01$.
***$p < .001$.

TABLE 4 Prediction accuracies for the yield of the new advanced lines using their close relatives as training sets

Training sets ($N = 128$)	Target sets ($N = 187$)	RR-BLUP	BL	BRR
2012BB	2013BB	0.33***	0.34***	0.31***
2012BB	2013HM	0.08†	0.10†	0.06†
2012BB	2013MO	0.38***	0.39***	0.37***
2012HM	2013HM	−0.02†	−0.01†	−0.02†
2012HM	2013BB	−0.32***	−0.06 ns	−0.28***
2012HM	2013MO	−0.34***	−0.04 ns	−0.28***
2012MO	2013MO	0.22**	0.17**	0.19***
2012MO	2013BB	0.19*	0.14*	0.17*
2012MO	2013HM	0.10†	0.12†	0.07†

Note. BB, HM, and MO represent yield trials at Billa Billa (QLD), Hermitage (QLD), and Moree (NSW), respectively. 2012 and 2013 are the years that the yield trials were conducted. The training sets consist of 128 breeding lines from 2012 Stage 1 and 2; the target sets consist of 187 new breeding lines from 2013 Stage 1. Two-sided test of correlations different from zero. RR-BLUP, ridge regression best linear unbiased estimation; BL = Bayesian least absolute shrinkage and selection operation; BRR = Bayesian ridge regression.
*$p < .05$.
**$p < .01$.
***$p < .001$.
†ns, nonsignificant at $p > 0.05$.

To further investigate the G × E effect, training sets (2012_BB_HM_MO and 2012_BB_MO), comprising the adjusted means (BLUE) of genotypes from the three 2012 trials (2012BB, 2012HM, and 2012MO), were used to predict 2013 trials 2013BB, 2013HM, 2013MO, and 2013_BB_HM_MO (Table 5, Table 6). Depending on the environment used, prediction accuracies either increased or decreased compared with analysis using a single 2012 trial as a training set. Prediction accuracies increased when 2012_BB_MO (compared with 2012_BB or 2012_HM or 2012_MO) was used as a training set to predict 2013BB, 2013HM, and 2013MO. However, prediction accuracies decreased when using 2012_BB_HM_MO (compared with 2012_BB or 2012_HM or 2012_MO) as a training set to predict 2013BB, 2013HM, and 2013MO. This observation can be explained by G × E interaction from the 2012HM trial (rainfed *vs* irrigated). This finding has important implications for the process of applying a training set across multiple environments. Evaluation of a training set in multiple environments can increase prediction accuracy as long as the environments have at least an intermediate phenotypic correlation (Table 3).

To test the effect of SNP functions (based on their location in the genome) on prediction accuracy, the 298,154 SNPs were divided into different classes: (a) exome+3′&5′UTR+MNase-HS, (b) exome, and (c) missense + alternative splicing in the exome. Prediction accuracies were estimated using a cross-validation approach based on these different classes and yield data from three yield trials. Prediction accuracies using all SNPs in different trials ranged from 0.30 in 2013HM to 0.58 in 2013MO (Table 6). Prediction accuracies using the three different classes did not differ

TABLE 5 Prediction accuracy of the yield of new advanced lines using close relatives in the presence of G × E effects

Training set (N = 128)	Target set (N = 187)	RR-BLUP	BL	BRR
2012_BB_HM_MO	2013_BB_HM_MO	0.23**	0.27***	0.22**
2012_BB_HM_MO	2013BB	0.18*	0.23**	0.16*
2012_BB_HM_MO	2013HM	0.10†	0.11†	0.11†
2012_BB_HM_MO	2013MO	0.25***	0.30***	0.23
2012_BB_MO	2013_BB_MO	0.48***	0.49***	0.45***
2012_BB_MO	2013BB	0.45***	0.46***	0.41***
2012_BB_MO	2013HM	0.15†	0.17†	0.12†
2012_BB_MO	2013MO	0.49***	0.49***	0.46***

Note. BB, HM, and MO represent yield trials at Billa Billa (QLD), Hermitage (QLD), and Moree (NSW), respectively. 2012 and 2013 are the years that the yield trials were conducted. The training sets consist of 128 breeding lines from 2012 Stage 1 and Stage 2; the target sets consist of 187 new breeding lines from 2013 stage 1. Two-sided test of correlations different from zero. RR-BLUP, ridge regression best linear unbiased estimation; BL = Bayesian least absolute shrinkage and selection operation; BRR = Bayesian ridge regression.

*$p < .05$.

**$p < .01$.

***$p < .001$.

†ns, nonsignificant at $p > .05$.

TABLE 6 The effects of single nucleotide polymorphism (SNP) functional annotations on prediction accuracies of grain yield using a five-fold cross-validation approach[a]

	SNP locations in the genome (no. of SNPs)			
Trials	Whole-genome (298K)	Exome+3′&5′UTR+MNase-HS[b] (117K)	Exome (29K)	Missense & alternative splicing in exome (11K)
2013BB	0.57 ± 0.01	0.57 ± 0.01	0.58 ± 0.00	0.56 ± 0.01
2013HM	0.30 ± 0.01	0.27 ± 0.01	0.32 ± 0.01	0.28 ± 0.01
2013MO	0.58 ± 0.01	0.57 ± 0.01	0.58 ± 0.01	0.55 ± 0.00

[a]Five-fold cross-validation: the dataset was divided into five subsets; one of which was used as a target set while the rest of four were used as a training set to train the ridge regression best linear unbiased estimation model. This process was repeated 10 times resulting in 50 cross-validation. Prediction accuracy was then estimated as Pearson's correlation coefficient between the predicted values and observed phenotypic values of the target set. [b]Micrococcal nuclease hypersensitive regions (MNase-HS) are open chromatin regions that are often associated with the level of gene expression and recombination hotspots, thus explaining a large proportion of heritable variance (Rodgers-Melnick et al., 2016).

significantly compared with using all SNPs, regardless of trial datasets.

4 | DISCUSSION

Chickpea was first introduced to Australia in the late 1970s and has gradually become the most widely grown pulse in recent years. Being a relatively newly cultivated crop in Australia, chickpea is a good example to study how crop species adapt to new environments. Genetic diversity is essential for a plant breeding program, which is particularly true for inbreeding crops such as chickpea, that are often characterized by narrow genetic diversity following domestication. Compared with historical Australian chickpea cultivars, we were able to demonstrate a substantial genetic diversity present in the Australian chickpea breeding program, a result based on genetic analysis of 315 advanced breeding lines with whole-genome resequencing data. This has important implications for implementing GS in the chickpea breeding program, as genetic diversity was shown to be rapidly depleted when a GS breeding scheme was implemented in wheat and maize breeding program (Daetwyler et al., 2015; Muller et al., 2018).

4.1 | Genome-wide association study of grain yield

In general, mapping resolution based on association study is determined mainly by population size, the number of DNA markers across the genome, and the extent of LD surrounding causal loci. The GWAS studies based on low-density marker often identify markers far away from causal genes and thus could not be validated/used in the application populations due to low levels of linkage disequilibrium between markers and causal genes. For a similar reason, it is recommended to use a

large number of markers in GS when resources allow (Lorenz, 2013; Riedelsheimer et al., 2013). In this study, more than 298K SNPs were discovered in 315 advanced breeding lines using a whole-genome resequencing method and subsequent GWAS identified 20 SNPs significantly associated with grain yield. These SNPs could be grouped into five regions based on the extent of surrounding LD (900 Kb, $r^2 = 0.4$). In total, 158 functionally annotated genes were found in these regions and 47 of them have been deposited in the KEGG database (Supplementary Table S4). They were classified into six major functional categories—protein families: genetic information processing (15), protein families: signaling and cellular processes (7), genetic information processing (6), carbohydrate metabolism (3), cellular processes (3), and other (13). A few candidate genes were involved in grain yield based on the literature. For example, sugar transporters are members of the major facilitator superfamily that play an important role in source sink partitioning, plant growth, and seed development (Doidy et al., 2012; Wobus et al., 1999). The E2FB transcription factor is a cell growth and cell division regulator in Arabidopsis and found to accelerate flowering and increase fruit yield in tomato (*Solanum lycopersicum* L.) (Abraham et al., 2012; Magyar et al., 2005). However, the involvement of those candidate genes in determining grain yield in chickpea should be interpreted with caution, and further functional validation such as RNAi or CRISPR-Cas9 is needed to confirm their influence on grain yield. In addition, a fine mapping approach using diverse material with more genetic recombinations could be used toward narrowing down the regions.

4.2 | Genomic selection based on whole-genome resequencing data

Many of the papers published to date regarding genomic selection in plants are based on DNA markers generated through GBS or SNP array platforms (Crossa et al., 2014; Norman et al., 2018; Roorkiwal et al., 2016). Although cost-effective, these platforms produce a relatively small number of markers and often have limitations, such as a large proportion of missing values in the case of GBS, or ascertainment bias towards SNPs present in the populations used for discovery during the SNP array design phase (Chu et al., 2020; Malomane et al., 2018). Ascertainment bias of SNP array is well documented and could affect estimating any population genetic parameters that rely on the site frequency spectrum, such as nucleotide diversity, Fst, and LD (Clark et al., 2005; Lachance et al., 2013). As the cost of sequencing technology continues to fall, it is becoming more practical to use a much larger number of DNA markers from whole-genome resequencing platforms in genomic selection, particularly for species with a small genome size such as chickpea and rice. However, beyond the associated higher genotyping cost, there are some challenges when implementing greater numbers of markers in a GS strategy. Most current GS software packages such as Synbreed, BGLR, and GAPIT are designed for low-density marker applications and not suitable for parallel computing. For example, it took approximately 4 h (a PC with 2.9-GHz CPU with 16 GB RAM) to perform a five-fold cross-validation for ~300 genotypes with 298K SNP using the BL model from the Synbreed package. The most time-consuming part of the Bayesian analysis is the Markov chain Monte Carlo process where the posterior distributions are approximated. A fast and efficient algorithm has been developed in other statistical applications and could be employed in genomic prediction (Calderhead, 2014; Quiroz et al., 2019; Waldmann et al., 2008).

Another challenge of using a large number of markers relates to overfitting a model, where residual variances are assigned to the markers unintentionally (Hickey et al., 2014). While overfitting a model could produce high prediction accuracy in the training population, it will likely perform poorly on the other datasets such as selection candidates. Some studies have shown that Bayesian models, employing either variable selection or shrinkage procedures, were able to ease the problems of overfitting a model and increased prediction accuracy while others failed to achieve that (Fikere et al., 2018; Wimmer et al., 2013). A few published studies present approaches to reduce the total number of markers by choosing a subset of markers with significant marker–trait associations (Y. Li et al., 2018; Spindel et al., 2016). For example, Y. Li et. al (2018) showed that using subsets of SNPs significantly associated (p-values between .05–.01) with yield and yield components in chickpea increased prediction accuracies up to two-fold compared with that using the full 144K SNP dataset.

Another approach involves preselection of markers based on functional annotation. A study using 28.3 million whole genome sequence variants from more than 16,000 dairy cattle found that sequence variants located in gene coding and regulatory regions (particularly the alternative splice site) explained a larger percentage of the total variance than expected by chance (Koufariotis et al., 2018). But other studies did not find significant improvement on prediction and suggested that this approach is likely trait-dependent due to different genetic architecture of each trait (de las Heras-Saldana et al., 2020; Do et al., 2015). In this study, 298K SNPs were subdivided into several classes according to their functional annotations. Compared with the 298K whole-genome SNPs, the subsets of markers did not influence prediction accuracy of gain yield significantly despite consisting of a much smaller number of markers. For example, the prediction accuracy based on 29K SNP located only in the exome of the chickpea genome was as good as accuracy derived from all SNP in all four trials. This suggests that an

exome-capture genotyping platform could generate prediction accuracy as high as WGS. This could potentially decrease the cost of genotyping significantly, particularly in crop species with a large and complex genome such as wheat and oat. It is worth pointing out that extensive LD in this germplasm could limit the ability to study prediction accuracies among various functional classes, however, the extent of LD can vary greatly in local regions and some SNPs within the same gene have very low LD.

4.3 | Updating the training set with phenotypes from relevant environments for genomic selection

In the context of genomic selection, a training set is a collection of germplasm with genotypic and phenotypic data. The marker effects are estimated/calibrated in the training set and used to predict the phenotypes or breeding values for a target set of germplasm using only genotypic data. Therefore, the training set should have a close relationship with the target set so that marker effects estimated in the training set are applicable to the target set where DNA markers are still in high LD with the causal genes. For example, genomic prediction of testcross values of maize was halved when less related material of target set was used (Albrecht et al., 2011). Similar results were observed in wheat, rice (Berro et al., 2019; Crossa et al., 2014; Norman et al., 2018), and livestock (Meuwissen et al., 2016; Toosi et al., 2010). In this study, advanced lines were used to predict the next cohort of advanced lines in the same breeding program with encouraging results.

Another important factor that needs to be considered when updating the training set is the environment where phenotypes are generated. For complex traits with low heritability, such as grain yield, environmental and G × E variances are usually large. In this study, we showed that adding another unrelated yield trial (rainfed vs. irrigated) in the training set did not increase but instead decreased prediction accuracy probably due to G × E interaction. A similar phenomenon was observed in barley where N. Heslot et al. (2013) addressed this issue by using a new method to remove less predictive environments from the training set. These examples suggest that careful consideration is needed when updating the training set with new phenotypes. The G × E interaction is a long-standing problem that complicates the plant breeding process. This is conventionally addressed in plant breeding through evaluation of METs so that these environments capture the true breeding values/phenotypes of the plants. Plant physiologists tend to view G × E interaction from the angle of phenotypic plasticity, where plant genotypes perform differently in response to environmental gradients and often use a reaction-norm model to handle G × E interaction (Jarquin et al., 2014; Sadras et al., 2016). X. Li et al. (2018) went a step further to make use of genomic information to reveal the interplay of genomic and environmental gradients on G × E interaction. If G × E is present in the target environment, the choice of which environments to include in the training set is critical for prediction, as demonstrated here in this study. Thanks to the rapid development of sensing and monitoring technology on-farm, environmental characterizations (climatic data, soil parameters, nitrogen level, etc.) in the field are becoming easily accessible and could be useful for choosing appropriate environments where the training set is to be evaluated. These environmental conditions could show differences in abiotic stresses such as drought, heat, and frost. N. Heslot et al. (2014) proposed an integrated model where environmental variables and crop growth modeling were used in a genomic selection framework to handle G × E interaction in wheat. Accuracy in predicting line performance in untested environments with climatic data increased by about 11% on average. Another recent study further improved this model by selecting a subset of environmental variables derived from a crop growth model and calculating the covariance matrices using an AMMI decomposition method (Rincent et al., 2019). X. Li et al. (2018) included an environmental index defined by photothermal time in the reaction-norm model and found that they can predict the flowering time of a sorghum recombinant inbred line mapping population in untested environments as high as $r = 0.74$ on average. It will be interesting to see if this approach also works well with a more complex trait such as yield in more diverse germplasm.

ACKNOWLEDGMENTS

This study was supported by grant GCF010013 through the Australia-India Strategic Research Fund (AISRF), Australian Government Department of Industry, Innovation and Science. The Pulse Breeding Australia chickpea breeding program (now Chickpea Breeding Australia) is funded by GRDC and NSW DPI. We thank Hans D. Daetwyler, John Harris, and Julie Hayes for their constructive comments on the manuscript.

AUTHOR CONTRIBUTIONS

Yongle Li: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Resources; Software; Visualization; Writing-original draft; Writing-review & editing. Pradeep Ruperao: Formal analysis; Software. Jacqueline Batley: Data curation; Resources; Supervision. David Edwards: Funding acquisition; Resources; Supervision; Writing-review & editing. William Martin: Data curation; Investigation; Resources. Kristy Hobson: Data curation; Funding acquisition; Investigation; Project administration; Resources; Supervision; Writing-review & editing. Tim Sutton: Conceptualization; Funding acquisition; Investigation; Project administration; Resources; Supervision; Writing-review & editing.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORCID

Yongle Li https://orcid.org/0000-0003-3966-8596
David Edwards https://orcid.org/0000-0001-7599-6760

REFERENCES

Abraham, Z., & del Pozo, J. C. (2012). Ectopic expression of E2FB, a cell cycle transcription factor, accelerates flowering and increases fruit yield in tomato. *Journal of Plant Growth Regulation*, *31*(1), 11–24. https://doi.org/10.1007/s00344-011-9215-y

Albrecht, T., Wimmer, V., Auinger, H. J., Erbe, M., Knaak, C., Ouzunova, M., Simianer, H., & Schon, C. C. (2011). Genome-based prediction of testcross values in maize. *Theoretical and Applied Genetics*, *123*(2), 339–350. https://doi.org/10.1007/s00122-011-1587-7

Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, *12*, 246. https://doi.org/10.1186/1471-2105-12-246

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Amalraj, A., Taylor, J., Bithell, S., Li, Y., Moore, K., Hobson, K., & Sutton, T. (2019). Mapping resistance to Phytophthora root rot identifies independent loci from cultivated (*Cicer arietinum* L.) and wild (*Cicer echinospermum* P.H. Davis) chickpea. *Theoretical and Applied Genetics*, *132*(4), 1017–1033. https://doi.org/10.1007/s00122-018-3256-6

Asoro, F. G., Newell, M. A., Beavis, W. D., Scott, M. P., Tinker, N. A., & Jannink, J. L. (2013). Genomic, marker-assisted, and pedigree-BLUP selection methods for beta-glucan concentration in elite oat. *Crop Science*, *53*(5), 1894–1906. https://doi.org/10.2135/cropsci2012.09.0526

Atieno, J., Li, Y., Langridge, P., Dowling, K., Brien, C., Berger, B., Varshney, R. K., & Sutton, T. (2017). Exploring genetic variation for salinity tolerance in chickpea using image-based phenotyping. *Scientific Reports*, *7*(1), 1300. https://doi.org/10.1038/s41598-017-01211-7

Barton, N. H., Etheridge, A. M., & Veber, A. (2017). The infinitesimal model: Definition, derivation, and implications. *Theoretical Population Biology*, *118*, 50–73. https://doi.org/10.1016/j.tpb.2017.06.001

Berro, I., Lado, B., Nalin, R. S., Quincke, M., & Gutierrez, L. (2019). Training population optimization for genomic selection. *Plant Genome*, *12*(3). https://doi.org/10.3835/plantgenome2019.04.0028

Calderhead, B. (2014). A general construction for parallelizing Metropolis-Hastings algorithms. *Proceedings of the National Academy of the Sciences of the United States of America*, *111*(49), 17408–17413. https://doi.org/10.1073/pnas.1408184111

Chu, J. T., Zhao, Y. S., Beier, S., Schulthess, A. W., Stein, N., Philipp, N., Roder, M. S., & Reif, J. C. (2020). Suitability of single-nucleotide polymorphism arrays versus genotyping-by-sequencing for genebank genomics in wheat. *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/fpls.2020.00042

Cingolani, P., Platts, A., le Wang, L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H., & Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research*, *15*(11), 1496–1502. https://doi.org/10.1101/gr.4107905

Crossa, J., Perez, P., Hickey, J., Burgueno, J., Ornella, L., Ceron-Rojas, J., Zhang, X., Dreisigacker, S., Babu, R., Li, Y., Bonnett, D., & Mathews, K. (2014). Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity*, *112*(1), 48–60. https://doi.org/10.1038/hdy.2013.16

Daetwyler, H. D., Hayden, M. J., Spangenberg, G. C., & Hayes, B. J. (2015). Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. *Genetics*, *200*(4), 1341. https://doi.org/10.1534/genetics.115.178038

de las Heras-Saldana, S., Lopez, B. I., Moghaddar, N., Park, W., Park, J. E., Chung, K. Y., Lim, D., Lee, S. H., Shin, D., & van der Werf, J. H. J. (2020). Use of gene expression and whole-genome sequence information to improve the accuracy of genomic prediction for carcass traits in Hanwoo cattle. *Genetics Selection Evolution*, *52*(1). https://doi.org/10.1186/s12711-020-00574-2

de Los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., & Calus, M. P. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, *193*(2), 327–345. https://doi.org/10.1534/genetics.112.143313

Do, D. N., Janss, L. L. G., Jensen, J., & Kadarmideen, H. N. (2015). SNP annotation-based whole genomic prediction and selection: An application to feed efficiency and its component traits in pigs. *Journal of Animal Science*, *93*(5), 2056–2063. https://doi.org/10.2527/jas.2014-8640

Doidy, J., Grace, E., Kuehn, C., Simon-Plas, F., Casieri, L., & Wipf, D. (2012). Sugar transporters in plants and in their interactions with fungi. *Trends in Plant Science*, *17*(7), 413–422. https://doi.org/10.1016/j.tplants.2012.03.009

Fikere, M., Barbulescu, D. M., Malmberg, M. M., Shi, F., Koh, J. C. O., Slater, A. T., MacLeod, I. M., Bowman, P. J., Salisbury, P. A., Spangenberg, G. C., Cogan, N. O. I., & Daetwyler, H. D. (2018). Genomic prediction using prior quantitative trait loci information reveals a large reservoir of underutilised blackleg resistance in diverse canola (*Brassica napus* L.) lines. *Plant Genome*, *11*(2). https://doi.org/10.3835/plantgenome2017.11.0100

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., & Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science*, *92*(2), 433–443. https://doi.org/10.3168/jds.2008-1646

Heslot, N., Akdemir, D., Sorrells, M. E., & Jannink, J. L. (2014). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theoretical and Applied Genetics*, *127*(2), 463–480. https://doi.org/10.1007/s00122-013-2231-5

Heslot, N., Jannink, J. L., & Sorrells, M. E. (2013). Using genomic prediction to characterize environments and optimize prediction accuracy in applied breeding data. *Crop Science*, *53*(3), 921–933. https://doi.org/10.2135/cropsci2012.07.0420

Heslot, N., Yang, H.-P., Sorrells, M. E., & Jannink, J.-L. (2012). Genomic selection in plant breeding: A comparison of models. *Crop Science*, *52*, 146. https://doi.org/10.2135/cropsci2011.06.0297

Hickey, J. M., Dreisigacker, S., Crossa, J., Hearne, S., Babu, R., Prasanna, B. M., Grondona, M., Zambelli, A., Windhausen, V. S., Mathews, K., & Gorjanc, G. (2014). Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. *Crop Science*, *54*(4), 1476–1488. https://doi.org/10.2135/cropsci2013.03.0195

Hill, W. G. (2014). Applications of population genetics to animal breeding, from wright, fisher and lush to genomic prediction. *Genetics*, *196*(1), 1–16. https://doi.org/10.1534/genetics.112.147850

Hoffstetter, A., Cabrera, A., Huang, M., & Sneller, C. (2016). Optimizing training population data and validation of genomic selection for economic traits in soft winter wheat. *G3 Genes|Genomes|Genetics*, *6*(9), 2919–2928. https://doi.org/10.1534/g3.116.032532

Jarquin, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Perez, P., Calus, M., Burgueno, J., & de los Campos, G. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics*, *127*(3), 595–607. https://doi.org/10.1007/s00122-013-2243-1

Khan, H. A., Siddique, K. H. M., & Colmer, T. D. (2016). Salt sensitivity in chickpea is determined by sodium toxicity. *Planta*, *244*(3), 623–637. https://doi.org/10.1007/s00425-016-2533-3

Kiran, A., Kumar, S., Nayyar, H., & Sharma, K. D. (2019). Low temperature-induced aberrations in male and female reproductive organ development cause flower abortion in chickpea. *Plant, Cell and Environment*, *42*(7), 2075–2089. https://doi.org/10.1111/pce.13536

Koufariotis, L. T., Chen, Y. P. P., Stothard, P., & Hayes, B. (2018). Variance explained by whole genome sequence variants in coding and regulatory genome annotations for six dairy traits. *BMC Genomics*, *19*. https://doi.org/10.1186/s12864-018-4617-x

Lachance, J., & Tishkoff, S. A. (2013). SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *Bioessays*, *35*(9), 780–786. https://doi.org/10.1002/bies.201300014

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Proc, G. P. D. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, X., Guo, T. T., Mu, Q., Li, X. R., & Yu, J. M. (2018). Genomic and environmental determinants and their interplay underlying phenotypic plasticity. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(26), 6679–6684. https://doi.org/10.1073/pnas.1718326115

Li, Y., Haseneyer, G., Schön, C. C., Ankerst, D., Korzun, V., Wilde, P., & Bauer, E. (2011). High levels of nucleotide diversity and fast decline of linkage disequilibrium in rye (*Secale cereale* L.) genes involved in frost response. *BMC Plant Biology*, *11*, 6. https://doi.org/10.1186/1471-2229-11-6

Li, Y., Ruperao, P., Batley, J., Edwards, D., Davidson, J., Hobson, K., & Sutton, T. (2017). Genome analysis identified novel candidate genes for ascochyta blight resistance in chickpea using whole genome re-sequencing data. *Frontiers in Plant Science*, *8*. https://doi.org/10.3389/fpls.2017.00359

Li, Y., Ruperao, P., Batley, J., Edwards, D., Khan, T., Colmer, T. D., Pang, J. Y., Siddique, K. H. M., & Sutton, T. (2018). Investigating drought tolerance in chickpea using genome-wide association mapping and genomic selection based on whole-genome resequencing data. *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.00190

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., Gore, M. A., Buckler, E. S., & Zhang, Z. (2012). GAPIT: Genome association and prediction integrated tool. *Bioinformatics*, *28*(18), 2397–2399. https://doi.org/10.1093/bioinformatics/bts444

Lorenz, A. J. (2013). Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: A simulation experiment. *G3 Genes|Genomes|Genetics*, *3*(3), 481–491. https://doi.org/10.1534/g3.112.004911

Magyar, Z., De Veylder, L., Atanassova, A., Bako, L., Inze, D., & Bogre, L. (2005). The role of the Arabidopsis E2FB transcription factor in regulating auxin-dependent cell division. *Plant Cell*, *17*(9), 2527–2541. https://doi.org/10.1105/tpc.105.033761

Malomane, D. K., Reimer, C., Weigend, S., Weigend, A., Sharifi, A. R., & Simianer, H. (2018). Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. *BMC Genomics*, *19*. https://doi.org/10.1186/s12864-017-4416-9

Meuwissen, T., Hayes, B., & Goddard, M. (2016). Genomic selection: A paradigm shift in animal breeding. *Animal Frontiers*, *6*(1), 6–14. https://doi.org/10.2527/af.2016-0002

Meuwissen, T., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829.

Muller, D., Schopp, P., & Melchinger, A. E. (2018). Selection on expected maximum haploid breeding values can increase genetic gain in recurrent genomic selection. *G3-Genes Genomes Genetics*, *8*(4), 1173–1181. https://doi.org/10.1534/g3.118.200091

Norman, A., Taylor, J., Edwards, J., & Kuchel, H. (2018). Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy. *G3 (Bethesda)*, *8*(9), 2889–2899. https://doi.org/10.1534/g3.118.200311

Norman, A., Taylor, J., Tanaka, E., Telfer, P., Edwards, J., Martinant, J. P., & Kuchel, H. (2017). Increased genomic prediction accuracy in wheat breeding using a large Australian panel. *Theoretical and Applied Genetics*, *130*(12), 2543–2555. https://doi.org/10.1007/s00122-017-2975-4

Pang, J., Turner, N. C., Khan, T., Du, Y. L., Xiong, J. L., Colmer, T. D., Devilla, R., Stefanova, K., & Siddique, K. H. (2016). Response of chickpea (*Cicer arietinum* L.) to terminal drought: Leaf stomatal conductance, pod abscisic acid concentration, and seed set. *Journal of Experimental Botany*. https://doi.org/10.1093/jxb/erw153

Quiroz, M., Kohn, R., Villani, M., & Tran, M. N. (2019). Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association*, *114*(526), 831–843. https://doi.org/10.1080/01621459.2018.1448827

Riedelsheimer, C., & Melchinger, A. E. (2013). Optimizing the allocation of resources for genomic selection in one breeding cycle. *Theoretical and Applied Genetics*, *126*(11), 2835–2848. https://doi.org/10.1007/s00122-013-2175-9

Rincent, R., Malosetti, M., Ababaei, B., Touzy, G., Mini, A., Bogard, M., Martre, P., Le Gouis, J., & van Eeuwijk, F. (2019). Using crop growth model stress covariates and AMMI decomposition to better predict genotype-by-environment interactions. *Theoretical and Applied Genetics*, *132*(12), 3399–3411. https://doi.org/10.1007/s00122-019-03432-y

Rodgers-Melnick, E., Vera, D. L., Bass, H. W., & Buckler, E. S. (2016). Open chromatin reveals the functional maize genome. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(22), E3177–3184. https://doi.org/10.1073/pnas.1525244113

Roorkiwal, M., Rathore, A., Das, R. R., Singh, M. K., Jain, A., Srinivasan, S., Gaur, P. M., Chellapilla, B., Tripathi, S., Li, Y., Hickey, J. M., Lorenz, A., Sutton, T., Crossa, J., Jannink, J. L., & Varshney, R.

K. (2016). Genome-enabled prediction models for yield related traits in chickpea. *Frontiers in Plant Science*, *7*, 1666. https://doi.org/10.3389/fpls.2016.01666

Sadras, V. O., Lake, L., Li, Y., Farquharson, E. A., & Sutton, T. (2016). Phenotypic plasticity and its genetic regulation for yield, nitrogen fixation and delta13C in chickpea crops under varying water regimes. *Journal of Experimental Botany*, *67*(14), 4339–4351. https://doi.org/10.1093/jxb/erw221

Spindel, J. E., Begum, H., Akdemir, D., Collard, B., Redona, E., Jannink, J. L., & McCouch, S. (2016). Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity*, *116*(4), 395–408. https://doi.org/10.1038/hdy.2015.113

Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., Su, Z., Pan, Y., Liu, D., Lipka, A. E., Buckler, E. S., & Zhang, Z. (2016). GAPIT Version 2: An enhanced integrated tool for genomic association and prediction. *Plant Genome*, *9*(2). doi:https://doi.org/10.3835/plantgenome2015.11.0120

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological*, *58*(1), 267–288.

Toosi, A., Fernando, R. L., & Dekkers, J. C. M. (2010). Genomic selection in admixed and crossbred populations. *Journal of Animal Science*, *88*(1), 32–46. https://doi.org/10.2527/jas.2009-1975

Van Belle, G., Fisher, L., Heagerty, P., & Lumley, T. (2004). *Biostatistics: A methodology for the health sciences* (2nd ed.). John Wiley & Sons, Inc.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*(1–3), 426–482.

Waldmann, P., Hallander, J., Hoti, F., & Sillanpaa, M. J. (2008). Efficient Markov chain Monte Carlo implementation of Bayesian analysis of additive and dominance genetic variances in noninbred pedigrees. *Genetics*, *179*(2), 1101–1112. https://doi.org/10.1534/genetics.107.084160

Wallace, T. C., Murray, R., & Zelman, K. M. (2016). The nutritional value and health benefits of chickpeas and hummus. *Nutrients*, *8*(12). https://doi.org/10.3390/nu8120766

Wang, Q., Tian, F., Pan, Y., Buckler, E. S., & Zhang, Z. (2014). A SUPER powerful method for genome wide association study. *Plos One*, *9*(9), e107684. https://doi.org/10.1371/journal.pone.0107684

Weigel, K. A., de los Campos, G., Vazquez, A. I., Rosa, G. J. M., Gianola, D., & Van Tassell, C. P. (2010). Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. *Journal of Dairy Science*, *93*(11), 5423–5435. https://doi.org/10.3168/jds.2010-3149

Wimmer, V., Lehermeier, C., Albrecht, T., Auinger, H. J., Wang, Y., & Schon, C. C. (2013). Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics*, *195*(2), 573–587. https://doi.org/10.1534/genetics.113.150078

Wobus, U., & Weber, H. (1999). Sugars as signal molecules in plant seed development. *Biological Chemistry*, *380*(7–8), 937–944. https://doi.org/10.1515/bc.1999.116

Zhang, W., Zhang, T., Wu, Y. F., & Jiang, J. M. (2012). Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in Arabidopsis. *Plant Cell*, *24*(7), 2719–2731. https://doi.org/10.1105/tpc.112.098061

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J. M., Arnett, D. K., Ordovas, J. M., & Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, *42*(4), 355–360. https://doi.org/10.1038/ng.546

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Li Y, Ruperao P, Batley J, et al. Genomic prediction of preliminary yield trials in chickpea: Effect of functional annotation of SNPs and environment. *Plant Genome*, 2021;e20166. https://doi.org/10.1002/tpg2.20166