the geneticssociety

www.nature.com/hdy

ARTICLE OPEN

Genomic prediction of cotton fibre quality and yield traits using Bayesian regression methods

Zitong Li[1]✉, Shiming Liu[2], Warren Conaty[2], Qian-Hao Zhu[1], Philippe Moncuquet[1], Warwick Stiller[2] and Iain Wilson[1]

© Crown 2022

Genomic selection or genomic prediction (GP) has increasingly become an important molecular breeding technology for crop improvement. GP aims to utilise genome-wide marker data to predict genomic breeding value for traits of economic importance. Though GP studies have been widely conducted in various crop species such as wheat and maize, its application in cotton, an essential renewable textile fibre crop, is still significantly underdeveloped. We aim to develop a new GP-based breeding system that can improve the efficiency of our cotton breeding program. This article presents a GP study on cotton fibre quality and yield traits using 1385 breeding lines from the Commonwealth Scientific and Industrial Research Organisation (CSIRO, Australia) cotton breeding program which were genotyped using a high-density SNP chip that generated 12,296 informative SNPs. The aim of this study was twofold: (1) to identify the models and data sources (i.e. genomic and pedigree) that produce the highest prediction accuracies; and (2) to assess the effectiveness of GP as a selection tool in the CSIRO cotton breeding program. The prediction analyses were conducted under various scenarios using different Bayesian predictive models. Results highlighted that the model combining genomic and pedigree information resulted in the best cross validated prediction accuracies: 0.76 for fibre length, 0.65 for fibre strength, and 0.64 for lint yield. Overall, this work represents the largest scale genomic selection studies based on cotton breeding trial data. Prediction accuracies reported in our study indicate the potential of GP as a breeding tool for cotton. The study highlighted the importance of incorporating pedigree and environmental factors in GP models to optimise the prediction performance.

Heredity (2022) 129:103–112; https://doi.org/10.1038/s41437-022-00537-x

INTRODUCTION

Cotton, primarily the tetraploid species *Gossypium hirsutum*, is widely grown in more than 70 countries. It is the world's leading renewable textile fibre crop (Paterson et al. 2012), as well as an important source of plant oil and protein (Jabran et al. 2019). The goal of cotton breeding is to achieve genetic gains for desirable fibre characteristics, disease resistance and yield traits in a time and cost-efficient manner. The Commonwealth Scientific and Industrial Research Organisation (CSIRO) cotton breeding program combines traditional field phenotyping and molecular marker assisted selection. Most economically important cotton traits are polygenic in nature making them difficult and expensive to manipulate and improve. As a result, a complete breeding cycle, from initial crossing to commercial release, takes a minimum of eight to ten years (Stiller and Wilson 2014). Genomic selection or genomic prediction (GP) is a relatively new molecular based breeding technology (Meuwissen et al. 2001). It has the potential to improve cotton breeding programs by efficiently allocating resources used in the breeding process by selecting individuals before being tested in field experiments based on genomic estimated breeding value (GEBV), reducing the breeding cycle interval (Crossa et al. 2017; Jannink et al. 2010) and the identification of genotypes that can be used as parents in future crosses to advance specific breeding objectives.

GP (Meuwissen et al. 2001) has been applied as a quantitative molecular breeding tool for crop improvement in various species including, but not limited to, wheat (Poland et al. 2012), maize (Millet et al. 2019), rice (Spindel et al. 2015), and soybean (Jarquín et al. 2014a). Briefly, GP uses a statistical predictive model based on a training population with known genotype and phenotype information to predict genomic breeding value of a related test population with available genotype information. The training population is used to estimate the model parameters, quantifying the association between the phenotypes and genotypes. Once the model is developed, those parameters are used to calculate GEBVs for test population with only known genotype information. Ideally, the prediction accuracy should be evaluated by comparing the GEBVs to true breeding values (TBVs). However, in practice TBVs are not available, thus phenotype scores are used as a surrogate to assess the accuracy of prediction. This evaluation process is important when developing and assessing a GP model.

Compared to conventional phenotype-based breeding approaches, GP may be less costly and more time effective than the traditional phenotype-based breeding approaches. Secondly, GP can be conducted in the early stages of plant development such as on the seed and/or early segregating generations of breeding populations so that a population can be enriched for desired plant characteristics. This process has the potential to reduce the breeding cycle (Crossa

[1]CSIRO Agriculture & Food, GPO Box 1600, Canberra, ACT 2601, Australia. [2]CSIRO Agriculture & Food, Locked Bag 59, Narrabri, NSW 2390, Australia. Associate editor Lindsey Compton ✉email: Zitong.Li@csiro.au

Received: 31 March 2021 Revised: 5 April 2022 Accepted: 7 April 2022
Published online: 6 May 2022

et al. 2017), resulting in more efficient use of costly phenotyping resources. Finally, unlike marker assisted selection which utilises a few large effect quantitative trait loci (QTL) for trait improvement, GP utilises genome-wide DNA variation. Therefore, it can capture many small QTL effects, which may cumulatively have a large contribution to the phenotype variation. Hence, GP is more appropriate to predict complex agronomic traits controlled by polygenic gene effects (Goddard and Hayes 2007).

Several factors may impact the power of GP to predict phenotype outcomes. These factors include the heritability and genetic architecture of quantitative traits under evaluation, population structure, the quality of phenotyping and genotyping, the density of markers, the size of the training population, the degree of relatedness between the training and test populations, and the statistical models being used to conduct prediction (Zhang et al. 2019).

Introduction of more data into the training population is usually beneficial for prediction accuracies. However, the inclusion of individuals that are unrelated to the test population in the training set may reduce the prediction accuracy (Wolc et al. 2016; Edwards et al. 2019). Hence, a training set optimisation procedure will improve prediction accuracies (Rincent et al. 2012; Akdemir et al. 2015; Berro et al. 2019). Alternatively, the relatedness between the training and test sets could also be considered by including the relationship among the individuals into the statistical model. Pairwise relationship coefficients can be inferred based on the known pedigree, and the corresponding relationship matrix can be incorporated as a random effect in the statistical predictive model. When pedigree information is lacking, unsupervised clustering approaches (Xu and Tian 2015; Pritchard et al. 2000) can be applied to infer population or family structure, and that information can also be incorporated into the models (Heslot and Jannink 2015; Vandenplas et al. 2018).

GP in cotton breeding is largely still under development as there have been limited GP or pedigree-based studies in cotton and largely, exiting studies have concentrated on fibre quality traits. Gapare et al. (2018) conducted a small-scale GP study on 215 historical varieties collected from the CSIRO breeding program. Cross validation (CV) results revealed that a number of prediction methods including genomic best linear unbiased prediction (Endelman 2011) and Bayesian AlphaBeta methods (Meuwissen et al. 2001) could provide promising prediction accuracies (i.e. up to 0.7 in terms of Pearson correlation between GEBVs and phenotypes) for fibre length and strength. The study also highlighted the importance of taking account of environmental factors in the prediction models. Islam et al. (2020) evaluated similar predictive methods as Gapare et al. (2018) on a multiple parental cross population comprising 550 lines with six fibre quality traits measured using CV. The study also yielded high prediction accuracies up to 0.69 for several fibre quality traits. Liu et al. (2020) proposed an alternative strategy using the sequence variations of 474 fibre length genes and their expression data during fibre development to conduct prediction for fibre length. Using a training population of 128 recombinant inbred lines, the prediction accuracy for fibre length was up to 0.83. Another relevant study by Pérez et al. (2015) used pedigree-based relationship matrix as a basis to predict yield using multiple environmental trials. Although it is not a GP study, it used pedigree data in a genotype or gene-environmental interaction model to achieve prediction accuracy of around 0.5.

The objective of this study is to build on previous cotton GP research to enhance the capability of a cotton GP model using fibre quality properties (fibre length, strength, SFI, elongation, micronaire, and uniformity), lint percentage and lint yield from 1385 cotton breeding lines collected from the CSIRO cotton breeding program. The study hypothesised that: (1) prediction accuracies will be improved through the combination of genomic and pedigree information in cotton GP models. Prediction analyses were conducted using parametric regression methods including Bayesian genomic best linear unbiased predictor (BG-BLUP), Bayesian LASSO and Bayes C, as well as combining these three models with a random effect to account for pedigree. In addition, a non-parametric Bayesian additive regression tree (BART) approach (Chipman et al. 2010; Waldmann 2016) was also applied to our data sets. All the models include year and trial information as covariates to account for the environmental effects. The performance of these models was evaluated in three prediction scenarios. In scenario 1, CV was adopted to randomly divide samples into multiple parts, and then used one part of the data in turn as the test population, and the rest as the training population. The scenario 2 used the latest 2017 data as the test population, and the previous data as the training, to mimic predictions based on unknown genotypes in unknown environments. The scenario 3 considered nine separate biparental families collected in 2017 as the test population, and evaluated whether using the whole training set, or only a subset of the training data that is relevant to the test set could lead to better prediction. This study is important as adopting novel approaches to improve prediction accuracies is essential for deploying GP models in a commercial breeding program. Only once accuracies reach a sufficient level (e.g. equivalent or better than phenotype selection) can an increase in the rate of genetic gain predicted by GP be realised.

MATERIALS AND METHODS

Phenotype data and analysis

The phenotype data used in this study included lint yield (LY; kg ha^{-1}), lint percent (LP; percentage of lint of seed cotton, %), and the fibre quality parameters of fibre length (LEN; upper half-mean length of sample), uniformity (UNI; the ratio of the mean fibre length to the upper half-mean length, expressed as %), short fibre index (SFI; the proportion by weight of fibre shorter than 12.7 mm), strength (STR; the force required at the breaking point for a bundle of fibres of a given weight and fineness, g tex^{-1}), elongation (EL; the extension ability of a bundle of fibres up to its breaking point, expressed as a % increase over its original length), and micronaire (MIC, a measure of air permeability of compressed fibre samples, which is a composite indication of fibre linear density and maturity, unitless).

All phenotype data were collected from experiments conducted under fully irrigated conditions at the CSIRO cotton breeding program's core research base at the Australian Cotton Research Institute (ACRI, 30° 12′S, 149°36′E) located at Myall Vale, Narrabri NSW, Australia. The climate at Myall Vale is semi-arid, characterised by mild winters, hot summers and summer-dominant rainfall patterns, with an annual average precipitation of 646 mm (Aust. BOM 2018). The soil of the site is a uniform grey cracking clay (USDA soil taxonomy: Typic Haplustert; Australian soil taxonomy: Grey Vertosol). Plant available soil water to 1.2 m at the site is between 160 and 180 mm (Tennakoon and Hulugalle 2006).

Experiments were laid out in row-column designs with four replicates, generated from CycDesigN software (VSN International, Hemel Hempstead, UK). Each plot consisted of three 10–12 m rows of cotton (depending on the individual experiment). A row spacing of 1 m was used with a planting density of about 10–12 plants m^{-2}. Management for all field experiments followed then or current high-input commercial practices, e.g. fully irrigated conditions with careful weed and insect control. Plots were furrow irrigated every 10–14 days (approximately 1 ML ha^{-1} applied at each irrigation) from December through to March, according to crop requirements. Each experiment was managed according to its individual requirements for irrigation, weed and pest control, with all plots receiving the same management regime. At approximately 60% open bolls, crops were defoliated with thidiazuron, and mature un-opened bolls were opened with ethephon. A second application of thidiazuron and ethephon was applied 7–10 days later.

Phenotype data were collected from 1385 lines across 42 experiments conducted between 1993 and 2017 (Table 1). Lines were predominantly at the F_4 to F_6 generation, depending on the initial self-generation of breeding families used for single plant selection and then derived breeding lines are tested in the stage-by-stage performance test trials (Liu and Constable 2017). The lines were both conventional and genetically modified (GM); including a mix of released cultivars as well as breeding

Table 1. Details of the lines studied.

Year	No. Lines		No. Biparental crosses	No. Experiments	Notes
	Conventional	Transgenic			
1993/2013	215	–	87	21	37 released cultivars, 178 breeding lines
2014/2015	171	–	11	3	Preliminary to advanced stage material
2015/2016	244	145[a]	37	7	Preliminary to advanced stage material
2016/2017	216	60[b]	15	5	Preliminary to advanced stage material
2017/2018	274	60[b]	17	6	Preliminary to advanced stage material

[a]Transgenic lines with B3F traits.
[b]Transgenic lines with B3XFlex traits.

lines undergoing different stage performance testing. Most of the 1385 lines (~85%) were phenotyped post-2014, and 215 lines included previously published phenotype data (Gapare et al. 2018). Note that some of the 215 lines collected pre-2014 were phenotyped in multiple years and experiments, and all the lines collected in or after 2014 were only phenotyped once. In total, this results in 1907 phenotype observations.

At harvest, seed cotton was mechanically harvested from the middle row of each plot with a spindle picker (modified Case International 1822) and weighed. The outside rows were not harvested and acted as buffers to minimise the edge effect and inter-plot competition. LP was determined from a 300 g sub-sample of the seed cotton that was ginned in a 20 saw gin with a pre-cleaner (Continental Eagle, Prattville, AL U.S.A.), and was subsequently used to calculate lint yield (kg ha^{-1}). Lint samples were collected and tested for fibre quality using a Spinlab High Volume Instrument (HVI) model 1000 (Uster Technologies AG, Uster, Switzerland).

Individual experiment's phenotype data were analysed using linear mixed models taking account of dimensional spatial variation, see Liu et al. (2015) for details. Briefly, the model is described as [Eq. 1]:

$$y = X\tau + Z_b\beta_b + Z_t\beta_r + Z_c\beta_c + \epsilon, \quad (1)$$

where $\boldsymbol{y}$ is a vector of plot observation and τ represents a vector of fixed genotypic (i.e. line) effects, $\boldsymbol{\beta}_b$ is a vector of random effects of replicates (i.e. complete block), $\boldsymbol{X}_\tau$ and $\boldsymbol{Z}_b$ are the corresponding design matrices. $\boldsymbol{\beta}_r$ and $\boldsymbol{\beta}_c$ are the vectors of random effects for rows and columns of the experiment with their corresponding design matrices of $\boldsymbol{Z}_r$ and $\boldsymbol{Z}_c$. Finally, $\boldsymbol{\varepsilon}$ is a vector of plot errors. Plot errors in the model are assumed to be autocorrelated along experiment dimensions, i.e. row and column, and modelled by the first order separable autoregressive process (AR1) covariance model. The best linear unbiased estimates of test lines from individual trial analysis were pooled together and used as the phenotype value in the GP analyses.

Genotype data

DNA isolation and SNP genotyping and calling were performed as per Gapare et al. (2018). Leaves from 10–12 plants from each line were combined for DNA extraction using the DNeasy PlantMini Kit (Qiagen) according to the manufacturer's instructions. All DNA samples were quantified using a NanoDrop 1000 (Thermo Scientific) and normalised to the same concentration (Zhu et al. 2016). DNA at 50 ng/μL for each of lines was processed according to Illumina protocols and hybridised to the CottonSNP63K array at CSIRO Agriculture and Food (Brisbane, Australia) according to the manufacturer's instructions. Chips were scanned using the Illumina iScan and analysed using the GenomeStudio Genotyping Module (v2.0, Illumina). Genotype calls for each SNP were performed based on the cluster file generated specifically for the CottonSNP63K array (Hulse-Kemp et al. 2015). The SNP calling was performed as for a diploid species so at each locus there are three possible genotypes—AA, AB, and BB. Filtering was performed to return polymorphic SNPs with call rate above 85% and minor allele frequency higher than 2.5%. A set of 12296 polymorphic SNPs were used for model training and GPs. These SNPs were distributed across all the 26 chromosomes of cotton with a density of 6 SNPs/Mbp. The missing genotype values at each marker were imputed based on known genotypes at its flanking markers using a Hidden Markov Chain model (Browning and Browning 2007), implemented in the R package "Synbreed" (Wimmer et al. 2012). Principal component analysis (PCA) was conducted on the genotype data using our own R code to investigate the genetic diversity among the samples.

The BG-BLUP model

The linear mixed effect (LMM) model can be specified as [Eqs. 2–4]

$$y_i = \beta_0 + u_i + W\alpha + Z\gamma + e_i, \quad (2)$$

where y_i is the phenotype record of the ith individual ($i = 1,\ldots, n$; n is the total number of individuals), β_0 is the model intercept, e_i is the residual error: $e = [e_1, \ldots, e_n] \sim N(0, I\sigma_e^2)$ (mutually independent for $i = 1,\ldots,n$), σ_e^2 is the residual variance, u_i is the random effect of SNP markers which follows a normal distribution Eq. 3

$$\mathbf{u} = [u_1, \ldots, u_n] \sim N\left(0, \sigma_g^2 G\right), \quad (3)$$

where σ_g^2 is the additive genetic variance, **G** is the genomic relationship matrix (GRM) Eq. 4 (referred as Method 1 in Van Raden 2008) estimated by

$$G_{ik} = \frac{1}{p}\sum_{j=1}^{p}\frac{(x_{ij} - 2p_j)(x_{kj} - 2p_j)}{2p_j(1 - p_j)}, \quad (4)$$

where x_{ij} is the genotype value of the ith individual and jth SNP, coded as −1, 0, and 1 for the three genotypes AA, AB and BB, respectively, p_j represents the minor allele frequency of the jth marker. In the Eq. (2), W and Z are the design matrices of experiments and years (Table S1) from which those individuals are collected, and α and γ are the corresponding random effects of experiments and years, with both following normal distributions: $\alpha \sim N(0, I\sigma_\alpha^2)$ and $\gamma \sim N(0, I\sigma_\gamma^2)$.

In Frequentist statistics, the model parameters including the variance components σ_0^2 and σ_g^2 could be estimated by the restricted maximum likelihood algorithm (Harville 1977). Alternatively, the LMM could be formulated into a Bayesian posterior model as [Eq. 5–6]

$$\begin{aligned} P(\beta_0, \mathrm{u}, \alpha, \gamma, \sigma_e^2|y) = {} & P(\boldsymbol{y}|\beta_0, \mathrm{u}, \alpha, \gamma, \sigma_e^2)P(\beta_0) \\ & \times P(\mathrm{u})P(\alpha)P(\gamma)P(\sigma_e^2) \end{aligned} \quad (5)$$

where $p(\boldsymbol{y}|\beta, \sigma_0^2)$ is the likelihood, specifying the Eq. (2) in the probability form:

$$P(\boldsymbol{y}|\beta_0, \mathrm{u}, \alpha, \gamma, \sigma_e^2) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\alpha_0^2}}\exp\left(-\frac{(y_i - \beta_0 - u_i - W\alpha - Z\gamma)^2}{2\sigma_e^2}\right), \quad (6)$$

and $P(\beta_0, \mathrm{u}, \sigma_e^2) = P(\beta_0)P(\mathrm{u})P(\sigma_e^2)P(\alpha)P(\gamma)$ is the prior distribution of the parameters. The prior of the random effect **u**: $P(\mathbf{u})$ is as specified in (3), where the variance component σ_g^2 together with the residual variance σ_e^2 are further assigned with Scaled inverse chi-squared distribution hyper priors, as suggested in Pérez and de los Campos (2014, File S1). Briefly, the degree of freedom (df) parameter is set to be 5, and the scale parameter is specified as $S_0 = \mathrm{var(y)} \times (1 - R^2) \times (\mathrm{df} + 2)$ and $S_g = \mathrm{var(y)} \times R^2 \times (\mathrm{df} + 2)/\mathrm{mean(diag(G))}$ for σ_e^2 and σ_g^2, respectively, where R^2 is specified as 0.5. These settings correspond to the assumption that 50% of the phenotype variance is explained by the genomic variance component, which is suggested by Pérez and de los Campos (2014). Based on our experiment (results not shown), using alternative R^2 values would not lead to any drastic change in heritability estimation and the GP results.

On the basis of the variance components estimated from the Eq. (2), the genomic heritability of a trait can be calculated

$$h^2 = \frac{\widehat{\sigma_g^2}}{\widehat{\sigma_g^2} + \widehat{\sigma_e^2} + \widehat{\sigma_a^2} + \widehat{\sigma_\gamma^2}}$$

This is a Bayesian alternative to the popular GCTA approach (Yang et al. 2011) for estimating genomic heritability of quantitative traits.

The BG-BLUP combined with pedigree

To incorporate the pedigree information, the BG-BLUP model (2) can be extended as [Eq. 7–8]

$$y_i = \beta_0 + u_i + v_i + W\alpha + Z\gamma + e_i, \tag{7}$$

where all terms are defined in the same way as in Eq. (2), except that v_i is a newly added random effect that has the covariance structure representing the pedigree-based relationship matrix **A**, so that the prior of v_i is specified as

$$P(\mathbf{v}) \sim N(0, \sigma_a^2 A) \tag{8}$$

The pedigree-based matrix **A** was calculated based on the genealogical information of the lines in the study tracing back to five generations (Fig. S1), using the R package "pedigree" (Coster 2015). Like the variance component σ_g^2, the variance component σ_a^2 can also be assigned with a scaled inverse chi-squared distribution prior (Pérez and de los Campos 2014). Now, the scale parameters for variance components are σ_e^2, σ_g^2 and σ_a^2 that are specified as $S_0 = \mathrm{var(y)} \times \left(1 - R_a^2 - R_g^2\right) \times (\mathrm{df} + 2)$, $S_g = \mathrm{var(y)} \times R_g^2 \times (\mathrm{df} + 2)/\mathrm{mean(diag(G))}$ and $S_a = \mathrm{var(y)} \times R_a^2 \times (\mathrm{df} + 2)/\mathrm{mean(diag(A))}$, respectively, where $R_g^2 = 0.4$ and $R_a^2 = 0.1$, corresponding to the model assumption that 40% of phenotype variance explained by genetics, and 10% explained by the pedigree information, based on the assumption that the genetic relationship would capture more phenotypic variation than the pedigree relationship (Fraimout et al. 2021). In addition, based on our numerical experiment (results not shown), changing those prior values would not lead to drastic change in the prediction results.

Bayesian LASSO

Additional to the BG-BLUP model building the relation between phenotypes and GRM, a multiple locus model could also be applied directly to evaluate the association between phenotypes and different markers as

$$y_i = \beta_0 + \sum_j^p x_{ij}\beta_j + W\alpha + Z\gamma + e_i,$$

or alternatively, in a likelihood form [Eq. 9]:

$$P(y|\beta, \sigma_e^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left(-\frac{\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j - W\alpha - Z\gamma\right)^2}{2\sigma_e^2}\right), \tag{9}$$

where β_j ($j = 1,\ldots, p$) is the regression coefficient representing the additive genetic effect of the marker j, and the other symbols including y_i, x_{ij}, β_0 and σ_e^2 are defined the same way as in Eq. (2). Given the fact that the number of SNPs is larger (i.e. $p > n$), it is essential to keep only the SNPs with non-negligible effects, and to exclude SNPs with small effects out of the model. In Bayesian statistics, this can be achieved by using shrinkage prior λ on regression parameters (O'Hara and Sillanpää 2009).

The regression parameter β_j (j=1,…, p) is the additive genetic effect of marker j, which is assumed to follow a double exponential (DE) prior distribution [Eq. 10]:

$$P(\beta_j) = \lambda \exp(-\lambda|\beta_j|) \tag{10}$$

The DE distribution has a heavier tail than the normal distribution, and it can shrink the effects of unimportant markers towards zero. The parameter λ determines the degree of shrinkage, i.e. how many markers will be excluded from the model. The DE distribution could be further represented as a scale mixture distribution of a normal distribution of β and an exponential distribution, which inspires the following hierarchical prior setting of the regression parameter β_j in Bayesian LASSO [Eq. 11–12]:

$$P(\beta_j) = N\left(\beta_j|0, \sigma_j^2\right), \tag{11}$$

$$P\left(\sigma_j^2\right) = \mathrm{Exp}\left(\sigma_j^2 \middle| \frac{\lambda^2}{2}\right) \tag{12}$$

The shrinkage factor λ^2 is further assigned with a hyper prior of a Gamma distribution Gamma (λ^2|s, r), and so it can be estimated as other model parameters. The shape and rate parameters of the Gamma prior was specified to s = 1.1 and $r = \frac{(s-1)}{2\times(1-R^2)/R^2\times MSx}$ (Pérez and de los Campos 2014), where MSx represents the sum of the variances of genotype values of each SNP, and $R^2 = 0.5$.

Bayes C

Another popular way to achieve shrinkage estimation is to assign a spike and slab prior (Ishwaran and Rao 2005) to the regression parameters as follows [Eq. 13]:

$$P(\beta_j|\lambda_j) \propto (1 - \gamma_j)\mathrm{I}_{(\beta_j=0)} + \gamma_j \mathrm{N}\left(\beta_j|0, \sigma_b^2\right), \tag{13}$$

where γ_j is a binary indicator variable to tell whether the genetic effect of SNP j should be non-negligible and follow a normal distribution, or whether the effect is small and assigned with a zero value. In formula (13), the indicator variable γ_j and the variance component σ_b^2 are further assigned with priors of Bernoulli: $Bern(\gamma_j|\pi)$ and Inverse chi-squared: $IG(\sigma_b^2|df, S_0)$, respectively. In the Inverse gamma prior $IG(\sigma_b^2|df, S_0)$, the parameters $\mathrm{df} = 5$ and $S_0 = \mathrm{var}(y) \times R^2 \times (df + 2)/MSx$, with $R^2 = 0.5$. In the Bernoulli prior Bern($\gamma_j|\pi$), the parameter π was further assigned with a Beta prior Beta($\pi \mid p_0, \pi_0$), with $p_0 = 50$, and $\pi_0 = 0.5$. The spike and slab prior (13) is often referred as the Bayes C model (Habier et al. 2011) in the GP literature.

Adding pedigree into the Bayesian LASSO and Bayes C model

To account for the pedigree information, a random effect **u** could be further added into the likelihood (9) as [Eq. 14]

$$P(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\sigma_e^2}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j - W\alpha - Z\gamma - u_i\right)^2}{2\sigma_e^2}\right), \tag{14}$$

where the random effect **u** following a normal prior as:

$$\boldsymbol{u} \sim N(0, \sigma_a^2 \mathbf{A}),$$

where A is the pedigree-based relationship matrix. The prior information for all other parameters in (14) could be assigned in the exact same way as in Bayesian LASSO or Bayes C.

Thus, both the Bayesian LASSO (or Bayes C, which has a different prior setting for marker effects) and BG-BLUP model utilise the same covariance matrix **A** constructed from the pedigree analysis, to account for the family structure. However, the two model classes use different way to model the dependency between genotype and phenotype data. Bayesian LASSO or Bayes C used a multiple locus model with a shrinkage prior to estimate the additive effects of SNPs. The BG-BLUP model used the SNP data to estimate the GRM to study the genotype-phenotype association.

The posterior distribution of all the three models can be evaluated using the Markov Chain Monte Carlo (MCMC) algorithm, in particular the Gibbs sampling, as presented in de los Campos et al. (2009) and Pérez and de los Campos. (2014). Practically, the MCMC was implemented by the R package BGLR (Pérez and de los Campos. 2014). The algorithm generated 50,000 posterior samples, with the first 10,000 samples as burn-in, and every 20th of the rest were stored to reduce the serial correlation.

Bayesian additive regression tree

Bayesian additive regression tree (BART) (Chipman et al. 2010; Hill et al. 2020) is a more flexible model structure compared to the linear regression, presented as [Eq. 15]

$$\mathrm{y_i} = \mathrm{f}(\mathrm{x_i}) + e_i, \tag{15}$$

where f represents a summation of many regression trees as

$$f(x) = \sum_{k=1}^{m} g(x, T_k, M_k),$$

where T_k represents a binary decision tree with a set of terminal nodes and interior decision rules, $\mu_{kl} \in M_k (l = 1, \ldots, b_k)$ are a set of parameter values associated with each terminal nodes of T_k, and e_i is the Gaussian residual error as in (2).The decision rules associated with T_k are binary splits of the genotypes **x** to $(\boldsymbol{x} \in A_{kl})$ and $(\boldsymbol{x} \notin A_{kl})$, where A is a subset of **x**. The function g then assigns each μ_{kl} to **x** as

$$g(\boldsymbol{x}, T_k, M_k) = \mu_{kl} \text{ if } \boldsymbol{x} \in A_{kl}$$

In BART, a regularised prior is assigned to each tree T_k and its terminal notes M_k, and assumed independence between different components:

$$\begin{aligned} p((T_1, M_1), (T_2, M_2), \ldots, (T_m, M_m)) &= \prod_{k=1}^{m} p(T_k, M_k) \\ &= \prod_{k=1}^{m} p(T_k) p(M_k | T_k) \\ &= \prod_{k=1}^{m} p(T_k) \prod_{l}^{b_k} (\mu_{kl} | T_k), \end{aligned}$$

where the prior $p(T_j)$ consists of three parts: (i) the probability that a node at the depth d is non-terminal given by $\alpha(1+d)^{-\beta}$ where $\boldsymbol{\alpha} = 0.95$, and $\beta = 2$ as suggested in Chipman et al. (2010) and (ii) a uniform distribution specified for the variables x_{ij} (j=1,…,p) which are assigned at each interior node for splitting, and (iii) a uniform distribution on the splitting rule assignment in each interior node. The $p(\mu_{kl}|T_k)$ is a normal distribution $N(0, \sigma_\mu^2)$, with the variance fixed to be $\sigma_\mu^2 = \frac{0.5}{k\sqrt{M}}$, and k = 2, and the number of trees M is fixed to be 200 as Waldmann (2016).

The prior tends to generate a lot of small trees with simple structure, and therefore can avoid the over-fitting problem. Compared to the G-BLUP, Bayes LASSO or Bayes C methods, the benefit of BART is that it can implicitly model not only the additive effects, but also the non-additive genetic effects such as dominance effects and gene-gene interaction effects (Waldmann 2016).

The BART model could be evaluated using Gibbs sampling with a few Metropolis Hasting sampling steps, which can be implemented using the R package BayesTree (Chipman and McCulloch 2016). Here we generated equivalent amount of MCMC samples as for the other three models.

Assessment of prediction accuracy

Three different scenarios were considered to evaluate the prediction accuracy of different approaches. *In Scenario 1*, a fivefold-CV and an additional 50-fold-CV were used by randomly dividing the samples into multiple parts (i.e. either 5 or 50) with equivalent sizes, and in turn using each part of the data (of 277 lines) as the test population, and the rest (of 1108 lines) as the training data. In fivefold-CV, the training and test set comprise 1108 and 277 lines, respectively. And in 50-fold-CV, the training and test sets comprise 1357 and 28 lines, respectively. In CV, the average predictive performance over different folds is considered as prediction accuracy. *In Scenario 2*, all the lines phenotyped at seasons 1993–2016 were used as the training population to build the model. The lines phenotyped in the most recent (2017/2018) season were used as the test population to calculate the prediction accuracy. This scenario reflects a likely approach of utilising GP in a breeding program where samples collected in previous years are used as the training data to generate GEBVs from the most recent season. *Scenario 3* used the same training population as in Scenario 1, but it focused on predicting separately on each single biparental family collected at 2017/2018 (Schopp et al. 2017; Brauner et al. 2020). For the training data set, we considered either using the whole set of samples up to 2016/2017, or only a subset of samples that are closely relevant to each test population defining by the relationship coefficients between samples calculated based on pedigree. We used relationship coefficients thresholds of both 0.125 and 0.25, representing the first cousin (i.e. sharing one grandparent) and half-sibling relations (i.e. sharing one parent), respectively. This approach explored whether an appropriate training population existed for single biparental family in terms of attaining a better prediction accuracy.

In all these three scenarios, the Pearson correlation coefficient between the GEBV and the phenotypes was considered as the measurement of prediction accuracy.

RESULTS

In this study, we conducted GP analyses on a total of 1385 lines using eight statistical methods. The methods were BG-BLUP, Bayesian LASSO, Bayes C and a non-parametric BART, used either only genomic data or combined genomic data with pedigree information. The prediction accuracies were then evaluated in three different scenarios.

Phenotype and genomic variation

A summary of phenotype variation including mean, standard error, minimum value, maximum value of phenotypes for each trait was given in Table 2. Genomic heritabilities estimated by BG-BLUP were 0.59 for LEN, 0.35 for UNI, 0.30 for SFI, 0.59 for STR, 0.46 for EL, 0.42 for MIC, 0.26 for LY, and 0.41 for LP (Table 2).

Principal component analysis (PCA) results (Fig. S2) of genomic data revealed no clear separation between the samples collected in different years.

Prediction scenario 1 (Cross validation)

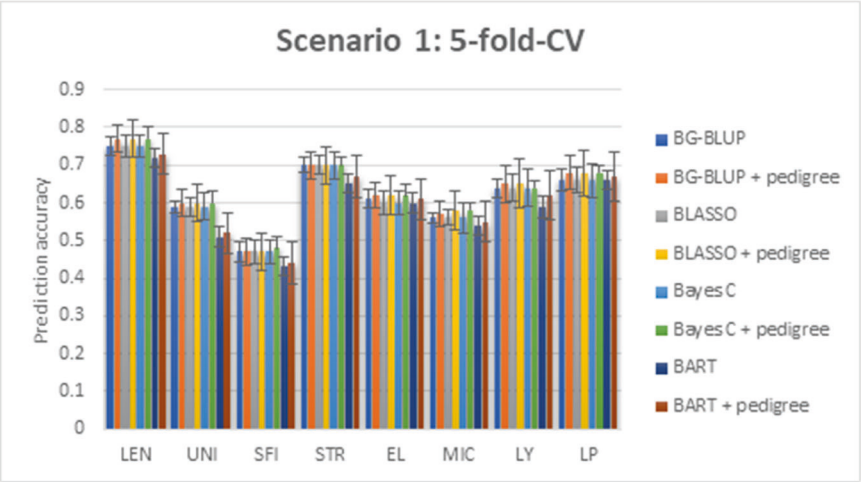
When using the fivefold-CV to evaluate the predictability of the models, the accuracies were 0.72–0.77 for LEN, 0.47–0.60 for UNI, 0.43–0.48 for SFI, 0.65–0.71 for STR, 0.60–0.62 for EL, 0.50–0.58 for MIC, 0.59–0.65 for LY, and 0.66–0.68 for LP (Fig. 1a; Table S2). The performance of BG-BLUP, Bayesian LASSO, Bayes C was comparable to each other. BART's performance was worse than the other three methods for UNI, but was comparable to other traits. Inclusion of pedigree information in all the methods improved the prediction accuracies by 1–3% for different traits (Table S2).

The 50-fold-CV's accuracies were 0.78–0.80 for LEN, 0.49–0.60 for UNI, 0.51–0.53 for SFI, 0.67–0.71 for STR, 0.55–0.60 for EL, 0.65–0.67 for MIC, 0.58–0.6 for LY, and 0.59–0.64 for LP. Similar as

Table 2. Summary of phenotype variation across traits including the mean, standard error, minimum value, and maximum value of original phenotypes and the adjusted phenotypes by checks in each trial to reduce the phenotype variation caused by management and environment.

	LEN	UNI	SFI	STR	EL	MIC	LY	LP
Mean	1.23	84.2	5.54	31.4	14.6	5.2	2683	49.3
SD	0.05	1.21	1.95	1.73	1.74	0.36	525	2.35
Min	1.10	80.6	2.3	26.3	0.9	4.25	1326	32.4
max	1.41	88.5	9.9	49.6	14.6	5.2	3693	49.3
Mean_adjust	1.03	1.00	0.94	1.01	0.96	0.99	1.00	0.99
SD_adjust	0.04	0.01	0.16	0.05	0.10	0.06	0.12	0.04
Min_adjust	0.93	0.96	0.35	0.88	0.35	0.70	0.44	0.76
Max_adjust	1.20	1.05	1.7	1.69	1.38	1.17	1.63	1.12
Genomic H^2	0.59	0.35	0.30	0.59	0.46	0.42	0.26	0.41
95% Credible intervals for H^2	(0.53, 0.66)	(0.28, 0.41)	(0.22, 0.37)	(0.53, 0.65)	(0.39, 0.53)	(0.35, 0.49)	(0.19, 0.33)	(0.29, 0.52)

(a)



(b)

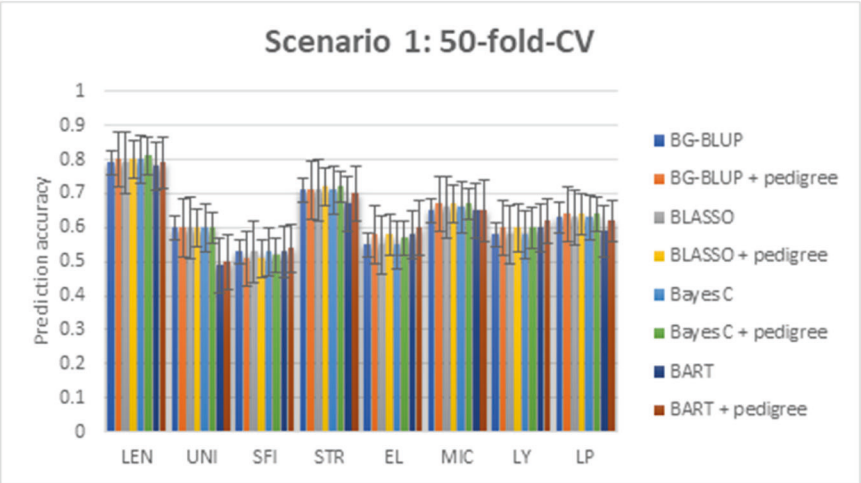


Fig. 1 The prediction accuracies and standard errors of scenario 1. (**a** Fivefold cross validation; and **b** 50-fold cross validation). Methods under evaluation were Bayesian genomic best linear unbiased predictor (BG-BLUP), Bayesian LASSO, Bayes C, Bayesian additive regression tree (BART), and these four models further adding pedigree or structure information as random effects. Traits being analysed included fibre length (LEN), uniformity (UNI), short fibre index (SFI), fibre strength (STR), fibre elongation (EL), fibre micronaire (MIC), lint yield (LY) and lint percentage (LP).

in the fivefold-CV, the inclusion of pedigree in all the Bayesian approaches slightly improved the prediction accuracies (Fig. 1b; Table S3).

Prediction scenario 2

By using all samples sourced prior to the 2017/18 (1051 lines) as the training population to build the model and the 2017/18 season samples (334 new lines) to evaluate the prediction, the prediction accuracies of the eight methods ranged 0.39–0.42 for LEN, 0.08–0.14 for UNI, 0.14–0.20 for SFI, 0.35–0.38 for STR, 0.41–0.48 for EL, 0.19–0.28 for MIC, 0.14–0.25 for LY, 0.30–0.36 for LP (Fig. 2; Table S4). BG-BLUP and Bayesian LASSO, Bayes C and BART achieved almost identical accuracies for LEN, STR and LP (Table S4), though they are formulated under different model assumptions. The BART method gave higher prediction accuracies for EL, MIC and LY. Bayesian BLUP performed better for UNI. Bayes LASSO performed better for SFI. Adding pedigree information as a random effect to BG-BLUP, Bayesian LASSO, Bayes C or BART, resulted in the highest accuracies for SFI, STR and LP, but had compromised the accuracies of EL and MIC prediction; however, all these changes were small in magnitude (Table S4).

Prediction scenario 3

Here nine biparental families collected from the year 2017/18 were treated as separate test populations to evaluate whether using the families which are closely related to the test population in the training population could improve the prediction.

The prediction accuracies for the Bayesian LASSO based on the whole training population (average over the 9 biparental families) were 0.23 for LEN, 0.25 for UNI, 0.14 for SFI, 0.4 for STR, 0.3 for EL, 0.22 for MIC, 0.13 for LY, and 0.38 for LP (Fig. 3; Table S5). Adding pedigree to Bayesian LASSO produced better prediction accuracies for all of the traits except EL. Using only the families that have relationship coefficients at least 0.125 (i.e. equivalent as sharing one common grandparent) with the test population in the training population yielded best prediction accuracies for MIC and LP (Table S5). Using the families that have relationship coefficients at least 0.25 (i.e. as sharing one parent) produced best prediction

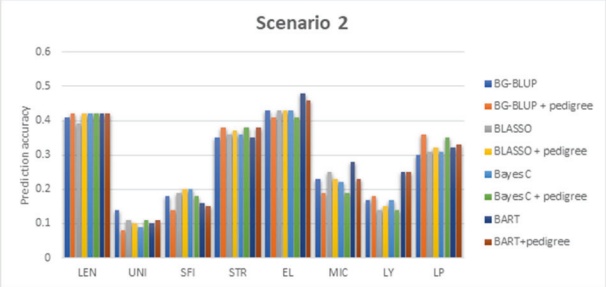


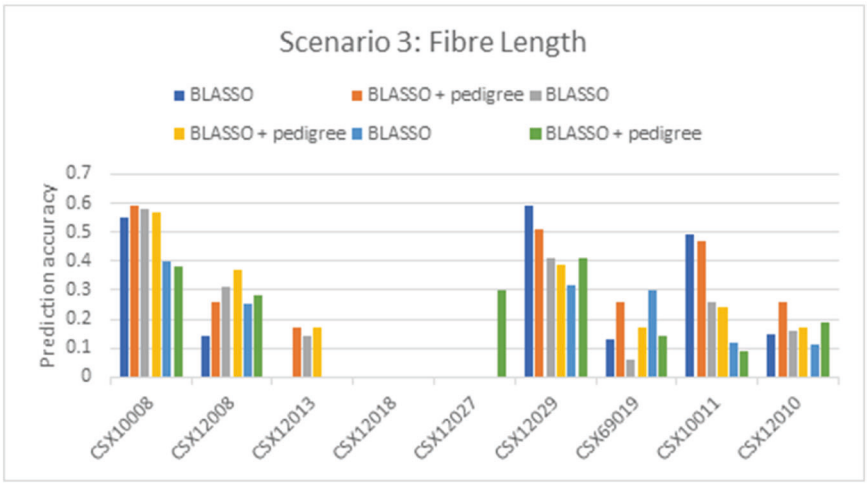
Fig. 2 The prediction accuracies of the scenario 2: the lines phenotyped at seasons 1993–2016 were used as the training population, and the data collected in the 2017/2018 season were used as the test population. Methods under evaluation were Bayesian G-BLUP, Bayesian LASSO, Bayes C, BART, and these three models further adding pedigree or structure information as random effects, and BART. Traits being analysed included fibre length (LEN), uniformity (UNI), short fibre index (SFI), fibre strength (STR), fibre elongation (EL), fibre micronaire (MIC), lint yield (LY) and lint percentage (LP).

accuracies for UNI, SFI, STR, and LY. The whole training data set performed best for LEN and EL.

DISCUSSION

In this paper, we presented a GP study on both fibre qualities and yield traits of cotton (*G. hirsutum*) with 1385 samples collected from the CSIRO cotton breeding program. This study is on a much larger scale than the previous two GP studies on the same species (Gapare et al. 2018; Islam et al. 2020). From a statistical modelling perspective, we focused on evaluating several Bayesian regression methods including BG-BLUP, Bayesian LASSO, Bayes C, and BART. In all the methods, year and experiment information were added as covariates to account for the environmental effects. The former three methods have already been widely used in the GP literature (Crossa et al. 2017; Wang et al. 2018), while the BART approach is less well known in the field (Waldmann 2016). Since the trait data were collected from multiple biparental families, we calculated the relationship matrix based on pedigree information, and used that matrix in conjunction with the genomic data in the model and observed an improvement in prediction accuracies.

(a)



(b)

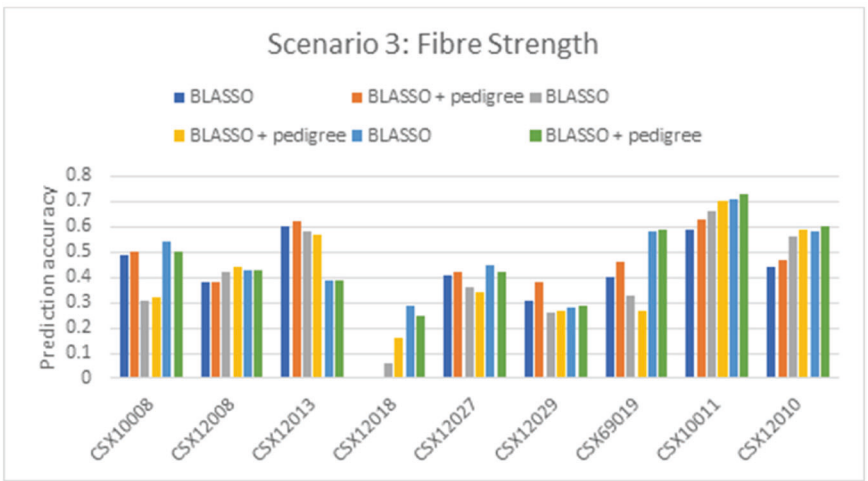


Fig. 3 The prediction accuracies of the scenario 3. a fibre length (LEN), and **b** strength (STR). This approach used each biparental family from season 2017/2018 as the separate test population. The training population was either all the lines phenotyped before 2017, or the families phenotyped before 2017 which are closely relevant to the target population (i.e. the related coefficient no < 0.125 or 0.25).

Prediction accuracies over three scenarios

Scenario 1 aims to evaluate the predictive ability of different predictive approaches by using cross-validation (CV), which are also used in many other GP studies (Runcie and Cheng 2019). In contrast, Scenario 2 reflects a likely approach when utilising GP in a breeding program, where samples collected in the past are used as the training population to generate GEBVs for the new samples from the most recent season. The prediction accuracies for traits in Scenario 1 were significantly higher than those in Scenario 2, which can be explained by how the training and test population being defined. In Scenario 2, the training and test data comprised samples from different biparental families generated in different years, and additionally their phenotypes were subject to different environments (e.g. climate variability across testing seasons). But in Scenario 1, the training and test populations were randomly defined so both have samples collected from the same families and phenotypes measured at the same environments, resulting in more accurate predictions. Our results are aligned with results from existing literature shown that the prediction accuracies of new genotypes in new environments (i.e. Scenario 2 were considerably lower than the prediction by using CV (i.e. randomly defining the training and test population, i.e. Scenario 1) (Jarquín et al. 2014b; Gillberg et al. 2019). Another observation was that the prediction accuracies in Scenarios 1 and 2 were both highly correlated with the square root of the genomic heritabilities across traits. For example, the Pearson correlation coefficients between prediction accuracies by GB-BLUP in Scenario 1 (50-fold-CV) and Scenario 2 and the squared root of heritabilities are 0.80 and 0.82, respectively. Note that the square root of the heritability of a trait was considered as the theoretically expected value of the prediction accuracies could be achieved by GP (Estaghvirou et al. 2013).

The Scenario 3 is related to Scenario 2 but with a focus on calculating GEBVs for within family samples instead of between family samples. Results highlight that the prediction accuracies vary dramatically across different biparental families. This may reflect the complexity of the relationship structure among those families. The pedigree-based relationship matrix had two roles here. First, it was used to tune the training population to select training samples that are closely relevant to the target family. Second, it was also used as a random effect in Bayesian regression models for the predictive analysis. Conducting a training set selection by using samples that are closely relevant to the target population showed an improvement of prediction accuracies for most of the traits, but the optimal threshold to determine the level of relatedness cannot be determined based on our results.

The results of Scenario 3 may be limited by the sample size within each biparental family, i.e. ~20 individuals. Such a limited number of individuals in the test population may result in some alleles which are helpful in explaining the phenotypic variation in the training populations not to be included in the test population due to genetic drift, which may result in some redundant SNPs and reduce the predictive power for some families. Moreover, samples were collected from populations which had undergone significant phenotype-based selection (i.e. truncation). This reduces the phenotypic range of traits, making it difficult to obtain accurate predictions using genomic data (Table S6).

Comparison to previous cotton genomic prediction studies

Gapare et al. (2018) used CV to evaluate prediction accuracies on 215 historical lines (a subset of the data set used in this study). They obtained prediction accuracies of 0.67 and 0.35 for LEN and STR, respectively. The prediction accuracies of these two traits in our study were considerably higher. This may be explained by the larger size of the training population in our study. Another explanation is that unlike Gapare et al. (2018) who used only genomic information to generate predictions, we have also incorporated pedigree into the models. We have shown that pedigree data provides complimentary information to the models. Thus, improved prediction accuracies can be achieved when genomics and pedigree are used in combination, as they are able to jointly describe the relatedness of individuals composited in both the training and predictive population (Velazco et al. 2019). From a breeding point of view, the training populations used in this study are predominantly recently developed CSIRO breeding lines. Given the continuity of germplasm enhancement activities in our breeding program, many elite individuals in these training populations were used as parents for new breeding crosses, from which lines in the testing populations are derived. Therefore, the relatedness between the training and testing populations is higher in this study compared to Gapare et al. (2018)'s study. This higher relatedness between the training and testing populations is likely to have influenced the improved prediction accuracies.

Another study (Islam et al. 2020) also conducted CV on 550 lines of a multiple parental cross (MAGIC) population produced in the US, and obtained maximum accuracies of 0.50 for LEN, 0.48 for UNI, 0.50 for SFI, 0.55 for STR, 0.68 for EL, and 0.35 for MIC (Fig. 2 in Islam et al. 2020). Except EL, all other traits had higher accuracies in our study. However, it is important to note that the study using a MAGIC population may not be comparable to ours due to the nature of the populations studied and the testing environments, as well as dissimilar genotyping and phenotyping methods.

Comparison between predictive models and between CV strategies

The four Bayesian regression methods being considered in this study have different model assumptions. The BG-BLUP model assumes the genetic effects of different markers to follow a normal distribution with a common variance. Bayesian LASSO and Bayes C assume markers to follow a prior distribution with individual variance. Accordingly, BG-BLUP is most suitable to analyse a polygenic trait with all markers having small genetic effects. Alternatively, Bayesian LASSO and Bayesian C work most efficiently for oligogenic traits with small number of markers having major effects, and the rest having small effects (Li and Sillanpää 2012). The fact that these methods have similar predictive performance on our data set may indicate that the traits being analysed here have a polygenic genetic architecture so that the Bayesian LASSO and Bayes C methods do not show advantage over the BG-BLUP model.

The BART model is a non-parametric method similar to other machine learning methods which have been proposed for GP such as random forest and boosting (Li et al. 2018). BART implicitly models the genetic effects in regression decision trees. Compared to the three parametric methods, the benefit of BART is that it can also account for non-additive effects such as dominance effects and gene-gene interaction effects.

BART's performance varies across traits and scenarios. From our analyses in Scenario 1, although BART did not provide the best accuracies among the methods for most of the traits, it indeed outperformed others for EL, MIC and LY, indicating some non-additive effects may influence those traits. However, because it is a non-parameterised method, BART cannot be used to identify the SNPs associated with those non-additive effects.

Furthermore, it must be also highlighted that adding the pedigree-based relationship matrix into the regression models improves predictions in all three scenarios. Using pedigree as a random effect is useful to account for the complex correlation structure between multiple families in the data, which cannot be fully explained by genetic data. This observation supports results published in GP research in other crop species (Velazco et al. 2019).

Another interesting technical perspective is to determine an optimal number of folds in CV (i.e. prediction Scenario 1). Our results showed that the prediction accuracies estimated by 50-fold-CV is systematically higher than the accuracies estimated by

fivefold-CV, although the difference is small in magnitude. The accuracies estimated by both 50-fold-CV and fivefold-CV are all highly correlated with the square root of genomic heritability (as shown above), which indicates both approaches are suitable for evaluating prediction accuracies for GP. One major difference is however that the standard error of prediction accuracies over folds in 50-fold-CV is much larger than that of the fivefold-CV. Moreover, the computation cost of 50-fold-CV is much more expensive. Hence, the use of small number of folds in CV should be a better choice for evaluating a model's predictability.

CONCLUSION

This research presents the first GP study using samples collected from multiple years and locations from a commercial cotton breeding program. It highlights that GP models, particularly when combined with pedigree information, provide significant potential to predict accurate GEBVs (i.e. maximum prediction accuracies of 0.50–0.76, depending on the target trait). However, as no prediction model constantly outperformed all other models across the prediction scenarios and traits presented in this work, it is important to apply a set of different models to new data sets. In addition, it must be acknowledged that the circumstances where GP could be deployed in a commercial breeding program (i.e. Scenarios 2 and 3) the prediction accuracies were not consistently high. We expect that the inclusion of environmental covariates and other 'omics' data may help improve the accuracy of GPs. Particularly when considering complex, polygenic traits where interactions between genes and environment have significant effects on phenotype outcomes. The study can be further extended by including environmental factors such as climate variables into the statistical models; building a genotype (gene)-by-environmental interaction model to conduct prediction analyses (Jarquín et al. 2014b; Crossa et al. 2017; Rogers et al. 2021), and; the extension of prediction models to simultaneously predict multiple correlated traits (Moeinizade et al. 2020).

DATA AVAILABILITY

The genotype and phenotype data as well as the pedigree-based relationship matrix used in this study are available from the CSIRO Data Access Portal https://doi.org/10.25919/k18n-nk98.

REFERENCES

Aust. BOM (2018) Australian Bureau of Meteorology: Climate Data Online. Commonwealth of Australia Bureau of Meteorology Web.

Akdemir D, Sanchez JI, Jannink J-L (2015) Optimization of genomic selection training populations with a genetic algorithm. Genet Selection Evol 47:38

Berro I, Lado B, Rafael SN, Quincke M, Gutiérrez L (2019) Training population optimization for genomic selection. Plant Genome 12:190028

Brauner PC, Müller D, Molenaar WS, Melchinger AE (2020) Genomic prediction with multiple biparental families. Theor Appl Genet 133:133–147

Browning BL, Browning SR (2007) Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. Am J Hum Genet 81:1084–1097

de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E et al. (2009) Predicting Quantitative Traits with Regression Models for Dense Molecular Markers and Pedigree. Genetics 182:375–385

Chipman HA, George EI, McCulloch RE (2010) BART: Bayesian Additive Regressive Trees. Ann Appl Stat 4:266–298

Chipman H, McCulloch R (2016). BayesTree: Bayesian Additive Regression Trees. R package. version 0.3-1.3, https://CRAN.R-project.org/package=BayesTree.

Coster A (2015) R Package 'pedigree'. https://cran.r-project.org/web/packages/pedigree/index.html

Crossa J, Perez-Rodriguez P, Cuevas J, Montesinos-Löpez O, Jarquín D, de los Campos G et al. (2017) Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends Plant Sci 22:961–975

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4:250–255

Edwards SM, Buntjer JB, Jackson R, Bentley AR, Lage J, Byrne E et al. (2019) The effects of training population design on genomic prediction accuracy in wheat. Theor Appl Genet 132:1943–1952

Estaghvirou SBO, Ogutu JO, Schulz-Streeck T, Knaak C, Ouzunova M, Gordillo A, Piepho H-P (2013) Evaluation of approaches for estimating the accuracy of genomic prediction in plant breeding. BMC Genom 14:860

Fraimout A, Li Z, Sillanpää MJ, Rastas P, Merilä J (2021) Dissecting the genetic architecture of quantitative traits using genome-wide identity-by-descent sharing among full-sibs. Molecular Ecology (submitted). https://doi.org/10.1101/2021.03.01.432833v1.full.pdf.

Gapare W, Liu S, Conaty W, Zhu Q-H, Gillepie V, Llewellyn D, Stiller W, Wilson I (2018) Historical datasets support genomic selection models for the prediction of Cotton Fiber Quality Phenotypes Across Multiple Environments. G3 8:1721–1732

Gillberg J, Marttinen P, Mamitsuka H, Kaski S (2019) Modelling G×E with historical weather information improves genomic prediction in new environments. Bioinformatics 35:4045–4052

Goddard ME, Hayes BJ (2007) Genomic selection. Anim Breed Genet 124:323–330

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the Bayesian alphabet for genomic selection. BMC Bioinforma 12:186

Harville DA (1977) Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. J Am Stat Assoc 72:320–338

Helsot N, Jannink JL (2015) An alternative covariance estimator to investigate genetic heterogeneity in populations. Genet Selection Evol 47:93

Hill J, Linero A, Murray J (2020) Bayesian additive regression trees: a review and look forward. Annu Rev Stat Its Application 7:251–278

Hulse-Kemp AM, Lemm J, Plieske J, Ashrafi H, Buyyarapu R, Fang DD et al. (2015) Development of a 63K SNP array for cotton and high-density mapping of intraspecific and interspecific populations of Gossypium spp. G3 5:1187–1209

Islam MS, Fang DD, Jenkins JN, Guo J, McCarty JC, Jones DC (2020) Evaluation of genomic selection methods for predicting fiber quality traits in Upland cotton. Mol Genet Genom 295:67–79

Ishwaran H, Rao JS (2005) Spike and Slab variable selection: frequentist and Bayesian strategies. Ann Stat 33:730–773

Jabran K, Ul-Allah S, Chauhan BS, Bakhsh A (2019) An introduction to global production trends and uses, history and evolution, and genetic and biotechnological improvements in cotton. In: Jabran K, Chauhan BS Eds. Cotton Production, 1st ed. Wiley, Hoboken, NJ, USA, p 1–5

Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genom 9:166–177

Jarquín D, Kyle K, Posadas L, Hyma K, Jelicka J, Graef G, Lorenz A (2014a) Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC Genom 15:740

Jarquín D, Crossa J, Lacaze X, Cheyron PD, Daucourt J, Lorgeu J et al. (2014b) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor Appl Genet 127:595–607

Li B, Zhang N, Wang Y-G, George AW, Reverter, Li Y (2018) Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. Front Genet 9:237

Li Z, Sillanpää MJ (2012) Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. Theor Appl Genet 125:419–435

Liu S, Constable GA, Cullis BR, Stiller WN, Reid PE (2015) Benefit of spatial analysis for furrow irrigated cotton breeding trials. Euphytica 201:253–264

Liu SM, Constable GA (2017) Effect of self-generation for initial selection on breeding better cotton. Euphytica 213:17

Liu Y, Xu Y, Zhang M, Cui Y, Sze S-H, Smith CW, Xu S, Zhang H-B (2020) Accurate prediction of a quantitative trait using the genes controlling the trait for gene-based breeding in cotton. Front Plant Sci 11:583277

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Millet EJ, Kruijer W, Coupel-Ledru A, Prado SA, Cabrera-Bosquet L, Lacube S et al. (2019) Genomic prediction of maize yield across European environmental conditions. Nat Genet 51:952–956

Moeinizade S, Kusmec A, Hu G, Wang L, Schnable PS (2020) Multi-trait Genomic Selection Methods for Crop Improvement. Genetics 4:931–945

O'Hara RB, Sillanpää MJ (2009) A review of Bayesian variable selection methods: what, how and which. Bayesian Anal 4:85–117

Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin DC et al. (2012) Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature 492:423–427

Pérez P, de los Campos G (2014) Genome-Wide Regression and Prediction with the BGLR Statistical Package. Genetics 198:483–495

Pérez P, Crossa J, Bondalapati K, Meyer GD, Pita F, de los Campos (2015) A Pedigree-Based Reaction Norm Model for Prediction of Cotton Yield in Multi-environment Trials. Crop Sci 55:1143–1151

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y et al. (2012) Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. Plant Genome 5:103–113

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Rincent R, Laloë, Nicolas S, Altmann T, Brunel D, Revilla P et al. (2012) Maximizing the Reliability of Genomic Selection by Optimizing the Calibration Set of Reference Individuals: Comparison of Methods in Two Diverse Groups of Maize Inbreds (Zea mays L.). Genetics 192:715–728

Rogers AR, Dunne JC, Romay C, Bohn M, Buckler ES, Ciampitti IA et al. (2021) The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize experiment. G3 11: jkaa050.

Runcie D, Cheng H (2019) Pitfalls and Remedies for Cross Validation with Multi-trait Genomic Prediction Methods. G3 9:3727–3741. G3, Jkaa050

Schopp P, Müller D, Wientjes YCJ, Melchinger AE (2017) Genomic prediction within and across biparental families: means and variances of prediction accuracy and usefulness of deterministic equations. G3 7:3571–3586

Spindel J, Begum H, Deniz A, Virk P, Collard B, Redona E et al. (2015) Genomic selection and association mapping in rice (Oryza sativa): effect of trait Genetic Architecture, Training Population Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Lines. PLoS Genet 11:e1005350

Stiller WN, Wilson IW (2014) Australian Cotton Germplasm Resources, World Cotton Germplasm Resources, edited by Abdurakhmonov I. InTech, Rijeka, Croatia, 10.5772/58414

Tennakoon SB, Hulugalle NR (2006) Impact of crop rotation and minimum tillage on water use efficiency of irrigated cotton in a Vertisol. Irrig Sci 25:45–52

Vandenplas J, Calus MPL, Gorjanc G (2018) Genomic prediction using individual-level data and summary statistics from multiple populations. Genetics 210:53–69

Waldmann P (2016) Genome-wide prediction using Bayesian additive regression trees. Genet Selection Evol 48:42

Wang X, Xu Y, Hu Z, Xu C (2018) Genomic selection methods for crop improvement: current status and prospects. Crop J 6:330–340

Wimmer V, Albrecht T, Auinger H-J, Schön C-C (2012) Synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics 28:2086–2087

Wolc A, Ktanis A, Arango J, Settar P, Fulton JE, O'Sullivan NP et al. (2016) Implementation of genomic selection in poultry industry. Anim Front 6:23–31

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423

Velazco JG, Malosetti M, Hunt CH, Mace ES, Jordan DR, van Eeuwijk FA (2019) Combing pedigree and genomic information to improve prediction quality: an example in sorghum. Theor Appl Genet 132:2055–2067

Xu D, Tian Y (2015) A comprehensive survey of clustering algorithms. Ann Data Sci 2:165–193

Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88:76–82

Zhang H, Yin L, Wang M, Yuan X, Liu X (2019) Factors Affecting the Accuracy of Genomic Selection for Agricultural Economic Traits in Maize, Cattle, and Pig Populations. Front Genet 10:189

Zhu Q-H, Zhang J, Liu D-X, Stiller WN, Liu D-J, Zhang Z-S et al. (2016) Integrated mapping and characterization of the gene underlying the okra leaf trait in Gossypium hirsutum L. J Exp Bot 67:763–774

ACKNOWLEDGEMENTS

This study was financially supported by Cotton Breeding Australia, a joint venture between CSIRO and Cotton Seed Distributors Ltd. We are grateful for the technical expertise of Vanessa Gillespie and Melanie Soliveres as well as the invaluable contribution to this work made by technical staff of the CSIRO cotton breeding group, who undertook the phenotyping and tissue sampling for genotyping. We also thank Russell McCulloch and Dr Bill Barendse for processing and help with the analysis of the Illumina CottonSNP63K assays. The authors thank Dr Klara Verbyla, Dr Shannon Dillon and Dr Hawlader Abdullah Al-Mamun for constructive discussions on GP and reviews of the initial drafts. We also thank for the constructive comments from the associate editor Dr Lindsey Compton and four anonymous reviewers.

AUTHOR CONTRIBUTIONS

ZL and IW conceived the study questions and designed the research. WS, SL and WC developed and collected breeding lines, and conducted phenotype analysis. QZ, IW and PM developed SNPchip genotype data. SL and ZL performed pedigree analysis. ZL performed the genomic prediction analysis. PM performed bioinformatics analysis and pedigree visualisation. ZL, WC and IW wrote the paper with contributions from all other authors.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41437-022-00537-x.

Correspondence and requests for materials should be addressed to Zitong Li.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© Crown 2022